

# Glaucoma detection: Binocular approach and clinical data in machine learning

Oleksandr Kovalyk-Borodyak<sup>ID\*</sup>, Juan Morales-Sánchez<sup>ID\*</sup>, Rafael Verdú-Monedero<sup>ID\*</sup>,  
José-Luis Sancho-Gómez<sup>ID</sup>

Universidad Politécnica de Cartagena, 30202 Cartagena, Spain

## ARTICLE INFO

### Keywords:

Glaucoma diagnosis  
Ocular fundus imaging  
CNN  
Deep learning  
Gradient-boosting decision trees  
Both eyes  
Clinical data  
SHAP

## ABSTRACT

In this work, we present a multi-modal machine learning method to automate early glaucoma diagnosis. The proposed methodology introduces two novel aspects for automated diagnosis not previously explored in the literature: simultaneous use of ocular fundus images from both eyes and integration with the patient's additional clinical data. We begin by establishing a baseline, termed *monocular mode*, which adheres to the traditional approach of considering the data from each eye as a separate instance. We then explore the *binocular mode*, investigating how combining information from both eyes of the same patient can enhance glaucoma diagnosis accuracy. This exploration employs the PAPILA dataset, comprising information from both eyes, clinical data, ocular fundus images, and expert segmentation of these images. Additionally, we compare two image-derived data modalities: direct ocular fundus images and morphological data from manual expert segmentation. Our method integrates Gradient-Boosted Decision Trees (GBDT) and Convolutional Neural Networks (CNN), specifically focusing on the MobileNet, VGG16, ResNet-50, and Inception models. SHAP values are used to interpret GBDT models, while the Deep Explainer method is applied in conjunction with SHAP to analyze the outputs of convolutional-based models. Our findings show the viability of considering both eyes, which improves the model performance. The binocular approach, incorporating information from morphological and clinical data yielded an AUC of 0.796 ( $\pm 0.003$  at a 95% confidence interval), while the CNN, using the same approach (both eyes), achieved an AUC of 0.764 ( $\pm 0.005$  at a 95% confidence interval).

## 1. Introduction

Glaucoma is the leading cause of irreversible blindness worldwide [1–5]. The risk of glaucoma increases with age [6], affecting one in 200 people under the age of 50 and one in 10 people over the age of 80. Due to the latest medical advances, human life expectancy has increased, meaning that glaucoma is expected to become a major public health problem. It is estimated that, in 2040, 111.8 million people between 40–80 years old will suffer from glaucoma [1]. Glaucoma is a disease that affects the Optic Nerve Head (ONH). It is characterized by a progressive and irreversible loss of the visual field, usually caused by high intraocular pressure. In the later stages of the disease, the visual field decreases and, if left unattended, the disease can cause severe damage to the visual field and even total blindness [7].

Given that there is no cure for the damage glaucoma causes, early detection is crucial to minimize the risk of vision loss. At present, the early and accurate detection of glaucoma poses numerous challenges. First, experts are required to perform adequate tests and evaluate the results. Secondly, glaucoma is a silent, initially asymptomatic disease,

often detected only in its advanced stages. It is estimated that 90% of glaucoma patients in the world are undiagnosed [8–12]. Therefore, developing new tools that improve the efficiency of the current diagnosis methods is vital for the early detection of glaucoma.

Although glaucoma is asymptomatic, indicators of the disease appear years before damage to the visual field occurs. The main procedures for the diagnosis of glaucoma are tonometry, which measures Intraocular Pressure (IOP), campimetry for the study of the visual field, Optical Coherence Tomography (OCT) to study the thickness of retinal nerve fiber layers and the Ocular Fundus Image (OFI) to study the morphology of the optic nerve. Both campimetry and optical tonometry are very expensive tests in terms of time and resources. On the other hand, the OFI presents several advantages over tonometry and campimetry, such as lower cost, and fast and non-invasive testing. Ophthalmologists rely on features extracted from the OFIs (Fig. 1(a)), such as the Cup-to-Disc Ratio (CDR) [13], the Rim-to-Disc Ratio (RDR) [14], and the

\* Corresponding authors.

E-mail addresses: [oleksandr.kovalyk@edu.upct.es](mailto:oleksandr.kovalyk@edu.upct.es) (O. Kovalyk-Borodyak), [juan.morales@upct.es](mailto:juan.morales@upct.es) (J. Morales-Sánchez), [rafael.verdu@upct.es](mailto:rafael.verdu@upct.es) (R. Verdú-Monedero), [josel.sancho@upct.es](mailto:josel.sancho@upct.es) (J.-L. Sancho-Gómez).

<https://doi.org/10.1016/j.artmed.2024.103050>

Received 5 April 2024; Received in revised form 28 November 2024; Accepted 5 December 2024

Available online 12 December 2024

0933-3657/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).

Inferior (I), Superior (S), Nasal (N), and Temporal (T) parameters, to assess and diagnose conditions. The CDR is analyzed by segmenting<sup>1</sup> the disc and the cup in the image, as shown in Fig. 1(b). The smaller the difference between the cup and the disc sizes, the greater the potential damage from glaucoma, as the condition typically causes the cup to enlarge. The RDR is defined as the minimum thickness of the neuroretinal rim relative to the disc diameter [14]. The smaller the relative thickness of the rim, the greater the deformation of the ONH. The ISNT rule, based on measurements of the neuroretinal rim thickness in four directions (Inferior, Superior, Nasal, and Temporal), typically satisfies the condition  $I > S > N > T$  in healthy individuals [15,16]. In clinical practice, ophthalmologists often perform both the segmentation and the calculation of the CDR and ISNT parameters manually, a process that can be time-consuming.

Due to all of the above, numerous machine learning-based techniques that utilize OFIs for diagnosis employ manual or automatic segmentation proposals to consider, implicitly or explicitly, parameters such as ISNT, CDR, and RDR [17–22]. These approaches are generally known as feature extraction methods, and include a second stage that performs an automatic glaucoma diagnosis. However, the current trend is to take advantage of Deep Learning (DL) capabilities to achieve diagnosis in a single stage, in proposals known as end-to-end methods [23–28].

On the other hand, it should be noted that in many cases ophthalmologists resort to complementary tests to confirm a diagnosis, such as OCT or other clinical information. For this reason, recent publications incorporate a variety of data sources, including OCT [29,30] and clinical data [31,32], to improve the robustness of the overall method. However, all the aforementioned articles use per-eye-data as an instance to train the machine learning models. Furthermore, the hypothesis regarding asymmetry between the anatomical characteristics of the optic nerve in both eyes has recently been revisited as an indicator for early diagnosis of glaucoma [33–36].

For a more extensive review of state-of-the-art automatic techniques for glaucoma diagnosis, we can refer to a recent survey [37], which compiles and classifies a large number of current relevant machine learning methods. The authors categorize feature extraction techniques into three main types: structural features, statistical features, and hybrid features. In the structural features category, techniques range from adaptive threshold frameworks and K-means clustering to advanced DL models like Faster R-CNN and M-Net. Moreover, the categories in which the authors classify end-to-end neural networks designed for glaucoma prediction are convolutional-based, autoencoder-based, attention-based, generative adversarial, geometric and hybrid.

In this study, we tackle the following three main objectives, particularly in the context of automatic early glaucoma detection: (1) the examination of the simultaneous contribution of data from both eyes, (2) the evaluation of the benefit of clinical information as supplementary data on OFI-derived features, and (3) a comparative analysis between end-to-end DL approaches and machine methods fed with extracted morphological features in the context of glaucoma assessment. For this purpose, the PAPILA [38] dataset has been used, an open access glaucoma dataset which contains clinical data and OFIs, including the segmentation of the optic disc and the optic cup depicted by two different experts. PAPILA is a population sample of glaucoma prevalence in the Region of Murcia, Spain, which permits us to evaluate the method in a real scenario. Based on the aforementioned hypothesis, our study aims to investigate the advantage of including data from both eyes of the same patient for the diagnosis of glaucoma, with a specific focus on its impact on early-stage glaucoma, which is deemed the most critical of the three objectives. Additionally, we seek to explore whether clinical data can offer novel insights when combined with morphological and OFI extracted data, and to quantify the significance of each type of data.

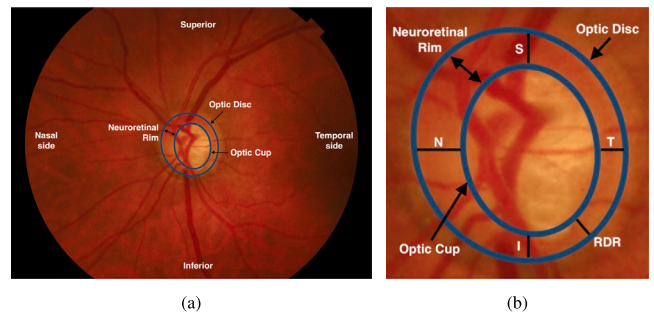


Fig. 1. Fundus image with the region of interest (ROI) around the optic disc. (a) Ocular Fundus Image (OFI). (b) ROI extracted from an OFI.

## 2. Materials and methods

### 2.1. Dataset selection criteria

There are a variety of datasets for glaucoma diagnosis, such as RIGA [39], ORIGA [40], RIMONE [41,42], PAPILA [38], Drishti-GS [43], ACRIMA [44], G1020 [45], and REFUGE [46], among others. All these datasets are publicly available, but only one of them meets the necessary requirements for this study (see Table 2). RIGA does not present a glaucoma diagnosis label, ACRIMA lacks disc and cup segmentations, and none of them contain information from both eyes or clinical data. Although it is possible to use other OFI datasets to (pre)train the OFI-based models, several issues emerge in practice. The most important one is that we aim to detect early glaucoma, and it is clear (see Table 1) that the PAPILA dataset is biased toward early-stage glaucoma. Severe and mid-stage glaucoma show significant VF/MD and ONH alterations that can be easily detected by machine and deep learning algorithms. Including other datasets with a high separability between severe glaucoma and healthy instances could introduce bias into the study. In addition, we found that adding other datasets during the training process did not show any improvement over the results when testing with the PAPILA test set.

### 2.2. Dataset

PAPILA [38] is a free and publicly available dataset that includes OFIs, optic disc and optic cup segmentations delineated by two experts, the diagnosis labels for the glaucoma disease, and additional data coming from patient clinical tests. The instances of the dataset are divided into three categories: *healthy*, *suspicious* and *glaucoma*. The PAPILA dataset is a population sample, which allows the performance of the machine learning models to be analyzed in a realistic scenario. The dataset is primarily composed of information on early stages of glaucoma, including patients of different ages and genders. Both data (OFIs with their corresponding segmentation and clinical data) and labels are provided for both eyes of the same patient. The dataset includes information from 244 patients, with a total of 488 samples (one for each eye), where 333 samples correspond to healthy eyes, 87 to eyes diagnosed with glaucoma, and the remaining 68 samples were suspected of glaucoma. Every sample includes clinical data, OFI and two expert segmentations of the optic disc and the optic cup. OFIs are centered on the papilla with a field of view of 30 degrees, and were acquired with a size of  $2576 \times 1934$  pixels and a 24-bit color depth. The images are provided in JPEG RGB format. In this work, we selected patients with healthy and glaucoma diagnoses, discarding the *suspicious* category because it is considered an unresolved situation that mixes both previous classes. Regarding the clinical data, among all the information that PAPILA offers, we selected the following: age (Ag), gender (G), refractive defect (D1, D2, A), phakic/pseudophakic (PPs), intraocular pressure (IOP), pachymetry (P), and axial length

<sup>1</sup> Image segmentation is the process of partitioning an image into distinct regions, or sets of pixels, to analyze specific areas separately.

**Table 1**  
Number of glaucomatous instances in the PAPILA dataset, distributed by glaucoma grade.

|                    | Early Glaucoma     |                                 | Mid Glaucoma                        | Severe Glaucoma          |
|--------------------|--------------------|---------------------------------|-------------------------------------|--------------------------|
| VF/MD intervals    | $VF/MD \geq -3$ dB | $-3 \text{ dB} > VF/MD \geq -6$ | $-6 \text{ dB} > VF/MD \geq -12$ dB | $-12 \text{ dB} > VF/MD$ |
| Number of eyes     | 32                 | 18                              | 19                                  | 18                       |
| Number of patients | 9                  | 9                               | 13                                  | 16                       |

**Table 2**  
Summary of publicly available datasets.

| Dataset            | Instance number | Ground truth labels |                       |                      | Clinical data | Both eyes<br>of the same<br>patient |
|--------------------|-----------------|---------------------|-----------------------|----------------------|---------------|-------------------------------------|
|                    |                 | Glaucoma<br>label   | Optic disc<br>contour | Optic cup<br>contour |               |                                     |
| RIGA [39,47]       | 750             | ✓                   | ✓                     | ✓                    | ✗             | ✗                                   |
| ORIGA [40]         | 650             | ✓                   | ✓                     | ✓                    | ✗             | ✗                                   |
| RIMONE [42,48]     | 485             | ✓                   | ✓                     | ✓                    | ✗             | ✗                                   |
| Drishti-GS [43,49] | 101             | ✓                   | ✓                     | ✓                    | ✗             | ✗                                   |
| ACRIMA [44,50]     | 705             | ✓                   | ✗                     | ✗                    | ✗             | ✗                                   |
| G1020 [45,51]      | 1020            | ✓                   | ✓                     | ✓                    | ✗             | ✗                                   |
| REFUGE [46,52]     | 1200            | ✓                   | ✓                     | ✓                    | ✗             | ✗                                   |
| PAPILA [38]        | 488             | ✓                   | ✓                     | ✓                    | ✓             | ✓                                   |

(AL). Although the PAPILA dataset also provides the Visual Field/Mean Defect (VF/MD) value, we excluded it from the experiments, as this clinical test was not performed for healthy patients, and accordingly the feature is not present in the dataset. Therefore, using VF/MD would introduce significant bias in the learning process.

### 2.2.1. Data conditioning and augmentation

For all data sources (OFI, clinical data, and morphological data extracted from segmentations) the same split strategy was used. The dataset was split by patient using the  $k$ -fold method with  $k = 5$  folds, ensuring that samples corresponding to both eyes of the same patient were always in the same set (train or test). This process was repeated 100 times, plus one additional iteration for hyperparameter tuning, with the overall process resulting in a total of 101 different splits to compute the metrics at a narrow 95% confidence interval.

Since the gradient-boosted method requires no special data preprocessing, the input data were left in a regular and tabular structure.

In the case of OFIs, the images were cropped to remove a few surrounding pixels and to obtain a square region of interest (ROI), that was resized to  $299 \times 299 \times 3$  and normalized in a range  $[0, 1]$ .

Regarding the expert segmentations, the relevant morphological features that medical experts use for characterizing the ONH (CDRA, CDRV, CDRH, RDR, I, S, N, and T) were automatically computed from the masks of the optical disc and optical cup.

Additionally, data augmentation was performed on the OFI data during training, including random flips, rotations, and zooms. No data augmentation was performed on the tabular data.

### 2.3. Machine learning models and training procedures

To fulfill the proposed research goals, a complete set of experiments was designed, consisting of two main parts (see Fig. 2): the monocular approach, which will serve as a reference baseline, and the binocular approach. In both cases, from the data sources point of view, the experiments could be categorized into three types: tabular, imaging, and data-ensemble.

In the monocular strategy data from each eye were considered as an input instance to the model for all three experiments, while for the binocular scenario, the information from both eyes of a single patient was considered as a unified input instance from the model perspective. Therefore, the result of the monocular approach set of experiments will serve as the baseline against which to compare the results obtained from the binocular approach.

In addition, to study the impact of clinical tests on glaucoma diagnosis, a data-ensemble dataset that combines clinical information with morphological and/or OFI data was curated. The aforementioned

experiments were used to compare the performance of clinical experts and deep learning algorithms in early-glaucoma diagnosis, considering both monocular and binocular approaches. In this context, the morphological data represent expert medical insights, derived from ONH imaging by segmenting the optical disc and optical cup. A gradient-boosted algorithm was employed to predict glaucoma scores from morphological data. In contrast, for deep learning algorithms, the input was the pure OFI. Finally, to improve our understanding of the results and identify the most influential data features, the SHapley Additive exPlanations (SHAP) technique [53] was used. Further details about experiments and models used for each experiment will be provided in the following sections.

#### Model fine-tuning and hardware

All models were fine-tuned using the Ray Python library [54]. Our tests were conducted on an RTX 3090 with 24 GB of VRAM, powered by an AMD EPYC 7643 processor with 256 GB of RAM.

#### Model selection criteria

Two distinct approaches were employed to effectively handle the diverse nature of the data from the PAPILA dataset, which includes both tabular data (clinical data, morphological data, and data-ensemble) and image data (OFI).

For the tabular data, a Gradient-Boosted Decision Tree (GBDT) algorithm was used, and for the image data, diverse CNN-based architectures such as ResNet-50, Inception, MobileNet and VGG16 were explored. All the CNN models were implemented in PyTorch, using the Python programming language.

**Tabular data:** The GBDT algorithm often provides state-of-the-art performance on tabular data [55]. Also, GBDT is known for being highly interpretable due to the simple structure of decision trees and the greedy nature of the gradient-boosting algorithm. This allows for easy understanding of how the model is making its predictions and the ability to extract the criteria used to classify instances, which is very important in medical problems. For the implementation of the GBDT algorithm, the XGBoost [56] Python library was used.

**Image data:** CNN-based models have proven to be highly effective in medical image classification, consistently achieving competitive results in a wide variety of medical imaging classification tasks. We investigate CNN-based algorithms that incorporate significant architectural modifications to enhance their performance in this field:

- VGG16 [57]: it is known for its simplicity and depth (16 layers).
- ResNet-50 [58]: it stands out for its ability to effectively train very deep neural networks using residual connections.

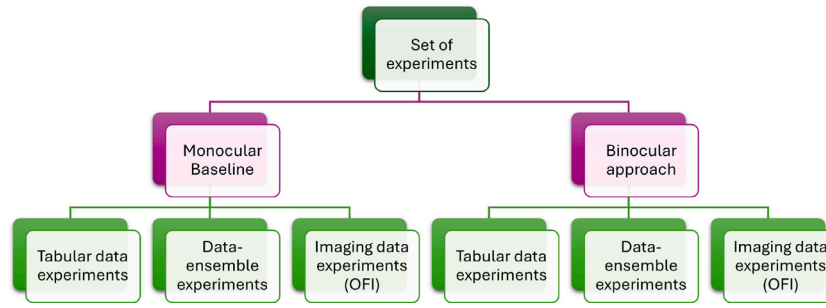


Fig. 2. Diagram of the overall experiment set.

- Inception [59]: use of the Inception modules, which are blocks of layers that perform multiple different types of convolutions in parallel.
- MobileNet [60]: designed to be efficient on mobile and embedded devices. It uses depth-wise separable convolutions to reduce the number of parameters and computational requirements.

### 2.3.1. Monocular approach

Firstly, several models were trained using a traditional approach, in which the data from each eye represents an individual instance within the dataset. We refer to this as the monocular approach or single-eye mode. This will serve as a baseline for further comparison with the binocular approach, to determine the benefits of using the data from both eyes simultaneously. The same splits were used for each mode to ensure comparability between the monocular and binocular modes. In the monocular mode, we randomly dropped one eye's information for each patient, with a probability  $p = 0.5$ , ensuring that the number of instances remained the same for both monocular and binocular modes. This random drop was applied across all data sources in the monocular mode test regime. Additionally, the random seed (one seed per fold) of this drop process was saved to further ensure data compatibility for the ensemble mode. The models can be categorized into three groups depending on the data source they are fed from: models fed by the tabular dataset, which includes clinical and morphological data; models fed by OFIs; and models fed by *data-ensemble*, that combine clinical data with morphological data in one hand and the clinical data with OFI features in the other. The configuration of each model is reviewed below.

**Tabular data models baseline configurations:** For the tabular data in monocular mode, we constructed three GBDT models. One model is dedicated to isolated clinical data, while the two remaining models are for morphological data, with one model per expert. The algorithm was configured without any constraints on the number of nodes or the maximum delta step allowed for each leaf output. No subsampling of the training instances was applied. Finally, from the class frequencies, the optimization process was weighted with a ratio of 3.47. These hyperparameters will remain the same for the rest of the of GBDT-based models. The rest of the hyperparameters are as follows:

- *Clinical model*: colsample\_bylevel 1, learning\_rate 0.002, max\_depth 13, gamma 0.16, alpha 1, lambda 0.12, colsample\_bytree 0.55 and min\_child\_weight 0.
- *Expert 1 model*: colsample\_bylevel 1, learning\_rate 0.06, max\_depth 13, gamma 1.11, alpha 1, lambda 0.33, colsample\_bytree 0.62 and min\_child\_weight 0.
- *Expert 2 model*: colsample\_bylevel 1, learning\_rate 0.024, max\_depth 15, gamma 2.42, alpha 1, lambda 0.19, colsample\_bytree 0.93 and min\_child\_weight 2.

**Imaging data models baseline configurations:** As mentioned above, four different CNN models were tested. In addition, pre-trained weights from *Imagenet* [61] were used to accelerate the training process. For each

model, the default output layer was removed to obtain a raw CNN head with an output dimension of  $1 \times L_M$ , which works as a feature extractor.  $L_M$  represents the latent dimension length for a model  $M$ , and varies depending on the model. Then, an *Multi-Layer Perceptron* (MLP) of variable hyperparameters was fed with the output of the previous CNN head, allowing us to adjust the number of hidden layers and the neural units within each MLP layer. The number of neural units in the first MLP layer depends on the length  $L_M$ , and the number of units in the output layer matches the number of classes (two in this binary classification scenario).

Each model was trained minimizing the standard cost function for image classification, the *Cross Entropy Loss*, and using *AdamW* [62] optimizer. Additionally, to address class imbalance, the sample loss is weighted inversely proportional to the frequency of each class. The configuration employed for each model is shown below:

- *VGG16*: batch size 8, number of CNN frozen layers 10, label smoothing 0.0026, learning rate  $2.58 \cdot 10^{-5}$ , weight decay  $2.52 \cdot 10^{-6}$ , drop out 0.17, MLP hidden units 512, MLP hidden layers 3, and epochs 20.
- *ResNet-50*: batch size 2, number of CNN frozen layers 40, label smoothing 0.15, learning rate  $2.26 \cdot 10^{-5}$ , weight decay  $1.47 \cdot 10^{-6}$ , drop out 0.3, MLP hidden units 512, MLP hidden layers 2, and epochs 10.
- *Inception*: batch size 2, number of CNN frozen layers 50, label smoothing 0.11, learning rate  $6.10 \cdot 10^{-5}$ , weight decay  $3.06 \cdot 10^{-6}$ , drop out 0.25, MLP hidden units 256, MLP hidden layers 3, and epochs 20.
- *MobileNet*: batch size 16, number of CNN frozen layers 10, label smoothing 0.19, learning rate  $8.8 \cdot 10^{-5}$ , weight decay  $1.5 \cdot 10^{-6}$ , drop out 0.22, MLP hidden units 128, MLP hidden layers 2, and epochs 10.

**Data-ensemble baseline models configurations:** Finally, to assess the contribution of clinical data, two strategies are adopted. On one hand, the clinical and morphological data from both experts were combined, and on the other hand, the clinical data are used jointly with the non-softmax output (without applying the softmax activation function) of the best-performing CNN model.

For the integration of clinical and morphological data, we concatenated some feature columns from both datasets, resulting in two different and specific datasets: *Clinical plus Morphological Expert 1* and *Clinical plus Morphological Expert 2*, each one comprising a total of 16 feature columns (eight from clinical data and eight from morphological data).

To merge the clinical data with the extracted OFI features, the output of the last layer of the CNN was taken before the softmax was applied, incorporating them into the clinical dataset as two new columns: OFI-G for the glaucoma score, and OFI-H for the health score. Although softmax probabilities were first considered for this purpose, experiments showed that GBDT models yield better performance when non-softmax output is used.

The hyperparameters configuration is as follows:



- *Clinical plus Morphological Expert 1*: colsample\_bylevel 1, learning\_rate 0.055, max\_depth 17, gamma 0.78, alpha 1, lambda 0.23, colsample\_bytree 0.61 and min\_child\_weight 4.
- *Clinical plus Morphological Expert 2*: colsample\_bylevel 1, learning\_rate 0.058, max\_depth 14, gamma 0.28, alpha 2, lambda 0.61, colsample\_bytree 0.90 and min\_child\_weight 0.
- *Clinical plus CNN-OFI*: colsample\_bylevel 1, learning\_rate 0.044, max\_depth 16, gamma 2.42, alpha 1, lambda 0.23, colsample\_bytree 0.50 and min\_child\_weight 1.

### 2.3.2. Binocular approach

In the binocular mode, we considered information from both eyes as a single input instance for the model. In this approach, an equivalent number of models were constructed, categorizing them into the same three groups: tabular data (clinical and morphological), imaging data (OFIs), and data-ensemble. Below, we further detail how data is processed across these categories.

**Tabular models configurations:** For tabular data, the GBDT algorithm was also used. To incorporate data from both eyes of the same patient, the information was concatenated column-wise. As with the monocular approach, three models were built: one for clinical data and two for morphological data, one per expert. The objective function and hyperparameters remain unchanged from the monocular approach. The remaining configuration details are as follows:

- *Clinical model*: colsample\_bylevel 1, learning\_rate 0.073, max\_depth 11, gamma 1.75, alpha 1, lambda 0.024, colsample\_bytree 0.58 and min\_child\_weight 0.
- *Expert 1 model*: colsample\_bylevel 1, learning\_rate 0.09, max\_depth 8, gamma 0.081, alpha 3, lambda 0.91, colsample\_bytree 0.81 and min\_child\_weight 1.
- *Expert 2 model*: colsample\_bylevel 1, learning\_rate 0.058, max\_depth 16, gamma 0.01, alpha 4, lambda 0.75, colsample\_bytree 0.91 and min\_child\_weight 0.

**Imaging models configurations:** To enable the model to access data from both eyes simultaneously, a model was built that allows for an input of two images at once. Internally, a sequential and individual feature extraction from each image was performed. The output features of each image were then concatenated and processed through an MLP, with an output layer of two neurons. Data augmentation in this approach was individually applied to each image; that is, all mentioned augmentations were applied randomly to each image, rather than to both images simultaneously. Like the monocular approach, we leverage the model's weights from *ImageNet*. Each model was configured as follows:

- *VGG16*: batch size 8, number of CNN frozen layers 50, label smoothing 0.0015, learning rate  $3.67 \cdot 10^{-5}$ , weight decay  $9.88 \cdot 10^{-6}$ , drop out 0.36, MLP hidden units 512, MLP hidden layers 2, and epochs 50.
- *ResNet-50*: batch size 4, number of CNN frozen layers 20, label smoothing 0.14, learning rate  $1.28 \cdot 10^{-5}$ , weight decay  $5.51 \cdot 10^{-6}$ , drop out 0.24, MLP hidden units 512, MLP hidden layers 1, and epochs 10.
- *Inception*: batch size 8, number of CNN frozen layers 230, label smoothing 0.0, learning rate  $1.19 \cdot 10^{-5}$ , weight decay  $9.88 \cdot 10^{-6}$ , drop out 0.43, MLP hidden units 1024, MLP hidden layers 1, and epochs 30.
- *MobileNet*: batch size 2, number of CNN frozen layers 20, label smoothing 0.003, learning rate  $1.55 \cdot 10^{-5}$ , weight decay  $5.35 \cdot 10^{-6}$ , drop out 0.11, MLP hidden units 1024, MLP hidden layers 1, and epochs 40.

**Data-ensemble models configurations:** In the ensemble mode, we adhered to the same strategy employed in the monocular mode. The morphological data were concatenated with the clinical data by columns.

When combining clinical data with OFI-extracted features, two additional columns, which correspond to the CNN model outputs for health and glaucoma scores, were added to the clinical data. The configuration for each model is outlined as follows:

- *Clinical plus Morphological Expert 1*: colsample\_bylevel 1, learning\_rate 0.091, max\_depth 8, gamma 0.96, alpha 1, lambda 0.22, colsample\_bytree 0.56 and min\_child\_weight 4.
- *Clinical plus Morphological Expert 2*: colsample\_bylevel 1, learning\_rate 0.061, max\_depth 12, gamma 0.071, alpha 3, lambda 0.34, colsample\_bytree 0.65 and min\_child\_weight 1.
- *Clinical plus CNN-OFI*: colsample\_bylevel 1, learning\_rate 0.051, max\_depth 10, gamma 2.34, alpha 1, lambda 0.72, colsample\_bytree 0.58 and min\_child\_weight 3.

### 2.3.3. Comprehensive evaluation

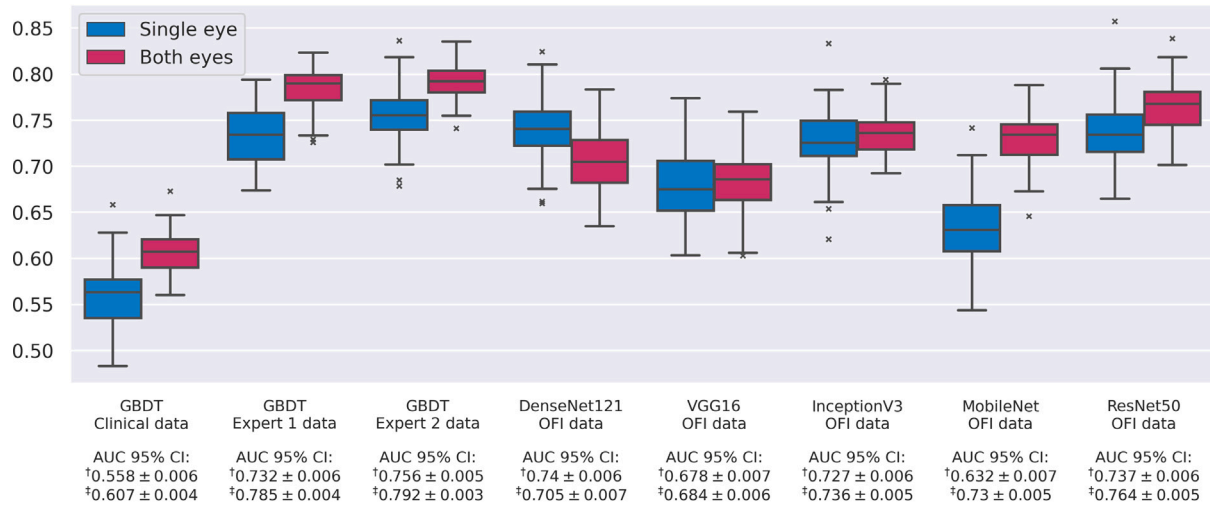
To evaluate the performance of proposed models, several metrics are used, such as the Area Under Curve (AUC) score, which stands for the area under Receiver Operating Characteristic (ROC), and the *Sensitivity* at a fixed *Specificities* of 85%, 90%, and 95% for ensemble models. Each metric is associated with a 95% confidence interval (CI), derived from training 100 models that share the same hyperparameters configuration across 100 different *k*-fold splits. To understand the model's decision-making for tabular data, we used SHAP technique [53] to determine feature importance and interactions. For OFI data, we use the Class Activation Map (CAM) technique, informed by SHAP values, to identify which OFI zones contributed to each class prediction.

## 3. Results

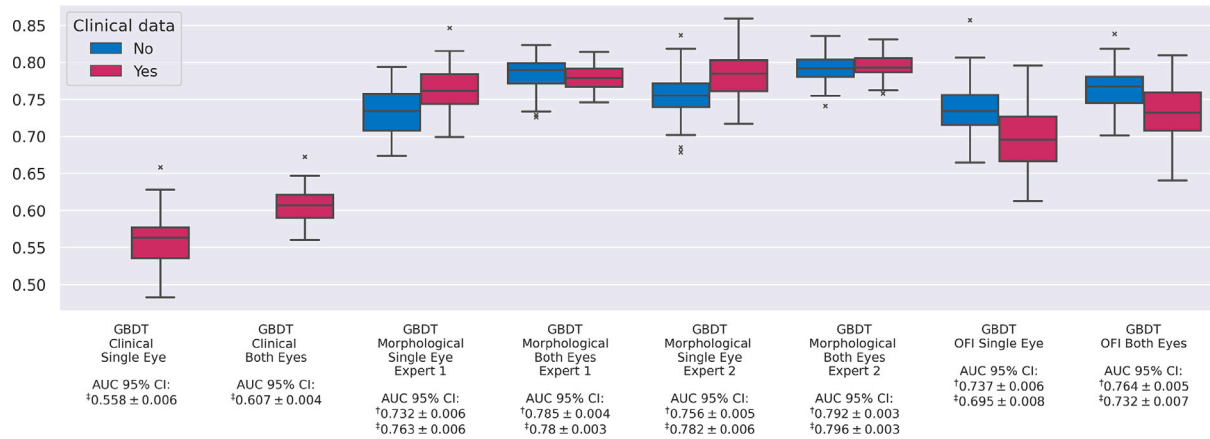
### 3.1. Binocular and monocular approaches

With regard to monocular and binocular comparison, all test data were collected in Figs. 3, 4, and 5, as well as in Table 3. In Fig. 3, a box plot illustrates a comparison between binocular and monocular approaches based on the AUC score for each *no-data-ensemble* models. This plot assesses the impact of the binocular approach on various data modalities (clinical, morphological and pure OFIs) and algorithms, particularly focusing on CNN-based architectures. Fig. 3 illustrates the mean AUC scores from one hundred experiments. The mean AUC with the monocular approach is represented by blue boxes, while the binocular approach is depicted in red. Each experiment includes two pairs of boxes, one for each approach, with CI at 95% displayed below each pair. The results indicate a general trend of higher mean AUC scores for the binocular approach, except for the DenseNet121 architecture. Notably, the MobileNet architecture shows a significant difference between the single-eye and binocular approaches. Among the CNN architectures, ResNet-50 achieved the best mean results, reaching an AUC score of  $0.764 \pm 0.005$  at a 95% CI in binocular mode. The highest overall results were obtained with morphological data from both experts using the binocular approach, with Expert 2 achieving an AUC score of  $0.792 \pm 0.003$  at a 95% CI, which is 0.007 points higher than Expert 1.

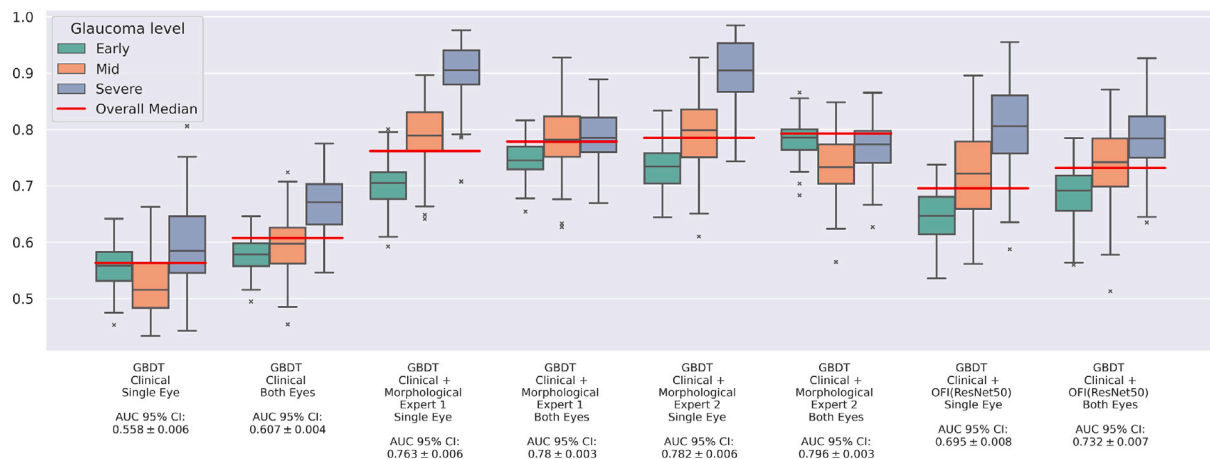
Fig. 4 presents the results of the *data-ensemble* analysis comparing clinical and morphological data with clinical and CNN output. The figure illustrates the AUC scores from one hundred experiments, with each pair of boxes representing the AUC score at a 95% CI. To aid comparison, boxes with the same *data-ensemble* combination are grouped in sets of four (without considering the pure clinical baseline). This categorization resulted in three distinct groups: clinical-morphological Expert 1, clinical-morphological Expert 2, and clinical-OFI, which integrates ResNet-50 output with clinical data. Across all three groups, a consistent trend was observed where the binocular mode reduced box variance and yielded higher scores. The highest performing *data-ensemble* combination was Expert 2 in binocular mode with clinical data, achieving an AUC score of  $0.796 \pm 0.003$  at a 95% CI.



**Fig. 3.** Box plot comparison of different no-data-ensemble models (without combining any data source with clinical data) in terms of AUC score. Below each model, the AUC score is indicated along for a 95% CI. The symbols  $\dagger$  and  $\ddagger$  represent scores for a single eye and both eyes, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



**Fig. 4.** Box plot comparing the impact of clinical data with the data-ensemble approach using AUC scores. Below each model, the AUC score is displayed along with a 95% CI. The symbol  $\dagger$  denotes scores without clinical data, and  $\ddagger$  represents scores with clinical data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

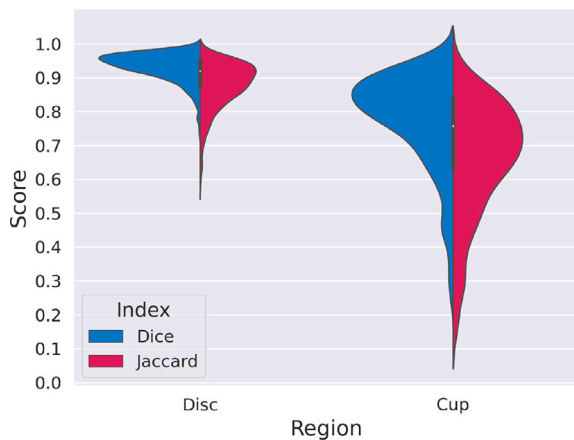


**Fig. 5.** Box plot comparing the impact of clinical data with the data-ensemble approach using AUC scores, stratified by stages of glaucoma progression. The red horizontal bar indicates the median AUC score across all stages, with each model's AUC score presented below with a 95% CI. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 3**

Results in terms of *Sensitivity* (in %) at *Specificity* levels of 85%, 90%, and 95% for ensemble models combining clinical data with morphological and CNN-OFI data, presented at a 95% CI.

|                         |                | Single eye                             |                |                | Both eyes      |                |                |                |
|-------------------------|----------------|----------------------------------------|----------------|----------------|----------------|----------------|----------------|----------------|
|                         |                | <i>Sensitivity at a Specificity of</i> |                |                |                |                |                |                |
|                         |                | 85%                                    | 90%            | 95%            | 85%            | 90%            | 95%            |                |
| Clinical<br>+<br>Morph. | Exp. 1         | <i>Overall</i>                         | 69.235 ± 1.103 | 63.241 ± 1.184 | 55.272 ± 1.231 | 74.298 ± 0.931 | 67.830 ± 0.870 | 57.298 ± 0.778 |
|                         |                | <i>Early</i>                           | 58.540 ± 1.641 | 51.074 ± 1.664 | 42.624 ± 1.684 | 61.556 ± 1.500 | 52.611 ± 1.370 | 40.278 ± 1.044 |
|                         |                | <i>Mid</i>                             | 73.193 ± 2.220 | 68.763 ± 2.263 | 60.665 ± 2.521 | 85.462 ± 1.146 | 79.846 ± 1.368 | 68.923 ± 1.622 |
|                         |                | <i>Severe</i>                          | 93.265 ± 1.609 | 90.008 ± 1.851 | 83.238 ± 2.541 | 79.562 ± 1.232 | 75.188 ± 1.050 | 67.000 ± 0.967 |
|                         | Exp. 2         | <i>Overall</i>                         | 71.503 ± 1.071 | 67.239 ± 1.119 | 59.900 ± 1.205 | 71.277 ± 0.583 | 68.489 ± 0.532 | 63.894 ± 0.691 |
|                         |                | <i>Early</i>                           | 62.771 ± 1.542 | 57.728 ± 1.490 | 49.556 ± 1.564 | 69.222 ± 1.198 | 64.444 ± 1.074 | 56.333 ± 1.375 |
|                         |                | <i>Mid</i>                             | 70.093 ± 2.390 | 66.841 ± 2.591 | 62.544 ± 2.499 | 68.615 ± 0.639 | 67.615 ± 0.696 | 65.308 ± 0.767 |
|                         |                | <i>Severe</i>                          | 95.060 ± 1.370 | 91.530 ± 1.822 | 83.291 ± 2.592 | 75.750 ± 0.642 | 73.750 ± 0.636 | 71.250 ± 0.727 |
| Clinical<br>+<br>OFI    | <i>Overall</i> | 54.531 ± 1.931                         | 42.445 ± 2.017 | 25.003 ± 1.969 | 59.660 ± 1.751 | 49.915 ± 1.849 | 31.915 ± 2.049 |                |
|                         | <i>Early</i>   | 44.701 ± 2.073                         | 32.815 ± 1.935 | 18.608 ± 1.728 | 40.889 ± 2.514 | 30.167 ± 2.098 | 15.000 ± 1.555 |                |
|                         | <i>Mid</i>     | 59.890 ± 3.313                         | 46.386 ± 3.571 | 26.245 ± 3.355 | 63.769 ± 1.740 | 56.615 ± 2.041 | 38.231 ± 2.616 |                |
|                         | <i>Severe</i>  | 75.613 ± 3.258                         | 64.859 ± 3.604 | 41.258 ± 3.694 | 77.438 ± 2.222 | 66.688 ± 2.555 | 45.812 ± 3.279 |                |



**Fig. 6.** Experts segmentation comparison in terms of Jaccard and Dice similarity coefficients.

Finally, the performance of both monocular and binocular approaches in the context of different glaucoma progression levels is presented. In Fig. 5, box plots of the AUC scores obtained by the GBDT algorithm trained on *data-ensemble* and clinical data are displayed. The results are categorized based on glaucoma levels (early, mid, and severe), with each box triplet representing an overall mean (mean across all glaucoma levels) indicated by a red bar and a mean at 95% CI under each triplet. For standalone clinical data and the combination of clinical and ResNet-50 output (referred to as clinical-OFI data), it is evident that the binocular approach yields benefits across all glaucoma progression levels, albeit with slightly inferior results in cases of severe glaucoma within the clinical-OFI dataset. Conversely, when considering clinical and morphological data-ensemble, the advantages of the binocular approach were primarily observed in early glaucoma, with no discernible benefits for mid and severe glaucoma levels, and in some instances, even worse outcomes.

Additionally, we reported a common diagnostic performance metric in Table 3, detailing sensitivity at 85%, 90%, and 95% specificities for our *data-ensemble* methods. This table provides overall performance and stage-specific scores for glaucoma, each with a 95% CI.

### 3.2. Clinical data contribution

To assess the influence of clinical data, we will examine Fig. 4, which illustrates the impact of combining clinical data with morphological data from experts and clinical-OFI data. The results show that incorporating clinical data and morphological data in monocular mode

provides clear advantages. However, using data from both eyes does not yield any additional benefits, as the outcomes remain consistent with those obtained without clinical data. Conversely, integrating clinical data with the output from ResNet-50 (GBDT OFI, and single and both eyes) leads to notably inferior results, as observed in the last four box plots of Fig. 4.

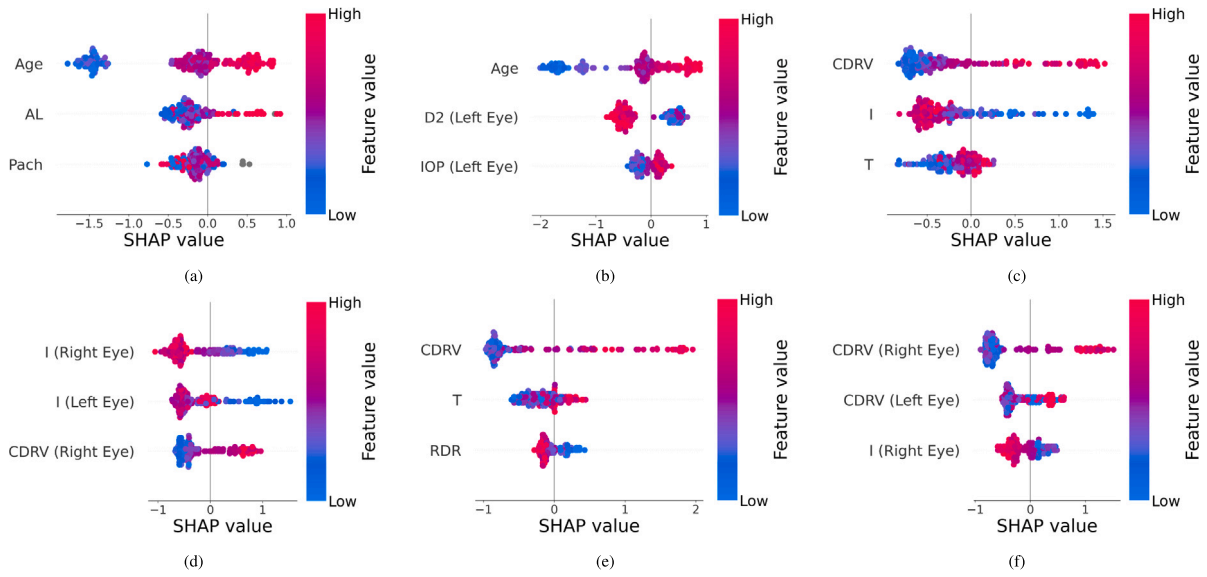
### 3.3. Morphological and CNN-OFI data

If we compare the performance of models trained on morphological data against models trained solely on OFI data, both in isolation and in conjunction with clinical data, the morphological data consistently outperforms the OFI data. This trend is evident across all three figures (Fig. 3, 4, and 5), illustrating the superior positioning of models trained on morphological data.

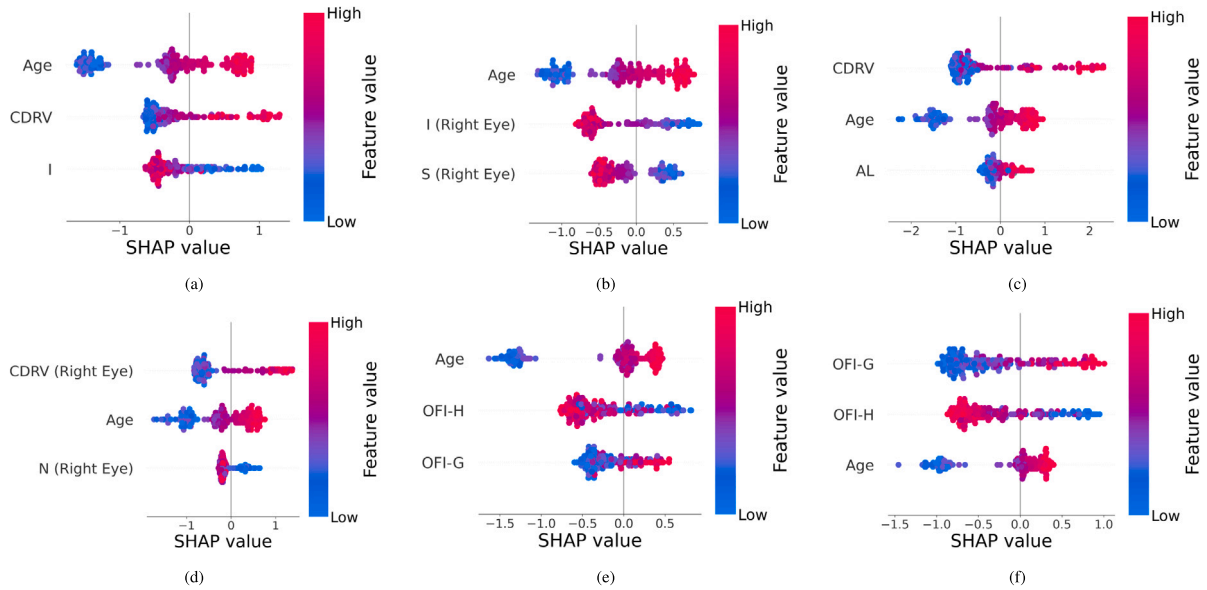
### 3.4. Interpretability of machine learning models

Figs. 7 and 8 depict the feature importance for the top three values in each ensemble model, for both monocular and binocular approaches. Feature importance was assessed using SHAP [63], a tool that comes from game theory [64]. SHAP quantitatively estimates the contribution of each feature to the positive class. A wider feature value distribution on the x-axis indicates a greater contribution to the model's importance in prediction. In these figures (Figs. 7 and 8), each row point represents an instance; the color gradient shows the feature value, and the vertical position reflects SHAP value density. Features are ordered from most (top) to least (bottom) important. In Figs. 7(a) and 7(b), age is consistently the most significant feature, with lower ages negatively influencing and higher ages positively influencing the glaucoma prediction score. Except for D2, which consistently ranks in the top three, other features show variable relevance across the monocular and binocular approaches. Morphological data from Experts 1 and 2 show that CDRV is predominantly the most important feature, except in Fig. 7(d), where it is second to the Inferior thickness I. CDRV, Inferior I, and Temporal T thicknesses are frequently identified as the most relevant features. In binocular mode, the right eye data consistently appears more influential than the left eye data for each feature.

Fig. 8 shows the SHAP values for the ensemble model combining clinical and morphological data, as well as clinical and ResNet-OFI data. In the case of the first expert in single-eye mode (Fig. 8(a)), the feature importance is very similar to that in Fig. 7(c), with CDRV and the Inferior I zone being the top morphological features. However, this time the clinical feature Age takes the first position. As shown in Fig. 4, the ensemble model improves the baseline by incorporating clinical data.



**Fig. 7.** Feature importance with SHAP values for GBDT models. The horizontal axis displays the SHAP values, while the vertical axis lists the features, sorted by importance. The color bar represents the numerical value of each feature. Negative SHAP values indicate that a feature negatively impacts the glaucoma class, positive SHAP values suggest a positive influence on the glaucoma class, and values near zero imply that the feature has no impact. (a) Clinical data in single-eye mode. (b) Clinical data in both-eyes mode. (c) Morphological data from Expert 1 in single-eye mode. (d) Morphological data from Expert 1 in both-eyes mode. (e) Morphological data from Expert 2 in single-eye mode. (f) Morphological data from Expert 2 in both-eyes mode. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



**Fig. 8.** Feature importance with SHAP values for GBDT ensemble-data models. The horizontal axis displays the SHAP values, while the vertical axis lists the features, sorted by importance. The color bar represents the numerical value of each feature. Negative SHAP values indicate that a feature negatively impacts the glaucoma class, positive SHAP values suggest a positive influence on the glaucoma class, and values near zero imply that the feature has no impact. (a) Clinical and morphological data from Expert 1 in single-eye mode. (b) Clinical and morphological data from Expert 1 in both-eyes mode. (c) Clinical and morphological data from Expert 2 in single-eye mode. (d) Clinical and morphological data from Expert 2 in both-eyes mode. (e) Clinical and ResNet-OFI output data in single-eye mode. (f) Clinical and ResNet-OFI output data in both-eyes mode. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In Fig. 8(b), we see a decrease in the significance of CDRV to the fourth place (considering I from the right eye and I from left eye as the same feature), but Age remains the most important feature. As indicated in Fig. 4, the overall AUC value decreases by 0.005 when using clinical data, but the error margin narrows by 0.001.

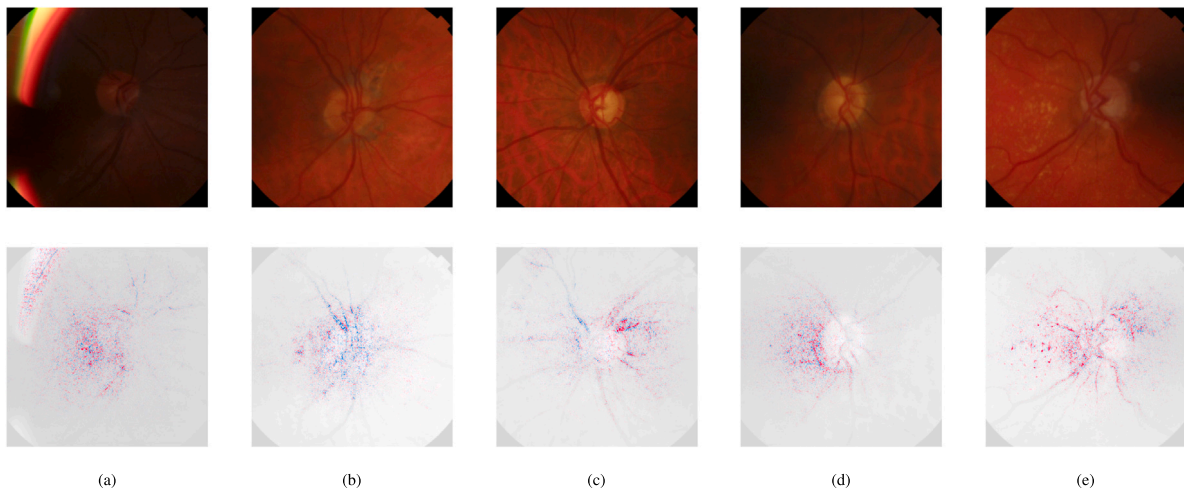
In Figs. 8(c) and 8(d) for Expert 2, CDRV is the most relevant feature in both approaches, followed by Age.

In both cases, adding clinical data positively impacted the diagnosis score.

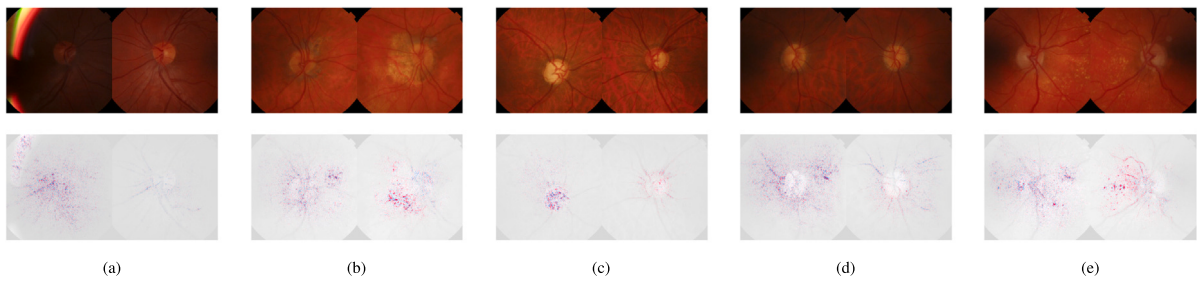
For the clinical and ResNet-OFI data (Figs. 8(e) and 8(f)), we observe an expected behavior: the features Age, OFI Health score (OFI-H), and OFI Glaucoma score (OFI-G) are the top three. It is worth mentioning that in the monocular approach (Fig. 8(e)), Age was the most important feature, while in the binocular approach (Fig. 8(f)), OFI-G and OFI-H were prioritized over Age.

Finally, we provide insights from ResNet-50 predictions, presenting a CAM approach for both single-eye and both-eyes modes in Figs. 9 and 10. These figures show the activations using the Deep Explainer method [65], where red indicates high activation values





**Fig. 9.** SHAP values for the monocular approach are calculated using the Deep Explainer method. In the visualization, red dots scattered across the image indicate pixels that positively contribute to the glaucoma class, while blue dots denote pixels that negatively impact the glaucoma class. (a) Right eye from patient #39; ground truth were tagged as *healthy*. Prediction score: 29.6% (toward glaucoma). (b) Right eye from patient #69; ground truth were tagged as *glaucoma* with a VF/MD of  $-0.25$  dB. Prediction score: 42.3% (toward glaucoma). (c) Left eye from patient #14; ground truth were tagged as *glaucoma* with a VF/MD of  $-4.65$  dB (Mid glaucoma). Prediction score: 71.3% (toward glaucoma). (d) Right eye from patient #88; ground truth were tagged as *glaucoma* with a VF/MD of  $-6.32$  dB. Prediction score: 74.7% (toward glaucoma). (e) Left eye from patient #19; ground truth were tagged as *glaucoma* with a VF/MD of  $-19.99$  dB. Prediction score: 62.5% (toward glaucoma). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



**Fig. 10.** SHAP values for the binocular approach, calculated using the Deep Explainer method. Red dots scattered across the image indicate pixels that positively contribute to the glaucoma class, while blue dots denote pixels that negatively impact the glaucoma class. (a) Both (right & left) from patient #39; ground truth were tagged as *healthy*. Prediction score: 29.6% (toward glaucoma). (b) Both (right & left) from patient #69; ground truth were tagged as *glaucoma* with a VF/MD of  $-0.25$  dB and  $-2.48$  dB respectively. Prediction score: 44.0% (toward glaucoma). (c) Both (right & left) from patient #14; ground truth were tagged as *glaucoma* with a VF/MD of  $-11.37$  dB and  $-4.65$  dB respectively. Prediction score: 95.3% (toward glaucoma). (d) Both (right & left) from patient #88; ground truth were tagged as *glaucoma* with a VF/MD of  $-6.32$  dB and  $-9.63$  dB respectively. Prediction score: 52.2% (toward glaucoma). (e) Both (right & left) from patient #19; ground truth were tagged as *glaucoma* with a VF/MD of  $-6.14$  dB and  $-19.99$  dB respectively. Prediction score: 85.4% (toward glaucoma). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

contributing positively to the glaucoma prediction score, and blue represents low activations contributing to the healthy class prediction score. For each mode, five instances were selected, representing healthy patients, glaucoma with VF/MD  $\geq -3$  dB, glaucoma with  $-3$  dB  $>$  VF/MD  $\geq -6$  dB, glaucoma with  $-6$  dB  $>$  VF/MD  $\geq -12$  dB, and glaucoma with VF/MD  $< -12$  dB. In single-eye mode, as seen in Fig. 9, activations around ONH indicate significant information contribution from this region. Additionally, Fig. 9(e) shows an activation map highlighting vessel segmentation near the ONH and areas indicating potential pathological features. In the binocular mode (Fig. 10), activations often associated with glaucoma are predominantly located near the ONH. The patterns of the monocular mode, such as vessel highlighting and hard exudates, are mirrored in the binocular mode, indicating consistency in the model's response to anomalies in OFIs.

#### 4. Discussion

Early detection of glaucoma is an extremely important task as it can prevent irreversible eye damage and thus preserve the patient's vision. In this study, we present a novel approach for diagnosing glaucoma by combining information from both eyes. To our knowledge, this is the first work that uses information from both eyes of the same patient

to automatically diagnose glaucoma. We provide a comparative analysis highlighting the improvements in diagnosis when using bilateral eye information versus an eye-by-eye approach. The research demonstrates the efficacy of using data from both eyes to enhance early-stage glaucoma diagnosis in both GBDT and CNN models. Additionally, we investigate the impact of incorporating patients' clinical test results on the accuracy of glaucoma diagnosis.

##### 4.1. Binocular approach

From Fig. 3, we note that the median performance in the binocular mode generally surpasses the single eye mode, particularly with clinical data, morphological data from Expert 1 and Expert 2, MobileNet, and ResNet-50. These models show improvements of 0.049, 0.053, 0.036, 0.098, and 0.027, respectively, at a 95% CI. Additionally, Fig. 5 shows that, for clinical-OFI data, the binocular approach significantly outperforms the monocular, increasing its median by 0.037, unlike the result with morphological and clinical data. This improvement likely stems from the binocular models ability to simultaneously access data from both eyes, enabling it to learn inter-eye relationships that may be imperceptible otherwise. These findings advocate for the binocular approach as a viable method, offering access to more complex information and enhancing overall performance.

#### 4.2. Morphological and pure OFI data

The superior performance of the morphological data (see Fig. 3) may be due to having a more constrained feature space in contrast to pure OFI data. Each image is represented in a high-dimensional feature space, within which glaucoma may or may not be present, with each new instance further constraining that space. The more constrained the space, the easier it becomes to identify or extract richer glaucoma-related features. This indicates that we have insufficient training data for a direct end-to-end DL approach. The CAM in Figs. 9 and 10 illustrate the network's sensitivity to image anomalies and its awareness of where glaucoma-related information resides. However, manually extracted data are richer and less noisy. This situation is consistent with the understanding that computer vision problems typically require large volumes of data to achieve optimal performance. The rapid evolution of the field from LeNet-5 [66] to the present, characterized by constant changes in model architecture, underscores the ongoing need for more data. In scenarios involving small datasets, like ours, we recommend exploring automated segmentation methods, though such explorations were beyond the scope of our current research.

#### 4.3. Contribution of clinical data

Adding clinical data to morphological data from Expert 2 resulted in an improvement of the AUC score, as shown in Fig. 4. For Expert 1, enhancement was seen only in the monocular approach, while the binocular approach had a similar performance.

Nevertheless, the best results were obtained from the ensemble model that combined clinical and morphological data from Expert 2 in binocular mode, achieving an AUC of  $0.796 \pm 0.003$ . This produces a sensitivity of  $63.894\% \pm 0.691$  at a specificity of 95% (see Table 3), both with a 95% CI. In comparing the CNN and morphological approaches, considerations include data acquisition cost and reliability. Morphological data acquisition, reliant on manual segmentations, is time-consuming and costly. Early diagnosis in glaucoma patients is crucial to prevent severe vision loss. Segmentation reliability is often subjective, particularly in determining the optic cup for early-stage or healthy patients. Fig. 6 shows the segmentation variance between the two experts, highlighting differences in the Dice coefficient and Jaccard score. The variance in cup segmentations reflects the realism of the PAPIA dataset, especially its early-stage cases with minimal morphological changes.

With GBDT OFI-based models, there was a decrease in AUC by 0.042 and 0.012 for monocular and binocular approaches, respectively. The decline is attributed to the limited predictive power of clinical data (except for Age) and the nature of the GBDT algorithm. Given the realism of the PAPIA dataset with many early-stage glaucoma cases, detection becomes more challenging, reflecting the multifactorial nature of the disease [67] where some forms, like normal tension glaucoma [68,69], occur without an elevated intraocular pressure. This is reflected in Figs. 7(a), 7(b), and 8, where IOP is seldom among the top three features, in contrast to Age, which consistently appears.

In our analysis, Age consistently emerged as the prominent feature, unlike IOP, underscoring its predictive value in age-related prevalence of glaucoma [3–5]. Gao et al. [70] reported an AUC of 0.731 using only age and sex data, further emphasizing the predictive significance of Age. Although Age is crucial, the monocular and binocular approaches yielded AUCs of  $0.558 \pm 0.006$  and  $0.607 \pm 0.004$ , respectively, indicating limited overall predictive power.

On the other hand, GBDT's basis in decision trees allows for easy interpretability through decision tracing, a crucial aspect in medical applications. In the case of combining clinical and OFI data, we added two new columns to the clinical dataset, which, as noted, are not as strong predictors as the complete set of morphological features. Figs. 8(e) and 8(f) show that the primary features often involve root splits, with subsequent splits relying on the remaining clinical data, which

ultimately leads to poorer predictions. Although GBDT outperformed the CNN-based models, we noted in Fig. 5 that those improvements in early-stage glaucoma detection always came at the expense of reduced AUC in severe cases. This suggests that, despite its interpretability, the GBDT method may not be ideal for a binocular approach, as its sequential splitting nature struggles to capture intra-ocular relationships, a task potentially better suited for a neural network algorithm. Interpretability, particularly in machine learning, involves complex considerations [71]. deep neural networks, potentially better at discerning these relationships, lack the clear interpretability provided by GBDT, highlighting a fundamental trade-off between model performance and transparency.

#### 4.4. Study limitations

The limitations of this study include the reliance on a small dataset, which may hinder the effectiveness of direct end-to-end DL approaches. Additionally, the subjective nature of manual segmentation in morphological data can introduce variability and inconsistency. The predictive power of clinical data is limited, with only Age showing significant relevance, while other factors like IOP are less predictive. Furthermore, the GBDT algorithm, while interpretable, may not effectively capture intra-ocular relationships due to its sequential splitting nature, potentially limiting its performance in binocular approaches.

This work studies how incorporating data from both eyes and additional clinical information can enhance the detection and analysis of glaucoma. While the study does not aim to produce a final clinical tool for immediate use, the results suggest that future glaucoma detection tools [72] should integrate these combined data sources for greater performance.

#### CRedit authorship contribution statement

**Oleksandr Kovalyk-Borodyak:** Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Juan Morales-Sánchez:** Writing – review & editing, Validation, Supervision, Project administration, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Rafael Verdú-Monedero:** Writing – review & editing, Writing – original draft, Resources, Funding acquisition, Data curation, Conceptualization. **José-Luis Sancho-Gómez:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Funding acquisition, Conceptualization.

#### Code availability

The code used to generate the results presented in this study is publicly available at GitHub (<https://github.com/TDAM-UPCT/glaucoma-bin>) under the MIT License.

#### Declaration of competing interest

The authors declare no competing interests.

#### Acknowledgments

This work has been partially funded by Instituto de Salud Carlos III, Spain (co-funded by European Regional Development Fund; “A way to make Europe”) through the national projects AES2017-PI17/00771 and AES2017-PI17/00821, and by Fundación Séneca, Spain through regional project 20901/PI/18.

## References

- [1] Tham Y-C, Li X, Wong TY, Quigley HA, Aung T, Cheng C-Y. Global prevalence of glaucoma and projections of glaucoma burden through 2040 : A systematic review and meta-analysis. *Ophthalmology* 2014;121(11):2081–90. <http://dx.doi.org/10.1016/j.ophtha.2014.05.013>.
- [2] Shaikh Y, Yu F, Coleman AL. Burden of undetected and untreated glaucoma in the United States. *Am J Ophthalmol* 2014;158(6):1121–9.e1. <http://dx.doi.org/10.1016/j.ajo.2014.08.023>, URL <https://www.sciencedirect.com/science/article/pii/S000293941400525X>.
- [3] Ryskulova A, Turczyn K, Makuc DM, Cotch MF, Klein RJ, Janiszewski R. Self-reported age-related eye diseases and visual impairment in the United States: results of the 2002 national health interview survey. *Am J Public Health* 2008;98(3):454–61.
- [4] Varma R, Ying-Lai M, Francis BA, Nguyen BB-T, Deneen J, Wilson MR, Azen SP, Group LALES, et al. Prevalence of open-angle glaucoma and ocular hypertension in latinos: the los angeles latino eye study. *Ophthalmology* 2004;111(8):1439–48.
- [5] Group EDP, et al. Prevalence of open-angle glaucoma among adults in the United States. *Arch Ophthalmol* 2004;122(4):532.
- [6] Institute NE. Us age-specific prevalence rates for glaucoma by age and race/ethnicity. 2019, URL <https://www.nei.nih.gov/learn-about-eye-health/resources-for-health-educators/eye-health-data-and-statistic/glaucoma-data-and-statistics>.
- [7] Jonas JB, Aung T, Bourne RR, Bron AM, Ritch R, Panda-Jonas S. Glaucoma. *Lancet* 2017;390(10108):2183–93. [http://dx.doi.org/10.1016/S0140-6736\(17\)31469-1](http://dx.doi.org/10.1016/S0140-6736(17)31469-1), URL <https://www.sciencedirect.com/science/article/pii/S0140673617314691>.
- [8] Dirani M, Crowston J, Taylor P, Moore P, Rogers S, Pezzullo M, Keefe J, Taylor H. Economic impact of primary open-angle glaucoma in australia. *Clin Exp Ophthalmol* 2011;39:623–32. <http://dx.doi.org/10.1111/j.1442-9071.2011.02530.x>.
- [9] Foster PJ, Oen FTS, Machin D, Ng T-P, Devereux JG, Johnson GJ, Khaw PT, Seah SKL. The prevalence of glaucoma in Chinese residents of Singapore: A cross-sectional population survey of the tanjong pagar district. *Arch Ophthalmol* 2000;118(8):1105–11. <http://dx.doi.org/10.1001/archophth.118.8.1105>, arXiv:<https://jamanetwork.com/journals/jamaophthalmology/articlepdf/413488/eeb90022.pdf>.
- [10] Glaucoma research foundation: Glaucoma facts and stats. URL <https://www.glaucoma.org/glaucoma/glaucoma-facts-and-stats.php>.
- [11] Shen SY, Wong TY, Foster PJ, Loo J-L, Rosman M, Loon S-C, Wong WL, Saw S-M, Aung T. The prevalence and types of glaucoma in malay people: the singapore malay eye study. *Invest Ophthalmol Vis Sci* 2008;49(9):3846–3851. <http://dx.doi.org/10.1167/iovs.08-1759>, URL <http://intl.iovs.org/cgi/content/full/49/9/3846>.
- [12] Susanna R, De Moraes CG, Cioffi GA, Ritch R. Why do people (still) go blind from glaucoma? *Transl Vis Sci Technol* 2015;4(2). 1–1.
- [13] Armaly MF, Sayegh RE. The cup/disc ratio: The findings of tonometry and tonography in the normal eye. *Arch Ophthalmol* 1969;82(2):191–6. <http://dx.doi.org/10.1001/archophth.1969.00990020193008>.
- [14] Kumar JRH, Seelamantula CS, Kamath YS, Jampala R. Rim-to-disc ratio outperforms cup-to-disc ratio for glaucoma prescreening. *Sci Rep* 2019;9(1). <http://dx.doi.org/10.1038/s41598-019-43385-2>.
- [15] Jonas JB, Gusek GC, Naumann GO. Optic disc, cup and neuroretinal rim size, configuration and correlations in normal eyes. *Invest Ophthalmol Vis Sci* 1988;29(7):1151–8.
- [16] Harizman N, Oliveira C, Chiang A, Tello C, Marmor M, Ritch R, Liebmann JM. The ISNT rule and differentiation of normal from glaucomatous eyes. *Arch Ophthalmol* 2006;124(11):1579–83. <http://dx.doi.org/10.1001/archophth.124.11.1579>.
- [17] Fan R, Alipour K, Bowd C, Christopher M, Brye N, Proudfoot JA, Goldbaum MH, Belghith A, Kirkin CA, Fazio MA, Liebmann JM, Weinreb RN, Pazzani M, Kriegman D, Zangwill LM. Detecting glaucoma from fundus photographs using deep learning without convolutions: Transformer for improved generalization. *Ophthalmol Sci* 2023;3(1):100233. <http://dx.doi.org/10.1016/j.xops.2022.100233>, URL <https://www.sciencedirect.com/science/article/pii/S2666914522001221>.
- [18] Elmannai H, Hamdi M, Meshoul S, Alhussan AA, Ayadi M, Ksibi A. An improved deep learning framework for automated optic disc localization and glaucoma detection. *CMES Comput Model Eng Sci* 2024;140(2):1429–57. <http://dx.doi.org/10.32604/cmesc.2024.048557>, URL <http://www.techscience.com/CMES/v140n2/56552>.
- [19] P S, J R, P. An automatic recognition of glaucoma in fundus images using deep learning and random forest classifier. *Appl Soft Comput* 2021;109:107512. <http://dx.doi.org/10.1016/j.asoc.2021.107512>.
- [20] Wang L, Gu J, Chen Y, Liang Y, Zhang W, Pu J, Chen H. Automated segmentation of the optic disc from fundus images using an asymmetric deep learning network. *Pattern Recognit* 2021;112:107810. <http://dx.doi.org/10.1016/j.patcog.2020.107810>.
- [21] Latif J, Tu S, Xiao C, Ur Rehman S, Imran A, Latif Y. Odgnet: a deep learning model for automated optic disc localization and glaucoma classification using fundus images. *SN Appl Sci* 2022;4(4). <http://dx.doi.org/10.1007/s42452-022-04984-3>.
- [22] Shyamalee T, Meedeniya D. Glaucoma detection with retinal fundus images using segmentation and classification. *Mach Intell Res* 2022;19(6):563–80. <http://dx.doi.org/10.1007/s11633-022-1354-z>.
- [23] Dixit A, Yohannan J, Boland MV. Assessing glaucoma progression using machine learning trained on longitudinal visual field and clinical data. *Ophthalmology* 2021;128(7):1016–26. <http://dx.doi.org/10.1016/j.ophtha.2020.12.020>.
- [24] Hervella AS, Rouco J, Novo J, Ortega M. End-to-end multi-task learning for simultaneous optic disc and cup segmentation and glaucoma classification in eye fundus images. *Appl Soft Comput* 2022;116:108347. <http://dx.doi.org/10.1016/j.asoc.2021.108347>.
- [25] Guo Y, Peng Y, Sun J, Li D, Zhang B. Dslin: Dual-tutor student learning network for multiracial glaucoma detection. *Neural Comput Appl* 2022;34(14):11885–910. <http://dx.doi.org/10.1007/s00521-022-07078-8>.
- [26] Zhao R, Chen X, Chen Z, Li S. Diagnosing glaucoma on imbalanced data with self-ensemble dual-curriculum learning. *Med Image Anal* 2022;75:102295. <http://dx.doi.org/10.1016/j.media.2021.102295>.
- [27] Das D, Nayak DR, Bhandary SV, Acharya UR. Cdam-net: Channel shuffle dual attention based multi-scale cnn for efficient glaucoma detection using fundus images. *Eng Appl Artif Intell* 2024;133. <http://dx.doi.org/10.1016/j.engappai.2024.108454>, cited by: 0 URL <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85191241510&doi=10.1016%2fj.engappai.2024.108454&partnerID=40&md5=024f4f10023c7eae3e55d1f134a8891f3>.
- [28] Chiang Y-Y, Chen C-L, Chen Y-H. Deep learning evaluation of glaucoma detection using fundus photographs in highly myopic populations. *Biomed* 2024;12(7). <http://dx.doi.org/10.3390/biomed12071394>, URL <https://www.mdpi.com/2227-9059/12/7/1394>.
- [29] Lee J, Kim YK, Park KH, Jeoung JW. Diagnosing glaucoma with spectral-domain optical coherence tomography using deep learning classifier. *J Glaucoma* 2020;29(4):287–94. <http://dx.doi.org/10.1097/jig.0000000000001458>.
- [30] Juneja M, Minhas JS, Singla N, Thakur S, Thakur N, Jindal P. Fused framework for glaucoma diagnosis using optical coherence tomography (oct) images. *Expert Syst Appl* 2022;201:117202. <http://dx.doi.org/10.1016/j.eswa.2022.117202>.
- [31] Xiong J, Li F, Li F, Song D, Tang G, Tang G, He J, Gao K, Zhang H, Cheng W, Song Y, Lin F, Hu K, Wang P, Li J-PO, Aung T, Qiao Y, Zhang X, Ting DS. Multimodal machine learning using visual fields and peripapillary circular oct scans in detection of glaucomatous optic neuropathy. *Ophthalmology* 2021. <http://dx.doi.org/10.1016/j.ophtha.2021.07.032>.
- [32] Mehta P, Petersen CA, Wen JC, Banitt MR, Chen PP, Bojikian KD, Egan C, Lee S-I, Balazinska M, Lee AY, Rokem A. Automated detection of glaucoma with interpretable machine learning using clinical data and multimodal retinal images. *Am J Ophthalmol* 2021;231:154–69. <http://dx.doi.org/10.1016/j.ajo.2021.04.021>.
- [33] Mahmudi T, Kafieh R, Rabbani H, Mehri A, Akhlaghi M-R. Evaluation of asymmetry in right and left eyes of normal individuals using extracted features from optical coherence tomography and fundus images. *J Med Signals Sens* 2021;11(1):12–23. [http://dx.doi.org/10.4103/jmss.JMSS\\_67\\_19](http://dx.doi.org/10.4103/jmss.JMSS_67_19).
- [34] Berenguer-Vidal R, Verdú-Monedero R, Morales-Sánchez J, Sellés-Navarro I, Kovalyk O, Sancho-Gómez J-L. Decision trees for glaucoma screening based on the asymmetry of the retinal nerve fiber layer in optical coherence tomography. *Sensors* 2022;22(13). <http://dx.doi.org/10.3390/s22134842>, URL <https://www.mdpi.com/1424-8220/22/13/4842>.
- [35] Rodríguez-Robles F, Verdú-Monedero R, Berenguer-Vidal R, Morales-Sánchez J, Sellés-Navarro I. Analysis of the asymmetry between both eyes in early diagnosis of glaucoma combining features extracted from retinal images and octs into classification models. *Sensors* 23(10). <http://dx.doi.org/10.3390/s23104737>, cited by: 1; All Open Access, Gold Open Access, Green Open Access (2023) URL <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85160587140&doi=10.3390%2fs23104737&partnerID=40&md5=3249e330556a2c1ca420ca0d8508ce2a>.
- [36] Plečáček D, Gotovac Đogaš V, Polašek O, Škunca Herman J. Determinants of human asymmetry: Does asymmetrical retinal vasculature predict asymmetry elsewhere in the body? *Life* 14(8). <http://dx.doi.org/10.3390/life14080929>, cited by: 0; All Open Access, Gold Open Access (2024). URL <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85202641346&doi=10.3390%2flife14080929&partnerID=40&md5=2229b3543c59e0adc0c52b442d92f4bd>.
- [37] Ashtari-Majlan M, Dehshibi MM, Masip D. Glaucoma diagnosis in the era of deep learning: A survey. *Expert Syst Appl* 2024;256:124888. <http://dx.doi.org/10.1016/j.eswa.2024.124888>, URL <https://www.sciencedirect.com/science/article/pii/S095741742401755X>.
- [38] Kovalyk O, Morales-Sánchez J, Verdú-Monedero R, Sellés-Navarro I, Palazón-Cabanes A, Sancho-Gómez J-L. Papila: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. *Sci Data* 2022;9(1):1–12.
- [39] Almazroa A, Alodhayb S, Osman E, Ramadan E, Hummadi M, Dlaim M, Alkatee M, Raahemifar K, Lakshminarayanan V. Retinal fundus images for glaucoma analysis: the RIGA dataset. In: Zhang J, Chen P-H, editors. *Medical imaging 2018: imaging informatics for healthcare, research, and applications*. International society for optics and photonics, vol. 10579, SPIE; 2018, p. 55–62. <http://dx.doi.org/10.1117/12.2293584>.



- [40] Zhang Z, Yin FS, Liu J, Wong WK, Tan NM, Lee BH, Cheng J, Wong TY. ORIGA-light: an online retinal fundus image database for glaucoma analysis and research. In: Int conf of the IEEE engineering in medicine and biology. 2010, p. 3065–8. <http://dx.doi.org/10.1109/iembs.2010.5626137>.
- [41] Fumero F, Alayón S, Sanchez JL, Sigut J, Gonzalez-Hernandez M. RIM-ONE: an open retinal image database for optic nerve evaluation. In: 2011 24th int. symposium on computer-based medical systems. CBMS, IEEE; 2011, p. 1–6. <http://dx.doi.org/10.1109/CBMS.2011.5999143>.
- [42] Batista FJF, Diaz-Aleman T, Sigut J, Alayon S, Arnay R, Angel-Pereira D. Rim-one dl: A unified retinal image database for assessing glaucoma using deep learning. Image Anal Stereol 2020;39(3):161–7. <http://dx.doi.org/10.5566/ias.2346>.
- [43] Sivaswamy J, Krishnadas SR, Datt Joshi G, Jain M, Syed Tabish AU. Drishti-GS: Retinal image dataset for optic nerve head (ONH) segmentation. In: 2014 IEEE 11th international symposium on biomedical imaging. ISBI, 2014, p. 53–6. <http://dx.doi.org/10.1109/ISBI.2014.6867807>.
- [44] Diaz-Pinto A, Morales S, Naranjo V, Kohler T, Mossi J, Navea A. CNNs for automatic glaucoma assessment using fundus images: an extensive validation. BioMed Eng OnLine 2019;18:1–19. <http://dx.doi.org/10.1186/s12938-019-0649-y>.
- [45] Bajwa MN, Singh GAP, Neumeier W, Malik MI, Dengel A, Ahmed S. G1020: A benchmark retinal fundus image dataset for computer-aided glaucoma detection. In: 2020 int joint conf on neural networks. IJCNN, 2020, p. 1–7. <http://dx.doi.org/10.1109/IJCNN48605.2020.9207664>.
- [46] Orlando JI, Fu H, Barbosa Breda J, van Keer K, Bathula DR, Diaz-Pinto A, Fang R, Heng P-A, Kim J, Lee J, Lee J, Li X, Liu P, Lu S, Murugesan B, Naranjo V, Phayé SSR, Shankaranarayana SM, Sikka A, Son J, van den Hengel A, Wang S, Wu J, Wu Z, Xu G, Xu Y, Yin P, Li F, Zhang X, Xu Y, Bogunović H. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Med Image Anal 2020;59:101570. <http://dx.doi.org/10.1016/j.media.2019.101570>.
- [47] Retinal fundus images for glaucoma analysis: RIGA dataset, [https://deepblue.lib.umich.edu/data/concern/data\\_sets/3b591905z](https://deepblue.lib.umich.edu/data/concern/data_sets/3b591905z).
- [48] RIMONE database, <https://medimrg.webs.ull.es/research/downloads/>.
- [49] Drishti-GS database, <http://cvit.iitit.ac.in/projects/mip/drishti-gs/mip-dataset2/Home.php>.
- [50] ACRIMA database, <https://figshare.com/s/c2d31f850af14c5b5232>.
- [51] G1020 dataset, <https://www.dfki.uni-kl.de/g1020>.
- [52] REFUGE: Retinal fundus glaucoma challenge, <https://refuge.grand-challenge.org/>.
- [53] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Advances in neural information processing systems. vol. 30, Curran Associates, Inc.; 2017, URL <https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>.
- [54] Liaw R, Liang E, Nishihara R, Moritz P, Gonzalez JE, Stoica I. Tune: A research platform for distributed model selection and training. 2018, arXiv preprint [arXiv:1807.05118](https://arxiv.org/abs/1807.05118).
- [55] Friedman JH, Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Statist 2001. <http://dx.doi.org/10.1214/aos/1013203451>.
- [56] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. KDD '16, New York, NY, USA: Association for Computing Machinery; 2016, p. 785–94. <http://dx.doi.org/10.1145/2939672.2939785>.
- [57] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2015, <http://dx.doi.org/10.48550/arxiv.1409.1556>.
- [58] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2015, <http://dx.doi.org/10.48550/arxiv.1512.03385>.
- [59] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. 2015, <http://dx.doi.org/10.48550/arxiv.1512.00567>.
- [60] Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. 2017, <http://dx.doi.org/10.48550/arxiv.1704.04861>.
- [61] Deng J, Socher R, Fei-Fei L, Dong W, Li K, Li L-J. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. CVPR, vol. 00, 2009, p. 248–55. <http://dx.doi.org/10.1109/CVPR.2009.5206848>, URL <https://ieeexplore.ieee.org/abstract/document/5206848/>.
- [62] Loshchilov I, Hutter F. Decoupled weight decay regularization. 2019, [arXiv:1711.05101](https://arxiv.org/abs/1711.05101).
- [63] Lundberg S, Lee S-I. A unified approach to interpreting model predictions. 2017, [arXiv:1705.07874](https://arxiv.org/abs/1705.07874).
- [64] The Shapley value: Essays in honor of Lloyd S. Shapley. Cambridge University Press; 1988.
- [65] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. 2019, [arXiv:1704.02685](https://arxiv.org/abs/1704.02685).
- [66] Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE 1998;86(11):2278–324. <http://dx.doi.org/10.1109/5.726791>.
- [67] Blumberg D, Skaat A, Liebmman JM. Emerging risk factors for glaucoma onset and progression. Prog Brain Res 2015;221:81–101.
- [68] Mallick J, Devi L, Malik PK, Mallick J. Update on normal tension glaucoma. J Ophthalmic Vis Res 2016;11(2):204.
- [69] Group CN-TGS, et al. Natural history of normal-tension glaucoma. Ophthalmology 2001;108(2):247–53.
- [70] Gao XR, Huang H, Kim H. Polygenic risk score is associated with intraocular pressure and improves glaucoma prediction in the UK Biobank cohort. Transl Vis Sci Technol 2019;8(2): 10–10.
- [71] Lipton ZC. The mythos of model interpretability. 2017, [arXiv:1606.03490](https://arxiv.org/abs/1606.03490).
- [72] Shyamalee T, Meedeniya D, Lim G, Karunaratne M. Automated tool support for glaucoma identification with explainability using fundus images. IEEE Access 2024;12:17290–307. <http://dx.doi.org/10.1109/ACCESS.2024.3359698>.