frontiers | Research Topics

Big data and artificial intelligence in ophthalmology - clinical application and future exploration

Edited by
Tae-im Kim, Darren Shu Jeng Ting, Yi-Ting Hsieh and Tyler Hyungtaek Rim

Published in
Frontiers in Medicine

ISSN 1664-8714
ISBN 978-2-8325-4171-5
DOI 10.3389/978-2-8325-4171-5

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Big data and artificial intelligence in ophthalmology - clinical application and future exploration

Topic editors

Tae-im Kim — Yonsei University, Republic of Korea

Darren Shu Jeng Ting — University of Nottingham, United Kingdom

Yi-Ting Hsieh — National Taiwan University Hospital, Taiwan

Tyler Hyungtaek Rim — Mediwhale Inc, Republic of Korea

Citation

Kim, T.-i., Ting, D. S. J., Hsieh, Y.-T., Rim, T. H., eds. (2023). *Big data and artificial intelligence in ophthalmology - clinical application and future exploration*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4171-5

Table of contents

frontiers | Frontiers in Medicine

TYPE Editorial
PUBLISHED 06 December 2023
DOI 10.3389/fmed.2023.1339280

OPEN ACCESS

EDITED AND REVIEWED BY
Jodhbir Mehta,
Singapore National Eye Center, Singapore

*CORRESPONDENCE
Tae-im Kim
✉ tikim@yuhs.ac

RECEIVED 15 November 2023
ACCEPTED 23 November 2023
PUBLISHED 06 December 2023

CITATION
Tan YY, Rim TH, Ting DSJ, Hsieh Y-T and Kim T-i (2023) Editorial: Big data and artificial intelligence in ophthalmology - clinical application and future exploration.
Front. Med. 10:1339280.
doi: 10.3389/fmed.2023.1339280

Editorial: Big data and artificial intelligence in ophthalmology - clinical application and future exploration

Yong Yu Tan[1], Tyler Hyungtaek Rim[2,3], Darren S. J. Ting[4,5,6], Yi-Ting Hsieh[7,8] and Tae-im Kim[9,10]*

[1]Cork University Hospital, Cork, Ireland, [2]Department of Ocular Epidemiology, Singapore Eye Research Institute, Singapore, Singapore, [3]Mediwhale Inc., Seoul, Republic of Korea, [4]Academic Unit of Ophthalmology, Institute of Inflammation and Ageing, University of Birmingham, Birmingham, United Kingdom, [5]Birmingham and Midland Eye Centre, Birmingham, United Kingdom, [6]Academic Ophthalmology, School of Medicine, University of Nottingham, Nottingham, United Kingdom, [7]Department of Ophthalmology, National Taiwan University Hospital, Taipei, Taiwan, [8]Department of Ophthalmology, College of Medicine, National Taiwan University, Taipei, Taiwan, [9]Department of Ophthalmology, The Institute of Vision Research, Yonsei University College of Medicine, Seoul, Republic of Korea, [10]Department of Ophthalmology, Corneal Dystrophy Research Institute, Yonsei University College of Medicine, Seoul, Republic of Korea

KEYWORDS

big data, ophthalmology, artificial intelligence, deep learning, machine learning

Editorial on the Research Topic
Big data and artificial intelligence in ophthalmology - clinical application and future exploration

Introduction

Artificial Intelligence (AI) stands at the forefront of innovation in ophthalmology, harnessing vast datasets to redefine diagnostics and treatment strategies. This Research Topic collates pioneering insights from global experts, emphasizing AI's transformative impact on ophthalmic healthcare. Contributors have adeptly navigated the challenges, offering novel algorithms and applications poised to elevate patient care, streamline service delivery, and broaden healthcare access. From intricate retinal imaging to expansive electronic health record analyses, the papers within this topic not only underscore AI's potent capabilities but also chart a course for its future roles in enhancing ophthalmological practice.

Diagnostic and predictive analytics

Won et al. presented a groundbreaking deep learning (DL) based classification system adept at distinguishing between bacterial and fungal keratitis through anterior segment photographs. Utilizing a dataset comprising 684 images from 107 patients confirmed with either bacterial or fungal keratitis, the study introduced two novel modules—the Lesion Guiding Module and the Mask Adjusting Module —which, when integrated with the ResNet-50 classifier, significantly outperformed the baseline model with an accuracy leap from 81.1 to 87.8%. The system's proficiency was further validated on an external set of 98 images, solidifying its potential as a rapid, reliable diagnostic tool in clinical settings.

Li et al. compared DL with human graders for evaluation against 300 fundus photographs. The AI's accuracy for diagnosing diabetic retinopathy and macular degeneration was on par with ophthalmologists, achieving an AUC of 0.990 and 0.945, respectively. It excelled in identifying glaucomatous optic neuropathy with an AUC of 0.994, better than human graders.

Liu et al. created ONION, a DL tool that discerns optic neuritis from optic neuropathy in acute phases, demonstrating an AUC of 0.903. Trained with EfficientNet-B0 on 871 eyes from 547 patients, ONION matched a retinal specialist's diagnostic ability, showing 79.6% sensitivity and 86.5% specificity in validation. It processed results in 23 s, highlighting its potential for rapid, accurate eye condition diagnosis in various healthcare settings.

Wang et al. introduced a novel DL approach to forecast postoperative visual outcomes in patients undergoing cataract surgery. Leveraging a dataset of 2051 eyes, their Model V achieved the lowest mean absolute error of 0.1250 and 0.1194 logMAR, and RMSE of 0.2284 and 0.2362 logMAR in the validation and test datasets, respectively. It achieved up to 91.7% precision and 93.8% sensitivity, indicating high predictive reliability and marking progress in preoperative patient assessments.

Shivananjaiah et al.'s study offerred a novel approach to predicting the likelihood of glaucoma progression to surgical intervention within 1 year, using DL. The researchers curated a cohort from electronic health records at Stanford University, capturing both structured data and free-text clinical notes from 2008 to 2020. The DL model was fed a blend of text embeddings from patient notes and structured clinical data, resulting in an impressive model performance—most notably, the multimodal fusion model exhibited an AUC of 0.899 and an F1 score of 0.745. This work demonstrates the potential role of DL in improving glaucoma treatment predictions using comprehensive patient data.

Image analysis and segmentation

Imran et al. introduced Feature Preserving Mesh Network (FPM-Net), a network that segments retinal vasculature semantically without preprocessing, crucial for supporting the analysis of ophthalmic diseases, achieving exceptional accuracy (96.92% on DRIVE, 97.28% on CHASE-DB, 97.27% on STARE) and efficiency with only 2.45 million parameters. The research showcases FPM-Net's proficiency in preserving detailed spatial features for improved segmentation performance, essential for accurate retinal vessel analysis, making it a valuable tool for early diagnosis and management of ophthalmic diseases.

Bai et al. developed DME-DeepLabV3+ model, a lightweight and proficient model for the extraction of diabetic macular oedema (DME) from optical coherence tomography (OCT) images. This study harnesses the DeepLabV3+ architecture to address the complexity of OCT images, where varied image quality and the blurred boundaries of DME regions pose a significant challenge. Evaluated using a dataset of 1711 OCT images and validated by experienced clinicians, the DME-DeepLabV3+ achieves remarkable performance metrics, including a mean Intersection over Union (MIoU) of 91.18% and high precision and recall rates. This innovation promises to streamline the diagnostic process, offering a rapid, automated, and accurate tool for DME extraction.

Innovation in disease detection and management

Gibson et al. harnessed the power of latent diffusion augmentation to enhance the DL analysis of neuro-morphology in limbal stem cell deficiency (LSCD). The study showcased a residual U-Net model, informed by the InceptionResNetV2 transfer learning model, to classify neuron morphology across various stages of LSCD compared to healthy controls. The model achieved accuracy in determining nerve fiber number (*R*-squared of 0.63), branching (*R*-squared of 0.63), and length (*R*-squared of 0.80). This method outperformed the same model trained only on original images, particularly in distinguishing LSCD with an AUC of 0.867. The results suggest that supplementing training data with latent diffusion-generated images can effectively enhance model performance.

Privacy and continual learning in AI

Verma et al. explored privacy-preserving methods in continual learning for medical image classification, focusing on retinal disease detection from OCT images and histology-based colon cancer classification. Their study revealed that Brain-Inspired Replay (BIR) excelled in retinal disease classification with notable accuracy, while Efficient Feature Transformations (EFT) were most accurate for colon cancer detection. They found that these methods, though slightly outperformed by joint retraining models, offer significant benefits for long-term clinical use by reducing catastrophic forgetting and enabling ongoing model updates without compromising patient privacy.

Conclusion

The breadth of this Research Topic, with its collection of varied and insightful studies, underscores the remarkable potential of big data and AI within ophthalmology. The contributions from the authors, with their innovative algorithms and forward-thinking perspectives, hold significant promise for transforming the fabric of eye care. These studies serve as cornerstones upon which future collaborative endeavors can be built, driving the advancement of ophthalmological practices into a new era. As we stand on the precipice of this technological revolution, the integration of these AI-driven tools and methodologies heralds a progressive shift toward enhanced patient care.

Author contributions

YT: Conceptualization, Writing – original draft, Writing – review & editing. TR: Writing – review & editing, Conceptualization. DT: Writing – review & editing. Y-TH: Writing – review & editing. T-iK: Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

The authors would like to thank the hundreds of colleagues who contributed to this Research Topic. The authors would also like to thank the board and staff of the Frontiers Publishing House for their continuous and unflinching support.

Conflict of interest

TR was employed by Mediwhale Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor JM declared a shared affiliation with the author TR.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 13 January 2023
DOI 10.3389/fmed.2022.1040562

OPEN ACCESS

EDITED BY
Steven Fernandes,
Creighton University, United States

REVIEWED BY
Usman Ali,
Sungkyunkwan University, Republic of Korea
Muhammad Owais,
Dongguk University Seoul, Republic of Korea

*CORRESPONDENCE
Rizwan Ali Naqvi
✉ rizwanali@sejong.ac.kr
Seung Won Lee
✉ swleemd@g.skku.edu

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Ophthalmology, a section of the journal Frontiers in Medicine

RECEIVED 09 September 2022
ACCEPTED 20 December 2022
PUBLISHED 13 January 2023

CITATION
Imran SMA, Saleem MW, Hameed MT, Hussain A, Naqvi RA and Lee SW (2023) Feature preserving mesh network for semantic segmentation of retinal vasculature to support ophthalmic disease analysis.
Front. Med. 9:1040562.
doi: 10.3389/fmed.2022.1040562

Feature preserving mesh network for semantic segmentation of retinal vasculature to support ophthalmic disease analysis

Syed Muhammad Ali Imran[1†], Muhammad Waqas Saleem[2], Muhammad Talha Hameed[2], Abida Hussain[1], Rizwan Ali Naqvi[3*†] and Seung Won Lee[4*]

[1]Faculty of CS and IT, Superior University, Lahore, Pakistan, [2]Department of Primary and Secondary Healthcare, Lahore, Pakistan, [3]Department of Unmanned Vehicle Engineering, Sejong University, Seoul, Republic of Korea, [4]School of Medicine, Sungkyunkwan University, Suwon, Republic of Korea

Introduction: Ophthalmic diseases are approaching an alarming count across the globe. Typically, ophthalmologists depend on manual methods for the analysis of different ophthalmic diseases such as glaucoma, Sickle cell retinopathy (SCR), diabetic retinopathy, and hypertensive retinopathy. All these manual assessments are not reliable, time-consuming, tedious, and prone to error. Therefore, automatic methods are desirable to replace conventional approaches. The accuracy of this segmentation of these vessels using automated approaches directly depends on the quality of fundus images. Retinal vessels are assumed as a potential biomarker for the diagnosis of many ophthalmic diseases. Mostly newly developed ophthalmic diseases contain minor changes in vasculature which is a critical job for the early detection and analysis of disease.

Method: Several artificial intelligence-based methods suggested intelligent solutions for automated retinal vessel detection. However, existing methods exhibited significant limitations in segmentation performance, complexity, and computational efficiency. Specifically, most of the existing methods failed in detecting small vessels owing to vanishing gradient problems. To overcome the stated problems, an intelligence-based automated shallow network with high performance and low cost is designed named Feature Preserving Mesh Network (FPM-Net) for the accurate segmentation of retinal vessels. FPM-Net employs a feature-preserving block that preserves the spatial features and helps in maintaining a better segmentation performance. Similarly, FPM-Net architecture uses a series of feature concatenation that also boosts the overall segmentation performance. Finally, preserved features, low-level input image information, and up-sampled spatial features are aggregated at the final concatenation stage for improved pixel prediction accuracy. The technique is reliable since it performs better on the DRIVE database, CHASE-DB1 database, and STARE dataset.

Results and discussion: Experimental outcomes confirm that FPM-Net outperforms state-of-the-art techniques with superior computational efficiency. In addition, presented results are achieved without using any preprocessing or postprocessing scheme. Our proposed method FPM-Net gives improvement results which can be observed with DRIVE datasets, it gives Se, Sp, and Acc as 0.8285, 0.98270, 0.92920, for CHASE-DB1 dataset 0.8219, 0.9840, 0.9728 and STARE datasets it produces 0.8618, 0.9819 and 0.9727 respectively. Which is a remarkable difference and enhancement as compared to the conventional methods using only 2.45 million trainable parameters.

KEYWORDS

ophthalmic diseases, retinal vasculature, retinal image segmentation, semantic segmentation, computer–aided diagnosis

1. Introduction

Ophthalmic diseases are increasing at an alarming rate. Early and automated diagnosis can help in preventing chronic ophthalmic disorders. Ophthalmic diseases include glaucoma, macular degeneration, Sickle cell retinopathy (SCR), and hypertensive and diabetic retinopathy. All of these are common but serious ophthalmic diseases and can lead to vision loss if not diagnosed at an early stage. An ophthalmological image assessment is commonly used for retinal disease analysis which shows retinal vessel changes that can lead to vision loss problems (1). Another vision loss syndrome that is affected by retinal ischemia is Sickle cell retinopathy (SCR). Reduced vessel density and altered vasculature shape are symptoms of sickle cell retinopathy (SCR) illness. Important biomarkers for early SCR identification include retinal vessels (1). A high blood sugar level causes the retinal illness known as diabetic retinopathy, which causes retinal vessels to enlarge or leak (2). A retinal condition called hypertensive retinopathy causes restricted retinal vessels as a result of elevated blood pressure which can be especially noticeable in the micro-vasculature (3). The location of the retinal vascular blockage can be determined using retinal vascular changes, which are often seen in bigger arteries. These retinal vascular illnesses are strongly related to the retinal morphologies of arteries and some other vessel diseases (1). Aimed at the early finding of chronic ophthalmic disorders by using different fundus images are retinal vessels which are a vital biomarker.

Precise retinal image analysis is necessary for early ophthalmic diagnosis. The complicated nature of the retinal blood vessels makes them essential biomarkers for diagnosing and analyzing many retinal disorders. However, it can be difficult to detect little changes in retinal vessels. Retinal vascular morphology includes location, thickness, tortuosity, formation, and removal, and is linked to several ocular illnesses (4). Ophthalmologists assess and record changes in the retinal vasculature manually. This procedure is time-consuming and labor-intensive. Additionally, the diagnosis of the aforementioned disorders can be made using the size of the retinal vessels, which is a distinct change that is difficult to find and evaluate using manual image analysis (4) by medical practitioners. Automatic illness inquiry is becoming more prevalent as deep learning technology progresses to help doctors make quicker and more accurate diagnoses (1). As the analysis of medical images is a crucial component of computer-aided disease diagnosis. Due to their dependability and adaptability, artificially intelligence-based approaches are more well-known in syndrome investigation than traditional image processing techniques. Deep learning-based algorithms help medical specialists to analyze various diseases using computer vision approaches (1–8).

Computer vision has an immense potential to evaluate these retinal disorders through image analysis for premature diagnosis. Ophthalmologists and other medical professionals are dealing with a variety of diagnostic challenges with the use of deep learning techniques like medical image segmentation. Semantic segmentation using deep learning is a cutting-edge technology for medical image segmentation that helps to avoid the manual processing of images for disease or symptom diagnosis (7). Most of the work done already for the retinal vessels segmentation is based on general image processing schemes; in which several image augmentation patterns were used to enhance the image contrast and detection process, which is usually based on some specific threshold. In such a case, a specific threshold cannot perform better with changes in the image acquisition system. Therefore, to incorporate the portability of the method, learning-based-segmentation algorithms are famous.

The process of semantic segmentation entails giving class labeling to each pixel of the image. Semantic segmentation may be thought of as the process of identifying an image class and isolating it from the other image classes by overlaying a

segmentation mask on top of it. Features extraction features and representations are frequently necessary for semantic segmentation to obtain an optimal correlation of the image, effectively reducing the noise. The suggested study explains the deep-learning-based semantic segmentation technique called Feature Preserving Mesh Network (FPM-Net) for the detection of precise retinal vasculature in fundus images. Here, we use multiple convolution layers with a combination of depth-wise separable convolutions to lessen the overall trainable parameters. Due to the spatial information being lost as a result of the pooling of layers, we employed feature-preserving blocks to maintain feature map sizes that were large enough to handle the lost spatial information. The dense connection prevents the vanishing gradient issue that plagues traditional networks' feature latency (9), leading to improved training. This feature-preserving block results in enhanced sensitivity of the suggested FPM Network without using costly preprocessing techniques. Finally, preserved features, low-level input image information, and up-sampled spatial features are aggregated at the final concatenation stage for improved prediction accuracy.

The suggested FPM-Net method was applied to the fundus images in three different publically available databases (5), The technique is reliable since it performs better even after being trained on the DRIVE database (2), STARE database (10), and CHASE-DB1 (10), making it appropriate for images captured under various situations without retraining. After experiments, the outcomes of segmentation concluded a meliorated performance with accuracy (Acc), sensitivity (Se), specificity (SP), and area under the curve (AUC) for retinal vasculature segmentation. The suggested method FPM-Net has a much better performance than conventional methods.

The structure of this paper is as follows. Some conventional and automated methods relevant to this work will be presented in Section 2. The embedding strategy and method are given in Section 3. Results can be found in Section 4 and discussions in Section 5. In Section 6, a conclusion is provided.

1.1. Research motivation

An increasing rate of growth is being observed in ophthalmic illnesses. Chronic ocular problems can be avoided with early and automated diagnosis. Retinal vascular alterations, which are frequently observed in larger arteries, can be used to pinpoint the exact location of the retinal vascular occlusion. The retinal morphology of arteries and a few other vessel diseases are closely related to these retinal vascular diseases (1). Retinal vessels, an important biomarker, are used to detect chronic retinal problems early by employing various fundus image observations. However, it could be challenging to spot slight variations in retinal vessels. The location, thickness, tortuosity, creation, and removal of retinal vessels all affect their morphology and are associated with several retinal diseases (4). Ophthalmologists manually evaluate and document changes to the retinal vasculature. This process takes a lot of time and effort. Additionally, the size of the retinal vessels, which is a unique alteration that is challenging to discover and analyze using manual image analysis (4), can be used to diagnose the aforementioned illnesses.

The evaluation of these retinal illnesses by image processing for early diagnosis has enormous potential for computer vision. Ophthalmologists and other medical practitioners are using deep learning methods like medical image segmentation to address a range of diagnostic issues. Deep learning-based semantic segmentation is an absolute technique for medical image segmentation that eliminates the need for manual image processing for the identification of illness or symptom (7).

2. Related work

Automated approaches are important for lowering the diagnostic workload of medical specialists, and the detection of retinal vasculature can be helpful for the premature investigation of a variety of eye-related diseases. There are two basic methods for segmenting retinal vessels: feature-based deep learning techniques and traditional image processing approaches. Various studies have been conducted using traditional techniques and common image-processing algorithms. Here we describe recent advances in image analysis and deep functionality learning techniques. Traditional image processing techniques have been studied recently, and deep learning-based techniques have grown with great constancy and performance (1). Researchers have previously developed a variety of machine-learning methods to separate the blood vessels from imaging the retinal fundus. When handling testing conditions such as recognized low-contrast micro-vessels, vessels with focal reflexes, and vessels within the sight of diseases, a significant number of visible retinal vessel division techniques are prone to more unfavorable results (2).

Numerous image-enhancement techniques are frequently used before thresholding in traditional image processing-based vessel segmentation approaches. In addition to using contrast-limited adaptive histogram equalization (CLAHE) to rise the divergence of fundus images, Alhussein et al. developed a segmentation method centered on Wiener and morphological filtering (3). The primary vascular region was located using the detector-based vessel identification approach developed by Zhou et al., and after the noise was removed, a Markov model was used to locate retinal vasculatures (11). In a similar vein, Ahamed et al. reported segmenting the autonomic vasculature multiscale line detection-based approach. To increase contrast, they added CLAHE toward the green channel and for the final segmentation, they combined morphological thresholding and hysteresis (4). For the segmentation of retinal vessels, Shah et al. employed a multiscale line-detection technique.

The images aimed at vessel segmentation were made better on the green channel using Gabor wavelet superposition and multiscale line detection (4). Using top hat with homomorphic filtering, Soto et al. presented a three-stage method. Following the initial stage of visual smoothing for image enhancement, two phases were employed to separately segment both thin and thick vessels. The segmentation findings were improved in the final stage with the application of morphological post-processing (5). Li et al. introduced an unsupervised technique in which integrated-tube marked point processes were applied to extract the vascular network from the images and to preprocess the images, image-enhancing techniques were applied. Utilizing the discovered tube width expansion, the final segmentation was carried out (7). Aswini et al. introduced an un-supervised technique consisting of hysteresis thresholding with two folds to identify retinal vessels. In their approach, morphological smoothness and background reduction were used to improve the fundus images before thresholding (8). Another approach based on image processing segmented the vasculature using the curvelet transform and line operation after pre-processing using anisotropic diffusion filtering, adaptive histogram equalization, and color space translation (11). Sundaram et al. suggested a hybrid strategy based on bottom-hat transform and multiscale image augmentation, where the segmentation work was carried out using morphological procedures (10). To reduce the aggravating noise that prevents vessel segmentation, using a probabilistic patch-based denoiser was recommended by Khawaja et al. (2) that combines a customized Frangi filter with a denoiser. After the CLAHE procedure, images are enhanced using an aggregated block-matching 3-D speckled filter, Naveed et al. suggested an unsupervised technique. Multiscale line detectors along with Frangi detectors were used in their model to segment data (12).

All the above-discussed methods are traditional image processing and some deep-feature-based learning techniques are used to investigate retinal vasculature segmentation. Learning-based approaches are increasingly well-known because, through feature-based learning, they may imitate the expertise of medical professionals. Furthermore, techniques for image augmentation make it possible to complete the task with lesser training samples. For supervised vessel segmentation, Oliveira et al. suggested an entirely convolutional deep-learning technique. They employed a multiscale convolutional network in a patch-based scenario, which was investigated by some kind of stationary wavelet transform (13).

Fraz et al. integrated the vessel centerlines identification method with the morphological bit plane slicing technique. They coupled bit plane slicing with vessel centerline on the enhanced gray-level images of retinal blood vessels (14). In addition to performing a mathematical morphological procedure on the image, Ghoshal et al. suggested an enhanced vascular extraction method from retinal images. They made negative grayscale images from the original and the image that had been removed from the vessels, then they excised to balance the image and then improved to produce thin vessels by turning the produced image into a binary image. To produce the vessel-extracted image, they finally combined the thin vessel image and binary image. They claimed that their performance results were satisfactory (15). The answers from the two-dimensional Gabor wavelet transform at various scales of each pixel were utilized as features by Soares et al. after they used this transform with supervised learning. They rapidly categorized a complicated model using a Bayesian classifier (16). To determine the properties necessary for segmenting retinal blood vessels, Ricci and Perfetti suggested a technique based on line operators. Because their model uses a line detector to analyze the green channel of retinal images, it is quicker and requires fewer features than prior approaches (17). A multi-layered forward-oriented artificial neural network was trained using the suggested artificial neural network approach by Marin et al. using a seven-dimensional feature vector. They employed the sigmoid activation function in each neuron of the three-layer network. They claimed that additional datasets are also successfully used by the trained network (18). A technique using a CNN architecture was created by Melinscak et al. to determine if each pixel is a vessel or a backdrop (19). According to Wang et al. proposal for a new retinal vascular segmentation approach that uses patch-based learning and Dense U-net, the approach seems attractive in terms of standard performance criteria (20). For segmenting retinal vessels, Guo et al. developed a CNN-based two-class classifier comprising two convolution layers and pooling layers, one dropout layer, and one loss layer. They concluded that the suggested approach had good accuracy and was quick to teach (21). Concerning the information loss brought on by image scaling during preprocessing, Leopold et al. proposed PixelBNN, an effective deep learning system for automatically segmenting fundus morphologies, and reported that it had a reduced test time and reasonably high performance (9). Technology advancements have produced images with a higher pixel density, sharp features, and a lot of data. As a result, good image quality can satisfy the requirements for actual application in image analysis and image comprehension (22). CNN is effective in classifying images and detecting objects, although the results vary depending on the network design, activation function chosen, and input picture quality. Poor quality input images have a detrimental impact on a CNN's performance, according to research (23), even if it is not immediately apparent. IterNet, a novel model based on UNet that can uncover hidden vessel information from the segmented vessel image rather than the raw input image, was proposed by Li et al. IterNet is made up of several mini-UNet iterations that can be up to four times deeper than a typical UNet (24). A new approach for segmenting blood vessels in retinal images was put out by Tchinda et al. The artificial neural networks and conventional edge detection filters are the foundation of this approach. The features vector is first extracted using edge detection filters. An artificial neural network is trained using the

obtained characteristics to determine whether or not each pixel is a part of a blood artery (25).

According to the properties of the retinal vessels in fundus images, a residual convolution neural network-based retinal vessel segmentation technique is presented. The encoder-decoder network structure is built by joining the low-level and high-level feature graphs, and atrous convolution is added to the pyramid pooling. The improved residual attention module and deep supervision module are used. The results of the trials performed using the fundus image data set from DRIVE and STARE demonstrate that this algorithm can successfully segment all retinal vessels and identify related vessel stems and terminals. This approach can identify more capillaries and is viable and successful for segmenting retinal vessels in fundus images (11). One of the most serious infectious diseases in the world, tuberculosis causes 25% of all preventable deaths in underdeveloped nations. This cross-sectional descriptive research set out to assess the effects of ocular TB on visual acuity both before and after 2 months of vigorous anti-tubercular treatment. Three individuals with pleural TB, seven with disseminated tuberculosis, and 133 with pulmonary tuberculosis comprised the sample. Every patient got a standard eye examination, which included measuring visual acuity and performing necessary indirect ophthalmoscopes, biomicroscopy, applanation tonometry, and fluorescence angiography. None of the patients exhibited tuberculosis-related vision impairment. The incidence of ocular involvement was determined to be 4.2% (6/143). Five of the six individuals with ocular involvement and one of the suspected ocular lesions satisfied the diagnostic criteria for probable ocular lesions. Two individuals showed bilateral findings of different ocular lesions: one had sclera uveitis and the other had choroidal nodules. The remaining four patients all had unilateral lesions, including unilateral choroidal nodules in the right eye, unilateral choroidal nodules in the left eye, and unilateral peripheral retinal artery blockage in the right eye (two cases). After 2 months of rigorous therapy, patients made favorable improvements with no discernible visual loss (26).

3. Suggested methodology

3.1. Suggested FPM-Net's outline

As explained in section 2, retinal vessels are assumed as an important potential biomarker for the diagnosis of many ophthalmic diseases. A very growing number of ophthalmic illnesses are found in a large number of people around the globe. Preventing persistent ocular problems can be aided by early and automated diagnosis. Precise retinal image analysis is necessary for early ophthalmic diagnosis. Numerous AI-based techniques provide intelligent solutions for automatic retinal vessel recognition. However, segmentation performance, complexities, and computing efficiency were significantly constrained by previous approaches. Due to the vanishing gradient issue, and conventional architectural design, the majority of the currently used approaches specifically failed to achieve a higher true positive rate. Figure 1 provides an outline of the suggested technique. The suggested technique simply uses fundus images as input deprived of applying the requirement of any pre-processing scheme. FPM-Net is applied to the input image for pixel-wise classification. The suggested network categorizes each pixel into two major categories: "vessel" (for vessel pixel) and "background" (for pixels other than vessels). Because of this, it provides a binary segmentation mask with values of "1" on vessels as well as "0" on the other classes. FPM-Net incorporates a feature-preserving block for enhanced performance and fast convergence.

3.2. Architecture of suggested FPM-Net

A suggested network for segmenting vessels that was created especially to improve the sensitivity (a better true positive rate) of retinal vascular detection is called a Feature Preserving Mesh Network (FPM-Net). The suggested FPM-Net is shown in Figure 2. Observe (Figure 2) that FPM-Net is a dense network composed of multiple convolution operations, and a shallow feature up-sampling block (FUB) followed by mesh-connected dense feature down-sampling block (FDB), and this overall architecture differs from conventional semantic segmentation networks like Seg-Net, U-Net, and DeepLabV3 in terms of encoder-decoder architecture where the decoder is same as an encoder.

To address above mentioned issues with conventional networks, FPM-Net is following four design principles. First, multiple uses of convolution layers in deep networks (e.g., VGG16) cause spatial loss if they are used without a feature reuse policy and the overall performance deteriorates (27). Following Dense-Net (22), to cover the spatial loss, dense connections are used between the convolution layers available in the network which guarantees the immediate feature transfer without latency. Secondly, the convolution layers with a larger number of channels contribute to increasing the number of learnable parameters substantially. To reduce the network cost, we use depth-wise separable convolution on the deep side of the network. Third, the spatial information that is available in the initial layers is very important as it contains the low-level features to represent the edges. The FPM-Net is utilizing a dense mesh that is connecting all the convolutional layers and transfers this valuable low-level information from FDB to FUB directly. This ensures the immediate edge information transfer without latency which results in better segmentation performance and quicker convergence of the network. Fourth, the multiple pooling operation causes severe spatial information loss that inevitably leads to a deterioration in performance (28).

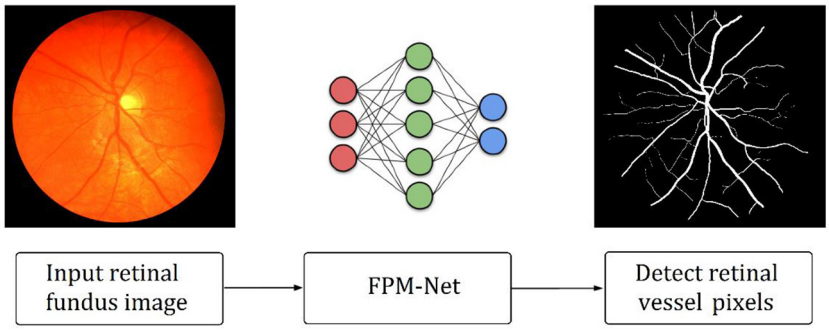

FIGURE 1
Outline of the suggested FPM-Net approach.

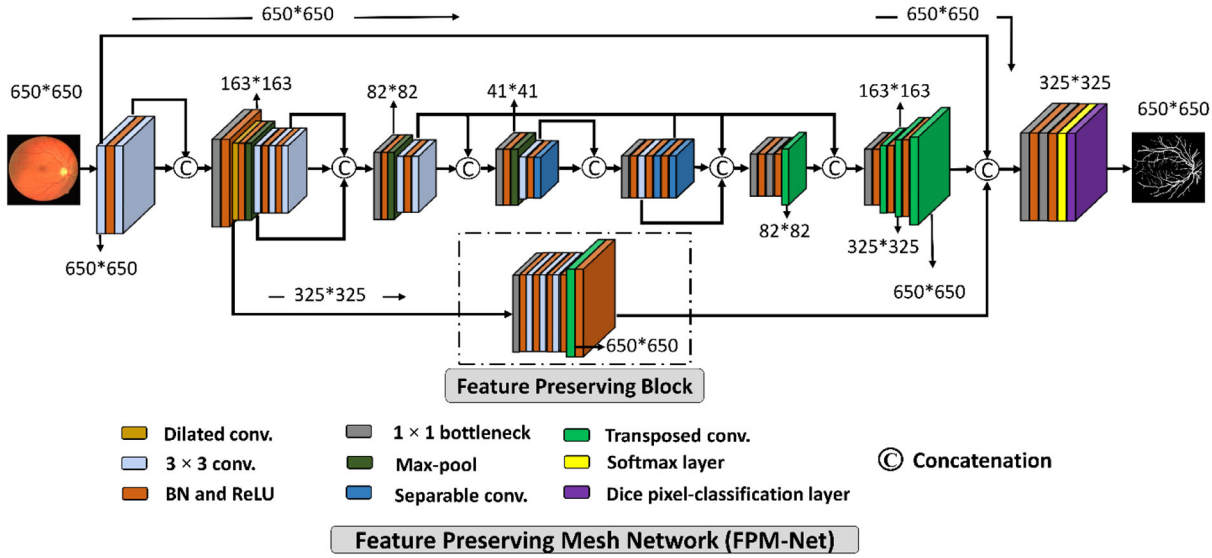

FIGURE 2
The suggested Feature Preserving Mesh Network (FPM-Net) architecture for retinal vessels segmentation.

Traditional convolutional neural networks employ excessive pooling operations for reducing the feature map size which is equally important to control memory usage. To cover the issues created by multiple pooling layers (minor information loss due to small feature map size), FPM-Net is using the feature preserving block (FPB) which keeps the feature map size larger to represent approximately all the valued features that can signify the vessel pixels. FPB is composed of a few low-cost convolution layers, and it is responsible to transfer a large feature map to the FUB. This FPM-Net provides better segmentation accuracy and is computationally efficient because it does not require a huge number of parameters for its training. This structure is completely diverse from traditional structures like Segmentation Networks (SegNet) (29) and U-Shaped Network (U-Net) (30), which employ a decoder similar to an encoder to produce an architecture that is excessively deep and has a lot of trainable parameters with many channels. Figure 2 explains the connectivity pattern of FPM-Net.

Figure 3 represents a schematic diagram for FPM-Net interconnection and the solid feature concatenation standards. The input convolution block uses the fundus images as input, runs them through many convolutional layers in FDB to extract significant features F_{ed} for the investigation of the retinal vasculatures, and then sends the enhanced dense features F_{ed} to the UB-A of FUB. K (F_{ed}) is created by concatenating the enhanced dense features T(F_{ed}) and intermediate feature information F_{if} that were acquired by the DFB-B and DFB-D, respectively. The K(F_{ed}) feature, represented by Equation (1), is produced *via* depth-wise concatenation using both T(F_{ed}) and F_{if}, where © represented depth-wise concatenation in green

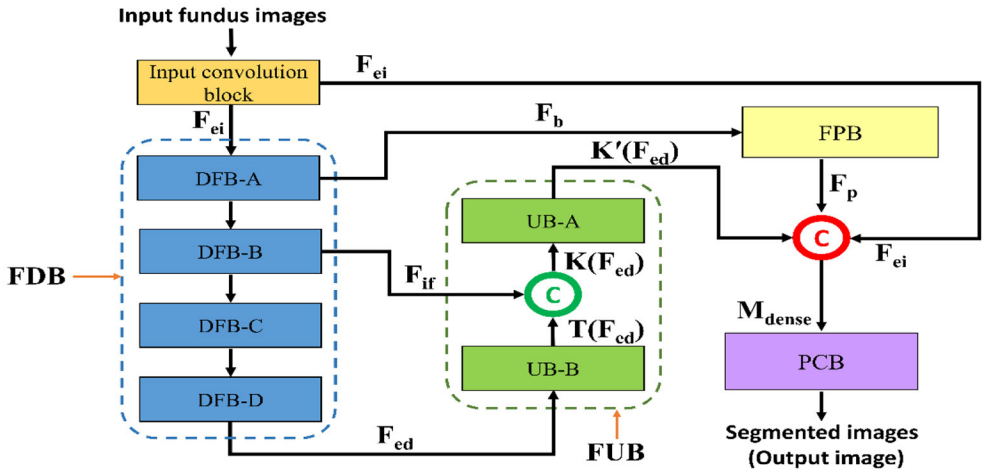

FIGURE 3
The schematic diagram for FPM-Net connectivity, FDB, FUB, FPB, and PCB represents the feature down-sampling block, feature up-sampling block, feature preserving block, and pixel classification block, respectively.

color. The F_b feature is being added to the feature-preserving block from the DFB-A. Since there haven't been any significant pooling operations, the feature F_p originating from the feature-preserving block (FPB) contains rich feature information that corresponds to the majority of the vessels in the images transfer to the final concatenation represented in red color.

$$K(F_{ed}) = T(F_{ed}) © F_{ei} \quad (1)$$

$$M_{dense} = K^{'}(F_{ed}) © F_p © F_{if} \quad (2)$$

Here, M is a densely concatenated feature made through the $K'(F_{ed})$, a feature after the up-sampling block, F_p preserved features, upcoming from the feature preserving block, and edge information f_{ei}, upcoming from the input convolution block. Where © denotes depth-wise concatenation. After final concatenation represented in red color concluded the output result having Equation (2).

3.3. Structure of feature preserving block

As shown in Figure 2, the suggested FPM-Net uses a feature-preserving block (FPB) to preserve valuable spatial information and disseminates it for the final concatenation. FPB takes the input from the dilated convolution, performs its function, and provides the feature results for the concatenation to the later layer. Because in the initial layer there is potential spatial information and features that can signify most of the vasculatures which will be helpful in the final prediction. The main problem that occurs while segmenting the image, the small objects were lost called the vanishing gradient but in FPB this vanishing gradient issue is solved. It simply uses three convolution layers and one transposed convolution to increase feature map size the feature map is resized to its original size using transposed convolution. As discussed above, the edge information from the initial layer, preserved features from FPB, and enhanced dense features are concatenated in the final stage which will boost the segmentation performance and improve the overall accuracy. After the final concatenation, softmax and pixel classification layers are utilized. The schematic structure of the feature-preserving block is mentioned in Figure 4.

3.4. Structure of suggested pixel classification block

The final concatenation has shown in red before the pixel classification block is given the rich features, K, from the up-sampling block. The PCB encompasses a 1 × 1 bottleneck (used to reduce the number of channels for pixel classification block), softmax, and dice pixel classification layer. The image pixels are categorized using a dice pixel classification layer that uses dice loss to solve the class imbalance and give improved segmentation. In this instance, "vessel" and "background" are two segmentation classes with values of "1" and "0," respectively. The pixel classification block is made up of a convolution whose filters are matched to the number of classes. The image pixels are identified using a pixel classification layer that uses dice loss to solve the class imbalance (31) and give improved segmentation. The dice loss (L_{DL}) is represented mathematically as,

$$LDL = 1 - \left(\frac{2 \times \sum_i^j Q_{p-i} R_{T-i}}{\sum_i^j Q_{p-i}^2 + R_{T-i}^2}\right) \quad (3)$$

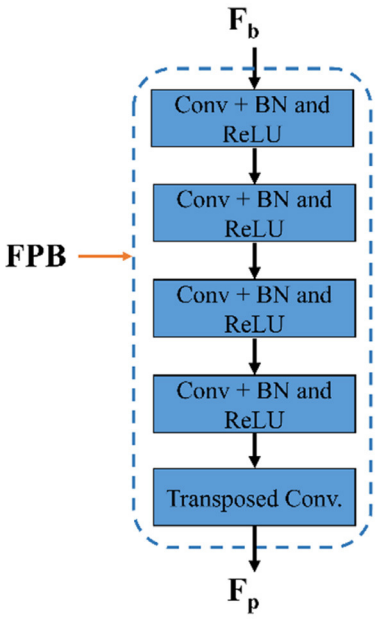

FIGURE 4
The structure of feature preserving block.

Where j refers to all of the image's observable pixels, i is the pixel under consideration, Q refers to the predicted labels, and R refers to the actual ground truth labels. R_{T-i} is the actual ground truth label, and Q_{p-i} is the expected possibility that pixel i belongs to a certain class.

4. Experimental results

4.1. Datasets

Intend to find results, vessels analysis was done on the DRIVE (2), CHASE-DB1 (10), and STARE (10) datasets for the suggested technique and additional studies for overall evaluation. These datasets are publicly accessible, and pixel-wise expert annotations on the photographs allow researchers to assess the algorithms. The following describes these datasets.

In the DRIVE dataset, 40 red, green, and blue fundus images in total are included in the collection. The dataset comes with carefully separated ground truths for analysis. The images have a 565 x 584-pixel resolution and a 45° field of view (FOV). For improved training, the 20 training images are enhanced. Examples of expertly annotated images on or after the DRIVE dataset are displayed in Figure 5A. In the CHASE-DB1 dataset with 28 images using a fundus camera (Nidek NM-200D) with a typical FOV of 30°. Complying with the validation requirement, with a total of 28 images, 20 images (with augmentation) were used in our studies for training purposes and the remaining eight for testing purposes. Examples of image pairings with professional annotations are shown in Figure 5B. The STARE dataset is a collection of 20 retinal images taken by a TopCon TRV-50 with a FOV of 35°. For assessment reasons, professional image annotations are given per image. We used cross-validation using the leave-one-out method in our studies, in which training is done on 19 images and just one left for testing. Similarly to this, each image in the 20 studies was chosen specifically for testing. Twenty experiments on average were used to get the data. Examples of image pairings with professional annotations from the STARE dataset are shown in Figure 5C. The training and testing image descriptions for each dataset are displayed in Table 1.

4.2. Experimental environment and augmented data

The suggested FPM-Net was developed using Microsoft Windows 10, MathWorks MATLAB R2022a, with a laptop having specifications. An Intel Core i7-11800H processor and RAM of 16 GB. The tests were performed using an NVIDIA GeForce RTX 3070 8GB GDDR6 graphics processing unit. Without using any method for weight initialization, migration, sharing, or fine-tuning from previous networks, the suggested models were trained from scratch. Tables 3A–C lists the important training hyperparameters.

Deep learning's segmentation effectiveness is closely correlated with the capacity of training data with labels; effective training requirements, and a substantial amount of training data with labels. To boost the quantity of data, we used image flipping and translation. The modified augmentation method involved flipping 20 original images in both vertical direction and horizontal directions to produce a total of 60 images. Then, the total images produced after the flipping procedure are 3,000, from the DRIVE dataset were produced by repeatedly translating these 60 images into (x, y) values and then continuing to flip them. A training set is prepared using a random image generation procedure, where the points (x, y) satisfy the conditions. The CHASE-DB and STARE databases were similarly enhanced to provide 1,500 and 1,300 images, respectively.

Considering the training details FPM-Net utilized an epsilon of 0.000001, and the initial learning rate of 0.00005 was applied. Global L2 normalization is utilized for training due to the benefits of quicker convergence and robustness over rising variation. To train the FPM-Net, a mini-batch size of 16 images is used because it is a dense network and requires less GPU memory due to bottleneck layers. In 25 epochs, both networks converge (5,000 iterations).

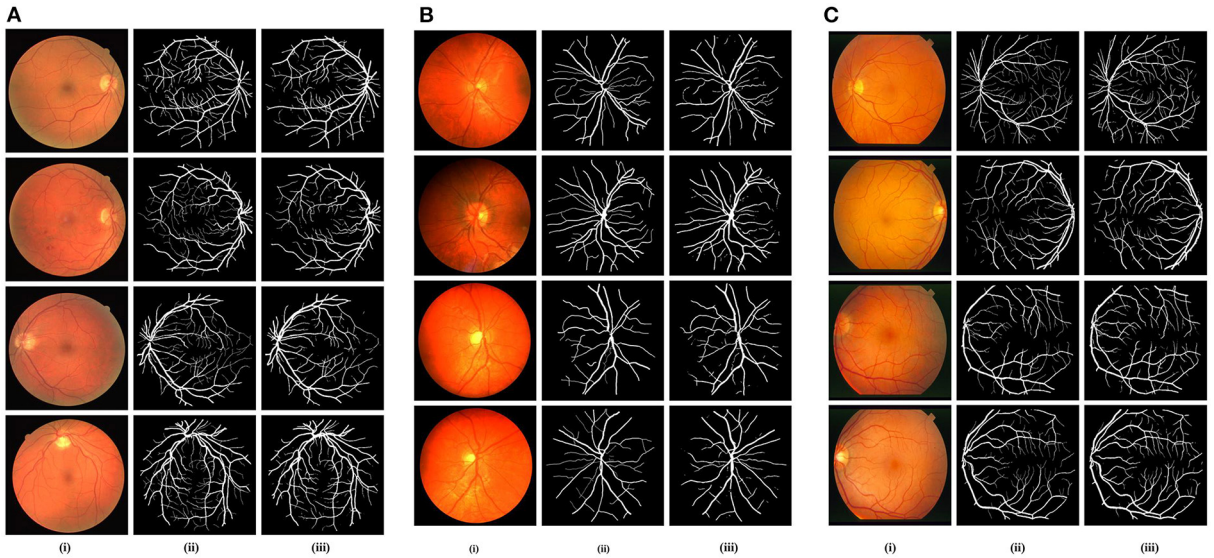

FIGURE 5
(A) DRIVE Dataset visualizations of the suggested FPM-Net: (i) Input original image, (ii) Expert annotation (Ground truth), and (iii) Predicted mask by FPM-Net. **(B)** CHASE-DB1 Dataset visualizations of the suggested FPM-Net (i) Input original image, (ii) Expert annotation (Ground truth), and (iii) Predicted mask by FPM-Net. **(C)** STARE Dataset visualizations of the suggested FPM-Net (i) Input original image, (ii) Expert annotation (Ground truth), and (iii) Predicted mask by FPM-Net.

TABLE 1 Details of the testing and specifications of all three used datasets in our method.

Name of dataset	Total images	Images division (training, testing)	Experimentation
DRIVE (2)	40 images	20, 20	One experiment
CHASE-DB1 (10)	28 images	20, 8	One experiment
STARE (10)	20 images	19, 1	20 experiments

TABLE 2 Performance measures with ablation study.

Method	SE	SP	Acc	AUC	Parameters
FPM-Net (without FPB)	0.8035	0.9801	0.9591	0.9790	2.44M
FPM-Net (with FPB)	0.8285	0.9827	0.9692	0.9851	2.45M

FPM-Net, Feature preserving mesh network; FPB, Feature preserving block; SE, sensitivity; SP, specificity; Acc, Accuracy; AUC, Area under Curve; ms, microseconds.

4.3. Ablation study for the suggested FPM-Net

The rich edge information is found in the starting layers by the network detection. By minimizing the vanishing gradient problem, the network's convergence is aided by the import of this data through skip connections (44). To investigate the efficacy of preserved features and dense connectivity for the suggested FPM-Net, an ablation study was conducted. In the ablation study, the training was done on FPM-Net architecture with and without FPB. Table 2 shows that, while maintaining the almost same number of parameters, FPB with preserved feature outperformed FPM-Net with dense connectivity in terms of true positive rate (SE), with a greater true positive rate. Table 2 clearly shows that feature concatenation caused a significant performance difference.

4.4. Evaluation of suggested network

For the suggested network output, FPM-Net offers a mask that displays all of the background and vessel pixels as "0" and "1," respectively. Sensitivity (SE), Specificity (SP), Accuracy (Acc), and area under curve AUC, to measure the performance of segmentation which are frequently utilized to assess how well-retinal images are segmented, were computed using the output mask of the suggested network and expert annotations (16). SE is denoted as a true positive rate, which illustrates how well the network can find vessel pixels. The SP as a true negative rate demonstrates the capacity to identify non-vessel pixels. The whole percentage of accurate predictions made thru the approach is represented by Acc. Equations (4)–(6) give the respective expressions for SE, SP, and Acc. A pixel with the prefix TP is identified in the expert's annotation as a vessel pixel and is projected to be one. FN denotes a pixel that the expert

annotation classifies as a vessel pixel even if it is expected to be a background pixel. A pixel with the prefix TN is identified in the expert's annotation as a vessel pixel and is expected to be one. FP denotes a pixel that the expert annotation classifies as a background pixel but which is expected to be a vessel pixel.

$$SN = \frac{TP}{TP + FN} \tag{4}$$

$$SP = \frac{TN}{TN + FP} \tag{5}$$

$$Acc = \frac{TP + TN}{TP + FN + FP + TN} \tag{6}$$

4.5. Comparison with other conventional techniques

To evaluate and compare the suggested FPM network with the conventional techniques, vessel analysis was done on the publicly accessible DRIVE CHASE-DB1, and STARE datasets. For the vessel category and the background category, the network generates a mask with both the corresponding grayscale values of "1" and "0," respectively. The visual outcomes of the suggested strategy for the three datasets stated above are shown in Figure 5. The suggested FPM-Net network's segmented image with the mask overlapped is shown in the figures along with the original images that were used as input into the network, experts provided the expert annotated image to evaluate research methods, the predicted mask at the network's production, and the predicted mask itself. The Numerical Comparison of Suggested FPM-Net utilizing the most recent technique is described in Tables 3A–C. By using our proposed method FPM-Net, there is significant improvement can be observed with DRIVE datasets, it gives S_e, S_p, and A_{cc} as 0.8285, 0.98270, 0.92920, for CHASE-DB1 dataset 0.8219, 0.9840, 0.9728 and STARE datasets it produces 0.8618, 0.9819 and 0.9727 respectively. Which is a remarkable difference and enhancement as compared to old and conventional methods.

TABLE 3A The comparison of the DRIVE data set's segmentation results using various segmentation techniques.

Method	Year	S_e	S_p	A_{cc}
Cross modality learning (17)	2015	0.7569	0.9816	0.9527
GMM classifier (18)	2015	0.7249	0.9830	0.9620
SP model (19)	2016	0.7811	0.9807	0.9535
CRF model (20)	2016	0.7897	0.9684	–
VS method (21)	2017	0.7779	0.9780	0.9521
RU-Net and R2U-Net (9)	2018	0.7792	0.9813	0.9556
LadderNet (22)	2018	0.7856	0.9810	0.9561
U-Net+joint losses (23)	2018	0.7653	0.9818	0.9542
CTF-Net (24)	2018	0.7979	0.9857	0.9685
Three-stage DL Model (25)	2019	0.7631	0.9820	0.9538
SD-Unet (32)	2019	0.7891	0.9848	0.9674
Dilated Conv. (33)	2019	0.7903	0.9813	0.9567
GFM (15)	2020	0.7614	0.9837	0.9604
DL methods (10)	2020	0.7979	0.9794	0.9563
AA-UNet (34)	2020	0.7941	0.9798	0.9558
EDC-Net (35)	2020	0.7092	0.9820	0.9447
Iternet (36)	2020	0.7735	0.9838	0.9673
MLC scheme (37)	2021	0.7761	0.9792	0.9519
LAC network (38)	2021	0.7921	0.9810	0.9568
ResDo-UNet (39)	2021	0.7985	0.9791	0.9561
FPM-Net (proposed)	2022	0.8285	0.98270	0.96920

"–" means the value is not available in the relevant research study.

4.6. Visual outcomes of suggested FPM-Net

In this instance, the suggested method's graphical outcomes for the identification of retinal vessels on the datasets of fundus image e.g., DRIVE, CHASE-DB1, and STARE are shown. (i) input original image, (ii) expert annotation (Ground truth), and (iii) FPM-Net mask are shown in Figures 5A–C.

5. Discussion

Precise retinal image analysis is necessary for early ophthalmic diagnosis. The complicated nature of the retinal blood vessels makes them essential biomarkers for diagnosing and analyzing many retinal disorders. However, it can be difficult to detect little changes in retinal vessels. Ophthalmologists assess and record changes in the retinal vasculature manually. To evaluate these retinal disorders through image investigation for premature diagnosis, computer vision has immense potential. Ophthalmologists and other medical professionals are dealing with a variety of diagnostic challenges with the use of deep learning techniques like medical image segmentation. Semantic segmentation using deep learning is a cutting-edge technology for medical image segmentation that helps to avoid the manual processing of images for disease or symptom diagnosis. With the advancement of supervised learning, autonomous sickness analysis is becoming more prevalent to help doctors make a quicker and more precise diagnosis. This semantic segmentation technique using deep learning will help ophthalmologists in this regard. The suggested study suggests the deep-learning-based semantic segmentation technique called FPM-Net for the detection of precise retinal vasculature in fundus images. Here, we use multiple convolution layers with a combination of

TABLE 3B The comparison of the CHASE-DB1 data set's segmentation results using various segmentation techniques.

Method	Year	S_e	S_p	A_{cc}
U-Net (40)	2015	0.7841	0.9701	0.9578
Cross modality learning (17)	2016	0.7507	0.9793	0.9581
RU-Net and R2U-Net (9)	2018	0.7756	0.9820	0.9634
U-Net+joint losses (23)	2018	0.7633	0.9809	0.9610
LadderNet (22)	2018	0.7978	0.9818	0.9656
U-Net+joint losses (23)	2018	0.7633	0.9809	0.9610
Three-stage DL Model (25)	2019	0.7641	0.9806	0.9607
GNN (41)	2019	0.9463	0.9364	0.9373
MCP-EM (42)	2019	0.8106	0.9807	0.9654
Ipn-v2 and octa-500 (27)	2019	0.8155	0.9725	0.9610
AA-UNet (34)	2020	0.8167	0.9704	0.9608
HAnet (28)	2020	0.8239	0.9813	0.9670
Iternet (36)	2020	0.7970	0.9823	0.9655
CTF-Net (29)	2020	0.7948	0.9842	0.9648
LAC network (38)	2021	0.7818	0.9819	0.9635
HDS-Net (30)	2020	0.8176	0.9776	0.9632
ResDo-UNet (39)	2021	0.8020	0.9794	0.9672
FPM-Net (proposed)	2022	0.8219	0.9840	0.9728

TABLE 3C The comparison of the STARE data set's segmentation results using various segmentation techniques.

Method	Year	S_e	S_p	A_{cc}
ECB method (43)	2012	0.7548	0.9763	0.9543
SP model (19)	2016	0.7867	0.9754	0.9566
CRF model (20)	2016	0.7680	0.9738	–
Cross modality learning (17)	2016	0.7726	0.9844	0.9628
DSM-UNet (44)	2018	0.7673	0.9901	0.9712
U-Net+joint losses (23)	2018	0.7581	0.9846	0.9612
CRF-Net (45)	2018	0.7543	0.9814	0.9632
SD-UNet (32)	2019	0.7548	0.9899	0.9725
Three-stage DL Model (25)	2019	0.7735	0.9857	0.9638
Ipn-v2 and octa-500 (27)	2019	0.7595	0.9878	0.9641
AA-UNet (34)	2020	0.7598	0.9878	0.9640
Iternet (36)	2020	0.7715	0.9886	0.9701
NFN+ Net (46)	2020	0.7963	0.9863	0.9672
D-GaussianNet (47)	2021	0.7904	0.9843	0.9837
HDS-Net (30)	2021	0.7946	0.9821	0.9626
ResDo-UNet (39)	2021	0.7963	0.9792	0.9567
FPM-Net (proposed)	2022	0.8618	0.9819	0.9727

depth-wise separable convolutions to lessen the overall trainable parameters. Due to the spatial information being lost as a result of the pooling of layers, we employed feature-preserving blocks to maintain feature map sizes that were large enough to handle the lost spatial information. The dense connection prevents the vanishing gradient issue that plagues traditional networks' feature latency (9), leading to improved training. This feature preserves block outcomes in improved sensitivity of the suggested FPM-Net deprived of using costly preprocessing techniques. Finally, preserved features, low-level input image information, and up-sampled spatial features are aggregated at the final concatenation stage for improved prediction accuracy. In previous studies, researchers used different networks such as AA-UNet (34), Iternet (36), NFN+ Net (46), D-GaussianNet (47), HDS-Net (30), and ResDo-Net (39) for the identification of Sensitivity (SE), Specificity (SP), Accuracy (Acc), and area under curve AUC, to measure the performance of segmentation which are frequently utilized to assess how well retinal images are segmented. But in this paper, our proposed FPM-Net produced more accurate results for SE, SP, Acc, and AUC than the rest of the research done by others. In this paper, a solid architecture is shown that enables precise semantic segmentation of the retinal blood vessels. The central ideas are discussed below.

- An efficient semantic segmentation network may give precise vessel detection deprived of the need for costly preprocessing.
- The network can learn adequate features for enhanced segmentation and quicker convergence because it delivers enhanced spatial information from the initial layers.
- Creating a shallow architecture can save many trainable parameters and it is not necessary to make feature up-sampling and feature down-sampling blocks identical. To reduce the network cost, we use depth-wise separable convolution on the deeper side of the network.
- While considering vessel segmentation, a shallower architecture with fewer layers and a smaller quantity of trainable parameters performs superior to robust architecture.
- The size of the ultimate feature map is essential. In contrast to existing architectures that significantly down-sample the image, FPM-Net avoids pooling layers and maintains enough feature map size which contains valuable features and offers better performance.
- Those techniques which are based on deep learning could help ophthalmologists do analysis more quickly and offer numerous approaches for analyzing diseases.

The original images used as input into the network, the expert-annotated image provided by experts to assess research

methodologies, the predicted mask at the network's production, and the predicted mask itself are all displayed in the figures along with the suggested FPM-Net network's segmented image with the mask overlapped. Tables 3A–C describes the Numerical Comparison of the Suggested FPM-Net using the most recent method. By using our proposed method FPM-Net, there is significant improvement can be observed with DRIVE datasets, it gives S_e, S_p, and A_{cc} as 0.8285, 0.98270, 0.92920, for CHASE-DB1 dataset 0.8219, 0.9840, 0.9728 and STARE datasets it produces 0.8618, 0.9819 and 0.9727 respectively. Which is a remarkable difference and enhancement in results as compared to old and conventional methods.

5.1. Limitations and future work

Even though the suggested FPM-Net recognizes retinal vessels with better segmentation performance, the suggested technique still has certain limitations. A learning-based segmentation technique, the suggested FPM-Net largely depends on the input training data. Medical data for disease analysis are extremely challenging to organize in large quantities. The amount of training data must thus be artificially increased by data augmentation. Additionally, the learning-based approaches produce output masks depending on the knowledge they have acquired, and the network's ultimate prediction may contain pixels that are both false positive and false negative.

We want to minimize the network's overall cost in the future by efficiently reducing the number of convolutions. The proposed technique is based on deep learning, as well as its efficiency solely depends on excellent training with sufficient training data. Additionally, the accuracy of the labeling generated by an ophthalmologist directly affects the precision of learning-based techniques. This will make it feasible to evaluate how well-upcoming deep-learning techniques screen for these particular disorders. We also want to develop a little system for mobile applications that run instantly. The medical sector will subsequently utilize these networks for more semantic segmentation purposes.

6. Conclusion

The goal of this study was to develop a network for segmenting shallow vessels that might effectively be used to support computer-aided diagnostics in the identification and diagnosis of retinal disease. The proposed method utilized the FPM-Net shallow network, which provides a successful remedy for retinal vasculature for computer-aided diagnostics. The recommended FPM network uses less memory, has more trainable parameters, and fewer layers, and can be trained with larger mini-batch sizes. A separate network with the name of FPM-Net is used to maintain a reduced final feature map during its convolutional phase. FPM-Net contains an improved portion of FPB that incorporates an external path that saves and delivers essential spatial information to increase the accuracy and robustness of the technique. As a result, when compared to other traditional approaches for detecting retinal vessels, our suggested vessel segmentation networks are more reliable and perform better without preprocessing, and they may be utilized to help medical professionals to diagnose and analyze diseases.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: Khawaja et al. (2) and Sundaram et al. (10).

Author contributions

SI: methodology and writing—original draft. MS, MH, and AH: validations. RN and SL: supervision and writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by a National Research Foundation (NRF) grant funded by the Ministry of Science and ICT (MSIT) and South Korea through the Development Research Program (NRF2022R1G1A1010226 and NRF2021R1I1A2059735).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Owais M, Arsalan M, Choi J, Mahmood T, Park KR. Artificial intelligence based classification of multiple gastrointestinal diseases using endoscopy videos for clinical diagnosis. *J Clin Med*. (2019) 8:986. doi: 10.3390/jcm8070986

2. Khawaja A, Khan TM, Naveed K, Naqvi SS, Rehman NU, Junaid Nawaz S. An improved retinal vessel segmentation framework using frangi filter coupled with the pobabilistic patch based denoiser. *IEEE Access*. (2019). 7:164344–61.

3. Alhussein M, Aurangzeb K, Haider SI. An unsupervised retinal vessel segmentation using Hessian and intensity based approach. *IEEE Access*. (2020) 8:165056–70. doi: 10.1109/ACCESS.2020.3022943

4. Ahamed ATU, Jothish A, Johnson G, Krishna SBV. Automated system for retinal vessel segmentation. In: *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*. IEEE (2018).

5. Ramos-Soto O, Rodr-Esparza E, Balderas-Mata SE, Oliva D, Hassanien AE, Meleppat RK, et al. An efficient retinal blood vessel segmentation in eye fundus images by using optimized top-hat and homomorphic filtering. *Comput Methods Programs Biomed*. (2021) 201:105949. doi: 10.1016/j.cmpb.2021.105949

6. Shah SAA, Shahzad A, Khan MA, Li C-K, Tang TB. Unsupervised method for retinal vessel segmentation based on gabor wavelet and multiscale line detector. *IEEE Access*. (2019) 7:167221–8.

7. Li T, Comer M, Zerubia J. An unsupervised retinal vessel extraction and segmentation method based on a tube marked point process model. In: *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE (2020).

8. Aswini S, Suresh A, Priya S, Krishna BVS. Retinal vessel segmentation using morphological top hat approach on diabetic retinopathy images. In: *2018 Fourth International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*. IEEE (2018).

9. Leopold HA, Orchard J, Zelek JS, Lakshminarayanan V. PixelBNN: Augmenting the PixelCNN with batch normalization and the presentation of a fast architecture for retinal vessel segmentation. *J Imaging*. (2019) 5:26. doi: 10.3390/jimaging5020026

10. Sundaram R, Ravichandran KS, Jayaraman P. Extraction of blood vessels in fundus images of retina through hybrid segmentation approach. *Mathematics* (2019) 7:169. doi: 10.3390/math7020169

11. Zhou C, Zhang X, Chen H. A new robust method for blood vessel segmentation in retinal fundus images based on weighted line detector and hidden Markov model. *Comput Methods Programs Biomed*. (2020) 187:105231. doi: 10.1016/j.cmpb.2019.105231

12. Naveed K, Abdullah F, Madni HA, Khan MAU, Khan TM, Naqvi SS. Towards automated eye diagnosis: An improved retinal vessel segmentation framework using ensemble block matching 3D filter. *Diagnostics (Basel)*. (2021) 11:114. doi: 10.3390/diagnostics11010114

13. Oliveira A, Pereira S, Silva CA. Retinal vessel segmentation based on fully convolutional neural networks. *Exp Syst Appl*. (2018) 112:229–42. doi: 10.1016/j.eswa.2018.06.034

14. Fraz MM, Barman SA, Remagnino P, Hoppe A, Basit A, Uyyononvara B, et al. An approach to localize the retinal blood vessels using bit planes and centerline detection. *Comput Method Pgm Biomed*. (2012) 108:600–16. doi: 10.1016/j.cmpb.2011.08.009

15. Ghoshal R, Saha A, Das S. An improved vessel extraction scheme from retinal fundus images. *Multimedia Tools Appl*. (2019) 78:25221–39. doi: 10.1007/s11042-019-7719-9

16. Soares JVB, Leandro JJG, Ceser RM, Jelinek HF, Cree MJ. Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification. *IEEE Trans Med Imag*. (2006) 25:1214–22.

17. Ricci, E, Perfetti R. Retinal blood vessel segmentation using line operators and support vector classification. *IEEE Trans Med Imag*. (2007) 26:1357–65.

18. Marin D, Aquino A, Gegundez-Arias ME, Bravo JM.A new supervised method for blood vessel segmentation in retinal images by using gray-level and moment invariants-based features. *IEEE Trans Med Imag*. (2010) 30:146–58.

19. Melinscak M, Prentasic P, Loncaric S. *Retinal Vessel Segmentation using Deep Neural Networks*. VISAPP (2015).

20. Wang C, Zhao Z, Ren Q, Xu Y, Yu Y.Dense U-net based on patch-based learning for retinal vessel segmentation. *Entropy*. (2019) 21:168. doi: 10.3390/e21020168

21. Guo C, Szemenyei M, Pei Y, Yi Y, Zhou W. SD-UNet: A structured dropout U-Net for retinal vessel segmentation. In: *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE (2019).

22. Zhuang J. LadderNet: Multi-path networks based on U-Net for medical image segmentation. *arXiv [Preprint]*. (2018). arXiv: 1810.07810. doi: 10.48550/arXiv.1810.07810

23. Yan Z, Yang X, Cheng K-T. Joint segment-level and pixel-wise losses for deep learning based retinal vessel segmentation. *IEEE Trans Biomed Eng*. (2018) 65:1912-23. doi: 10.1109/TBME.2018.2828137

24. Li L, Verma M, Nakashima Y, Nagahara H, Kawaski R. Iternet: Retinal image segmentation utilizing structural redundancy in vessel networks. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. IEEE (2020).

25. Tchinda BS, Tchioptop S, Noubom M, Louis-Dorr V, Wolf D. Retinal blood vessels segmentation using classical edge detection filters and the neural network. *Inform Medicine Unlocked*. (2021) 23:100521. doi: 10.1016/j.imu.2021.100521

26. de Oliveira SBV, Passos F, Hadad DJ, Zbyszynski L, de Almeida JPS, Castellani LGS, et al. The impact of ocular tuberculosis on vision after two months of intensive therapy. *Braz J Infect Dis*. (2018) 22:159–65. doi: 10.1016/j.bjid.2018.03.005

27. Li M, Zhang Y, Ji Z, Xie K, Yuan S, Liu Q, et al. Ipn-v2 and octa-500: Methodology and dataset for retinal image segmentation. *arXiv [Preprint]*. (2020). arXiv: 2012.07261. doi: 10.48550/arXiv.2012.07261

28. Wang D, Haytham A, Pottenburgh J, Saeedi O, Tao Y. Hard attention net for automatic retinal vessel segmentation. *IEEE J Biomed Health Inform*. (2020) 24:3384–96. doi: 10.1109/JBHI.2020.3002985

29. Wang K, Zhang X, Huang S, Wang Q, Chen F. Ctf-net: Retinal vessel segmentation via deep coarse-to-fine supervision network. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. Iowa City, IA: IEEE (2020). p. 1237–41.

30. Yang L, Wang H, Zeng Q, Liu Y, Bian G, A. hybrid deep segmentation network for fundus vessels via deep-learning framework. *Neurocomputing*. (2021) 448:168–78. doi: 10.1016/j.neucom.2021.03.085

31. Sudre CH, Li W, Vercauteren T, Ourselin S, Jorge Cardoso M. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Quebec City, QC: Springer (2017). p. 240–8.

32. Guo C, Szemenyei M, Pei Y, Yi Y, Zhou W. SD-UNet: A structured dropout UNet for retinal vessel segmentation. In: *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*. Athens: IEEE (2019). p. 439–44.

33. Lopes AP, Ribeiro A, Silva CA. Dilated convolutions in retinal blood vessels segmentation. In: *2019 IEEE 6th Portuguese Meeting on Bioengineering (ENBENG)*. Lisbon: IEEE (2019). p. 1–4.

34. Lv Y, Ma H, Li J, Liu S. Attention guided U-Net with atrous convolution for accurate retinal vessels segmentation. *IEEE Access*. (2020) 8:32826–39. doi: 10.1109/ACCESS.2020.2974027

35. Sule O, Viriri S. Enhanced convolutional neural networks for segmentation of retinal blood vessel image. In: *2020 Conference on Information Communications Technology and Society (ICTAS)*. Durban: IEEE (2020). p. 1–6.

36. Li L, Verma M, Nakashima Y, Nagahara H, Kawasaki R. Internet: Retinal image segmentation utilizing structural redundancy in vessel networks. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE (2020). p. 3656–65.

37. Zou B, Dai Y, He Q, Zhu C, Liu G, Su Y, et al. Multi-label classification scheme based on local regression for retinal vessel segmentation. *IEEE/ACMTrans Comput Biol Bioinform*. (2020) 18:2586–97. doi: 10.1109/TCBB.2020.2980233

38. Li X, Jiang Y, Li M, Yin S. Lightweight attention convolutional neural network for retinal vessel image segmentation. *IEEE Trans Indust Inform*. (2020) 17:1958–67. doi: 10.1109/TII.2020.2993842

39. Liu Y, Shen J, Yang L, Bian G, Yu H. ResDO-UNet: A deep residual network for accurate retinal vessel segmentation fromfundus images. *Biomed Signal Process Control*. (2023) 79:104087. doi: 10.1016/j.bspc.2022.104087

40. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Munich: Springer (2015). p. 234–241.

41. Shin SY, Lee S, Yun ID, Lee KM. Deep vessel segmentation by learning graphical connectivity. *Med Image Anal*. (2019) 58:101556. doi: 10.1016/j.media.2019.101556

42. Tang P, Liang Q, Yan X, Zhang D, Coppola G, Sun W. Multiproportion channel ensemble model for retinal vessel segmentation. *Comput Biol Med*. (2019) 111:103352. doi: 10.1016/j.compbiomed.2019.103352

43. Fraz MM, Remagnino P, Hoppe A, Uyyanonvara B, Rudnicka AR, Owen CG, et al. An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Trans Biomed Eng*. (2012) 59:2538–48. doi: 10.1109/TBME.2012.2205687

44. Zhang Y, Chung A. Deep supervision with additional labels for retinal vessel segmentation task. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Hong Kong: Springer (2018). p. 83–91.

45. Hu K, Zhang Z, Niu X, Zhang Y, Cao C, Xiao F, et al. Retinal vessel segmentation of color fundus images using multiscale convolutional neural network with an improved cross-entropy loss function. *Neurocomputing*. (2018) 309:179–91. doi: 10.1016/j.neucom.2018.05.011

46. Wu Y, Xia Y, Song Y, Zhang Y, Cai W. NFN+: a novel network followed network for retinal vessel segmentation. *Neural Netw*. (2020) 126:153–62. doi: 10.1016/j.neunet.2020.02.018

47. Alvarado-Carrillo DE, Ovalle-Magallanes E, Dalmau-Cede1o OS. DGaussianNet: Adaptive distorted Gaussian matched filter with convolutional neural network for retinal vessel segmentation. In: *International Symposium on Geometry and Vision*. Auckland: Springer (2021). p. 378–92.

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 01 March 2023
DOI 10.3389/fmed.2023.1115032

OPEN ACCESS

EDITED BY
Tyler Hyungtaek Rim,
Mediwhale Inc., Republic of Korea

REVIEWED BY
Tae Keun Yoo,
B&VIIT Eye Center/Refractive Surgery & AI Center, Republic of Korea
Wen Fan,
Nanjing Medical University, China

*CORRESPONDENCE
Mingguang He
✉ mingguang_he@yahoo.com

SPECIALTY SECTION
This article was submitted to Ophthalmology, a section of the journal Frontiers in Medicine

RECEIVED 03 December 2022
ACCEPTED 03 February 2023
PUBLISHED 01 March 2023

CITATION
Li Z, Guo X, Zhang J, Liu X, Chang R and He M (2023) Using deep leaning models to detect ophthalmic diseases: A comparative study.
Front. Med. 10:1115032.
doi: 10.3389/fmed.2023.1115032

Using deep leaning models to detect ophthalmic diseases: A comparative study

Zhixi Li[1], Xinxing Guo[1,2], Jian Zhang[1], Xing Liu[1], Robert Chang[3] and Mingguang He[1]*

[1]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-Sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangdong Provincial Clinical Research Center for Ocular Diseases, Guangzhou, China, [2]Wilmer Eye Institute, Johns Hopkins University, Baltimore, MD, United States, [3]Department of Ophthalmology, Byers Eye Institute at Stanford University, Palo Alto, CA, United States

Purpose: The aim of this study was to prospectively quantify the level of agreement among the deep learning system, non-physician graders, and general ophthalmologists with different levels of clinical experience in detecting referable diabetic retinopathy, age-related macular degeneration, and glaucomatous optic neuropathy.

Methods: Deep learning systems for diabetic retinopathy, age-related macular degeneration, and glaucomatous optic neuropathy classification, with accuracy proven through internal and external validation, were established using 210,473 fundus photographs. Five trained non-physician graders and 47 general ophthalmologists from China were chosen randomly and included in the analysis. A test set of 300 fundus photographs were randomly identified from an independent dataset of 42,388 gradable images. The grading outcomes of five retinal and five glaucoma specialists were used as the reference standard that was considered achieved when ≥50% of gradings were consistent among the included specialists. The area under receiver operator characteristic curve of different groups in relation to the reference standard was used to compare agreement for referable diabetic retinopathy, age-related macular degeneration, and glaucomatous optic neuropathy.

Results: The test set included 45 images (15.0%) with referable diabetic retinopathy, 46 (15.3%) with age-related macular degeneration, 46 (15.3%) with glaucomatous optic neuropathy, and 163 (55.4%) without these diseases. The area under receiver operator characteristic curve for non-physician graders, ophthalmologists with 3–5years of clinical practice, ophthalmologists with 5–10years of clinical practice, ophthalmologists with >10years of clinical practice, and the deep learning system for referable diabetic retinopathy were 0.984, 0.964, 0.965, 0.954, and 0.990 (p=0.415), respectively. The results for referable age-related macular degeneration were 0.912, 0.933, 0.946, 0.958, and 0.945, respectively, (p=0.145), and 0.675, 0.862, 0.894, 0.976, and 0.994 for referable glaucomatous optic neuropathy, respectively (p<0.001).

Conclusion: The findings of this study suggest that the accuracy of this deep learning system is comparable to that of trained non-physician graders and general ophthalmologists for referable diabetic retinopathy and age-related macular degeneration, but the deep learning system performance is better than that of trained non-physician graders for the detection of referable glaucomatous optic neuropathy.

KEYWORDS

deep learning, diabetic retinopathy, age-related macular degeneration, glaucomatous optic neuropathy, fundus photograph

Introduction

Diabetic retinopathy (DR), glaucomatous optic neuropathy (GON), and age-related macular degeneration (AMD) are responsible for more than 18% of visual impairment and blindness cases globally (1–6). While it is estimated that 80% of vision loss is avoidable through early detection and intervention (7–9), approximately 50% of cases remain undiagnosed (10, 11). High rates of undiagnosed disease can be attributed to these conditions being asymptomatic in their early stages, coupled with a disproportionately low availability of eye care services, particularly within developing countries and under-served populations (12).

Previous research has demonstrated that color fundus photography is an effective tool for the diagnosis of AMD, GON, and DR (13–15). Despite this, accurate interpretation of the optic nerve and retina is highly dependent on clinical experts, limiting the utility in low recourse settings. Deep learning represents an advancement of artificial neural networks that permits improved predictions from raw image data (16). Recently, several studies have investigated the application of deep learning algorithms for the automated classification of common ophthalmic disorders (17–21), with promising results for disease classification (sensitivity and specificity range = 80–95%). Thereby, these systems offer great promise to improve the accessibility and cost-effectiveness of ocular disease screening in developing countries.

Despite this, most previous systems could only detect a single ocular disorder, thus would omit severe blinding eye diseases. In addition, previous studies have evaluated on retrospective datasets, and there is a paucity of data directly comparing the performance of deep learning system (DLS) capable to detect common blindness diseases to that of general ophthalmologists or non-physician graders. Given the fact that in real world screening programs, human graders or general ophthalmologists may also make mistakes, a robust study to directly compare DLS and general ophthalmologists or non-physician graders is of paramount importance for healthcare decision makers and patients to make informed decisions relating to the deployment of these systems.

Therefore, in the present study, we investigated the diagnostic agreement between ophthalmologists with varying levels of experience, non-physician graders, and validated deep learning models (22) for DR, GON, and AMD on an independent dataset in China.

Abbreviations: DLS, Deep learning systems; DR, Diabetic retinopathy; AMD, Age related macular degeneration; GON, Glaucomatous optic neuropathy; AUC, Area under receiver operator characteristic curve; GONE, Glaucomatous optic neuropathy evaluation; VCDR, Vertical cup to disc ratio; RNFL, Retinal nerve fiber layer; NHS, English national health screening; DESP, Diabetic eye screening program; DME, Diabetic macular edema.

Methods

This study was approved by the Institutional Review Board of the Zhongshan Ophthalmic Center, China (2017KYPJ049) and conducted in accordance with the Declaration of Helsinki. All graders and ophthalmologists have been informed that their data will be compared with the DLS. Informed consent for the use of fundus photographs was not required as images were acquired retrospectively and were fully anonymized.

Test set development, reference standard, and definitions

A total of 300 fundus photographs were randomly selected from a subset of 42,388 independent gradable images from the online LabelMe dataset (http://www.labelme.org, Guangzhou, China) (22, 23). The LabelMe dataset includes images from 36 hospital ophthalmology departments, optometry clinics, and screening settings in China that include various kinds of eye diseases, such as DR, glaucoma, and AMD. The data will be available upon request. Retinal photographs were captured using a variety of common conventional desktop retinal cameras, including Topcon, Canon, Heidelberg, and Digital Retinography System. The LabelMe dataset was graded for DR, GON, and AMD by 21 ophthalmologists who previously achieved an unweighted kappa of ≥0.70 (substantial) on a test set of images. Images were randomly assigned to a single ophthalmologist for grading and were returned to the pooled dataset until three consistent grading outcomes were achieved. Once an image was given a reference standard label it was removed from the grading dataset. This process has been described in detail elsewhere (22, 23).

Stratified random sampling was used to select 50 images of each disease category and an additional 150 images classified as normal or a disease other than DR, AMD, and GON. Poor quality images (defined as ≥50% of the fundus photograph area obscured) were excluded. Images that were included in the training and internal validation datasets of the deep learning models were not eligible for inclusion. Following the selection of images, experienced retinal ($n = 5$) specialists independently labeled all 300 images to establish a reference standard for DR and AMD. Similarly, glaucoma specialists ($n = 5$) independently graded all images to determine the GON reference standard. Specialists were blinded to any previous medical history or retinal diagnosis for the included images. Once all images were graded, they were converted to a two-level classification for each disease: non-referable and referable. Each image was only assigned a

conclusive label if more than 50% of the specialists reported a consistent grading outcome.

A website[1] was developed to allow human graders to log in and interpret images. Diabetic retinopathy severity was classified as none, mild non-proliferative DR (NPDR), moderate NPDR, severe NPDR, and proliferative DR using the International Clinical Diabetic Retinopathy scale (24). Diabetic macular edema (DME) was defined as any hard exudates within one-disk diameter of the fovea or an area of hard exudates in the macular area at least 50% of the disk area (25). Referable DR was defined as moderate NPDR or worse with or without the presence of DME. The severity of AMD was graded according to the clinical classification of AMD, which has been described elsewhere (26). For the purpose of this study, referable AMD was defined as late wet AMD as it was the only subtype of AMD that could be managed with effective therapy currently. Glaucomatous optic neuropathy was classified as absent or referable GON according to definitions utilized by previous population-based studies (27–29). The definition of referable GON included the presence of any of the following: vertical cup to disk ratio (VCDR) $\geq$0.7; rim width $\leq$0.1 disk diameter; localized notches; and presence of retinal nerve fiber layer (RNFL) defect and/or disk hemorrhage.

Development of the deep learning system

The development and validation of the DR, GON, and AMD models have been described in detail elsewhere (22, 30–32). In brief, referable GON, DR, and AMD deep learning algorithms were developed using a total of 210,473 fundus photographs (referable DR, 106,244; referable GON, 48,116; referable AMD 56,113). Several pre-processing steps were performed for normalization to control for variations in image size and resolution. This included augmentation to enlarge heterogeneity, applying local space average color for color constancy and downsizing image resolution to 299 × 299 pixels (33). Finally, eight convolutional neural networks were contained within the DLS (Version 20,171,024), all adopting Inception-v3 architecture (34). The development of the networks was described in our previous studies (22, 23, 32). Briefly, the networks were downsized to 299 × 299, and local space average color and data augmentation were adopted. These networks were trained from scratch and included (1) classification for referable DR, (2) classification of DME, (3) classification of AMD, (4) classification of GON, and (5) assessment of the availability of the macular region and rejection of non-retinal photographs.

Graders and ophthalmologists identification and recruitment

Five trained non-physician graders, who also previously received training for DR, AMD, and GON classification, usually graded images from 50 to 100 participants for common blindness diseases every workday and underwent tests per quarter, from Zhongshan Ophthalmic Center Image Grading Center with National Health Screening (NHS) DR grader certification were recruited to grade all these images.

We also invited general ophthalmologists from four provincial hospitals and five county hospitals in seven provinces in China (Guangdong, Guangxi, Fujian, Jiang Su, Yunnan, Xinjiang, and Inner Mongolia province). General ophthalmologists who had at least 3 years clinical practice including residency were eligible to participate.

Selected ophthalmologists were sent an invitation to participate *via* email or mobile phone text message. Those who did not respond were followed up with a telephone call. The clinical practice characteristics of invited ophthalmologists were obtained from publicly available resources or personally *via* telephone.

Of the 330 ophthalmologists who were eligible to participate, 66 (20%) were randomly selected and subsequently invited to participate in the study. Nineteen ophthalmologists (28.8%) declined or did not respond and 47 ophthalmologists (71.2%) agreed to participate. A flow chart outlining the recruitment of ophthalmologists is shown in Figure 1.

Test set implementation

Participants independently reviewed all 300 images in a random order. They were blinded to the reference standard and the grades assigned by other participants. Due to the variability in existing classification criteria for GON, a standardized grading criteria was provided to all participants. Participants were not provided with details of the comprehensive grading criterion utilized for the grading of DR and AMD, as it was assumed that the participants' experience would be sufficient to enable them to classify these disorders into the specific categories (DR: mild, moderate, severe NPDR and proliferative DR; AMD: early or moderate AMD, late dry AMD, and late wet AMD). There was no time limit for the interpretation of each image. All grading results were converted to a two-level classification for each disease (referable and non-referable disorders) and then compared against the reference standard. The eight deep learning models were also tested using the same images.

In order to characterize the features of misclassified images by DLS and human graders, an experienced ophthalmologist (Z.X.L.) reviewed misclassified fundus photographs and classified them into categories arbitrarily developed by a consensus meeting by investigators.

Statistical analysis

The area under the receiver operating characteristic curve (AUC), rate of agreement and unweighted kappa were calculated. Agreement was defined as the proportion of images that were correctly classified by participants or the DLS models using the gold standard label as a reference standard. Firstly, data from all participants were used and in this situation, the CIs accounting for within and between subject variability by estimating the variance using the form; $\{\text{var.}(\text{parameter}_p) + [\text{avg}(\text{parameter}_p) \times (1-\text{avg}(\text{parameter}_p))]/n_c\}/n_p$, where $\text{avg.}(\text{parameter}_p)$ denotes the average corresponding parameter (AUC, agreement rate or kappa) among participants, $\text{var.}(\text{parameter}_p)$ denotes the sample variance of parameter among participants, n_c denotes the number of images interpreted by each participant, and n_p denotes the number of participants.

1 http://v.labelme.org

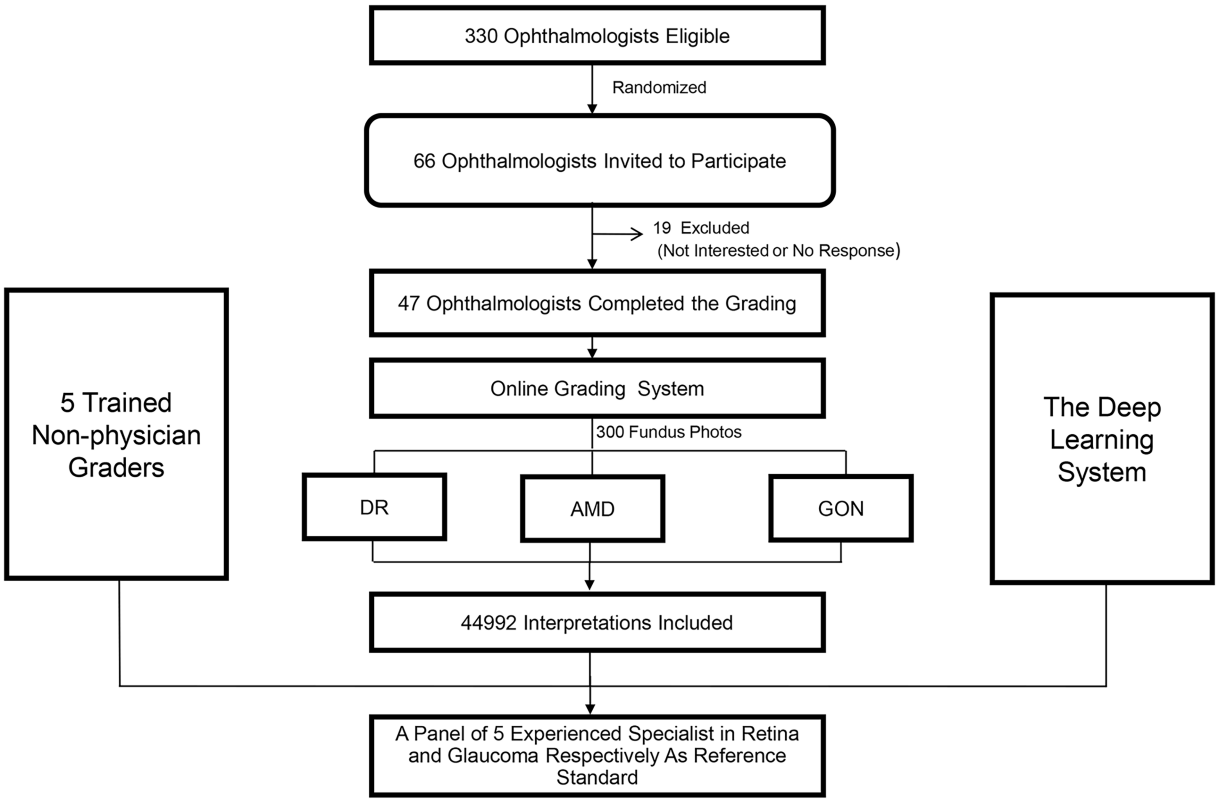

FIGURE 1
Recruitment, workflow, and grading of ophthalmologists and non-physician graders.

Then, a representative grading result for graders and ophthalmologists was made when more than 50% of group members achieved consistent grading outcomes. As the DLS can generate a continuous probability between 0 and 1 for referable disorders, AUC for DLS was calculated using these continuous probabilities to compared with reference standard, whereas the agreement rate and unweighted kappa were dichotomized by assigning a certain probability when reaching the highest accuracy. The AUCs of graders, ophthalmologists, and DLS were calculated by comparing with reference standard for two-level classification (referable and non-referable).

We investigated the extent to which the clinical experience of ophthalmologists was associated with agreement. Logistic regression models of ophthalmologist agreement that simultaneously incorporated several ophthalmologist characteristics (hospital level, academic affiliation, clinical practice years, and clinical expertise) were modeled. Non-physician graders were not included in this analysis due to the relatively small sample size ($n = 5$).

Sensitivity analyses was used to explore whether the grading results would change by using an alternate reference standard instead of the specialist-derived standard. Firstly, cases where the reference standard was different from the most frequent (≥80.0%) grading result of the participants were identified (8 of 300 images). Then, the results were reanalyzed by substituting the most frequent grading outcome of participants as the reference standard for the eight images, or just excluding the eight images. A p value of less than 0.05 was regarded as statistically significant. Stata statistical software (version 14; College Station, Texas, United States) was used.

Results

Reference dataset

Of the 300 images included in the dataset, the total number of images labeled as referable DR, AMD, and GON according to the final specialist grading were 45 (15.0%), 46 (15.3%), and 46 (15.3%), respectively. The remaining 163 (54.4%) images were classified as normal or a disease other than DR, AMD, and GON.

Graders and ophthalmologists characteristics

The five trained non-physician graders were all females with a mean age of 30.4 ± 2.2 years (range, 27–34 years) and an average of 3.6 ± 0.6 years (range, 2–5 years) of grading experience in DR screening support and research image grading. There were 6, 23, 12, and 6 general ophthalmologists aged <30, 30–40, 40–50, and ≥50 years, respectively. Among these ophthalmologists, there were 22 males and 25 females. Twenty-seven were from affiliated hospitals and the other were from nonaffiliated hospitals. Their lengths of clinical practice were 5 years ($n = 13$), 5–10 years ($n = 16$), and ≥10 years ($n = 18$).

Diagnostic agreement among deep learning models, trained non-physician graders, and ophthalmologists

Table 1 displays the agreement distribution by individual grading outcomes of specialists performing initial reference standard grading compared to the final reference standard. The overall agreement rate of the initial independent specialist diagnoses was 96.5% for referable DR, 98.1% for referable AMD, and 92.8% for referable GON.

Table 2 provides a comparison between the DLS and general ophthalmologists. The sensitivity and specificity of the DLS for referable DR were 97.8% (44/45) and 92.5% (236/255), respectively. The results for general ophthalmologists for referable DR were 91.1% (41/45) and 99.6% (254/255), respectively.

TABLE 1 Comparison of the five specialist ophthalmologist's independent gradings vs. final expert consensus reference standard for 300 fundus photographs.[a]

	Specialist ophthalmologists independent gradings			
Final reference standard	Absent	Present	Missing	Total
Referable DR[b]				
Absent	1,269	6	0	1,275
Present	45	178	2	225
Total	1,314	184	2	1,500
Late wet AMD[c]				
Absent	1,258	12	0	1,270
Present	16	214	0	230
Total	1,274	226	0	1,500
Referable GON[d]				
Absent	1,176	94	0	1,270
Present	14	216	0	230
Total	1,190	310	0	1,500

[a]The overall all agreement rate for referable DR, late wet AMD, and GON were 96.5, 98.1, and 92.8%, respectively.
[b,c]The members to make reference standard were consisted of five retina specialists, and each disorder was graded for multiple categories and then converted to two levels for analysis.
[d]The members were consisted of five glaucoma specialists.
DR, diabetic retinopathy; AMD, age-related macular degeneration; GON, glaucomatous optic neuropathy.

Table 3 compares the grading agreement of trained non-physician graders, ophthalmologists, and the DLS versus the reference standard. There were no significant differences in the AUC of non-physician graders, general ophthalmologists with different levels of clinical experience, and the DLS for the interpretation of referable DR ($p = 0.415$, compared with expert consensus reference diagnosis) and referable AMD ($p = 0.145$, compared with expert consensus reference diagnosis). For the classification of GON, the DLS achieved a superior AUC result compared to non-physician graders ($p < 0.001$).

Ophthalmologist characteristics related with image interpretation agreement

The agreement between general ophthalmologists' image grading and the reference standard is shown in Table 4. Table 4 shows that the overall agreement was higher for referable DR in ophthalmologists with greater clinical experience ($p = 0.009$) and those who were specialists ($p = 0.040$). Agreement was significantly higher for referable AMD in ophthalmologists from provincial level hospitals ($p = 0.017$), adjunct academic affiliations ($p = 0.002$), ophthalmologists with more years of clinical practice ($p = 0.009$), and those who were glaucoma or retinal specialist ophthalmologists ($p = 0.006$). Similarly, the level of agreement for referable GON was greater among ophthalmologists from provincial level hospitals ($p < 0.001$), those from adjunct academic affiliations ($p < 0.001$), those with more years of clinical experience ($p < 0.001$) and those who were glaucoma or retinal specialist ophthalmologists ($p < 0.001$).

Image disagreement characteristics

The interpretations of non-physician graders, ophthalmologists, and the DLS compared with the reference standard for each of the 300 fundus photographs for diabetic retinopathy are shown in Figure 2. This figure also demonstrates that several images caused mistakes common to nonphysician graders, ophthalmologists, and the DLS; for example, images #1 and #87 triggered consistent false positives. In the same way, images #71, #97, #140, #181, #232, and #239 displayed consistent false negatives. These images are shown in Figure 3. The general features of images that were misclassified by human participants (trained non-physician graders and ophthalmologists) are summarized in Table 5.The primary reason

TABLE 2 Comparison of deep learning system and general ophthalmologists to the expert consensus reference standard.

	Reference standard	Deep learning system			Ophthalmologists		
		Agreement (%)	Misclassification (%)	Total	Agreement (%)	Misclassification (%)	Total
Diabetic Retinopathy	Referable	44 (97.8)	1 (2.2)	45	41 (91.1)	4 (8.9)	45
	Non-referable	236 (92.5)	19 (7.5)	255	254 (99.6)	1 (0.4)	255
Age related macular degeneration	Referable	39 (83.0)	8 (7.0)	47	43 (91.5)	4 (8.5)	47
	Non-referable	245 (96.8)	8 (3.2)	253	248 (98.0)	5 (2.0)	253
Glaucomatous optic neuropathy	Referable	45 (97.8)	1 (2.2)	46	42 (91.3)	4 (8.7)	46
	Non-referable	252 (99.2)	2 (0.8)	254	249 (98.0)	5 (2.0)	254

A representative grading result for graders and ophthalmologists were made when more than 50% of group members achieved a consistent grading.

TABLE 3 Agreement of image interpretation by trained non-physician graders, general ophthalmologists, and deep learning system versus the expert consensus reference standard.[a]

	Trained non-physician graders (95% CI)	Ophthalmologists (95% CI)				Deep learning system[a] (95% CI)	*p* value
		Clinical experience 3–5years	Clinical experience 5–10years	Clinical experience >10years	Total		
Referable DR							
Model 1							
AUC	0.984 (0.960–1.000)	0.964 (0.926–1.000)	0.965 (0.927–1.000)	0.954 (0.911–0.996)	0.954 (0.911–0.995)	0.990 (0.982–0.999)	0.415
Kappa	0.959 (0.845–1.000)	0.946 (0.832–1.000)	0.947 (0.834–1.000)	0.933 (0.820–1.000)	0.933 (0.820–1.000)	0.775 (0.665–0.886)	
Agreement rate	0.989 (0.971–0.998)	0.983 (0.961–0.996)	0.987 (0.966–0.996)	0.983 (0.961–0.995)	0.983 (0.962–0.995)	0.933 (0.899–0.959)	
Referable AMD							
Model 1							
AUC	0.912 (0.859–0.964)	0.933 (0.887–0.979)	0.946 (0.904–0.987)	0.958 (0.922–0.995)	0.948 (0.906–0.989)	0.945 (0.903–0.986)	0.145
Kappa	0.823 (0.710–0.936)	0.851 (0.738–0.964)	0.876 (0.762–0.989)	0.901 (0.788–1.000)	0.887 (0.774–1.000)	0.798 (0.685–0.911)	
Agreement rate	0.953 (0.923–0.974)	0.960 (0.931–0.979)	0.967 (0.940–0.983)	0.973 (0.948–0.988)	0.970 (0.944–0.986)	0.947 (0.915–0.969)	
Referable GON							
Model 1							
AUC	0.675 (0.604–0.746)	0.862 (0.797–0.926)	0.894 (0.836–0.953)	0.976 (0.946–1.000)	0.953 (0.911–0.994)	0.994 (0.988–0.999)	<0.001
Kappa	0.445 (0.341–0.549)	0.779 (0.666–0.891)	0.825 (0.712–0.938)	0.961 (0.848–1.000)	0.922 (0.809–1.00)	0.926 (0.813–1.00)	
Agreement rate	0.887 (0.845–0.920)	0.947 (0.914–0.969)	0.957 (0.927–0.977)	0.990 (0.971–0.998)	0.980 (0.957–0.993)	0.980 (0.956–0.993)	

DR, diabetic retinopathy; AMD, age-related macular degeneration; GON, glaucomatous optic neuropathy; AUC, area under receiver operator characteristic curve; CI, confidence interval.
[a]The AUC, kappa, and agreement rate of graders and ophthalmologists were calculated using a representative grading result for each group when there was at least 50% of group members reached consistent grading.

for false negative of referable DR was the presence of DME ($n = 10$, 58.9%), while two cases (100.0%) with microaneurysm/s and artifacts resulted in false positive by human participants. For referable AMD, false negative cases were mostly related to the presence of subtle subretinal hemorrhage ($n = 6$, 50.0%). False positives resulted from misclassification of earlier forms of AMD ($n = 9$, 75.1%). Among human participants, the most common reason for false negative of referable GON were those images with borderline VCDR ($n = 8$, 27.7%), while false positives occurred in those images which displayed physiological cupping ($n = 14$, 93.3%).

One fundus image demonstrated coexisting intraretinal microvascular abnormality and DME that were not identified by the DLS. The most common reason for false positives by the DLS was the presence of microaneurysm/s only ($n = 10$, 55.5%; Table 6). For referable AMD, the presence of subretinal hemorrhage ($n = 5$, 71.4%) was the primary reason for false negative and other diseases ($n = 7$, 87.5%) including DR or GON. For referable GON, the DLS under-interpreted one image with VCDR less than 0.7, while two images with physiological large cupping ($n = 2$, 40%) and three images with other diseases ($n = 3$, 60%) were incorrectly classified as positive.

Discussion

In this study, we prospectively compared the diagnostic agreement of trained non-physician graders and ophthalmologists using three validated deep learning models for the detection of referable DR, late wet AMD, and GON from color fundus photographs. Our results suggest that the performance of the deep learning models for referable DR and AMD are comparable to non-physician graders and ophthalmologists. As for referable GON, the DLS outperformed non-physician graders.

There was no difference among the non-physician graders, ophthalmologists with different years of clinical practice, and the DLS for the diagnostic accuracy of referable DR. The non-physician graders included in this study all had grader certification from the NHS DR screening program, underwent regular assessments every month, and routinely interpreted fundus photographs of diabetic patients from nationwide screening programs, which may explain their relatively high agreement compared to the gold standard. While the DLS also exhibited comparably good performance when compared with non-physician graders and general ophthalmologists.

Comparison of the DLS with general ophthalmologists found that the DLS had higher sensitivity (97.8 vs. 91.1%) and lower specificity (92.5 vs. 99.6%) for the classification of referable DR. However, nearly half of the false positive cases identified by the DLS included ($n = 8$, 44.5%) other disorders, for example, late wet AMD and retinal degeneration. The remaining false positive images ($n = 10$, 55.5%) had mild NPDR. Those images identified as false positive by the DLS would receive a referral and be identified during confirmatory examination conducted by a specialist.

Previous studies have shown that the majority of referral cases for DR (73%) are as a result of DME (35). There are 100 million patients with DR worldwide which corresponds to 7.6 million DME patients (36). However, our results showed that images that were characterized

TABLE 4 Ophthalmologist characteristics for image interpretation versus expert consensus reference standard.

Characteristics	Referable diabetic retinopathy				Referable age-related macular degeneration				Referable glaucomatous optic neuropathy			
	n	AUC (95% CI)	Agreement rate (95% CI)	*p*	*n*	AUC (95% CI)	Agreement rate (95% CI)	*p*	*n*	AUC (95% CI)	Agreement rate (95% CI)	*p*
Hospital												
County level (*n* = 20)	5,794	0.929 (0.929–0.930)	0.955 (0.955–0.956)		5,878	0.871 (0.871–0.872)	0.929 (0.929 0.930)		5,868	0.818 (0.818–0.820)	0.903 (0.902–0.903)	
Provincial level (*n* = 27)	7,894	0.932 (0.931–0.932)	0.956 (0.956–0.957)		7,971	0.872 (0.871–0.872)	0.929 (0.929–0.930)		8,030	0.875 (0.875–0.876)	0.933 (0.932–0.933)	
				0.808[a]				0.017[a]				<0.001[a]
Academic affiliation												
None (*n* = 18)	5,196	0.926 (0.925–0.926)	0.954 (0.954–0.955)		5,281	0.867 (0.867–0.868)	0.926 (0.926–0.927)		5,269	0.810 (0.810–0.811)	0.897 (0.897–0.898)	
Adjunct affiliation (*n* = 29)	8,492	0.934 (0.934–0.935)	0.957 (0.957–0.958)		8,568	0.891 (0.891–0.892)	0.941 (0.941–0.942)		8,629	0.877 (0.876–0.878)	0.934 (0.934–0.935)	
				0.343[b]				0.002[b]				<0.001[b]
Clinical practice (yrs)												
≤5 (*n* = 13)	3,718	0.925 (0.924–0.925)	0.951 (0.950–0.951)		3,780	0.875 (0.875–0.876)	0.928 (0.927–0.928)		3,782	0.806 (0.805–0.807)	0.569 (0.892–0.893)	
5–10 (*n* = 16)	4,637	0.929 (0.928–0.929)	0.953 (0.953–0.954)		4,703	0.876 (0.876–0.877)	0.934 (0.934–0.935)		4,743	0.848 (0.847–0.849)	0.919 (0.919–0.920)	
>10 (*n* = 18)	5,333	0.937 (0.937–0.938)	0.839 (0.838–0.840)		5,366	0.892 (0.891–0.892)	0.942 (0.942–0.943)		5,373	0.887 (0.886–0.888)	0.941 (0.940–0.041)	
				0.009[c]				0.009[c]				<0.001[c]
Expertise in ophthalmology												
Nonexpert (*n* = 27)	7,797	0.929 (0.929–0.930)	0.953 (0.953–0.954)		7,919	0.873 (0.873–0.874)	0.930 (0.930–0.931)		7,934	0.817 (0.816–0.817)	0.902 (0.901–0.902)	
Expert (*n* = 20)	5,891	0.933 (0.933–0.934)	0.960 (0.960–0.961)		5,930	0.894 (0.894–0.895)	0.942 (0.942–0.943)		5,964	0.898 (0.898–0.899)	0.944 (0.944–0.945)	
				0.040[d]				0.006[d]				<0.001[d]

[a]A test for trend based on logistic regression model which diagnostic agreement for corresponding disorder was considered as the outcome variable and a two-category variable for hospital level was regarded as independent variable.
[b]A test for trend based on logistic regression model which diagnostic agreement for corresponding disorder was considered as the outcome variable and a two-category variable for whether to be an adjunct affiliation was regarded as independent variable.
[c]A test for trend based on logistic regression model which diagnostic agreement for corresponding disorder was considered as the outcome variable and a three-category variable for clinical practice years was regarded as independent variable.
[d]A test for trend based on logistic regression model which diagnostic agreement for corresponding disorder was considered as the outcome variable and a two-category variable for expertise in ophthalmology was regarded as independent variable.
CI, confidence interval.

10.3389/fmed.2023.1115032

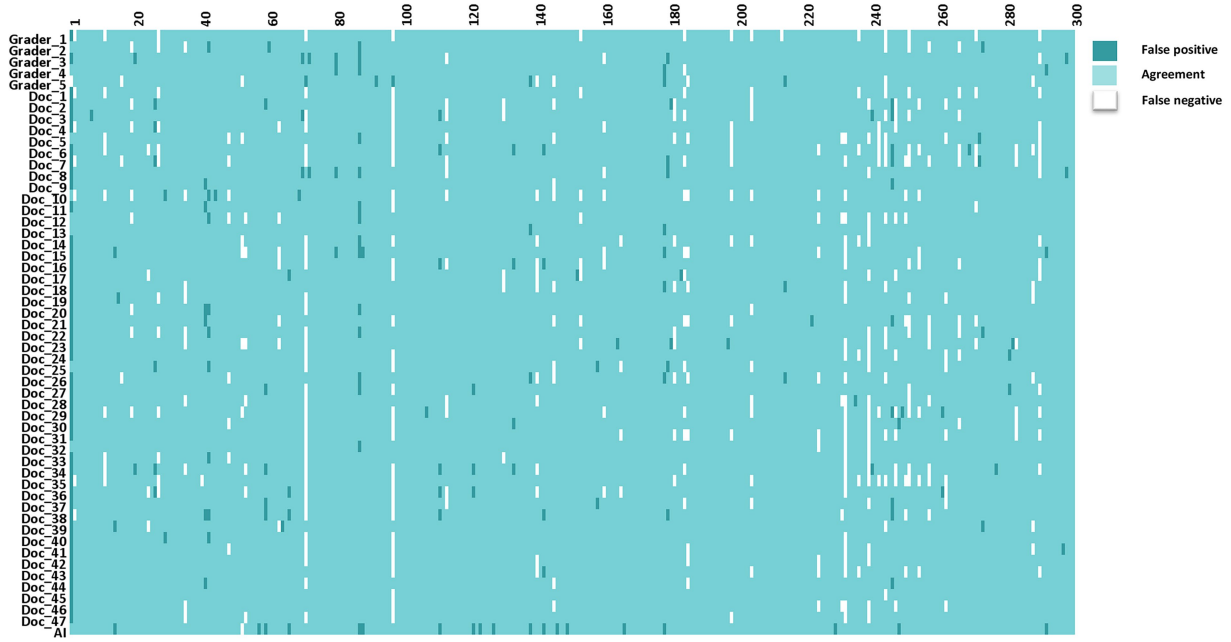

FIGURE 2
The interpretations of graders, ophthalmologists, and artificial intelligence compared with the reference standards for each of the 300 fundus photographs for diabetic retinopathy.

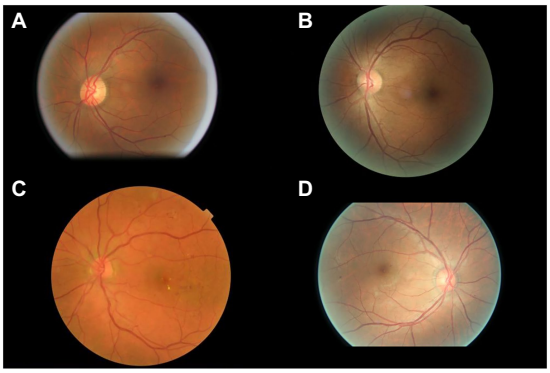

FIGURE 3
Sample images consistently misclassified by human participants. **(A,B)** Images with only microaneurysm misclassified as referable diabetic retinopathy. **(C)** Images of diabetic macular edema misclassified as non-referable diabetic retinopathy. **(D)** Microaneurysm and dot hemorrhage misclassified as non-referable diabetic retinopathy.

as DME ($n = 10$, 58.9%) were under interpreted by human graders more often than other DR lesions. DR changes related to DME displayed considerable variation among graders and ophthalmologists, with an overall agreement rate of 71% when compared with the reference standard. Therefore, the importance of not overlooking the diagnosis of DME among graders and ophthalmologists should be emphasized.

The DLS outperformed non-physician graders in the classification of referable GON in this study. The variability in inter-assessor agreement among non-physician graders and ophthalmologists for the classification of ocular disorders is well known, especially glaucoma (37, 38). The Glaucomatous optic neuropathy evaluation (GONE) project previously reported that ophthalmology trainees underestimated glaucoma likelihood in 22.1% of optic disks and overestimated 13.0% of included optic disks. This has been similar in our study where general ophthalmologists underestimated 23.8% and underestimated 8.9% of included optic disks (37). Furthermore, Breusegem et al. (38) reported that non-expert ophthalmologists had significantly lower accuracy compared with experts in the diagnosis of glaucoma. Our results are in agreement with previous studies and showed that ophthalmologists with more clinical experience and specialist training in ophthalmology achieve higher inter-assessor agreement. The experience and knowledge obtained through years of clinical practice is likely to play a significant role in interpretation and performance accuracy. In contrast, the DLS is easily able to adopt labels from experienced ophthalmologists to learn the most representative characteristics of GON. Fundus photography is an important method to evaluate GON, however, the diagnosis of glaucoma requires the results of visual field analysis, optical coherence tomography, and intra ocular pressure measurements to make an accurate diagnosis. Thus, further studies to compare DLS with ophthalmologists using multi-modality clinical data is warranted.

The main strength of our study was to prospectively compare the performance of a DLS for the detection of three common blinding eye diseases to non-physician graders and ophthalmologists of varying levels of experience and with different specialties. Our study is also distinctly different from previous reports (19, 39–42). First, we evaluated three ocular diseases at the same time. Second, no prospective comparison of ophthalmologists with varying levels of clinical experience and trained non-physician graders with a DLS for common ocular disorders has been reported. Previous authors have compared the performance of the DLS with that of graders or specialists; this is often considered the gold standard for the development of the DLS (39,

TABLE 5 Characteristics of the disagreement images by human participants.[a]

Reason	No.	Proportion (%)
Referable DR		
False negative		
MA, hemorrhage, DME	10	58.9
Dot hemorrhage, MA	4	23.5
MA, hemorrhage, HEs, CWS	3	17.6
Subtotal	17	100.0
False positive		
Microaneurysm/s, Artifacts	1	100.0
Subtotal	1	100.0
Referable AMD		
False negative		
Subretinal Hemorrhage	6	50.0
Sub-retinal/Sub-RPE fibrovascular proliferation	3	25.0
Serous detachment of the sensory retina or RPE	3	25.0
Sub-total	12	100.0
False positive		
Other macular degeneration	9	75.1
Myopic maculopathy	1	8.3
Choroidal osteoma	1	8.3
Other diseases (Pre-macular hemorrhage)	1	8.3
Sub-total	12	100.0
Referable GON		
False negative		
Borderline VCDR	8	27.7
Borderline VCDR with RNFL defect	6	20.7
Optic disk with tilt or rotation	5	17.2
With other diseases	3	10.3
Rim < 0.1	3	10.3
Notch	2	6.9
Linear hemorrhage around optic disk	2	6.9
Sub-total	29	100.0
False positive		
Physiological large cupping (0.5 ≤ VCDR < 0.7)	14	93.3
Juxtapapillary capillary hemangioma	1	7.7
Sub-total	15	100.0

[a]The cases included in this analysis were those with more than 20% of the individual human participants (graders and ophthalmologists) inconsistent with the reference standard.
DR, diabetic retinopathy; MA, microaneurysm; HEs, hard exudates; CWS, cotton-wool spot; DME, diabetic macular edema; AMD, age-related macular degeneration; RPE, retina pigment epithelium; and GON, glaucomatous optic neuropathy; VCDR, vertical cup to disc ratio; RNFL, retinal nerve fiber layer.

TABLE 6 Characteristics of the disagreement images by deep learning system.

Reason	No.	Proportion (%)
Referable DR		
False negative		
MA, IRMA, DME	1	100.0
Sub-total	1	100.0
False positive		
MA only	10	55.5
Other diseases		
Late wet AMD	4	22.2
Retinal degeneration	3	16.7
RVO	1	5.6
Subtotal	18	100.0
Referable AMD		
False negative		
Subretinal hemorrhage	5	71.4
Serous detachment of the sensory retina or RPE	2	28.6
Subtotal	7	100.0
False positive		
Other diseases		
DR	7	87.5
GON	1	12.5
Subtotal	8	100.0
Referable GON		
False negative		
VCDR < 0.7 with notch	1	100.0
Sub-total	1	100.0
False positive		
Physiologic large cupping (0.5 ≤ VCDR < 0.7)	2	40.0
Other diseases		
AMD	2	40.0
Juxtapapillary capillary hemangioma	1	20.0
Subtotal	5	100.0

DR, diabetic retinopathy; MA, microaneurysm; IRMA, intra-retinal microvascular abnormality; DME, diabetic macular edema; AMD, age-related macular degeneration; VRO, retinal vein occlusion; RPE, retina pigment epithelium; and GON, glaucomatous optic neuropathy.

41, 43). Non-physician graders and ophthalmologists are susceptible to making diagnostic mistakes. Our study included independent graders and ophthalmologists to evaluate the performance of the DLS. Therefore, the current study will provide information on the accuracy of the DLS, as well as a more comprehensive understanding and acceptance of how AI systems might work or contribute.

There are several limitations of this study which warrant further consideration. On one hand, human participants included in this study

were recruited from China. This has the potential to affect the generalizability of these results to other human graders, especially those in developed countries. In the future, similar studies should be attempted in other countries with different physician or specialist training system. On the other hand, the use of single-field, non-stereoscopic fundus photographs without the inclusion of optical coherence tomography may lead to a reduced sensitivity for DR and particularly DME detection for human participants and the DLS.

In conclusion, our DLS demonstrated sufficient agreement with non-physician graders and general ophthalmologists when compared to the reference standard diagnosis agreement for referable DR and AMD. The DLS performance was better than non-physician graders and ophthalmologists with ≤10 years of clinical experience for referable GON. Further investigation is required to validate the performance in real-world, clinical settings which display the full spectrum and distribution of lesions and manifestations encountered in clinical practice.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by the Institutional Review Board of the Zhongshan Ophthalmic Center, China. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

ZL and MH were involved in the concept, design, and development of the deep learning algorithm. ZL, XG, JZ, XL, RC, and MH contributed to the acquisition, analysis, and interpretation of data. ZL wrote the manuscript. All authors revised and edited the manuscript. MH is the guarantor of this work and as such has full access to all the data in the study and takes responsibility for data integrity and the accuracy of the data analysis. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by National Key R&D Program of China (2018YFC0116500), the Fundamental Research Funds of the State Key Laboratory in Ophthalmology, National Natural Science Foundation of China (81420108008), and Science and Technology Planning Project of Guangdong Province (2013B20400003).

Acknowledgments

The authors wish to thank the staffs from the Image Grading Center, Zhongshan Ophthalmic Center, Sun Yat-Sen University for their grading assistance during this project.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Cheung, N, Mitchell, P, and Wong, TY. Diabetic retinopathy. *Lancet*. (2010) 376:124–36. doi: 10.1016/S0140-6736(09)62124-3

2. Bressler, NM. Age-related macular degeneration is the leading cause of blindness. *JAMA*. (2004) 291:1900–1. doi: 10.1001/jama.291.15.1900

3. Pascolini, D, Mariotti, SP, Pokharel, GP, Pararajasegaram, R, Etya'ale, D, Négrel, AD, et al. 2002 global update of available data on visual impairment: a compilation of population-based prevalence studies. *Ophthalmic Epidemiol*. (2004) 11:67–115. doi: 10.1076/opep.11.2.67.28158

4. Stevens, GA, White, RA, Flaxman, SR, Price, H, Jonas, JB, Keeffe, J, et al. Global prevalence of vision impairment and blindness: magnitude and temporal trends, 1990–2010. *Ophthalmology*. (2013) 120:2377–84. doi: 10.1016/j.ophtha.2013.05.025

5. Bourne, RR, Stevens, GA, White, RA, Smith, JL, Flaxman, SR, Price, H, et al. Causes of vision loss worldwide, 1990-2010: a systematic analysis. *Lancet Glob Health*. (2013) 1:e339–49. doi: 10.1016/S2214-109X(13)70113-X

6. Tham, YC, Li, X, Wong, TY, Quigley, HA, Aung, T, and Cheng, CY. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology*. (2014) 121:2081–90. doi: 10.1016/j.ophtha.2014.05.013

7. Frick, KD, and Foster, A. The magnitude and cost of global blindness: an increasing problem that can be alleviated. *Am J Ophthalmol*. (2003) 135:471–6. doi: 10.1016/S0002-9394(02)02110-4

8. Armstrong, KL, Jovic, M, Vo-Phuoc, JL, Thorpe, JG, and Doolan, BL. The global cost of eliminating avoidable blindness. *Indian J Ophthalmol*. (2012) 60:475–80. doi: 10.4103/0301-4738.100554

9. Pizzarello, L, Abiose, A, Ffytche, T, Duerksen, R, Thulasiraj, R, Taylor, H, et al. VISION 2020: the right to sight: a global initiative to eliminate avoidable blindness. *Arch Ophthalmol*. (Chicago, Ill: 1960). (2004) 122:615–20. doi: 10.1001/archopht.122.4.615

10. Tapp, RJ, Shaw, JE, Harper, CA, de Courten, MP, Balkau, B, McCarty, DJ, et al. The prevalence of and factors associated with diabetic retinopathy in the Australian population. *Diabetes Care*. (2003) 26:1731–7. doi: 10.2337/diacare.26.6.1731

11. Weih, LM, Nanjan, M, McCarty, CA, and Taylor, HR. Prevalence and predictors of open-angle glaucoma: results from the visual impairment project. *Ophthalmology*. (2001) 108:1966–72. doi: 10.1016/S0161-6420(01)00799-0

12. Subburaman, GB, Hariharan, L, Ravilla, TD, Ravilla, RD, and Kempen, JH. Demand for tertiary eye Care Services in Developing Countries. *Am J Ophthalmol*. (2015) 160:619–627.e1. doi: 10.1016/j.ajo.2015.06.005

13. Scanlon, PH. The english national screening programme for diabetic retinopathy 2003–2016. *Acta Diabetol*. (2017) 54:515–25. doi: 10.1007/s00592-017-0974-1

14. Klein, R, Klein, BE, Neider, MW, Hubbard, LD, Meuer, SM, and Brothers, RJ. Diabetic retinopathy as detected using ophthalmoscopy, a nonmydriatic camera and a standard fundus camera. *Ophthalmology*. (1985) 92:485–91. doi: 10.1016/S0161-6420(85)34003-4

15. Chan, HH, Ong, DN, Kong, YX, O'Neill, EC, Pandav, SS, Coote, MA, et al. Glaucomatous optic neuropathy evaluation (GONE) project: the effect of monoscopic versus stereoscopic viewing conditions on optic nerve evaluation. *Am J Ophthalmol*. (2014) 157:936–944.e1. doi: 10.1016/j.ajo.2014.01.024

16. LeCun, Y, Bengio, Y, and Hinton, G. Deep learning. *Nature*. (2015) 521:436–44. doi: 10.1038/nature14539

17. Hassan, SS, Bong, DB, and Premsenthil, M. Detection of neovascularization in diabetic retinopathy. *J Digit Imaging*. (2012) 25:437–44. doi: 10.1007/s10278-011-9418-6

18. Abramoff, MD, Lou, Y, Erginay, A, Clarida, W, Amelon, R, Folk, JC, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Invest Ophthalmol Vis Sci*. (2016) 57:5200–6. doi: 10.1167/iovs.16-19964

19. Chakrabarty, L, Joshi, GD, Chakravarty, A, Raman, GV, Krishnadas, SR, and Sivaswamy, J. Automated detection of glaucoma from topographic features of the optic nerve head in color fundus photographs. *J Glaucoma*. (2016) 25:590–7. doi: 10.1097/IJG.0000000000000354

20. Issac, A, Partha Sarathi, M, and Dutta, MK. An adaptive threshold based image processing technique for improved glaucoma detection and classification. *Comput Methods Prog Biomed*. (2015) 122:229–44. doi: 10.1016/j.cmpb.2015.08.002

21. Zheng, Y, Hijazi, MH, and Coenen, F. Automated "disease/no disease" grading of age-related macular degeneration by an image mining approach. *Invest Ophthalmol Vis Sci*. (2012) 53:8310–8. doi: 10.1167/iovs.12-9576

22. Li, Z, He, Y, Keel, S, Meng, W, Chang, RT, and He, M. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology*. (2018) 125:1199–206. doi: 10.1016/j.ophtha.2018.01.023

23. Li, Z, Keel, S, Liu, C, He, Y, Meng, W, Scheetz, J, et al. An automated grading system for detection of vision-threatening referable diabetic retinopathy on the basis of color fundus photographs. *Diabetes Care*. (2018) 41:2509–16. doi: 10.2337/dc18-0147

24. Ophthalmology AAo. International clinical diabetic retinopathy disease severity scale detailed table. (2002)

25. Programme D. Revised grading definitions for the NHS diabetic eye screening Programme. (2012)

26. Ferris, FL 3rd, Wilkinson, CP, Bird, A, Chakravarthy, U, Chew, E, Csaky, K, et al. Clinical classification of age-related macular degeneration. *Ophthalmology*. (2013) 120:844–51. doi: 10.1016/j.ophtha.2012.10.036

27. Iwase, A, Suzuki, Y, Araie, M, Yamamoto, T, Abe, H, Shirato, S, et al. The prevalence of primary open-angle glaucoma in Japanese: the Tajimi study. *Ophthalmology*. (2004) 111:1641–8. doi: 10.1016/S0161-6420(04)00665-7

28. He, M, Foster, PJ, Ge, J, Huang, W, Zheng, Y, Friedman, DS, et al. Prevalence and clinical characteristics of glaucoma in adult Chinese: a population-based study in Liwan District. *Guangzhou Invest Ophthalmol Vis Sci*. (2006) 47:2782–8. doi: 10.1167/iovs.06-0051

29. Topouzis, F, Wilson, MR, Harris, A, Anastasopoulos, E, Yu, F, Mavroudis, L, et al. Prevalence of open-angle glaucoma in Greece: the Thessaloniki eye study. *Am J Ophthalmol*. (2007) 144:511–519.e1. doi: 10.1016/j.ajo.2007.06.029

30. Zhixi Li, SK, Liu, C, He, Y, Meng, W, Scheetz, J, Lee, PY, et al. An automated grading system for vision-threatening referable diabetic retinopathy detection based on color fundus photographs. *Diabetes Care*. (2018) 41:2509–16. doi: 10.2337/dc18-0147

31. Keel, S, Lee, PY, Scheetz, J, Li, Z, Kotowicz, MA, MacIsaac, RJ, et al. Feasibility and patient acceptability of a novel artificial intelligence-based screening model for diabetic retinopathy at endocrinology outpatient services: a pilot study. *Sci Rep*. (2018) 8:4330. doi: 10.1038/s41598-018-22612-2

32. Keel, S, Li, Z, Scheetz, J, Robman, L, Phung, J, Makeyeva, G, et al. Development and validation of a deep-learning algorithm for the detection of neovascular age-related macular degeneration from colour fundus photographs. *Clin Exp Ophthalmol*. (2019) 47:1009–18. doi: 10.1111/ceo.13575

33. Ebner, M. Color constancy based on local space average color. *Mach Vis Appl*. (2009) 20:283–301. doi: 10.1007/s00138-008-0126-2

34. Christian Szegedy, VV. Rethinking the inception architecture for computer vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2016).

35. Looker, HC, Nyangoma, SO, Cromie, DT, Olson, JA, Leese, GP, Black, MW, et al. Rates of referable eye disease in the Scottish National Diabetic Retinopathy Screening Programme. *Br J Ophthalmol*. (2014) 98:790–5. doi: 10.1136/bjophthalmol-2013-303948

36. Yau, JW, Rogers, SL, Kawasaki, R, Lamoureux, EL, Kowalski, JW, Bek, T, et al. Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care*. (2012) 35:556–64. doi: 10.2337/dc11-1909

37. O'Neill, EC, Gurria, LU, Pandav, SS, Kong, YX, Brennan, JF, Xie, J, et al. Glaucomatous optic neuropathy evaluation project: factors associated with underestimation of glaucoma likelihood. *JAMA Ophthalmol*. (2014) 132:560–6. doi: 10.1001/jamaophthalmol.2014.96

38. Breusegem, C, Fieuws, S, Stalmans, I, and Zeyen, T. Agreement and accuracy of non-expert ophthalmologists in assessing glaucomatous changes in serial stereo optic disc photographs. *Ophthalmology*. (2011) 118:742–6. doi: 10.1016/j.ophtha.2010.08.019

39. Gulshan, V, Peng, L, Coram, M, Stumpe, MC, Wu, D, Narayanaswamy, A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. (2016) 316:2402–10. doi: 10.1001/jama.2016.17216

40. Silva, PS, Horton, MB, Clary, D, Lewis, DG, Sun, JK, Cavallerano, JD, et al. Identification of diabetic retinopathy and ungradable image rate with ultrawide field imaging in a national teleophthalmology program. *Ophthalmology*. (2016) 123:1360–7. doi: 10.1016/j.ophtha.2016.01.043

41. Gargeya, R, and Leng, T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*. (2017) 124:962–9. doi: 10.1016/j.ophtha.2017.02.008

42. Burlina, PM, Joshi, N, Pekala, M, Pacheco, KD, Freund, DE, and Bressler, NM. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmol*. (2017) 135:1170–6. doi: 10.1001/jamaophthalmol.2017.3782

43. Ting, DSW, Cheung, CY, Lim, G, Tan, GSW, Quang, ND, Gan, A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*. (2017) 318:2211–23. doi: 10.1001/jama.2017.18152

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 13 April 2023
DOI 10.3389/fmed.2023.1157016

OPEN ACCESS

EDITED BY
Tyler Hyungtaek Rim,
Mediwhale Inc., Republic of Korea

REVIEWED BY
Tae Keun Yoo,
B&VIIT Eye Center/Refractive Surgery &AI Center, Republic of Korea
Lisa Zhuoting Zhu,
Centre for Eye Research Australia, Australia

*CORRESPONDENCE
Sophia Y. Wang
✉ sywang@stanford.edu

SPECIALTY SECTION
This article was submitted to Ophthalmology, a section of the journal Frontiers in Medicine

RECEIVED 02 February 2023
ACCEPTED 15 February 2023
PUBLISHED 13 April 2023

CITATION
Jalamangala Shivananjaiah SK, Kumari S, Majid I and Wang SY (2023) Predicting near-term glaucoma progression: An artificial intelligence approach using clinical free-text notes and data from electronic health records. *Front. Med.* 10:1157016. doi: 10.3389/fmed.2023.1157016

Predicting near-term glaucoma progression: An artificial intelligence approach using clinical free-text notes and data from electronic health records

Sunil K. Jalamangala Shivananjaiah, Sneha Kumari, Iyad Majid and Sophia Y. Wang*

Department of Ophthalmology, Byers Eye Institute, Stanford University, Palo Alto, CA, United States

Purpose: The purpose of this study was to develop a model to predict whether or not glaucoma will progress to the point of requiring surgery within the following year, using data from electronic health records (EHRs), including both structured data and free-text progress notes.

Methods: A cohort of adult glaucoma patients was identified from the EHR at Stanford University between 2008 and 2020, with data including free-text clinical notes, demographics, diagnosis codes, prior surgeries, and clinical information, including intraocular pressure, visual acuity, and central corneal thickness. Words from patients' notes were mapped to ophthalmology domain-specific neural word embeddings. Word embeddings and structured clinical data were combined as inputs to deep learning models to predict whether a patient would undergo glaucoma surgery in the following 12 months using the previous 4-12 months of clinical data. We also evaluated models using only structured data inputs (regression-, tree-, and deep-learning-based models) and models using only text inputs.

Results: Of the 3,469 glaucoma patients included in our cohort, 26% underwent surgery. The baseline penalized logistic regression model achieved an area under the receiver operating curve (AUC) of 0.873 and F1 score of 0.750, compared with the best tree-based model (random forest, AUC 0.876; F1 0.746), the deep learning structured features model (AUC 0.885; F1 0.757), the deep learning clinical free-text features model (AUC 0.767; F1 0.536), and the deep learning model with both the structured clinical features and free-text features (AUC 0.899; F1 0.745).

Discussion: Fusion models combining text and EHR structured data successfully and accurately predicted glaucoma progression to surgery. Future research incorporating imaging data could further optimize this predictive approach and be translated into clinical decision support tools.

KEYWORDS

artificial intelligence, glaucoma, electronic health records, natural language processing, explainability, glaucoma surgery

1. Introduction

Glaucoma is a chronic progressive disease of the optic nerve and is one of the leading causes of irreversible blindness (1). Many patients remain at an early asymptomatic stage for long periods, while others progress to vision loss and require surgery (2). Although some factors contributing to progression are relatively easy to identify and measure, such as elevated intraocular pressure (IOP) or decreased central corneal thickness, many other factors, such as medication adherence, are less easily characterized (3). It is often difficult for doctors to predict whose glaucoma will worsen. However, now that the digitization of health records has created vast collections of information about patients (including medication and diagnosis information, demographic information, and free-text clinical notes), artificial intelligence (AI) techniques can be developed to analyze patient records and predict ophthalmic outcomes, including glaucoma progression.

Previous efforts have been undertaken to build machine-learning and deep-learning classification algorithms to predict glaucoma progression (4). Many studies have focused on structured information from electronic health records (5–7). Our previous studies have also explored methods for incorporating clinical free-text progress notes into AI prediction algorithms using natural language processing techniques (8, 9). A common challenge has been that these efforts typically have not considered the temporal element of prediction, as most AI prediction algorithms are simple classification algorithms with no specific time horizon. An outcome prediction is most useful when attached to a specific time horizon so that appropriate clinical steps can be taken. Similarly, algorithms trained only on baseline (presenting) information can only be used in limited circumstances and only for new patients. Algorithms that can be deployed at any point in a patient's clinical course would have broader utility.

The present study aims to develop artificial intelligence models that can predict glaucoma progression to the point of requiring surgery within 1 year, using inputs from electronic health records (EHRs) that are both structured and free-text. The present models would thus be able to be used on glaucoma patients at any time during their treatment course, overcoming a key limitation of previous work. Furthermore, unlike previous models, these models would incorporate temporal information by providing predictions over a fixed time horizon. We compared 3 types of models: a model incorporating information from clinical notes (clinical free text), models using only structured data inputs, and a multimodal fusion model that used both clinical free text and structured data as inputs.

2. Methods

2.1. Study population and cohort construction:

The overall objective of our algorithm was to predict whether a patient with glaucoma will require surgery within 12 months following a designated encounter visit, given at least 4 months of medical history prior to the encounter date. We narrowed the timeframe under consideration to the near-term future because, although a patient might have surgery at any time in the future (including many years after the initial glaucoma diagnosis), the most relevant prediction is whether the patient will need surgery within the next year. Thus, we carefully constructed a cohort to suit these prediction needs.

We first identified, from the Stanford Research Repository (10), all encounters for patients seen by the Department of Ophthalmology at Stanford University since 2008. We included all patients with at least two encounters with a glaucoma-related diagnosis as determined by the International Classification of Disease Codes (ICD10: H40, H42, or Q15.0; not including glaucoma suspect codes starting with H40.0 and ICD9 equivalents). Theoretically, a model could perform a prediction at any date in a patient's treatment timeline; for the purposes of our model training, we defined a single prediction date for each patient, on which the model predicts whether that patient would progress to glaucoma surgery within 12 months of that defined date.

This prediction date effectively divides the patient's information into a historical look-back period and a future look-forward period. We required a minimum of 4 months and up to 12 months of look-back period during which the clinical information used for the prediction was gathered. The model then predicts whether the patient will progress to require surgery over the following 12 months of look-forward period. Patients without at least 4 months of look-back data or without any clinical progress notes within the look-back period were excluded. For patients who underwent surgery, the prediction date was defined as either 12 months prior to surgery or after the initial 4 months of follow-up (whichever was later). For patients who did not undergo surgery, a minimum of 12 months of follow-up after the initial lookback period was required to ascertain that no surgery was performed over the entire look-forward period; the prediction date was defined as 12 months prior to the last follow-up visit, with the caveat that only patients with at least 4 months of historical lookback were included. A summary of cohort construction timelines with example patients is given in Figure 1. This formulation of the cohort and the prediction date is similar to that used in a previous study predicting near-term palliative care needs among inpatients (11).

With our inclusion/exclusion criteria, the final cohort included 3,469 patients. A cohort identification flow diagram is shown in Figure 2. We randomly divided our cohort into training, validation, and test groups in 70% ($N = 2,429$), 15% ($N = 520$), and 15% ($N = 520$) proportions, respectively. The proportion of patients who progressed to surgery in each of these groups was 25.9% ($N = 629$), 26.0% ($N = 135$), and 26.9% ($N = 140$), respectively. The validation set was used for model and hyperparameter tuning, and the final evaluation was performed on the test set.

2.2. Feature engineering

2.2.1. Text data preprocessing

Ophthalmology clinical progress notes from the look-back period were identified and concatenated such that the most recent notes appeared first. All notes were lower-cased and tokenized. Punctuation and stop words [Nltk library (12)] were removed. We mapped each word of the document with previously trained 300-dimensional ophthalmology domain-specific neural word embeddings (13) for input into models. To understand the general characteristics of the words associated with patients who progressed

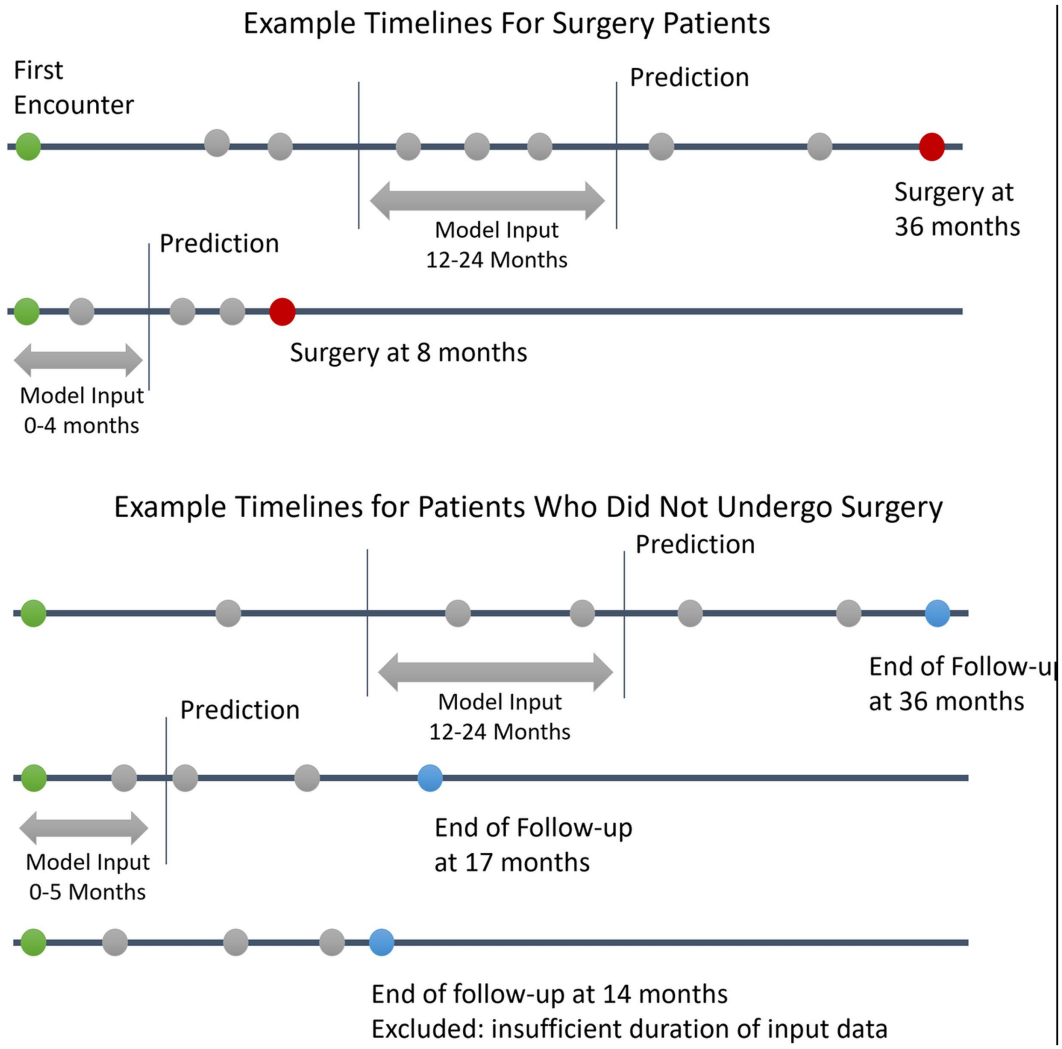

FIGURE 1
Example Patient Electronic Health Records Timelines. This figure depicts example timelines of patients who did and did not progress to glaucoma surgery. Green circles represent the first patient encounter. Red circles represent surgery dates. Blue circles represent the last follow-up encounter for patients who did undergo surgery. Grey circles represent other encounters with associated clinical data in the electronic health record.

to surgery and patients who did not, we calculated pointwise mutual information (14) for words that occurred in the notes of at least 20 different patients.

2.2.2. Structured data preprocessing

Information on the patient's demographics (age, gender, race/ethnicity), diagnoses (International Classification of Disease codes), medications, and eye examination results (visual acuity, intraocular pressure, and central corneal thickness of both eyes) was obtained from each patient's look-back period. Visual acuity was converted to mean logarithm of the minimum angle of resolution (logMAR). Numeric data were standardized to a mean of 0 and standard deviation of 1. We identified the low, medium, high, and most recent values for each eye examination feature. For the medication and diagnosis data, we filtered out features with <1% variance. For missing values in numeric features, we created a missing indicator column for the feature after performing column mean imputation. Categorical variables were converted to a series of Boolean dummy variables. After the final preprocessing, 127 structured features remained.

2.3. Modeling

2.3.1. Text model

To create a model that uses free text from clinical notes as the input, we built a one-dimensional convolutional network model (15), a similar style of which was previously been demonstrated to work well on ophthalmology notes (9). Figure 3 depicts the architecture of the text model. The free-text clinical notes were input into the model after padded or truncation to a length of 770 tokens (the 80th percentile length of notes). We applied a set of 25 one-dimensional convolutional filters, each of sizes 2, 3, 4, and 5, followed by max pooling. The outputs of these filters were concatenated and passed through 2 additional fully-connected layers with dropout to obtain our final prediction. For training, we used the Adam optimization

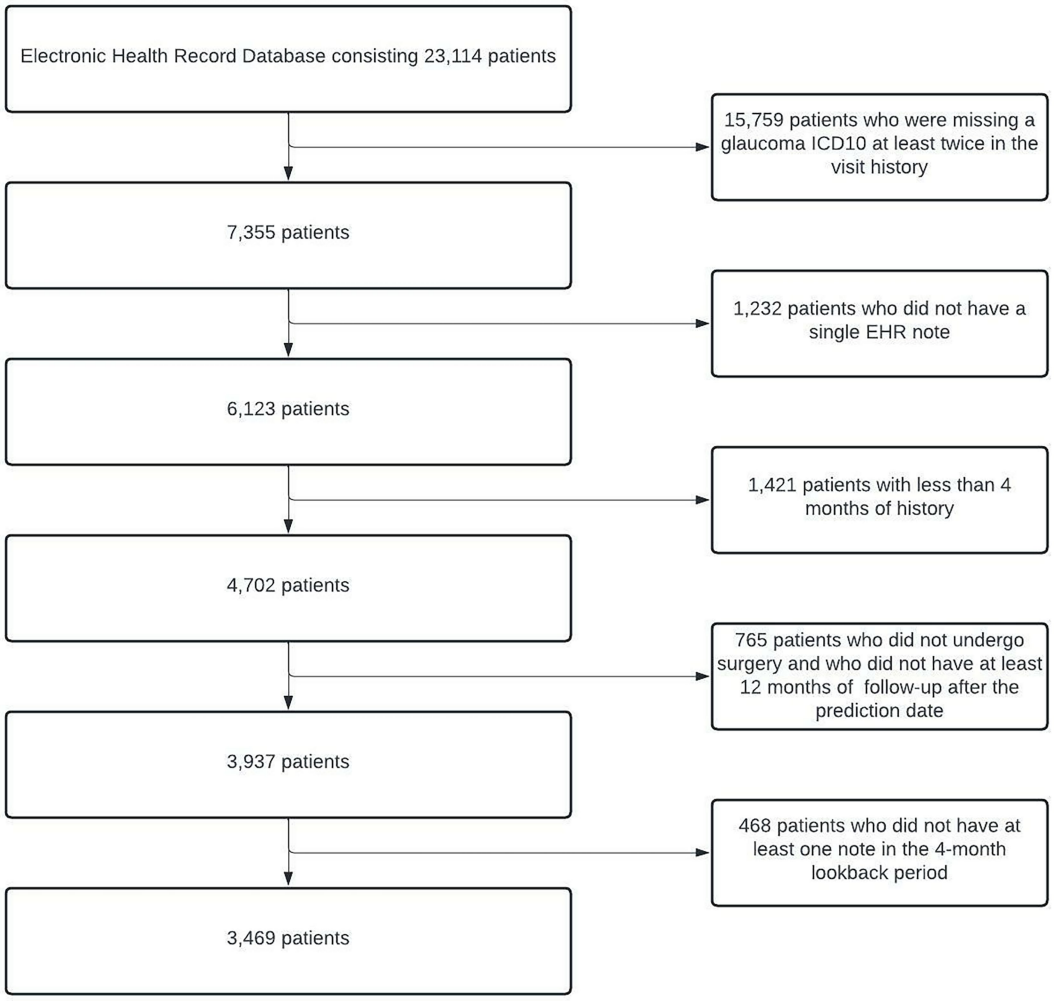

FIGURE 2
Cohort Description and Construction. Flowchart depicting the process of identifying eligible glaucoma patients from EHRs. At the end of all of the processing steps, 3,469 patients were included in the study. EHRs, electronic health records.

algorithm. We tuned the model's architecture, learning rate (0.0003), weight decay factor (0.01), and batch size (32) to optimize validation loss.

2.3.2. Structured EHR data model

To create a model that uses structured EHR data as the input, we started by building a basic L1 penalized logistic regression model with the structured data. We also trained several tree-based models, including random forest and extreme gradient boosting (XGboost), and saw that these models performed better on our dataset than our baseline model. We tuned the maximum depth, minimum samples per leaf node, and the number of trees for the tree-based models. We then built a fully connected neural network model based on the structured data, as follows: The input features were fed into the neural network with 2 hidden layers of 60 and 30 nodes. The first hidden layer used an input of 60 nodes, followed by ReLU activation and a dropout layer with a probability of 0.5. The second hidden layer had 30 nodes, also followed by ReLU activation and a dropout layer with a probability of 0.5. Then, a final prediction layer used softmax activation for classification. Models' hyperparameters were tuned using 5-fold cross validation (Table 1).

2.3.3. Multimodal fusion model

To create a model that considers both clinical free text and structured EHR data as the input, we created a neural network-based multimodal model. Each layer of the neural network can be thought of as feature engineering, wherein the model is automatically engineering these layers. We extracted the final layer of the text model and the deep learning-based structured model (as these are features curated by the individual models), combined the features, and then applied L1 penalized logistic regression to predict the outcome. The model architecture is depicted in Figure 4.

2.4. Evaluation

We evaluated the performance of all our models on the held-out test dataset (data set aside for testing the model after training and validation) using precision (also known as positive predictive value), recall (also known as sensitivity), F1 score (the harmonic mean of the precision and recall), the area under the receiver operating curve (AUROC), and the area under the precision-recall curve (AUPRC). For metrics of precision, recall, and F1 score, which are reported at a

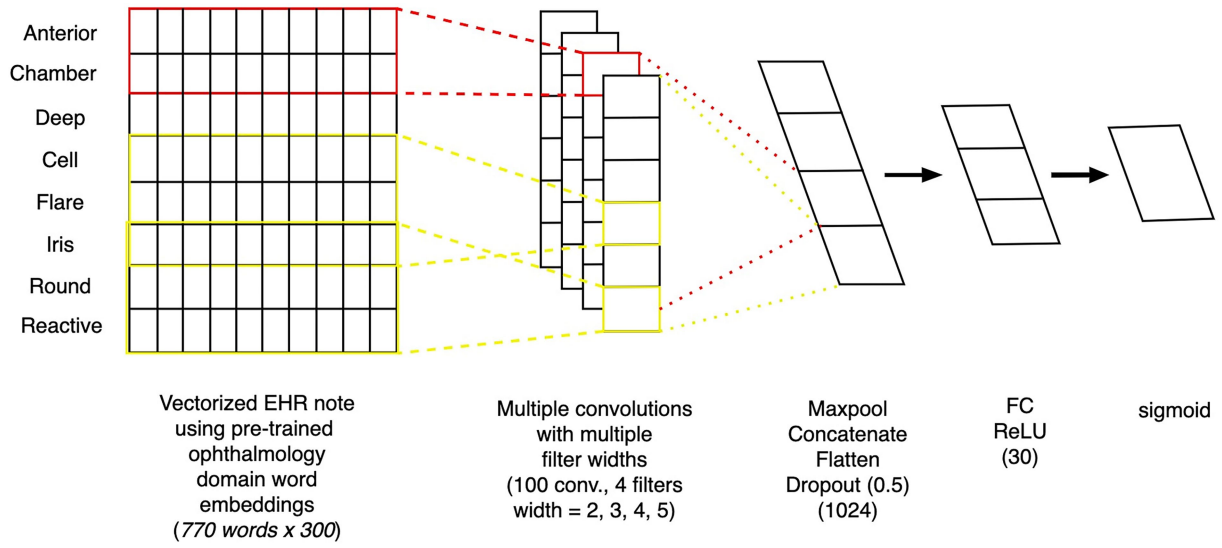

FIGURE 3
Architecture of convolutional neural network text model. Depiction of the architecture of a convolutional neural network model, which takes as inputs free-text from clinical notes and outputs the probability of whether the patient will progress to surgery within 12 months.

TABLE 1 Hyperparameters for the tuned models.

Modeling method	Data	Hyperparameters
Logistic regression	Structured data	Penalty: L1
		Max iteration: 100
		Scoring: roc_auc
XGboost	Structured data	N_estimators: 100
		Max depth: 3
		Learning rate: 0.1
		Reg_lambda: 0
Gradient boosted trees	Structured data	Learning rate: 0.1
		Max depth: 3
		N_estimators: 100
		Subsample: 0.5
Random forest	Structured data	Max depth: 50
		Min samples per leaf node: 30
		N_estimators: 1200
Deep learning	Structured data	Learning rate: 0.0005
		Weight Decay: 0.1
		Batch Size: 32

single classification threshold, the optimal threshold was tuned on the validation set for the best F1 score; the final precision, recall, and F1 scores were evaluated on the test set using this optimized threshold.

2.5. Explainability

2.5.1. Explainability for the structured features

To understand which structured features had more predictive power, we used SHapley Additive exPlanations (SHAP) values (16). This technique calculates the importance of the features based on the magnitude of feature attributions, using a game theory approach to explain the results of any machine learning model and make them interpretable by measuring the feature contribution to individual predictions. We used SHAP TreeExplainer (17), which estimates the SHAP values for tree-and ensemble-based models, on the best random-forest model.

2.5.2. Explainability for the text model

GradCAM, a class-discriminative localization technique originally proposed by Selvaraju et al. (18) that generates visual explanations for any convolutional neural network (CNN) without requiring architectural changes or re-training. The technique was further adapted for Text-CNN (19). We used this technique to investigate the key phrases that led our Text-CNN model to predict that a given patient would require surgery within 12 months.

3. Results

3.1. Population characteristics:

Population characteristics of the subjects included in the study are summarized in Table 2. Of these patients, 26% went on to require glaucoma surgery. The mean age of the patients in the study was 67 years, and the mean IOP for both eyes was close to 15 mmHg. The mean logarithm of the minimum angle of resolution visual acuity for both eyes (mean logMAR) was around 0.55 for the right eye and 0.65 for the left eye (Snellen equivalent approximately

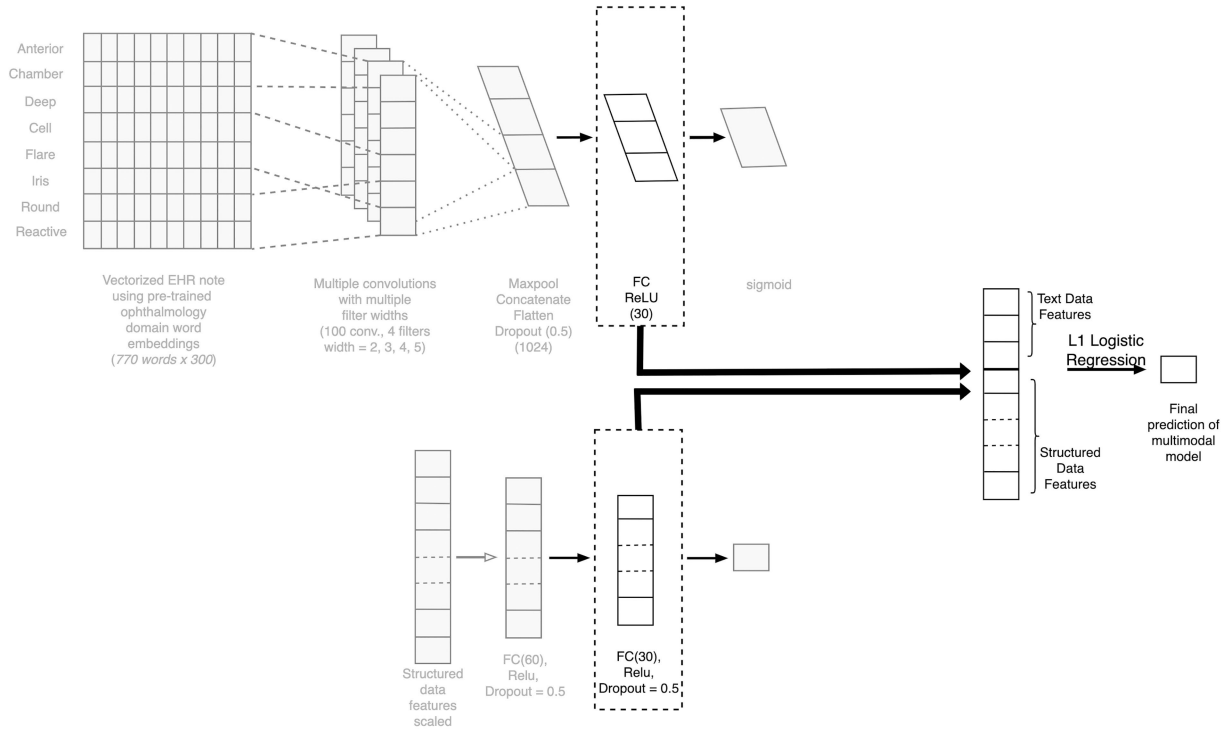

FIGURE 4
Architecture of the fusion model combining text and structured data. Depiction of the architecture of our multimodal fusion model, which fuses both structured and free-text modalities of data. The final layers of the TextCNN is concatenated with the final layer of the structured model neural network to create a combined feature vector for final prediction. CNN, competitive neural network; EHR, electronic health record; FC, fully connected.

TABLE 2 Population characteristics.

Characteristics	Total (*n*=3,469)	No surgery (*n*=2,565)	Surgery (*n*=904)
Age (years)	67 ± 18	69 ± 17	66 ± 18
Intraocular pressure (mmHg)			
Right eye	15.4 ± 6.1	15.1 ± 5.0	15.5 ± 6.8
Left eye	15.4 ± 6.2	15.3 ± 5.7	15.4 ± 6.6
Visual acuity (logMAR)			
Right eye	0.55 ± 0.84	0.48 ± 0.81	0.61 ± 0.86
Left eye	0.65 ± 0.92	0.59 ± 0.93	0.70 ± 0.92
Sex			
Female sex	1766 (50.9)	1,330 (51.8)	436 (48.2)
Male sex	1703(49.1)	1,235 (48.2)	468 (51.8)
Race and ethnicity			
Asian	997 (28.7)	707 (27.5)	290 (32.1)
White	1,444 (41.6)	1,138 (44.4)	306 (33.8)
Hispanic	378 (10.9)	251 (9.8)	127 (14.0)
Black	142 (4.2)	98 (3.8)	44 (4.9)
Other	508 (14.6)	371 (14.5)	137 (15.2)

logMAR = logarithm of the minimum angle of resolution. Data are presented as mean ± standard deviation or number (%). Numeric variables are reported here as conventionally seen, for ease of interpretation. They were standardized to a mean of 0 and variance of 1 before being applied to the modeling approaches.

20/70 and 20/90, respectively). The population in the study was predominantly Asian and White. To gain an overview of words most highly associated with patients who progressed to surgery and those who did not, pointwise mutual information was calculated for words that occurred in at least 20 patients. The top 10 words with the highest pointwise mutual information scores are presented in Supplementary Table S1.

3.2. Model results

For the structured data models, the L1 penalized logistic regression model resulted in an AUC score of 0.873 and F1 score of 0.750. Tree-based models resulted in AUC and F1 of 0.870 and 0.757 (XGboost), 0.871 and 0.749 (gradient boosted trees), and 0.876 and 0.746 (random forest). The deep learning structured model resulted in AUC of 0.885 and F1 score of 0.757. For all models, the classification threshold was tuned to optimize F1 score on the validation set. Receiver operating characteristic curves and precision-recall curves for the best structured, text, and combined deep learning models are shown in Figure 5. The combined model, which included both structured and free-text features, outperformed the structured-data-only model and free-text-only model. The free-text-only model resulted in an AUC of 0.767, and the combined model had an AUC of 0.899. Table 3 presents the F1 score and the corresponding precision and recall scores for each of the modalities of data based on deep learning models.

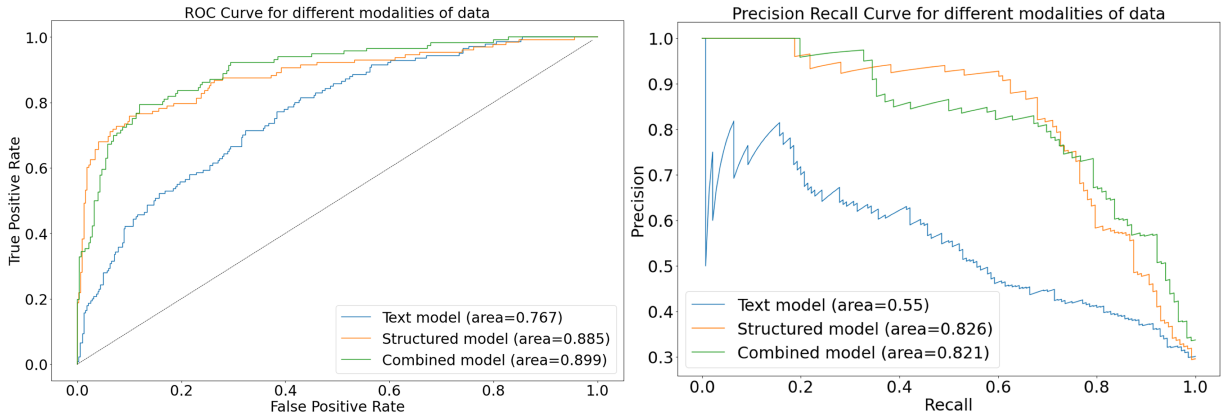

FIGURE 5
Receiver-operating and precision-recall curves for models. This figure depicts receiver operating characteristic curves and precision recall curves for models predicting glaucoma progression to surgery. The free-text model uses only clinical notes as inputs into a convolutional neural network; the structured model uses only structured electronic health records data as inputs into a deep learning model; and the combined model fuses both structured and text inputs into a fusion deep learning model.

TABLE 3 Performance metrics.

Modality	Precision	Recall	F1 score	Threshold
Text model	0.3959	0.8285	0.5357	0.4
Structured model	0.8000	0.7187	0.7572	0.55
Multimodal model	0.7022	0.7931	0.7449	0.35

3.3. Explainability

3.3.1. Explainability for structured data

Figure 6 depicts the mean absolute Shapley values for the top 20 most important features from the structured data for predicting which patients will require surgery, using the random forest model. The most important features include the use or nonuse of glaucoma medications as per the medication lists, IOP, VA, and refraction spherical equivalent. These features are similar to the factors considered by glaucoma specialists when they predict a glaucoma patient's prognosis with respect to the need for surgery.

3.3.2. Explainability for text data

Figure 7 highlights words and phrases in example clinical progress notes that were identified by GradCAM-text as most important for model predictions. For a high-risk patient, these explainability methods highlight clinical features that tend to indicate acute risk of surgery ("Outside ophthalmologist performed laser," "referred urgently for cataract and glaucoma surgery" etc.), while for a low-risk patients, highlighted clinical features are generic or low risk ("glaucoma suspect," "intraocular pressure was normal" etc.), which is in alignment with the expectations of glaucoma specialists.

4. Discussion

In the present study, we developed an AI approach that successfully predicts whether glaucoma patients would require surgery in the following 12 months based on EHR structured data and clinical progress notes. The study compared the results from 3 different approaches: (1) a deep-learning model which used doctor's free-text progress notes as input; (2) traditional machine-learning and deep-learning modeling approaches, which used structured EHR data as input; and (3) fusion deep-learning models, which used both structured EHR data and free-text notes as input. The resulting predictions indicated that fusion models trained using both structured EHR data and free-text notes as features performed better than models using either structured EHR only or free-text notes only. Explainability studies showed that models relied upon clinically relevant features in both the structured and text inputs.

Our work expands upon previous work which has generally focused on using single modalities of data. Baxter et al. explored many deep-learning and tree-based models for structured EHR data inputs but concluded that a logistic regression model had the best performance, with an AUC of 0.67 (4). Hu et al. demonstrated that using massively pre-trained language models for clinical free-text notes could improve the AUC to 0.70 (20). Wang et al. achieved an AUC of 0.73 on structured data only and an AUC of 0.70 using text features only. Using our combined model to predict glaucoma surgery in the near term, we were able to achieve an AUC of 0.899.

In addition to achieving significantly higher AUC, our model has inherent flexibility in terms of input that is in line with the needs of real-life patients and physicians in the clinic. Previous studies were limited to predicting a patient's prognosis regarding the need for surgery using data from their initial visit or data included from the baseline period. Some studies focused on predicting future surgery over all time, which sometimes meant predicting surgery even 10 years into the future (9, 20). A unique strength of our study was the formulation of a model that could be used for any glaucoma patient at

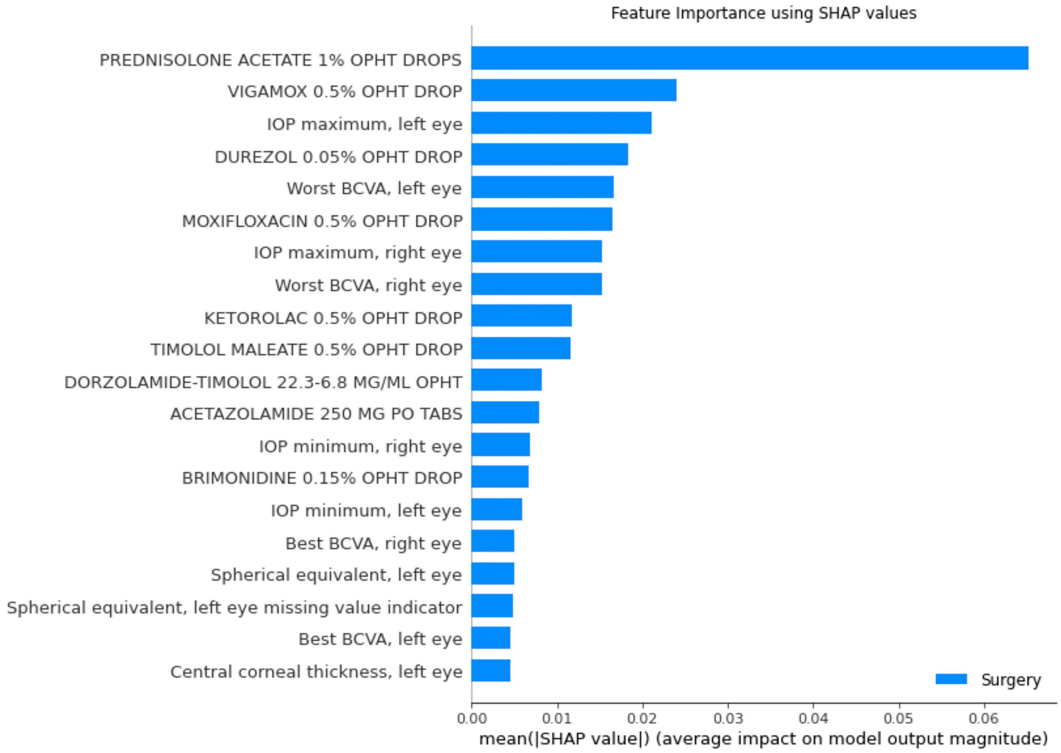

FIGURE 6
Shapley feature importance for structured data. The figure depicts the mean absolute Shapley value for the top 20 most important features in the structured data for predicting whether a patient would progress to the point of requiring surgery within the next year, using the random forest model. The mean absolute Shapley value was calculated for all patients in the test set.

Example High-Risk Patient Note
HPI – 75 yo man referred for acute angle closure glaucoma right eye. Outside ophthalmologist performed laser peripheral iridotomy earlier this week for IOP in the 5s with subsequent reduction of IOP to 40s. Started the patient on latanoprost, dorzolamide/timolol, and brimonidine with lowering of IOP to 30s. Started on Diamox as well. Patient is phakic and is referred urgently for cataract and glaucoma surgery.

Example Low-Risk Patient Note
35 year old M referred by optometry for evaluation of glaucoma suspect. Patient reports that he was at the optometrist getting routine evaluation for glasses when he was noted to be a glaucoma suspect. No family history for glaucoma. Intraocular pressure was normal. Of note patient is highly myopic and has worn glasses all of his life. He does not remember ever being evaluated for glaucoma previously. He had a normal visual field test by report. No history of previous eye surgery, lasers, or trauma. Not using any eye medications.

FIGURE 7
Feature importance for text data. This figure shows example notes for glaucoma patients at high and low risk for progressing to glaucoma surgery. The GRAD-CAM-Text method was used to identify words important to the model prediction, highlighted in red for predicting surgery and in green for predicting no surgery.

any point during their follow-up, rather than just at their initial visit or within a restricted baseline period, to predict the dynamic probability of whether the patient will require a surgical procedure within 1 year from the prediction date. Furthermore, considering more recent or updated information in the present models likely caused the prediction performance to improve.

We also investigated what types of information models were relying upon for prediction to improve transparency and trustworthiness for these AI models, a common criticism of which is that they are "black boxes" difficult for clinicians to understand. For the free-text model, explainability studies using GradCAM-Text showed the key phrases from the notes that were most important for the predictions, which were aligned with clinical expectations: for example, "referred urgently." Similarly, for the structured-EHR-data models, analysis of Shapley values showed the most important features included the use of various glaucoma medications, intraocular pressure, refraction, and visual acuity. These are similar features to those clinicians would take into consideration when making a decision (2).

This study has several remaining limitations and challenges. Our models are built and validated on patients who visited a single academic center, which may limit generalizability. Additionally, the input text length was limited, which is a challenge of deep learning architectures for incorporating text. To develop a model that can predict prognosis for any patient at any point in their treatment trajectory, it must be taken into consideration that every patient will have different amounts, and differing complexity, of data as input. Structured data can be easily summarized (e.g., most recent measurement values, highs, lows, medians, or even presence of specific conditions) but raw inputs of text require every word to be input, and models must have a standardized input length. To solve

this problem, some sort of meaningful summary representation of a variable amount of text history would be required, to reduce the text to a standardized input size. Furthermore, although we could perform explainability studies for the text and structured data models, there are no commonly used methods to investigate explainability in the multimodal models, which could be an area for future research.

In conclusion, we used multimodal electronic health records data to develop models to predict which glaucoma patients were likely to progress to surgery in the following 12 months, significantly outperforming previous models built for similar tasks. We showed that both text-based and structured-data-based models relied upon clinically relevant information to make predictions. Fusion models relying on both structured data and text notes, while lacking explainability, may improve model performance as compared with models relying on only structured data and only free-text notes. To take the next step towards translation into clinical decision support tools, further research is needed to improve explainability and performance, potentially by incorporating larger data sources and imaging data modalities.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: The dataset contain protected health information of individuals and cannot be publicly shared; for those wishing to collaborate, please contact the corresponding author. Requests to access these datasets should be directed to SW, sywang@stanford.edu.

Ethics statement

The studies involving human participants were reviewed and approved by Stanford University Institutional Review Board. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

SJ, SK, and SW contributed to data analysis. SW contributed to data acquisition and supervised the study. All authors contributed to the interpretation of the data, drafting and critical revision of the manuscript, and approval of the final manuscript.

Funding

This work was supported by National Eye Institute K23EY03263501 (SW); Career Development Award from Research to Prevent Blindness (SW); unrestricted departmental grant from Research to Prevent Blindness; and departmental grant National Eye Institute P30-EY026877.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2023.1157016/full#supplementary-material

References

1. Quigley, HA, and Broman, AT. The number of people with glaucoma worldwide in 2010 and 2020. *Br J Ophthalmol*. (2006) 90:262–7. doi: 10.1136/bjo.2005.081224

2. Weinreb, RN, Aung, T, and Medeiros, FA. The pathophysiology and treatment of glaucoma: a review. *JAMA*. (2014) 311:1901–11. doi: 10.1001/jama.2014.3192

3. Pantalon, AD, Feraru, C, and Chiseliță, D. Risk factors and long term progression in open angle glaucoma patients. *Rom J Ophthalmol*. (2016) 60:174–80.

4. Baxter, SL, Marks, C, Kuo, T-T, Ohno-Machado, L, and Weinreb, RN. Machine learning-based predictive modeling of surgical intervention in glaucoma using systemic data from electronic health records. *Am J Ophthalmol*. (2019) 208:30–40. doi: 10.1016/j.ajo.2019.07.005

5. Brigatti, L, Nouri-Mahdavi, K, Weitzman, M, and Caprioli, J. Automatic detection of glaucomatous visual field progression with neural networks. *Arch Ophthalmol*. (1997) 115:725–8. doi: 10.1001/archopht.1997.01100150727005

6. Bowd, C, Chan, K, Zangwill, LM, Goldbaum, MH, Lee, T-W, Sejnowski, TJ, et al. Comparing neural networks and linear discriminant functions for glaucoma detection using confocal scanning laser ophthalmoscopy of the optic disc. *Invest Ophthalmol Vis Sci*. (2002) 43:3444–54. PMID: 12407155

7. Weinreb, RN, Zangwill, L, Berry, CC, Bathija, R, and Sample, PA. Detection of glaucoma with scanning laser polarimetry. *Arch Ophthalmol*. (1998) 116:1583–9. doi: 10.1001/archopht.116.12.1583

8. Barrows, RC Jr, Busuioc, M, and Friedman, C. Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches. *Proc AMIA Symp*. (2000) 7:51–5.

9. Wang, SY, Tseng, B, and Hernandez-Boussard, T. Deep learning approaches for predicting glaucoma progression using electronic health records and natural language processing. *Ophthalmol Sci*. (2022) 2:100127. doi: 10.1016/j.xops.2022.100127

10. Lowe, HJ, Ferris, TA, Hernandez, PM, and Weber, SC. STRIDE--an integrated standards-based translational research informatics platform. *AMIA Annu Symp Proc*. (2009) 2009:391–5. PMID: 20351886

11. Avati, A, Jung, K, Harman, S, Downing, L, Ng, A, and Shah, NH. Improving palliative care with deep learning. *BMC Med Inform Decis Mak*. (2018) 18:122. doi: 10.1186/s12911-018-0677-8

12. Nltk. *Github*. (2021) Available at: https://github.com/nltk/nltk (Accessed June 2, 2021).

13. Wang, S, Tseng, B, and Hernandez-Boussard, T. Development and evaluation of novel ophthalmology domain-specific neural word embeddings to predict visual prognosis. *Int J Med Inform*. (2021) 150:104464. doi: 10.1016/j.ijmedinf.2021.104464

14. Church, KW, and Hanks, P. Word association norms, mutual information, and lexicography. *Comput Linguist*. (1990) 16:22–9.

15. Kim, Y. *Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Stroudsburg, PA, USA: Association for Computational Linguistics (2014).

16. Lundberg, SM, Erion, G, Chen, H, DeGrave, A, Prutkin, JM, Nair, B, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. (2020) 2:56–67. doi: 10.1038/s42256-019-0138-9

17. Lundberg, S. *Shap, Github*. (2021) Available at: https://github.com/slundberg/shap (Accessed June 2, 2021).

18. Selvaraju, RR, Cogswell, M, Das, A, Vedantam, R, Parikh, D, and Batra, D. *Grad-CAM: Visual Explanations from deep Networks via Gradient-based Localization*. arXiv [csCV] (2016) Available at: http://arxiv.org/abs/1610.02391.

19. Samek, W, Wiegand, T, and Muller, K-R. Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. *ITU J*. (2018) 1:39–48. doi: 10.48550/arXiv.1708.08296

20. Hu, W, and Wang, SY. Predicting glaucoma progression requiring surgery using clinical free-text notes and transfer learning with transformers. *Transl Vis Sci Technol*. (2022) 11:37. doi: 10.1167/tvst.11.3.37

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 12 May 2023
DOI 10.3389/fmed.2023.1165135

OPEN ACCESS

EDITED BY
Darren Shu Jeng Ting,
University of Nottingham, United Kingdom

REVIEWED BY
Tae Keun Yoo,
B & VIIT Eye Center/Refractive Surgery & AI Center, Republic of Korea
Xiangjia Zhu,
Fudan University, China

*CORRESPONDENCE
Haochao Ying
haochaoying@zju.edu.cn
Wen Xu
xuwen2003@zju.edu.cn

†These authors have contributed equally to this work and share first authorship

RECEIVED 13 February 2023
ACCEPTED 14 April 2023
PUBLISHED 12 May 2023

CITATION
Wang J, Wang J, Chen D, Wu X, Xu Z, Yu X, Sheng S, Lin X, Chen X, Wu J, Ying H and Xu W (2023) Prediction of postoperative visual acuity in patients with age-related cataracts using macular optical coherence tomography-based deep learning method.
Front. Med. 10:1165135.
doi: 10.3389/fmed.2023.1165135

Prediction of postoperative visual acuity in patients with age-related cataracts using macular optical coherence tomography-based deep learning method

Jingwen Wang[1†], Jinhong Wang[2†], Dan Chen[1], Xingdi Wu[1], Zhe Xu[1], Xuewen Yu[1,3], Siting Sheng[1], Xueqi Lin[1], Xiang Chen[1], Jian Wu[4], Haochao Ying[5*] and Wen Xu[1*]

[1]Eye Center of the Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, Zhejiang, China, [2]College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang, China, [3]Department of Ophthalmology, The First People's Hospital of Xiaoshan District, Xiaoshan Affiliated Hospital of Wenzhou Medical University, Hangzhou, Zhejiang, China, [4]Second Affiliated Hospital School of Medicine, School of Public Health, and Institute of Wenzhou, Zhejiang University, Hangzhou, Zhejiang, China, [5]School of Public Health, Zhejiang University, Hangzhou, Zhejiang, China

Background: To predict postoperative visual acuity (VA) in patients with age-related cataracts using macular optical coherence tomography-based deep learning method.

Methods: A total of 2,051 eyes from 2,051 patients with age-related cataracts were included. Preoperative optical coherence tomography (OCT) images and best-corrected visual acuity (BCVA) were collected. Five novel models (I, II, III, IV, and V) were proposed to predict postoperative BCVA. The dataset was randomly divided into a training (n=1,231), validation (n=410), and test set (n=410). The performance of the models in predicting exact postoperative BCVA was evaluated using mean absolute error (MAE) and root mean square error (RMSE). The performance of the models in predicting whether postoperative BCVA was improved by at least two lines in the visual chart (0.2LogMAR) was evaluated using precision, sensitivity, accuracy, F1 and area under curve (AUC).

Results: Model V containing preoperative OCT images with horizontal and vertical B-scans, macular morphological feature indices, and preoperative BCVA had a better performance in predicting postoperative VA, with the lowest MAE (0.1250 and 0.1194LogMAR) and RMSE (0.2284 and 0.2362LogMAR), and the highest precision (90.7% and 91.7%), sensitivity (93.4% and 93.8%), accuracy (88% and 89%), F1 (92% and 92.7%) and AUCs (0.856 and 0.854) in the validation and test datasets, respectively.

Conclusion: The model had a good performance in predicting postoperative VA, when the input information contained preoperative OCT scans, macular morphological feature indices, and preoperative BCVA. The preoperative BCVA and macular OCT indices were of great significance in predicting postoperative VA in patients with age-related cataracts.

KEYWORDS

cataract surgery, visual acuity, deep learning, macula, optical coherence tomography

1. Introduction

Cataract, defined as the opacity of the lens, is one of the leading causes of visual impairment worldwide and a primary cause of blindness, estimated to be responsible for 15.2 million cases of blindness in 2020 (1). Many factors lead to the formation of cataracts, including age, diabetes, and ultraviolet irradiation, and age remains the major risk factor for cataracts (2). The only effective treatment is surgery. Most patients can gain excellent visual acuity (VA) after cataract surgeries. However, some patients may fail to obtain satisfying visual outcomes due to complicated fundus diseases. Predicting visual outcomes before cataract surgeries can help patients adjust their expectations appropriately and aid doctors in making reasonable decisions for patients whose vision may not be improved. This can avoid the waste of medical resources and contradictions between doctors and patients.

As the sharpest part of the retina for vision, the macula remains one of the most important factors in determining VA after cataract surgery. Optical coherence tomography (OCT) is a non-invasive, high-resolution cross-sectional imaging modality of the structural retina *in vivo*. The introduction of OCT helps ophthalmologists qualitatively and quantitatively assess the subtle structural changes in the macular region (3, 4). Previous studies have reported that abnormal morphological changes in OCT images can lead to worse visual outcomes in patients with retinal diseases (5–8). Most clinicians currently analyze OCT images empirically to judge the function of the retinal macula and thus roughly estimate the postoperative VA of patients with cataracts. However, there is no standardized evaluation system based on large samples to quantify the relationship between macular morphological changes and postoperative VA of patients with cataracts.

Artificial intelligence (AI), especially deep learning (DL), has been widely used to analyze retinal images in the past few decades. Some studies have achieved satisfactory results in applying DL algorithms to predict postoperative visual outcomes in retinal diseases (9–11). Mao et al. have investigated the predictive factors of VA in patients with retinitis pigmentosa after cataract surgery (12). The preoperative best-corrected visual acuity (BCVA), the status of the external limiting membrane, and central macular thickness are found to be important parameters to predict postoperative VA. Recently, Wei et al. have constructed an OCT-based DL approach to predict the postoperative VA of patients with high myopia (13). Xiang et al. have developed an intelligent system based on OCT images for long-term BCVA prediction in 3 and 5 years after surgery in patients with congenital cataracts (14). All of the above studies have yielded good results. However, these studies are hard to explain the underlying mechanism due to the lack of anatomic and morphological features integrated with the study. It is essential since the microstructure of the macula are closely correlated with the postoperative VA of patients with cataracts (15).

In the present study, AI models were developed based on preoperative macular OCT images and BCVA to predict the postoperative VA in patients with age-related cataracts. The AI models were then compared to evaluate the prediction performances of certain postoperative BCVA and whether postoperative BCVA was improved by at least two lines in the visual chart (0.2LogMAR).

2. Materials and methods

2.1. Participants

The study was performed on 2,051 eyes from 2,051 patients with cataracts who underwent uneventful cataract surgeries operated by the same experienced cataract surgeon in the Eye Centre at the Second Affiliated Hospital of Zhejiang University, School of Medicine, from December 2018 to June 2020. The dataset consisting of 2051 eyes from 2051 patients with cataracts was randomly divided into a training ($n = 1,231$), validation ($n = 410$), and test set ($n = 410$). Collected clinical data included gender, laterality, and surgical age, as well as BCVA measured preoperatively and 1 month postoperatively. Image data included horizontal and vertical B-scan macular OCT images of the patient at the same preoperative visit.

Inclusion criteria were as follows: (1) age 50–90 years old, diagnosed with senile cataract, the degree of lens opacity was graded by the Lens Opacities Classification System III: cortical opacity at grade 4 and below, posterior subcapsular opacity at grade 4 and below. The hardness of the nucleus was graded by the Emery and Little classification: grade IV and below, (2) reliable OCT measurements of the macula were performed before cataract surgery, (3) underwent peaceful cataract surgeries, and (4) had reliable BCVA measured preoperatively and 1 month postoperatively.

Exclusion criteria were as follows: (1) Amblyopia; (2) Congenital ocular anomalies; (3) Cataracts caused by trauma or congenital anomalies; (4) Refractive media opacities that seriously affected macular OCT image clarity or visual prognosis, such as severe lens opacities, centered corneal opacities, severe vitreous opacities; (5) Poor quality macular OCT images that affected image analysis; (6) Combined with retinal detachment, retinitis pigmentosa, and fundus lesions such as optic nerve and choroid that might affect VA; (7) Combined with nystagmus or head tremor and other diseases that were susceptible to interference during macular OCT examination; (8) Combined with consciousness or intellectual impairment that affected the accuracy of visual acuity test results; The patients are excluded if they meet the any of the above exclusion criteria.

This study was approved by the Institutional Review Board of the Second Affiliated Hospital of Zhejiang University, School of Medicine. Written informed consents was obtained from all the participants. This study complied with the Declaration of Helsinki and was registered at[1] (accession number NCT04887909).

2.2. Macular OCT images

OCT images were acquired from Spectrialis OCT (Heidelberg Engineering, Heidelberg, Germany), Cirrus OCT (Carl Zeiss Meditec, Dublin, California, United States), and FD-OCT (RTVue; Optovue Inc., Fremont, California, USA). Images from Spectrialis OCT had a resolution of 768 by 496 pixels, with a scan width of 10,000 μm and a scan depth of 2,000 μm in the air. Images from Cirrus OCT had a resolution of 938 by 625 pixels, with a scan width of 6,000 μm and a scan depth of 2,000 μm in the air. Images from FD-OCT had a

1 www.clinicaltrials.gov

resolution of 1,020 by 960 pixels, with a scan width of 10,000 μm and a scan depth of 2,000 μm in the air.

To fully obtain the information from OCT images, morphological features of the macula were extracted and analyzed. The process of image analysis is presented in Figure 1. Firstly, irrelevant signal-to-noise was reduced by the denoising algorithm. The pixels were smoothly connected using the dilate algorithm, and the edges were detected to obtain the layered boundaries. The internal limiting membrane (ILM) layer was probed from top to bottom (Step 1). A horizontal correction was then performed based on the curve of the ILM layer to obtain a profile of the total retinal thickness (Step 2). Similarly, by creating a gradient graph to filter out the hazy features next to the retinal pigment epithelium (RPE) and highlight the RPE layer, the boundary of the retinal pigment epithelial cell layer was probed (Step 3). Five marks of the fovea were automatically recognized (Step 4). There was no slope in the temporal and nasal rims of the fovea in the horizontal meridian (Figures 1A,E). The pseudocode of this pre-processing process is in Algorithm 1. Additionally, the pit of the fovea (Figure 1C) had no slope. A maximum slope for the temporal and nasal foveal walls of the horizontal meridian was also detected (Figures 1B,D). Five salient features of the foveal pit were extracted from these five marks, including foveal thickness, pit depth, diameter, maximum thickness, and foveal slope (16, 17).

Algorithm 1 Layer Segmentation

Input: raw image I
Output: output five salient features A,B,C,D,E

img: $I(H, W)$

# Denoising iteration
for $i = 1, 2, \ldots, k$ **do**
    img = cv2.fastNlMeanDenosing(img, 10, λ_i, λ_i/2)
# Edge detection by Canny algorithm
img = cv2.canny(img, α, β)
# Dilate operation
img = cv2.dilate(img)

# ILM layer segmentation
for $i = 0, 1, \ldots, W-1$ **do**
    for $j = 0, 1, \ldots, H-1$ **do**
        ILM ← get the ILM layer by finding the first pixel img[i,j] with value 1 from top to bottom for each i
θ ← get the angle based on first five pixel of ILM layer
# Rotate image for horizontal correction
M = cv2.getRotationMatrix2D((W/2, H/2), θ)
img = cv2.warpAffine(img, M, (W, H))

# Get the gradient graph
grad = gradientImagecreate(img)

# RPE layer segmentation
for $i = 0, 1, \ldots, W-1$ **do**
    for $j = H-1, H-2, \ldots, 0$ **do**
        RPE ← get the RPE layer by finding the pixel img[i,j] with maximum gradient from bottom to top for each i

# Determine C
for $c = W/2 - 100, \ldots, W/2 + 100$ **do**
    C ← find the pixel (c, ILM[c]) of ILM layer that near the middle area with minimum slope as point C
# Determine A
for $a = 0, 1, \ldots, c-1$ **do**
    A ← find the pixel (a, ILM[a]) of ILM layer between left boundary and point C with maximum value of difference |ILM[a] - RPE[a]| as point A
# Determine E
for $e = c+1, c+2, \ldots, W-1$ **do**
    E ← find the pixel (e, ILM[e]) of ILM layer between C and right boundary with maximum value of difference |ILM[e] - RPE[e]| as point E
# Determine B
for $b = a+1, a+2, \ldots, c-1$ **do**
    B ← find the pixel (b, ILM[b]) between point A and point C with maximum absolute slope as point B
# Determine D
for $d = c+1, c+2, \ldots, e-1$ **do**
    D ← find the pixel (d, ILM[d]) between point C and point E with maximum absolute slope as point D
return [A, B, C, D, E]

2.3. DL models

The models consisted of three submodules based on a DL algorithm. The first was the CNN module used to extract features from OCT images. The Second was the encoding module used to encode the feature and output the embeddings of different models. The third module was the transformer module used to fuse each model's embeddings and predict postoperative BCVA. Specifically, Resnet-18 was selected as a CNN module whose depth was fit for this task, avoiding overfitting and poor efficiency. It flattened the output feature into a 1D vector with 512 dim. In the encoding module, the vector of the image feature became a token with 128 dim, while the vector of preoperative VA (1 dim) and external morphological feature (7 dim) became tokens with 32 dim. Since the Transformer has powerful capacities in multi-modal fusion (18, 19), we applied Transformer with multi-head self-attention to learn the dependence between different models. In the Transformer module, the token of each modal and a prediction token, which had the same size, were concatenated to a token sequence, and the prediction token represented the fusion result. Follow the vanilla Transformer architecture (20), each Transformer encoder layer contained Multi-Headed Self-Attention (MSA), Layer Normalization (LN), and Feed-Forward Net-work (FFN) blocks using residual connections. Specifically, the MSA was defined as follows:

$$MSA(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where the Q, K and V denoted the linear result on input feature X. Based on the training dataset, the SGD optimizer was used to optimize the model to minimize the root mean square error (RMSE) loss function. We assume that the VA prediction is $\tilde{y}_i$ and true postoperative VA is y_i. The RMSE loss function is formulated as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\tilde{y}_i - y_i)^2}$$

The maximum number of training epochs was set to be 100. The initial learning rate was set to 0.01 and attenuated by 0.1 every 40 epochs. Figure 2 shows the workflow of the DL models.

The DL models were built based on different input information as follows:

Model I: an OCT image of horizontal B-scan,

Model II: an OCT image of vertical B-scan,

Model III: two OCT images of horizontal and vertical B-scans,

Model IV: two OCT images of horizontal and vertical B-scans, preoperative BCVA,

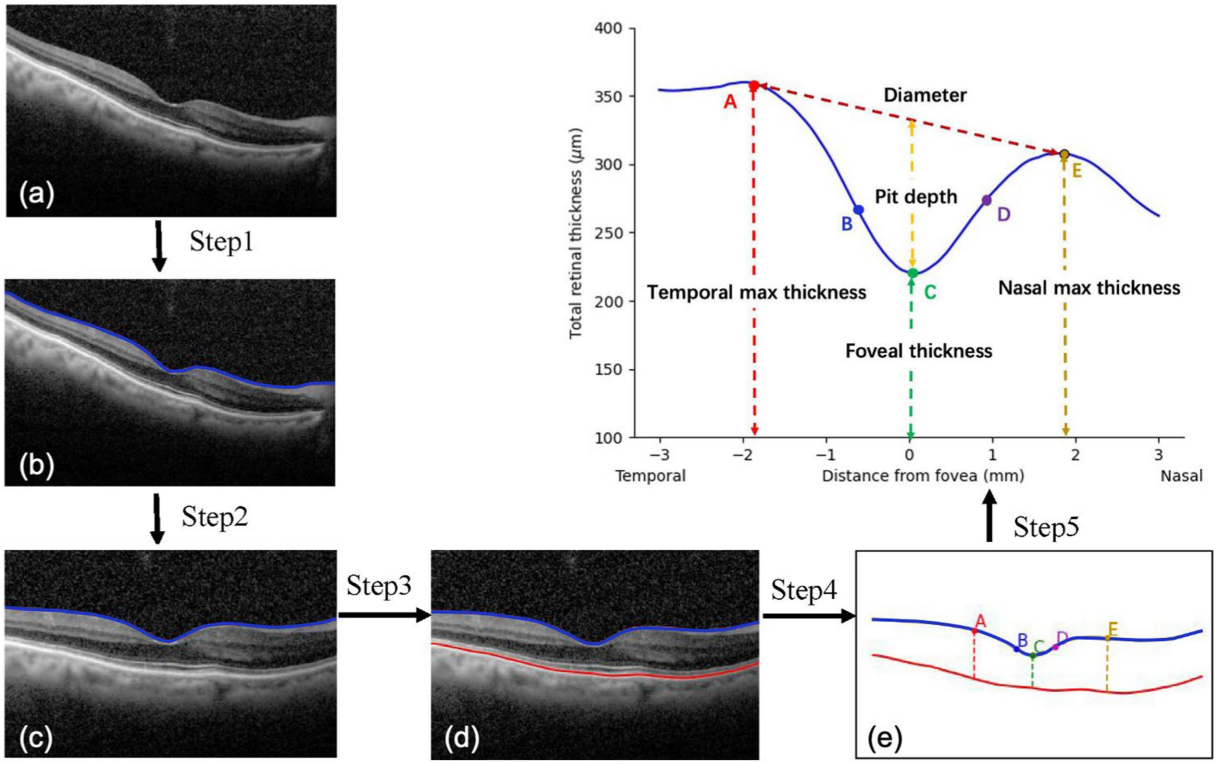

FIGURE 1
Processing of macular images. **(A,E)** A zero slope and the peak of the temporal and nasal foveal rims; **(C)** A zero slope and the center of the fovea; **(B,D)** The maximum slopes of the foveal wall.

Model V: two OCT images of horizontal and vertical B-scans, preoperative BCVA, and the indices of macula morphological features.

2.4. Model performance

Two evaluation metrics, mean absolute error (MAE) and RMSE, were applied to quantitatively measure the difference between the predicted VA and the true postoperative VA. The MAE represented the mean absolute error of the prediction values, which showed the difference between the predicted and actual values. The formula for MAE was as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\tilde{y}_i - y_i|$$

The RMSE was the square root of the mean square error (MSE). The MSE was the mean of the squared error of the prediction values. In terms of unit agreement with the original variables, the RMSE was more interpretable.

Three general classification metrics, including precision, sensitivity, and accuracy, were used to estimate the models' performance in predicting whether postoperative BCVA was improved by at least two lines in the visual chart (0.2LogMAR). Their methods of calculation were as follows:

Precision = TP/(TP + FP).

Sensitivity = TP/(TP + FN).

Accuracy = (TP + TN) / (TP + FP + TN + FN).

F1 = 2·Sensitivity·Precision/ (Sensitivity + Precision).

Here, TP is the true positive, FP is the false positive, TN is the true negative, and FN is the false negative. Receiver operating characteristic (ROC) curves for five models were calculated to obtain the area under the ROC curves (AUCs).

2.5. Statistics

Statistical analysis was performed using a commercial statistical software package (SPSS Statistics 26.0; IBM, Armonk, NY). Continuous variables were described as the mean ± standard deviation. Normal distributions for all datasets were assessed using Shapiro–Wilk normality tests. Normally distributed data were analyzed using one-way analysis of variance (ANOVA). Nonparametric data were analyzed by the Kruskal-Wallis test. The Chi-square test was used to test for categorical variables. $p < 0.05$ was considered statistically significant. The TRIPOD statement was followed.

3. Results

3.1. Patient characteristics

A total of 2,051 eyes from 2,051 patients with cataracts were included in this study. Table 1 summarizes the demographic information. No difference was found in all the clinical characteristics among the training, validation, and test datasets ($p > 0.05$).

Data were shown as the mean ± standard deviation. Abbreviation: LogMAR-logarithm of the minimum angle of resolution; BCVA-best corrected distance visual acuity.

3.2. Indices of macular morphological features

Table 2 illustrates the indices of macular morphological features. There was no significant difference among the training, validation and test datasets ($p > 0.05$). Figure 3 showed the OCT images with different changes in macular morphology. It demonstrated the normal

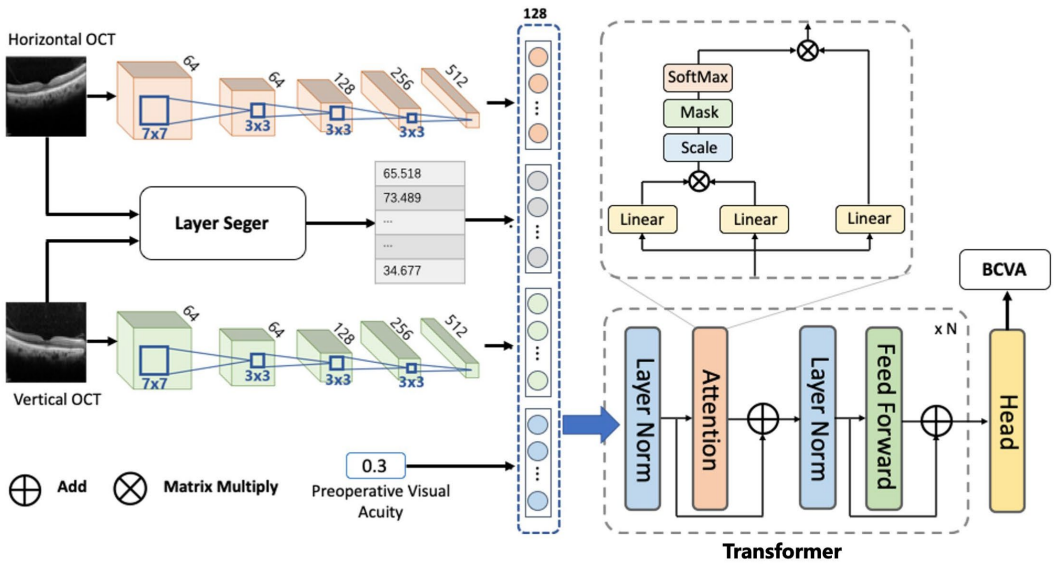

FIGURE 2
The workflow of the DL model.

(Figure 3A), macular epiretinal membrane (Figure 3B), edema (Figure 3C), and retinoschisis (Figure 3D), and the values of indices were also shown. The Grad-CAM results were shown in Figure 4, showing the highly discriminative region of OCT scans when predicting the VA.

Data were shown as the mean ± standard deviation.

TABLE 1 Demographic and clinical characteristics of the patients.

	Training (*n*=1,231)	Validation (*n*=410)	Test (*n*=410)
Number of eyes	1,231	410	410
Female gender (%)	768 (62.4%)	242 (59.0%)	246 (60.0%)
Age (years)	69.94 ± 11.1	69.33 ± 10.78	69.35 ± 10.65
Preoperative BCVA (LogMAR)	0.66 ± 0.52	0.65 ± 0.53	0.62 ± 0.51
Postoperative BCVA (LogMAR)	0.17 ± 0.32	0.17 ± 0.32	0.17 ± 0.32
Difference between postoperative BCVA and preoperative BCVA (LogMAR)	−0.48 ± 0.45	−0.47 ± 0.43	−0.44 ± 0.4

3.3. Prediction of postoperative BCVA

Table 3 presents the performance of all five models in predicting exact postoperative BCVA in the validation and test datasets. Compared with the model I-III, model IV showed better predictive performance in the validation (MAE = 0.1355logMAR, RMSE = 0.2307logMAR) and test (MAE = 0.1303logMAR, RMSE = 0.2566logMAR) datasets. When the detection and analysis of macular morphology indices were added to OCT images, the performance of model V was greatly promoted, with the lowest MAE

TABLE 2 The macular morphology detection values.

	Training	Validation	Test
Foveal thickness (μm)	277.4 ± 310.32	259.62 ± 267.34	265.16 ± 288.4
Foveal pit depth (μm)	89.53 ± 76.38	89.95 ± 74.87	89.3 ± 81.36
Foveal pit diameter (μm)	2,199.19 ± 999.12	2,204.41 ± 1004.19	2,218.88 ± 945.39
Temporal max thickness (μm)	371.64 ± 317.16	352.68 ± 276.84	358.78 ± 294.41
Nasal max thickness (μm)	370.27 ± 314.69	351.71 ± 277.75	357.4 ± 296.17
Temporal foveal slope (°)	10.06 ± 5.83	10.02 ± 6.68	10.14 ± 6.67
Nasal foveal slope (°)	10.41 ± 5.79	10.41 ± 6.97	10.47 ± 6.71

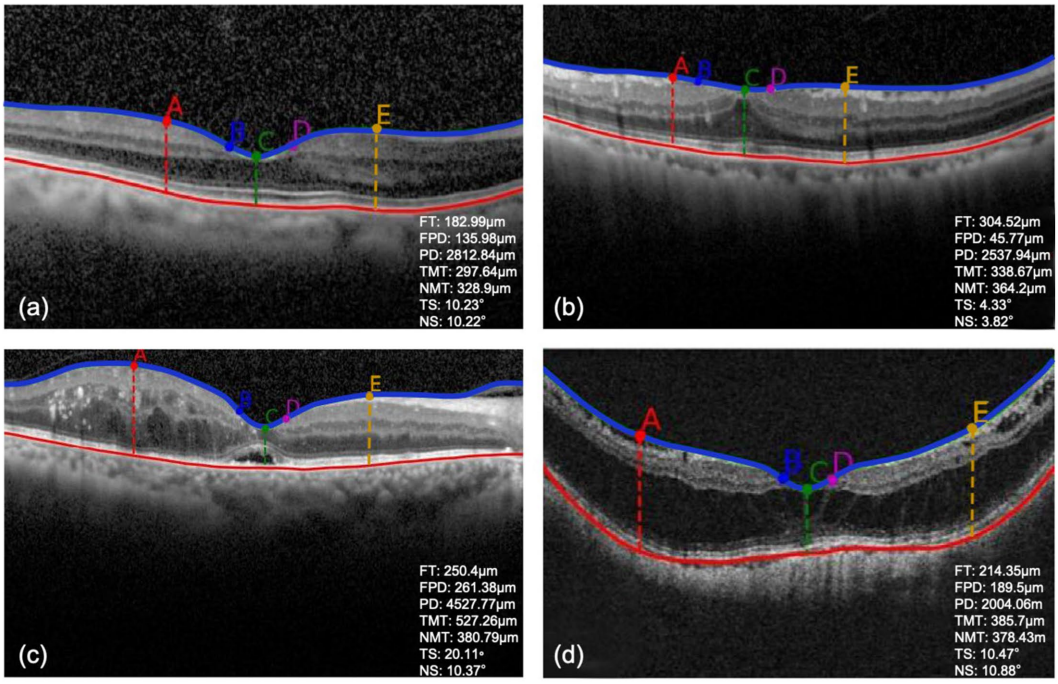

FIGURE 3
(A–D) The detection results of OCT images with different macular morphological changes. FT, foveal thickness; FPD, foveal pit depth; PD, pit diameter; TMT, temporal max thickness; NMT, nasal max thickness; TS, temporal foveal slope; NS, nasal foveal slope.

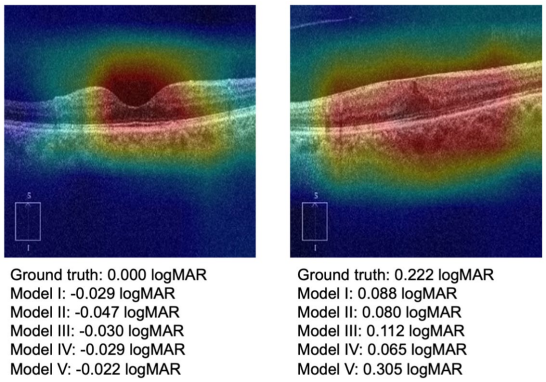

FIGURE 4
Grad-CAM results of normal macular and macular with epiretinal membrane.

TABLE 3 The performance of five models in predicting postoperative BCVA in the validation and test datasets.

Models	Validation dataset		Test dataset	
	MAE	RMSE	MAE	RMSE
I	0.1499	0.2761	0.1390	0.2624
II	0.1335	0.2504	0.1377	0.2778
III	0.1392	0.2613	0.1356	0.2730
IV	0.1355	0.2307	0.1303	0.2566
V	0.1250	0.2284	0.1194	0.2362

(0.1250 and 0.1194logMAR) and RMSE (0.2284 and 0.2362logMAR) in the validation and test datasets, respectively.

3.4. Prediction of postoperative BCVA improvement (0.2logMAR)

Table 4 presents the performance of all five models in predicting postoperative BCVA improvement (0.2logMAR) in the validation and test datasets. Compared with the model I-III, model IV showed better prediction performance in the validation (precision = 89.4%, sensitivity = 91.7%, accuracy = 85.7%, F1 = 90.5%, AUC = 0.816) and test (precision = 90.9%, sensitivity = 91.2%, accuracy = 86.6%, F1 = 91%, AUC = 0.804) datasets. Model V provided the highest precision (90.7 and 91.7%), sensitivity (93.4 and 93.8%), accuracy (88 and 89%), F1 (92 and 92.7%) and AUCs (0.856 and 0.854) in the validation and test datasets, when macular morphology indices on OCT images were detected and analyzed ($p < 0.05$).

4. Discussion

In the present study, we constructed AI prediction models for postoperative BCVA of patients with age-related cataracts based on macular OCT images and preoperative BCVA using a DL method. AI models could help doctors judge the visual outcomes of cataract surgery and aid patients in setting their surgical expectations to a reasonable level.

Ever since the first cataract surgery was performed, the evaluation of postoperative VA has been a major concern for both doctors and patients (21). Many ophthalmological examinations have been used to predict postoperative VA, such as potential acuity meter (PAM) (22–24), laser interferometer (LI) (25–27), critical flicker frequency (28–30), electrophysiological examination (31) and so on. PAM is a device combined with a slit lamp that projects a light source containing a visual chart. Light is projected through the opacified refractive media to the retina, thus providing a prediction of the patient's postoperative VA. Gus et al. have compared the accuracy of PAM in predicting VA in patients with cataracts with different degrees of lens opacities (22). VA at 3 months post-operatively was considered to be accurate if it was between the upper and lower rows of the predicted VA. It was found that for mild to moderate cataracts, the accuracy of PAM ranged from 50 to 58.3%, for patients with severe opacities, the accuracy was only 27.8%, and for patients with extremely severe opacities, the accuracy was only 6.7%. LI projects two coherent beams from a He-Ne laser into the pupil to produce interference fringes, the width of which depends on the distance between the two beams and can be varied to correspond to the visual acuity chart by changing the width of the interference fringes (25). Similar to PAM, studies have found the accuracy of LI also needs to be enhanced (25–27). The photoreceptor cells of the retina produce a complex series of electrical responses upon light stimulation that can be recorded by visual electrophysiological examination. Based on the characteristics of its waveform, it can basically reflect the functional condition of the retina and the status of the optic nerve. Salvador et al. have performed visual electrophysiological examinations on mature cataract patients and found that the magnitude of each parameter in the visual electrophysiological examination was not affected by the degree of lens opacities (31). Analysis of waveform amplitude and latency prolongation time could indirectly reflect whether the postoperative visual prognosis was good to a certain extent. In addition, some studies have tried to explore the correlation between massive preoperative biological parameters and postoperative BCVA in patients with cataracts (15, 32–34). The preoperatively observed macular disease is found to be the factor most strongly associated with poor visual outcomes (15). However, the accuracy of the methods mentioned above needs to be further improved, and some require the subjective cooperation of patients, which is difficult in some cases. In the present study, we extracted the macular morphological features on OCT images and developed AI models to predict the postoperative VA in patients with cataracts. The prediction performance of the models was evaluated, and satisfactory prediction results were achieved.

With the rapid development of computational power and learning algorithms, AI is widely used in the field of ophthalmology, and it has also been used in predicting postoperative VA in patients with cataracts. Alexeeff et al. have compared the accuracy of three machine learning models for predicting BCVA following cataract surgery using data recorded in the electronic health system (35). Preoperative BCVA, age, and age-related macular degeneration are found to be the most critical variables in the final model, which are the key factors of our research. However, they just roughly distinguish patients with better or worse postoperative BCVA than 20/50. None of the three algorithms can accurately predict postoperative VA. Wei et al. have

TABLE 4 The performance of five models in predicting postoperative BCVA improvement in the validation and test dataset.

	Model	Precision (%)	Sensitivity (%)	Accuracy (%)	F1 (%)	AUC	*p* value
Validation dataset	I	90.1	86.7	83.2	88.4	0.813	0.016*
	II	90.3	89.4	85.1	89.8	0.815	0.025*
	III	88.6	92.1	85.4	90.3	0.826	0.083
	IV	89.4	91.7	85.7	90.5	0.816	0.015*
	V	90.7	93.4	88	92	0.856	–
Test dataset	I	90.8	83.4	81.2	86.9	0.802	0.001*
	II	90.5	84	81.5	87.1	0.786	<0.001*
	III	90	90.6	85.4	90.3	0.81	0.017*
	IV	90.9	91.2	86.6	91	0.804	0.003*
	V	91.7	93.8	89	92.7	0.854	–

*Statistically significant versus Model V.

developed an OCT-based DL approach to predict postoperative BCVA in patients with high myopic (13). The ensemble model is found to show stably outstanding performance in internal and external test datasets. Xiang et al. have designed a system based on OCT images to predict the postoperative long-term BCVA of children with congenital cataracts (14). Six machine learning algorithms are applied. For 3-year predictions, the MAEs and RMSEs are 0.1482–0.2117 logMAR and 0.1916–0.2942 logMAR, and for 5-year predictions, they are 0.1198–0.1845 logMAR and 0.1692–0.2537 logMAR. Nevertheless, no anatomic or morphological macular features are incorporated into the study, and these data are less explicable. In our current study, we developed AI models based on preoperative OCT images and BCVA to predict the postoperative BCVA in patients with cataracts. The prediction performances of the models were further evaluated to clarify whether the model could accurately predict exact postoperative BCVA and whether the improvement (0.2logMAR) of postoperative BCVA could be predicted precisely. Promising results in the validation and test datasets were achieved, when the input information contained preoperative OCT images with horizontal and vertical B-scans, macular morphological feature indices, and preoperative BCVA. AI models that integrate large sample sizes of preoperative VA and macular OCT image morphological parameters are promising for postoperative VA prediction in patients with cataracts.

Further, the performance of the models was compared. When preoperative BCVA was added as input information, model IV performed better than Model I-III. It suggested that preoperative VA, affected by both cataract and fundus diseases, was a meaningful predictor of postoperative VA in patients with cataracts. Studies have shown that preoperative VA is related to postoperative VA to some extent, which is consistent with our study (15, 35). Model V containing preoperative OCT images with horizontal and vertical B-scans, macular morphological feature indices, and preoperative BCVA had a better performance in predicting postoperative BCVA, with the lowest MAE and RMSE, as well as the highest precision, sensitivity, and accuracy in the validation and test datasets, respectively. Geng et al. (36) have predicted the visual outcomes in patients undergoing macular hole surgery with several macular morphological parameters on OCT, including macular hole index, tractional hole index, hole form factor, area ratio factor (ARF), and volume ratio factor. ARF is found to efficiently express three-dimensional characteristics of the macular hole and has achieved good prediction results (sensitivity = 0.769, specificity = 0.786, AUC = 0.806). Sacconi et al. (37) have identified that structural OCT features are associated with BCVA outcomes in patients with type 3 macular neovascularization secondary to age-related macular degeneration after 3-year treatment with anti-VEGF injections. The presence of subretinal fluid at baseline is found to be the most significant independent negative predictor of functional outcomes. These studies have proved that the morphological abnormalities of the macula are closely associated with vision in ophthalmic diseases. In our present study, compared with a single macular OCT image, the more specific macular morphological indices, the higher accuracy of the model was revealed in predicting VA. After integrating macular morphological parameters, the prediction performance of BCVA was significantly improved, which was in consistent with the previous studies. These results suggested that OCT images, macular morphological features, and preoperative BCVA were all helpful for predicting postoperative BCVA in patients with cataracts.

Additionally, deep learning has yielded fresh perspectives on the formerly elusive correlation between retinal morphology and physiological parameters. Avinash et al. have trained a DL model to predict the refractive error from fundus images using two different datasets with high accuracy (38). For all types of refractive errors, both individual and mean attention maps, emphasizing the features that are indicative of refractive error, exhibited a distinct focus on the fovea. Yoo et al. (39) have evaluated a DL model for estimating uncorrected refractive error using retinal OCT images containing the retina and optic disc. It has been discovered that morphological features in OCT images contribute to detecting eyes with refractive errors. These studies suggest that the retina contains a wealth of previously unknown information, and that the combination of AI and retinal images may potentially lead to unexpected breakthroughs in the future.

Our study has several limitations. A study including data from multiple medical centers with a larger sample size will be helpful for AI model training. Besides, in the current study, we extracted and

analyzed the external morphological features of the macula, and stratification within the retina may further improve the prediction performance of AI models. The model primarily focused on the external morphology of the macula and may not be optimal for diagnosing specific macular pathologies such as subretinal fluid or macular edema, without large-scale manually-labeled lesion data. Since the opacification of refractive media can interfere with the quality of macular OCT image and further affect the observation and extraction analysis of macular area morphology, patients with relatively good quality macular OCT images were mainly selected for this study, and the prediction accuracy of the model needs further study for patients with severe dense cataract. Further investigation and exploration are needed to integrate with more ophthalmic examinations, such as fundus photography and visual sensitivity, to establish a more informative database for a comprehensive assessment of the patient's eyes, leading to a more accurate prediction of postoperative VA.

5. Conclusion

In summary, our study constructed a novel DL model to predict postoperative BCVA, showing a satisfying result in predicting postoperative BCVA in patients with cataracts. A combination AI model of OCT images, macular morphological feature indices, and preoperative BCVA was helpful for predicting postoperative BCVA in patients with cataracts.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Ethics statement

The studies involving human participants were reviewed and approved by the Institutional Review Board of the Second Affiliated Hospital of Zhejiang University, School of Medicine. The patients/participants provided their written informed consent to participate in this study.

Author contributions

WX, JW, and HY created the study design. JWW and JHW performed data analysis. JWW, JHW, HY, and WX performed drafting and critical revisions of the manuscript. JWW, JHW, DC, XW, ZX, XY, SS, XL, XC, JW, HY, and WX participated in data collection. All authors contributed to the article and approved the submitted version.

Funding

This study was funded by the National Key Research and Development Program of China (2020YFE0204400).

Acknowledgments

The authors appreciate the doctors and nurses from the Eye Center of the Second Affiliated Hospital, School of Medicine, Zhejiang University, for their support of the study. College of Computer Science and Technology from Zhejiang University is acknowledged for providing technical support for the study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Steinmetz JD, Bourne R, Briant P, Flaxman S, Taylor H, Jonas J, et al. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the right to sight: an analysis for the global burden of disease study. *Lancet Glob Health*. (2021) 9:e144–60. doi: 10.1016/S2214-109X(20)30489-7

2. Liu YC, Wilkins M, Kim T, Malyugin B, Mehta JS. *Cataracts Lancet*. (2017) 390:600–12. doi: 10.1016/S0140-6736(17)30544-5

3. Puliafito CA, Hee MR, Lin CP, Reichel E, Schuman JS, Duker JS, et al. Imaging of macular diseases with optical coherence tomography. *Ophthalmology*. (1995) 102:217–29. doi: 10.1016/s0161-6420(95)31032-9

4. Wilkins JR, Puliafito CA, Hee MR, Duker JS, Reichel E, Coker JG, et al. Characterization of epiretinal membranes using optical coherence tomography. *Ophthalmology*. (1996) 103:2142–51. doi: 10.1016/s0161-6420(96)30377-1

5. Kawczynski MG, Bengtsson T, Dai J, Hopkins JJ, Gao SS, Willis JR. Development of deep learning models to predict best-corrected visual acuity from optical coherence tomography. *Transl Vis Sci Technol*. (2020) 9:51. doi: 10.1167/tvst.9.2.51

6. Endo H, Kase S, Tanaka H, Takahashi M, Katsuta S, Suzuki Y, et al. Factors based on optical coherence tomography correlated with vision impairment in diabetic patients. *Sci Rep*. (2021) 11:3004. doi: 10.1038/s41598-021-82334-w

7. Orlando JI, Gerendas BS, Riedl S, Grechenig C, Breger A, Ehler M, et al. Automated quantification of photoreceptor alteration in macular disease using optical coherence tomography and deep learning. *Sci Rep*. (2020) 10:5619. doi: 10.1038/s41598-020-62329-9

8. Sandberg MA, Brockhurst RJ, Gaudio AR, Berson EL. The association between visual acuity and central retinal thickness in retinitis pigmentosa. *Invest Ophthalmol Vis Sci*. (2005) 46:3349–54. doi: 10.1167/iovs.04-1383

9. Chen SC, Chiu HW, Chen CC, Woung LC, Lo CM. A novel machine learning algorithm to automatically predict visual outcomes in intravitreal ranibizumab-treated patients with diabetic macular edema. *J Clin Med*. (2018) 7:475. doi: 10.3390/jcm7120475

10. Rohm M, Tresp V, Müller M, Kern C, Manakov I, Weiss M, et al. Predicting visual acuity by using machine learning in patients treated for neovascular age-related macular degeneration. *Ophthalmology*. (2018) 125:1028–36. doi: 10.1016/j.ophtha.2017.12.034

11. Huang CY, Kuo RJ, Li CH, Ting DS, Kang EYC, Lai CC, et al. Prediction of visual outcomes by an artificial neural network following intravitreal injection and laser therapy for retinopathy of prematurity. *Br J Ophthalmol*. (2020) 104:1277–82. doi: 10.1136/bjophthalmol-2019-314860

12. Mao J, Fang D, Chen Y, Tao J, Wu M, Wu S, et al. Prediction of visual acuity after cataract surgery using optical coherence tomography findings in eyes with retinitis pigmentosa. *Ophthalmic Surg Lasers Imaging Retina*. (2018) 49:587–94. doi: 10.3928/23258160-20180803-06

13. Wei L, He W, Wang J, Zhang K, du Y, Qi J, et al. An optical coherence tomography-based deep learning algorithm for visual acuity prediction of highly myopic eyes after cataract surgery. *Front Cell Dev Biol*. (2021) 9:652848. doi: 10.3389/fcell.2021.652848

14. Xiang Y, Chen J, Xu F, Lin Z, Xiao J, Lin Z, et al. Longtime vision function prediction in childhood patients with cataracts based on optical coherence tomography images. *Front Bioeng Biotechnol*. (2021) 9:646479. doi: 10.3389/fbioe.2021.646479

15. Yoeruek E, Deuter C, Gieselmann S, Saygili O, Spitzer MS, Tatar O, et al. Long-term visual acuity and its predictors after cataract surgery in patients with uveitis. *Eur J Ophthalmol*. (2010) 20:694–701. doi: 10.1177/112067211002000409

16. Liu XT, Shen MX, Chen C, Huang SH, Zhuang XR, Ma QK, et al. Foveal pit morphological changes in asymptomatic carriers of the G11778A mutation with Leber's hereditary optic neuropathy. *Int J Ophthalmol*. (2020) 13:766–72. doi: 10.18240/ijo.2020.05.11

17. Dubis AM, McAllister JT, Carroll J. Reconstructing foveal pit morphology from optical coherence tomography imaging. *Br J Ophthalmol*. (2009) 93:1223–7. doi: 10.1136/bjo.2008.150110

18. Dai Y, Gao Y, Liu F. TransMed: transformers advance multi-modal medical image classification. *Diagnostics*. (2021) 11:1384. doi: 10.3390/diagnostics11081384

19. Huang Z, Zeng Z, Liu B, Fu D, Fu J. (2020). Pixel-bert: aligning image pixels with text by deep multi-modal transformer. arXiv [Preprint]. doi: 10.48550/arXiv.2004.00849

20. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. (2017). Attention is all you need. arXiv [Preprint]. doi: 10.48550/arXiv.1706.03762.

21. Jaffe NS. History of cataract surgery. *Ophthalmology*. (1996) 103:S5–S16. doi: 10.1016/s0161-6420(96)30760-4

22. Gus PI, Kwitko I, Roehe D, Kwitko S. Potential acuity meter accuracy in patients with cataracts. *J Cataract Refract Surg*. (2000) 26:1238–41. doi: 10.1016/s0886-3350(00)00409-0

23. Devereux CJ, Rando A, Wagstaff CM, Story IH. Potential acuity meter results in patients with cataracts. *Clin Exp Ophthalmol*. (2000) 28:414–8. doi: 10.1046/j.1442-9071.2000.00349.x

24. Cuzzani OE, Ellant JP, Young PW, Gimbel HV, Rydz M. Potential acuity meter versus scanning laser ophthalmoscope to predict visual acuity in patients with cataracts. *J Cataract Refract Surg*. (1998) 24:263–9. doi: 10.1016/s0886-3350(98)80209-5

25. Datiles MB, Edwards PA, Kaiser-Kupfer MI, McCain L, Podgor M. A comparative study between the PAM and the laser interferometer in cataracts. *Graefes Arch Clin Exp Ophthalmol*. (1987) 225:457–60. doi: 10.1007/BF02334176

26. Miller ST, Graney MJ, Elam JT, Applegate WB, Freeman JM. Predictions of outcomes from cataract surgery in elderly persons. *Ophthalmology*. (1988) 95:1125–9. doi: 10.1016/s0161-6420(88)33049-6

27. Lasa MS, Datiles MB, Freidlin V. Potential vision tests in patients with cataracts. *Ophthalmology*. (1995) 102:1007–11. doi: 10.1016/s0161-6420(95)30921-9

28. Vianya-Estopà M, Douthwaite WA, Pesudovs K, Noble BA, Elliott DB. Development of a critical flicker/fusion frequency test for potential vision testing in media opacities. *Optom Vis Ence*. (2004) 111:2317–8. doi: 10.1016/j.ophtha.2004.09.012

29. Romo GBD, Douthwaite WA, Elliott DB. Critical flicker frequency as a potential vision technique in the presence of cataracts. *Invest Ophthalmol Vis Sci*. (2005) 46:1107–12. doi: 10.1167/iovs.04-1138

30. Douthwaite WA, Vianya-Estopà M, Elliott DB. Predictions of postoperative visual outcome in subjects with cataract: a preoperative and postoperative study. *Br J Ophthalmol*. (2007) 91:638–43. doi: 10.1136/bjo.2006.093401

31. Perez Salvador GE, Perez Salvador JL. Variability of electro-physiological readings in mature cataract. *Arch Soc Esp Oftalmol*. (2002) 77:543–51.

32. Chak M, Wade A, Rahi JS. British congenital cataract interest group. Long-term visual acuity and its predictors after surgery for congenital cataract: findings of the British congenital cataract study. *Invest Ophthalmol Vis Sci*. (2006) 47:4262–9. doi: 10.1167/iovs.05-1160

33. Bonaparte LA, Trivedi RH, Ramakrishnan V, Wilson ME. Visual acuity and its predictors after surgery for bilateral cataracts in children. *Eye (Lond)*. (2016) 30:1229–33. doi: 10.1038/eye.2016.166

34. Lim ME, Minotti SC, D'Silva C, Reid RJ, Schlenker MB, Ahmed IK. Predicting changes in cataract surgery health outcomes using a cataract surgery appropriateness and prioritization instrument. *PLoS One*. (2021) 16:e0246104. doi: 10.1371/journal.pone.0246104

35. Alexeeff SE, Uong S, Liu L, Shorstein NH, Carolan J, Amsden LB, et al. Development and validation of machine learning models: electronic health record data to predict visual acuity after cataract surgery. *Perm J*. (2020) 25:1. doi: 10.7812/TPP/20.188

36. Geng XY, Wu HQ, Jiang JH, Jiang K, Zhu J, Xu Y, et al. Area and volume ratios for prediction of visual outcome in idiopathic macular hole. *Int J Ophthalmol*. (2017) 10:1255–60. doi: 10.18240/ijo.2017.08.12

37. Sacconi R, Forte P, Tombolini B, Grosso D, Fantaguzzi F, Pina A, et al. OCT predictors of 3-year visual outcome for type 3 macular neovascularization. *Ophthalmol Retina*. (2022) 6:586–94. doi: 10.1016/j.oret.2022.02.010

38. Varadarajan AV, Poplin R, Blumer K, Angermueller C, Ledsam J, Chopra R, et al. Deep learning for predicting refractive error from retinal fundus images. *Invest Ophthalmol Vis Sci*. (2018) 59:2861–8. doi: 10.1167/iovs.18-23887

39. Yoo TK, Ryu IH, Kim JK, Lee IS. Deep learning for predicting uncorrected refractive error using posterior segment optical coherence tomography images. *Eye*. (2022) 36:1959–65. doi: 10.1038/s41433-021-01795-5

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 18 May 2023
DOI 10.3389/fmed.2023.1162124

OPEN ACCESS

EDITED BY
Tae-im Kim,
Yonsei University, Republic of Korea

REVIEWED BY
Tae Keun Yoo,
B&VIIT Eye Center / Refractive Surgery & AI Center, Republic of Korea
Tyler Hyungtaek Rim,
Mediwhale Inc., Republic of Korea

*CORRESPONDENCE
Dong Hui Lim
✉ ldhlse@gmail.com
Yong Man Ro
✉ ymro@kaist.ac.kr

†PRESENT ADDRESS
Tae-Young Chung,
Renew Seoul Eye Center, Seoul, Republic of Korea

‡These authors have contributed equally to this work and share first authorship

RECEIVED 09 February 2023
ACCEPTED 24 April 2023
PUBLISHED 18 May 2023

CITATION
Won YK, Lee H, Kim Y, Han G, Chung T-Y, Ro YM and Lim DH (2023) Deep learning-based classification system of bacterial keratitis and fungal keratitis using anterior segment images.
Front. Med. 10:1162124.
doi: 10.3389/fmed.2023.1162124

Deep learning-based classification system of bacterial keratitis and fungal keratitis using anterior segment images

Yeo Kyoung Won[1‡], Hyebin Lee[2‡], Youngjun Kim[1], Gyule Han[1], Tae-Young Chung[1†], Yong Man Ro[2*] and Dong Hui Lim[1,3*]

[1]Department of Ophthalmology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea, [2]Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea, [3]Department of Digital Health, Samsung Advanced Institute for Health Sciences and Technology, Sungkyunkwan University, Seoul, Republic of Korea

Introduction: Infectious keratitis is a vision threatening disease. Bacterial and fungal keratitis are often confused in the early stages, so right diagnosis and optimized treatment for causative organisms is crucial. Antibacterial and antifungal medications are completely different, and the prognosis for fungal keratitis is even much worse. Since the identification of microorganisms takes a long time, empirical treatment must be started according to the appearance of the lesion before an accurate diagnosis. Thus, we developed an automated deep learning (DL) based diagnostic system of bacterial and fungal keratitis based on the anterior segment photographs using two proposed modules, Lesion Guiding Module (LGM) and Mask Adjusting Module (MAM).

Methods: We used 684 anterior segment photographs from 107 patients confirmed as bacterial or fungal keratitis by corneal scraping culture. Both broad- and slit-beam images were included in the analysis. We set baseline classifier as ResNet-50. The LGM was designed to learn the location information of lesions annotated by ophthalmologists and the slit-beam MAM was applied to extract the correct feature points from two different images (broad- and slit-beam) during the training phase. Our algorithm was then externally validated using 98 images from Google image search and ophthalmology textbooks.

Results: A total of 594 images from 88 patients were used for training, and 90 images from 19 patients were used for test. Compared to the diagnostic accuracy of baseline network ResNet-50, the proposed method with LGM and MAM showed significantly higher accuracy (81.1 vs. 87.8%). We further observed that the model achieved significant improvement on diagnostic performance using open-source dataset (64.2 vs. 71.4%). LGM and MAM module showed positive effect on an ablation study.

Discussion: This study demonstrated that the potential of a novel DL based diagnostic algorithm for bacterial and fungal keratitis using two types of anterior

segment photographs. The proposed network containing LGM and slit-beam MAM is robust in improving the diagnostic accuracy and overcoming the limitations of small training data and multi type of images.

KEYWORDS

anterior segment image, bacterial keratitis, convolutional neural network (CNN), deep learning (DL), fungal keratitis, infectious keratitis, lesion guiding module (LGM), mask adjusting module (MAM)

1. Introduction

Infectious keratitis is a common cause of permanent blindness worldwide and can cause serious complications such as corneal perforation, corneal opacification, and endophthalmitis if not properly treated (1–6). Approximately 2,300,000 cases of microbial keratitis (including those caused by bacteria, fungi, viruses, and *Acanthamoeba*) occur annually in South Korea, where bacteria still dominate as the causative organisms of the disease (5). It is known to show various patterns depending on the region, climate, and country. For example, in temperate climates, fungal and mixed infections are more common than in tropical and semi-tropical areas. From an epidemiological point of view, ocular trauma and contact lens-associated keratitis have been increasing in recent years (7, 8).

The selection of an effective antimicrobial agent requires the identification of the causative microorganism. The gold standard for diagnosis is corneal scraping and culture, but it is not always available, and bacterial or fungal growth on culture plates takes several days or weeks (9–11). Even if it is actually microorganism positive, the result may be negative and the lesion may worsen while waiting for the result. Therefore, empirical therapy with broad-spectrum antibiotics, antifungals, and antiviral agents should be initiated based on the clinical experience of the ophthalmologist, based on the shape, size, depth, and location of the lesion, before culture results are obtained (9–12). However, bacterial and fungal keratitis are not completely distinct from each other. If patients receive unnecessary or late treatment due to an incorrect diagnosis, it may result in poor outcomes for the sufferer's vision, poor quality of life, and increased medical expenses.

Because the deep learning approach has shown remarkable performance in various image processing tasks such as classification and object detection, it has been applied in numerous research fields. Deep learning, using various types of medical images, is also used for the accurate diagnosis and treatment of many ocular diseases. As a result of the development of the methods based on deep learning, the diagnostic performance has been equivalent to or even surpassed the diagnostic ability of clinicians (13–15). Therefore, we expect that the application of deep learning in keratitis diagnosis can assist clinicians in reducing misdiagnoses and improving medical equity and accessibility to medical care.

In this context, we propose a deep learning-based computer-aided diagnosis (CAD) network that classifies and diagnoses bacterial and fungal keratitis combining with two novel modules which can improve keratitis diagnosis accuracy and predict more accurate lesion areas than conventional models.

2. Materials and methods

2.1. Study approval

This study was performed at Samsung Medical Center (SMC) and Korea Advanced Institute of Science and Technology (KAIST) according to the tenets of the Declaration of Helsinki. The Institutional Review Board of SMC (Seoul, Republic of Korea) approved this study (SMC 2019-01-014).

2.2. Participants and data collection

A retrospective analysis of the medical records of patients who had been diagnosed and treated for infectious keratitis (bacterial and fungal keratitis) at the SMC between January 1, 2002, and December 31, 2018, was conducted. All the patients underwent corneal scraping and culture; other forms of keratitis, such as viral or acanthamoeba keratitis, were excluded in this study.

Anterior segment image dataset, called the SMC dataset, is a set of anterior segment images collected from 107 patients. It consists of broad-beam and slit-beam anterior segment images (**Figure 1**) (16). A total of 594 images from 88 patients were collected for training splits, and 90 images from 19 patients were collected for the test split. The training set comprised 361 images of 64 bacterial keratitis and 233 images of 24 fungal keratitis. The test set comprised 46 images of 13 bacterial keratitis and 24 images of 6 fungal keratitis. None of the patients belonged to both the training and test splits simultaneously. For the experiment, each image was resized to 500 pixels × 750 pixels. Three ophthalmologists (Y.K., T-YC., and D.H.L.) annotated lesions on images related to the diagnosis of keratitis.

To verify the performance of the proposed network, we made the open source dataset consisted of 98 anterior segment images which were collected from Google image search and ophthalmology textbooks (17–19), and used it only as a test split.

The distribution of the images is shown in **Table 1**.

2.3. Proposed network

An overview of the entire study design and the proposed network framework is shown in **Figure 2**. It contains two proposed modules: the lesion guiding module (LGM) and slit-beam mask adjusting module (MAM). Each module was attached to the main

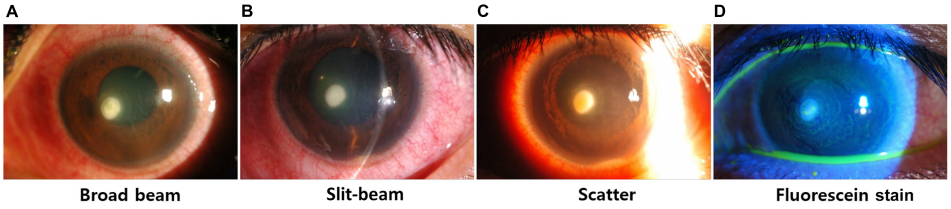

FIGURE 1
Examples of various types of anterior segment images. **(A)** Broad-beam image, **(B)** slit-beam image, **(C)** scatter image, **(D)** fluorescein stain.

TABLE 1 The distribution of images in SMC dataset and open source dataset.

	SMC dataset				Open source dataset	
	Training split		Test split		Test split	
	Broad-beam	Slit-beam	Broad-beam	Slit-beam	Broad-beam	Slit-beam
Bacterial keratitis	149	212	25	21	58	4
Fungal keratitis	99	134	19	25	32	4

SMC, Samsung Medical Center.

classifier. These two modules were introduced to overcome the aforementioned limitations for classifying the cause of keratitis. In the training stage, LGM makes the network attend to the lesion instead of other details in the anterior segment image, such as reflected light. Because its output has the form of a heat map, the detected lesion location can be obtained. MAM finely generates an optimal mask-pointing slit-beam and small parts that have less impact on the diagnosis. By comparing the masked and unmasked input images in the learning process, the network distinguishes between the necessary and unnecessary parts for diagnosis in the anterior segment image and acquires the ability to not pay attention to the unnecessary parts. We set the baseline classifier to ResNet-50 (15), and the architecture of our proposed network was based on ResNet-50. Three LGMs were inserted between each residual block of the ResNet-50.

2.3.1. LGM

Lesion Guiding Module is designed for deep learning-based diagnostic systems to learn the location information of lesions annotated by ophthalmologists in the anterior segment image. In a classifier with convolutional layers, the n LGMs are inserted between the layers, as shown in Figure 3. The bounding boxes annotated by ophthalmologists are converted to a binary mask before being input to LGM. In LGM, the intensity of the intermediate feature maps is multiplied with a binary mask during the training stage. Following this process, an important part of the diagnosis has a negative value and a relatively unrelated part of the diagnosis has a positive value. Therefore, LGM is trained to minimize the loss function and can point to lesions that are correlated to the diagnosis.

2.3.2. Slit-beam MAM

In contrast to LGM learning information about areas to be focused on, the slit-beam MAM learns information about areas that should not be focused on. By using MAM in a training phase, the single network can efficiently learn slit-beam images and broad-beam images without paying attention to the slit-beam portion of the slit-beam image (Figure 4). In addition, it can prevent the network from focusing on complex textures, such as eyelashes or blood vessels, in the anterior segment image. MAM is a module that allows the main classification network to focus on the important parts for diagnosis, so it is used only in the training phase and not in inference.

Details of the proposed network and training procedure are provided in the Supplementary material and Supplementary Figures 1–3.

2.4. Evaluation metrics

We compared the diagnostic performance of our deep learning network system with that of the baseline classifier, ResNet-50. To evaluate the accuracy of the diagnosis, the simple accuracy when the cause had the higher probability score was assumed as the final decision and areas under the receiver operating characteristic curve (AUC) were calculated. In addition, the accuracy of the detected lesion location during the diagnosis process was measured using the intersection over union (IOU) metric. The IOU is the ratio of the overlap between the predicted bounding box and the ophthalmologist's manually labeled bounding box. The closer the IOU value is to 1, the more accurate is the location of the detected lesion. We also visualized the spatial attention map of LGM. To verify the individual effects of the proposed modules on diagnostic accuracy, we conducted an ablation study.

3. Results

3.1. Performance of the metrics

Details of the performance of baseline (ResNet-50) and the proposed method are shown in Table 2. The proposed method showed the higher values in all classification performances

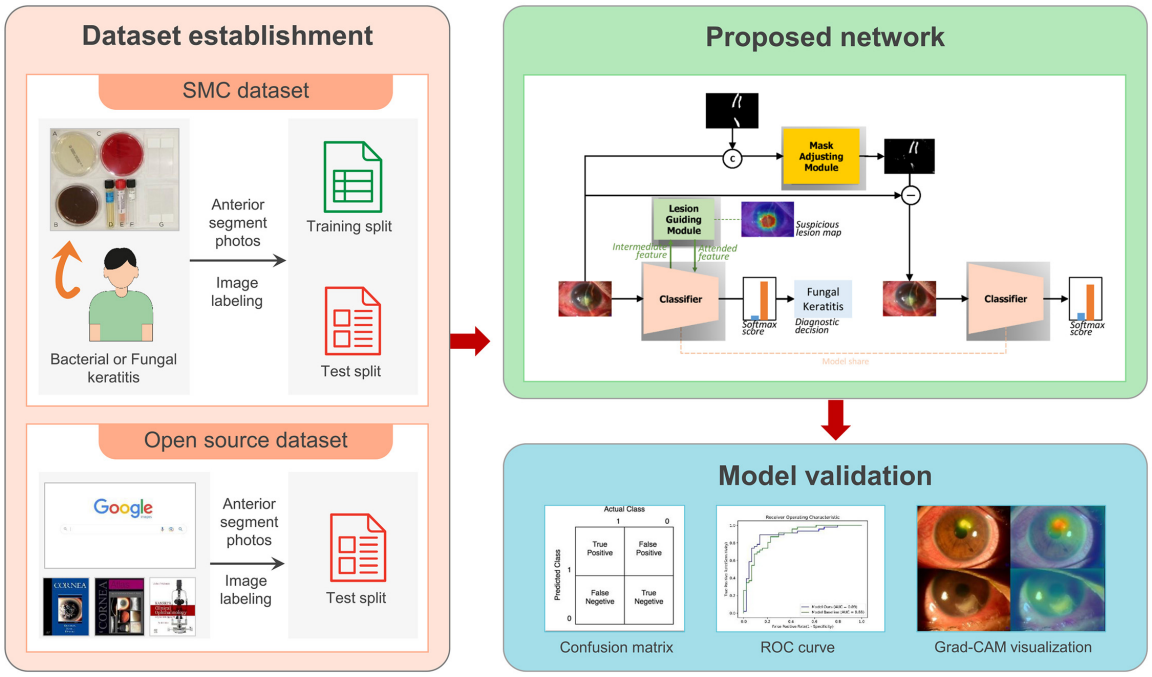

FIGURE 2
An overview of the entire study design and the proposed network framework.

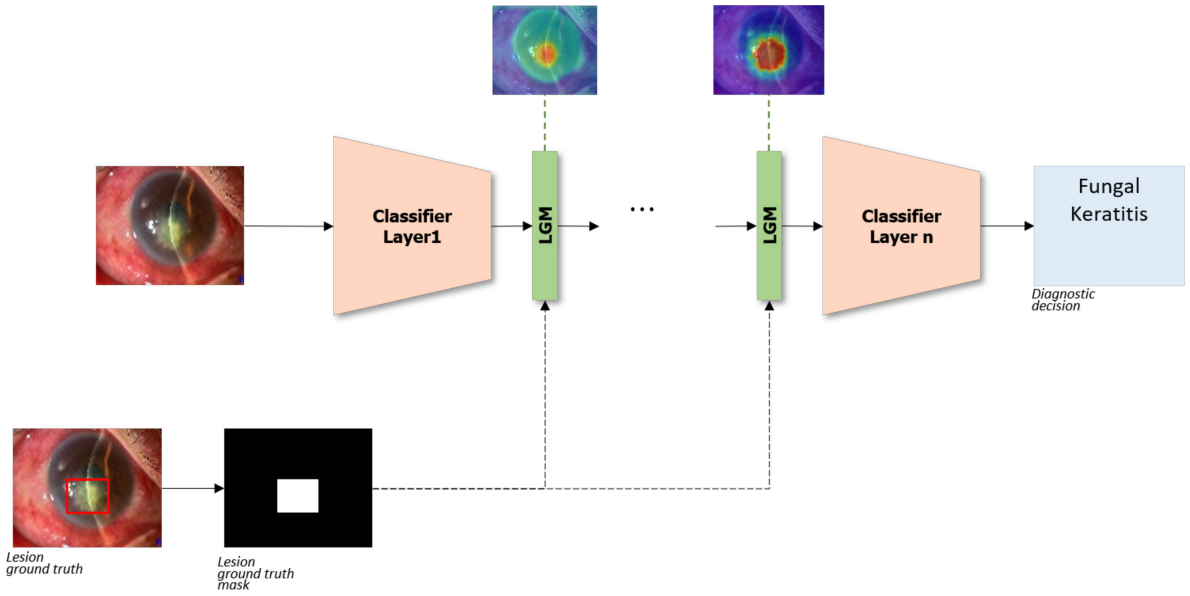

FIGURE 3
Lesion guiding modules applied on classifier. LGM, lesion guiding module.

including accuracy than baseline with both SMC dataset and open source dataset.

3.2. Diagnostic accuracy

By inferencing the two datasets, we obtained the diagnostic accuracy, as shown in Table 3. Comparing the diagnostic accuracy of the baseline network when the training image type was only a broad-beam and both a broad-beam and slit-beam, the accuracy of the broad-beam image decreased even though the number of images was more than doubled (B:0.818 → B:0.795). This result means that for different image types, simply increasing the number of images does not work properly to improve the performance. Furthermore, the network trained with broad-beam images show low diagnostic accuracy in slit-beam images (S:0.587).

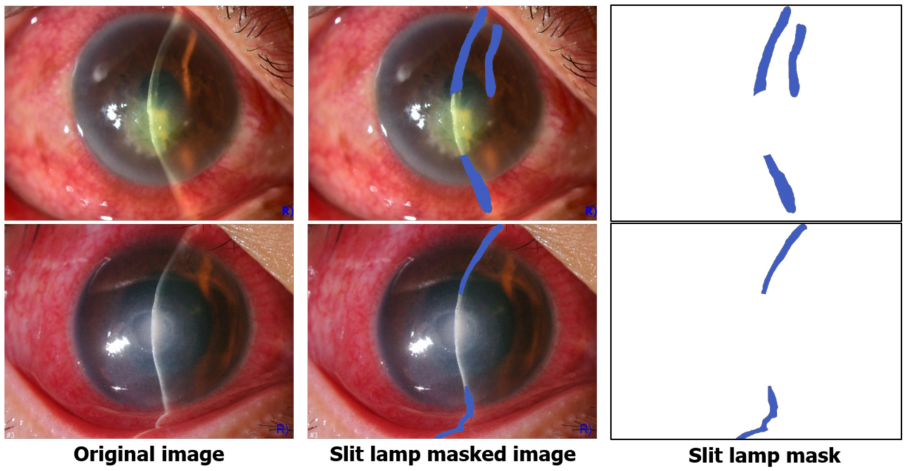

FIGURE 4
Examples of the slit-beam mask annotation.

TABLE 2 Classification performance of baseline and proposed model.

Networks	Dataset	Classification performance				
		Sensitivity	Specificity	Accuracy	Precision	F1 score
Baseline (Resnet-50)	SMC dataset	0.750	0.870	0.811	0.784	0.825
	Open source dataset	0.333	0.823	0.643	0.680	0.745
Proposed method	SMC dataset	0.864	0.891	0.878	0.872	0.882
	Open source dataset	0.417	0.887	0.714	0.724	0.797

SMC, Samsung Medical Center.

TABLE 3 Diagnostic accuracy on dataset.

Networks	Image types used in training	Accuracy	
		SMC dataset	Open source dataset
Baseline (Resnet-50)	Broad-beam	0.700 (B:0.818/S:0.587)	0.653
	Broad-beam, Slit-beam	0.811 (B:0.795/S:0.826)	0.642
Proposed method	Broad-beam, Slit-beam	0.878 (B:0.909/S:0.848)	0.714

SMC, Samsung Medical Center. B represents accuracy for only broad-beam images and S represents accuracy for only slit-beam images.

TABLE 4 Localization performance with IOU metric on the SMC dataset.

Networks	Localization method	Mean IOU (Threshold = 0.45)
Baseline (ResNet-50)	Grad-CAM	0.175
Proposed method	LGM (between ResNet layer 3 and 4)	0.489

IOU, intersection over union; SMC, Samsung Medical Center; Grad-CAM, gradient-weighted class activation mapping; LGM, lesion guiding module.

In contrast, the network with the proposed modules showed an approximately 8% increase in the SMC dataset and 7% in the open source dataset on diagnostic accuracy compared to those of the baseline network trained with both a broad-beam and slit-beam images.

3.3. Lesion localization

Supplementary Figure 4 showed the calculated IOU with various thresholds. Among these values, the best value occurred when the threshold was 0.45, as shown in Table 4. This shows that the spatial attention map of LGM, which was located between ResNet blocks 3 and 4, was more accurate than the baseline network with Grad-CAM (20).

3.4. Ablation study

Table 5 showed the results demonstrating that LGM and MAM modules had a positive effect on the accuracy of the SMC dataset and open source dataset both. Corresponding ROC curves are shown in Figure 5.

3.5. Qualitative results

Figure 6A showed that LGM points to more accurate lesion areas regardless of lesion size or shape, whereas the Grad-CAM

TABLE 5 Ablation study on dataset.

Networks	Accuracy	
	SMC dataset	Open source dataset
Baseline (ResNet-50)	0.811 (B:0.795/S:0.826)	0.642
Baseline + LGM	0.856 (B:0.864/S:0.848)	0.673
Baseline + MAM	0.833 (B:0.841/S:0.826)	0.653
Baseline + LGM + MAM	0.878 (B:0.909/S:0.848)	0.714

B represents accuracy for only broad-beam images and S represents accuracy for only slit-beam images. LGM, lesion guiding module; MAM, mask adjusting module; SMC, Samsung Medical Center.

method has high values for complex textures such as eyelashes and blood vessels (20). LGM shows that it does not attend to the reflected light in the anterior segment images. In the case of the slit-beam images, the Grad-CAM method tends to point not only to lesions but also to slit-beam (20), whereas LGM only attends to lesions without adjusting the slit-beam (**Figure 6B**). In the case of misdiagnosed images, as shown in **Figure 6C**, most of the anterior segment images were obtained from the stained eyes. At this time, it could be shown that LGM tends to point to an excessively wide area, including a lesion site or an area unrelated to diagnosis.

4. Discussion

We developed a novel deep learning algorithm that specializes in diagnosing bacterial and fungal keratitis by analyzing anterior segment images. A representative convolutional neural network (CNN), ResNet-50, was used as the backbone of the algorithm with the two proposed modules, LGM and MAM, resulting in high performance in distinguishing the images with bacterial keratitis from those with fungal keratitis.

The early and accurate diagnosis of infectious keratitis is essential for resolving the infection and minimizing corneal damage (12). Generally, the presence of an irregular/feathery border, satellite lesions, and endothelial plaque is associated with fungal keratitis, whereas a wreath infiltrate or epithelial plaque is associated with bacterial keratitis (21). However, it is difficult to distinguish exactly based on the characteristics of the infected lesions. Bacterial and fungal keratitis are often confused, especially in the early stages, but the medications used are different, and the prognosis for fungal keratitis is much worse (22–24). If patients with keratitis receive unnecessary or late treatment due to an incorrect diagnosis, complications such as corneal opacity, poor outcome for vision, or endophthalmitis may occur. Therefore, we focused on differentiating between bacterial and fungal keratitis among patients with infectious keratitis in this study.

A deep-learning approach was used to analyze a variety of medical images. Recent advances in deep learning technology in the ophthalmic field have also allowed rapid and accurate diagnosis of several ocular diseases (14, 25). However, unlike deep learning systems using optical coherence tomography and retinal fundus images, only a few studies dealing with anterior segment images have been published. In particular, in the case of infectious keratitis, it is difficult to apply a deep learning algorithm directly because it is related to lesions in various positions, and it is difficult to identify the causative pathogen without corneal culture.

In this study, we utilized a deep learning CAD network system for the differential diagnosis of bacterial and fungal keratitis based on the different shapes of corneal lesions. There are two major challenges in diagnosing the cause of keratitis by using anterior segment images. First, because of the corneal aspheric shape, the depth and extent of the infiltration lesions that were seen in the actual slit-lamp examination are not clearly visible in the image, and small lesions are often not represented in the image. Additional information such as trauma (dirty water, soil, soft contact lens,

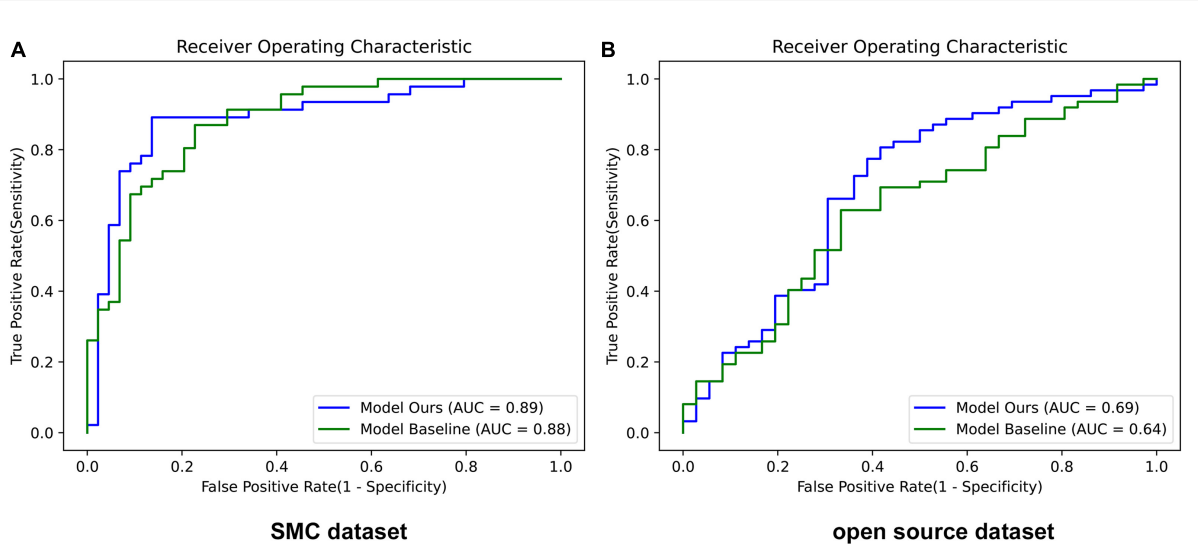

FIGURE 5
Performance of ROC curves in baseline and our proposed network. **(A)** SMC dataset and **(B)** Open source dataset. Model Ours is a new deep learning model that combines LGM and MAM with the baseline ResNet-50. Model Baseline is ResNet-50. AUC, area under the curve.

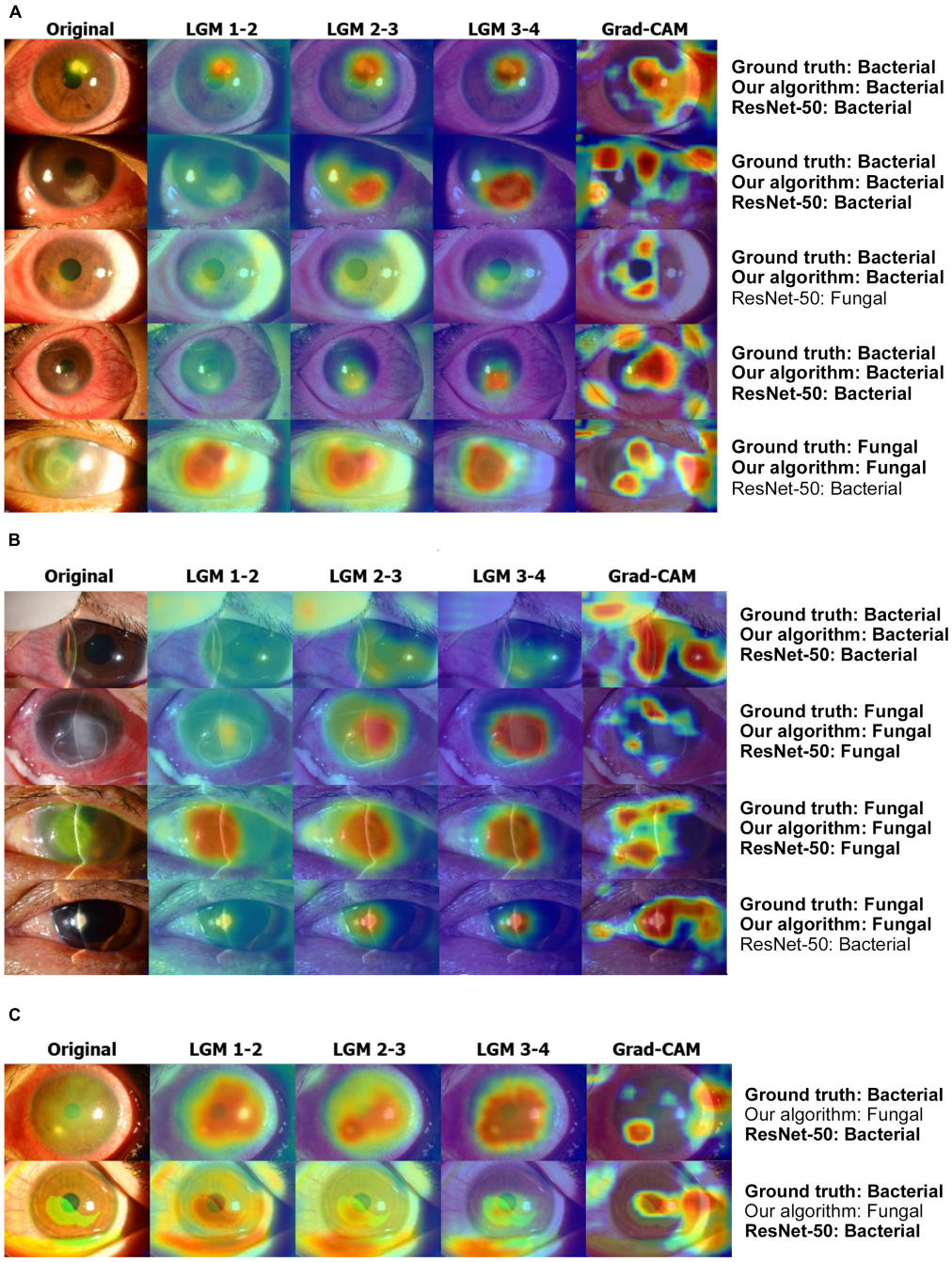

FIGURE 6
Examples of result visualization. **(A)** Broad-beam images, **(B)** slit-beam images, and **(C)** misdiagnosis cases. LGM 1-2 denotes LGM between ResNet blocks 1 and 2. LGM 2-3 denotes LGM between ResNet blocks 2 and 3. LGM 3-4 denotes LGM between ResNet blocks 3 and 4. LGM, Lesion guiding module; Grad-CAM, gradient-weighted class activation mapping.

etc.) or past medical history of the patient are very important for diagnosis. Furthermore, the appearance differs depending on the light source; therefore, keratitis is often misdiagnosed. Therefore, for an accurate diagnostic network, the knowledge of experienced ophthalmologists who can distinguish the lesion related to the diagnosis from other anatomical parts should be transferred to the diagnostic network. Second, it is difficult to obtain a large number of anterior segment images of keratitis from the various aspects of each pathogen. It is also problematic that anterior segment images can be captured in different ways. These diverse features make it difficult to learn a diagnosis network with a finite number of weights, reducing diagnostic accuracy. To solve these problems, a method for learning the features regardless of the type of anterior segment image is required. We propose two modules to solve these two challenges.

The two proposed modules, LGM and MAM, were combined with the main classifier. In the feature extraction process, LGM attends to a suspicious area related to diagnosis. Because its output

has the form of a heat map, the detected lesion location can be obtained. The slit-beam MAM is used only in the training stage; it generates an optimal mask pointing to the slit-beam and small parts that have less impact on diagnosis. The unnoticed part of the actual diagnosis process is also suppressed in the learning process of the network so that the anterior segment images with and without the slit-beam can be effectively learned together. Through training a network using this module in the proposed procedure, the network can learn different types of anterior segment image (broad-beam and slit-beam) efficiently, and we obtained a high diagnostic accuracy for infectious keratitis using different types of images.

Similar to our deep learning model, some recent other studies have applied deep learning models to distinguish patients with fungal keratitis from those with bacterial keratitis using anterior segment images. The algorithms by Hung et al. (26) based on DenseNet161 and ResNet-50 achieved average accuracies of 0.786 and 0.773, respectively. Ghosh et al. (27) constructed a model called deep keratitis based on ResNet-50 with Grad-CAM. However, its precision was 0.57 (95% CI: 0.49−0.65) which was lower than 0.878 in our results. The model with VGG19 exhibited the highest performance (0.88). Redd et al. (28) showed the highest AUC of 0.86 in MobileNet among 5 CNNs, which was nearly similar with our AUC result (0.89). Our model achieved an overall accuracy of approximately 88%, which is comparable to these previous models. Most previous studies just emphasized the application of deep learning techniques for the diagnosis of infectious keratitis. Rather than suggesting a new developed deep learning network, they analyzed the performance of each existing CNN such as ResNet, DenseNet, and ResNeXt in diagnosing the keratitis. Meanwhile, to our knowledge, we first presented a novel deep learning framework combined with two proposed modules, which is specialized in diagnosing bacterial and fungal keratitis. Furthermore, several studies have been published recently to distinguish the infectious keratitis by causative pathogens including bacterial, fungal, acanthamoeba, or viral keratitis. Zhang et al. (29) and Koyama et al. (14) provided the deep learning based diagnostic models for 4 types of infectious keratitis, and they also showed the lower accuracy in diagnosing bacteria and fungi than acanthamoeba or virus.

This study had several limitations. First, we only distinguished between bacterial and fungal keratitis. Indeed, the causes of infectious keratitis are diverse, and it is difficult to discriminate non-infectious immune keratitis in the early stages. Through subsequent studies, we need to develop a system to discriminate between the various types of keratitis. Second, the training process of LGM requires a hand-labeled bounding box annotation by specialists who are experienced in keratitis diagnosis. Therefore, a significant amount of time and resources are required to train the model. In MAM, a sophisticated pixel-level slit-beam mask is required for learning. Relatively less expertise is required than for lesion annotation, but it is also difficult to obtain masks in large quantities because of the higher accuracy required for pixel-level labeling than for bounding boxes. The expensive work required to obtain resources for learning can be a deterrent preventing the proposed network from being used in practice. Third, the usability of LGM is limited. The broad-beam anterior segment images with and without the slit-beam used in this experiment had a relatively high similarity to each other compared to other types (scatter and stain images). Therefore, by applying LGM to the original and image-level modified image, effective learning of the diagnostic system could be achieved regardless of the presence of the slit-beam in this study. However, to apply a similar mechanism to all types of anterior segment images, a method of dividing image features into parts having diagnostic information at the feature level and parts that are unnecessary for diagnosis is required. Fourth, although the LGM technique of our study allowed us to distinguish between bacterial keratitis and fungal keratitis by first accurately finding the lesion and looking at the characteristics of the lesion, exactly which part of the lesion was used to distinguish between bacterial and fungal keratitis is unknown with the results. However, LGM can identify the pathologic areas accurately by focusing only on lesions without any other noise than the Grad-Cam in the heatmaps. Finally, we combined our two modules with ResNet-50. Because some previous deep learning algorithms tend to show higher performance in other CNNs, not on ResNet-50, further studies comparing different CNNs applying our modules are required to enhance the diagnostic accuracy of infectious keratitis.

In conclusion, our deep-learning framework for the diagnosis of infectious keratitis was successfully developed and validated. LGM is presented for an accurate diagnosis by emphasizing the lesions associated with the diagnosis. To prevent the less-informative part from affecting the diagnostic result and to efficiently learn two different types of anterior segment images in a single network, we designed a new learning procedure using a masking module, MAM, to control masking in the training phase. The results showed that our proposed module had a meaningful effect in enhancing the diagnostic performance of bacterial and fungal keratitis on different anterior segment image datasets.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: The approval process from research data review committee is required and the dataset is not open to the public. Requests to access these datasets should be directed to YW, wyk900105@hanmail.net.

Ethics statement

This study was performed at Samsung Medical Center (SMC) and Korea Advanced Institute of Science and Technology (KAIST) according to the tenets of the Declaration of Helsinki. The Institutional Review Board of SMC (Seoul, Republic of Korea) approved this study (SMC 2019-01-014).

Author contributions

DL and YR designed the study, reviewed the design and results, and submitted the draft. YK, T-YC, and DL annotated lesions on images related to the diagnosis of keratitis. YW, HL, YK, and GH analyzed and interpreted the clinical data. YW and HL drafted the submitted manuscript draft. All authors have read and approved the final manuscript.

Funding

This research was supported by the Samsung Medical Center Research and Development Grant #SMO180231, #SMO1230241, and a National Research Foundation of Korea grant funded by the Korean Government's Ministry of Education (NRF-2021R1C1C1007795; Seoul, Republic of Korea) which was received by DL.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2023.1162124/full#supplementary-material

References

1. Green M, Apel A, Naduvilath T, Stapleton F. Clinical outcomes of keratitis. *Clin Exp Ophthalmol.* (2007) 35:421–6. doi: 10.1111/j.1442-9071.2007.01511.x

2. Austin A, Lietman T, Rose-Nussbaumer J. Update on the management of infectious keratitis. *Ophthalmology.* (2017) 124:1678–89. doi: 10.1016/j.ophtha.2017.05.012

3. Chirambo M, Benezra D. Causes of blindness among students in blind school institutions in a developing country. *Br J Ophthalmol.* (1976) 60:665–8. doi: 10.1136/bjo.60.9.665

4. Pascolini D, Mariotti S. Global estimates of visual impairment: 2010. *Br J Ophthalmol.* (2012) 96:614–8. doi: 10.1136/bjophthalmol-2011-300539

5. Pleyer U, Behrens-Baumann W. [Bacterial keratitis. Current diagnostic aspects]. *Ophthalmologe.* (2007) 104:9–14. doi: 10.1007/s00347-006-1466-9

6. Thylefors B, Negrel A, Pararajasegaram R, Dadzie K. Global data on blindness. *Bull World Health Organ.* (1995) 73:115–21.

7. Alexandrakis G, Alfonso E, Miller D. Shifting trends in bacterial keratitis in south Florida and emerging resistance to fluoroquinolones. *Ophthalmology.* (2000) 107:1497–502. doi: 10.1016/S0161-6420(00)00179-2

8. Garg P, Sharma S, Rao G. Ciprofloxacin-resistant *Pseudomonas* keratitis. *Ophthalmology.* (1999) 106:1319–23. doi: 10.1016/S0161-6420(99)00717-4

9. Austin A, Schallhorn J, Geske M, Mannis M, Lietman T, Rose-Nussbaumer J. Empirical treatment of bacterial keratitis: An international survey of corneal specialists. *BMJ Open Ophthalmol.* (2017) 2:e000047. doi: 10.1136/bmjophth-2016-000047

10. Hsu H, Nacke R, Song J, Yoo S, Alfonso E, Israel H. Community opinions in the management of corneal ulcers and ophthalmic antibiotics: A survey of 4 states. *Eye Contact Lens.* (2010) 36:195–200. doi: 10.1097/ICL.0b013e3181e3ef45

11. McDonald E, Ram F, Patel D, McGhee C. Topical antibiotics for the management of bacterial keratitis: An evidence-based review of high quality randomised controlled trials. *Br J Ophthalmol.* (2014) 98:1470–7. doi: 10.1136/bjophthalmol-2013-304660

12. Mun Y, Kim M, Oh J. Ten-year analysis of microbiological profile and antibiotic sensitivity for bacterial keratitis in Korea. *PLoS One.* (2019) 14:e0213103. doi: 10.1371/journal.pone.0213103

13. Gulshan V, Peng L, Coram M, Stumpe M, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA.* (2016) 316:2402–10. doi: 10.1001/jama.2016.17216

14. Koyama A, Miyazaki D, Nakagawa Y, Ayatsuka Y, Miyake H, Ehara F, et al. Determination of probability of causative pathogen in infectious keratitis using deep learning algorithm of slit-lamp images. *Sci Rep.* (2021) 11:22642. doi: 10.1038/s41598-021-02138-w

15. Xu Y, Mo T, Feng Q, Zhong P, Lai M, Chang E. Deep learning of feature representation with multiple instance learning for medical image analysis. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Florence: (2014). p. 1626–30. doi: 10.1109/ICASSP.2014.6853873

16. Zhang K, Liu X, Liu F, He L, Zhang L, Yang Y, et al. An interpretable and expandable deep learning diagnostic system for multiple ocular diseases: Qualitative study. *J Med Internet Res.* (2018) 20:e11144. doi: 10.2196/11144

17. Bowling B. *Kanski's Clinical Ophthalmology: A Systematic Approach*. 9th ed. Philadelphia, PA: Saunders Ltd (2015).

18. Krachmer H. *Cornea*. Amsterdam: Elsevier (2005).

19. Krachmer H, David A. *Cornea Atlas*. Amsterdam: Elsevier (2007).

20. Selvaraju R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D editors. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*. Venice: (2017). p. 618–26. doi: 10.1109/ICCV.2017.74

21. Dalmon C, Porco T, Lietman T, Prajna N, Prajna L, Das M, et al. The clinical differentiation of bacterial and fungal keratitis: A photographic survey. *Invest Ophthalmol Vis Sci.* (2012) 53:1787–91. doi: 10.1167/iovs.11-8478

22. Schaefer F, Bruttin O, Zografos L, Guex-Crosier Y. Bacterial keratitis: A prospective clinical and microbiological study. *Br J Ophthalmol.* (2001) 85:842–7. doi: 10.1136/bjo.85.7.842

23. Toshida H, Kogure N, Inoue N, Murakami A. Trends in microbial keratitis in Japan. *Eye Contact Lens.* (2007) 33:70–3. doi: 10.1097/01.icl.0000237825.98225.ca

24. Yeh D, Stinnett S, Afshari N. Analysis of bacterial cultures in infectious keratitis, 1997 to 2004. *Am J Ophthalmol.* (2006) 142:1066–8. doi: 10.1016/j.ajo.2006.06.056

25. Xu C, Zhu X, He W, Lu Y, He X, Shang Z, et al. editors. Fully deep learning for slit-lamp photo based nuclear cataract grading. *Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Part IV 22*. Berlin: Springer (2019).

26. Hung N, Shih A, Lin C, Kuo M, Hwang Y, Wu W, et al. Using slit-lamp images for deep learning-based identification of bacterial and fungal keratitis: Model development and validation with different convolutional neural networks. *Diagnostics.* (2021) 11:1246. doi: 10.3390/diagnostics11071246

27. Ghosh A, Thammasudjarit R, Jongkhajornpong P, Attia J, Thakkinstian A. Deep learning for discrimination between fungal keratitis and bacterial keratitis: Deep Keratitis. *Cornea.* (2022) 41:616–22. doi: 10.1097/ICO.0000000000002830

28. Redd T, Prajna N, Srinivasan M, Lalitha P, Krishnan T, Rajaraman R, et al. Image-based differentiation of bacterial and fungal keratitis using deep convolutional neural networks. *Ophthalmol Sci.* (2022) 2:100119. doi: 10.1016/j.xops.2022.100119

29. Zhang Z, Wang H, Wang S, Wei Z, Zhang Y, Wang Z, et al. Deep learning-based classification of infectious keratitis on slit-lamp images. *Ther Adv Chronic Dis.* (2022) 13:20406223221136071. doi: 10.1177/20406223221136071

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 29 June 2023
DOI 10.3389/fmed.2023.1188542

OPEN ACCESS

EDITED BY
Darren Shu Jeng Ting,
University of Nottingham, United Kingdom

REVIEWED BY
Melinda Chang,
Children's Hospital Los Angeles, United States
Aristeidis Konstantinidis,
University Hospital of Alexandroupolis, Greece

*CORRESPONDENCE
Haotian Lin
✉ haot.lin@hotmail.com
Hui Yang
✉ yanghui9@hotmail.com

†These authors share first authorship

RECEIVED 17 March 2023
ACCEPTED 12 June 2023
PUBLISHED 29 June 2023

CITATION
Liu K, Liu S, Tan X, Li W, Wang L, Li X, Xu X, Fu Y, Liu X, Hong J, Lin H and Yang H (2023) Deep learning system for distinguishing optic neuritis from non-arteritic anterior ischemic optic neuropathy at acute phase based on fundus photographs.
Front. Med. 10:1188542.
doi: 10.3389/fmed.2023.1188542

Deep learning system for distinguishing optic neuritis from non-arteritic anterior ischemic optic neuropathy at acute phase based on fundus photographs

Kaiqun Liu[1†], Shaopeng Liu[2†], Xiao Tan[3], Wangting Li[4], Ling Wang[5], Xinnan Li[1], Xiaoyu Xu[1], Yue Fu[1], Xiaoning Liu[1], Jiaming Hong[6], Haotian Lin[1*] and Hui Yang[1*]

[1]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China, [2]School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China, [3]Department of Ophthalmology, Shenzhen Aier Eye Hospital Affiliated to Jinan University, Shenzhen, Guangdong, China, [4]Department of Ophthalmology, Shenzhen Eye Hospital, Shenzhen, Guangdong, China, [5]Department of Ophthalmology, the First Hospital of Nanchang, The Third Affiliated Hospital of Nanchang University, Nanchang, China, [6]School of Medical Information Engineering, Guangzhou University of Chinese Medicine, Guangzhou, China

Purpose: To develop a deep learning system to differentiate demyelinating optic neuritis (ON) and non-arteritic anterior ischemic optic neuropathy (NAION) with overlapping clinical profiles at the acute phase.

Methods: We developed a deep learning system (ONION) to distinguish ON from NAION at the acute phase. Color fundus photographs (CFPs) from 871 eyes of 547 patients were included, including 396 ON from 232 patients and 475 NAION from 315 patients. Efficientnet-B0 was used to train the model, and the performance was measured by calculating the sensitivity, specificity, and area under the receiver operating characteristic curve (AUC). Also, Cohen's kappa coefficients were obtained to compare the system's performance to that of different ophthalmologists.

Results: In the validation data set, the ONION system distinguished between acute ON and NAION achieved the following mean performance: time-consuming (23 s), AUC 0.903 (95% CI 0.827–0.947), sensitivity 0.796 (95% CI 0.704–0.864), and specificity 0.865 (95% CI 0.783–0.920). Testing data set: time-consuming (17 s), AUC 0.902 (95% CI 0.832–0.944), sensitivity 0.814 (95% CI 0.732–0.875), and specificity 0.841 (95% CI 0.762–0.897). The performance (κ = 0.805) was comparable to that of a retinal expert (κ = 0.749) and was better than the other four ophthalmologists (κ = 0.309–0.609).

Conclusion: The ONION system performed satisfactorily distinguishing ON from NAION at the acute phase. It might greatly benefit the challenging differentiation between ON and NAION.

KEYWORDS

artificial intelligence, acute phase, optic neuritis, non-arteritic anterior ischemic optic neuropathy, color fundus photographs

Introduction

Demyelinating optic neuritis (ON) and non-arteritic anterior ischemic optic neuropathy (NAION) are acute optic neuropathies with overlapping clinical profiles. While differentiation between ON and NAION can present challenges, not all cases are complex and there may be a certain rate of misdiagnosis in some clinical scenarios, especially in underdeveloped areas (1). Multiple examinations are required to facilitate the differential diagnosis in some patients. Despite this, a misdiagnosis rate of 59.8% has been reported, with 16% of the patients receiving excessive magnetic resonance imaging, 16% receiving needless lumbar puncture, and 11% receiving unnecessary intravenous steroids application (2). The pathological mechanisms of ON and NAION were distinct, requiring different treatment regimens (3). Moreover, diligent follow-up following treatment is necessary to finalize the diagnosis in a specific challenging situation. Early diagnosis may have a significant benefit on the visual prognosis. However, the prompt diagnosis was difficult due to the shortage of experienced neuro-ophthalmologists and advanced ophthalmic devices, resulting in irreversible vision loss (4, 5).

Artificial Intelligence (AI) based on fundus photography has shown incredible differential ability on retinal disease. It has been utilized to screen for diabetic retinopathy (6), age-related macular degeneration (7), and glaucoma (8). It could even distinguish between gender and an optic disc of the left or right eye (9, 10). Also, it was used to measure retinal nerve fiber layer thickness in ON and NAION (11). Fundus photography is a routine, cost-effective, non-invasive examination technique used to diagnose ON and NAION (1). However, previous studies have not reported using AI to distinguish ON and NAION based on fundus photographs.

Therefore, we describe the use of AI analysis to help differentiate between ON and NAION from the acute phase based solely on color fundus photographs (CFPs) and to generate guidelines for distinguishing these conditions when they overlap clinically. The ultimate goal is to develop an effective, convenient, and cost-effective AI-aided diagnostic technique for improving the differential diagnosis efficiency of ON and NAION.

Materials and methods

The research adhered to the tenets of the Declaration of Helsinki and has obtained the Ethics Committee approval from Zhongshan Ophthalmic Center (ZOC) (2021KYPJ002).

Abbreviations: AI, artificial intelligence; AUC, area under the receiver operating characteristic curve; CAM, class activation map; CI, confidence interval; CNN, convolutional neural network; CFPs, color fundus photographs; CON, compression optic neuropathy; COVID-19, the novel coronavirus disease 2019; HON, hereditary optic neuropathy; ION, infectious optic neuropathy; MRI, magnetic resonance imaging; N, number of fundus photographs; NAION, non-anterior ischemic optic neuropathy; OCT, optical coherence tomography; OCT-A, optical coherence tomography angiography; ON, optic neuritis; RNFL, retinal nerve fiber layer thickness; TON, toxic optic neuropathy; VA, visual acuity; VF, visual field; VEP, visual evoked potentials; ZOC, Zhongshan Ophthalmic Center; ONION, deep learning system for distinguishing optic neuritis from non-anterior ischemic optic neuropathy.

Data inclusion

A total of 871 eyes from 547 patients with ON and NAION were included, which were within 3 months of onset. And pediatric patients were not included in our study. When bilateral eyes were affected, photographs of bilateral eyes were included; if unilateral eye was affected, only the affected eye was included, while the unaffected eye was excluded; each CFP represented a unique eye. All CFPs with ON or NAION in the data set were obtained from the neuro-ophthalmological expert (HY), and were acquired using two retinal cameras (TRC-50DX, TOPCON YAMAGATA Co., Ltd., Japan, and FF 450plus, Carl Zeiss Meditec AG, Germany). The diagnosis of each CFP was established by a combination of detailed medical history, clinical examination, ophthalmic imaging, or neurological imaging, including visual acuity (VA, Snellen charts), VF (Humphrey Visual Field Analyzer, Carl Zeiss Meditec, Dublin, CA, USA), OCT (Carl Zeiss Meditec, Jena, Germany) or Spectral OCT (Heidelberg Engineering, Heidelberg, Germany), visual evoked potentials (VEP), MRI, autoimmune antibodies tests (cell-based assay to search for AQP4 antibody), and patients with 6-month follow-up records to clarify that these records were used to further confirm the diagnosis based on the patient's disease progression and response to medication over a 6-month period. CFPs that were blurry, grossly out of focus, or did not fully display the optic disc were excluded.

The deep learning system (ONION) development

The deep learning system distinguishing ON from NAION at the acute phase using original CFPs, named ONION, was developed. We trained the ONION using one of the state-of-the-art convolutional neural network (CNN) algorithms, Efficientnet-B0. The Efficientnet-B0 algorithm was a fine-designed deep CNN model that Google proposed in recent years. Firstly, the training data set was utilized in the ONION system to optimize its parameters and produce candidate models. During the model training process, the batch size was set to 16, the focus loss was set to 0.5, and the number of training iterations was set to 50. Secondly, we performed a 10-fold cross-validation test (training 60%; validation 20%; and testing 20%) using CFPs from our data set for internal validation and model development. After the 10-fold cross-validation was completed, we then tested the cross-validation model with the best performance against the testing data set.

Performance comparison between the ONION system and ophthalmologists

All study CFPs in the testing data set were evaluated independently by five trained graders [three fellowships with 2, 3, and 5 years of fellowships (LW, XT, and YF), one board certified ophthalmologists with over 10 years of clinical experience (XLiu) and one retinal expert (XX)] who were masked to the diagnosis in either eye. These graders were classified independently based on CFPs alone, and these CFPs were classified to ON or NAION. CFPs

grading was performed at a centralized reading center using high-resolution, high-definition LCD computer displays. These display monitors were regularly color-calibrated to a color temperature of 6,500 K and gamma setting of 2.2 (Spyder4PRO; Datacolor, Lawrenceville, NJ, USA).

Visualization of photographs features

We created heatmaps through class activation mapping (CAM) to identify the key regions in the CFPs used by the EfficientNet to classified ON and NAION. Briefly, the CAM was obtained by projecting back the weights of the EfficientNet model's output layer onto the convolutional feature, which could identify the class-specific discriminative regions, namely, the specific regions in the CFPs with the highest impact on the prediction outcomes (the "warmer" the color, e.g., red, the more highly activated a particular region is). We performed CAM analyses on all 348 photographs in the validation and testing data set, and these CAM images were reviewed and interpreted by one of the authors (KL).

Statistical analysis

To assess the performance of ONION, we calculated the AUC, sensitivity, specificity, and accuracy. The 95% confidence intervals (CIs) were estimated for all the performance metrics. Cohen's kappa coefficients, AUC, sensitivity, and specificity were applied to compare the agreement scores among the results of the ONION system and ophthalmologists to the ground truth in the binary classification of ON or NAION. The Kappa result was interpreted as follows: values 0–0.2 as indicating no agreement, 0.21–0.39 as minimal, 0.40–0.59 as weak, 0.60–0.79 as moderate, 0.80–0.90 as strong, and >0.90 as almost perfect. Statistical analyses were carried out using SPSS Statistics version 23 (SPSS Inc., Chicago, IL, USA). The results of the ROCs obtained from the model and the ophthalmologists have merged accordingly into a figure using Adobe Illustrator CS6 version 24.0.1 (Adobe Inc., USA).

Results

Data set labels

The data set from 547 patients covered 871 eyes and 871 CFPs. The NAION group involved 159 patients (50.48%) in the bilateral affected eyes, while the ON group involved 164 patients (45.46%) in the bilateral affected eyes. The basic characteristics and the sample size of each data set are shown in Table 1.

Performance of the ONION system

In the validation data set, the ONION system distinguished between acute ON and NAION achieved the following mean performance: Time-consuming (23 s), AUC 0.903 (95% CI 0.827–0.947), accuracy 0.833 (95% CI 0.746–0.895), sensitivity 0.796 (95% CI 0.704–0.864), and specificity 0.865 (95% CI 0.783–0.920). Testing data set: time-consuming (17 s), AUC 0.902 (95% CI 0.832–0.944), accuracy 0.830 (95% CI 0.750–0.889), sensitivity 0.814 (95% CI 0.732–0.875), and specificity 0.841 (95% CI 0.762–0.897). The performance of the ONION system is shown in Table 2.

TABLE 1 The basic characteristics of each patient group.

Group	NAION	ON	Total
Patients, *n*	315 (57.59)	232 (42.41)	547 (100)
Eyes, *n*	475 (54.54)	396 (45.46)	871 (100)
Age (mean ± SD)	53.67 ± 9.15 (31–83)	31.55 ± 12.08 (18–84)	
Female, *n* (%)	101 (32.06)	150 (64.66)	189 (45.89)
Bilateral affected, *n* (%)	159 (50.48)	164 (70.69)	323 (59.05)
Acute phase CFPs, *n* (%)	475 (54.54)	396 (45.46)	871 (100)
Disc edema, *n* (%)	396 (83.37)	117 (29.55)	561 (58.90)
Disc hemorrhages, *n* (%)	261 (54.95)	107 (27.02)	368 (42.25)
ONION			
Training data set, *n* (%)	285 (60.00)	238 (60.00)	523 (60.04)
Validation data set, *n* (%)	95 (20.00)	79 (20.00)	174 (19.98)
Testing data set, *n* (%)	95 (20.00)	79 (20.00)	174 (19.98)

ON, optic neuritis; NAION, non-arteritic anterior ischemic optic neuropathy; CFPs, color fundus photographs.

TABLE 2 Classification performance of the ONION system.

Data set	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)
Validation data set	0.903 (0.827–0.947)	0.796 (0.704–0.864)	0.865 (0.783–0.920)	0.833 (0.746–0.895)
Testing data set	0.902 (0.832–0.944)	0.814 (0.732–0.875)	0.841 (0.762–0.897)	0.830 (0.750–0.88)

AUC, area under the receiver operating characteristic curve; CI, confidence interval.

The ONION system had as good performance as a retinal expert

As shown in Figure 1, the ONION system was comparable to that of the retinal expert [time-consuming (30 min), AUC: 0.816, sensitivity: 0.830, and specificity: 0.803] in the testing data set when distinguishing acute ON from NAION, which surpassed all the other ophthalmologists. Compared to the ground truth of the testing data set, Cohen's kappa coefficients of ONION (κ = 0.805) exhibited slightly higher performance in disease identification than those of humans, even for the retinal expert (κ = 0.309–0.749) (Supplementary Table 1).

Visualization of the ONION system

In order to investigate the feature extraction location used by the ONION system during classification decisions, we performed CAM analysis on 348 images. As shown in Figure 2, typical CAM images with correct classification exhibited strong activation in the optic disc in 320 eyes (91.95%). Among these, 166 eyes

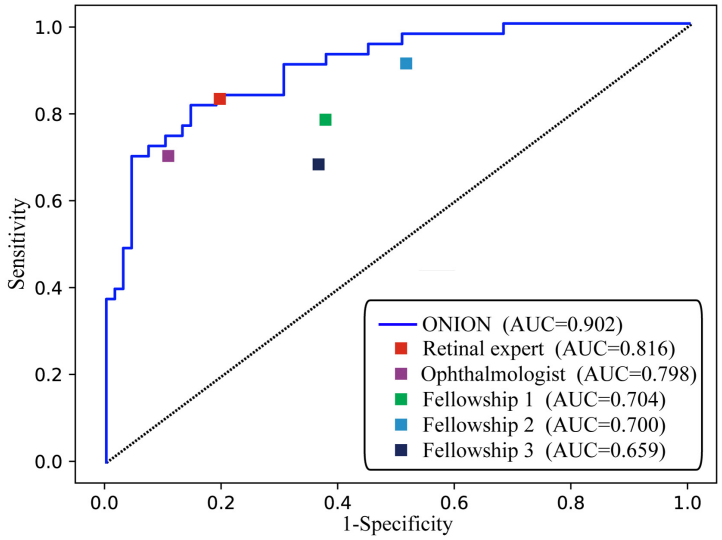

FIGURE 1
Performance of the ONION system and ophthalmologists. ROC of ONION system and five human graders for distinguishing acute ON from NAION. AUC, area under the receiver operating characteristic curve; NAION, non-arteritic anterior ischemic optic neuropathy; ON, optic neuritis.

(90.22%) were diagnosed with ON, while 154 eyes (93.90%) were diagnosed with NAION. By contrast, in typical CAM images with incorrect classification, we observed strong activation outside of the optic disc in 25 eyes (7.18%). Specifically, 18 eyes (9.78%) were diagnosed with ON and 10 eyes (6.09%) were diagnosed with NAION (**Supplementary Figure 1** and **Supplementary Table 2**).

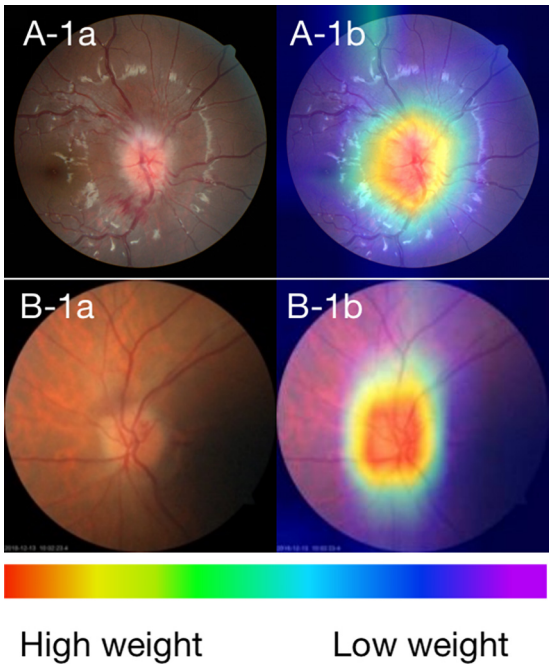

FIGURE 2
(A) Fundus photographs and **(B)** corresponding CAMs for ON and NAION. A-1a and A-1b: photographs of the ON 3 days after the onset. B-1a and B-1b: photographs of the NAION 1 week after the onset. The warmer the color, the higher its weight, and the more critical it is in a particular area.

Discussion

We attempted to distinguish between ON and NAION by AI based on CFPs alone from the acute phase, which achieved excellent performance with a mean AUC of 0.902. Its performance was comparable to that of a retinal expert and has surpassed that of other ophthalmologists. CAM revealed that the system produced a diagnosis using accurate distinguishing features, and the optic disc area was the most relevant to the diagnosis in the CFPs. These findings showed the encouraging application prospect of AI for differentiating these two common optic neuropathies in neuro-ophthalmology. In conditions where complex auxiliary examinations and neuro-ophthalmologist resources are not available, such as in underdeveloped areas, it is of unique advantage to make a reliable differentiation diagnosis based on CFPs alone.

This study contains the following advantages. Firstly, our study is the first to use AI to distinguish between ON and NAION during the acute phase. Secondly, our data is comprehensive and includes 6-month follow-up records for further diagnostic confirmation. In addition, our sample size is relatively large, which further ensures the accuracy of our model. Thirdly, we utilized the state-of-the-art EfficientNet algorithm to achieve expert-level performance on the CNN model and system implementation. The EfficientNet algorithm is widely regarded for achieving good classification accuracy through a model with moderate complexity, compared with other deep CNN algorithms. As demonstrated in the model evaluation results, our final output models achieved expert-level classification accuracy. Fourthly, we used CAM to further validate the accuracy of our AI model, which identified specific areas in the optic disc and to some extent demonstrated the interpretability of our AI model. Early diagnosis is an important issue of concern, especially for ophthalmic emergencies such as ON and NAION.

While treatment may not always improve prognosis, early diagnosis and timely treatment can still enhance the patient's quality of life.

Some attempts other than AI had been tried before to distinguish ON from NAION, such as B-scan ultrasonography (12), fluorescing angiography or laser speckle flowgraph (13), and MRI scanning (14). AI has been introduced into neuro-ophthalmology recently and has shown interesting results. AI could distinguish optic disc pallor from normal disc by machine learning and distinguish swollen discs of various optic neuropathies or pseudo-papilledema from normal discs using transfer learning based on CFPs alone (15, 16). Just recently, AI successfully distinguished papilledema caused by intracranial hypertension from other abnormalities and normal disc with an excellent AUC of 0.960 using a deep learning system based on CFPs alone (17, 18), and its performance was at least as good as two expert neuro-ophthalmologists (19). The same group further upgraded the system to distinguish mild/moderate/severe papilledema with an AUC of 0.930 (20). The differential diagnosis of ON and NAION is very challenging and essential in neuro-ophthalmology, but owing to the low incidence and lack of extensive and accurate training datasets, few studies that have used AI to distinguish between ON and NAION, and currently only Razaghi et al. have conducted research using optical coherence tomography scans as the basis for differentiating between these two conditions with the help of deep learning algorithm (11). In our study, we introduced AI into this field based on CFPs. A relatively extensive database and a reliable photograph label in our center enabled this ONION system to make an excellent performance for distinguishing ON from NAION in the acute phase. This was uncommon because even experienced ophthalmologists have difficulty identifying whether the cause of disc swelling is ON or NAION at the acute phase based on CFPs alone.

To evaluate the performance of the ONION system, we compared it to that of a retinal expert, an ophthalmologist, and three fellowships. The Cohen's kappa coefficient indicates that retinal specialist possess significant reliability in distinguishing between ON and NAION. However, less experienced professionals may face difficulty in discerning between these two conditions, which is consistent with common sense. The result showed that our ONION system could be an effective and helpful tool for assisting ophthalmologists in distinguishing ON from NAION in the acute phase. So, we expect that in the future, a rapid differential diagnosis of NAION and ON could be made even in underdeveloped areas through AI and telemedicine technology.

Class activation map visualized the learning procedure of the ONION system, showing that the optic disc area contributed the most to system detection (91.95%), whether it is centered or partially centered. our use of AI, based purely on CFPs, demonstrated good performance in distinguishing ON from NAION. This result indirectly indicates that even though both ON and NAION can lead to optic disc edema, there may still be significant or subtle differences in the surface appearance of the optic disc.

Our study has some limitations: firstly, it was a single-center study. All of the participants in our study were from China. Therefore, there was no ethnic or racial diversity within our study population. While the lack of diversity in our sample limits the generalizability of our findings to other populations. Secondly, as it is a binary classification of the ON and NAION, it could not satisfy the real-world clinical requirements where other optic neuropathies are needed to differentiate. We need the next generation of the ONION system to identify other optic neuropathies, such as papilledema caused by intracranial hypertension and hereditary optic neuropathy. Thirdly, the ONION system was based on CFPs alone, a combination of CFPs with other modality images, such as VF, OCT, and MRI, might help get a more reliable and robust diagnostic ability. Fourthly, our study only compared the AI model with retinal specialists and general ophthalmologists, without comparison to neuro-ophthalmologists. However, we note that in some regions, the distribution of neuro-ophthalmologists is limited, making it difficult for many patients with ON and NAION to receive diagnosis from a neuro-ophthalmologist. Finally, AI interpretability is a topic that requires further exploration and attention in practical applications. We are committed to improving and enriching our model and seeking more interpretable and practical algorithms, so as to provide doctors with more accurate and reliable diagnostic and treatment recommendations in a wider range of medical environments.

Conclusion

We developed a deep learning system ONION that could distinguish ON from NAION at the acute phases, with high sensitivity and specificity based on CFPs alone. Its efficiency was comparable to that of a retinal specialist. It required further prospective validation studies to prove its diagnostic ability in real-world clinical settings.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by the Zhongshan Ophthalmic Center. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this manuscript.

Author contributions

HY and HL conceived and designed the study. SL and JH developed and validated the deep learning system. KL did the literature, designed the statistic analysis, and drafted the manuscript. HY finalized the diagnoses and labeled the fundus photographs. LW, XT, YF, XLiu, and XX participated in the data testing. KL and XLi collected the data. HY, WL, and JH critically revised the manuscript. All authors reviewed and approved the final version to be published.

Funding

This research was supported by the National Natural Science Foundation of China (81870656 and 82171035), the High-Level Science and Technology Journals Projects of Guangdong Province (2021B1212010003), and Science and Technology Program of Guangzhou (202201020337).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2023.1188542/full#supplementary-material

References

1. Bianchi Marzoli S, Martinelli V. Optic neuritis: differential diagnosis. *Neurol Sci.* (2001) 22(Suppl 2):S52–4. doi: 10.1007/s100720100034

2. Stunkel L, Kung N, Wilson B, McClelland C, Van Stavern G. Incidence and causes of overdiagnosis of optic neuritis. *JAMA Ophthalmol.* (2018) 136:76–81. doi: 10.1001/jamaophthalmol.2017.5470

3. Toosy A, Mason D, Miller D. Optic neuritis. *Lancet Neurol.* (2014) 13:83–99. doi: 10.1016/S1474-4422(13)70259-X

4. Frohman L. The human resource crisis in neuro-ophthalmology. *J Neuroophthalmol.* (2008) 28:231–4. doi: 10.1097/WNO.0b013e318185e084

5. Stunkel L, Sharma R, Mackay D, Wilson B, Van Stavern G, Newman N, et al. Patient harm due to diagnostic error of neuro-ophthalmologic conditions. *Ophthalmology.* (2021) 128:1356–62. doi: 10.1016/j.ophtha.2021.03.008

6. Grossman S, Calix R, Tow S, Odel J, Sun L, Balcer L, et al. Neuro-ophthalmology in the era of COVID-19: future implications of a public health crisis. *Ophthalmology.* (2020) 127:e72–4. doi: 10.1016/j.ophtha.2020.05.004

7. Burlina P, Joshi N, Pekala M, Pacheco K, Freund D, Bressler N. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmol.* (2017) 135:1170–6. doi: 10.1001/jamaophthalmol.2017.3782

8. Li Z, He Y, Keel S, Meng W, Chang R, He M. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology.* (2018) 125:1199–206. doi: 10.1016/j.ophtha.2018.01.023

9. Korot E, Pontikos N, Liu X, Wagner S, Faes L, Huemer J, et al. Predicting sex from retinal fundus photographs using automated deep learning. *Sci Rep.* (2021) 11:10286. doi: 10.1038/s41598-021-89743-x

10. Liu T, Ting D, Yi P, Wei J, Zhu H, Subramanian P, et al. Deep learning and transfer learning for optic disc laterality detection: implications for machine learning in neuro-ophthalmology. *J Neuroophthalmol.* (2020) 40:178–84. doi: 10.1097/WNO.0000000000000827

11. Razaghi G, Hedayati E, Hejazi M, Kafieh R, Samadi M, Ritch R, et al. Measurement of retinal nerve fiber layer thickness with a deep learning algorithm in ischemic optic neuropathy and optic neuritis. *Sci Rep.* (2022) 12:17109. doi: 10.1038/s41598-022-22135-x

12. Dehghani A, Giti M, Akhlaghi M, Karami M, Salehi F. Ultrasonography in distinguishing optic neuritis from nonarteritic anterior ischemic optic neuropathy. *Adv Biomed Res.* (2012) 1:3. doi: 10.4103/2277-9175.94425

13. Kim M, Kim U. Analysis of fundus photography and fluorescein angiography in nonarteritic anterior ischemic optic neuropathy and optic neuritis. *Korean J Ophthalmol.* (2016) 30:289–94. doi: 10.3341/kjo.2016.30.4.289

14. Rizzo J III, Andreoli C, Rabinov J. Use of magnetic resonance imaging to differentiate optic neuritis and nonarteritic anterior ischemic optic neuropathy. *Ophthalmology.* (2002) 109:1679–84. doi: 10.1016/s0161-6420(02)01148-x

15. Yang H, Oh J, Han S, Kim K, Hwang J. Automatic computer-aided analysis of optic disc pallor in fundus photographs. *Acta Ophthalmol.* (2019) 97:e519–25. doi: 10.1111/aos.13970

16. Ahn J, Kim S, Ahn K, Cho S, Kim U. Accuracy of machine learning for differentiation between optic neuropathies and pseudopapilledema. *BMC Ophthalmol.* (2019) 19:178. doi: 10.1186/s12886-019-1184-0

17. Milea D, Najjar R, Zhubo J, Ting D, Vasseneix C, Xu X, et al. Artificial intelligence to detect papilledema from ocular fundus photographs. *N Engl J Med.* (2020) 382:1687–95. doi: 10.1056/NEJMoa1917130

18. Liu T, Wei J, Zhu H, Subramanian P, Myung D, Yi P, et al. Detection of optic disc abnormalities in color fundus photographs using deep learning. *J Neuroophthalmol.* (2021) 41:368–74. doi: 10.1097/WNO.0000000000001358

19. Biousse V, Newman N, Najjar R, Vasseneix C, Xu X, Ting D, et al. Optic disc classification by deep learning versus expert neuro-ophthalmologists. *Ann Neurol.* (2020) 88:785–95. doi: 10.1002/ana.25839

20. Vasseneix C, Najjar R, Xu X, Tang Z, Loo J, Singhal S, et al. Accuracy of a deep learning system for classification of papilledema severity on ocular fundus photographs. *Neurology.* (2021) 97:e369–77. doi: 10.1212/WNL.0000000000012226

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 14 August 2023
DOI 10.3389/fmed.2023.1227515

OPEN ACCESS

EDITED BY
Yi-Ting Hsieh,
National Taiwan University Hospital, Taiwan

REVIEWED BY
Chia-Ying Tsai,
Fu Jen Catholic University Hospital, Taiwan
Jo-Hsuan Wu,
University of California, San Diego,
United States

*CORRESPONDENCE
Tanvi Verma
✉ Tanvi_Verma@ihpc.a-star.edu.sg

†These authors have contributed equally to this work and share first authorship

RECEIVED 23 May 2023
ACCEPTED 28 July 2023
PUBLISHED 14 August 2023

CITATION
Verma T, Jin L, Zhou J, Huang J, Tan M, Choong BCM, Tan TF, Gao F, Xu X, Ting DS and Liu Y (2023) Privacy-preserving continual learning methods for medical image classification: a comparative analysis. *Front. Med.* 10:1227515. doi: 10.3389/fmed.2023.1227515

Privacy-preserving continual learning methods for medical image classification: a comparative analysis

Tanvi Verma[1*†], Liyuan Jin[2,3†], Jun Zhou[1], Jia Huang[1], Mingrui Tan[1], Benjamin Chen Ming Choong[1], Ting Fang Tan[2,4], Fei Gao[1], Xinxing Xu[1], Daniel S. Ting[2,3,4] and Yong Liu[1]

[1]Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore, [2]Artificial Intelligence and Digital Innovation Research Group, Singapore Eye Research Institute, Singapore, Singapore, [3]Duke-NUS Medical School, Singapore, Singapore, [4]Singapore National Eye Centre, Singapore, Singapore

Background: The implementation of deep learning models for medical image classification poses significant challenges, including gradual performance degradation and limited adaptability to new diseases. However, frequent retraining of models is unfeasible and raises concerns about healthcare privacy due to the retention of prior patient data. To address these issues, this study investigated privacy-preserving continual learning methods as an alternative solution.

Methods: We evaluated twelve privacy-preserving non-storage continual learning algorithms based deep learning models for classifying retinal diseases from public optical coherence tomography (OCT) images, in a class-incremental learning scenario. The OCT dataset comprises 108,309 OCT images. Its classes include normal (47.21%), drusen (7.96%), choroidal neovascularization (CNV) (34.35%), and diabetic macular edema (DME) (10.48%). Each class consisted of 250 testing images. For continuous training, the first task involved CNV and normal classes, the second task focused on DME class, and the third task included drusen class. All selected algorithms were further experimented with different training sequence combinations. The final model's average class accuracy was measured. The performance of the joint model obtained through retraining and the original finetune model without continual learning algorithms were compared. Additionally, a publicly available medical dataset for colon cancer detection based on histology slides was selected as a proof of concept, while the CIFAR10 dataset was included as the continual learning benchmark.

Results: Among the continual learning algorithms, Brain-inspired-replay (BIR) outperformed the others in the continual learning-based classification of retinal diseases from OCT images, achieving an accuracy of 62.00% (95% confidence interval: 59.36-64.64%), with consistent top performance observed in different training sequences. For colon cancer histology classification, Efficient Feature Transformations (EFT) attained the highest accuracy of 66.82% (95% confidence interval: 64.23-69.42%). In comparison, the joint model achieved accuracies of 90.76% and 89.28%, respectively. The finetune model demonstrated catastrophic forgetting in both datasets.

Conclusion: Although the joint retraining model exhibited superior performance, continual learning holds promise in mitigating catastrophic forgetting and facilitating continual model updates while preserving privacy in healthcare deep

learning models. Thus, it presents a highly promising solution for the long-term clinical deployment of such models.

KEYWORDS

continual learning, medical image classification, model deployment, optical coherence tomography, comparative analysis

1. Introduction

Continual learning refers to the process of continually training and updating a deep learning model over time, as new data becomes available. This approach is particularly pertinent for medical image classification model deployment because any potentially deployed deep learning model could suffer from a gradual decline in performance from underlying distribution shifts over time. By continuously retraining the model with new data, the model could maintain high classification performance and adapt to changes in the data distribution. However, in continual learning scenarios, the conventional deep learning approach often leads to catastrophic forgetting, where the model experiences memory loss or a significant decline in performance on previous classes after being trained on new tasks or datasets (1). This is a commonly reported phenomenon in deep learning because the model prioritizes its weights and biases optimization for the new task, leading it to forget or overwrite previously learned information. Alternatively, to mitigate catastrophic forgetting in medical image classification, one potential solution is to retrain the model with cumulative data whenever a new dataset becomes available. However, retraining from scratch frequently in the model's deployment phase is not practical. Furthermore, data privacy is of utmost importance in the medical domain, and due to strict regulatory requirements, it may not always be possible to access old data (2). Additionally, medical data is often stored in dedicated servers, making it difficult to shuffle multiple datasets. Researchers have proposed a number of continual learning approaches to overcome catastrophic forgetting. However, there have been few studies on medical imaging using continual learning. Furthermore, achieving a trade-off between stability and plasticity remains another challenge in the continual learning scenario. Stability refers to retaining previously acquired knowledge, while plasticity pertains to the model's ability to learn new knowledge from the new data.

In the context of continual learning, "non-storage" refers to the absence of storing or retaining old data from previous tasks or classes. Continual learning approaches can be categorized into two broad groups: exemplar-based and exemplar-free approaches. Exemplar-based approaches store a small number of data or exemplars and reintroduce them with the new data during training to prevent the model from forgetting the old knowledge. Conversely, exemplar-free approaches (non-storage) rely on regularization, expansion, and generative replay to achieve similar goals without storing exemplar data. Given concerns about privacy and accessibility with medical data, the storage of previous exemplar data from old studies is not feasible. Therefore, exemplar-free continual learning approaches are preferable for medical image classification models and preserve healthcare privacy.

The training process for continual learning happens sequentially in a series of tasks, and during each training session, the model only has access to the data for the current task. *Task incremental*, *domain incremental* and *class incremental* are three common scenarios of continual learning. In task incremental learning, the model has access to the task identifier during inference, which obviates the need for the model to differentiate between images from different tasks. Domain incremental learning, on the other hand, does not require task identification at inference time, since the output space of each task is the same. Class incremental learning, the most clinically relevant yet challenging scenario, involves the model's inability to access the task identifier at inference time, and the model must therefore be capable of distinguishing between images from all tasks. The class incremental learning is more closely aligned with real-life scenarios and more suitable for real-world medical image classification.

Hence, in this research, privacy-preserving exemplar-free continual learning approaches were explored in class learning scenarios for medical imaging classification targeting significant and prevalent diseases using publically available datasets, namely optical coherence tomography (OCT) (3) and PathMNIST (4). As a benchmark for continual learning performance, CIFAR10 dataset (5) was included.

2. Background

Continual learning involves training machine learning models on data from a series of tasks $D = \{D_1, D_2, \cdots, D_T\}$. Task $D_t = \{(\boldsymbol{x}_i^t, y_i^t)\}_{i=1}^{n_t}$ is the t^{th} task where $\boldsymbol{x}_i^t \in X_t$ is an input, $y_i^t \in Y_t$ is the corresponding label and n_t is number of classes in the task. During training the t^{th} task, only training data D_t is available while the training data of previous tasks are no longer accessible. The tasks are generally assumed to be distinct from one another. The goal of continual learning is to train a parameterized model $f_\theta : X \rightarrow Y$ that can predict the correct label for an unseen test sample from any task. Here $X = \cup_{t=1}^{T} X_t$ is the input space and $Y = \cup_{t=1}^{T} Y_t$ is the output space. In continual learning, the marginal probability distribution of inputs varies across tasks, i.e., $P(X_1) \neq P(X_2)$. Based on probability distribution of output space and whether the task identity is provided at the inference time, there are three different continual learning scenarios (6).

2.1. Task incremental scenario

The probability distribution of output space varies between tasks ($P(Y_1) \neq P(Y_2)$) and task identifier is provided at the time of inference for task incremental scenario. Hence, it is possible to train models with task-specific components in this scenario. "Multi-headed" network architecture is commonly used for this scenario where each task has its own output units, but the rest of the network is shared among tasks.

2.2. Domain incremental scenario

The output space (and hence the corresponding probability distribution) remains same across the tasks for domain incremental scenario, i.e., $\{Y_1\} = \{Y_2\}$ and $P(Y_1) = P(Y_2)$. The problem setting looks similar to domain adaptation (7). However unlike domain adaptation, which focuses on achieving good performance on new task, the goal of continual learning is to maintain good performance on previously learned tasks while also achieving reasonable performance on new tasks.

2.3. Class incremental scenario

Similar to task incremental scenario, the probability distribution of output space varies between tasks in class incremental scenario. However, the model does not have access to task identifier at the time of inference which make it the most complex scenario of continual learning. The network architecture for class incremental scenario is generally "single-headed" where a single output layer is used to make predictions for all tasks. Sometimes "multi-headed" architectures are used for this scenario, but it needs prediction of task identifier before predicting the class label of the image at the time of inference. Figure 1 shows an example of a "single-headed" model which is being trained in class incremental scenario to classify OCT images into different retinal pathologies. In many real-world applications, it is not practical to assume that the task identifier will be available at the time of inference, especially in the medical domain where the model is often used to classify diseases. Therefore, an exclusive focus was placed on this scenario in this research.

3. Exemplar-free continual learning approaches

In this section, three main categories of exemplar-free continual learning approaches and associated algorithms were summarized.

3.1. Regularization-based methods

Regularization-based continual learning methods add a regularization term to the training loss function that encourages it to retain the knowledge it has learned from previous tasks while also allowing it to adapt to new tasks. Regularization-based methods can be further divided into weight-based and data-based regularization methods.

The first group of regularization-based approaches aims to prevent weight drift, which is considered to be crucial for previous tasks. This is achieved by estimating the importance of each parameter in the network after learning each task. When training new tasks, the importance of each parameter is taken into account and used to discourage its changes. For example, Elastic Weight Consolidation (EWC) (8) uses a quadratic penalty term to restrict modification of important weights. (9) proposed a quadratic penalty method for continual learning of neural networks that contain batch normalization layers. Synaptic Intelligence (SI) method (10) uses synapse to measure the weights' importance. Memory Aware Synapses (MAS) (11) determines the importance of weights using a Hebbian learning model, which is based on the sensitivity of the output function. Riemanian Walk (RWalk) (12) method uses Fisher Information Matrix approximation and online path integral to calculate the importance for each parameter. (13) defined a notion of uncertainty and made the variance of the incoming weights of each node trainable. To maintain stability-plasticity trade-off, they also included two regularization terms for stability and plasticity respectively. Similarly, (14) added a drifting regularization for stability and a Lasso regularization for plasticity. There are a few methods which regularize gradients of weights. For instance, Orthogonal Weights Modification (OWM) (15) maps the weights modification (gradients) onto a subspace generated by all the previous tasks in order to maintain the performance of previous tasks. Gradient Projection Memory (GPM) (16) regularizes the gradients by restricting the direction of gradient descent steps. Likewise, (17) proposed conceptor-aided backpropagation (CAB), in which gradients are shielded by conceptors (characterizes the linear subspace formed by activation in a layer) against degradation of previously learned tasks.

On the other hand, data-based regularization methods use knowledge distillation (18) to prevent the model from forgetting. Knowledge distillation refers to the technique of transferring the knowledge of a model trained on previous tasks to a new model that will learn new tasks. The basic idea is to use the output of the model trained on previous tasks as a "soft target" for the new model. The new model is trained to mimic the output of the old model on the previous tasks, as well as learn new tasks. Learning without Forgetting (LwF) (19) is one of the earliest methods to use knowledge distillation for continual learning. Extending on LwF, (20) proposed Learning without Memorizing (LwM), which adds attention distillation loss by Gradient-weighted Class Activation Mapping (Grad-CAM) along with knowledge distillation loss to mitigate catastrophic forgetting. There are a few methods such as PODNet (21), Attention Uncertainty (AU) (22) and GeoDL (23) which incorporate exemplar memory in addition to knowledge distillation techniques to address the challenge of continual learning.

A major drawback of regularization-based approaches is that they are highly sensitive to the selection of hyperparameters, making it challenging to achieve a suitable balance between stability and plasticity.

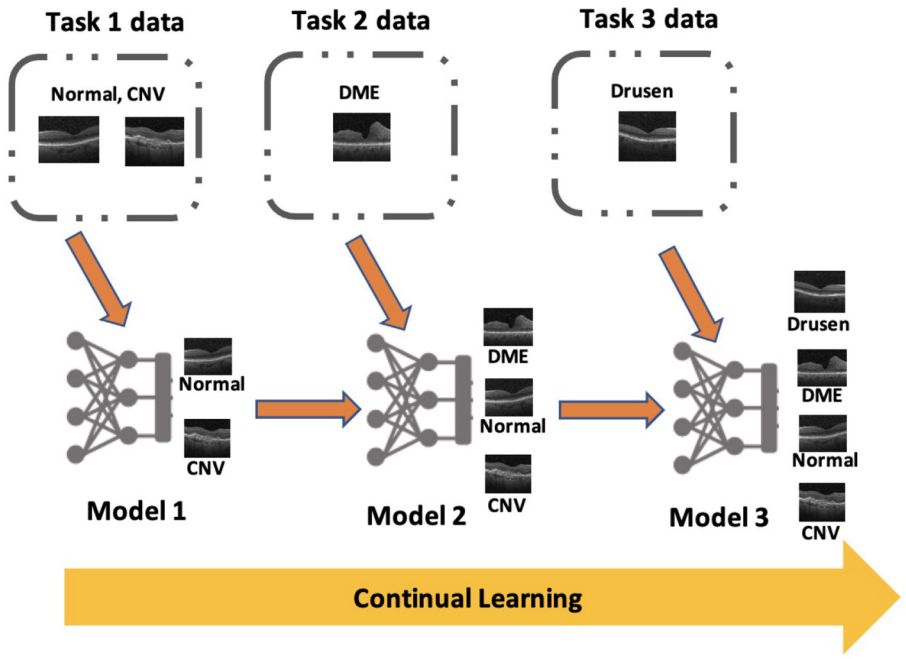

FIGURE 1
An example for class incremental scenario in medical imaging classification: Continual learning based deep learning algorithm sequentially learns a series of retinal pathologies in OCT such as CNV, DME, Drusen, and normal finding.

3.2. Expansion-based methods

In contrast to regularization-based methods, expansion-based continual learning approaches focus on expanding the capacity of the model to handle new tasks. This can be achieved by adding new parameters or neurons to the model, or by creating multiple versions of the model, each specialized for a specific task.

One popular approach of expansion-based continual learning is called dynamic expansion, where the model's capacity is expanded by adding new neurons or parameters to the model as new tasks are encountered. This approach allows the model to adapt to new tasks by creating new representations or features that are specific to the new task. Dynamically Expandable Network (DEN) (24), Reinforced Continual Learning (RCL) (25) and Compacting Picking Growing (CPG) (26) are some methods which fall under this category.

Another approach is called task-specific expansion, where either multiple versions of the model (each specialized for a specific task) are created, or task specific classifier is added to the model. This allows the overall system to handle multiple tasks simultaneously. Progressive Network (27), Additive Parameter Decomposition (APD) (28), Efficient Feature Transformation (EFT) (29) and Expert Gate (30) are some of the methods that utilize task-specific expansion to continuously learn new tasks without forgetting previous tasks.

Most of the existing expansion-based methods are suitable for the task-incremental scenario, i.e., they assume availability of task identity at the time of inference. However, certain methods such as Expert Gate, iTAML and EFT predict the task identifier prior to predicting the correct class. Expansion-based methods typically exhibit superior performance compared to regularization-based methods. However, the requirement for two levels of inferences, first determining the task identity and then the actual class label, may decrease the overall performance of the model.

3.3. Generative replay-based methods

The idea of generative replay-based methods in continual learning is to generate synthetic data that resembles the distribution of old tasks, and use them to train the model along with the data in new tasks. This allows the model to learn new tasks while maintaining knowledge of previously learned tasks, without the need to keep any exemplars from those tasks. Some popular generative methods used in continual learning include generative replay (GR) (31), Replay-through-Feedback (RtF) (32) and Brain Inspired Replay (BIR) (33). GR utilizes a generative adversarial network to generate previous data whereas RtF and BIR use a variational autoencoder as generator.

The main disadvantage of generative method is that it takes a long time to train the generative model. Furthermore, it has been shown that generative models may struggle when dealing with complex datasets (34).

4. Compared methods

In this study, pertinent algorithms from each category of exemplar-free continual learning methods were selected. Within the weight-based regularization methods category, EWC (8), SI

(10), MAS (11), MUC-MAS (35), RWalk (12), OWM (15), and GPM (16) were included. For data-based regularization, LwF (19) and LwM (20) were selected. EFT (29) was chosen as the expansion-based method, while GR (31) and BIR (33) were selected from the generative-replay methods. The lower baseline (finetune) and the upper baseline (joint) approaches were included for comparison. Finetuning involves adapting a pre-trained model to new data or tasks without starting from scratch, while joint training relies on retraining the model on cumulative data from all tasks. A comprehensive description of these selected algorithms is provided in the Supplementary material.

5. Methodology

The model underwent incremental training on three tasks. By definition, Model 1 was initially trained on the first task, followed by expanding Model 1 into Model 2 through sequential training on task 2 data. Similarly, Model 2 was further expanded into Model 3 after being trained on task 3 data. The detailed network architecture and hyperparameter values are provided in the Supplementary material.

5.1. Datasets

Two medical datasets: optical coherence tomography (OCT) (3) and PathMNIST (36) and one non-medical dataset: CIFAR10 (5) were selected. All images are in RGB color, normalized using the mean and standard deviation of ImageNet. All three datasets were split into three tasks and the set of classes in each task was fixed across all experiments. The specific class description for each dataset is specified below.

- **OCT:** This dataset contains over 108,309 publicly available OCT training images, including four classes regarding the condition of the retina: Normal (47.21%), Drusen (7.96%), Choroidal Neovascularization (CNV) (34.35%), Diabetic Macular Edema (DME) (10.48%). There are 250 testing images for each class. The first task contains two classes: Normal and CNV, the second task contains only DME and the last task contains only Drusen. It was selected for its highly imbalanced classes simulating a more realistic continual learning scenario, particularly when the model has already been trained on a large number of data in earlier tasks and the new task only contains a small number of training images. Furthermore, additional investigations on continuous learning sequences were explored for the significance of unbalanced data distribution, as reported in the Supplementary material. Those scenarios include task 1 containing a large amount of data while task 3 containing the least amount of data, and vice versa.
- **PathMNIST:** As part of MedMNIST dataset, PathMNIST dataset is selected due to its distinguishing feature of encompassing a greater number of disease classes. It consists of histology slides with 9 different colon pathology classes. It contains 89,996 training and 7,180 testing images. The number of training images in a class varies from 7,886 to 12,885, with an average of 10,000. The 9 classes are divided into three tasks containing three classes each.
- **CIFAR10:** As a traditional continual learning benchmark, CIFAR10 dataset was selected to compare the adaptability and robustness of different algorithms. It consists of 60,000 natural images in 10 classes, such as cat and truck. There are 5,000 training images and 1,000 testing images per class. The first task contains four classes while the two subsequent tasks contain three classes each.

5.2. Statistical analysis

Experiments were conducted using three different seed values. Following common practices in the field of continual learning (12, 37), average accuracy and average forgetting were selected as the evaluation metrics. Average accuracy measures the overall performance of the model after training on task t is complete. It is computed as $A_t = \frac{1}{t}\sum_{i=1}^{t} a_t^i$, where a_t^i is the accuracy of the model on task i after training on task t. On the other hand, task-wise accuracy was also introduced to highlight model adaptation and the balance between stability and plasticity during sequential training, measured by the intermediate and final accuracies of each task on the intermediate (Model 1 and Model 2) and final model (Model 3). The forgetting metric was included by measuring the decline in accuracy for each task by comparing the highest accuracy achieved during training with the final accuracy after training is completed. This provides an estimate of the extent model has been forgotten based on its current state. The forgetting on task i after the model has been trained on task t is given by $f_t^i = \max_{j\in\{1,\cdots,t-1\}} (a_j^i - a_t^i)$. The average forgetting of the final model is computed as $F = \frac{1}{T-1}\sum_{i=1}^{T-1} f_T^i$.

6. Results

The average accuracy and forgetting of the final model (Model 3) after it has been trained on the three tasks sequentially are calculated. The task-wise accuracy on each model is also measured for a better insight into the balance between stability and plasticity.

6.1. Average accuracy and forgetting

The average accuracy and forgetting of Model 3 are presented in Table 1, and the effect of sequential training on the average accuracy of the model is shown in Figure 2. For average accuracy, joint training obtains best performance of 90.76%, 89.28% and 88.01% respectively and it serves as the upper bound. Among all validated algorithms, BIR shows the best retinal disease classification on OCT and CIFAR10 performance with average accuracy of 62.00% (95% CI 59.36-64.64%) and 64.68% (95% confidence interval (CI) 63.76-65.59%) respectively. EFT obtains the best colorectal cancer histology classification on PathMNIST with average accuracy of 66.82% (95% CI 64.23-69.42%). Additionally, EFT demonstrates consistent better accuracy and lower forgetting across different datasets with average accuracy of 43.20% (95% CI 41.24-45.16%)

TABLE 1 Average accuracy and forgetting of Model 3 for the three datasets.

Category	Method	OCT		PathMNIST		CIFAR10	
		Accuracy	Forgetting	Accuracy	Forgetting	Accuracy	Forgetting
Baseline	Finetune	33.33	100	28.89	99.18	32.20	86.47
	Joint	90.76	-	89.28	-	88.01	-
Regularization	LwF	44.8	80.23	25.20	81.33	32.90	65.62
	LwM	41.75	41.58	23.02	35.43	44.66	41.48
	EWC	31.54	76.83	29.16	95.82	32.53	93.32
	SI	43.60	**21.27**	32.40	28.94	28.51	40.62
	MAS	31.73	83.69	29.97	74.54	36.18	80.07
	MUC-MAS	39.32	76.87	33.49	52.19	33.84	65.10
	GPM	41.16	26.41	39.51	24.96	36.47	36.18
	OWM	38.93	83.10	52.42	**16.34**	48.30	42.18
	RWalk	33.33	100	27.05	85.13	35.00	90.99
Expansion	EFT	43.20	38.13	**66.82**	29.39	60.65	**31.31**
Generative	GR	35.83	66.05	21.95	90.88	31.50	74.28
	BIR	**62.00**	51.31	35.17	88.17	**64.68**	42.30

The bold values indicate best results for each column.

on OCT and 60.65% (95% CI 58.57-62.73%) on CIFAR10. RWalk's performance is poor and is comparable to the performance of Finetune with catastrophic forgetting. Other algorithms such as LwF, EWC, MAS, MUC-MAS and GR appear to have comparable performance to Finetune in terms of average accuracy. However, they are still able to retain some knowledge of previous tasks (evident from their lower forgetting), in contrast to Finetune, which almost completely forgets previous tasks, as can be observed in Table 2. GR performs poorly on all datasets in contrast to similar generative replay method BIR.

In terms of forgetting, although the accuracy of SI is not high, it demonstrates the least forgetting (21.27% on OCT). For PathMNIST and CIFAR10 datasets, OWM and EFT show the least forgetting, measured by 16.34% and 31.31% respectively. We also note that the performance of exemplar-free continual learning methods is greatly influenced by the structure of the dataset. For instance, OWM, which performs relatively well on PathMNIST (52.42% accuracy, 16.34% forgetting) and CIFAR10 (48.30% accuracy, 42.18% forgetting), does poorly on the unbalanced data of OCT (38.93% accuracy, 83.10% forgetting). In contrast, SI performs better on OCT (43.6% accuracy, 21.27% forgetting) and performs relatively worse on PathMNIST (32.40% accuracy, 28.94% forgetting) and CIFAR10 (28.51% accuracy, 40.62% forgetting).

In summary, EFT and BIR exhibit better overall accuracy compared to the other selected exemplar-free continual learning methods.

6.2. Task-wise accuracy

Based on Figures 3–5, the accuracy of each task is affected after the model is trained on each task for OCT, PathMNIST and CIFAR10 datasets respectively. Plots in Figures 3A, 4A, 5A show the decline in accuracy of task 1 data after sequential training on the three tasks. The X-axis represents the three models. It is observed that although the initial accuracy on task 1 is almost similar for all methods, it drastically declines (except for a few well-performing methods such as EFT, BIR, etc.) after training on task 2. However, the accuracies of task 2 data vary widely (Figures 3B, 4B, 5B) depending on the stability-plasticity trade-off used by each method. The initial accuracy of task 2 for methods such as SI, LwM, and GPM is relatively low as, to maintain stability, the model is not flexible enough to incorporate new knowledge. The same behavior is observed for the task 3 data (Figures 3C, 4C, 5C), and the initial accuracy of task 3 data varies with the algorithm.

Results reported in Table 2 provides a better picture of the stability-plasticity trade-off employed by each algorithm where the accuracy of each task on Model 3 was reported. Algorithms such as LwF, EWC, MAS, RWalk and GR almost completely forget previous tasks and their average accuracy reported in Table 1 is mainly due to the high accuracy of task 3. Based on the relatively better performance of EFT across three different datasets, it supports the assertion that expansion-based methods, which add task-specific nodes to the model, are better able to retain the knowledge of previous tasks compared to regularization-based methods. As EFT maintains a task-specific classifier, it is able to maintain relatively better accuracies on each task.

It is noteworthy that for some algorithms such as LwM, SI, and GPM, the final accuracy of task 1 is higher than the final accuracy of task 2, which appears counterintuitive. However, as seen in Figures 3B, 4B, 5B, to maintain stability, the initial accuracies of task 2 are low to begin with. Therefore, when the model is subsequently trained on task 3, the accuracy of task 2 further declines and is lower than the final accuracy of task 1 (which was high to begin with). This explains the behavior of having task 2's final accuracy lower than task 1's final accuracy.

The majority of methods, except for EFT and OWM to some extent, exhibit poor performance on the PathMNIST dataset. This

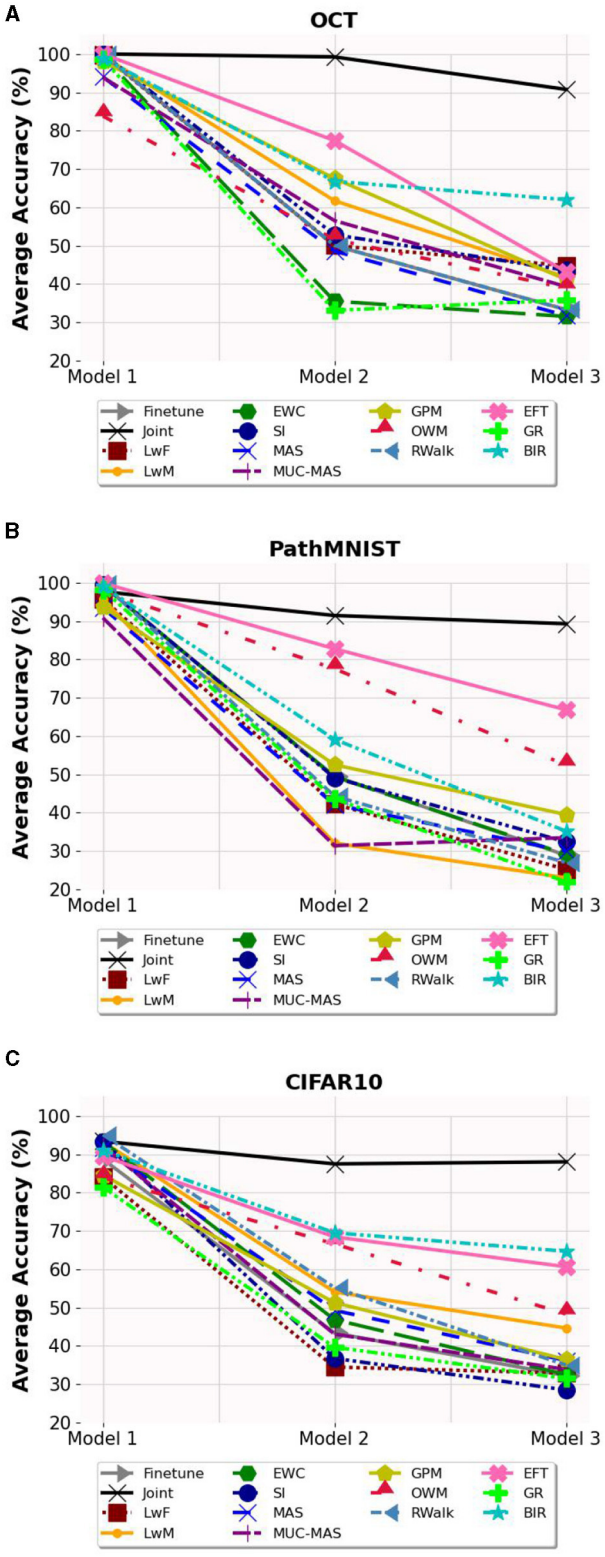

FIGURE 2
The average accuracy on the three models. **(A)** Average accuracy on OCT dataset. **(B)** Average accuracy on PathMNIST dataset. **(C)** Average accuracy on CIFAR10 dataset.

could be attributed to the presence of overlapping features among classes across different tasks. The limited success of OWM on this dataset highlights the existence of feature overlap, as even with an extensive training regime of 500 epochs and a learning rate of 1e7, OWM struggled to identify a suitable subspace in the feature space for the third task, despite its attempt to project the gradients of the new task orthogonally to the subspace of previous tasks.

Another notable observation is that while methods like LwF and BIR excel at retaining knowledge of the most recent task, they tend to exhibit a higher degree of forgetting as the number of tasks increases. For instance, although BIR demonstrates the highest average accuracy in both OCT and CIFAR10, it still experiences significant forgetting (51.31% on OCT and 42.30% on CIFAR10). This indicates that although these models achieve high accuracy on current tasks, they struggle to retain information about earlier tasks. Additionally, regularization-based methods are highly sensitive to hyperparameters, and even small changes to the regularization loss coefficient can lead to vastly different results.

In the context where the later tasks have more training data compared to the earlier tasks, BIR consistently maintains its performance on the OCT dataset, even when the class sequence is altered. This is demonstrated in the Supplementary material. However, while EFT maintains its overall accuracy, its task-wise accuracy exhibits variations for different sequences. GR's performance remains unaffected by data imbalance; however, its overall accuracy remains low, and forgetting remains high. Other regularization-based methods, including LwF, EWC, MAS and MUC-MAS demonstrate similar patterns. For the remaining regularization-based methods, such as LWM, SI, GPM, OWM and RWalk changes in the task sequence impact their task-wise accuracy. Despite this, their overall accuracy remains low.

In conclusion, striking a balance between stability and plasticity poses a significant challenge, especially for regularization-based algorithms. However, generative-based and expansion-based methods show promise in enhancing classification accuracy for tasks in the context of continual learning. Furthermore, generative-based methods are well-suited for scenarios where data is imbalanced across tasks. These approaches hold the potential for addressing the complexities of retaining previously learned knowledge while accommodating new information.

7. Discussion

In the field of medical imaging, deep learning models have been widely used for classification tasks across various medical imaging modalities. These models have been incorporated into real-world practice for decision-making in diagnosis and treatment, as evidenced by recent examples such as referable diabetic retinopathy screening in ophthalmology using color fundus photography (38). While deep learning models have provided a balance between healthcare burden and disease management, emerging imaging devices and new disease pathologies require further improvement of existing models. Continual learning has the potential to mitigate catastrophic forgetting and enable continual model updates,

TABLE 2 Task wise accuracy of Model 3 for the three datasets.

Category	Method	OCT			PathMNIST			CIFAR10		
		Task 1	Task 2	Task 3	Task 1	Task 2	Task 3	Task 1	Task 2	Task 3
Baseline	Finetune	0.0	0.0	100	0.0	0.0	86.67	0.0	2.53	94.09
	Joint	99.47	97.47	75.33	95.99	85.93	85.92	86.03	86.47	91.55
Regularization	LwF	0.53	39.33	94.53	0.63	15.09	59.83	0.14	21.76	76.79
	LwM	63.13	0.93	61.20	28.97	26.69	13.38	36.65	26.81	73.51
	EWC	0.01	0.13	94.47	1.01	5.20	81.27	0.01	0.01	97.57
	SI	63.33	14.53	52.93	49.36	24.84	22.99	29.85	0.42	55.26
	MAS	0.13	10.37	84.68	2.04	6.07	81.78	3.35	18.05	87.16
	MUC-MAS	9.28	20.99	87.69	18.81	4.14	77.51	5.58	18.69	77.26
	GPM	90.66	1.73	31.07	79.77	2.23	36.53	58.1	4.0	47.30
	OWM	14.40	13.60	88.80	77.72	73.18	6.35	32.35	49.61	62.94
	RWalk	0.0	0.0	100	0.22	0.06	80.88	0.81	6.33	97.86
Expansion	EFT	61.47	38.40	29.73	59.60	73.63	67.24	41.10	60.74	80.10
Generative	GR	2.57	16.81	88.11	0.26	1.82	63.77	2.74	7.53	84.22
	BIR	10.53	78.40	97.07	17.33	3.38	84.82	30.18	68.13	95.71

making it a promising solution for medical image classification models. In our study, we focused on comparing exemplar-free continual learning methods for medical image classification. We observed that regularization-based methods generally struggled in addressing catastrophic forgetting, while expansion-based methods and generative replay-based methods showed potential in retaining knowledge of earlier tasks. Although the AI field frequently reports on objective metrics like forgetting, it lacks clinical reliability and applicability. This is primarily due to its limited inter-model variability and weak correlation with model performance as demonstrated. Additionally, as the concept of continual learning is still in its early stages within the medical field, determining the average accuracy threshold for clinical deployment remains an unexplored area that necessitates careful consideration of balancing healthcare privacy concerns.

Our results show a considerable gap between best-performing continual learning algorithms with traditional joint training model which involves storing previous data and retraining the model from scratch. However, this conventional joint training model would not exist in the real world due to its storage for retraining strongly violating healthcare patient privacy. Despite challenges for current continual learning-based methods to achieve optimal classification performance, continual learning based deep learning model would likely be the next paradigm for medical image classification models with the aforementioned advantages. Another interesting topic is the balance between loss in the deep learning model's performance and breakage in individual patient privacy. As the next paradigm for deep learning in medical imaging, continual learning could potentially offer cross-institutional training for expanding model generalizability, learning about new diseases, and even enhancing limited data on rare diseases. Federated Learning (FL) (39) offers an alternative to deep learning when dealing with imbalanced healthcare data and healthcare privacy issues, relying on distributed localized model training and subsequent updating centralized models without exchanging raw input data. This approach has shown practical applicability in real-world scenarios, particularly in multi-center collaborations focused on COVID-19 detection during the recent pandemic (40) during the last pandemic. While continual learning may appear similar to FL, its distinct strength lies in the fact that knowledge is "memorized" within the model parameters, eliminating the need for complex adjunct security measures like differential privacy or blockchain integration in FL. Moreover, in FL, all the training data is assumed to be available simultaneously, albeit on different local clients. On the other hand, continual learning deals with the temporal factor, where training data arrives with time, and older training data may become unavailable due to privacy concerns.

As for continual learning techniques, although many original continual learning approaches achieve their best performance in task incremental scenarios, class incremental scenarios are more realistic in healthcare settings and continual learning deployment in medical image classification due to intrinsic input data complexity and uncertainty to task information. However, not all existing continual learning algorithms perform well in the class incremental learning scenario, due to their unique design and inconsistent performance in different seeds. Additionally, certain algorithms demonstrated robustness in handling data imbalance across tasks. The strength of our research lies in being the first comparative study that extensively analyzed all existing privacy-preserving continual learning algorithms on two medical imaging modalities. Such significant datasets were selected based on their application in diseases with high prevalence, morbidity and mortality. However, we acknowledge the limitation of not including many other forms of medical imaging. In future work, it is pertinent

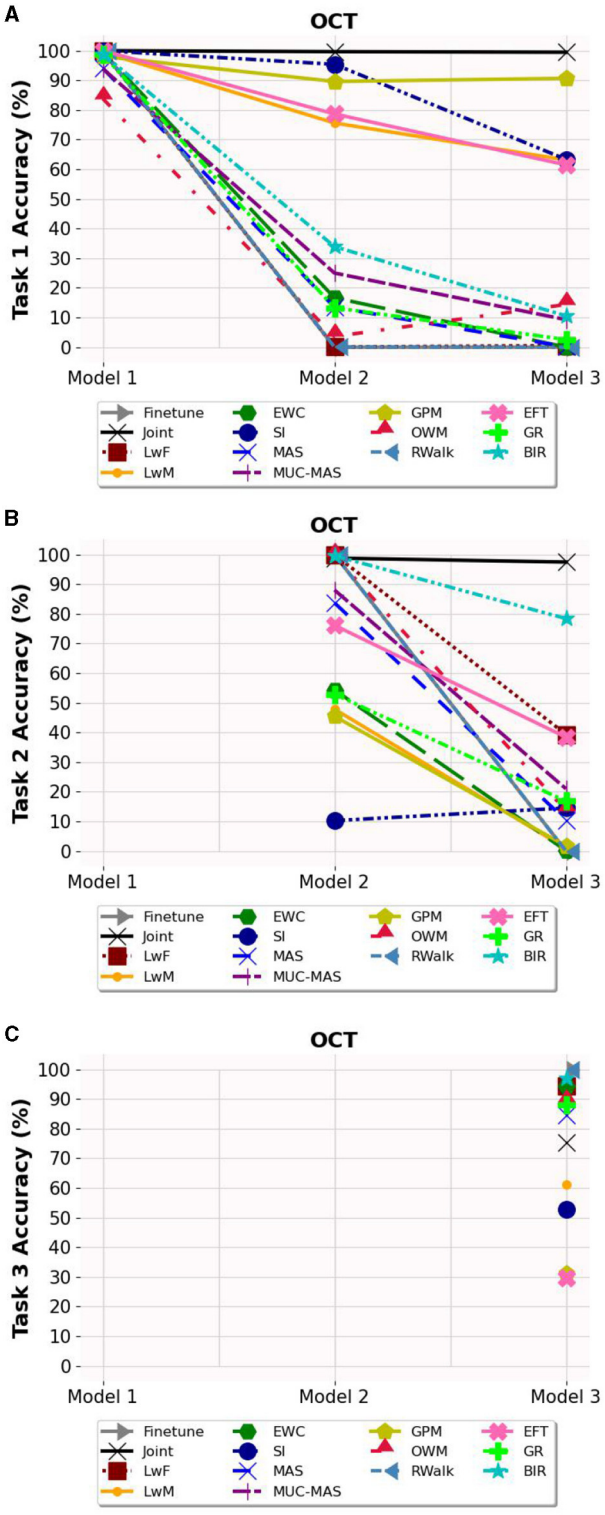

FIGURE 3
Task wise accuracy of OCT dataset on the three models. **(A)** Task 1 accuracy. **(B)** Task 2 accuracy. **(C)** Task 3 accuracy.

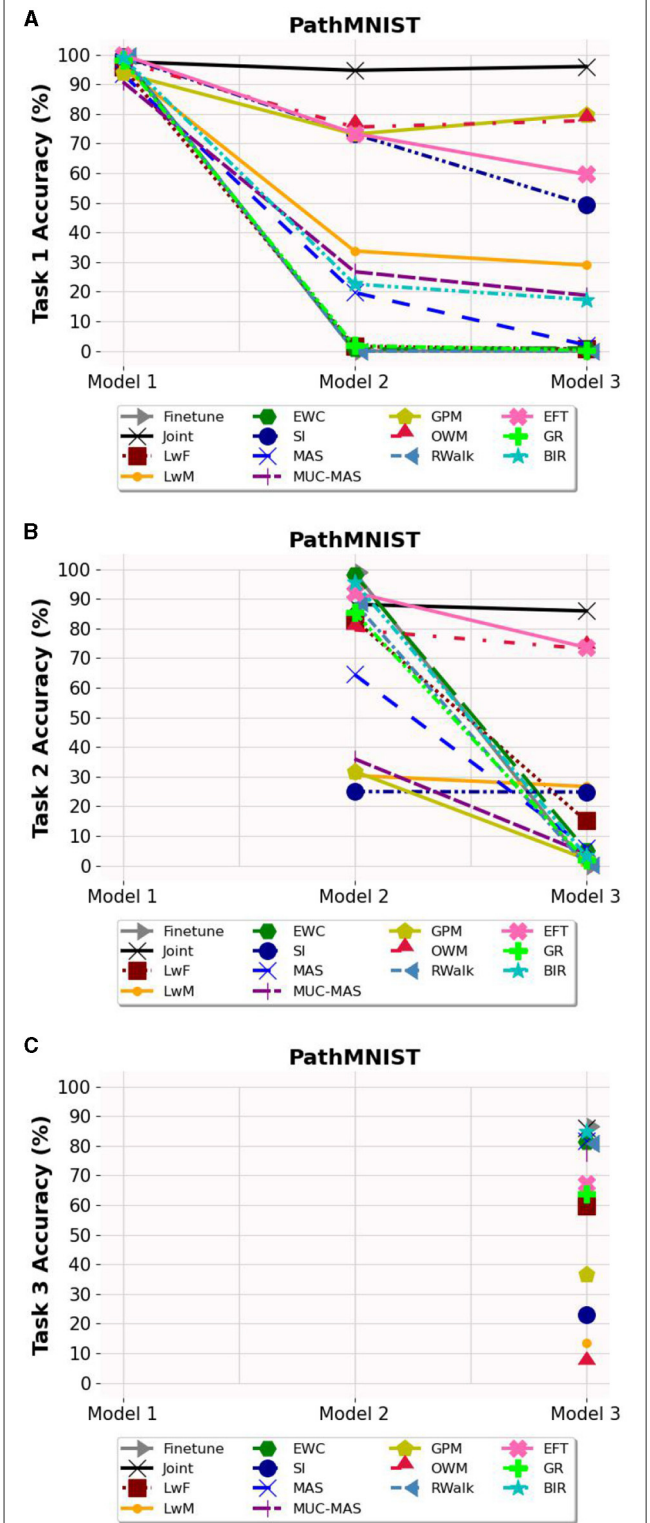

FIGURE 4
Task wise accuracy of PathMNIST dataset on the three models. **(A)** Task 1 accuracy. **(B)** Task 2 accuracy. **(C)** Task 3 accuracy.

to explore the influence of different input data types, for example, chest X-ray, color fundus photography, computed tomography, and magnetic resonance imaging.

8. Related work

With a growing interest in have been many approaches proposed in the literature. This has been followed by several surveys

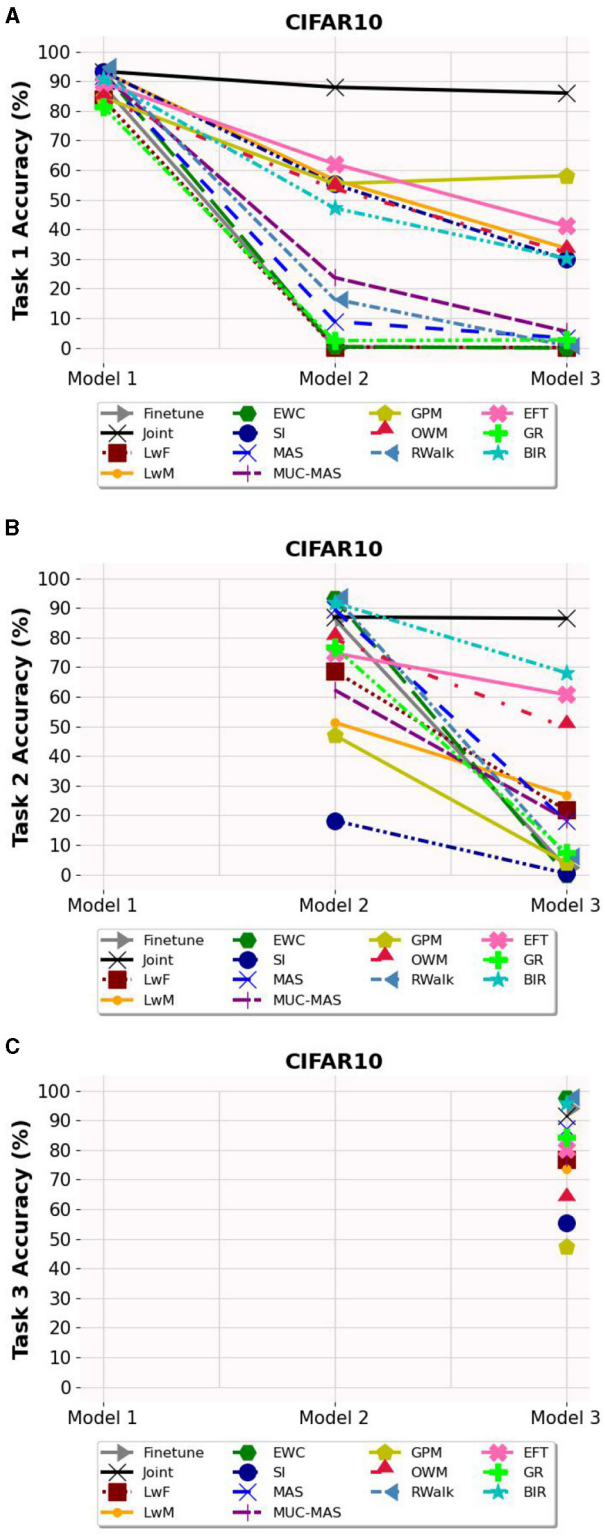

FIGURE 5
Task wise accuracy of CIFAR10 dataset on the three models. **(A)** Task 1 accuracy. **(B)** Task 2 accuracy. **(C)** Task 3 accuracy.

and empirical papers that aim to provide an overview of the field and evaluate the performance of these methods. To enable a structured comparison between continual learning methods, Van de Ven and Tolias (6) described the three scenarios of learning. Parisi et al. (41) discussed continual learning from the perspective of biological lifelong learning such as structural plasticity, memory replay, curriculum and transfer learning etc. Delange et al. (42) offered a comprehensive experimental comparison of 11 different continual learning methods, however, they focused on easier task incremental setting and assume that the task identifier is known at the time of inference. In their interesting work Farquhar and Gal (43) examined standard evaluation practices and observed that based on the selection of experimental design, some continual learning approaches look better than they are. Qu et al. (44) grouped continual learning methods by their representative techniques, including regularization, knowledge distillation, memory, generative replay, parameter isolation etc. The study conducted by Hayes et al. (45) presents a thorough comparison between replay in the mammalian brain and replay in artificial neural networks and identified the gaps between replay in these two fields. Most of the empirical surveys (46–48) cover all three scenarios of continual learning and select only a handful of approaches suitable for each scenario for comparing performances. Hence, these works do not deep dive into a more focused use case of exemplar-free class-incremental setting which is particularly relevant to the medical domain. Furthermore, they do not compare these state-of-the-art algorithms on medical datasets. Research conducted by Derakhshani et al. (49) is closest to our work, where the authors have established a benchmark for classifying diseases using the MedMNIST dataset. However, they have considered a limited selection of five continual learning methods across all three scenarios of continual learning. As exemplar-free methods, they chose EWC (8), MAS (11), and LwF (19), while iCarL (50) and EEIL (51) were selected as exemplar-based methods. They reported iCarL achieving the highest performance on the PathMNIST dataset with an accuracy of 58.46%. In contrast, EFT (29), an expansion-based exemplar-free method, performed the best (66.82%) on the PathMNIST dataset based on the current study.

9. Conclusion

Three major continual learning methods namely regularization-based, expansion-based, and generative replay-based methods, and relevant algorithms were explained and summarized in this research. Furthermore, twelve state-of-the-art privacy-preserving continual learning algorithms were investigated for medical imaging classification using deep learning models. BIR algorithm achieved the best average accuracy on OCT for retinal disease classification among all continual learning algorithms, and EFT is the best-performing algorithm on PathMNIST for colorectal cancer histology classification. It was suggested both expansion-based and generative replay-based methods, specifically EFT and BIR, show the greatest potential in continual learning for medical applications. Given the frequent model updates and the need for the integration of new medical knowledge, continual learning has become an increasingly important topic in healthcare model deployment. Nevertheless, the trade-off between performance loss and patient privacy remains a crucial consideration. Continual learning offers a promising avenue for

improving model performance while preserving patient privacy, and could potentially be the next paradigm for next-generation deep learning-based medical image classification.

Data availability statement

Publicly available datasets were analyzed in this study. OCT data can be found at: https://data.mendeley.com/datasets/rscbjbr9sj/3, PathMNIST data can be found at: https://zenodo.org/record/6496656#.ZGMJ--xByys and CIFAR10 data can be found at: https://www.cs.toronto.edu/~kriz/cifar.html.

Author contributions

TV and LJ: conceptualization. TV, LJ, JZ, JH, MT, and BC: methodology, implementation, and experiments. TV, LJ, JZ, JH, MT, and TT: draft preparation and visualization. TV, LJ, FG, and XX: review and editing. DT and YL: supervision. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the Agency for Science, Technology and Research (A*STAR) AME Programmatic Funds (Grant Number: A20H4b0141), National Medical Research Council, Singapore (Grant Number: NMRC/HSRG/0087/2018, MOH-000655-00, and MOH-001014-00), and Duke-NUS Medical School (Grant Number: Duke-NUS/RSF/2021/0018 and 05/FY2020/EX/15-A58).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2023.1227515/full#supplementary-material

References

1. Goodfellow IJ, Mirza M, Xiao D, Courville A, Bengio Y. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:13126211* (2013). doi: 10.48550/arXiv.1312.6211

2. American Health Information Management Association. Retention and destruction of health information. *J AHIMA*. (2013). Available online at: https://library.ahima.org/PB/RetentionDestruction

3. Kermany D, Zhang K, Goldbaum M. Large dataset of labeled optical coherence tomography (OCT) and chest x-ray images. *Mendeley Data*. (2018) 3:10–17632.

4. Yang J, Shi R, Ni B. "Medmnist classification decathlon: a lightweight autoML benchmark for medical image analysis," in: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE (2021). p. 191–5.

5. Krizhevsky A, Hinton G, et al. Learning multiple layers of features from tiny images (2009). Available online at: https://www.cs.toronto.edu/~kriz/cifar.html

6. Van de Ven GM, Tolias AS. Three scenarios for continual learning. *arXiv preprint arXiv:190407734* (2019). doi: 10.48550/arXiv.1904.07734

7. Pan SJ, Tsang IW, Kwok JT, Yang Q. Domain adaptation via transfer component analysis. *IEEE Trans Neural Netw*. (2010) 22:199–210. doi: 10.1109/TNN.2010.2091281

8. Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu AA, et al. Overcoming catastrophic forgetting in neural networks. *Proc Natl Acad Sci USA*. (2017) 114:3521–6. doi: 10.1073/pnas.1611835114

9. Lee J, Hong HG, Joo D, Kim J. Continual learning with extended kronecker-factored approximate curvature. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2020). p. 9001–10.

10. Zenke F, Poole B, Ganguli S. Continual learning through synaptic intelligence. In: *International Conference on Machine Learning*. PMLR (2017). p. 3987–95.

11. Aljundi R, Babiloni F, Elhoseiny M, Rohrbach M, Tuytelaars T. Memory aware synapses: learning what (not) to forget. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. (2018).

12. Chaudhry A, Dokania PK, Ajanthan T, Torr PH. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. (2018). p. 532–47.

13. Ahn H, Cha S, Lee D, Moon T. *Uncertainty-based continual learning with adaptive regularization*. Advances in Neural Information Processing Systems 32 (2019).

14. Jung S, Ahn H, Cha S, Moon T. Continual learning with node-importance based adaptive group sparse regularization. In: *Advances in Neural Information Processing Systems 33*. Curran Associates (2020). p. 3647–58.

15. Zeng G, Chen Y, Cui B, Yu S. Continual learning of context-dependent processing in neural networks. *Nat Mach Intell*. (2019) 1:364–72. doi: 10.1038/s42256-019-0080-x

16. Saha G, Garg I, Roy K. Gradient projection memory for continual learning. *arXiv preprint arXiv:210309762* (2021). doi: 10.48550/arXiv.2103.09762

17. He X, Jaeger H. Overcoming catastrophic interference using conceptor-aided backpropagation. In: *International Conference on Learning Representations*. (2018).

18. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:150302531* (2015). doi: 10.48550/arXiv.1503.02531

19. Li Z, Hoiem D. Learning without forgetting. *IEEE Trans Pattern Anal Mach Intell*. (2017) 40:2935–47. doi: 10.1109/TPAMI.2017.2773081

20. Dhar P, Singh RV, Peng KC, Wu Z, Chellappa R. Learning without memorizing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2019). p. 5138–46.

21. Douillard A, Cord M, Ollion C, Robert T, Valle E. PODNet: pooled outputs distillation for small-tasks incremental learning. In: *European Conference on Computer Vision*. Springer (2020). p. 86–102.

22. Kurmi VK, Patro BN, Subramanian VK, Namboodiri VP. Do not forget to attend to uncertainty while mitigating catastrophic forgetting. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. (2021). p. 736–45.

23. Simon C, Koniusz P, Harandi M. On learning the geodesic path for incremental learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2021). p. 1591–600.

24. Yoon J, Yang E, Lee J, Hwang SJ. Lifelong learning with dynamically expandable networks. *arXiv preprint arXiv:170801547* (2017). doi: 10.48550/arXiv.1708.01547

25. Xu J, Zhu Z. *Reinforced continual learning*. Advances in Neural Information Processing Systems 31 (2018).

26. Hung CY, Tu CH, Wu CE, Chen CH, Chan YM, Chen CS. *Compacting, picking and growing for unforgetting continual learning*. Advances in Neural Information Processing Systems 32 (2019).

27. Rusu AA, Rabinowitz NC, Desjardins G, Soyer H, Kirkpatrick J, Kavukcuoglu K, et al. Progressive neural networks. *arXiv preprint arXiv:160604671* (2016). doi: 10.48550/arXiv.1606.04671

28. Yoon J, Kim S, Yang E, Hwang SJ. Scalable and order-robust continual learning with additive parameter decomposition. *arXiv preprint arXiv:190209432* (2019). doi: 10.48550/arXiv.1902.09432

29. Verma VK, Liang KJ, Mehta N, Rai P, Carin L. Efficient feature transformations for discriminative and generative continual learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2021). p. 13865–75.

30. Aljundi R, Chakravarty P, Tuytelaars T. Expert gate: lifelong learning with a network of experts. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2017). p. 3366–75.

31. Shin H, Lee JK, Kim J, Kim J. *Continual learning with deep generative replay*. Advances in Neural Information Processing Systems 30 (2017).

32. van de Ven GM, Tolias AS. Generative replay with feedback connections as a general strategy for continual learning. *arXiv preprint arXiv:180910635* (2018). doi: 10.48550/arXiv.1809.10635

33. van de Ven GM, Siegelmann HT, Tolias AS. Brain-inspired replay for continual learning with artificial neural networks. *Nat Commun.* (2020) 11:1–14. doi: 10.1038/s41467-020-17866-2

34. Lesort T, Caselles-Dupré H, Garcia-Ortiz M, Stoian A, Filliat D. Generative models from the perspective of continual learning. In: *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE (2019). p. 1–8.

35. Liu Y, Parisot S, Slabaugh G, Jia X, Leonardis A, Tuytelaars T. More classifiers, less forgetting: a generic multi-classifier paradigm for incremental learning. In: Vedaldi A, Bischof H, Brox T, Frahm JM, editors. *Computer Vision – ECCV 2020*. Cham: Springer International Publishing (2020). p. 699–716.

36. Kather JN, Krisam J, Charoentong P, Luedde T, Herpel E, Weis CA, et al. Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. *PLoS Med.* (2019) 16:e1002730. doi: 10.1371/journal.pmed.1002730

37. Mirzadeh SI, Farajtabar M, Pascanu R, Ghasemzadeh H. Understanding the role of training regimes in continual learning. In: *Advances in Neural Information Processing Systems 33*. Curran Associates (2020). p. 7308–7320.

38. Ting DSW, Cheung CYL, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*. (2017) 318:2211–23. doi: 10.1001/jama.2017.18152

39. McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA. Communication-efficient learning of deep networks from decentralized data. In: *Artificial Intelligence and Statistics*. PMLR (2017). p. 1273–82.

40. Dou Q, So TY, Jiang M, Liu Q, Vardhanabhuti V, Kaissis G, et al. Federated deep learning for detecting COVID-19 lung abnormalities in CT: a privacy-preserving multinational validation study. *NPJ Digit Med.* (2021) 4:60. doi: 10.1038/s41746-021-00431-6

41. Parisi GI, Kemker R, Part JL, Kanan C, Wermter S. Continual lifelong learning with neural networks: a review. *Neural Netw.* (2019) 113:54–71. doi: 10.1016/j.neunet.2019.01.012

42. Delange M, Aljundi R, Masana M, Parisot S, Jia X, Leonardis A, et al. A continual learning survey: defying forgetting in classification tasks. *IEEE Trans Pattern Anal Mach Intell.* (2021) 44:3366–3385. doi: 10.1109/TPAMI.2021.3057446

43. Farquhar S, Gal Y. Towards robust evaluations of continual learning. *arXiv preprint arXiv:180509733* (2018). doi: 10.48550/arXiv.1805.09733

44. Qu H, Rahmani H, Xu L, Williams B, Liu J. Recent advances of continual learning in computer vision: an overview. *arXiv preprint arXiv:210911369* (2021). doi: 10.48550/arXiv.2109.11369

45. Hayes TL, Krishnan GP, Bazhenov M, Siegelmann HT, Sejnowski TJ, Kanan C. Replay in deep learning: current approaches and missing biological elements. *Neural Comput.* (2021) 33:2908–50. doi: 10.1162/neco_a_01433

46. Masana M, Liu X, Twardowski B, Menta M, Bagdanov AD, van de Weijer J. Class-incremental learning: survey and performance evaluation on image classification. *IEEE Trans Pattern Anal Mach Intell.* (2022) 45:5513–33. doi: 10.1109/TPAMI.2022.3213473

47. Belouadah E, Popescu A, Kanellos I. A comprehensive study of class incremental learning algorithms for visual tasks. *Neural Netw.* (2021) 135:38–54. doi: 10.1016/j.neunet.2020.12.003

48. Mai Z, Li R, Jeong J, Quispe D, Kim H, Sanner S. Online continual learning in image classification: an empirical survey. *Neurocomputing.* (2022) 469:28–51. doi: 10.1016/j.neucom.2021.10.021

49. Derakhshani MM, Najdenkoska I, van Sonsbeek T, Zhen X, Mahapatra D, Worring M, et al. LifeLonger: a benchmark for continual disease classification. In: *International Conference on Medical Image Computing and Computer Assisted Intervention*. (2022).

50. Rebuffi SA, Kolesnikov A, Sperl G, Lampert CH. ICARL: incremental classifier and representation learning. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. (2017). p. 2001–10.

51. Castro FM, Marín-Jiménez MJ, Guil N, Schmid C, Alahari K. End-to-end incremental learning. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. (2018). p. 233–48.

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 08 September 2023
DOI 10.3389/fmed.2023.1150295

OPEN ACCESS

EDITED BY
Darren Shu Jeng Ting,
University of Nottingham, United Kingdom

REVIEWED BY
Zizhong Hu,
Nanjing Medical University, China
Rizwan Ali Naqvi,
Sejong University, Republic of Korea

*CORRESPONDENCE
Biao Yan
✉ yanbiao1982@hotmail.com
Zhenhua Wang
✉ zh-wang@shou.edu.cn
Qin Jiang
✉ jiangqin710@126.com

†These authors have contributed equally to this work

RECEIVED 24 January 2023
ACCEPTED 25 August 2023
PUBLISHED 08 September 2023

CITATION
Bai Y, Li J, Shi L, Jiang Q, Yan B and Wang Z (2023) DME-DeepLabV3+: a lightweight model for diabetic macular edema extraction based on DeepLabV3+ architecture. *Front. Med.* 10:1150295. doi: 10.3389/fmed.2023.1150295

DME-DeepLabV3+: a lightweight model for diabetic macular edema extraction based on DeepLabV3+ architecture

Yun Bai[1†], Jing Li[1†], Lianjun Shi[2†], Qin Jiang[2*], Biao Yan[3*] and Zhenhua Wang[1*]

[1]College of Information Science, Shanghai Ocean University, Shanghai, China, [2]The Affiliated Eye Hospital, Nanjing Medical University, Nanjing, China, [3]Eye Institute, Eye and ENT Hospital, Shanghai Medical College, Fudan University, Shanghai, China

Introduction: Diabetic macular edema (DME) is a major cause of vision impairment in the patients with diabetes. Optical Coherence Tomography (OCT) is an important ophthalmic imaging method, which can enable early detection of DME. However, it is difficult to achieve high-efficiency and high-precision extraction of DME in OCT images because the sources of OCT images are diverse and the quality of OCT images is not stable. Thus, it is still required to design a model to improve the accuracy of DME extraction in OCT images.

Methods: A lightweight model (DME-DeepLabV3+) was proposed for DME extraction using a DeepLabV3+ architecture. In this model, MobileNetV2 model was used as the backbone for extracting low-level features of DME. The improved ASPP with sawtooth wave-like dilation rate was used for extracting high-level features of DME. Then, the decoder was used to fuse and refine low-level and high-level features of DME. Finally, 1711 OCT images were collected from the Kermany dataset and the Affiliated Eye Hospital. 1369, 171, and 171 OCT images were randomly selected for training, validation, and testing, respectively.

Conclusion: In ablation experiment, the proposed DME-DeepLabV3+ model was compared against DeepLabV3+ model with different setting to evaluate the effects of MobileNetV2 and improved ASPP on DME extraction. DME-DeepLabV3+ had better extraction performance, especially in small-scale macular edema regions. The extraction results of DME-DeepLabV3+ were close to ground truth. In comparative experiment, the proposed DME-DeepLabV3+ model was compared against other models, including FCN, UNet, PSPNet, ICNet, and DANet, to evaluate DME extraction performance. DME-DeepLabV3+ model had better DME extraction performance than other models as shown by greater pixel accuracy (PA), mean pixel accuracy (MPA), precision (Pre), recall (Re), F1-score (F1), and mean Intersection over Union (MIoU), which were 98.71%, 95.23%, 91.19%, 91.12%, 91.15%, and 91.18%, respectively.

Discussion: DME-DeepLabV3+ model is suitable for DME extraction in OCT images and can assist the ophthalmologists in the management of ocular diseases.

KEYWORDS

diabetic macular edema, optical coherence tomography, deep learning, DeepLabV3+, extraction model

1. Introduction

Diabetic macular edema (DME) is the major cause of vision loss in the patients with diabetic retinopathy. Increasing prevalence of DME is tightly correlated with the global epidemic of diabetes mellitus (1, 2). DME is usually caused by the rupture of retinal barrier and increased permeability of retinal vessels, which is characterized by the leakage of fluid and other plasma components. The effusion can accumulate in the macula, resulting in edema (3, 4). In the clinical work, the presence and severity of retinopathy are required to be determined according to the size of edema area.

Optical Coherence Tomography (OCT) is a non-contact, non-invasive, and highly sensitive ophthalmic imaging method, which can enable early detection of diabetic macular edema by observing the transverse section of macular degeneration (5). Normal OCT image is shown in Figure 1A and OCT image with DME is shown in Figure 1B. DMEs accumulated in typical relative positions within the main retinal layers. Based on OCT patterns of DME, DME can be classified into three different patterns, including diffuse retinal thickening (DRT), cystoid macular edema (CME), and serous retinal detachment (SRD). CME normally starts to manifest symptoms in the inner retina, while SRD and DRT typically appear in the outer retina. In the severe advanced stages of DR, CMEs can also proliferate from the inner to the outer retina and merge with DRT (6). Thus, rapid and accurate detection of all types of edemas is of great significance for evaluating the progression of diabetic retinopathy. In the clinical work, DME is usually segmented by the well-trained experts (7). However, manual extraction of DME edemas is time-consuming and labor-intensive. Moreover, there is inevitable variability in the extraction results by different experts. With increased prevalence of diabetes, an increasing number of patients require disease management based on OCT images in the clinical practices. Thus, it is highly required to design an automatic method for rapid and accurate detection of DME in OCT images.

Image extraction is processed and analyzed according to the features, including image color, spatial structure, and texture information (8). Image extraction models can divide an image into several specific regions, such as threshold-based extraction model (9, 10), region-based extraction model (11, 12), and edge detection-based extraction model (13). With the development of deep learning, several models have been developed to extract DME, such as fully convolutional network (FCN), U-Net, and PSPNet. Based on these deep learning models, several scholars have also developed the improved models for DME extraction. Table 1 showed the strengths and weaknesses of different models for DME extraction.

The sources of OCT images are diverse and the quality of OCT images is not always stable. Moreover, the size and distribution of DMEs are not uniform and the borders of DMEs are blurred. Thus, it is still required to design a novel model to improve the accuracy of DME extraction in OCT images. In this study, we proposed a lightweight automatic model (DME-DeepLabV3+) based on the DeepLabV3+ architecture. The major contributions of the proposed DME-DeepLabV3+ are shown below:

- Taking MobileNetV2 as the backbone, the ability of DME-DeepLabV3+ is improved in extracting the low-level features of DME.
- Improving ASPP by the sawtooth wave-like dilation rate, DME-DeepLabV3+ avoids grid effects, learns more local information, and extracts high-level features of DME better.
- Based on the decoder, DME-DeepLabV3+ fuses the low-level and high-level features of DME, and refines the results of DME extraction.

2. Materials and methods

The flowchart of the proposed model, DME-DeepLabV3+, was shown in Figure 2.

Low-level features of DME extraction by MobileNetV2; High-level features of DME extraction by the improved ASPP; Fusion and refinement of low-level and high-level features of DME by the decoder.

2.1. Low-level features of DME extraction by MobileNetV2

DeepLabV3+ is a deep learning model for image extraction with deep convolutional nets, which takes Xception as the backbone network (21). Xception uses numerous parameters, complicated operations, and high computer performance requirements (22), which leads to several challenges for DME extraction, such as fault-extraction and over-extraction problems. MobileNetV2 is a lightweight network, which shows a great advantage to solve the fault-extraction and

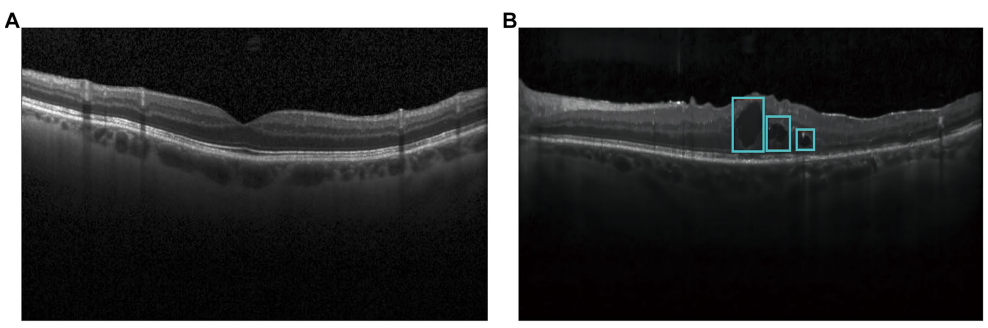

FIGURE 1
Optical coherence tomography images in diabetic patients and healthy controls. **(A)** Normal OCT image; **(B)** OCT image with DME.

TABLE 1 Strengths and weaknesses of different models for DME extraction.

Models	Strengths	Weaknesses
FCN (14)	End-to-end pixel-level classification without inputting size constraints	Ignore target boundary details and lack spatial consistency
U-Net (15)	Good extraction performance on small objects	Down-sampling operators cause spatial information loss during encoding
PSPNet (16)	Aggregate contextual information from different regions and improve the ability of obtaining global information	No effective fusion of shallow features and missing target boundary details
FCN + Sobel operator + Dijkstra algorithm (17)	Achieve better results in DICE index	Divide OCT extraction tasks into two stages, coarse and fine extraction, which makes OCT extraction cumbersome
FCN + multiphase level set (18)	Avoid overlapping phenomenon of boundary and reduce the need for large training datasets	
U-Net + Bayesian deep learning (19)	Improve the accuracy of OCT image extraction with better versatility and interpretability	Poor extraction performance for small-area objects
PSPNet + dual attention mechanism (20)	Aggregate context information of different regions	Insensitive to the information of fluid accumulation regions in DME

over-extraction problems (23). In DME-DeepLabV3+ model, we used MobileNetV2 as the backbone to simplify model structure, which could improve the extraction efficiency and reduce the problems of fault-extraction and over-extraction.

MobileNetV2 used depthwise separable convolution to reduce the number of parameters and complex operations. Depthwise separable convolution consisted of DepthWise (DW) and PointWise (PW), whereas DW performed convolution operations on each channel of the input layer and PW fused the features and obtained the feature information with stronger expressive ability. MobileNetV2 used Inverted Residual to improve the memory efficiency.

In Inverted Residual, the dimension of DME features was increased by 1×1 convolution. Next, DME features were extracted by 3×3 DW convolution, and the dimension of DME features was reduced by 1×1 convolution (Figure 3). When the stride was 1, DME output features were consistent with the input features and shortcuts were used to add the elements of DME input and output. When the stride was 2, no shortcut was required. At the same time, a linear bottleneck neural network was used in the last 1×1 convolutional layer of Inverted Residual, which could reduce the loss of low-dimensional feature of DME information.

Compared with DeepLabV3+ with Xception as the backbone network, DME-DeepLabV3+ with MobileNetV2 as the backbone network not only improved the accuracy but also improved the efficiency in DME extraction.

2.2. High-level features of DME extraction by the improved ASPP

ASPP consists of atrous convolution with different dilation rates, which strikes the best trade-off between multi-scale feature extraction and context assimilation, especially for small objects (24). DME has multi-scale features, especially with several small areas of edema. ASPP was then used to extract high-level features of DME. However, the dilation rate in ASPP had a grid effect, which not only lost the semantic information but also ignored the consistency of local information in edema regions (25). Here, we replaced the original dilation rate with the sawtooth wave-like dilation rate to improve ASPP for extracting the high-level features of DME. A sawtooth wave-like dilation rate was formed by the repeated combination of two sets of the same "rising edge" type dilation rate.

Figures 4, 5 show the illustration of the atrous convolution principle of DeepLabV3+ and DME-DeepLabV3+, respectively. Figures 4A,B show RF (receptive field) and the number of calculation times of DeepLabV3+. Figures 5A,B show RF and the number of calculation times of DME-DeepLabV3+. The results show that there was about 73% of information loss due to the grid effect in DeepLabV3+ model. In DME-DeepLabV3+ model, each pixel was effectively used and involved in further computations. Compared with DeepLabV3+ model, increased dilation rate in DME-DeepLabV3+ model can avoid the grid effects and learn more local information.

2.3. Fusion and refinement of low-level and high-level features by the decoder

Low-level and high-level features of DME were extracted by MobileNetV2 and the improved ASPP, respectively. All features of DME were fused and refined by the decoder. The decoder is mainly composed of ordinary convolution and fusion layers. It fuses the features extracted from the encoder, uses the up-sampling to restore the feature dimension, and outputs the prediction results of the same size with less information loss as possible (26). In the decoder, low-level features with fewer channels were obtained by 1×1 convolution. Bilinear up-sampling of high-level features were conducted by a factor of 4. The concatenation features were obtained by concatenating the low-level features and high-level features and a feature concatenation was refined by a few 3×3 convolutions. Finally, the results of DME extraction were output following another bilinear up-sampling by a factor of 4.

2.4. Ethical statement

The design and conduct of this study adhere to the intent and principles of the Declaration of Helsinki. The protocols were also reviewed and approved by the ethical committee of Eye Hospital (Nanjing medical university). Informed consents were obtained from all participants.

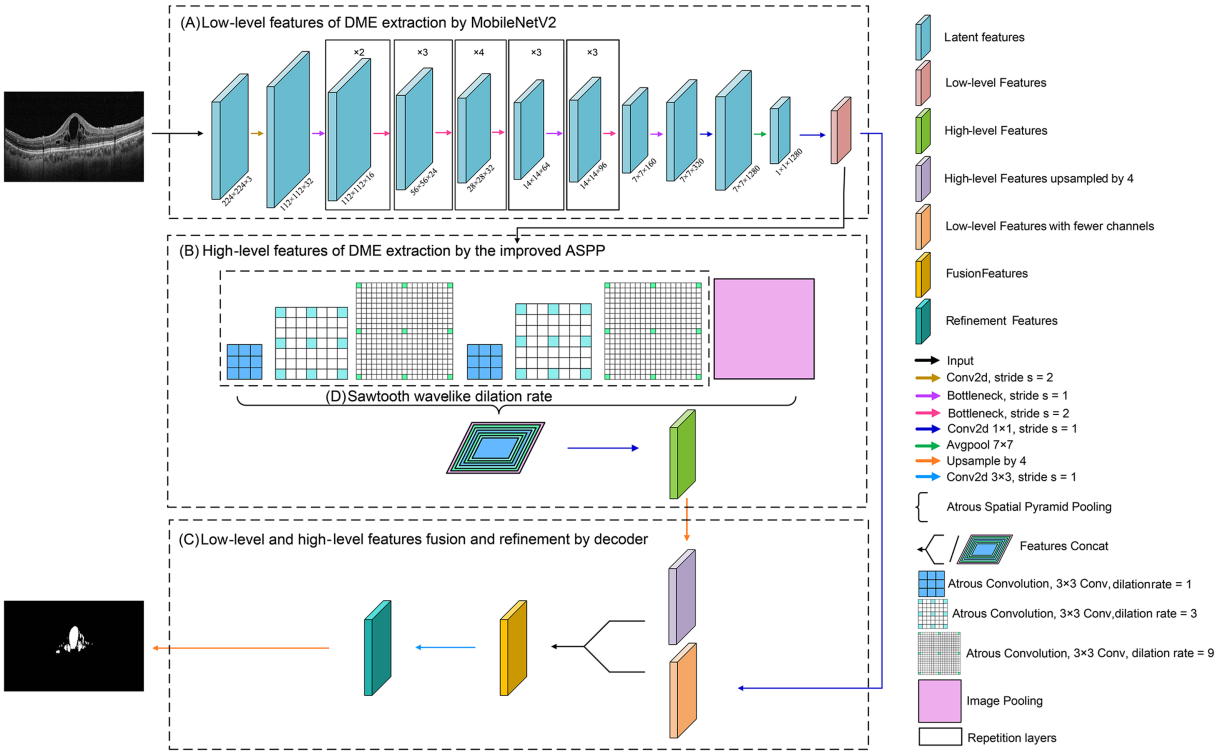

FIGURE 2
Flowchart of DME extraction by DME-DeepLabV3+ model.

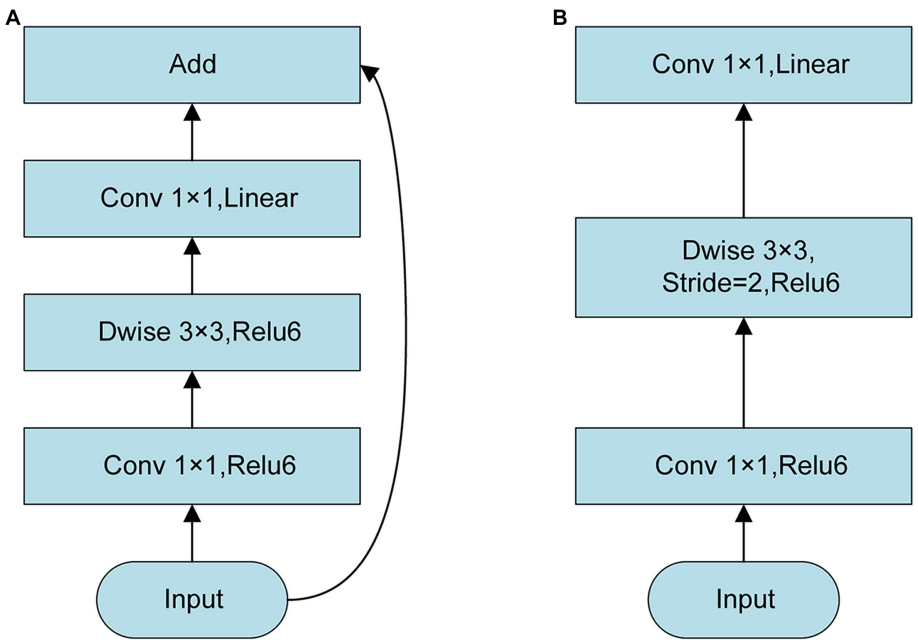

FIGURE 3
Structure of Inverted Residual. **(A)** Stride = 1; **(B)** Stride = 2.

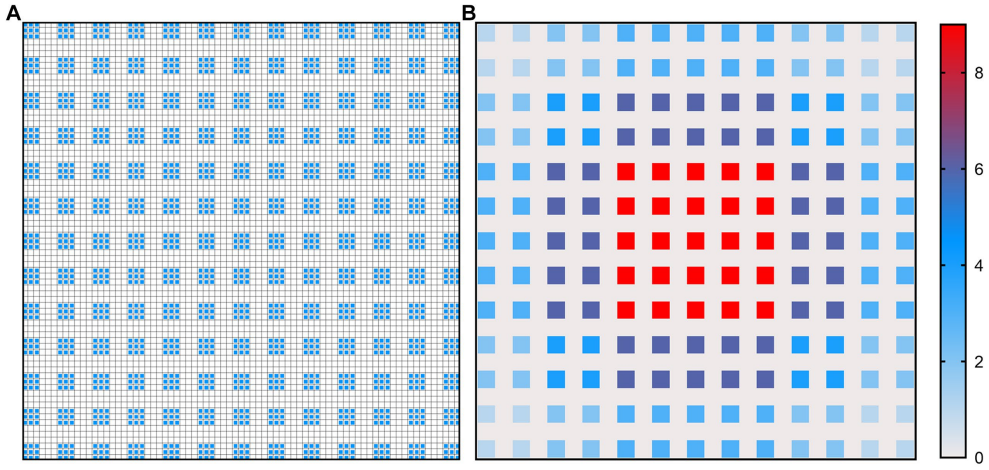

FIGURE 4
Illustration of atrous convolution principle of DeepLabV3+ with dilation rate = [1, 6, 12, 18] and RF = 75 × 75. **(A)** Effective pixels in RF, which were marked in blue; **(B)** The number of calculation times of each pixel.

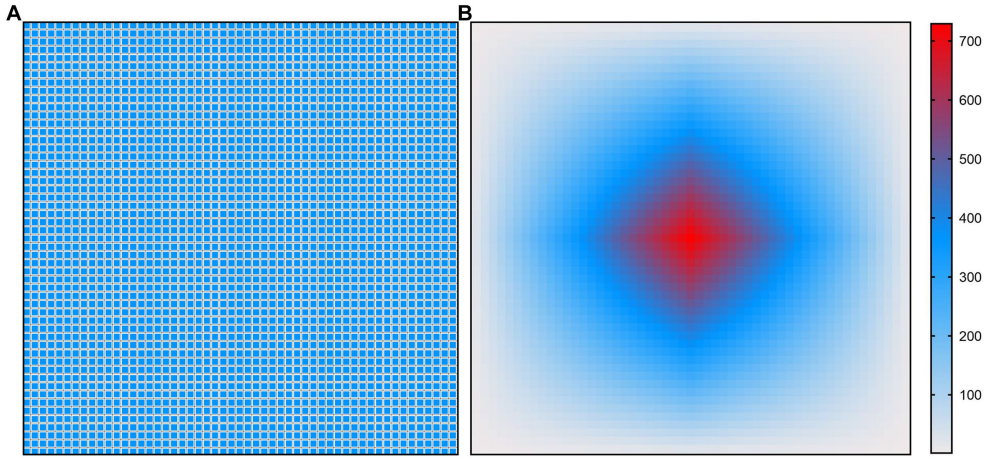

FIGURE 5
Illustration of atrous convolution principle of DME-DeepLabV3+ with dilation rate = [1, 3, 9, 1, 3, 9] and RF = 53 × 53. **(A)** Effective pixels in RF, which were marked in blue; **(B)** The number of calculation times of each pixel.

2.5. Datasets

The datasets contained 1711 OCT images, including 416 images (512×512 pixels) selected from the Kermany dataset (27) and 1,295 OCT datasets (938×420 pixels) collected from the Affiliated Eye Hospital, Nanjing Medical University. All patients were required to undergo OCT scanning by a spectral domain OCT (RTVue, Optovue Inc., United States). These OCT images were centered on the macula with an axial resolution of 10 μm and a 24-bit depth and were acquired in 2 s, covering 4×4 mm area. Inclusion criteria were as follows: the presence of macular edema in at least one eye and clear optical media allowing OCT imaging with good quality. Subsets of 1,369, 171, and 171 OCT images were randomly selected for training, validation, and testing, respectively. Each OCT image was individually labeled by three experienced clinicians who had more than 10-year clinical working experience. The annotation results were binarized by MATLAB software, where the background was labeled as 0 and the DME labeled as 1. Due to the limited human energy, some artificial deviations were inevitable. For these images, a senior expert was consulted and thorough rounds of discussion and adjudication were conducted to ensure the accuracy of the labeling. The original OCT images and ground truth are shown in Figure 6.

3. Implementation

The hardware configurations used for this study are shown below: Windows 10, NVIDIA GeForce RTX 3060. The software environment is the deep-learning framework PyTorch 1.10.0, CUDA 11.3, and the programming language Python 3.9.

4. Evaluation metrics

Seven metrics were calculated to estimate the extraction performance of DME-DeepLabV3+, including pixel accuracy (PA),

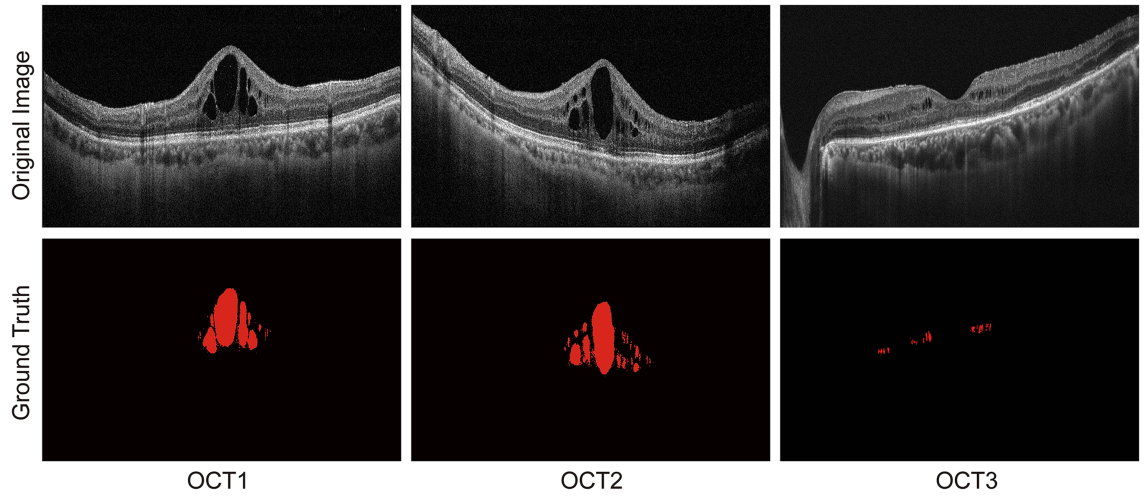

FIGURE 6
Original OCT images and DME labeling by three experienced clinicians.

mean pixel accuracy (MPA), precision (Pre), recall (Re), F1-score (F1), mean intersection over union (MIoU), and frames per second (FPS).

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}} \quad (1)$$

$$MPA = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \quad (2)$$

$$Pre = \frac{TP}{TP + FP} \quad (3)$$

$$Re = \frac{TP}{TP + FN} \quad (4)$$

$$F1 = \frac{2 \times Pre \times Re}{Pre + Re} \quad (5)$$

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \quad (6)$$

$$FPS = \frac{frameNum}{elapsedTime} \quad (7)$$

TP, FP, and FN denote the true positive region, false positive region, and false negative region, respectively. p_{ii} is the number of edema area pixels which was correctly classified as edema areas; p_{ij} is the number of background area pixels which are misclassified as edema areas; p_{ji} is the number of edema area pixels which are incorrectly classified as the background; k is the labeling results of different classes, where k = 0 expressed as background class and k = 1 as DME class; *frameNum* is the number of OCT images that are input to the model when performing inference; *elapsedTime* is the time consumed by the model when performing inference. PA is the overall pixel accuracy. MPA is the average pixel accuracy of DME and background. Pre and Re are the proportion of real DME regions in the samples predicted as DME and the proportion of correct predictions in all DME, respectively. F1-score (F1) is a balanced metric and determined by precision and recall. MIoU is a metric to measure the similarity of ground truth and prediction. FPS is the number of OCT images inferred per second.

5. Results

To evaluate the performance of DME extraction of DME-DeepLabV3+ model, two comparative experiments were performed. In experiment 1, DME extraction performance of DME-DeepLabV3+ model was evaluated by comparing against DeepLabV3+ model under different settings. In experiment 2, DME extraction performance of DME-DeepLabV3+ model was evaluated by comparing against other end-to-end models, including FCN, UNet, PSPNet, ICNet, and DANet.

5.1. Experiment 1 (ablation experiment)

To evaluate the effects of MobileNetV2 and the improved ASPP on DME extraction performance, the proposed DME-DeepLabV3+ model was compared against DeepLabV3+ model with different settings, including DeepLabV3+, DeepLabV3+ with MobileNetV2 (MobileNetV2-DeepLabV3+), DeepLabV3+ with the improved ASPP

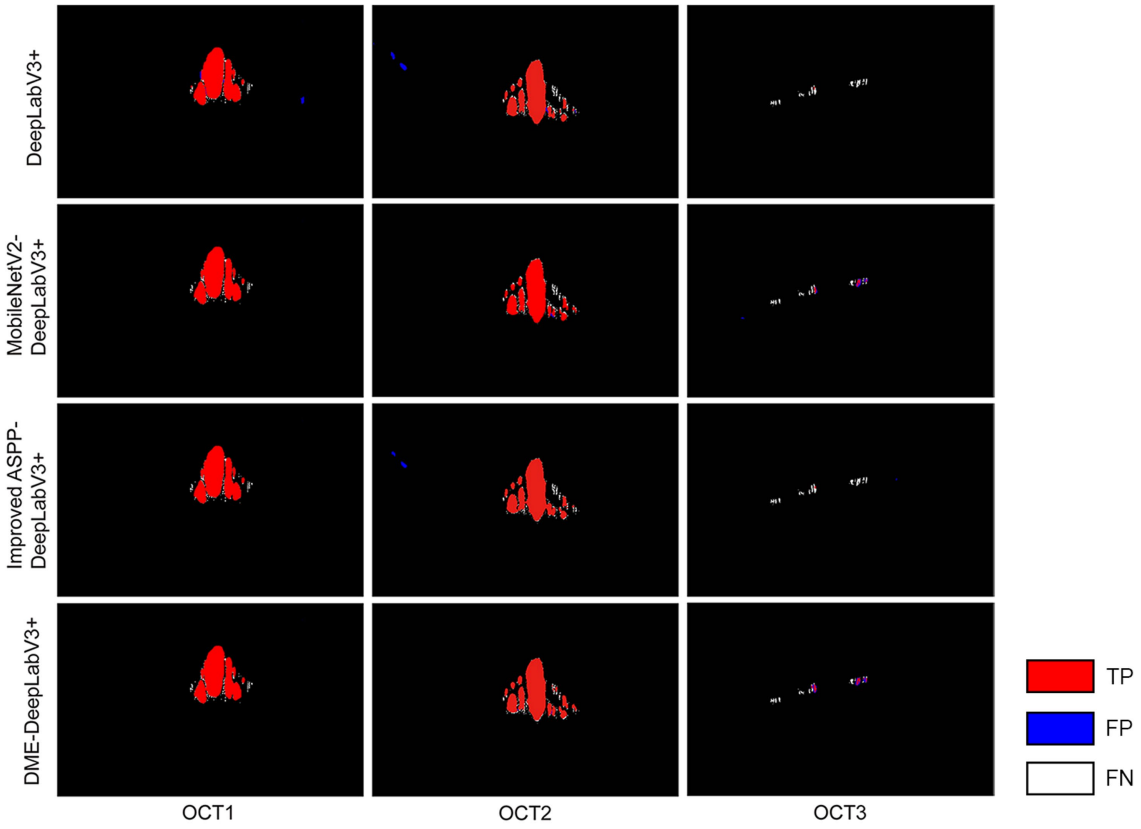

FIGURE 7
DME extraction results by DeepLabV3+ model with different settings Red, blue, and white DME regions represented TP regions, FP regions and FN regions.

TABLE 2 Evaluation metrics of DME extraction by DeepLabV3+ model with different settings.

Models	Evaluation metrics						
	PA(%)	MPA(%)	Pre(%)	Re(%)	F1(%)	MIoU(%)	FPS(it/s)
DeepLabV3+	98.41 ± 0.07	93.45 ± 0.36	90.45 ± 0.85	87.78 ± 0.62	89.09 ± 0.16	89.27 ± 0.20	5.84 ± 0.87
MobileNetV2-DeepLabV3+	98.69 ± 0.01	94.95 ± 0.14	91.02 ± 0.58	91.09 ± 0.43	91.06 ± 0.08	91.05 ± 0.05	9.24 ± 0.33
Improved ASPP-DeepLabV3+	98.47 ± 0.03	93.90 ± 0.40	90.43 ± 0.30	88.49 ± 0.68	89.45 ± 0.42	89.61 ± 0.25	5.59 ± 0.13
DME-DeepLabV3+	98.71 ± 0.02	95.23 ± 0.26	91.19 ± 0.44	91.12 ± 0.48	91.15 ± 0.02	91.18 ± 0.09	9.03 ± 0.43

(Improved ASPP-DeepLabV3+). Figure 7 showed the DME extraction results by DeepLabV3+ model with different settings, where red, blue, and white DME regions represented true positive (TP) regions, false positive (FP) regions and false negative (FN) regions, respectively. DeepLabV3+ model led to some missed and false extraction of DME. MobileNetV2-DeepLabV3+ and improved ASPP-DeepLabV3+ reduced the missed and false extraction of DME. However, the missed extraction still existed in small edematous regions as shown in OCT2. DME-DeepLabV3+ had better extraction performance, especially in small-scale macular edema regions. The extraction results of DME-DeepLabV3+ were close to the ground truth.

Table 2 showed the results of evaluation metrics for DeepLabV3+ under different settings. Compared with DeepLabV3+ model, the MobileNetV2-DeepLabV3+ enhanced the scores of PA, MPA, Pre, Re, F1, MIoU, and FPS of DME extraction results, which were 98.69(0.28↑), 94.95(1.50↑), 91.02(0.57↑), 91.09(3.31↑), 91.06(1.97↑), 91.05(1.78↑), and 9.24(3.40↑), respectively. FPS increased by about 58%. The improved ASPP-DeepLabV3+ enhanced the scores of PA, MPA, Re, F1, and MIoU of DME extraction results, which were 98.47(0.06↑), 93.90(0.45↑), 88.49(0.71↑), 89.45(0.36↑), and 89.61(0.34↑), respectively. DME-DeepLabV3+ enhanced the scores of PA, MPA, Pre, Re, F1, and MIoU of DME extraction results, which were 98.71(0.30↑), 95.23(1.78↑), 91.19(0.74↑), 91.12(3.34↑), 91.15(2.06↑), and 91.18(1.91↑), respectively. FPS was 9.03, which was lower than that of MobileNetV2-DeepLabV3+ (0.21↓).

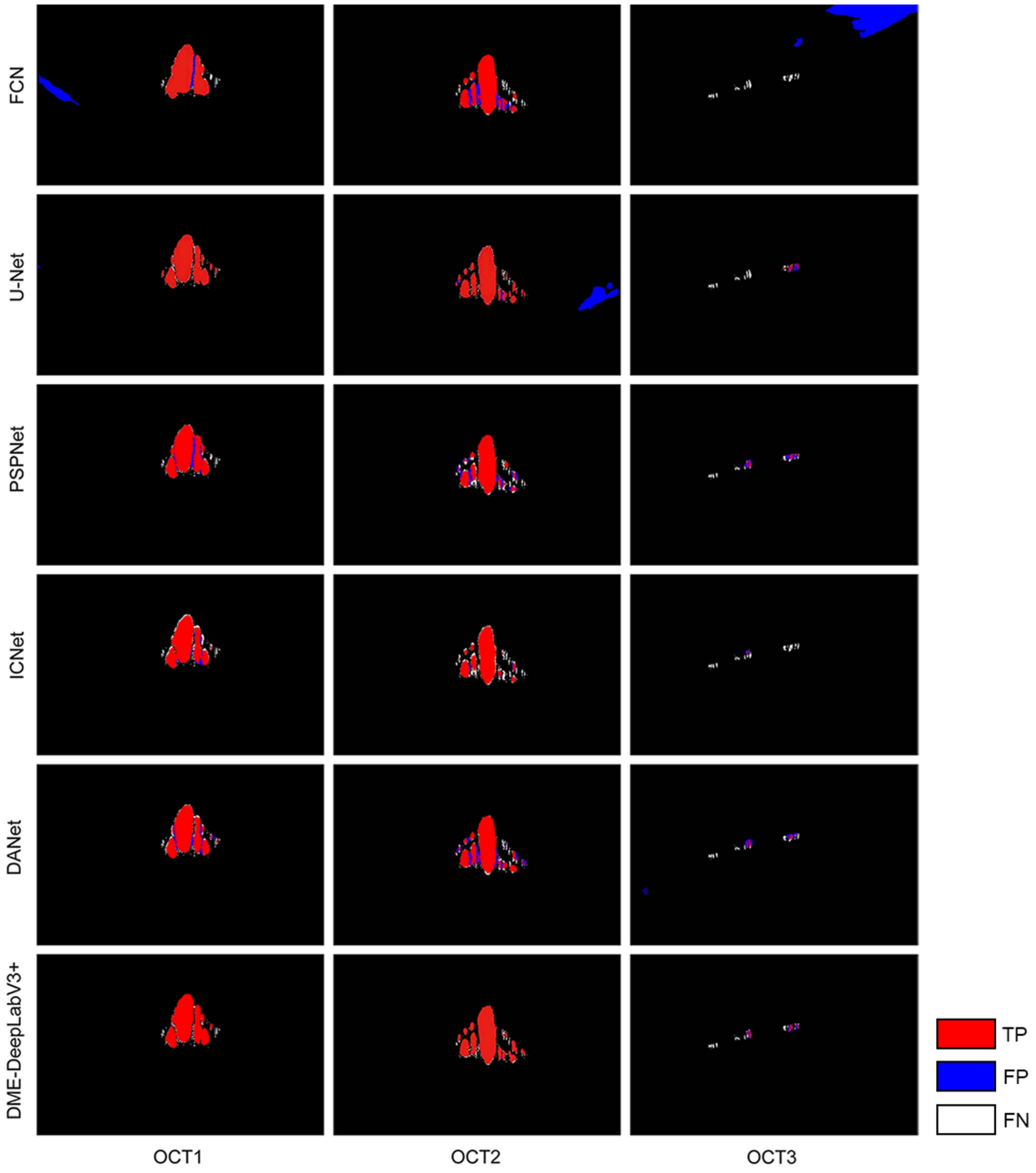

FIGURE 8
DME extraction results by different models Red, blue, and white DME regions represented TP regions, FP regions and FN regions.

5.2. Experiment 2 (comparative experiment)

We evaluate DME extraction performance of DME-DeepLabV3+ model by comparing against other models, including FCN, UNet, PSPNet, ICNet, and DANet. Figure 8 showed DME extraction results by different models, where red, blue, and white DME regions represented true positive (TP) regions, false positive (FP) regions, and false negative (FN) regions, respectively. As shown in Figure 8, DME extraction results of DME-DeepLabV3+ were close to the ground truth. A part of the background was extracted falsely by FCN, U-Net, or DANet models. Compared with FCN and U-Net, PSPNet and ICNet reduced the fault-extraction, but the small-scale macular edema was over-extracted. Table 3 showed the parameter configurations of different models and Table 4 showed the results of DME evaluation metrics. Compared with FCN, U-Net, PSPNet, and DANet models, DME-DeepLabV3+ achieved higher scores of PA, MPA, and FPS. As for Pre, Re, F1, and MIoU, DME-DeepLabV3+ substantially exceeded other models. Compared with ICNet, DME-DeepLabV3+ achieved a better trade-off in the accuracy and efficiency for DME extraction.

TABLE 3 Parameter configurations of different DME extraction models.

Models	Backbone	Learning rate	Total epochs	Batch size
FCN	ResNet50	0.01	200	1
U-Net	VGG16	0.01	200	2
PSPNet	ResNet50	0.01	200	2
ICNet	ResNet50	0.01	200	2
DANet	ResNet101	0.0001	200	2
DME-DeepLabV3+	MobileNetV2	0.01	200	2

6. Conclusion and discussion

With the increased incidence of diabetes, DME has become a major cause of visual impairment in diabetic patients (28). DME occurs as a result of the disruption of blood-retinal barrier and consequent increase in vascular permeability (29). OCT allows longitudinal, functional, and microstructural analysis of human macula (30). However, manual labeling DME is time-consuming and labor-intensive (31). Automatic extraction of DME based on machine learning can help physicians assess disease severity, determine treatment options, and improve life quality of patients (32). Thus, it is urgent to develop an efficient model for DME detection. In this study, we proposed a lightweight model based on DeepLabV3+, termed DME-DeepLabV3+, to extract DME in OCT images. MobileNetV2 architecture was used as the backbone to extract the low-level features of DME and reduce the model complexity to enhance DME detection accuracy. With the help of improved ASPP structure, DME-DeepLabV3+ avoided the grid effects and learned more local information. Finally, the decoder was used to fuse the low-level and high-level features of DME and refined the results of DME extraction.

OCT image modality has been widely used for detecting DME due to its non-invasive and high-resolution features. Considering the clinical characteristics that are present in OCT images such as thickness, reflectivity or intraretinal fluid accumulation, DMEs have been categorized into three different types: SRD, DRT, and CME. Traditional DME detection is based on the low-level hand-crafted features, which require significant domain knowledge and are sensitive to the variations of lesions. Given great variability of morphology, shape, and relative ME position, it is difficult to detect all three ME types simultaneously. Our proposed model can achieve automatic and simultaneous detection of all three types of ME (SRD, DRT, and CME) in the ophthalmological field. However, the accuracy of DRT detection is still not good as SRD or CME detection. DRT is characterized by a sponge-like retinal swelling of the macula with reduced intraretinal reflectivity. In addition, DRT is characterized by uniform thickening of inner retinal layers but without macroscopic optical empty spaces. Thus, further improvement of our proposed model is still required for enhancing the accuracy of the automatic detection of DRT edemas.

In clinical practice, layer segmentation and fluid area segmentation can provide qualitative information and visualization of retinal structure, which is important for DME assessment and monitoring. Although commercial OCT devices with on-board proprietary segmentation software are available, the definition of retinal boundaries varies between the manufacturers, making the quantitative retinal thickness difficult. In addition, proprietary software is difficult to be used for image analysis from other OCT devices, which poses a great challenge for effective diagnosis of DME (33). Although automated methods for layer segmentation have been proposed, most of them usually ignore the priority of mutually exclusive relationships between different layers, which can also affect the accuracy of DME assessment (34). In future study, we will improve our model to consider both layer segmentation and fluid area segmentation for better monitoring the progression of DME in retinal diseases.

Both microaneurysm (MA) formation and DME lesions are the important signs of DR. Early and accurate detection of DME and MAs can reduce the risk of DR. Due to the small size of MA lesions and low contrast between MA lesion and retinal background, automated MA detection is still challenging. Many imaging modalities have been used to detect MAs, including color fundus images, optical coherence tomography angiography (OCTA), and fluorescein fundus angiography (FFA). However, MAs are situated on the capillaries, which are not often visible in color fundus images. Although FFA can capture the small changes of retinal vessels, FFA is an invasive method compared with other imaging modalities. OCTA can provide the detailed visualization of vascular perfusion and allow for the examination of retinal vasculature in 3D (35). In future study, we would also improve our model by considering the segmentation of FFA for better monitoring the progression of DME in retinal diseases. We would design modules with better feature extraction capabilities, such as embedding attention mechanism to the model, strengthening key information, suppressing useless

TABLE 4 Evaluation metrics of DME extraction by different models.

Models	Evaluation metrics						
	PA(%)	MPA(%)	Pre(%)	Re(%)	F1(%)	MIoU(%)	FPS(it/s)
FCN	98.27 ± 0.11	89.88 ± 1.50	81.76 ± 1.20	79.44 ± 1.40	80.58 ± 0.85	82.66 ± 0.48	4.08 ± 0.20
U-Net	98.61 ± 0.01	90.73 ± 1.05	86.31 ± 0.41	81.15 ± 0.65	83.58 ± 0.15	85.32 ± 0.23	3.43 ± 0.24
PSPNet	98.69 ± 0.04	92.52 ± 0.41	84.61 ± 1.27	85.76 ± 0.90	85.17 ± 0.24	86.41 ± 0.21	7.96 ± 0.48
ICNet	98.07 ± 0.02	90.94 ± 0.20	90.73 ± 0.78	82.57 ± 0.93	86.45 ± 0.29	86.86 ± 0.13	15.86 ± 0.30
DANet	98.06 ± 0.01	92.15 ± 0.16	87.44 ± 0.77	85.68 ± 1.16	86.54 ± 0.21	87.11 ± 0.08	6.12 ± 0.50
DME-DeepLabV3+	98.71 ± 0.02	95.23 ± 0.26	91.19 ± 0.44	91.12 ± 0.48	91.15 ± 0.02	91.18 ± 0.09	9.03 ± 0.43

information, and better capturing contextual information, to improve the generalization of the model for the diagnosis of retinal diseases.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

BY, ZW, and QJ were responsible for the conceptualization and data collection. BY, ZW, and YB were responsible for the experiment design and manuscript writing. JL and LS conducted the data collection and data entry. BY and ZW were responsible for overall supervision and manuscript revision. All authors contributed to the article and approved the submitted version.

Funding

This research was generously supported by the grants from the National Natural Science Foundation of China (Grant Nos. 82171074 and 82070983).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer ZH declared a shared parent affiliation with the authors LS, QJ, and BY to the handling editor at the time of review.

Publisher's note

References

1. Schmidt-Erfurth U, Garcia-Arumi J, Bandello F, Berg K, Chakravarthy U, Gerendas BS, et al. Guidelines for the Management of Diabetic Macular Edema by the European Society of Retina Specialists (Euretina). *Ophthalmologica*. (2017) 237:185–222. doi: 10.1159/000458539

2. Steinmetz JD, Bourne RR, Briant PS, Flaxman SR, Taylor HR, Jonas JB, et al. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to vision 2020: the right to sight: an analysis for the global burden of disease study. *Lancet Glob Health*. (2021) 9:e144–60. doi: 10.1016/S2214-109X(20)30489-7

3. Tan GS, Cheung N, Simo R, Cheung GC, Wong TY. Diabetic Macular Oedema. *Lancet Diabetes Endocrinol*. (2017) 5:143–55. doi: 10.1016/S2213-8587(16)30052-3

4. Das A, McGuire PG, Rangasamy S. Diabetic macular edema: pathophysiology and novel therapeutic targets. *Ophthalmology*. (2015) 122:1375–94. doi: 10.1016/j.ophtha.2015.03.024

5. Van Melkebeke L, Barbosa-Breda J, Huygens M, Stalmans I. Optical coherence tomography angiography in Glaucoma: a review. *Ophthalmic Res*. (2018) 60:139–51. doi: 10.1159/000488495

6. Wu Q, Zhang B, Hu Y, Liu B, Cao D, Yang D, et al. Detection of morphologic patterns of diabetic macular edema using a deep learning approach based on optical coherence tomography images. *Retina*. 41:1110–7. doi: 10.1097/iae.0000000000002992

7. Wang Z, Zhong Y, Yao M, Ma Y, Zhang W, Li C, et al. Automated segmentation of macular edema for the diagnosis of ocular disease using deep learning method. *Sci Rep*. (2021) 11:1–12. doi: 10.1038/s41598-021-92458-8

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

9. Makandar A, Halalli B. Threshold based segmentation technique for mass detection in mammography. *J Comput*. (2016) 11:472–8. doi: 10.17706/jcp.11.6.472-478

10. Zebari DA, Zeebaree DQ, Abdulazeez AM, Haron H, Hamed HNA. Improved threshold based and trainable fully automated segmentation for breast Cancer boundary and pectoral muscle in mammogram images. *IEEE Access*. (2020) 8:203097–116. doi: 10.1109/ACCESS.2020.3036072

11. Biratu ES, Schwenker F, Debelee TG, Kebede SR, Negera WG, Molla HT. Enhanced region growing for brain tumor Mr image segmentation. *J Imaging*. (2021) 7:22. doi: 10.3390/jimaging7020022

12. Liu J, Yan S, Lu N, Yang D, Fan C, Lv H, et al. Automatic segmentation of foveal avascular zone based on adaptive watershed algorithm in retinal optical coherence tomography angiography images. *J Innov Opt Health Sci*. (2022) 15:2242001. doi: 10.1142/S1793545822420019

13. Chatterjee S, Suman A, Gaurav R, Banerjee S, Singh AK, Ghosh BK, et al. Retinal blood vessel segmentation using edge detection method. *J Phys Conf Ser*. (2021) 1717:012008. doi: 10.1088/1742-6596/1717/1/012008

14. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. *IEEE Trans Pattern Anal Mach Intell*. (2017) 39:640–51. doi: 10.1109/cvpr.2015.7298965

15. Ronneberger O, Fischer P, Brox T. *U-net: convolutional networks for biomedical image segmentation*. International conference on medical image computing and computer-assisted intervention. Berlin: Springer (2015).

16. Zhao H, Shi J, Qi X, Wang X, Jia J. *Pyramid scene parsing network. Proceedings of the IEEE conference on computer vision and pattern recognition*. (2017).

17. Ben-Cohen A, Mark D, Kovler I, Zur D, Barak A, Iglicki M, et al. *Retinal layers segmentation using fully convolutional network in Oct images*. Israel: RSIP Vision, 1–8. (2017).

18. Ruan Y, Xue J, Li T, Liu D, Lu H, Chen M, et al. Multi-phase level set algorithm based on fully convolutional networks (Fcn-Mls) for retinal layer segmentation in Sd-Oct images with central serous Chorioretinopathy (Csc). *Biomed Opt Express*. (2019) 10:3987–4002. doi: 10.1364/BOE.10.003987

19. Orlando JI, Seeböck P, Bogunović H, Klimscha S, Grechenig C, Waldstein S, et al., *U2-net: a Bayesian U-net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological Oct scans*. 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019) (2019).

20. Wu J, Chen J, Xiao Z, Geng L. *Automatic layering of retinal Oct images with dual attention mechanism*. 2021 3rd international conference on intelligent medicine and image processing (2021).

21. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H. *Encoder-decoder with Atrous separable convolution for semantic image segmentation*. Proceedings of the European conference on computer vision (ECCV) (2018).

22. Zhang R, Du L, Xiao Q, Liu J. Comparison of backbones for semantic segmentation network. *J Phys Conf Ser*. (2020) 1544:012196. doi: 10.1088/1742-6596/1544/1/012196

23. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. *Mobilenetv2: Inverted Residuals and Linear Bottlenecks*. Proceedings of the IEEE conference on computer vision and pattern recognition (2018).

24. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected Crfs. *IEEE Trans Pattern Anal Mach Intell*. (2017) 40:834–48. doi: 10.1109/tpami.2017.2699184

25. Wang P, Chen P, Yuan Y, Liu D, Huang Z, Hou X, et al. *Understanding convolution for semantic segmentation*. 2018 IEEE winter conference on applications of computer vision (WACV) (2018).

26. Badrinarayanan V, Kendall A, Cipolla R. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intell*. (2017) 39:2481–95. doi: 10.1109/tpami.2016.2644615

27. Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cells*. (2018) 172:1122–1131.e9. doi: 10.1016/j.cell.2018.02.010

28. Kim EJ, Lin WV, Rodriguez SM, Chen A, Loya A, Weng CY. Treatment of diabetic macular edema. *Curr Diab Rep*. (2019) 19:1–10. doi: 10.1007/s11892-019-1188-4

29. Noma H, Yasuda K, Shimura M. Involvement of cytokines in the pathogenesis of diabetic macular edema. *Int J Mol Sci*. (2021) 22:3427. doi: 10.3390/ijms22073427

30. Daruich A, Matet A, Moulin A, Kowalczuk L, Nicolas M, Sellam A, et al. Mechanisms of macular edema: beyond the surface. *Prog Retin Eye Res*. (2018) 63:20–68. doi: 10.1016/j.preteyeres.2017.10.006

31. Pekala M, Joshi N, Liu TA, Bressler NM, DeBuc DC, Burlina P. Deep learning based retinal Oct segmentation. *Comput Biol Med*. (2019) 114:103445. doi: 10.1016/j.compbiomed.2019.103445

32. de Moura J, Samagaio G, Novo J, Almuina P, Fernández MI, Ortega M. Joint diabetic macular edema segmentation and characterization in Oct images. *J Digit Imaging*. (2020) 33:1335–51. doi: 10.1007/s10278-020-00360-y

33. Alex V, Motevasseli T, Freeman WR, Jayamon JA, Bartsch D-UG, Borooah S. Assessing the validity of a cross-platform retinal image segmentation tool in Normal and diseased retina. *Sci Rep*. (2021) 11:21784. doi: 10.21203/rs.3.rs-396609/v1

34. Wei H, Peng P. The segmentation of retinal layer and fluid in Sd-Oct images using Mutex Dice loss based fully convolutional networks. *IEEE Access*. (2020) 8:60929–39. doi: 10.1109/ACCESS.2020.2983818

35. Hervella ÁS, Rouco J, Novo J, Ortega M. Retinal microaneurysms detection using adversarial pre-training with unlabeled multimodal images. *Inf Fusion*. (2022) 79:146–61. doi: 10.1016/j.inffus.2021.10.003

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 16 October 2023
DOI 10.3389/fmed.2023.1270570

OPEN ACCESS

EDITED BY
Darren Shu Jeng Ting,
University of Nottingham, United Kingdom

REVIEWED BY
Kunpeng Pang,
Shandong University, China
Federico Castro-Muñozledo,
National Polytechnic Institute of Mexico (CINVESTAV), Mexico

*CORRESPONDENCE
William Speier
✉ speier@ucla.edu
Sophie X. Deng
✉ deng@jsei.ucla.edu

RECEIVED 31 July 2023
ACCEPTED 02 October 2023
PUBLISHED 16 October 2023

CITATION
Gibson D, Tran T, Raveendran V, Bonnet C, Siu N, Vinet M, Stoddard-Bennett T, Arnold C, Deng SX and Speier W (2023) Latent diffusion augmentation enhances deep learning analysis of neuro-morphology in limbal stem cell deficiency.
Front. Med. 10:1270570.
doi: 10.3389/fmed.2023.1270570

Latent diffusion augmentation enhances deep learning analysis of neuro-morphology in limbal stem cell deficiency

David Gibson[1], Thai Tran[1], Vidhur Raveendran[2], Clémence Bonnet[3,4], Nathan Siu[1,4,5], Micah Vinet[2,5], Theo Stoddard-Bennett[6], Corey Arnold[1,2,5], Sophie X. Deng[4,5,7]* and William Speier[1,2,5]*

[1]Medical Informatics Home Area, Graduate Programs in Bioscience, University of California, Los Angeles, Los Angeles, CA, United States, [2]Department of Bioengineering, University of California, Los Angeles, Los Angeles, CA, United States, [3]Ophthalmology Department, Cochin Hospital and Paris Cité University, AP-HP, Paris, France, [4]Stein Eye Institute, University of California, Los Angeles, Los Angeles, CA, United States, [5]Computational Diagnostics Lab, University of California, Los Angeles, Los Angeles, CA, United States, [6]David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, United States, [7]Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA, United States

Introduction: Limbal Stem Cell Deficiency (LSCD) is a blinding corneal disease characterized by the loss of function or deficiency in adult stem cells located at the junction between the cornea and the sclera (i.e., the limbus), namely the limbal stem cells (LSCs). Recent advances in *in vivo* imaging technology have improved disease diagnosis and staging to quantify several biomarkers of *in vivo* LSC function including epithelial thickness measured by anterior segment optical coherence tomography, and basal epithelial cell density and subbasal nerve plexus by *in vivo* confocal microscopy. A decrease in central corneal sub-basal nerve density and nerve fiber and branching number has been shown to correlate with the severity of the disease in parallel with increased nerve tortuosity. Yet, image acquisition and manual quantification require a high level of expertise and are time-consuming. Manual quantification presents inevitable interobserver variability.

Methods: The current study employs a novel deep learning approach to classify neuron morphology in various LSCD stages and healthy controls, by integrating images created through latent diffusion augmentation. The proposed model, a residual U-Net, is based in part on the InceptionResNetV2 transfer learning model.

Results: Deep learning was able to determine fiber number, branching, and fiber length with high accuracy (R2 of 0.63, 0.63, and 0.80, respectively). The model trained on images generated through latent diffusion on average outperformed the same model when trained on solely original images. The model was also able to detect LSCD with an AUC of 0.867, which showed slightly higher performance compared to classification using manually assessed metrics.

Discussion: The results suggest that utilizing latent diffusion to supplement training data may be effective in bolstering model performance. The results of the model emphasize the ability as well as the shortcomings of this novel deep learning approach to predict various nerve morphology metrics as well as LSCD disease severity.

KEYWORDS

deep learning, ophthalmology, machine learning, limbal stem cell deficiency, *in vivo* confocal microscopy

Introduction

Limbal stem cells (LSCs) are adult stem cells located at the junction between the cornea and the sclera which are responsible for continuous corneal epithelial renewal (1). Limbal stem cell deficiency (LSCD) is a potentially blinding corneal disease caused by a loss of function functional LSCs (2). This condition presents with debilitating symptoms including photophobia, burning, irritation, and loss of vision potentially leading to blindness. Without LSCs, conjunctival epithelial cells invade the corneal surface leading to decreased vision as a result of conjunctivalization of the cornea (3). Recent guidelines have clarified the disease definition, diagnosis, staging, and management of LSCD (2, 4).

As clinical presentation does not always correlate with the level of LSCD or *in vivo* LSC function (5, 6), it is recommended to perform additional diagnostic tests including *in vivo* imaging such as anterior segment optical coherence tomography (AS-OCT) and *in vivo* laser scanning confocal microscopy (ICVM) to evaluate *in vivo* biomarkers of the disease (2, 7). A composite score correlating with disease severity can then be generated by combining these biomarkers (7). One of the biomarkers is the central corneal sub-basal nerve density (SND) (8, 9). A decrease in SND correlates with the severity of LSCD. Other nerve parameters correlating with the severity of LSCD include central corneal sub-basal nerve branching number, fiber number, and fiber tortuosity (7, 8). Quantification of these nerve parameters requires highly trained personnel to manually annotate images, is time-consuming, and is open to interrater variability.

There have been many other attempts at automating the process of neuro-morphological classification outside of the cornea. Most of the current literature focuses on the segmentation of neuro-images and classifying neurons through the use of various deep learning algorithms (10, 11). While many approaches use convolutional neural networks (CNN), recent benchmarking efforts have revealed that linear discriminant analysis (LDA) can serve as a promising discriminatory classifier (10). Prior research in other domains of ophthalmology has used deep learning models to aid in the diagnosis of diabetic neuropathy and fungal keratitis using IVCM images (12–14). Diagnostic challenges can be remedied by integrating recent advances in computational approaches into the current clinical workflow. This approach can increase diagnosis precision and reduce time-to-treatment and clinician burden.

To address these challenges, our morphological classifier automates the process of diagnosing LSCD using nerve morphology features from ICVM images. Total corneal nerve fiber length, corneal nerve fiber density, corneal nerve branch density, and tortuosity coefficient are among the nerve morphology biomarkers used for disease staging clinically (7). These biomarkers have been shown to correlate significantly with LSCD as well as other biomarkers such as basal cell density (8, 9). Examining these biomarkers will elucidate how these quantifiable morphology features relate to LSCD disease progression. We employed deep learning to classify neuron morphology. To maximize the effectiveness of these models we developed a novel pre-processing pipeline for use prior to training and testing. To overcome potential hurdles with the size of our data set we employed random sampling with replacement, as well as image augmentation and enhancement. Stable diffusion (SD) is a latent diffusion model which is generally used as a text-to-image model (15). Deep learning requires large datasets and diffusion models present an opportunity for more robust data augmentation beyond typical image transformations like rotations and flips (16). A number of recent studies have explored diffusion models for specific tasks in medical imaging, including synthesizing magnetic resonance imaging and computed tomography volume scans (17). To date, no other work in neuro-morphological classification has incorporated artificially generated images into training datasets using SD. The current study is the first to demonstrate the use of this approach in subbasal nerve analysis.

Materials and methods

Dataset

Appropriate consent was obtained from study subjects in accordance to IRB protocol (UCLA IRB #10-001601). The study was compliant with HIPAA regulations and adhered to the Declaration of Helsinki.

LSCD diagnosis was based on a comprehensive examination including history, slit lamp examination, and fluorescein-staining pattern, and confirmed in all cases by IVCM and/or AS-OCT and impression cytology (2). The control group included patients without any ocular or systemic morbidities and a normal ocular examination. The stage of LSCD was classified as mild (2–4 points), moderate (5–7 points), or severe (8–10 points) based on a clinical scoring system previously published (18). IVCM volume scans of the central cornea were obtained from 133 patients clinically presenting with LSCD and 54 healthy controls. Of the 187 volume scans, 62 were Mild (Class I), 55 were Moderate (Class II), and 16 were Severe (Class III). Figure 1 shows an example case of each severity. In total, 641 individual scans were obtained. IVCM scans were performed using HRT III (Heidelberg Engineering GmBH, Germany) with the Rostock cornea module at the Stein Eye Institute, University of California, Los Angeles. A minimum of three high-quality Z-scans were acquired in the central cornea from the superficial epithelium down to the anterior stroma (40 scans of 400 µm × 400 µm, one every 2 microns, representing 8-bit grayscale 384 × 384 pixels). For each eye, up to 4 individual scans from the volume were identified by a senior cornea specialist (CB) as clinically relevant for nerve morphology identification and quantified by two trained readers for disease severity, fiber number, fiber length, branch number, and nerve tortuosity. The average of the provided labels by the two readers was used as ground truth labels for the quantification task.

Image preprocessing

To enhance the visibility of the structures of interest, we applied a combination of Top Hat and Black Hat transformations. The Top Hat transform is designed to find bright objects on a dark background, while the Black Hat transform does the opposite. By subtracting the Black Hat transform from the Top Hat transform and adding it back to the original image, we obtain an enhanced image that highlights the neurons in each image. The resulting transformed image provides improved contrast for quantification of biomarkers (Figure 2).

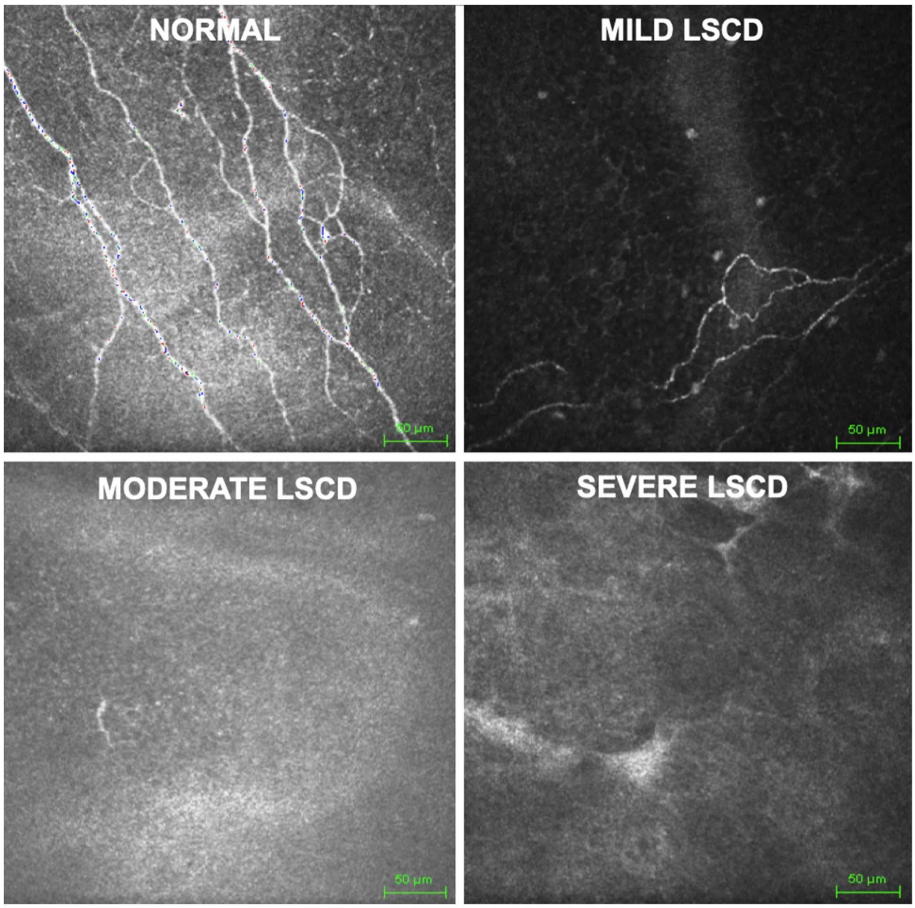

FIGURE 1
Grayscale *in vivo* confocal microscopy images of central cornea used as data source. Classification (control, mild, moderate, severe) based on presence, density, branching, and tortuosity of visible nerves.

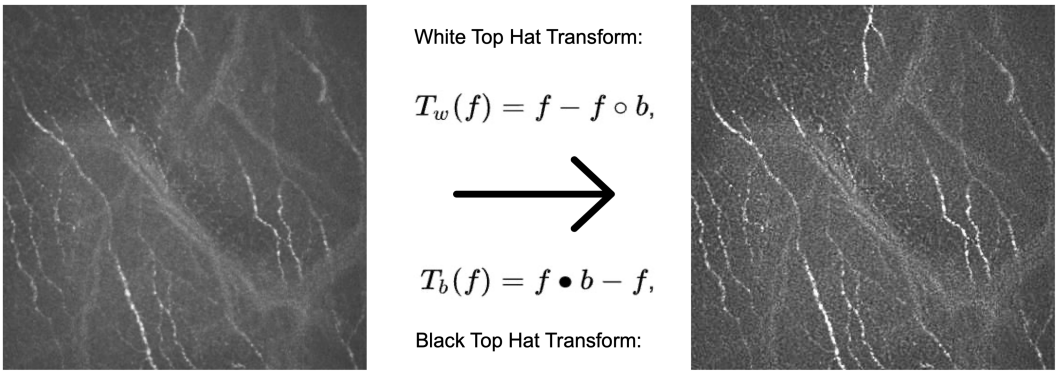

FIGURE 2
Example of an original image (left) subject to the Top Hat and Black Hat transformation, resulting in a contrast enhanced image (right) with highlighted neurons compared to the original.

Stable diffusion

Training set images were passed through Stable diffusion (SD), a model used to create additional artificial images by extending patterns. The model's architecture, comprising a latent diffusion model (LDM) developed by the CompVis Group at Ludwig Maximilian University of Munich, consisted of two main components: (1) a variational autoencoder (VAE), a machine learning model that can infer and create new data relationships within images, and (2) a U-Net, a CNN designed for biomedical image segmentation by analyzing the image

at different scales (15). Each image from our original dataset underwent 16 augmentation runs, with varying strengths of Gaussian noise infusion ranging from 0.2 to 0.28 (Figure 3). An example of the different strength hyper-parameter outputs can be seen in Figure 4.

At each iteration, the latent diffusion model generated new images by iteratively denoising random noise, guided by the CLIP text encoder pretrained on relevant concepts (19). This process allowed the generation of diverse representations of images while maintaining their original disease severity. The augmented images created through SD were then interlayed with our original training dataset before performing normalization, transformation, and finally, model training.

Data Split and normalization

Image arrays and their corresponding metric values were read in after the image data was split into training and validation in a proportion of 70/30 by patient for model development. Separate lists were created to store the image features and the corresponding target variables for severity, fiber number, branch, fiber length, and tortuosity. These target variables represented the labels and metrics used for the subsequent analysis. Before feeding the image data into the model, normalization was applied to ensure consistent and standardized input. The image features were normalized to the range of [0, 1]. This normalization step ensured that all images had consistent intensity ranges and facilitated the convergence and stability of the model during training. Images were lastly feature scaled to ensure all input features have a similar scale or range.

Model architecture and training

Machine learning models are often constructed on an existing model architecture, pre-trained on an existing dataset, and applied to a domain-specific task. A model architecture consists of a stack of layers that take an input image, perform a series of transformations on the image, and output a prediction based on the features learned from the image. The model architecture designed and tested in this study is based on InceptionResNetV2, a model pre-trained on the ImageNet dataset (20, 21), which provides a solid foundation for image feature extraction (Figure 5). A CNN is a machine learning model that feeds images through a series of convolutions to understand features within images. The InceptionResNetV2 model is a state-of-the-art CNN with residual connections that feed information to later layers of the model to aid with model training. To adapt the InceptionResNetV2 for our specific task, we appended it with additional layers, creating a Residual U-Net (ResUNet) architecture. The model was trained to predict multiple nerve morphology metrics and was composed of several key elements. First, residual blocks (a sequence of layers that take the output of a layer and add it to another layer) are integrated into the architecture, featuring skip connections that facilitate gradient flow and promote the effective extraction of both low-level and high-level features from the input images. These residual connections allow the model to bypass certain layers during training, enabling efficient network propagation and reducing the vanishing gradient problem, which is an issue where learned features are lost during model training.

Encoders and decoders are key components of a machine learning model architecture that break down an image to learn features and subsequently output these learned features to make a prediction. The model contains a decoder architecture which plays a critical role in reconstructing feature maps to their original spatial dimensions, allowing for image information to be more easily analyzed. This reconstruction is achieved through the incorporation of upsampling layers, which facilitate the restoration of high-resolution spatial details lost during the downsampling process in the encoder portion of the model. The upsampling layers work by replicating existing feature values to upscale the feature maps, effectively enlarging them to match their original dimensions. The decoder is composed of residual blocks, which leverage skip connections (connections that skip layers to deliver information to layers further within the architecture) to preserve essential feature information while upscaling the feature maps. Each residual block consists of two convolutional blocks with activation functions and batch normalization, which act as information processing layers. The skip connections within the residual blocks allow the decoder to directly access the original input features, facilitating the propagation of gradients and preventing the degradation of feature information during upsampling.

To enable the model to predict multiple metrics, we incorporated separate output branches for each target metric, namely severity, fiber number, branch characteristics, fiber length, and tortuosity. These

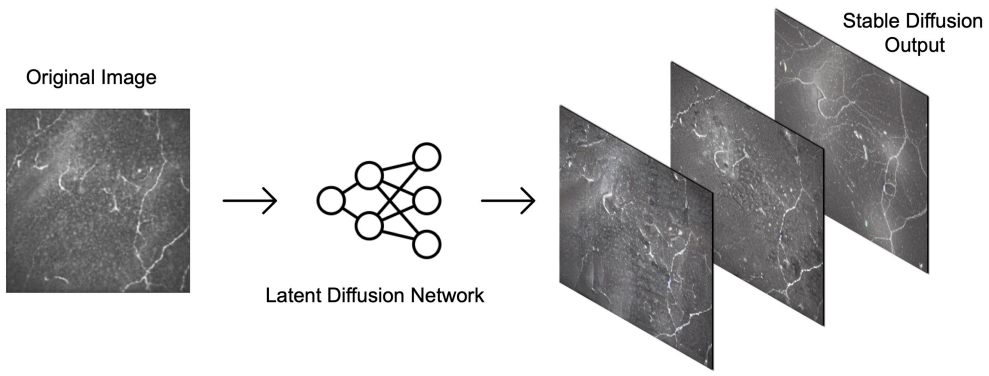

FIGURE 3
Examples of output possibilities from using stable diffusion on the confocal microscopy dataset. The latent diffusion network can be applied at various strengths to generate varying degrees of similarity between output images and the original image.

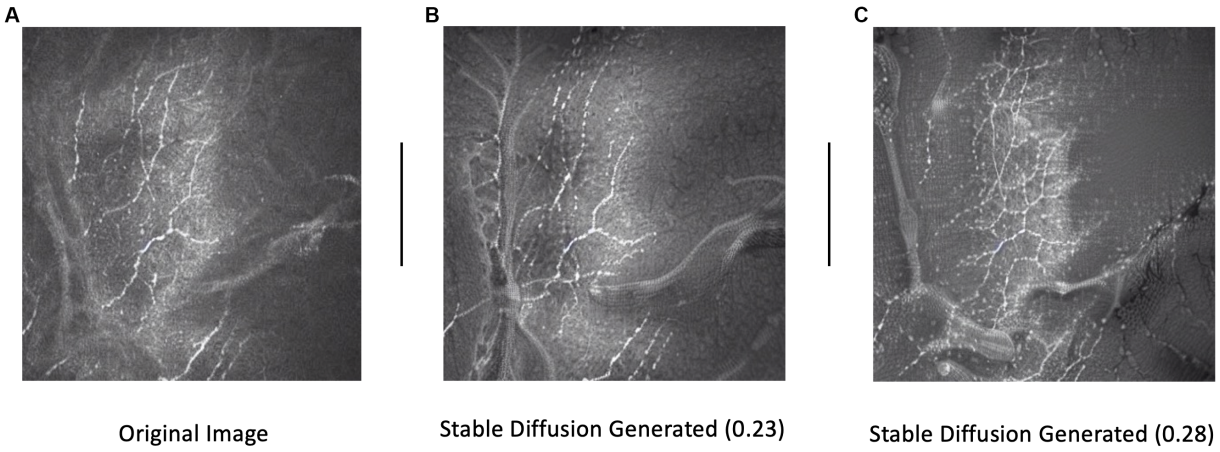

FIGURE 4
Example of an original image **(A)** and subsequent stable diffusion generated images at 0.23 **(B)** and 0.28 **(C)** strengths. Strength dictates the level of Gaussian noise infused into the original image.

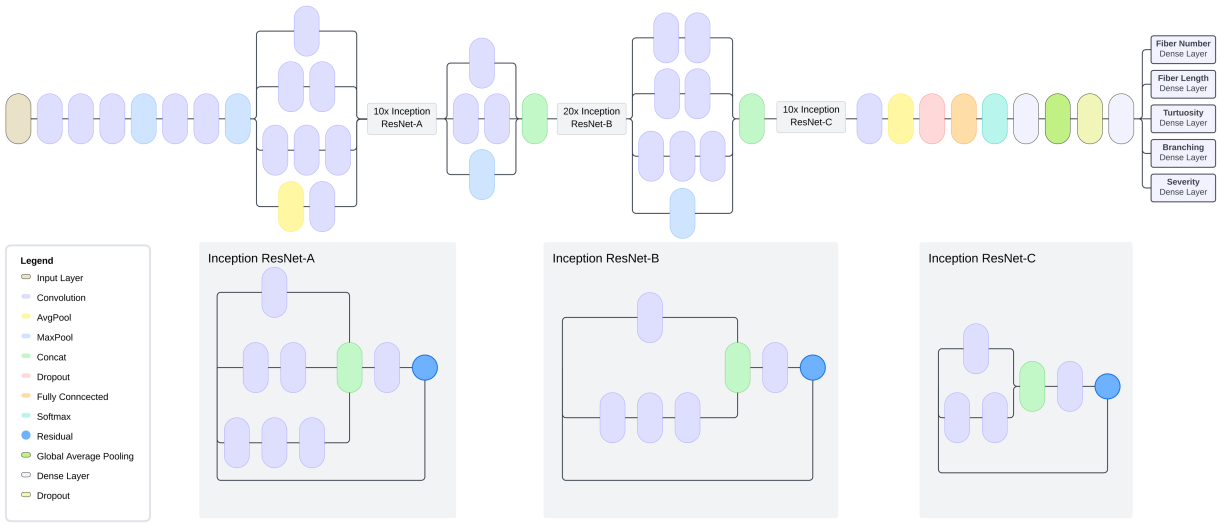

FIGURE 5
Model Architecture: Multi-task model based on InceptionResNetV2 transfer learning and U-Net architecture. Model is composed of a global average pooling layer followed by two fully connected layers. The model has a total of 54,764,136 parameters, of which 427,400 were trainable.

output branches consist of convolutional layers with exponential linear unit (ELU) activation functions and L2 regularization. Activation functions allow predictions by applying a mathematical transformation to a model's predicted output. L2 regularization is applied to calibrate the model during training by adjusting loss term values. To further optimize model training, we compiled it with the Adam optimizer using a learning rate of 0.001. Our choice of mean squared error (MSE) loss functions tailored for each output branch allowed us to effectively quantify the discrepancy between the predicted values and the ground truth. Additionally, we included mean absolute error (MAE) and root mean squared error (RMSE) as evaluation metrics to comprehensively assess the model's performance during training and validation.

During model training there is a possibility for a model to overfit, where the model cannot generalize to data outside of the training set, leading to poor predictions. Augmentations of the images in the training set can be performed to help counter a model from overfitting to the training set. To enhance the model's generalizability and prevent overfitting, we applied rigorous data augmentation techniques aside from and in combination with SD-generated images during training. We augmented the training images with rotations, translations, shearing, zooming, and horizontal flipping. Furthermore, we implemented a learning rate scheduler during training to dynamically adjust the learning rate, which is a mathematical value that determines the rate at which a model learns and processes image features, based on the validation loss. The scheduler reduced the learning rate by a factor of 0.1 if the validation loss did not improve after a certain number of epochs (the number of times that a machine learning model is trained), thus facilitating the model's exploration of different areas in the loss landscape and preventing it from getting

stuck in local minima. The final model was trained for 10 epochs, determined empirically on the training set, and tested on our held-out test set.

Evaluation

The model's performance was evaluated using root mean squared error (RMSE) and coefficient of determination (R^2) as metrics for predicting fiber number, branch, fiber length, and tortuosity parameters. RMSE was employed to assess the accuracy of the model by quantifying the average difference between predicted values and ground truth from the validation dataset. R^2 reported the proportion of variance in the prediction for each metric that was explained by the model. To evaluate the model's predictive performance for severity, Area Under the Receiver Operating Characteristic Curve (AUC) was used to measure the ability of the model to distinguish between true positives and true negatives, and F1 score was used as a metric of the balance between performance and recall.

Results

The results of the predictive models for various metrics related to limbal stem cell deficiency are summarized in Table 1.

When training the ResUNet on a combination of SD-augmented images and original images, the predictive model exhibited a notable performance increase across all morphology metrics aside from tortuosity when compared to the model trained solely on original images. For fiber number prediction, the model achieved an RMSE of 4.44 and an R^2 of 0.63. Similarly, for branch prediction, the model obtained an RMSE of 4.54 with an R^2 of 0.63. Most notably, the model demonstrated high accuracy in predicting fiber length, with an RMSE of 5.56 and a high R^2 of 0.80. However, in the case of tortuosity prediction, the model's performance showed room for improvement, yielding an RMSE of 11.33 and a relatively low R^2 of 0.03. Distribution of predictions and ground truth values can be found in Figure 6.

In contrast, when training the model on only original images as input, the model's predictive capabilities were noticeably lower compared to the SD-augmented model. The RMSE for fiber number was 5.04, and the corresponding R^2 was 0.53, indicating a moderate level of accuracy. For branch prediction, the model obtained an RMSE of 5.13 and an R^2 of 0.52. For fiber length estimation, the model achieved an RMSE of 6.3 and an R^2 of 0.74. Finally, regarding tortuosity prediction, the model showed an RMSE of 4.31 and an R^2 of 0.13.

Disease severity was assessed individually in a single metric variation of the residual U-net model. This single task version of the model was also trained on the SD and non-SD supplemented data sets. Disease severity was stratified and assessed as disease vs. no disease (control), control/mild vs. moderate/severe, and severe vs. non-severe. Performance was compared against a classifier based on the morphological features that were assessed manually during routine clinical practice. A summary of disease severity statistics can be found in Table 2.

In all three scenarios, the ResUnet model with SD outperformed classification using manual morphological metrics in terms of F1 (0.839 vs. 0.805, 0.743 vs. 0.703, and 0.255 vs. 0.222, respectively). The accuracy using the ResUnet with SD produced higher accuracy than the manual metrics for classifying the controls and severe cases (0.789 vs. 0.758 and 0.577 vs. 0.531, respectively), and they had the same accuracy when classifying between mild and moderate cases (0.778). The Area Under the Receiver Operator Characteristic Curve (AUC) for the ResUnet with SD was higher in the moderate and severe cases (0.810 vs. 0.803 and 0.765 vs. 0.733, respectively), but slightly lower in the control classification (0.855 vs. 0.857; Figure 7).

When comparing the classification with and without SD in the classifier, the model with SD again had higher accuracy (0.789 vs. 0.747, 0.778 vs. 0.768, and 0.577 vs. 0.515) and F1 (0.839 vs. 0.797, 0.743 vs. 0.734, and 0.255 vs. 0.230). The model without SD had higher AUC values for the control (0.855 vs. 0.867) and mild classifications (0.810 vs. 0.816), while the SD model had the highest AUC when classifying the severe cases (0.765 vs. 0.746).

Discussion

This study presents a novel approach to enhancing the performance of multi-task nerve morphology prediction by incorporating images generated through latent diffusion models as a form of training set bolstering. Specifically, SD-generated images were introduced as an augmentation technique to complement the original dataset in training the residual U-Net architecture. The results demonstrate the potential of this approach, as the inclusion of SD-generated images led to a notable improvement in the model's nerve morphology prediction performance when compared to the same model trained solely on original images. The introduction of SD-generated images allowed the residual U-net model we created to leverage additional synthetic samples, resulting in enhanced generalization and predictive capabilities. It is important however to acknowledge the inherent limitations of training machine learning models on generated data.

TABLE 1 Multi-task neuro-morphology results: root mean squared error and R-squared values for stable diffusion trained model and model trained on solely original images.

Models	Fiber number RMSE	Fiber number R^2	Branch RMSE	Branch R^2	Fiber length RMSE	Fiber length R^2	Tortuosity RMSE	Tortuosity R^2
Stable diffusion + Original images	**4.44**	**0.63**	**4.54**	**0.63**	**5.56**	**0.80**	11.33	0.03
Original images only	5.04	0.53	5.13	0.52	6.3	0.74	**4.31**	**0.13**

Bold values represent the model with the best performance for each metric.

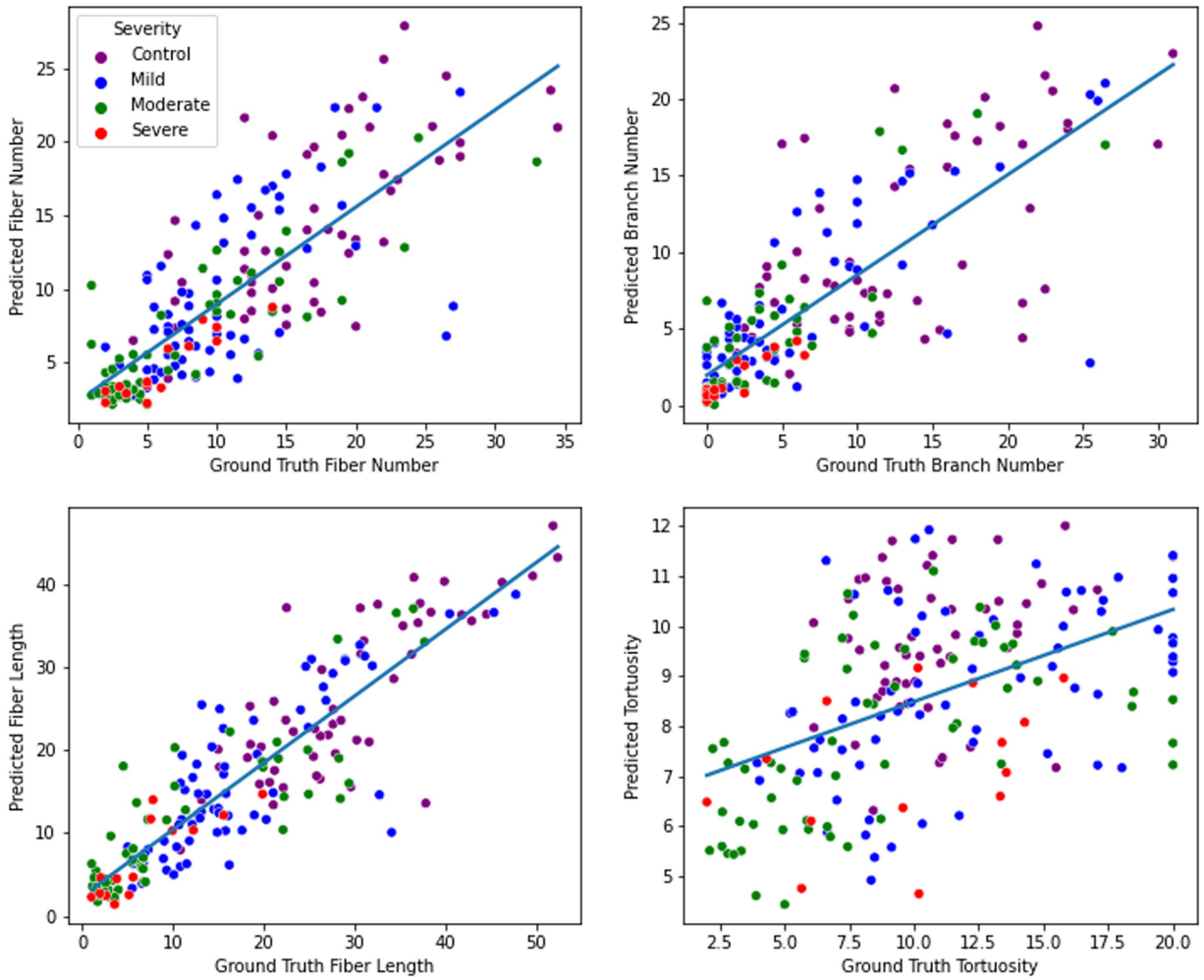

FIGURE 6
Nerve feature prediction performance. From top left to bottom right: fiber number, branch number, fiber length, and tortuosity. For each metric, a regression line was fitted to illustrate the direction and magnitude of the correlation between ground truth and predicted values.

TABLE 2 Severity comparisons and corresponding AUC, Precision, Recall, F1, and accuracy values with and without stable diffusion (SD) images and using manually assessed metrics only.

		AUC	Precision	Recall	F1	Accuracy
Control vs. mild, moderate, severe	Manual	0.857	**0.980**	0.683	0.805	0.758
	SD	0.855	0.947	**0.754**	**0.839**	**0.789**
	No SD	**0.867**	0.970	0.676	0.797	0.747
Control, mild vs. moderate, severe	Manual	0.803	0.718	0.689	0.703	**0.778**
	SD	0.810	**0.838**	0.667	**0.743**	**0.778**
	No SD	**0.816**	0.653	**0.838**	0.734	0.768
Control, mild, moderate vs. severe	Manual	0.733	0.126	0.929	0.222	0.531
	SD	**0.765**	**0.146**	**1.00**	**0.255**	**0.577**
	No SD	0.746	0.130	**1.00**	0.230	0.515

Bold values represent the model with the best performance for each metric in each severity classification.

Previous work on using nerve features for severity prediction has theorized that morphological changes to nerves become more pronounced as disease severity increases, which is supported by the results shown (22). The high AUC of 0.855 supports the claim that nerve morphology can be used as an effective predictor of LSCD severity. In a larger pipeline, the disease severity prediction by nerve morphology can be supplemented by other predictors such as cell morphology and basal cell density. While in the context of this study

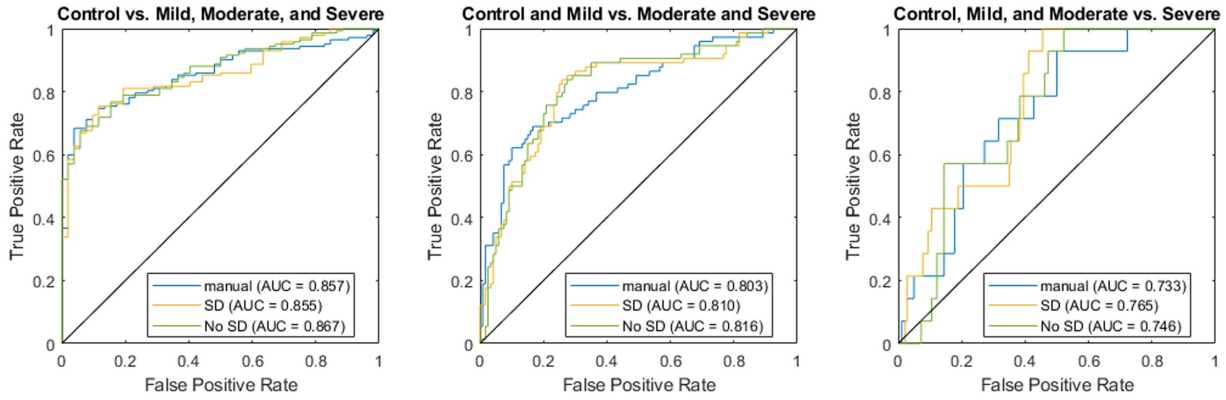

FIGURE 7
Receiver operator characteristic curves for regression using manually derived features (blue), deep learning using stable diffusion (SD; orange), and deep learning without SD (green).

nerve density was a sufficient predictor, some confocal microscopy images contain less pronounced nerve morphology. Thus, the predictions shown would likely be most beneficial as part of a larger pipeline that examines multiple features of corneal images for disease severity prediction.

Overall, the inclusion of SD-augmented images in the ResUnet model led to substantial improvements in predicting fiber number, branch, and fiber length metrics associated with limbal stem cell deficiency. The enhanced predictive accuracy demonstrated the model's capacity to better capture intricate spatial features and relationships, resulting in more reliable and robust estimations. The models had more difficulty learning tortuosity; this difficulty is understandable as the clinical evaluation can be highly variable in images with few nerves present (i.e., severe cases). While certain challenges remain in predicting tortuosity accurately, the promising outcomes highlight the potential of the proposed approach for comprehensive and precise assessments of limbal stem cell deficiency metrics.

The inclusion of SD training images was less helpful in classifying severity than individual morphological features. While the SD model had consistently higher accuracy and F1, the differences were relatively small and largely resulted from a different classification threshold as demonstrated by the similar ROC curves (Figure 7). The highest difference between the two was in classifying severe cases, which had the fewest training examples, which could indicate that SD helped overcome this lack of data. The overall similarity in predictions could be a result of the classification using nerve images alone whereas clinical diagnosis utilizes multiple other data sources and modalities. The utility of these models is demonstrated in the fact that the deep learning model outperforms a classifier trained only on the metrics manually assessed from these images through standard clinical practice. Future directions can extend these methods to incorporate these other data sources to create a more holistic classifier that better represents the full spectrum of information available to a clinician.

SD images may be beneficial for training the model in terms of increasing the dataset size, although it potentially may limit the features that can be learned. SD inherently alters the structure of an image, although the degree to which this affects the model's ability to learn features is unknown. The addition of SD images in the training dataset may improve model predictions as the feature identification task becomes more difficult. Conversely, the inclusion of SD images may limit the model performance by forcing the model to learn extraneous features that were created as a result of including the SD-generated images.

Despite the limitations of SD, our findings highlight the potential benefits of incorporating latent diffusion-generated images as a means of data augmentation in the context of nerve morphology regression. The approach not only offers an avenue to expand the training dataset but also provides an opportunity to explore diverse representations of the same images while preserving their semantic meaning. Further investigations into mitigating overfitting and optimizing the augmentation process are warranted to unlock the full potential of this novel technique. The combination of real and generated data may lead to more robust and accurate predictive models when faced with limited training data. This can serve to facilitate advancements in the diagnosis and treatment of nerve-related pathologies and beyond.

Future directions

In the future, additional processing methods such as bootstrapping, creating larger SD image sets, and modifying the noise parameters and weights in our latent diffusion model can be performed to improve results and further validate efficacy of the model. Additionally, it is possible that employing SD images changes the patterns or number of neurons in an image, which could make it more representative of a different disease state from the original image. Thus, including a step that segments the neurons in both the raw and SD images, compares the pixel volume of each, and assigns a weight to the SD image based on the difference when compared to its unaltered counterpart would help to correct this issue. Alternatively, once a model is trained to be proficiently accurate on raw data, it could be used to assign predicted metric values on SD images. Further research is required to validate the use case for including SD images in training data as a method of data augmentation.

Conclusion

This study demonstrates the effectiveness of incorporating SD-generated images as an augmentation technique to enhance the performance of a multi-task nerve morphology prediction model for limbal stem cell deficiency (LSCD). The inclusion of SD images significantly improved the model's predictive capabilities for fiber number, branch, and fiber length metrics associated with LSCD, showcasing its potential for precise assessments of this disorder. However, challenges remain in accurately predicting tortuosity and disease severity, warranting further investigation. While this approach shows promise in leveraging synthetic data to bolster training sets, careful consideration of overfitting and model convergence is essential. By refining the preprocessing methods and exploring additional augmentation techniques, this novel approach may lead to more robust and accurate predictive models for nerve morphology analysis and disease severity prediction, potentially improving clinical workflows in the future.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by UCLA Institutional Review Board. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

DG: Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. TT: Investigation, Software, Writing – original draft, Writing – review & editing. VR: Investigation, Software, Writing – original draft, Writing – review & editing. CB: Data curation, Writing – review & editing. NS: Formal Analysis, Writing – review & editing. MV: Formal Analysis, Writing – review & editing. TS-B: Data curation, Writing – review & editing. CA: Methodology, Resources, Supervision, Writing – review & editing. SD: Conceptualization, Funding acquisition, Resources, Writing – review & editing. WS: Conceptualization, Data curation, Formal Analysis, Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work is supported in part by an unrestricted grant from Research to Prevent Blindness to the Department of Ophthalmology at the University of California, Los Angeles. SD received grant support from the National Eye Institute (R01 EY021797 and R01 EY028557) and from the California Institute for Regenerative Medicine (TR2-01768, CLIN1-08686, and CLIN2-11650) to study limbal stem cells.

Conflict of interest

SD is a consultant for Novartis US, Amgen, Cellusion, Kala Pharmaceuticals, and Claris Biotherapeutics, Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Tseng SCG. Concept and application of limbal stem cells. *Eye*. (1989) 3:141–57. doi: 10.1038/eye.1989.22

2. Deng SX, Borderie V, Chan CC, Dana R, Figueiredo FC, Gomes JAP, et al. Global consensus on definition, classification, diagnosis, and staging of limbal stem cell deficiency. *Cornea*. (2019) 38:364–75. doi: 10.1097/ICO.0000000000001820

3. Deng SX, Sejpal KD, Tang Q, Aldave AJ, Lee OL, Yu F. Characterization of Limbal stem cell deficiency by in vivo laser scanning confocal microscopy: a microstructural approach. *Arch Ophthalmol*. (2012) 130:440–5. doi: 10.1001/archophthalmol.2011.378

4. Deng SX, Kruse F, Gomes JAP, Chan CC, Daya S, Dana R, et al. Global consensus on the management of limbal stem cell deficiency. *Cornea*. (2020) 39:1291–302. doi: 10.1097/ICO.0000000000002358

5. Chan E, Le Q, Codriansky A, Hong J, Xu J, Deng SX. Existence of normal limbal epithelium in eyes with clinical signs of total limbal stem cell deficiency. *Cornea*. (2016) 35:1483–7. doi: 10.1097/ICO.0000000000000914

6. Le Q, Samson CM, Deng SX. A case of corneal neovascularization misdiagnosed as total limbal stem cell deficiency. *Cornea*. (2018) 37:1067–70. doi: 10.1097/ICO.0000000000001631

7. Le Q, Chauhan T, Cordova D, Tseng CH, Deng SX. Biomarkers of *in vivo* limbal stem cell function. *Ocul Surf*. (2022) 23:123–30. doi: 10.1016/j.jtos.2021.12.005

8. Chuephanich P, Supiyaphun C, Aravena C, Bozkurt TK, Yu F, Deng SX. Characterization of the corneal subbasal nerve plexus in limbal stem cell deficiency. *Cornea*. (2017) 36:347–52. doi: 10.1097/ICO.0000000000001092

9. Bhattacharya P, Edwards K, Harkin D, Schmid KL. Central corneal basal cell density and nerve parameters in ocular surface disease and limbal stem cell deficiency: a review and meta-analysis. *Br J Ophthalmol*. (2020) 104:1633–9. doi: 10.1136/bjophthalmol-2019-315231

10. Vasques X, Vanel L, Villette G, Cif L. Morphological neuron classification using machine learning. *Front Neuroanat*. (2016) 10:102. doi: 10.3389/fnana.2016.00102/full

11. Zeng H, Sanes JR. Neuronal cell-type classification: challenges, opportunities and the path forward. *Nat Rev Neurosci*. (2017) 18:530–46. doi: 10.1038/nrn.2017.85

12. Scarpa F, Colonna A, Ruggeri A. Multiple-image deep learning analysis for neuropathy detection in corneal nerve images. *Cornea*. (2020) 39:342–7. doi: 10.1097/ICO.0000000000002181

13. Preston FG, Meng Y, Burgess J, Ferdousi M, Azmi S, Petropoulos IN, et al. Artificial intelligence utilising corneal confocal microscopy for the diagnosis of peripheral neuropathy in diabetes mellitus and prediabetes. *Diabetologia*. (2022) 65:457–66. doi: 10.1007/s00125-021-05617-x

14. Lv J, Zhang K, Chen Q, Chen Q, Huang W, Cui L, et al. Deep learning-based automated diagnosis of fungal keratitis with in vivo confocal microscopy images. *Ann Transl Med*. (2020) 8:706. doi: 10.21037/atm.2020.03.134

15. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. (2022). High-resolution image synthesis with latent diffusion models. arXiv; Available at: http://arxiv.org/abs/2112.10752

16. Trabucco B, Doherty K, Gurinas M, Salakhutdinov R. (2023). Effective data augmentation with diffusion models [internet]. arXiv; Available at: http://arxiv.org/abs/2302.07944

17. Khader F, Mueller-Franzes G, Arasteh ST, Han T, Haarburger C, Schulze-Hagen M, et al. (2023). Medical diffusion: Denoising diffusion probabilistic models for 3D Medical Image Generation [Internet]. arXiv; 2023. Available at: http://arxiv.org/abs/2211.03364

18. Aravena C, Bozkurt K, Chuephanich P, Supiyaphun C, Yu F, Deng SX. Classification of Limbal stem cell deficiency using clinical and confocal grading. *Cornea*. (2019) 38:1–7. doi: 10.1097/ICO.0000000000001799

19. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. (2021). Learning transferable visual models from natural language supervision [internet]. arXiv; Available at: http://arxiv.org/abs/2103.00020

20. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. (2016). Inception-v4, inception-ResNet and the impact of residual connections on learning [internet]. arXiv; Available at: http://arxiv.org/abs/1602.07261

21. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. (2009). "ImageNet: a large-scale hierarchical image database." in *2009 IEEE conference on computer vision and pattern recognition*. p. 248–255.

22. Liu P, Panchavati S, Pleasure M, Siu N, Bonnet C, Deng S, et al. (2022). "Mobile net V2 based diagnosis and grading of limbal stem cell deficiency." in *2022 IEEE 22nd international conference on bioinformatics and bioengineering (BIBE)*. p. 174–179.

Frontiers in Medicine

Translating medical research and innovation into improved patient care

A multidisciplinary journal which advances our medical knowledge. It supports the translation of scientific advances into new therapies and diagnostic tools that will improve patient care.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact