www.nature.com/scientificreports

scientific reports

OPEN

Automated diagnosing primary open-angle glaucoma from fundus image by simulating human's grading with deep learning

Mingquan Lin[1], Bojian Hou[1], Lei Liu[2], Mae Gordon[3], Michael Kass[3], Fei Wang[1,5]✉, Sarah H. Van Tassel[4,5]✉ & Yifan Peng[1,5]✉

Primary open-angle glaucoma (POAG) is a leading cause of irreversible blindness worldwide. Although deep learning methods have been proposed to diagnose POAG, it remains challenging to develop a robust and explainable algorithm to automatically facilitate the downstream diagnostic tasks. In this study, we present an automated classification algorithm, GlaucomaNet, to identify POAG using variable fundus photographs from different populations and settings. GlaucomaNet consists of two convolutional neural networks to simulate the human grading process: learning the discriminative features and fusing the features for grading. We evaluated GlaucomaNet on two datasets: Ocular Hypertension Treatment Study (OHTS) participants and the Large-scale Attention-based Glaucoma (LAG) dataset. GlaucomaNet achieved the highest AUC of 0.904 and 0.997 for POAG diagnosis on OHTS and LAG datasets. An ensemble of network architectures further improved diagnostic accuracy. By simulating the human grading process, GlaucomaNet demonstrated high accuracy with increased transparency in POAG diagnosis (comprehensiveness scores of 97% and 36%). These methods also address two well-known challenges in the field: the need for increased image data diversity and relying heavily on perimetry for POAG diagnosis. These results highlight the potential of deep learning to assist and enhance clinical POAG diagnosis. GlaucomaNet is publicly available on https://github.com/bionlplab/GlaucomaNet.

Primary open-angle glaucoma (POAG) is one of the leading causes of blindness in the US and worldwide[1]. It has been projected to affect approximately 111.8 million people by 2040. Among these patients, 5.3 million may be bilaterally blind[2]. In the United States, POAG is the most common form of glaucoma and is the leading cause of blindness among African-Americans[3] and Hispanics[4]. POAG is asymptomatic until it reaches an advanced stage when visual field (VF) loss occurs. Therefore, accurately identifying individuals with glaucoma is critical to clinical decision-making, which can help inform the need for medical and surgical treatments and patient monitoring[5,6].

Screening is an efficient and effective way to detect POAG to prevent blindness[7]. However, screening needs expensive facilities and experienced ophthalmologists, which causes low prevalence[7]. Meanwhile, fundus photography is convenient and inexpensive for recording optic nerve head structure. But detecing POAG on fundus photographs still requires human labor and severely depends on experienced ophthalmologists. Therefore, it is important to develop an automatic model to detect POAG with high accuracy and low cost from fundus photographs.

Developments in artificial intelligence have provided potential opportunities for automatic POAG diagnosis using fundus photographs. Singh et al. segmented optic disc and optic cup images to obtain the vertical cup-to-disc ratio (VCDR) and extracted the handcrafted features from the VCDR to perform the classification[8]. Acharya et al. extracted texture features and higher-order spectral features to detect POAG based on a Support Vector Machine (SVM) and Naïve Bayesian (NB) classifier[9]. Dua et al. also used SVM and NB to classify the

[1]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA. [2]Institute for Public Health, Washington University School of Medicine, St. Louis, MO, USA. [3]Department of Ophthalmology and Visual Sciences, Washington University School of Medicine, St. Louis, MO, USA. [4]Department of Ophthalmology, Weill Cornell Medicine, New York, NY, USA. [5]These authors jointly supervised this work: Fei Wang, Sarah H. Van Tassel and Yifan Peng. ✉email: few2001@med.cornell.edu; sjh2006@med.cornell.edu; yip4002@med.cornell.edu

Dataset	OHTS	LAG
#Participants	1636	4855
#Images	37,339	4915
#POAG	2321	1711
#Normal	35,018	3144

Table 1. Characteristics of the OHTS and LAG datasets.

wavelet-based energy features[10]. An adaptive threshold-based image processing method was adopted by Issac et al. for POAG classification[11]. However, these methods are influenced by the segmentation accuracy of the optic disc and optic cup or have relatively low classification accuracy because they only consider handcrafted features. Recently, deep learning methods have demonstrated promising results in biology and medicine[12–27]. For POAG detection, several works directly used Convolutional Neural Networks (CNN) to detect glaucoma[28–33]. Fu et al. provided a novel dis-aware ensemble network for glaucoma screening from fundus images[34]. Li et al. put forward a framework that integrates holistic and local deep features for glaucoma classification[35].

While these studies have brought significant improvements, several limitations hinder their translation into clinical practice. First, most models were developed on small datasets from a single institution[28,29,34,35]. It makes the methods less generalizable to different populations and settings. Second is the requirement of refined image annotation such as attention maps[30], image cropping[32], and optic nerve head segmentation[33]. Today, it remains challenging to build highly-accurate systems using image-level labels. Finally, all these works lack interpretability which reduces their utility for medical applications. Thus, it is critical to develop an interpretable AI system for ophthalmologists besides providing a fast and accurate POAG diagnosis. Specifically, interest has recently grown in designing clinical systems that can reveal why models make specific predictions[36–38]. To the best of our knowledge, few works in POAG diagnosis have been studied in this direction.

This work addresses these issues by proposing a joint fusion network (GlaucomaNet) to accurately and robustly diagnose POAG using variable fundus photographs. Different from previous studies, GlaucomaNet consists of two convolutional neural networks to simulate the human grading process: one performing preliminary grading and the other detailed grading. It is trained in an end-to-end strategy without manual cropping so that the whole training process is fast and less labor-intensive and can be quickly deployed to provide a first-line assessment. In addition, an ensemble of network architectures further improved diagnostic accuracy. We validated the models on two independent datasets (Table 1). One is the Ocular Hypertension Treatment Study (OHTS) cohort, one of the largest longitudinal clinical trials in POAG (1636 participants and 37,399 images) from 22 centers in the United States[39]. The other is the Large-scale Attention-based Glaucoma (LAG) database, a publicly available database collected at the Chinese Glaucoma Study Alliance and Beijing Tongren Hospital[30]. GlaucomaNet achieved the highest AUC for POAG diagnosis on these two datasets. To further evaluate the rationale of our model, we consider the prediction from the model for the POAG diagnosis once the supporting rationales (optic disc) are stripped. To this end, we constructed a new dataset by masking the optic disc portion of the fundus photographs. When stripping the bounding box of optic discs and random bounding boxes were masked, we showed that the rationales (optic disc) are influential in the POAG diagnosis in our network.

Methods

Data acquisition.

In this study, we include two independent datasets. These two databases are large, cross-sectional, longitudinal, and population-based studies. All participants provided informed consent at original study entry. This study does not need institutional review board approval because it does not constitute human subjects research. All experiments were performed in accordance with relevant guidelines and regulations.

Ocular hypertension treatment study. The first dataset was obtained from the Ocular Hypertension Treatment Study (OHTS)[39]. It is one of the largest longitudinal clinical trials in POAG from 22 centers in the United States. The study protocol was approved by an independent Institutional Review Board at each clinical center.

The participants in this dataset were selected according to both eligibility and exclusion criteria[39]. Briefly, the eligibility criteria include intraocular pressure (IOP) (between 24 and 32 mmHg in one eye and between 21 and 32 mmHg in the fellow eye) and age (between 40 and 80 years old). The visual field tests were interpreted by the Visual Field Reading Center, and stereoscopic photographs of the optics discs were interpreted by the Optic Disc Reading Center. Exclusion criteria included previous intraocular surgery, visual acuity worse than 20/40 in either eye, and diseases that may cause optic disc deterioration and visual field loss (such as diabetic retinopathy).

The reading center workflow has been described in detail in Kass et al.[39]. In brief, two certified readers independently assessed optic disc deterioration. If there was disagreement between two readers, a senior reader adjudicated in a masked fashion. The POAG diagnosis in a quality control sample of 86 eyes (50 normal eyes and 36 with progression) showed test–retest agreement at $\kappa = 0.70$ (95% confidence interval [CI], 0.55–0.85).

Figure 1 shows the creation of the OHTS dataset in our study. Because of the inherent redundancy in a pair of stereoscopic photographs, we used only one of the photographs for each eye. There are 7964 images (11.9%) that contain the stereo pair in a single image file. We split them into two images. In the end, we obtained 39,339 fundus photographs from 3272 eyes of 1636 subjects. On the image level, 6.2% (2321) of fundus photographs were diagnosed with POAG (Table 1).

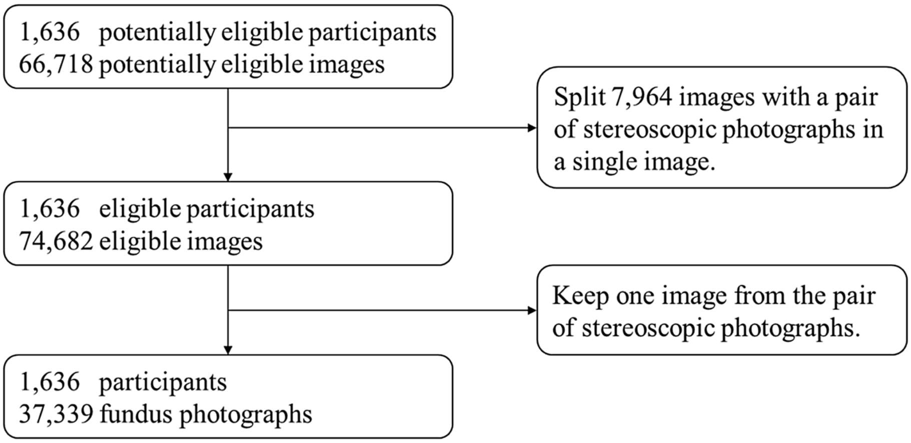
1,636 potentially eligible participants
66,718 potentially eligible images
Split 7,964 images with a pair of stereoscopic photographs in a single image.
1,636 eligible participants
74,682 eligible images
Keep one image from the pair of stereoscopic photographs.
1,636 participants
37,339 fundus photographs

Figure 1. Creation of the OHTS dataset.

Large-scale Attention-based Glaucoma database. The second dataset was obtained from the Large-scale Attention-based Glaucoma (LAG) database, a publicly available database collected at the Chinese Glaucoma Study Alliance and Beijing Tongren Hospital[40] (https://github.com/smilell/AG-CNN). LAG contains 4855 fundus images, of which 35% (1711) have POAG. This dataset has a high proportion of glaucoma cases (41.2%), exceeding what would be expected when screening for glaucoma, which was designed intentionally to create a balanced sample. All research adhered to the tenets of the Declaration of Helsinki. Qualified glaucoma specialists diagnosed each fundus image by considering both morphologic and functional analyses such as intraocular pressure, visual field loss, and manual optic disc assessment. Finally, all fundus images were confirmed with positive or negative glaucoma, seen as the gold standard.

Model development.

Overall architecture. GlaucomaNet comprises two convolutional blocks followed by seven layers (Fig. 2). First, a single fundus image x_i is passed through two convolutional neural networks, DenseNet-201[41] and ResNet-152[42]. For DenseNet-201, we used the output of the 200 layers. For ResNet-152, we used the output of the 151st layer. We denote the outputs as F_d and F_r, respectively. We then concatenated F_d and F_r and used 1×1 convolution (conv1d) followed by a batch normalization layer (BN)[43] and rectified linear units (ReLU)[44] as transition layers. In the end, a global average pooling (AvgPooling)[45] with a fully connected layer is performed, and a sigmoid classifier is attached.

$$F_d = DenseNet201(x_i)$$

$$F_r = ResNet152(x_i)$$

$$F_1 = conv1d(concat(F_d, F_r))$$

$$F_2 = ReLU(BN(F_1))$$

$$F_3 = AvgPooling(F_2)$$

$$o = sigmoid(W_3 F_3 + b_3)$$

Loss function. On both datasets, there is a severe class imbalance between non-POAG and POAG images. For example, only 6.22% of images in the OHTS dataset contain POAG. To overcome this data imbalance issue, we applied weighted cross-entropy[46], a commonly used loss function in classification. The adopted weighted cross-entropy was as follows:

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}\left[w_1 y_n log\left(\widehat{y}_n(x_n, \theta_s)\right) + w_2(1 - y_n)log\left(1 - \widehat{y}_n(x_n, \theta_s)\right)\right]$$

where N is the number of training examples, w_1 and w_2 are controlling hyperparameters, y_n is the ground truth while $\widehat{y}_n$ is the likelihood predicted by the classifier, and θ_s represents the parameters of the neural network.

Image augmentation. A stochastic image augmentation was randomly applied to transform a fundus photograph into an augmented view. In this work, we sequentially appled three simple augmentation operations: (1) random rotation between 0° and 10°, (2) random translation: an image was translated randomly along the x- and y-axes by distances ranging from 0 to 10% of width or height of the image, and (3) random flipping. These data

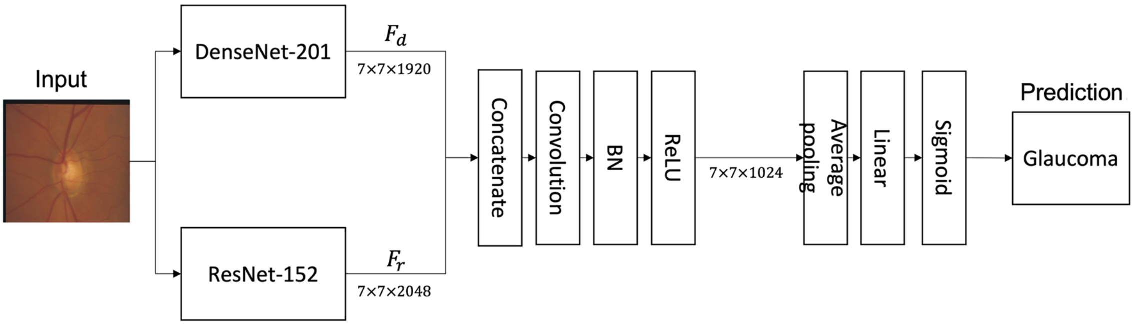

Figure 2. The architecture of the proposed GlaucomaNet.

augmentation operations are crucial in increasing the diversity of the dataset and thus yielding effective and robust representations[24].

Training strategy. We first trained two different single models, DenseNet-201[41] and ResNet-152[42]. Then, we selected the convolutional part of the DenseNet-201and ResNet-152 as the convolutional part of the proposed model. Afterward, we trained the entire network in an end-to-end manner. During the training, we updated the entire network parameters. Therefore, the loss was propagated back to the individual neural networks, creating better feature representations for each training iteration. Since we learned the feature representations from intermediate layers, our training strategy is closed to a joint fusion strategy[47].

To use the pre-trained model on ImageNet, all images were resized to 224 × 224. All models were implemented by Keras with a backend of Tensorflow. The proposed network was optimized using the Adam optimizer method[48]. The learning rate was 5×10^{-5}. The experiments were performed on Intel Core i9-9960 X 16 cores processor and NVIDIA Quadro RTX 6000 GPU.

Ensemble methods. In machine learning, ensemble methods combine multiple models to produce improved results, surpassing single models. This study also evaluated the performance of three ensemble methods: macro averaging, random forest[49], and linear regression. The macro averaging model averages the predicted class probabilities of the seven models together. The random forest model computes a number of decision trees. The final classification corresponds to the majority vote among the individual trees. Here, we used the predicted class probabilities of the seven models and trained the random forest using 500 trees on the development set. For linear regression, we also used the predicted class probabilities of the seven models as input.

Model's clinical rationales. State-of-the-art models for POAG diagnosis, including ours, are now predominantly deep neural networks that are opaque in terms of how they come to make predictions. This limitation has increased interest in designing more interpretable deep models that reveal the 'reasoning' behind model outputs[50]. One common way to explain the deep learning model is through a saliency map where important regions on the images are highlighted to correspond to the disease of interest[22,24,32]. For POAG diagnosis, segmented optic disc and optic cup are typically used. However, such spatial annotation requires substantial effort and is not generally available in large quantities. In addition, such approaches do not establish whether the model relied on these regions to make a prediction.

In this work, we proposed a different way to assess the plausibility of rationales by measuring rationale faithfulness—rationales ought to have meaningfully influenced its prediction[51,52]. Following DeYoung et al.[53], we conducted a contrasting experiment to explore the rationales of the optic disc. Figure 3 provides two demos for the experiment.

First, we manually masked the optic disc r_i from the fundus photograph with a bounding box. Let $m(x_i)$ be the original prediction provided by a model m. We then consider the predicted class from the model once the supporting bounding box is stripped. Intuitively, the model should generate a less "correct" class. We then measure this as comprehensiveness

$$comprehensiveness = \frac{1}{n}\sum\nolimits_{n}(m(x_i) - m(x_i/r_i)).$$

In particular, if we selected glaucomatous images where their labels were predicted by the model correctly ($m(x_i) \equiv 1$), a score of 1.0 indicates that the optic discs (rationale) are indeed influential in the prediction, while 0.0 indicates that the rationales are not the reasons for the prediction.

For comparison, we also randomly masked the image without completely overlapping the optic disc's bounding box. The overlapping ratio (i.e., the intersection of the union (IOU)) is no larger than a threshold. Specifically, we randomly selected the IOU's threshold from 1/2, 2/5, 1/3, and 1/4. The ratio of the area of the original bounding box can also be variant (1, 1/4, 1/9, and 1/16). Finally, we generated ten images with bounding boxes of different thresholds and areas.

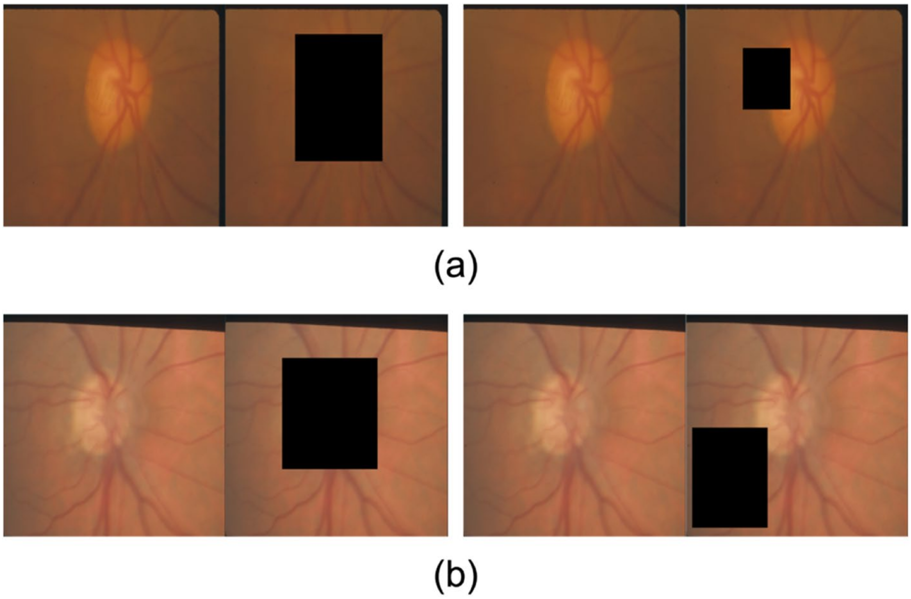

Figure 3. Examples of fundus photographs and their corresponding masked images: (**a**) POAG due to VF only (the left one is based on the optic disc mask and the right one is based on a random mask) and (**b**) POAG due to glaucomatous disc criteria (the left one is based on the optic disc mask and the right one is based on a random mask).

Evaluation metrics.

Our experiments report accuracy, precision, sensitivity (recall), specificity, and F1-score.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, Precision = \frac{TP}{TP + FP}, Sensitivity = \frac{TP}{TP + FN},$$

$$Specificity = \frac{TN}{TN + FP}, F1 - score = 2\frac{precision \bullet recall}{precision + recall}$$

TP, TN, FP, and FP denote true positive, true negative, false positive, and false negative. In addition, we report the AUC (Area Under the Curve) ROC (Receiver Operating Characteristics) curve. A ROC curve plots true positive rate (TPR, also called sensitivity) vs. false positive rate (FPR) at different classification thresholds. These metrics were assessed by the Mann–Whitney U test. The Mann–Whitney U test is associated with a p-value, which is the probability of obtaining a mean difference between two methods as extreme as observed in the test, assuming the null hypothesis is correct (i.e., there is no difference between the results generated by two models).

In this study, we used the fivefold cross-validation to obtain a distribution of the experimental metrics and reported standard deviations. For the OHTS dataset, we split the entire dataset randomly into five groups at the patient level. This ensured that no participant was in more than one group to avoid cross-contamination between the training and test datasets. In each fold of the cross-validation, we took one group (20% of total subjects) as the hold-out test set and the remaining 4 groups as the training set. For the LAG dataset, we conducted fivefold cross-validation using the same strategy.

Results

POAG diagnosis on the OHTS dataset.

The performance was compared with DenseNet-201 and ResNet-152 (Fig. 4) and four state-of-the-art neural networks, VGG-16[54], NASNetMobile[55], Xception[56], MobileV2[57], and ResNet-50[42] (Supplementary Table S1). Our model achieved the best results, with an accuracy of 0.930, an F1-score of 0.490, and an AUC of 0.904. Compared to the best baseline (DenseNet-201), GlaucomaNet has higher accuracy (4.80%), F1-score (7.50%), and AUC (1.70%). The p-value indicates that accuracy, precision, specificity, F1-score, and AUC obtained by GlaucomaNet are statically higher than state-of-the-art neural networks. From Supplementary Table S1, we also found that DenseNet-201 and ResNet-152 achieved better results than other baselines (1.7–14.9%). It is sufficient to compare our model to these two models hereafter.

By combining the different network architectures in a model ensemble, we found that the random forest method improved accuracy to 0.939 and AUC to 0.910 (Fig. 5 and Supplementary Table S2). Meanwhile, the F1-score decreased, leading to reduced sensitivity. However, these changes were not statistically significant.

To evaluate the model's rationales, we randomly selected 100 "correctly predicted" glaucomatous fundus photographs from the OHTS dataset. After stripping the bounding box of optic discs, 97% of images (comprehensiveness score of 97%) were detected as "normal" by GlaucomaNet (Supplementary Table S3). On the contrary, 36% of images (comprehensiveness score of 36%) were detected as "normal" when the random bounding boxes were masked. The results demonstrate that our model is dramatically less confident in its POAG prediction once

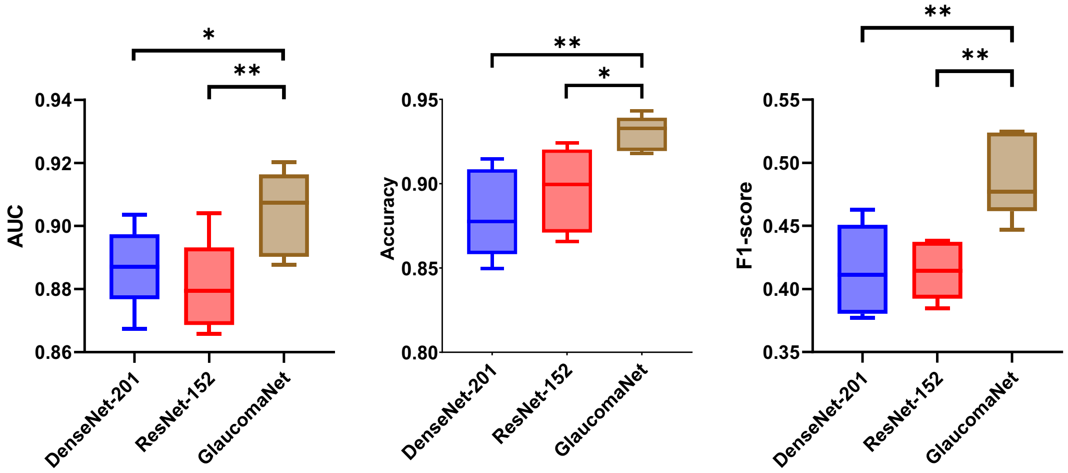

Figure 4. Comparison of different metrics (standard deviation) for different model architectures in the OHTS dataset. p-values are calculated between GlaucomaNet and other models. *p-value ≤ 0.05, **p-value ≤ 0.01.

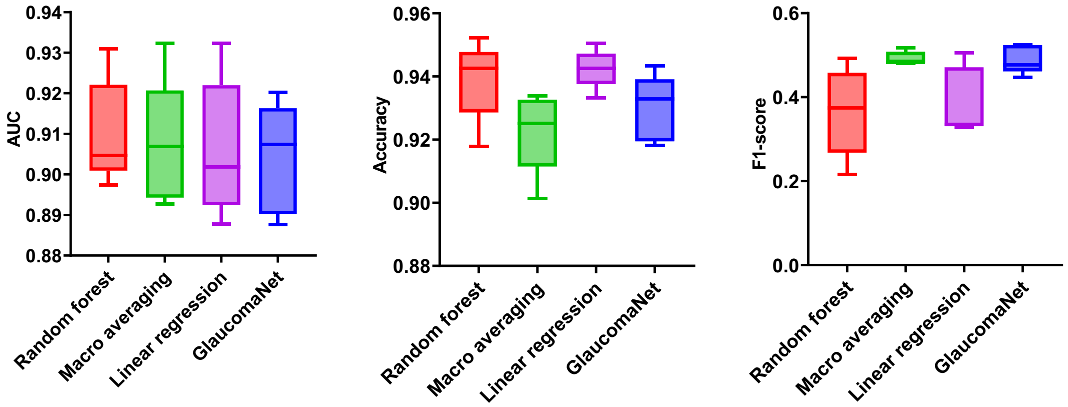

Figure 5. Comparison of different metrics (standard deviation) for proposed and ensemble methods in the OHTS dataset.

the optic discs are masked. We applied DenseNet-201 to the same experimental settings and obtained comprehensiveness scores of 82% and 20%, respectively. The comparison between GlaucomaNet and DenseNet-201 suggests that optic discs are more needed in our model to predict POAG from fundus photographs—that is, our model is more rational than DenseNet.

Supplementary Figure S1 presents more specific results by histograms. The scores for the IOU's thresholds 1/2, 2/5, 1/3, and 1/4 are 40.3%, 38.4%, 33.0%, and 30.7%, and 25.4%, 23.6%, 18.0%, and 12.9% for GlaucomaNet and DenseNet-201, respectively; the scores for the area ratios 1, 1/4, 1/9, and 1/16 are 65.7%, 32.8%, 26.4%, and 21.1%, and 47.0%, 13.4%, 11.7%, and 8.8% for GlaucomaNet and DenseNet-201, respectively.

POAG diagnosis on the LAG dataset. Figure 6 compares the results of GlaucomaNet with DenseNet-201 and ResNet-152 on the LAG dataset. Our model obtained the best results, with an accuracy of 0.969, an F1-score of 0.955, and an AUC of 0.997 (Supplementary Table S4). GlaucomaNet also outperforms the state-of-the-art methods for all evaluation metrics as shown in Supplementary Table S4. In addition, we include the results reported in Li et al.[30], although they used a much larger LAG dataset to train the model. GlaucomaNet achieves superior results even though we only used 41.3% of the LAG dataset for training.

T-distributed stochastic neighbor embedding (t-SNE) method. In this study, the internal features learned by GlaucomaNet were studied using t-distributed Stochastic Neighbor Embedding, which is well suited for visualizing high-dimensional datasets[58]. We first obtained the 1024-dimensional features from the output of averaging pooling layer based on the proposed model. Then, we applied the t-SNE technique to reduce the vector into 2 dimensions for visualization. Figure 7 demonstrates that glaucomatous images can be separated from non-glaucomatous images.

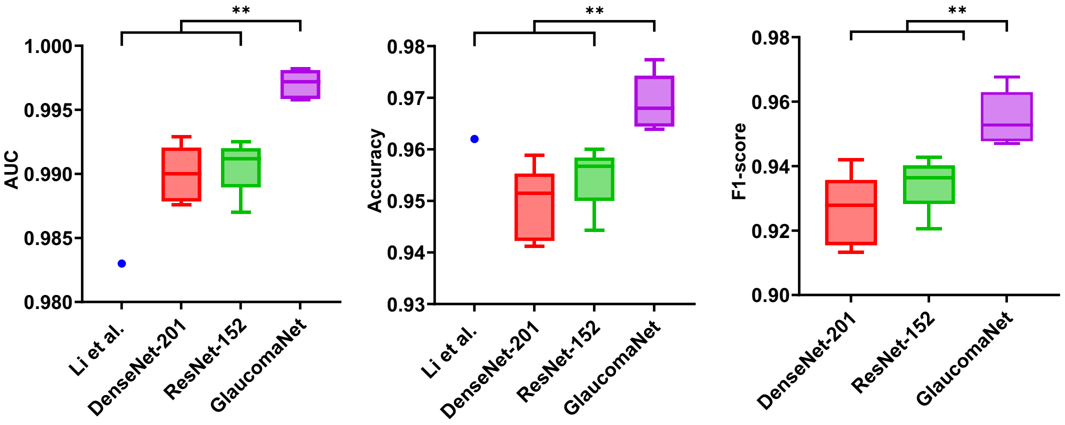

Figure 6. Comparison of different metrics for different model architectures in the LAG dataset. The model reported in Li et al. was trained on 10,928 images, with 4528 having POAG. There is no reported F1-score for Li et al. P-values are calculated between GlaucomaNet and other models. *p-value ≤ 0.05, **p-value ≤ 0.01.

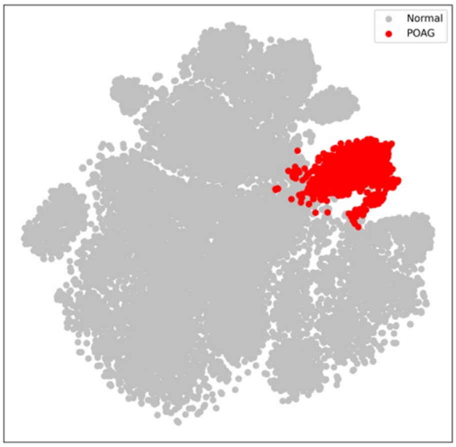

Figure 7. The t-SNE visualization of the proposed model on the OHTS dataset. Each point represents a fundus image. Red and gray dots represent the glaucomatous and non-glaucomatous images.

Discussion

The primary aim of this study is to develop a fully automatic deep learning network for POAG diagnosis using fundus photographs. On two datasets, our proposed model was superior to the state-of-the-art model. Unlike single neural networks such as DenseNet, our method fused the features from two state-of-the-art networks. This simulates the fusion of two readers who first learn the discriminative features separately and then fuse the features for grading.

Our experiment also shows that ensemble methods produce more accurate solutions than a single model. On the OHTS dataset, the random forest achieves the highest AUC. There might be two reasons behind this. First, each tree in the random forests is built from a sample drawn with a replacement from the training set. Second, random forests select a subset of features, rather than all features to train the model. As can be seen from Supplementary Table S2, its variance decreases, resulting in an overall better model.

Of note, although most evaluation metrics of our proposed model were superior, the sensitivity was lower than other models on the OHTS dataset. One potential reason is that the positive (glaucomatous) examples for model training in the OHTS dataset were as low as 11.9% of the total images. We postulate that further training our model with a larger number of POAG images may improve its performance. We also noticed two studies conducted on OHTS. One was performed by Thakur et al.[32], which achieved an AUC of 0.945. There are two main differences between our approach and theirs. First, we used the whole OHTS dataset while Thakur et al.[32] discarded 24% of the fundus photographs with poor image quality. We believe our experimental settings allow us to more closely access the model's real-world clinical performance. Second, we did not manually crop the images in the data preprocessing stage, which minimizes the labor required to deploy our model to healthcare centers with a quick turn-around time and again more closely resembles a real-world workflow. To make a fair comparison, we applied the same method in Thakur et al.[32] to our dataset (MobileV2 in Supplementary Table S1). The AUC was 0.799, lower than the GlaucomaNet under the same setting. The other study was performed by Fan et al.[59], which achieved an AUC of 0.88 based on optic disc or VF change attributable to POAG by the Endpoint Committee (the same criteria as us). We applied their model (ResNet-50) to our data split under the same setting and achieved the AUC of 0.863, which is similar to the result obtained by Fan et al. Because we used different data split and preprocessing methods, the results are not strictly comparable.

We also observed that the performance of GlaucomaNet varies dramatically across two datasets. There are potentially three reasons. First, digital fundus photographs have a higher resolution than digitized film fundus photographs. The LAG dataset contains digital fundus photographs, while OHTS contains digitized film fundus photographs. However, GlaucomaNet can still get an AUC of 0.904 on the OHTS dataset. Second, the images in the LAG datasets were carefully selected by the ophthalmologists, and all glaucomatous images are due to structural glaucomatous optic nerve abnormalities. In contrast, some of POAG fundus photographs in OHTS are due to visual field changes in the absence of glaucomatous disc changes over time. Therefore, the task on LAG is less challenging than OHTS.

Limitation and future work. One limitation of our proposed model arises from the imbalance in the dataset used for its training, particularly the extremely low proportion of images with POAG. As described previously, this likely contributed to the relatively lower sensitivity. However, as evidenced by the results on more balanced LAG, this limitation may be addressed by further training using image datasets with a higher proportion of POAG cases.

Another limitation lies in detecting POAG eyes due to visual field abnormality without any obvious structural sign of Glaucomatous Optic Neuropathy (GON). These images had a higher tendency to be missed by our model. It would be interesting for future studies to differentiate POAG due to VF abnormality and GON.

In conclusion, this study proposed a new end-to-end deep learning network for automatic POAG detection from fundus photographs. Two datasets were used to evaluate the proposed model. The results demonstrated that the proposed network has a good performance on POAG diagnosis. We also analyzed the performance of several ensemble methods; lessons from these approaches may have applicability to developing deep learning models for other retinal diseases, such as diabetic retinopathy, age-related macular degeneration, and even for image-based deep learning systems outside of ophthalmology. Although deep learning models are often considered "black-box" entities, we aimed to improve the transparency of our algorithm by constructing a new dataset by masking the optic disc on the fundus photographs. These "contrasting" examples help us understand if the rationales (optic disc) influence the POAG diagnosis. These efforts to demystify deep learning models may help improve levels of acceptability to patients and adoption by ophthalmologists.

The deep learning model and data partition are publicly available (https://github.com/bionlplab/GlaucomaNet). By making them available, we aim to maximize this study's transparency and reproducibility and provide a benchmark for further refinement and development of the algorithm. In addition, this deep learning model, trained on one of the largest POAG fundus photograph datasets, may allow for future deep learning studies on POAG diagnosis where only smaller datasets are available.

In the future, we aim to improve the model by incorporating multi-modal data. It would also be interesting to compare the model's accuracy with different groups of ophthalmologists (e.g., graders at Reading Centers, glaucoma specialists, general ophthalmologists, and trainee ophthalmologists). These results could then be tested and validated by further studies from different countries and populations. In that case, it is possible that the integration of deep learning models into clinical practice might become increasingly acceptable to patients and ophthalmologists and may ultimately enhance clinical decision-making.

Data availability

The Ocular Hypertension Treatment Study (OHTS) dataset is available upon request due to patient protection (https://ohts.wustl.edu/), please contact the author in[39] (Michael Kas: kass@wustl.edu). The large-scale attention-based glaucoma (LAG) database is available in https://github.com/smilell/AG-CNN (please contact liliu1995@buaa.edu.cn or liliu419@foxmail.com to get the password to download the dataset).

Received: 25 February 2022; Accepted: 30 July 2022
Published online: 18 August 2022

References

1. Bourne, R. R. *et al.* Causes of vision loss worldwide, 1990–2010: a systematic analysis. *Lancet Glob. Health* **1**, e339–e349 (2013).
2. Quigley, H. A. & Broman, A. T. The number of people with glaucoma worldwide in 2010 and 2020. *Br. J. Ophthalmol.* **90**, 262–267 (2006).
3. Sommer, A. *et al.* Racial differences in the cause-specific prevalence of blindness in east Baltimore. *N. Engl. J. Med.* **325**, 1412–1417 (1991).
4. Jiang, X., Torres, M., Varma, R., Group, L. A. L. E. S. Variation in intraocular pressure and the risk of developing open-angle glaucoma: The Los Angeles Latino eye study. *Am. J. Ophthalmol.* **188**, 51–59 (2018).
5. Doshi, V., Ying-Lai, M., Azen, S. P., Varma, R., Group, L. A. L. E. S. Sociodemographic, family history, and lifestyle risk factors for open-angle glaucoma and ocular hypertension: The Los Angeles Latino Eye Study. *Ophthalmology* **115**, 639-647.e632 (2008).
6. Quigley, H. A., Katz, J., Derick, R. J., Gilbert, D. & Sommer, A. An evaluation of optic disc and nerve fiber layer examinations in monitoring progression of early glaucoma damage. *Ophthalmology* **99**, 19–28 (1992).
7. Kolomeyer, N. N. *et al.* Lessons learned from two large community-based glaucoma screening studies. *J. Glaucoma* **21**, 2123 (2021).
8. Singh, A., Dutta, M. K., ParthaSarathi, M., Uher, V. & Burget, R. Image processing based automatic diagnosis of glaucoma using wavelet features of segmented optic disc from fundus image. *Comput. Methods Programs Biomed.* **124**, 108–120 (2016).
9. Acharya, U. R., Dua, S., Du, X. & Chua, C. K. Automated diagnosis of glaucoma using texture and higher order spectra features. *IEEE Trans. Inf. Technol. Biomed.* **15**, 449–455 (2011).
10. Dua, S., Acharya, U. R., Chowriappa, P. & Sree, S. V. Wavelet-based energy features for glaucomatous image classification. *IEEE Trans. Inf Technol. Biomed.* **16**, 80–87 (2011).

11. Issac, A., Sarathi, M. P. & Dutta, M. K. An adaptive threshold based image processing technique for improved glaucoma detection and classification. *Comput. Methods Programs Biomed.* **122**, 229–244 (2015).
12. Lin, M. *et al.* Fully automated segmentation of brain tumor from multiparametric MRI using 3D context deep supervised U-Net. *Med. Phys.* **48**, 4365 (2021).
13. Zhang, Y., Li, X., Lin, M., Chiu, B. & Zhao, M. Deep-recursive residual network for image semantic segmentation. *Neural Comput. Appl.* **32**, 12935–12947 (2020).
14. Tang, Y.-X. *et al.* Automated abnormality classification of chest radiographs using deep convolutional neural networks. *NPJ Digit. Med.* **3**, 1–8 (2020).
15. Wang, Z., Jiang, X., Liu, J., Cheng, K.-T. & Yang, X. Multi-task siamese network for retinal artery/vein separation via deep convolution along vessel. *IEEE Trans. Med. Imaging* **39**, 2904–2919 (2020).
16. Wang, Z. *et al.* Automated detection of clinically significant prostate cancer in mp-MRI images based on an end-to-end deep neural network. *IEEE Trans. Med. Imaging* **37**, 1127–1139 (2018).
17. Lin, M. *et al.* Cascaded triplanar autoencoder m-net for fully automatic segmentation of left ventricle myocardial scar from three-dimensional late gadolinium-enhanced mr images. *IEEE Journal of Biomedical and Health Informatics*. **26**(6), 2582–2593 (2022).
18. Wanyan, T. *et al.* Supervised pretraining through contrastive categorical positive samplings to improve COVID-19 mortality prediction. *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. (2022).
19. Gulshan, V. *et al.* Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).
20. Peng, Y. *et al.* Predicting risk of late age-related macular degeneration using deep learning. *NPJ Digit. Med.* **3**, 1–10 (2020).
21. Keenan, T. D. *et al.* Deep learning automated detection of reticular pseudodrusen from fundus autofluorescence images or color fundus photographs in AREDS2. *Ophthalmology* **127**, 1674–1687 (2020).
22. Peng, Y. *et al.* DeepSeeNet: A deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs. *Ophthalmology* **126**, 565–575 (2019).
23. Keenan, T. D. *et al.* A deep learning approach for automated detection of geographic atrophy from color fundus photographs. *Ophthalmology* **126**, 1533–1540 (2019).
24. Grassmann, F. *et al.* A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. *Ophthalmology* **125**, 1410–1420 (2018).
25. Poplin, R. *et al.* Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* **2**, 158–164 (2018).
26. Lin, M, *et al.* Artificial intelligence in tumor subregion analysis based on medical imaging: A review. *J. Appl. Clin. Med. Phys.* **22**(7), 10–26 (2021).
27. Ghahramani, G. *et al.* Multi-task deep learning-based survival analysis on the prognosis of late AMD using the longitudinal data in AREDS. *AMIA Annual Symposium Proceedings.* Vol. **2021**. (American Medical Informatics Association, 2021).
28. Chen, X. *et al.* Glaucoma detection based on deep convolutional neural network. *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).* 715–718 (2015).
29. Raghavendra, U. *et al.* Deep convolution neural network for accurate diagnosis of glaucoma using digital fundus images. *Inf. Sci.* **441**, 41–49 (2018).
30. Li, L. *et al.* A large-scale database and a CNN model for attention-based glaucoma detection. *IEEE Trans. Med. Imaging* **39**, 413–424 (2019).
31. Li, Z. *et al.* Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology* **125**, 1199–1206 (2018).
32. Thakur, A. *et al.* Predicting glaucoma before onset using deep learning. *Ophthalmol. Glaucoma* **3**, 262–268 (2020).
33. Christopher, M. *et al.* Performance of deep learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs. *Sci. Rep.* **8**, 1–13 (2018).
34. Fu, H. *et al.* Disc-aware ensemble network for glaucoma screening from fundus image. *IEEE Trans. Med. Imaging* **37**, 2493–2501 (2018).
35. Li, A. *et al.* Integrating holistic and local deep features for glaucoma classification. *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).* 1328–1331 (2016).
36. Yang, X. *et al.* Co-trained convolutional neural networks for automated detection of prostate cancer in multi-parametric MRI. *Med. Image Anal.* **42**, 212–227 (2017).
37. Zoabi, Y., Deri-Rozov, S. & Shomron, N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *npj Digit. Med.* **4**, 1–5 (2021).
38. Monaghan, C. *et al.* Machine Learning for Prediction of Patients on Hemodialysis with an Undetected SARS-CoV-2 Infection. *Kidney360* **13**, 456–468 (2021).
39. Kass, M. *et al.* The ocular hypertension treatment study: A randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma. *Arch. Ophthalmol* **120**, 701–713 (2002).
40. Li, L. *et al.* Attention Based Glaucoma Detection: A Large-Scale Database and CNN Model. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 10571–10580 (2019).
41. Huang, G. *et al.* Densely Connected Convolutional Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 4700–4708 (2017).
42. He, K. *et al.* Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 770–778 (2016).
43. Ioffe, S. *et al.* Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *International Conference on Machine Learning.* 448–456 (2015).
44. Glorot, X. *et al.* Deep Sparse Rectifier Neural Networks. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics.* 315–323 (2011).
45. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
46. Ho, Y. & Wookey, S. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE Access* **8**, 4806–4813 (2019).
47. Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I. & Lungren, M. P. Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines. *NPJ Digit. Med.* **3**, 1–9 (2020).
48. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint* arXiv:1412.6980 (2014).
49. Ho, T. K. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**, 832–844 (1998).
50. Hou, B.-J. & Zhou, Z.-H. Learning with interpretable structure from gated RNN. *IEEE Trans. Neural Netw. Learn. Syst.* **31**, 2267–2279 (2020).
51. Zaidan, O. *et al.* Using "annotator rationales" to improve machine learning for text categorization. *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference.* 260–267 (2007).

52. Yu, M., Chang, S., Zhang, Y. & Jaakkola, T. S. Rethinking cooperative rationalization: Introspective extraction and complement control. *arXiv preprint* arXiv:1910.13294 *(2019).*
53. DeYoung, J. *et al.* ERASER: A benchmark to evaluate rationalized NLP models. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* arXiv:1911.03429. 4443–4458 (2019).
54. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint* arXiv:1409.1556 *(2014).*
55. Zoph, B. *et al.* Learning Transferable Architectures for Scalable Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 8697–8710 (2018).
56. Chollet, F. *et al.* Xception: Deep Learning with Depthwise Separable Convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 1251–1258 (2017).
57. Sandler, M. *et al.* MobileNetV2: Inverted Residuals and Linear Bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 4510–4520 (2018).
58. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 11 (2008).
59. Fan, R. *et al.* Detecting glaucoma in the ocular hypertension study using deep learning. *JAMA Ophthalmol.* **140**, 383–391 (2022).

Acknowledgements

This project was supported by the National Library of Medicine under award number 4R00LM013001. This work was also supported by awards from the National Eye Institute, the National Center on Minority Health and Health Disparities, National Institutes of Health (grants EY09341, EY09307), Horncrest Foundation, awards to the Department of Ophthalmology and Visual Sciences at Washington University, the NIH Vision Core Grant P30 EY 02687, Merck Research Laboratories, Pfizer, Inc., White House Station, New Jersey, and unrestricted grants from Research to Prevent Blindness, Inc., New York, NY.

Author contributions

M.L. implemented the methods, conducted the experiments and wrote the paper. Y.P. advised on all aspects of the work involved in this project and assisted in the paper writing. S.H.V.T. provided knowledge about POAG, suggestions for experiments, and edited the paper. F.W. advised the overall direction of the project and edited the paper. B.H. advised the methods and edited the paper. L.L., M.G., and M.K. provided ideas about the POAG detection based on the dataset and edited the paper. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-17753-4.

Correspondence and requests for materials should be addressed to F.W., S.H.T. or Y.P.