参赛队员姓名：谢昕然

中学：中国人民大学附属中学____________

省份：北京市__________________________

国家/地区：中国_______________________

指导教师姓名:徐恪，施一宁______________

指导教师单位：清华大学，中国人民大学附属中学________________________

论文题目：AI-based Glaucoma Diagnoses Based on Phone-taken Colored Fundus Retinal Images

AI-based Glaucoma Diagnoses Based on Phone-taken Colored Fundus Retinal Images

Xinran Xie[1]
[1] High School Affiliated to Renmin University of China, Beijing 100086, China
E-mail: selinaxie2007@163.com

***Abstract*—Early detection and diagnosis of glaucoma are crucial to prevent irreversible eye damage. This paper introduces an online, AI-based APP for early glaucoma diagnoses, improving the conveniency and effectiveness of glaucoma prevention. It expands diagnostic settings by enabling users to upload their phone-taken colored retinal fundus images on Cloud and receive AI-based diagnostic results directly. This eliminates the two major obstacles in current glaucoma diagnosis: high equipment costs and the necessity of ophthalmologists. The APP's realization involves two image preprocessing modules and a prediction module. The preprocessing phase, encompassing the rectification and noise-removal modules, innovatively employs matrix algorithms, including a perspective transform for rectification and the FTDTV model for denoising, in which intelligent matrix operations ensure superior computational efficiency. To guarantee accurate and robust glaucoma diagnosis, the prediction module features innovative components. These include a Polar Transform layer and MobileNet 1.0 for image focus and feature extraction, an Attention Module for handling imbalanced tags, an Overfitting Prevention strategy, and Diversity Learning to enhance model robustness against unpredictable image capture processes. When applied to real colored retinal fundus image datasets, the prototype application showed promising results. The ACC and WKappa values reach 0.9301 and 0.9221, respectively, when testing the effectiveness of this paper's proposed system in real-world glaucoma image datasets, demonstrating a high potential for wide-scale, real-world application in early glaucoma detection and diagnosis.**

***Index Terms*—Glaucoma diagnosis, perspective transform, FTDTV.**

I. Introduction

Glaucoma, positioned as the second leading cause of blindness globally by the World Health Organization (WHO), inflicts progressive damage to the optic nerve, usually linked with elevated intraocular pressure and visual field losses [1], [2]. If left undiagnosed or untreated, the damage or resultant blindness is irreversible [3]. As of 2020, an alarming 5.9 million out of 79.6 million patients worldwide suffer from irreversible bilateral blindness due to glaucoma [4]. Despite the critical need for early glaucoma detection, several challenges exist, particularly the latency of symptoms, which often leads to delayed diagnosis and treatment. Open-angle glaucoma, the most prevalent type, is characteristically silent in its early stages, with patients typically unaware of the initial symptoms including painless intraocular pressure increase and peripheral vision losses [5].

Globally, approximately 6.9 million glaucoma patients suffer from preventable visual impairments due to delayed diagnosis and treatment [1]. This issue is especially prevalent in developing countries, where technological and socioeconomic barriers often lead to delayed diagnoses and consequential visual impairments [6], [7]. For instance, in Egypt, delayed diagnosis accounts for 43.03% of glaucoma cases [8], while in China, out of 9.4 million glaucoma patients, 5.2 million (55%) are blind in at least one eye, and 1.7 million (18.1%) suffer from bilateral blindness [9]. Given the irreversible damage caused by glaucoma and the current diagnosis delays, accurate and early detection of glaucoma is of paramount importance.

Currently, the primary methods for early glaucoma detection include optic coherence tomography (OCT) diagnoses, visual field (VF) tests, and colored retinal fundus image (CRFI) diagnoses. Despite their prevalence, these methods present significant limitations. On the one hand, high equipment costs of OCT diagnoses and VF tests restrict access for smaller hospitals and financially limited patients. On the other hand, CRFI diagnoses, despite being less expensive and more widely available, rely on skilled medical professionals' diagnoses and other technologies' assistance, which are often inaccessible to individuals in remote regions with limited medical resources.

Based on existing literature, there is a relatively limited number of studies exploring alternative methods for early glaucoma detection beyond direct analyses of OCT, VF images, and colored retinal fundus images. This is particularly true for highly applicable methods that utilize user-friendly platforms, such as online apps for the general public.

This paper proposes a mobile app designed to facilitate early glaucoma diagnosis while alleviating limitations inherent in currently prevalent methods. The app enables users to photograph their paper-version colored fundus images using their smartphones and upload these photos for analysis by cloud-based machine learning algorithms. This process eliminates the need for professional medical workers, making the app an ideal tool for glaucoma early detection. The app accomplishes four key objectives: economic feasibility, convenience, prevalence, and accuracy.

- **Economic Feasibility and Convenience:** By employing AI algorithms to analyze inexpensive colored fundus images, the app offers an economical solution that doesn't require direct involvement from medical professionals, thus ensuring convenience.
- **Prevalence:** The app's prevalence is facilitated by the increasingly widespread internet infrastructure, which has been equipped in remote rural regions of many developing countries. In 2022, the number of Chinese internet users reached 1.05 billion, with an internet penetration rate of 74.4%, according to the China Internet Network Information Center (CNNIC) [10].
- **Accuracy:** The app's intelligent machine learning algorithms are robust and can deliver reliable results even when

analyzing low-quality phone-taken photos.

However, the development of this app comes with its own set of challenges. The first challenge is ensuring the correct positioning of images and managing the noise associated with them. Second, the subtlety of early glaucoma signs can lead to difficulties in accurately classifying samples. Lastly, the challenges are compounded by the existence of highly unbalanced sample tags and the unpredictable nature of the image capturing process.

To address these problems, the app first rectifies images using a perspective transform. After the rectification, a noise-removal module efficiently denoises the uploaded image to improve its overall quality. With these two steps, the quality of the images themselves is significantly enhanced. Furthermore, a prediction module is designed with AI-powered components, improving the accuracy and efficiency of the analysis. The main technique contributions are listed as follows.

- To ensure images are optimally rectified for AI analysis and diagnosis, this paper implements a **perspective transform algorithm**. This efficient method rectifies images via straightforward matrix operations, thus creating a reliable basis for subsequent AI analysis and diagnosis.
- To effectively and efficiently remove noises from images, this paper proposes a method to **recover noise-free images using low-rank tensor recovery**. Specifically, the objective images' low-rank features are demonstrated through SVD operations, and this paper utilizes the FTDTV model [11] in conjunction with the Alternating Direction Method of Multipliers (ADMM) algorithm. The FTDTV model, with its low-rank factor prior, reduces computational burdens and eliminates the need to predetermine the Tucker rank. By leveraging the strengths of the FTDTV model, the corresponding noise-free image can be extracted from the observed image, effectively isolating the noise.
- To ensure **accurate and robust glaucoma diagnosis** despite the challenges of imbalanced tags and unpredictable image capture processes, this paper presents a developed **prediction module**. The module features innovative components, including a Polar Transform layer and MobileNet 1.0 for image feature extraction, an Attention Module to handle imbalanced tags, an Overfitting Prevention strategy, and Diversity Learning to fortify model robustness against unpredictable image capture processes. Integrating these strategies results in a highly effective tool for glaucoma diagnoses.
- Extensive experiments are conducted on four datasets containing diversified colored fundus retinal images, i.e. G1020, ORIGA, LAG-dataset, Real dataset, to verify the effectiveness of the proposed glaucoma diagnostic system. Ablation studies show that the two image preprocessing modules, the rectification module and the noise-removal module, and the optimizations, including polar transform, the overfitting prevention module, and the diversity learning, in the prediction module significantly improve the diagnostic accuracy and generalization ability. Experiments show that this paper's proposed online glaucoma diagnostic system achieves satisfactory diagnostic results, reaching the ACC and WKappa of 0.9301 and 0.9221, respectively, outperforming some of the existing diagnostic methods: CABNet, ResMLP, and UQ.

II. Related work

For early glaucoma detection, there are currently three mainstream approaches: OCT diagnoses, VF tests, and CRFI diagnoses. This section analyzes and discusses these three approaches.

A. OCT diagnoses

In recent years, OCT diagnosis has become one of the most common glaucoma tests. It is a non-contact, non-invasive diagnostic tool that provides cross-sectional imaging of the anterior and posterior eye, using light in an approach similar to computed tomography [12]. The OCT diagnosis offers fine details of each retinal layer and blood vessels, thus providing efficient and quick results [13]. Some studies have combined machine learning models with OCT images for improved diagnoses. For example, [14] uses a 3D Convolutional Neural Network (CNN) to identify diagnostic regions associated with glaucoma classifications. In [15], automated machine classifiers are investigated to better distinguish glaucomatous eyes from non-glaucomatous ones based on OCT images. [16] identifies the Random Forest (RF) model as the best machine learning model for detecting glaucomatous symptoms on OCT images. However, despite the high quality of OCT images and the innovative machine-learning approaches, two significant problems exist. Technologically, artifacts in OCT images often resemble glaucomatous signs, leading to inaccurate detection results [17]. In terms of the economic feasibility, OCT devices are expensive and bulky, making them inaccessible for widespread use especially in remote regions [18].

B. VF tests

The VF test is another mainstream glaucoma test that can detect visual loss in peripheral vision, which is an important early sign of glaucoma [19]. Similar to OCT diagnoses, VF tests provide high-quality images and have been investigated using machine learning techniques. For instance, [20] uses linear regression analysis of the Visual Field Index (VFI) to predict whether uploaded VF images are glaucomatous. In [21], the Recurrent Neural Network (RNN) is used to provide robust labels and predictions of visual loss, aiding the diagnosis of glaucoma. However, VF tests face technological and economic challenges. Technologically, glaucomatous signs in VF tests are often detectable only after retinal nerve fiber layer loss has occurred, leading to delayed diagnosis [22]. Economically, similar to OCT tests, the high cost of VF test devices makes them inaccessible to many patients in remote or financially struggling regions.

C. CRFI diagnoses

CRFI diagnosis is a common eye test that records colored images of the interior surface of the eye using a fundus camera to aid in the diagnosis of eye health [23]. In glaucoma diagnoses, CRFI focuses on the optic cup and disc region and calculates the cup-to-disc ratio, an important glaucomatous sign [24]. Unlike

OCT diagnoses and VF tests, CRFI diagnoses are relatively low-cost and can be found in almost all hospitals and community health care service centers. Machine learning algorithms are also applied to emphasize glaucomatous signs and aid in diagnoses. For example, [25] applies automatic image processing to retinal fundus images, offering pixel-level segmentation of optic cups and discs to calculate the cup-to-disc ratio. [26] employs unsupervised Anomaly Detection (AD) models for detection based on colored fundus images. [25] uses image transforms and Gabor filters to maximize glaucomatous signs and utilizes Artificial Neural Network (ANN) to extract features for later diagnoses. [27] employs the U-Net architecture for glaucomatous sign identification and uses SVM, neural network, and Adaboost classifiers to provide diagnostic aids for doctors. However, since patients normally only have paper-version colored fundus retinal images, a significant drawback of CRFI diagnoses is their dependency on the final diagnoses made by professional medical workers. Further advancements are needed to offer automated diagnoses for people in regions with limited medical resources.

Additionally, some studies devise an AI-based image identification system to return glaucoma diagnostic results to users, based on cheap colored fundus retinal photography technology. For example, [28] proposes GLIM-Net, a diagnostic system consisting of a time positional encoding module and a time-sensitive multi-head self-attention (MSA) module. Thus, GLIM-Net provides the users with the predicted probability that they will develop glaucoma in the future based on users' uploaded colored fundus retinal images in digital forms. Despite the high accuracy of the GLIM-Net prediction system, it doesn't apply in some real-world diagnostic settings patients are facing, since users hardly have access to the digital format of their colored fundus retinal images but only have the images in printed paper format. Given the situation that patients often get their paper-version colored fundus retinal images directly through their community health care service without having the chance to visit ophthalmologists, the GLIM-Net, which doesn't take the phone-taken images of users' paper-version results into account, still can't alleviate patients' diagnostic problems. Namely, patients have difficulties getting glaucoma diagnostic results either from AI diagnoses or professional ophthalmologists' diagnoses.

III. Solution overview

As shown in Fig.1, this app consists of two parts: the client side and the server side. The client side captures fundus images and uploads them to the server side. Once there, the images are processed, features are extracted, and a final diagnostic result (positive or negative for glaucoma) is provided.

As illustrated in Fig.2, the server side is composed of three modules: the **rectification module**, the **noise-removal module**, and the **prediction module**. These modules are effectively integrated to deliver a straightforward diagnostic result to the users.

- In the **rectification module**, the uploaded image undergoes a perspective transform for optimal rectification. This method efficiently rectifies the image using matrix manipulations, without incurring significant computational complexity and costs.
- The **noise-removal module** separates the noise-free image from the observed noisy image, leveraging the low-rankness of images. Specifically, the FTDTV model [11] in conjunction with the Alternating Direction Method of Multipliers (ADMM) algorithm are utilized. This ensures high accuracy while maintaining high computational efficiency.
- The **prediction module** processes denoised images from the noise-removal module to complete the diagnostic process, guaranteeing accurate and robust results. This is achieved by integrating various sophisticated techniques to highlight critical regions of the image, extract features, address imbalanced tags, and handle unpredictable image capture processes. These combined strategies create a highly accurate and robust tool for glaucoma diagnosis.

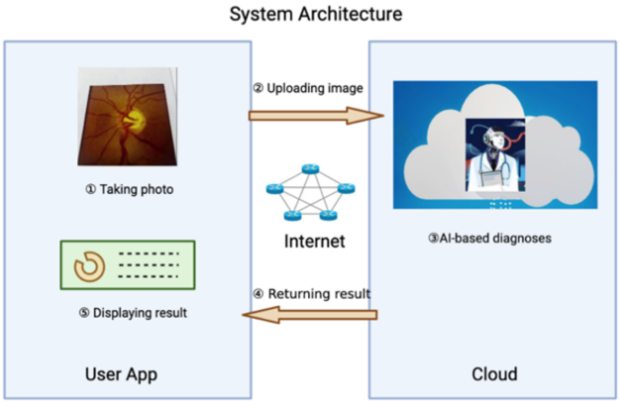


Fig. 1: System architecture.

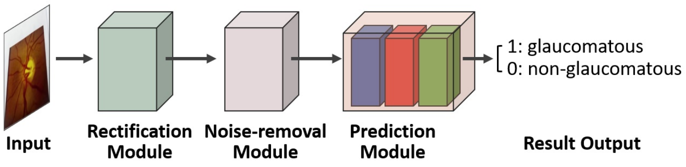


Fig. 2: Server-side structure.

After all processing and diagnostic procedures are completed on the cloud, the diagnostic result is returned to the client side of the app. The subsequent sections of this paper will provide detailed information on these three major modules of the app.

IV. Rectification module

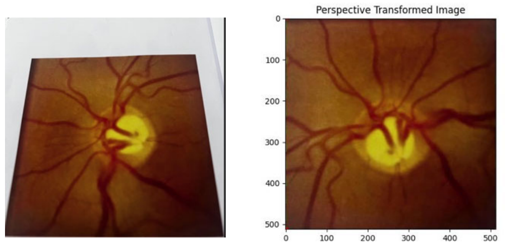
A. Problem

When users upload fundus images taken with their phones, it can be challenging to ensure that the images are properly centered and captured with the phone's camera parallel to the image. This often leads to images being in inappropriate positions for subsequent feature extraction and analysis, thus hindering accurate interpretations and diagnoses. Therefore, it is crucial to have a convenient and effective rectification module that can output an image with the correct position and fixed size, enabling reliable feature extraction and analysis.

The main function of the rectification module is illustrated in Fig. 3. The input image, captured by the phone's camera (Fig. 3(a)), is transformed into the rectified image (Fig. 3(b)).

B. Method

While several neural network-based rectification methods have been proposed recently, they typically involve complex



(a) The input image is not parallel to the phone's camera.

(b) The image after rectification with the fixed size 512×512.

Fig. 3: Rectification module's function.

neural network models and require pairs of images for training the alignments. However, obtaining such image pairs can be difficult. For example, the RANSAC-Flow method [29] produces good rectification results through a two-stage process involving a coarse rectification using RANSAC based on existing features and a fine rectification using a deep network. However, the RANSAC-Flow method requires two images of the same scene from different views to establish a reference.

Instead of requiring pairs of images for reference, this paper adopts the perspective transform method, leveraging its low computational complexity through the use of simple matrix manipulations with no need for referencing pair images.

The perspective transform involves transforming the 2D Cartesian coordinates of the uploaded image into 3D homogeneous coordinates using homography matrix operations. The conversion is performed using the following equations, where $\mathbf{X}$ and $\mathbf{Y}$ represent the horizontal and vertical Cartesian coordinates in the 2D image, respectively, and $\widetilde{\mathcal{U}}, \widetilde{\mathcal{V}}, \widetilde{\mathcal{W}}$ represent the homogeneous coordinates obtained by multiplying the Cartesian coordinates with the 3×3 homography matrix, which includes eight unknown elements denoted as $\mathbf{H}_{ij}$.

$$\begin{pmatrix} \widetilde{\mathcal{U}} \\ \widetilde{\mathcal{V}} \\ \widetilde{\mathcal{W}} \end{pmatrix} = \begin{pmatrix} H_{11} & H_{12} & H_{13} \\ H_{21} & H_{22} & H_{23} \\ H_{31} & H_{32} & 1 \end{pmatrix} \begin{pmatrix} X \\ Y \\ 1 \end{pmatrix} \quad (1)$$

The rectification module can be divided into four main steps:

- **Extraction of Actual Cartesian Coordinates**: The actual Cartesian coordinates ($\mathbf{X}$ and $\mathbf{Y}$) of the four vertices of the uploaded square-shaped image are extracted. These coordinates represent the current position and orientation of the image.
- **Determination of Expected Homogeneous Coordinates**: Based on the desired output size of the image, the expected homogeneous coordinates ($\widetilde{\mathcal{U}}, \widetilde{\mathcal{V}}, \widetilde{\mathcal{W}}$) of the vertices after rectification are determined. These coordinates define the desired position and size of the rectified image.
- **Calculation of Homography Matrix Elements**: The determined homogeneous coordinates are used in matrix operations to calculate the values of the eight unknown elements (H_{ij}) in the homography matrix. The homography matrix represents the transformation needed to rectify the image.
- **Transformation and Conversion**: Using the calculated homography matrix, all points in the original image are transformed to their rectified positions in homogeneous coordinates. This transformation is achieved by multiplying the coordinates with the homography matrix. Finally, the rectified image in homogeneous coordinates is converted back to 2D Cartesian coordinates using the equations:

$$\mathbf{X}' = \frac{\widetilde{\mathcal{U}}}{\widetilde{\mathcal{V}}} \quad \textit{and} \quad \mathbf{Y}' = \frac{\widetilde{\mathcal{V}}}{\widetilde{\mathcal{W}}} \quad (2)$$

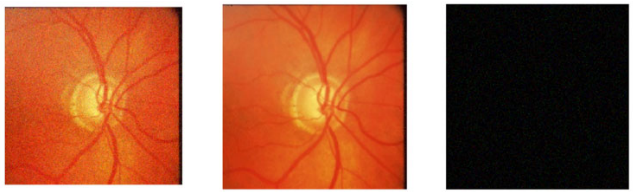
These equations convert the rectified image from homogeneous coordinates to its final 2D Cartesian representation, where $\mathbf{X}'$ and $\mathbf{Y}'$ represent the rectified horizontal and vertical coordinates, respectively.

The perspective transform method **effectively performs rectification using simple matrix operations while maintaining low computational burdens and without requiring any reference images**.

V. Noise-removal module

A. Problem

One major challenge in accurately analyzing glaucoma in users' uploaded images taken by phones is the presence of noise. The noise in uploaded images can be attributed to factors including camera shake, original movements of objects, and out-of-focus optics [30]. Typical types of noise include Gaussian and impulse noises [31]. The quality of colored retinal fundus images is significantly degraded by these noises, making it difficult to detect glaucoma.



(a) Observation

(b) Noise-free image

(c) Noise

Fig. 4: Noise-removal module: separate the noise-free image from the observed image.

In order to address this issue, a noise-removal module, as shown in Fig. 4, is introduced with the objective of efficiently generating a noise-free image from the observed image.

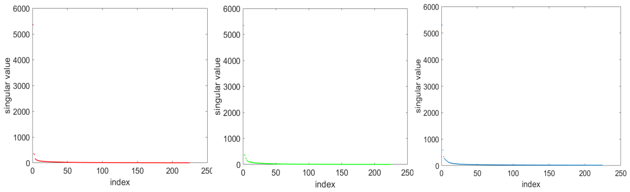
B. Method

Currently, there are various denoising methods available, and one representative method is the low-rank tensor recovery approach. Among existing tensor recovery approaches, the Tucker decomposition (TKD) is an effective denoising tool due to its strong representation ability. TKD links factor matrices with a core tensor to capture hidden data and achieve satisfactory image recovery performance [32]. However, existing TKD-based models suffer from low computational efficiency when dealing with large-scale tensors due to complex SVD operations. Additionally, predetermining the Tucker rank beforehand adds to the overall complexity of the problem-solving process.

To overcome these limitations, this paper adopts the FTDTV model [11] and proposes the ADMM algorithm. The FTDTV model, with the assistance of low-rank factor prior, helps alleviate computation burdens and eliminates the need for determining the Tucker rank in advance. By leveraging the advantages of the FTDTV model, the corresponding noise-free image can be separated from the observed image, effectively isolating the noise.

The proposed noise-removal method treats a noisy image as a combination of its corresponding noise-free image and noise, both represented in tensor form. This allows for direct denoising from the input image itself by separating the noise-free image. Since the input colored retinal fundus images are represented in the RGB form, they can be represented as fourth-order tensors. Therefore, the noise-free image is represented by $\mathbf{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3 \times I_4}$, the original observation by $\mathbf{Y} \in \mathbb{R}^{I_1 \times I_2 \times I_3 \times I_4}$, and the noise by $\mathbf{Z} \in \mathbb{R}^{I_1 \times I_2 \times I_3 \times I_4}$, where I_1 and I_2 are the length and width of the image, respectively, I_3 defaults to 3 for RGB images with 3 channels, and I_4 represents the number of images included in the dataset.

In order to determine the appropriateness of using low-rank separation to extract the noise-free image, SVD is conducted on some digital-version original colored fundus retinal images to validate their low-rankness.



Mode-1 utfolding Matrix

Mode-2 utfolding Matrix

Mode-3 utfolding Matrix

Fig. 5: Singular values of the mode 1, mode 2, and mode 4 utfolding matrices.

Fig. 5 illustrates the curves depicting the change in singular values of the unfolding matrices for mode 1, mode 2, and mode 4. The declining nature of these curves clearly indicates that the majority of singular values are close to zero, while only a few large singular values dominate. This observation provides strong evidence for the low-rankness of the image.

Leveraging the low-rankness of the images, the separation of the noise-free image can be achieved using Robust Principle Component Analysis (RPCA), as depicted in Figure 5. The basic and conceptual noise-removal model is outlined as follows:

$$\mathbf{Y} = \mathbf{X} + \mathbf{Z} \tag{3}$$

The conceptual model aims to extract the low-rank noise-free image $\mathbf{X}$ from the observed image $\mathbf{Y}$, which can be formulated as an inverse problem. To ensure the problem is well-posed and avoids instability, prior regularization is incorporated for both the original observation and the noise. In this context, the FTDTV model is utilized, which is a low-rank tensor denoising model that combines factors prior and total variation regularization. The modeling of the FTDTV model is as follows:

$$\begin{gathered}\min \lambda_1 \sum_{n=1}^{N} \beta_n \left|F_n X_{(n)}\right| + \alpha_n \sum_{n=1}^{N} \|U_n\|_* + \lambda_2 \|\mathcal{G}\|_F^2 + \lambda_3 \Phi_2(Z) \\ \textit{s.t.}\ \ X = \mathcal{G} \times {}_1\mathrm{U}_1 \times {}_2\mathrm{U}_2 \ldots \times {}_\mathrm{N}\mathrm{U}_\mathrm{N}\ \ \textit{and}\ \ Y = X + Z\end{gathered} \tag{4}$$

The regularization coefficients $\{\alpha_n\}_{n=1}^{\mathrm{N}}, \lambda_1, \lambda_2, \lambda_3$ are introduced in the model. The term $\sum_{\mathrm{n}=1}^{\mathrm{N}} \beta_\mathrm{n} \left|\mathrm{F}_\mathrm{n}\mathrm{X}_{(\mathrm{n})}\right|$ represents the total variation regularization, where β_n can take values of either 1 or 0. The L1 norm of a matrix is denoted by $| \bullet |$, and F_n is a matrix of dimensions $(I_n - 1) \times I_{n'}$ with all elements being 0 except for $[F_n]_{i,i} = 1$ and $[F_n]_{i,i+1} = -1$. The total variation regularization term is employed to promote local piecewise smoothness in the images.

The low-rank properties of the noise-free image $\mathbf{X}$ are extracted using $\sum_{n=1}^{N} \|\mathbf{U_n}\|_*$ and Tucker decomposition, where $\mathbf{X}$ is represented as $\mathbf{X} = \mathcal{G} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \cdots \times_N \mathbf{U}_N$, and $\mathcal{G}$ serves as an overfitting-preventing term.

The term $\Phi_2(Z)$ denotes the sparse regularizer for Z, and in this model, a non-convex and non-smooth MCP (Minimax Concave Penalty) function [11] is employed to overcome the limitations of the commonly used L1 norm in existing literature.

In order to solve the aforementioned model, auxiliary variables need to be introduced, leading to the following equivalent model:

$$\begin{aligned}&\min \lambda_1 \sum_{n=1}^{N} \beta_n \left|Q_n\right| + \alpha_n \sum_{n=1}^{N} \|U_n\|_* + \lambda_2 \|\mathcal{G}\|_F^2 + \lambda_3 \Phi_2(Z) \\ &\textit{s.t.}\ \ \{Q_n = F_n R_n, R_n = X_{(n)}, V_n = U_n\}_{n=1}^{N} \\ &X = \mathcal{G} \times {}_1\mathrm{U}_1 \times {}_2\mathrm{U}_2 \ldots \times {}_\mathrm{N}\mathrm{U}_\mathrm{N} \\ &Y = X + Z\end{aligned} \tag{5}$$

To handle the "min" constraint in the model and simplify the optimization process, this paper adopts Alternating Direction Method of Multipliers (ADMM). The augmented Lagrangian function of the proposed model is as follows:

$$\begin{aligned}&\mathcal{L}\left(\mathrm{Q_n}, \mathrm{R_n}, \mathrm{U_n}, \mathrm{V_n}, \mathrm{X}, \mathcal{G}, \mathrm{Z}; \Lambda_\mathrm{n}, \Omega_\mathrm{n}, \Gamma_\mathrm{n}, \mathcal{W}, \mathcal{K}\right) = \\ &\lambda_1 \sum_{\mathrm{n}=1}^{\mathrm{N}} \beta_\mathrm{n} \left|\mathrm{Q_n}\right| + \alpha_\mathrm{n} \sum_{\mathrm{n}=1}^{\mathrm{N}} \|\mathrm{U_n}\|_* + \lambda_2 \|\mathcal{G}\|_\mathrm{F}^2 + \\ &\lambda_3 \Phi_2(\mathrm{Z}) + \sum_{\mathrm{n}=1}^{\mathrm{N}} \beta_\mathrm{n} \frac{\varrho}{2} \left\| \mathrm{Q_n} - \mathrm{F_n R_n} + \frac{\Lambda_\mathrm{n}}{\varrho} \right\|_\mathrm{F}^2 + \\ &\sum_{\mathrm{n}=1}^{\mathrm{N}} \beta_\mathrm{n} \frac{\varrho}{2} \left\| \ \mathrm{V_n} - \mathrm{U_n} + \frac{\Omega_\mathrm{n}}{\varrho} \right\|_\mathrm{F}^2 + \sum_{\mathrm{n}=1}^{\mathrm{N}} \beta_\mathrm{n} \frac{\varrho}{2} \|\mathrm{R_n} - \\ &\mathrm{X}_{(\mathrm{n})} + \frac{\Gamma_\mathrm{n}}{\varrho} \ \|_\mathrm{F}^2 + \frac{\varrho}{2} \left\| \ \mathrm{Y} - \mathrm{X} - \mathrm{Z} + \frac{\mathcal{K}}{\varrho} \ \right\|_\mathrm{F}^2 + \frac{\varrho}{2} \left\| \ X - \right. \\ &\mathcal{G} \times {}_1\ \mathrm{V}_1 \times {}_2\ \mathrm{V} \ldots \times_\mathrm{N} \mathrm{V_N} + \frac{\mathfrak{w}}{\varrho} \|_\mathrm{F}^2,\end{aligned} \tag{6}$$

where $\Lambda_\mathrm{n}, \Omega_\mathrm{n}, \Gamma_\mathrm{n}, \mathcal{W}, \mathcal{K}$ are Lagrangian multipliers, and ϱ is the penalty term. Then the ADMM algorithms is as follows:

C. Analysis

The incorporation of the MCP function enables the noise-removal module to effectively extract sparse noises, while the ADMM algorithm efficiently computes the result for the proposed FTDTV model. The noise-removal module achieves high accuracy and will significantly improve the overall performance of the final prediction.

Notably, the noise-removal module exhibits satisfactory efficiency due to two key factors. Firstly, the low-rank prior is applied to the small-size factors, thereby reducing the computational costs associated with SVD operations. Secondly, the

Algorithm 1 ADMM algorithm

```
Require: observation $Y$, the parameters $\alpha_n, \beta_n, \lambda_1, \lambda_2, \lambda_3, \tau, \varrho, \mu, \gamma_1, \gamma_2$.
Initialize: $X^0, \mathcal{G}^0, Z^0, \mathcal{W}^0, \mathcal{K}^0, \{Q_n^0, R_n^0, \Lambda_n^0, \Omega_n^0, \Gamma_n^0,\}_{n=1}^N$, $k = 0$
while not converge do
    Update $Q_n^{k+1} = \operatorname{argmin} \mathcal{L}\left(Q_n, R_n^k, U_n^k, V_n^k, \mathcal{X}^k, \mathcal{G}^k, Z^k; \Lambda_n^k, \Omega_n^k, \Gamma_n^k, \imath$
    Update $R_n^{k+1} = \operatorname{argmin} \mathcal{L}\left(Q_n^{k+1}, R_n, U_n^k, V_n^k, Z^k, \mathcal{G}^k, \mathcal{S}^k; \Lambda_n^k, \Omega_n^k, \Gamma_n^k, w^k, \mathcal{K}^k\right)$
    Update $\mathrm{U_n^{k+1}} = \operatorname{argmin} \mathcal{L}\left(\mathrm{Q_n^{k+1}, R_n^{k+1}, U_n, V_{n^k}^{k}}, \mathcal{X}^\mathrm{k}, \mathcal{G}^\mathrm{k}, \mathrm{Z^k}; \Lambda_n^k, \Omega_n^k, \Gamma_n^k, w^k, \mathcal{K}^k\right)$
    Update $\mathrm{V_n^{k+1}} = \operatorname{argmin} \mathcal{L}\left(\mathrm{Q_n^{k+1}, R_n^{k+1}, U_n^{k+1},\ V_n}, \mathcal{X}^\mathrm{k}, \mathcal{G}^\mathrm{k}, \mathrm{Z^k}; \Lambda_n^k, \Omega_n^k, \Gamma_\pi^k, w^k, \mathcal{K}^k\right)$
    Update $\mathcal{X}^{k+1} = \operatorname{argmin} \mathcal{L}\left(\mathrm{Q_n^{k+1}, R_n^{k+1}, U_n^{k+1},\ V_n^{k+1}}, \mathcal{X}, \mathcal{G}^\mathrm{k}, \mathrm{Z^k}; \Lambda_n^k, \Omega_n^k, \Gamma_n^k, w^k, \mathcal{K}^k\right)$
    Update $\mathcal{G}^{k+1} = \operatorname{argmin} \mathcal{L}\left(\mathrm{Q_n^{k+1}, R_n^{k+1}, U_n^{k+1},\ V_n^{k+1}}, \mathcal{X}^\mathrm{k+1}, \mathcal{G}, \mathrm{Z^k}; \Lambda_n^k, \Omega_n^k, \Gamma_n^k, w^k, \mathcal{K}^k\right)$
    Update $Z^{k+1} = \operatorname{argmin} \mathcal{L}\left(Q_n^{k+1}, R_n^{k+1}, U_n^{k+1}, V_n^{k+1}, \mathcal{X}^{k+1}, \mathcal{G}^{k+1}, Z; \Lambda_n^k, \Omega_n^k, \Gamma_{n'}^k w^k, \mathcal{K}^k\right)$
    Update multipliers $\Lambda_n^{k+1}, \Omega_n^{k+1}, \Gamma_n^{k+1}, \mathcal{W}^{k+1}, \mathcal{K}^{k+1}$ and the penalty term $\varrho$
    $k := k + 1$
end while
```

module eliminates the need for determining ranks in advance, further enhancing its efficiency.

Considering both the denoising performance and efficiency, the noise-removal module proposed in this paper is highly applicable and supportive of the prevalent usage scenarios of the App.

VI. Prediction module

A. Problem and challenges

Despite the utilization of high-quality input images preprocessed by the preceding two modules, predicting glaucoma remains a significant challenge due to the following four reasons:

- **Subtle Glaucomatous Indications:** Certain crucial signs of glaucoma in colored fundus images are subtle and difficult to detect. This subtlety complicates the extraction and identification of these minor yet critical features.
- **Imbalanced Training Data:** The dataset used for training glaucoma classification models exhibits severe class imbalance. Non-glaucomatous images substantially outnumber glaucomatous ones, which can lead to biased model learning. In such cases, non-glaucomatous features may dominate, impeding the effective identification of glaucomatous images.
- **Limited Training Data:** Acquiring high-quality glaucoma fundus images is a challenging task, leading to a constrained dataset. Training neural network models with such limited data can potentially lead to overfitting, thereby affecting the model's performance in real-world scenarios.
- **Variability in Data Sources:** Glaucoma fundus image datasets can vary significantly due to differences in data acquisition equipment and procedures. It's difficult to predict the conditions under which patients obtain the images. The network model needs to learn to handle this diversity to accommodate a wide range of images in real-world application scenarios.

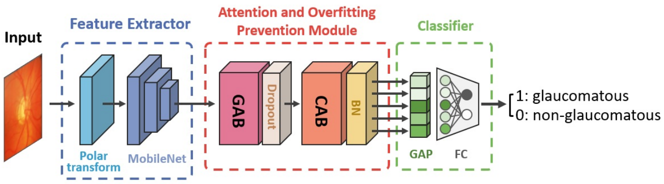


Fig. 6: Prediction module.

B. Method

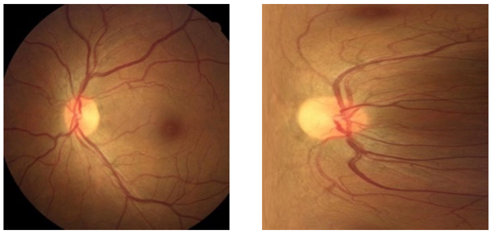
To address the challenges outlined above, this paper proposes a prediction framework specifically tailored for glaucoma prediction. The prediction module (as depicted in Fig.6) consists of three key components:

- **Feature Extractor:** Designed to overcome the challenge of detecting small and hard-to-see glaucomatous signs in input images, the Feature Extractor starts with a preprocessed image that has been through prior rectification and noise-removal modules. This paper first introduces a Polar Transform layer to preprocess the input image, focusing particularly on highlighting crucial regions of interest: the optic cup and disc. This paper then uses MobileNet 1.0 as the backbone to extract image features, selected based on extensive experimentation that has demonstrated its superior performance compared to other network architectures.
- **Attention and Overfitting Prevention Module:** This module effectively addresses the issue of imbalanced sampling tags and mitigates the risk of overfitting. It consists of two key parts: an Attention Module, designed specifically to handle imbalanced labels, which captures fine details and discriminative features using the Global Attention Block (GAB) and the Category Attention Block (CAB); and overfitting prevention mechanisms, including random dropout applied to GAB-produced feature maps and batch normalization implemented on CAB or the overall attention module-generated feature maps.
- **Classifier:** In the final stage of the prediction module, a classifier is utilized to generate the definitive binary diagnostic result. This classifier employs a Global Average Pooling (GAP) layer and a Fully Connected (FC) layer to process the enhanced features and make a glaucoma prediction. Furthermore, to ensure robust prediction, this paper employs the Entropy Loss Function and proposes a diverse learning strategy. This strategy aims to train the network model to recognize and learn diverse features from various types of image sources.

1) Feature Extractor

The feature extraction process in our model begins with the utilization of a **Polar Transform layer**. This layer preprocesses input images and emphasizes critical areas such as the optic cup and disc. Specifically, during the polar transformation, the center of the input (the colored fundus image) is set as the origin in the new polar coordinate system. Once the origin is set, every point with Cartesian coordinates is converted into polar coordinates using a specific formula. Here, (θ, r) represents the resulting polar coordinates, and (x, y) refers to points in the original rectangular Cartesian coordinates.

$$\begin{cases} \theta = \tan^{-1}\left(\frac{y}{x}\right) \\ r = \sqrt{x^2 + y^2} \end{cases} \tag{7}$$



(a) The input image, in which the optic cup and disc are in normal sizes.

(b) The preprocessed image, in which the optic cup and disc are highlighted.

Fig. 7: Polar transform.

Following the polar transformation, the processed input enters the **MobileNet 1.0 backbone feature extractor**. This backbone serves as a fundamental feature extractor within the prediction module. It was selected for its superior performance in feature extraction, a conclusion derived from rigorous experimentation with various models, including Vgg16, Resnet50, Xception, Densenet121, and Inceptionv2.

MobileNet 1.0, a lightweight deep neural network primarily designed for mobile vision applications, is particularly effective for this task [33]. Its architectural design, featuring a stack of multiple 1×1 and 3×3 convolutional layers, an average pooling layer, a Fully Connected (FC) layer, and a Softmax classifier, makes it an efficient and effective backbone for our prediction module.

2) *Attention and Overfitting Prevention Module*

Attention module for imbalanced learning. An existing prediction system that provides diabetic retinopathy grading results is CABNet [34], which incorporates an attention module. This module includes a Global Attention Block (GAB) to capture detailed features, such as subtle lesions, and a Category Attention Block (CAB) for category-level processing of discriminative features, treating them equally with other features [34].

Inspired by CABNet [34], this paper handles imbalanced training data by processing fine information in the attention module after coarse feature extraction through MobileNet 1.0 and input simplification by a 1×1 convolutional layer (as shown in Fig.8). In this module, the GAB learns the global fine details, while the CAB focuses on discriminative regions and refines the feature extraction done by GAB.

In the CAB, the input first undergoes channel attention. This feature selector indicates the importance of each feature channel by learning channel-wise attention weights. The operation process of channel attention is as follows, where F_{c-att} denotes the output of channel attention, σ presents the Sigmoid function, Conv2 refers to two 1×1 convolutional layers, GAP is the Global Attention Pooling layer, F_{GAB-1N} denotes GAB's input, and $\otimes$ signifies element-wise multiplication.

$$F_{c-att} = (\sigma(Conv2(GAP(F_{GAB-IN})))) \otimes F_{GAB-IN} \tag{8}$$

F_{c-att} is then used as the input for spatial attention, which signifies the importance of each spatial position by learning spatial attention weights. The spatial attention feature map is obtained through the following operation, in which C_{GAP} represents cross-channel average pooling.

$$F_{GAB-out} = F_{c-att} \otimes (\sigma(C\_GAP(F_{c-att}))) \tag{9}$$

The output of GAB, denoted as $F_{GAB-out}$ serves as the input for CAB. Within GAB, a 1×1 convolutional layer is first employed to produce feature maps, denoted as $F' \in \mathbb{R}^{H\times W\times 2K}$, where k is the number of channels needed to detect discriminative regions for two classes (glaucomatous and non-glaucomatous). To prevent overfitting, the optimization of random dropout is introduced, generating feature maps with reduced weights ($F'' \in \mathbb{R}^{H\times W\times 2K}$). To ensure that discriminative regions are learned, half of the features are randomly set to zero, resulting in new feature maps ($F''' \in \mathbb{R}^{H\times W\times 2K}$). The importance ($S_i$) of the feature maps for the class is evaluated using the following formula, where GMP stands for global max pooling and $f'''_{i,j}$represents the j^{th} feature map for the i^{th} class from the input feature maps after random dropout (F'''.).

$$S_i = \frac{1}{k}\sum_{j=1}^{k} GMP\left(f'''_{i,j}\right), i \in \{1, 2\} \tag{10}$$

Meanwhile, to prevent important information from being overlooked during the random dropout, all features undergo a category-wise cross-channel average pooling operation, as represented by the following formula. In this formula, $f''_{i,j}$ denotes the j^{th} feature map for the i^{th} class, which is derived from the original input feature maps prior to the random dropout operation, F'.

$$F''_{i\_\text{avg}} = \frac{1}{k}\sum_{j=1}^{k} f''_{i,j}, \quad i \in \{1, 2\} \tag{11}$$

The category attention, denoted as $ATT_{CAB} \in \mathbb{R}^{H\times W\times 1}$, is calculated by combining operations both with and without the random dropout operation. This approach is designed to focus specifically on meaningful discriminative regions and produce the output of CAB, where $F_{CAB-OUT}$ represents the output feature maps.

$$\begin{aligned} ATT_{CAB} &= \frac{1}{L}\sum_{i=1} LS_i F_{i\_avg''} \\ F_{CAB-OUT} &= F_{CAB-IN} \otimes ATT_{CAB} \end{aligned} \tag{12}$$

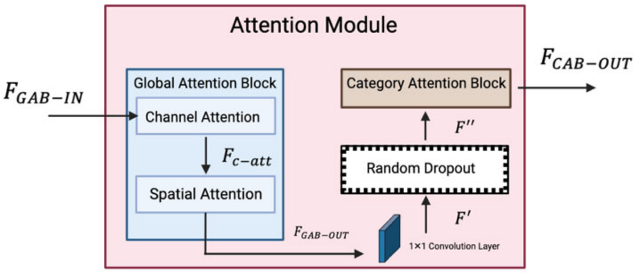
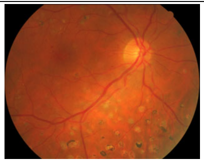
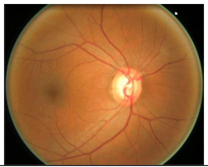
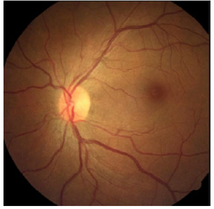
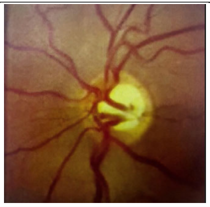


Fig. 8: Attention module.

TABLE I: Three model results diagram.

Model	*Image*	Illustration
G1020		1020 Glaucoma: 724 Negative glaucoma:296
ORIGA		*650 Glaucoma:482 Negative glaucoma:168*
LAG-dataset		*5824 Glaucoma:2392 Negative glaucoma:3432*
Real dataset		*826 Glaucoma:673 Negative glaucoma:153*

Overfitting Prevention Module. To bolster the prediction module's generalization capability, this paper has incorporated an overfitting prevention module consisting of two operational layers. This includes a random dropout, as previously mentioned in the attention module, situated between GAB and CAB to generate feature maps, denoted as F''. During this random dropout operation, the weights for all features are reduced to 0.5. Simultaneously, for the feature maps produced by the attention module, denoted as $F_{CAB-OUT}$, a batch normalization (BN) operation is applied, yielding the output $F'_{CAB-OUT}$ for the classifier.

*3) **Classifier:***

This classifier employs a Global Average Pooling (GAP) layer and a Fully Connected (FC) layer to process the enhanced features and make a glaucoma prediction.

Loss function

The loss function employed during the training of the prediction module to constantly evaluate the model's prediction performance is identified as an Entropy Loss Function. The loss is computed using the following formula, where N denotes the total number of samples, y_i represents the classification result (1 or 0) of the i^{th} sample, and p_i indicates the probability that the i^{th} sample is positive (1).

$$L = \frac{1}{N}\sum_{i=1}^{N} y_i \log\left(p_i\right) + \left(1-y_i\right)\log\left(1-p_i\right) \tag{13}$$

Diversity learning

Fundus image datasets for glaucoma can exhibit significant variability due to differences in data acquisition equipment and procedures. The sources of these variations may include, but are not limited to:

- Diversity in the types and models of fundus photography equipment used.
- Unique settings for parameters such as exposure and contrast on individual fundus photography cameras.
- Comorbidities in glaucoma patients, such as cataracts or vitreous body disorders, which can induce refractive media opacity, thereby affecting the imaging outcomes.

Table I provides a visual representation of the varied images derived from disparate datasets.

Predicting the precise acquisition process of a patient's images is challenging. Consequently, if the training of the network model only relies on one single dataset, the prediction accuracy may be compromised when the user inputs an image that deviates from the training set type.

To leverage the diverse information from various datasets, this paper proposes an integrative approach that combines multiple datasets into a comprehensive dataset. By training the network model using this large dataset, this paper aims to better accommodate the disparities across data sources. This, in turn, is expected to enhance the model's robustness in glaucoma identification tasks. In section VII, experimental results to corroborate the efficacy of the diversity learning approach will be presented and discussed.

VII. Performance evaluation

A. Experiment setup

This paper assess the proposed techniques using four datasets, as detailed in Table I: G1020 [35], ORIGA [36], LAG-dataset [37], and Real-dataset, which is a part of the SIFG-database [38]. Each dataset is split into two subsets: a training set and a testing set.

- The G1020 dataset: The training set comprises 875 images, including 655 glaucomatous and 220 non-glaucomatous. The testing set consists of 145 images, with 69 glaucomatous and 76 non-glaucomatous.
- The ORIGA dataset: The training set includes 520 images, of which 386 are glaucomatous and 134 are non-glaucomatous. The testing set includes 130 images, with 96 glaucomatous and 34 non-glaucomatous.
- The LAG-dataset: The training set includes 5244 images, composed of 2007 glaucomatous and 3237 non-glaucomatous.
- The Real-dataset: The training set includes 695 images, with 590 glaucomatous and 105 non-glaucomatous. The testing set consists of 131 images, which include 83 glaucomatous and 48 non-glaucomatous.

To simulate real-world scenarios where users upload colored retinal fundus images taken with their phones, this paper created a 'phone-taken dataset.' This dataset consists of images from the Real-dataset that were printed out and then photographed using mobile phones.

Accuracy (ACC) is a commonly employed metric for measuring the performance of a classification model. It represents the ratio of correctly classified samples to the total number of samples. In the implementation, this paper uses a result matrix to record the model's predictions for the two categories, and then compute the accuracy. A higher accuracy signifies better model performance.

The Weighted Kappa (WKappa) is a metric used to evaluate the performance of classification models, taking into account both the consistency and the importance of different categories. The WKappa is defined as follows: $W_{Kappa} = 1 - (\sum(w_{ij} \times O_{ij})/\sum(w_{ij} \times E_{ij}))$, where w_{ij} represents the weight given to the agreement between category i and category j. If the model's predictions align with the actual labels, the value of w_{ij} is 0. In contrast, if they are not in agreement, the value of w_{ij} is 1/4. O_{ij} denotes the observed frequency of consistency between the model's predictions and the actual labels for categories i and j. On the other hand, E_{ij} represents the expected frequency for categories i and j.

Weighted Kappa's value range is from -1 to 1. It's typically utilized to assess the consistency between model classifications and random classifications. A value closer to 1 indicates superior model performance.

In this paper's proposed method, several techniques aiming at improving prediction accuracy and robustness are listed below.

- Rectification module: Rectifies the images.
- Noise-removal module: Removes noise.
- Polar Transform layer: Emphasizes key regions such as the optic cup and disc.
- MobileNet 1.0: Extracts fundamental features from images.
- Attention Module: Manages imbalanced labels.
- Overfitting Prevention strategy: Mitigates overfitting.
- Diversity Learning: Ensures model robustness across varied datasets.

To validate the effectiveness of this paper's proposed main modules, this paper constructs a **Basic model** that consists of MobileNet 1.0, the Attention Module, and Overfitting Prevention strategy. This paper refers to the model that includes all modules as the **Whole model**.

B. Ablation Study on Two image preprocessing modules

To enhance the overall prediction efficiency, this paper proposes two image preprocessing modules for rectifying and denoising images before the final classification stage.

The experimental results validate the efficacy of these modules. The effectiveness of the rectification module is demonstrated in Table II, and the effectiveness of the noise-removal module is shown in Table III.

TABLE II: Ablation on the rectification module.

Dataset	Model	ACC	W-kappa
Photo-taken dataset	Basic	0.6415	0.1925
Photo-taken dataset	Basic + Rectification	0.6804	0.2392

TABLE III: Ablation on the noise-removal module.

Dataset	Model	ACC	W-kappa
Photo-taken dataset	Basic	0.8308	0.6046
Photo-taken dataset	Basic+Noise Remove	0.9033	0.9012

C. Ablation Study on the Prediction Module

Table IV presents an ablation study conducted on the targeted optimizations within the prediction module. The results highlight that the inclusion of the Polar Transform and Overfitting-Prevention modules leads to a significant improvement in ACC and Kappa metrics.

TABLE IV: Ablation on the prediction module.

Dataset	Model	ACC	W-kappa
Photo-taken dataset	Whole	0.9301	0.9221
Photo-taken dataset	Whole-Polar Transform	0.8527	0.6517
Photo-taken dataset	Whole-Polar Transform - Overfitting-prevention	0.7829	0.4728

TABLE V: Ablation on diversity learning I.

Training Set	Test Set	Model	ACC	W-kappa
Combined big dataset	Real_dataset	Basic	0.8092	0.5383
Real_dataset	Real_dataset	Basic	0.7786	0.4536
Combined big dataset	ORIGA	Basic	0.7615	0.3386
ORIGA	ORIGA	Basic	0.7154	0.1827
Combined big dataset	G1020	Basic	0.4897	0.0238
G1020	G1020	Basic	0.5179	0.0670
Combined big dataset	LAG-dataset	Basic	0.9569	0.9023
LAG-dataset	LAG-dataset	Basic	0.9362	0.5589

Due to the use of different types of equipment in various hospitals, the colored fundus retinal images in the four datasets exhibit significant differences. In order to enhance the robustness of the prediction module, enabling it to account for a wider variety of colored retinal images, we have employed diversity learning that combines images from all four datasets. According to Tables V and VI, this approach yields satisfactory results.

Table V shows the effectiveness of diversity learning by comparing the performance of the basic model on combined datasets and individual datasets. Except for the G1020 dataset, the ACC and Kappa metrics are higher when the basic model is trained on a diversified dataset rather than individual datasets.

Table VI reinforces the conclusion drawn from Table V, indicating that combined datasets perform better than individual ones. Specifically, when the basic model, trained only on a single dataset without diversity learning, is applied to each dataset to evaluate its generalization ability, the model performs well only on the test set from the same dataset. However, its performance decreases on test sets from other datasets, indicating a lack of

TABLE VI: Ablation on diversity learning II.

Training Set	Test Set	Model	ACC	W-kappa
Real_dataset	Real_dataset	Basic	0.7786	0.4536
	ORIGA		0.2923	-0.1671
	G1020		0.5172	-0.0039
	LAG-dataset		0.5103	-0.0196
ORIGA	Real_dataset	Basic	0.3664	-0.0264
	ORIGA		0.7154	0.3656
	G1020		0.5172	0.0512
	LAG-dataset		0.8328	0.6258
G1020	Real_dataset	Basic	0.3130	-0.4416
	ORIGA		0.7308	-0.0152
	G1020		0.5172	0.0671
	LAG-dataset		0.6207	-0.0327
LAG-dataset	Real_dataset	Basic	0.3817	-0.0084
	ORIGA		0.6692	0.2861
	G1020		0.4897	-0.0056
	LAG-dataset		0.9362	0.8524

TABLE VII: Comparisons to CABNet, MesMLP, and UQ.

Method	Dataset	ACC	W-kappa
CABNet	Photo-taken dataset	0.8076	0.5447
ResMLP	Photo-taken dataset	0.7385	0.6115
UQ	Photo-taken dataset	0.6521	0.3145
Whole	Photo-taken dataset	0.9301	0.9221

robustness when diversity learning is not implemented.

D. Performance Comparison

To evaluate the comprehensive performance of the AI-based early diagnosis system for glaucoma proposed in this paper, this paper compared its experimental results with a widely-used diagnostic systems: the CABNet [34]. Additionally, though not specifically designed for eye disease diagnosis, this paper implemented two recent image classification algorithms: ResMLP [39] and UQ [40]. The CABNet, initially used for grading different levels of diabetic retinopathy, consists of a backbone, an attention module, and a classifier. ResMLP is an architecture built entirely upon multi-layer perceptrons for image classification. It's a simple residual network that alternates between a linear layer, where image patches interact independently and identically across channels, and a two-layer feed-forward network, where channels interact independently per patch. UQ proposes the use of background classes to reduce class activation uncertainty without significantly increasing training time. Notably, neither the CABNet, ResMLP nor UQ include image-preprocessing modules. As shown in Table VII, the diagnostic system proposed in this paper outperforms in terms of both ACC and Wkappa metrics. Significant improvements, particularly in diagnosing phone-taken images, can be attributed to the image preprocessing modules and the targeted optimizations in the prediction module proposed in this paper. Specifically, the proposed system delivers approximately 11% higher ACC than CABNet. For the kappa value, this paper's proposed system surpasses CABNet by a remarkable 38%, demonstrating the proposed system's superior capability in providing accurate and robust glaucoma detection results based on phone-taken colored retinal fundus images. While ResMLP is a promising recent image classification algorithm, its ACC is inferior to both CABNet and this paper's proposed system, as it does not specialize in disease diagnosis.

VIII. Conclusion

This paper proposes an App prototype designed to provide online, early diagnostic results for glaucoma based on machine learning algorithms. To achieve this, three basic modules have been developed to perform rectification, denoising, and prediction tasks, respectively. The extensive experiments presented in this paper demonstrate that the techniques are effective. In the future, this online glaucoma early diagnosis App could prove particularly useful for people living in remote regions or those with limited financial resources. They could easily obtain AI-based diagnostic results and, if necessary, seek early treatment. Consequently, delayed diagnoses and subsequent deterioration can be largely alleviated, improving overall eye health outcomes.

REFERENCES

[1] W. H. Organization *et al.*, "World report on vision," 2019.

[2] E. M. Stone, J. H. Fingert, W. L. Alward, T. D. Nguyen, J. R. Polansky, S. L. Sunden, D. Nishimura, A. F. Clark, A. Nystuen, B. E. Nichols *et al.*, "Identification of a gene that causes primary open angle glaucoma," *Science*, vol. 275, no. 5300, pp. 668–670, 1997.

[3] D. A. Lee and E. J. Higginbotham, "Glaucoma and its treatment: a review," *American journal of health-system pharmacy*, vol. 62, no. 7, pp. 691–699, 2005.

[4] H. A. Quigley and A. T. Broman, "The number of people with glaucoma worldwide in 2010 and 2020," *British journal of ophthalmology*, vol. 90, no. 3, pp. 262–267, 2006.

[5] "GlaucomaFactsandStats|Glaucoma.org.Accessed:Feb1,2022[Glaucoma.org].Available:glaucoma.org/glaucoma-facts-and-stats/."

[6] R. Thomas, "Glaucoma in developing countries," *Indian journal of ophthalmology*, vol. 60, no. 5, p. 446, 2012.

[7] Y. Eslami, H. Amini, R. Zarei, G. Fakhraie, S. Moghimi, S. F. Mohammadi, M. Sheibani, and R. Daneshvar, "Socioeconomic factors and disease severity at glaucoma presentation," 2011.

[8] I. M. Eissa, N. B. Abu Hussein, A. E. Habib, Y. M. El Sayed *et al.*, "Examining delay intervals in the diagnosis and treatment of primary open angle glaucoma in an egyptian population and its impact on lifestyle," *Journal of Ophthalmology*, vol. 2016, 2016.

[9] P. J. Foster and G. J. Johnson, "Glaucoma in china: how big is the problem?" *British journal of ophthalmology*, vol. 85, no. 11, pp. 1277–1282, 2001.

[10] Cnnic, "The 49th statistical report on china's internet development," *China Internet Network Information Center*, 2021.

[11] X. Tian, K. Xie, and H. Zhang, "A low-rank tensor decomposition model with factors prior and total variation for impulsive noise removal," *IEEE Transactions on Image Processing*, vol. 31, pp. 4776–4789, 2022.

[12] J. S. Schuman, M. R. Hee, A. V. Arya, T. Pedut-Kloizman, C. A. Puliafito, J. G. Fujimoto, and E. A. Swanson, "Optical coherence tomography: a new tool for glaucoma diagnosis." *Current opinion in ophthalmology*, vol. 6, no. 2, pp. 89–95, 1995.

[13] "WhatIsOpticalCoherenceTomography?.Accessed:Apr2,2018, [AmericanAcademyofOphthalmology].Available:www.aao.org/eye-health/treatments/what-is-optical-coherence-tomography."

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

[15] Z. Burgansky-Eliash, G. Wollstein, T. Chu, J. D. Ramsey, C. Glymour, R. J. Noecker, H. Ishikawa, and J. S. Schuman, "Optical coherence tomography machine learning classifiers for glaucoma detection: a preliminary study," *Investigative ophthalmology & visual science*, vol. 46, no. 11, pp. 4147–4152, 2005.

[16] C.-W. Wu, H.-L. Shen, C.-J. Lu, S.-H. Chen, and H.-Y. Chen, "Comparison of different machine learning classifiers for glaucoma diagnosis based on spectralis oct," *Diagnostics*, vol. 11, no. 9, p. 1718, 2021.

[17] A. Kamalipour, S. Moghimi, H. Hou, R. C. Penteado, W. H. Oh, J. A. Proudfoot, N. El-Nimri, E. Ekici, J. Rezapour, L. M. Zangwill *et al.*, "Oct angiography artifacts in glaucoma," *Ophthalmology*, vol. 128, no. 10, pp. 1426–1437, 2021.

[18] J. Olson, P. Sharp, K. Goatman, G. Prescott, G. Scotland, A. Fleming, S. Philip, C. Santiago, S. Borooah, D. Broadbent *et al.*, "Improving the economic value of photographic screening for optical coherence tomography-detectable macular oedema: a prospective, multicentre, uk study," *Health Technology Assessment*, 2013.

[19] "Boyd,Kierstan,VisualFieldTest.Accessed:Jan27,2019, [AmericanAcademyofOphthalmology].Available:www.aao.org/eye-health/tips-prevention/visual-field-testing."

[20] B. Bengtsson, V. M. Patella, and A. Heijl, "Prediction of glaucomatous visual field loss by extrapolation of linear trends," *Archives of ophthalmology*, vol. 127, no. 12, pp. 1610–1615, 2009.

[21] K. Park, J. Kim, and J. Lee, "Visual field prediction using recurrent neural network," *Scientific reports*, vol. 9, no. 1, p. 8385, 2019.

[22] H. A. Quigley, E. M. Addicks, and W. R. Green, "Optic nerve damage in human glaucoma: Iii. quantitative correlation of nerve fiber loss and visual field defect in glaucoma, ischemic neuropathy, papilledema, and toxic neuropathy," *Archives of ophthalmology*, vol. 100, no. 1, pp. 135–146, 1982.

[23] "ColorFundusPhotography|DepartmentofOphthalmology.Accessed: 2019,[Med.ubc.ca].Available:ophthalmology.med.ubc.ca/patient-care/ophthalmic-photography/color-fundus-photography/."

[24] M. K. Dutta, A. K. Mourya, A. Singh, M. Parthasarathi, R. Burget, and K. Riha, "Glaucoma detection by segmenting the super pixels from fundus colour retinal images," in *2014 international conference on medical imaging, m-health and emerging communication systems (MedCom)*. IEEE, 2014, pp. 86–90.

[25] C. Muramatsu, Y. Hayashi, A. Sawada, Y. Hatanaka, T. Hara, T. Yamamoto, and H. Fujita, "Detection of retinal nerve fiber layer defects on retinal fundus images for early diagnosis of glaucoma," *Journal of biomedical optics*, vol. 15, no. 1, pp. 016 021–016 021, 2010.

[26] Y. Han, W. Li, M. Liu, Z. Wu, F. Zhang, X. Liu, L. Tao, X. Li, and X. Guo, "Application of an anomaly detection model to screen for ocular diseases using color retinal fundus images: design and evaluation study," *Journal of medical Internet research*, vol. 23, no. 7, p. e27822, 2021.

[27] R. Shinde, "Glaucoma detection in retinal fundus images using u-net and supervised machine learning algorithms," *Intelligence-Based Medicine*, vol. 5, p. 100038, 2021.

[28] X. Hu, L.-X. Zhang, L. Gao, W. Dai, X. Han, Y.-K. Lai, and Y. Chen, "Glim-net: chronic glaucoma forecast transformer for irregularly sampled sequential fundus images," *IEEE Transactions on Medical Imaging*, 2023.

[29] X. Shen, F. Darmon, A. A. Efros, and M. Aubry, "Ransac-flow: generic two-stage image alignment," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 2020, pp. 618–637.

[30] X. Mao, Y. Liu, W. Shen, Q. Li, and Y. Wang, "Deep residual fourier transformation for single image deblurring," *arXiv preprint arXiv:2111.11745*, vol. 2, no. 3, p. 5, 2021.

[31] A. Noor, Y. Zhao, R. Khan, L. Wu, and F. Y. Abdalla, "Median filters combined with denoising convolutional neural network for gaussian and impulse noises," *Multimedia Tools and Applications*, vol. 79, pp. 18 553–18 568, 2020.

[32] M. Farhadi and Y. Yang, "Tkd: Temporal knowledge distillation for active perception," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2020, pp. 942–951.

[33] "MobileNetV1.Accessed:Aug13.2023,[Huggingface.co].Available: huggingface.co/docs/transformers/model_doc/mobilenet_v1."

[34] A. He, T. Li, N. Li, K. Wang, and H. Fu, "Cabnet: Category attention block for imbalanced diabetic retinopathy grading," *IEEE Transactions on Medical Imaging*, vol. 40, no. 1, pp. 143–153, 2020.

[35] M. N. Bajwa, G. A. P. Singh, W. Neumeier, M. I. Malik, A. Dengel, and S. Ahmed, "G1020: A benchmark retinal fundus image dataset for computer-aided glaucoma detection," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–7.

[36] Z. Zhang, F. Yin, J. Liu, W. Wong, N. Tan, B. Lee, J. Cheng, and T. Wong, "Origa: An online retinal fundus image database for glaucoma analysis and research," in *Proceedings of the 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pp. 3065–3068.

[37] "PaperswithCode-LAGDataset..Available:Paperswithcode.com, paperswithcode.com/dataset/lag."

[38] "SIFG-database.Available:https://github.com/XiaofeiWang2018/DeepGF."

[39] H. Touvron, P. Bojanowski, M. Caron, M. Cord, A. El-Nouby, E. Grave, G. Izacard, A. Joulin, G. Synnaeve, J. Verbeek *et al.*, "Resmlp: Feedforward networks for image classification with data-efficient training," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 5314–5321, 2022.

[40] H. Dipu Kabir, "Reduction of class activation uncertainty with background information," *arXiv e-prints*, pp. arXiv–2305, 2023.

Acknowledgement

This project is inspired by one of Mr. Yining Shi's guided projects, in which the student designed an instrument with which retinal fundus images can be taken with phone cameras to aid eye-disease diagnoses. This project aims at proposing a glaucoma diagnosis method based on phone-taken colored retinal fundus images for alleviating the optical health problems due to the inequality in allocation of medical resources, especially in remote regions. In this project, I'm in charge of the overall system design, implementation, optimizations processes, and paper writing in the project. Thanks for the patient and meticulous guidance from Prof. Ke Xu, for his most valuable guidance and discussions about algorithm mechanism designs, system implementations, experimental analysis, and the revisions on the paper. I also thank Mr. Yining Shi for his instruction on the inspiration of the project's implementation scenarios, algorithm optimizations and the mechanism realizations. Prof. Ke Xu's and Mr. Yining Shi's guidance and helps are both unpaid. Finally, thanks for Prof. Yiqiang Chen and his team from University of Chinese Academy of Sciences for providing part of the testing image datasets, and thanks for Dr. Yi Zhou from Xiangya Hospital of Central South University for providing pathological instructions on glaucoma and diagnostic guidance based on colored retinal fundus images.

指导老师简历

徐恪，清华大学教授、博士生导师、计算机系副主任，国家杰出青年科学基金获得者，入选北京市卓越青年科学家计划，曾获国家技术发明二等奖、国家科技进步二等奖、中国电子学会电子信息科学技术奖一等奖。2011年获中国计算机学会青年科学家奖，2012年获中创软件人才奖，是中国电子学会理事和会士，曾在ACM SIGCOMM、Oarkland、ACM CCS、UNISEX Security、NDSS和IEEE/ACM TON等国际顶级会议和期刊发表论文100余篇。曾经获得Globecom 2015和IWQoS 2021的最佳论文奖和USENIX Security 2023的杰出论文奖。

施一宁，中国人民大学附属中学教师，担任年级组组长、通用技术教研组高二年级备课组组长、ICC部门通用技术备课组长，负责普通高中通用技术学科的教学组织及课程建设，负责人大附中特色课程《虚拟现实技术》课程的研发与教育教学。获得海淀区研究性学习优秀课题指导教师优秀奖、约翰霍普金斯CTY项目认证教师、京津冀创客教育优秀辅导教师、海淀区金鹏科技论坛优秀辅导教师、第十六届“全国中小学信息技术创新与实践活动”STEAM课程一等奖、第十六届“全国中小学信息技术创新与实践活动”教育信息化发明创新奖、海淀区青少年科技创新大赛科技辅导员一等奖、北京教育学院卓越教师工作室青年教师团队成员、海淀区“风采杯”通用技术学科教学设计二等奖、海淀区优秀班主任等奖励。