Machine Learning: Science and Technology

PAPER • OPEN ACCESS

Interpretable surrogate models to approximate the predictions of convolutional neural networks in glaucoma diagnosis

To cite this article: Jose Sigut *et al* 2023 *Mach. Learn.: Sci. Technol.* **4** 045024

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
8 June 2023

REVISED
6 October 2023

ACCEPTED FOR PUBLICATION
27 October 2023

PUBLISHED
6 November 2023

PAPER

Interpretable surrogate models to approximate the predictions of convolutional neural networks in glaucoma diagnosis

Jose Sigut[1,*], Francisco Fumero[1], Rafael Arnay[1], José Estévez[1] and Tinguaro Díaz-Alemán[2]

[1] Department of Computer Science and Systems Engineering, Universidad de La Laguna, Camino San Francisco de Paula, 19, La Laguna 38203, Santa Cruz de Tenerife, Spain
[2] Department of Ophthalmology, Hospital Universitario de Canarias, Carretera Ofra S/N, La Laguna 38320, Santa Cruz de Tenerife, Spain
[*] Author to whom any correspondence should be addressed.

E-mail: jfsigut@ull.edu.es

Keywords: interpretable surrogate models, glaucoma diagnosis, convolutional neuronal networks

Abstract

Deep learning systems, especially in critical fields like medicine, suffer from a significant drawback, their black box nature, which lacks mechanisms for explaining or interpreting their decisions. In this regard, our research aims to evaluate the use of surrogate models for interpreting convolutional neural network (CNN) decisions in glaucoma diagnosis. Our approach is novel in that we approximate the original model with an interpretable one and also change the input features, replacing pixels with tabular geometric features of the optic disc, cup, and neuroretinal rim. We trained CNNs with two types of images: original images of the optic nerve head and simplified images showing only the disc and cup contours on a uniform background. Decision trees were used as surrogate models due to their simplicity and visualization properties, while saliency maps were calculated for some images for comparison. The experiments carried out with 1271 images of healthy subjects and 721 images of glaucomatous eyes demonstrate that decision trees can closely approximate the predictions of neural networks trained on simplified contour images, with R-squared values near 0.9 for VGG19, Resnet50, InceptionV3 and Xception architectures. Saliency maps proved difficult to interpret and showed inconsistent results across architectures, in contrast to the decision trees. Additionally, some decision trees trained as surrogate models outperformed a decision tree trained on the actual outcomes without surrogation. Decision trees may be a more interpretable alternative to saliency methods. Moreover, the fact that we matched the performance of a decision tree without surrogation to that obtained by decision trees using knowledge distillation from neural networks is a great advantage since decision trees are inherently interpretable. Therefore, based on our findings, we think this approach would be the most recommendable choice for specialists as a diagnostic tool.

1. Introduction

Glaucoma is one of the main causes of blindness in the world and is characterized, in many cases, by the almost absence of symptoms until the disease is already quite advanced, hence it is known as silent blindness [1, 2]. Therefore, it is very important to diagnose it early to avoid its worst consequences.

Techniques based on Deep Learning have become, in recent years, a fundamental tool for automated disease diagnosis systems, especially those that use images as input data. In this work, we will focus on convolutional neural networks (CNN). The main reason for the proliferation of Deep Learning techniques is the high performance they provide compared to Machine Learning approaches based on tabular data, as long as a sufficient number of samples is available for training. On the contrary, these systems have a main drawback that can become critical in high stakes fields such as medicine and that is their black box nature with the consequent lack of mechanisms for explaining or interpreting the decisions made [3–9].

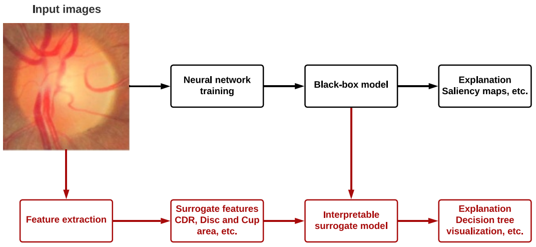

Input images
Neural network training
Black-box model
Explanation Saliency maps, etc.
Feature extraction
Surrogate features CDR, Disc and Cup area, etc.
Interpretable surrogate model
Explanation Decision tree visualization, etc.

Figure 1. Diagram showing the approach followed to obtain the surrogate model used in this work. A CNN is trained using the input fundus images to obtain a black-box model which can be explained using techniques like saliency maps. On the other hand, numerical clinical features are extracted from the images (surrogate features) and used to train an interpretable surrogate model such as a decision tree.

In an attempt to alleviate this drawback, different strategies have been proposed. One commonly used approach for interpreting CNN decisions from input images is the use of saliency or attribution maps [10]. Essentially, an attribution map algorithm assigns an importance measure to each pixel in the image by introducing perturbations and examining how the network response changes. Despite this popularity, several authors have raised concerns about their use in medical applications [11–15] due to their lack of consistency and robustness. In addition, since attribution maps offer only approximate insights of the effects of a neural network's inner workings, the information provided may be insufficient and of little use to the specialist.

A very different approach to the problem is based on the use of surrogate models [16]. A surrogate model within the scope of machine learning interpretability is an interpretable model that is trained to approximate the predictions of a black box model. Examples of algorithms which lead to interpretable models are linear regression or decision trees. Thus, we can draw conclusions about the black box model by interpreting the surrogate model. Surrogate models have been mostly used in problems with well-structured data because its application to Deep Learning models involving images is difficult since the inputs to these systems are arrays of thousands of pixels. As far as we know, the well-known LIME method [17] is the only case of surrogate model that has been used in this context for local interpretability after converting the image from a pixel representation to superpixels. Section 2 reviews work related to the use of surrogate models for interpretability in the field of medicine and, in particular, in the diagnosis of glaucoma.

In this paper, we propose a novel use of surrogate models applied to CNNs. The novelty lies in the fact that we not only change the original model for an interpretable one, but we also change the input features, replacing pixels by geometrical descriptive parameters of the optic disc, cup and neuroretinal rim commonly used by glaucoma specialists (see section 3.2). This approach to the problem has some similarities with the one described in [18] where dimensional reduction of the feature space is also carried out in order to improve interpretability. This is also in line with the well-known fact that, often with structured data, sparsity is a useful measure of interpretability given that humans can handle at most 3 to 5 meaningful items at a time [19]. Figure 1 shows an outline of the approach followed.

Consequently, the resulting interpretable surrogate model will attempt to mimic the predictions of the corresponding neural network model in such a way that its decisions can be interpreted in terms of features that make sense to the medical specialist. In contrast, explaining the exact mechanisms by which a CNN has arrived at a certain decision, where millions of operations and complex relations are involved, is unaffordable in practical cases, and these explanations would still be difficult to understand in terms of the usual medical criteria. In any case, it is important to keep in mind that even in the most common assumption of black box models based on tabular data, it cannot be guaranteed that the corresponding surrogate model will use the input features in the same way to make the prediction, especially when correlations exist between them [6].

As with all interpretability techniques, there exist advantages and disadvantages. In the case of surrogate models, one of the issues commonly raised is when the model is considered to be sufficiently close to the original black box model. In addition, it may happen that it approximates very well for one part of the data but not for another. In the case that it approximates the original model very well to the point of almost matching it, the question arises as to whether it still makes sense to use a black box model or to discard it and directly train the interpretable model with the original data [6]. Related to this, it will be shown that it is possible to make such a discard and even improve the performance of the interpretable model.

For this purpose, numerous experiments have been carried out with four different CNN architectures that have been trained with two types of input images. In one case the original images of the optic nerve head were used, and in the other a simplified version of these images was used showing only the disc and cup outlines on a uniform background set to a constant value, preserving only the geometry of these two anatomical structures. With regard to the surrogate models, most of the experiments were performed with decision trees because of their good interpretability properties in terms of simplicity, clarity and intuitiveness. In addition, to complement the information provided by the surrogate models, saliency maps have been calculated using some popular methods. All the detail about the neural network models, the input images, the extracted clinical features, and the surrogate models is provided in section 3. The description of the experiments and their discussion can be found in section 4.

In summary, the main contributions of this work are:

1. To determine to what extent it is possible to reproduce the decisions of CNNs with surrogate models whose inputs are geometric features of the optic disc, cup and rim, commonly used in glaucoma diagnosis.
2. Evaluate the influence on the above point of factors affecting the training of neural network models such as the type of input images used or the specific architecture of the network.
3. For those cases where the interpretable surrogate model is able to replicate what the neural network does, determine globally and locally the importance of the chosen features in the system predictions.
4. Critical evaluation of the explainability provided by surrogate models versus that obtained by calculating the corresponding saliency maps with different methods.
5. Obtaining an intrinsically interpretable model with performance equivalent to that of black box models based on neural networks.

The manuscript ends with some conclusions in section 5.

2. Related work

In this review of related work, we will mainly focus on interpretability techniques applied to CNNs for glaucoma diagnosis, with special emphasis on the use of surrogate models.

The use of tabular clinical data to fit interpretable machine learning models has been widely explored in glaucoma diagnosis, either alone [20, 21] or in combination with fundus images [22–24]. Another common approach is the extraction of relevant features from images. For example, Xu *et al* [25] developed a hierarchical system based on deep learning that, after a pre-diagnosis of a fundus image (also based on deep learning), apply another neural network to segment the disc, and two neural networks to segment the cup, depending on whether the pre-diagnosis assesses the image as glaucoma or not. Finally, they extract two features from the segmentation of these structures: mean cup-to-disc ratio (MCDR) and ISNT score. At the same time, another network segments the input full-fundus image to detect if there are RNFL defects (RNFLD). With these three features, they build a decision tree to produce the final diagnosis. This allows them to provide a feature-based explanation, i.e. to identify specific features such as the presence of RNFLD, the size of the MCDR, or a low ISNT score that contribute to a glaucoma diagnosis. Additionally, they are able to provide example-based explanations, which are side-by-side comparisons between the current case being analyzed and another case where the criteria for diagnosing glaucoma are not met. A similar approach is used in other studies as well, such as [26–28], where they segment the disc and cup and extract different features, such as vertical cup-to-disc ratio (vCDR), the shape of the disc and cup, the disc and cup area, rim size at inferior, superior, nasal, and temporal regions, etc and then train different algorithms to produce the final classification, with varying degrees of interpretability depending on the algorithm applied.

However, none of these works utilizes surrogate models for glaucoma diagnosis. As stated in the introduction and to the best of our knowledge, the LIME method is the only surrogate model used for interpretability of a CNN model for glaucoma diagnosis based on fundus images [17]. LIME is a post-hoc algorithm that is model-agnostic and can be applied to any black-box model to provide local interpretability. It does so by generating perturbed data in the neighborhood of the instance to be explained and fitting an interpretable linear model on the perturbed samples. To apply this algorithm to models that use images as input data, it is typical to convert the image from pixel representation to superpixels and apply LIME on the superpixel data. Following this approach, Kinger *et al* [29] proposed a CNN architecture to detect various ocular conditions, including glaucoma, and then used the LIME method to highlight which segments (superpixels) of the image contributed positively or negatively to the network's output. Gheisari *et al* [30] designed an architecture that combines a CNN with a recurrent neural network (RNN) to diagnose glaucoma on fundus videos and compared this approach to a CNN model that uses only fundus images. In this case, they used the LIME algorithm to show the regions of the image used to make the prediction of the combined CNN/RNN model. Since their input data is video, they averaged the intensity maps produced by LIME on sequential frames, demonstrating the importance of vascularized regions of the superior and inferior retina in the model prediction, which is consistent with previous studies.

In spite of its usefulness, the LIME method is not without its drawbacks. One problem is its instability and inconsistency between successive executions due to the random perturbations employed by the algorithm. This issue was highlighted by Li *et al* [31], who proposed G-LIME as a possible solution. Similarly, Visani *et al* [32] proposed OptiLIME for the same purpose and tested it on the NHANES I dataset, which models the risk of death using clinical measurements from 14 407 individuals across 79 features.

Another challenge of using LIME on image data is the selection of a suitable segmentation algorithm [33]. Additionally, as already mentioned, it is worth noting that LIME provides only local interpretations. To overcome this constraint, several methods have been proposed to use LIME for global interpretation by aggregating multiple local explanations, such as SP-LIME, developed by the same authors of LIME [17], GALE [34], or NormLime [35]. However, we could not find any application of these methods to medical diagnosis problems.

Apart from the above approaches, global surrogate models, such as decision trees, have been successfully employed to address other medical problems. In [36], the authors utilized various interpretability techniques, including global surrogate models, on a model that predicts the risk of developing hypertension based on cardiorespiratory fitness data. Similarly, [37] employed multiple AI models such as random forests, neural networks, and ensembles of neural networks to fit models for breast cancer prediction. Interpretability techniques, such as a global surrogate method based on a decision tree, individual conditional expectation plots, and Shapley values were used to analyze these models. Notably, a new random forest model was trained using the five most important features derived from the global surrogate method for the original random forest model, achieving better performance than using all the features. Moreover, it is worth mentioning that the WDBC dataset [38] used in this study provides a set of features extracted from digitized images of fine needle aspirates of breast masses, utilizing image processing algorithms.

As discussed in section 1, we have found only one work with some similarity to our proposal, namely [18], where the surrogate data models (SDMs) are introduced to interpret large-scale machine learning crisis prediction models. Their input data is a high-dimensional set of features, and their proposal is to approximate the black-box model using a low-dimensional set of features (surrogate features), some of which might not be present in the original feature set of the black-box model. They argue that by reducing the dimension of the feature set, the interpretation of the results becomes easier even if the SDMs do not use an intrinsically interpretable model. They also sustain that simpler surrogate models that use the original high-dimensional dataset do not improve substantially the interpretability of the model. A similar scenario arises in our problem when using images as input data, due to their inherently high-dimensionality nature, so that the reduction of dimensions is achieved by extracting from the images well-known clinical features that have been extensively validated for glaucoma diagnosis, as previously mentioned. However, in the work presented in [18], its application domain is totally different and no images are used as input data, hence the novelty of our approach.

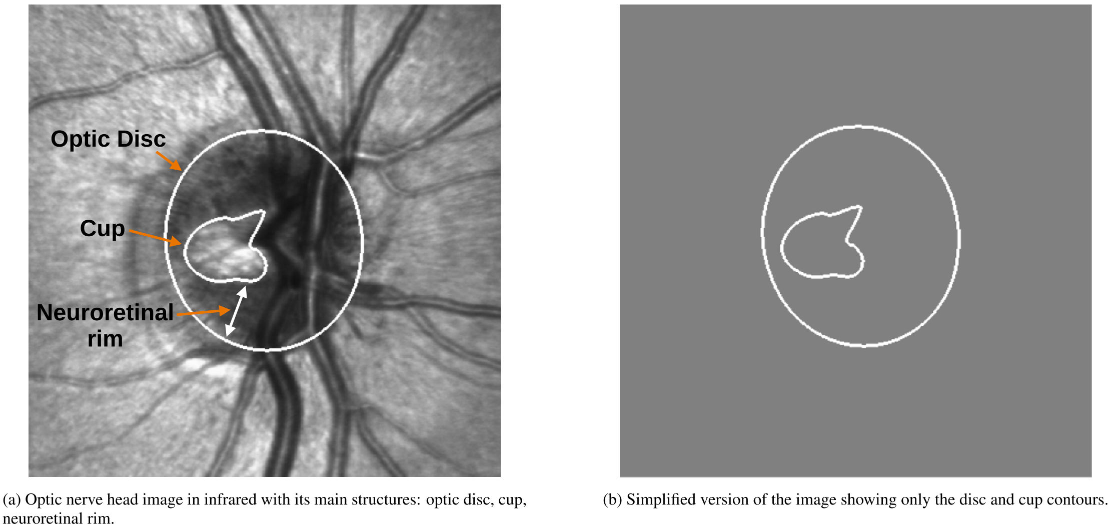
3. Materials and methods

3.1. Datasets

Two different datasets have been used in this work. Some experiments were carried out with our own dataset composed of 606 eye fundus images with which the Deep Learning models were trained and tested. From these images, 399 correspond to glaucoma cases and 207 to healthy subjects. The images were captured in infrared (815 nm) in the Hospital Universitario de Canarias with a Spectralis SD-OCT. The fact of using infrared images is only a matter of convenience to facilitate their acquisition due to the logistics of the hospital. Images in the visible could also have been used for the experiments we have performed. This dataset will be called 'Dataset IR' throughout the rest of this paper.

Patients with a diagnosis of primary or secondary open-angle glaucoma with untreated intraocular pressure greater than 21 mmHg were included in the study. The diagnosis of glaucoma was not only based on the observation of the eye fundus images but on the presence of reproducible defects in the white-white perimetry and/or morphological criteria based on a spectral domain optical coherence tomograph SD-OCT Spectralis with the Glaucoma Premium and Posterior Pole module. Both eyes were included if they met the inclusion criteria. Subjects with concomitant ocular pathology other than glaucoma, lower visual acuity of 20/40, refractive error greater than five diopters of spherical equivalent or three diopters of astigmatism, level of false positives, negatives and fixation errors equal to or greater than 25% in the visual field were excluded from the study. Patients with hypoplastic or oblique optic nerves were also excluded.

As mentioned in section 1, a simplified version of the eye fundus images showing only the contours of the disc and the cup on a gray background was also considered, as shown in an example in figure 2. More specifically, the pixels in the background of the image were assigned the value 128 and the pixels in the edges the value 255. These contours have been traced by an ophthalmologist specialized in the diagnosis of



(a) Optic nerve head image in infrared with its main structures: optic disc, cup, neuroretinal rim.

(b) Simplified version of the image showing only the disc and cup contours.

Figure 2. Example of an infrared (815 nm) image captured in the Hospital Universitario de Canarias with a Spectralis SD-OCT, showing the most relevant regions to diagnose glaucoma. Next, its corresponding simplified version displaying the contours of the disc and the cup on a gray background. The background is set to 128 and the contours to 255.

Table 1. Summary presenting the quantity of images within each dataset. It includes the total number of samples per class and the distribution of samples between the training and testing sets.

Dataset	Class	# of samples	Training samples	Test samples
Dataset IR	Glaucoma	399	319	80
Dataset IR	Healthy	207	166	41
Dataset C	Glaucoma	399	319	80
Dataset C	Healthy	207	166	41
Dataset XC	Glaucoma	721	511	210
Dataset XC	Healthy	1271	763	508

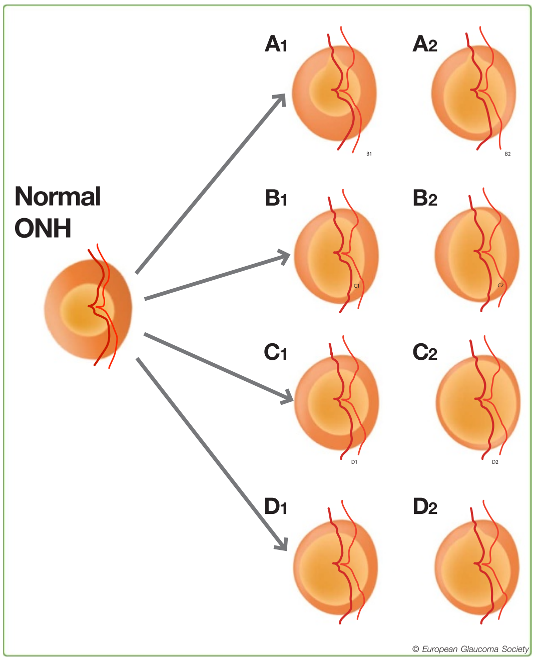
glaucoma, with more than 15 years of experience. In what follows, this variant of the first dataset will be called ‘Dataset C’.

The other dataset used in this work for training and testing the neural networks is an extended version of the previous one that includes, apart from our samples, other samples from very popular and publicly available image sets such as RIM–ONE DL, REFUGE and DRISHTI–GS. Thus, with this extension we have 1271 images corresponding to healthy subjects and 721 images corresponding to subjects with glaucoma. This set of images has only been used in its simplified version with the only information of the disc and cup contours, and will be called ‘Dataset XC’ in the following sections.

RIM–ONE DL [39] results from combining the three previous versions of the popular RIM-ONE dataset [40], [41]. This new version consists of 313 retinographies from normal subjects and 172 retinographies from patients with glaucoma. All of these images have been assessed by two experts and include a manual segmentation of the disc and cup. The REFUGE challenge database [42] consists of 1200 retinal images. About 10% of the dataset (120 samples) corresponds to glaucomatous subjects, including Primary Open Angle Glaucoma and Normal Tension Glaucoma. Ground truth segmentation of disc and cup is also included. This set of images is provided in three parts: training, validation and test, each with 400 images. In this work only the training and test partitions have been used, the validation one was ignored. DRISHTI-GS consists of 50 training and 51 testing images of which 70 correspond to subjects with glaucoma and 31 to healthy subjects [43]. For each image, manual segmentations were collected for both OD and cup region from four different human experts with varying clinical experience.

Each of the datasets used in this study underwent a partitioning process into training and testing sets. In the case of our dataset, across all its variations, we performed a random split, allocating 80% of the data for training and 20% for testing. For the other datasets, we adhered to their respective predefined splits, as published in the literature, to establish training and testing subsets.

Table 1 provides a concise overview of the datasets composition, detailing the number of samples per dataset and class. Additionally, it specifies the count of samples allocated for training and testing within each dataset.



© European Glaucoma Society

Figure 3. Progression of glaucomatous damage at the optic disc: early localized loss (A1), advancing to localized plus diffuse rim loss (A2). Early localized rim loss, polar notches (B1); more advanced polar notches (B2). Diffuse or concentric rim loss, early (C1); advanced (C2). Diffuse rim loss (D1), followed by localized rim loss (notch) (D2). Reproduced with permission from [44]. CC BY-NC-ND 4.0.

3.2. Disc, cup and rim geometrical features

Glaucoma diagnosis is a difficult problem due to the lack of reliable biomarkers so the diagnosis is usually based on a combination of information including medical history and both structural and functional longitudinal tests. In spite of this, in the case of diagnosis based on fundus imaging, some patterns have been identified that aid in the discrimination between healthy and glaucomatous eyes. As shown in figure 2, the optic disc usually appears in the image as a vertically oval shaped area. The cup is another important region inside the disc which usually appears as a brighter intensity area, as a result of the loss of nerve fibers caused by the disease. The part that remains between the cup and the disc is known as the neuroretinal rim.

As noted in [44, 45] glaucoma is characterized by progressive narrowing of the neuroretinal rim, taking the form of diffuse narrowing, localized notching, or a combination of both. Figure 3 shows the progression of glaucomatous damage at the optic disc. In this respect, there is a certain consensus when it comes to identifying certain areas of the rim as those that are usually the main affected in the case of suffering from glaucoma, depending on the state of the disease. Therefore, the geometry of the disc, cup and rim can provide very valuable information for the diagnosis of glaucoma.

Taking into account the above and taking as a reference what is exposed in [45], 14 features related to the shape and dimensions of the optic disc and cup have been calculated for this work, which are detailed in table 2. It should be noted that both the disc and the cup have been manually delimited by an expert with more than 15 years of experience in the field. Features have been extracted from these contours using standard image processing techniques.

The size of the disc (a_{OD}), cup (a_{OC}) and neuroretinal rim (a_{NR}) was calculated as the area in number of pixels of the corresponding region in each image. The optic cup size in relation to the optic disc size was also calculated:

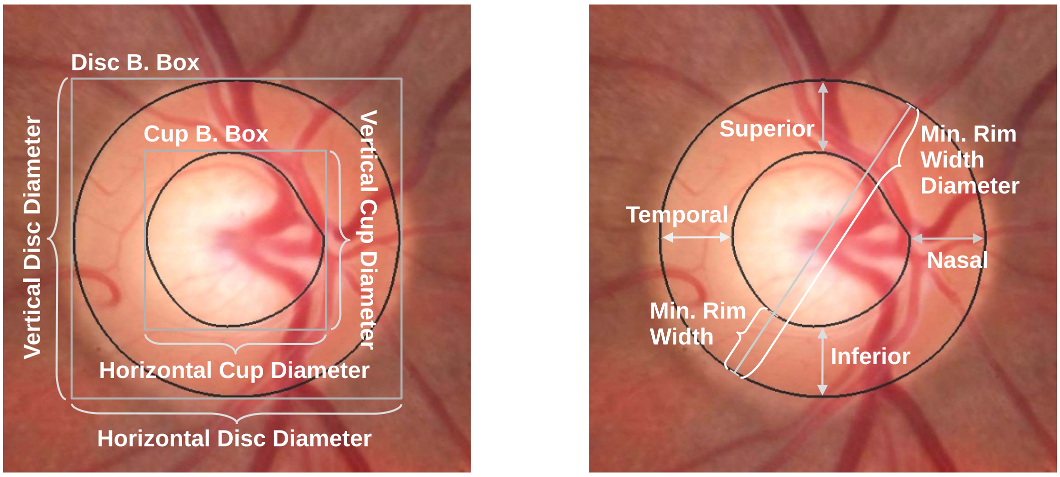
$$a_{\mathrm{CDR}} := \frac{a_{\mathrm{OC}}}{a_{\mathrm{OD}}}. \tag{1}$$

Quantifying the shape of an optic disc requires the delimitation of a bounding box (figure 4(a)) where the horizontal and vertical diameters are defined (h_d, v_d). The corresponding diameters for the optic cup are similarly defined (h_c, v_c). Then a shape parameter for the optic disc and the optic cup (s_{OD}, s_{OC}) can be calculated as:

$$s_{\mathrm{OD}} := \frac{h_d}{v_d} s_{\mathrm{OC}} \qquad := \frac{h_c}{v_c}. \tag{2}$$

Table 2. List of the 14 features used in this work related to the shape and dimensions of the optic disc, cup and neuroretinal rim. Their number, mathematical symbol and abbreviation are also indicated, as used throughout the rest of the article.

No.	Symbol (Abbr.)	Feature
1	a_{OD} (OD size)	Optic disc size
2	s_{OD} (OD shape)	Optic disc shape
3	a_{NR} (NR size)	Neuroretinal rim size
4	I-ISNT	Neuroretinal rim shape (I—ISNT rule)
5	S-ISNT	Neuroretinal rim shape (S—ISNT rule)
6	N-ISNT	Neuroretinal rim shape (N—ISNT rule)
7	T-ISNT	Neuroretinal rim shape (T—ISNT rule)
8	$w_{\mathrm{NR}}^{\mathrm{min.rim}}$ (NR width)	Neuroretinal rim shape (narrowest rim width)
9	w_{RDR} (RD ratio)	Rim-to-disc ratio
10	a_{OC} (OC size)	Optic cup size
11	s_{OC} (OC shape)	Optic cup shape
12	a_{CDR} (CDR size)	Optic cup size in relation to the optic disc size
13	v_{CDR} (V CDR)	Vertical cup-to-disc ratio
14	h_{CDR} (H CDR)	Horizontal cup-to-disc ratio



(a) Optic nerve head image depicting the bounding box of the disc and cup, and their horizontal and vertical diameters.

(b) Optic nerve head image representing the Inferior, Superior, Nasal, and Temporal sectors of the ISNT rule, and the minimum rim width and its corresponding disc diameter used to calculate the rim-to-disc ratio.

Figure 4. Sample optic nerve head images showing some of the most frequently used features to diagnose glaucoma. The ratio between the horizontal and vertical diameter of the disc provides information about its shape. Analogously for the cup. The ratio between the vertical diameter of the cup and the disc provides the vertical cup-to-disc ratio. Similarly for the horizontal cup-to-disc ratio. The ratio between the minimum rim width and its corresponding disc diameter results in the rim-to-disc ratio.

To capture the shape of the neuroretinal rim, different features were calculated. On the one hand, the thickness of the rim was determined at different points according to the ISNT rule [46] where I represents the Inferior sector of the optic disc, S the Superior sector, N the Nasal sector, and T the Temporal sector (figure 4(b)).

On the other hand, if $A_{\mathrm{OD}}^{\mathrm{min.rim}}$ represents the axis of the optic disc with the narrowest rim, its length ($L_{\mathrm{OD}}^{\mathrm{min.rim}}$) and the width of the neuroretinal rim in the same axis ($w_{\mathrm{NR}}^{\mathrm{min.rim}}$) (figure 4(b)) are obtained, and used to calculate the rim-to-disc ratio (w_{RDR}) [47]:

$$w_{\mathrm{RDR}} := \frac{w_{\mathrm{NR}}^{\mathrm{min.rim}}}{L_{\mathrm{OD}}^{\mathrm{min.rim}}}. \tag{3}$$

Finally, two other measures were calculated, the ratio of both vertical and horizontal cup and disc diameters ($v_{\mathrm{CDR}}, h_{\mathrm{CDR}}$) [48] defined as:

$$\begin{aligned} v_{\mathrm{CDR}} &:= \frac{v_c}{v_d} \\ h_{\mathrm{CDR}} &:= \frac{h_c}{h_d}. \end{aligned} \tag{4}$$

The vCDR is, without doubt, the most popular and studied geometrical feature for diagnosing glaucoma on fundus images. Orlando *et al* [42] show that using only this measure it is possible to train a classifier that achieves quite competitive results on the REFUGE dataset. On the other hand, the European Glaucoma Society [44] discourages its use as the sole measure for diagnosis.

3.3. Deep learning models

The datasets described in section 3.1 were used to train CNN models with different architectures. In this section, we present the methodology followed to train all these models. The results achieved are contained in section 4.1.

Four well–known CNN architectures were considered, VGG19 [49], ResNet50 [50], InceptionV3 [51], and Xception [52], available in the Keras module of the Tensorflow v2 package. The main motivation for using four different CNN architectures was to evaluate the influence of the specific network architecture on the results obtained. To adapt the models to the problem, we replaced the top layer of each network with a GlobalAveragePooling2D layer, followed by a DropOut layer and a Dense layer with 2 outputs using the SoftMax activation function. The input layer was set to $224 \times 224 \times 3$ for ResNet50 and VGG19 and $299 \times 299 \times 3$ for InceptionV3 and Xception. The DropOut value was set to 0.5 for VGG19 and 0.2 for InceptionV3, ResNet50, and Xception. Additionally, we kept the BatchNormalization layers in inference mode for InceptionV3, ResNet50, and Xception, to avoid updating the non-trainable weights of these layers during the training phase.

In all cases, to further ensure the robustness of our models, we employed a 5-fold cross-validation technique on the training set. We trained all the models using the same strategy. First, we froze the pre-trained base model and trained the new top layer for 200 epochs using an RMSprop optimizer with a learning rate of 1×10^{-6} and categorical cross-entropy loss. Second, we unfroze the base model and trained the entire model end-to-end for 250 epochs using the same optimizer, learning rate, and loss function. We used a batch size of 32 for all models, except for Xception, where we used a batch size of 16 due to memory constraints.

We applied the pre-processing function included in Keras for each of the models. For ResNet50 and VGG19, we transformed each image from RGB to BGR and centered each channel to zero by subtracting the mean values of the ImageNet dataset. For InceptionV3 and Xception, we scaled each image between -1 and 1.

To address class imbalance, we set class weights when training the models. The weight for each class was computed per fold using the formula in equation (5), where W_c is the weight for class c, N is the total number of data samples, and N_c is the number of samples of class c.

$$W_c = \frac{N}{2 * N_c} \tag{5}$$

To prevent overfitting, we applied data augmentation to the input samples, consisting of random rotations $(-15^\circ, 15^\circ)$, horizontal flip and random brightness adjustments in the range [0.5, 1.5].

We selected the best-performing model for each fold per network architecture based on the epoch that maximized the validation accuracy averaged across the 5 folds. In total, we obtained 20 models per dataset, 5 different models per architecture.

3.4. Evaluation methodology

The methodology used to evaluate how well the black box model f, in our case a neural network model, is approximated by the interpretable surrogate model g is the one proposed in [16]. Accordingly, the following steps have to be accomplished to obtain the surrogate model (see algorithm 1):

1. Choose a dataset X, which is a matrix where each column represents one of the geometrical features outlined in section 3.2 and each row an image. These features are extracted from the set of images that was used for training the neural network models.
2. Obtain predictions from the trained neural network models for the selected set of images. The ouput probability of being a glaucomatous eye was taken as the target for this problem.
3. Choose an interpretable model. In this work, we have focused on linear regression and decision tree regression models. Note the need to pose the problem as a regression since we are considering a continuous variable as the target, as mentioned in the previous step. Linear regression was chosen because of its simplicity, while decision tree regression was chosen for its ability to provide good explanations and natural visualization through its nodes and edges.
4. Train the chosen interpretable model on the dataset X and the predictions obtained in steps 1 and 2.

Algorithm 1. Pseudocode of the process to obtain a surrogate model.

```
# Step 1: Extract features from the images used to train the CNNs
I = load_image_dataset()
X = extract_features(I) # X is a matrix with geometrical features

# Step 2: Obtain predictions from the trained neural network models
Y = neural_network(I) # Y are the predictions by the CNN

# Step 3: Choose an interpretable model
selected_model = ''linear_regression'' | ''decision_tree_regression''

# Step 4: Train the chosen interpretable model
surrogate_model = train_interpretable_model(selected_model, X, Y)
```

Once the surrogate model *g* is trained, we can evaluate how well it approximates *f* by using the R-squared measure [16]:

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{\sum_{i=1}^{n}\left(\hat{y}_*^{(i)} - \hat{y}^{(i)}\right)^2}{\sum_{i=1}^{n}\left(\hat{y}^{(i)} - \bar{\hat{y}}\right)^2}. \tag{6}$$

In the equation above, $\hat{y}_*^{(i)}$ represents the prediction of the interpretable model *g* for the *i*th instance, $\hat{y}^{(i)}$ represents the prediction of *f*, the black box model, for the *i*th instance, and $\bar{\hat{y}}$ represents the mean of the predictions of *f*. SSE refers to the sum of squares error, while SST refers to the sum of squares total. As Molnar points out [16], the R-squared measure provides insight into how much variance can be accounted for by the surrogate model. A low SSE and an R-squared value close to 1 indicate that the surrogate model is an accurate approximation of the black box model. Conversely, a high SSE and an R-squared value close to 0 suggest that the surrogate model is unable to explain the behavior of the black box model.

4. Experimental results and discussion

Different experiments have been carried out following the methodology described in section 3. Details of these experiments can be found in the following subsections.

4.1. Evaluation of the deep learning models

Following the procedure described in section 3.3, we performed a series of experiments using the different datasets detailed in section 3.1. Our own dataset (Dataset IR), as already mentioned, comprises 606 images and was randomly partitioned into training and testing sets with an 80/20 ratio. We evaluated the performance of the trained models on this test set, consisting of 41 samples from healthy subjects and 80 samples from glaucoma subjects. The results obtained are presented in table 3, which reports the worst and best performance per network architecture in terms of balanced accuracy. Balanced accuracy, which is the arithmetic mean of sensitivity and specificity, has been shown to be a good metric for evaluating classifiers with imbalanced classes [53]. Other commonly used metrics are also shown.

The simplified version of our dataset (Dataset C) showing only the disc and cup contours was used to train another 20 models under the same conditions as before and using the same hyper-parameters except for the data augmentation technique which did not include random brightness adjustments because it makes no sense for this type of images. The results, as presented in table 4, show a slight decrease in the performance of the models with respect to those trained with the original images but still can be considered as competitive despite not having access to all the image content. This observation supports our hypothesis that the geometry of the disc and cup carries crucial diagnostic information our neural network models effectively exploit. We also compared the performance of these models with that of two expert ophthalmologists who were asked to make a diagnosis based on the simplified images alone. Interestingly, the CNN models outperformed the experts in all cases, suggesting that the neural networks do a excellent job with this limited information.

The second dataset (Dataset XC) includes only simplified images and, as discussed in section 3.1, merges our data with RIM-ONE DL, REFUGE, and DRISHTI-GS. In this case, as also described in that section, our dataset was partitioned as in the first two experiments, while the remaining datasets were split according to their respective published splits for training and testing. As a result, 763 images from healthy subjects and 511 images from subjects with glaucoma were available for training. The training of these models was also

Table 3. Summary of the results obtained by evaluating the models on our test dataset with the original infrared images (Dataset IR). For simplicity, only the results corresponding to the minimum and maximum balanced accuracy (B. Accuracy) are displayed.

Network	Fold	Sensitivity	Specificity	Accuracy	B. Accuracy	F1 score	AUROC	AUPR
VGG19	5	0.8750	0.8049	0.8512	0.8399	0.8861	0.9125	0.9489
VGG19	2	0.9375	0.8293	0.9008	0.8834	0.9259	0.9550	0.9721
ResNet50	5	0.8375	0.7805	0.8182	0.8090	0.8590	0.9200	0.9527
ResNet50	3	0.9250	0.8537	0.9008	0.8893	0.9250	0.9413	0.9624
InceptionV3	3	0.9250	0.8293	0.8926	0.8771	0.9193	0.9466	0.9591
InceptionV3	4	0.9625	0.8780	0.9339	0.9203	0.9506	0.9607	0.9749
Xception	2	0.9000	0.8780	0.8926	0.8890	0.9172	0.9530	0.9661
Xception	1	0.9500	0.8780	0.9256	0.9140	0.9441	0.9325	0.9584

Table 4. Summary of the results obtained by evaluating the models on our test dataset with the simplified images (Dataset C). For simplicity, only the results corresponding to the minimum and maximum balanced accuracy (B. Accuracy) are displayed. The results achieved by the glaucoma specialists on the same test set are also shown.

Architecture / Specialist	Fold	Sensitivity	Specificity	Accuracy	B. Accuracy	F1 score	AUROC	AUPR
VGG19	4	0.8375	0.8293	0.8347	0.8334	0.8701	0.9046	0.9330
VGG19	1	0.9125	0.8293	0.8843	0.8709	0.9125	0.9253	0.9519
ResNet50	4	0.8625	0.8293	0.8512	0.8459	0.8846	0.9183	0.9539
ResNet50	5	0.8625	0.8780	0.8678	0.8703	0.8961	0.9247	0.9569
InceptionV3	1	0.8625	0.8293	0.8512	0.8459	0.8846	0.9296	0.9642
InceptionV3	2	0.9000	0.8780	0.8926	0.8890	0.9172	0.9293	0.9643
Xception	4	0.8125	0.8537	0.8264	0.8331	0.8609	0.9226	0.9585
Xception	3	0.8250	0.8780	0.8430	0.8515	0.8742	0.9195	0.9533
Specialist 1	—	0.8000	0.8293	0.8099	0.8146	0.8477	—	—
Specialist 2	—	0.7875	0.8537	0.8099	0.8206	0.8456	—	—

Table 5. Summary of the results obtained per neural network architecture by evaluating the models according to different metrics on the extended test dataset (Dataset XC), which joins our test set with that of RIM–ONE DL, REFUGE, and DRISHTI–GS. For simplicity, only the results corresponding to the minimum and maximum balanced accuracy (B. Accuracy) are displayed.

Architecture	Fold	Sensitivity	Specificity	Accuracy	B. Accuracy	F1 score	AUROC	AUPR
VGG19	2	0.8667	0.9134	0.8997	0.8900	0.8349	0.9560	0.9172
VGG19	4	0.8857	0.9193	0.9095	0.9025	0.8513	0.9661	0.9280
ResNet50	2	0.8429	0.9311	0.9053	0.8870	0.8389	0.9571	0.9190
ResNet50	4	0.8810	0.9154	0.9053	0.8982	0.8447	0.9657	0.9377
InceptionV3	1	0.8238	0.9429	0.9081	0.8834	0.8398	0.9580	0.9218
InceptionV3	3	0.8714	0.9311	0.9136	0.9013	0.8551	0.9607	0.9250
Xception	1	0.8571	0.9094	0.8942	0.8833	0.8257	0.9555	0.9164
Xception	4	0.8667	0.9232	0.9067	0.8949	0.8445	0.9577	0.9228

carried out under the same conditions as described for the previous experiments. The results for the test set, consisting of 508 samples from healthy subjects and 210 samples from glaucoma subjects, are displayed in table 5. As can be seen on the table, in terms of balanced accuracy, the performance achieved by the models trained on the Dataset XC is generally higher than that of the models trained on the Dataset C alone.

Overall, the trained models performed well in all three scenarios evaluated. This is crucial as a poorly performing black box model renders the interpretation of the surrogate model pointless, making the black box model itself insignificant.

4.2. Evaluation of the surrogate models

Figure 5 shows the plots of the average R-squared values obtained for each of the architectures, both for our image dataset (Dataset IR) and for its simplified version (Dataset C), also with two different types of surrogate models, linear regression and decision trees. Decision tree models were trained using grid search in their parameter space, exploring all posible combinations of maximum depths of 3, 4 and 5 and minimum number of samples per leaf node of 1, 2, 3, 4, 5, 8, 10 and 15. Both linear regression and decision tree models were trained using cross validation. The averaged results of the 3 folds for each architecture are shown. On the other hand, figure 6 shows an example scatter plot to compare the predictions of one of the folds for the VGG19 architecture and the corresponding surrogate decision tree for both the original and contour images.

Important differences are observed between the models trained with the original images and those trained with the equivalent contour images. While with the original images, the average R-squared values

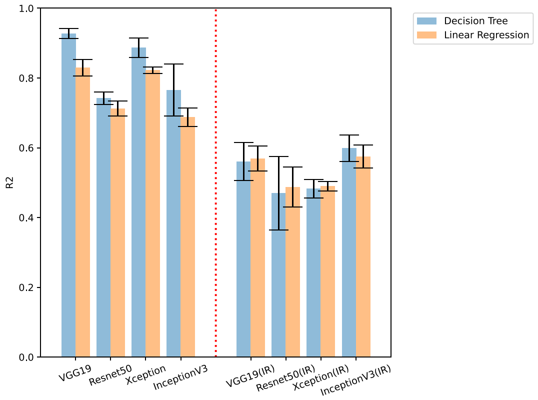


Figure 5. Average R-squared values obtained for each of the architectures, both for our image dataset (IR) and for its simplified version (Dataset C), with two different types of surrogate models, linear regression and decision trees.

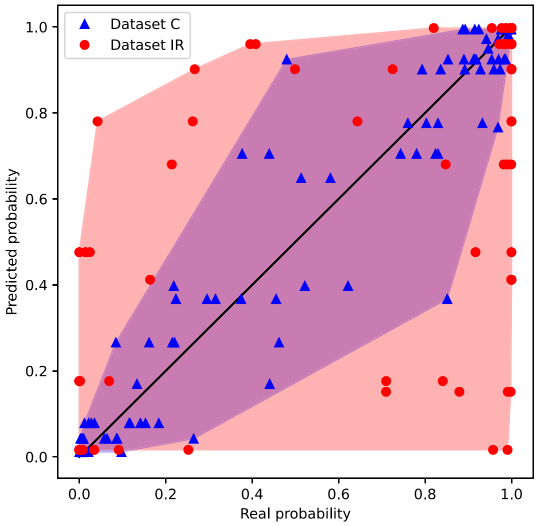


Figure 6. Scatter plot comparing the output probabilities of a VGG19 model (Real probability) against those of its corresponding surrogate decision tree (Predicted probability), for the Dataset IR and the Dataset C. As can be seen the predicted probabilities are closer to the real ones in the case of the Dataset C.

range between 0.5 and 0.6, in the case of the contour images, the values are considerably higher, even exceeding 0.9. This seems to show that the surrogate system, trained with the geometric features, finds it easier to mimic the prediction of the neural networks when the information with which they are trained is of the same nature, being only the contours of the disc and the cup. This is also reflected in the example in figure 6. On the other hand, there are slight differences between using a linear regressor or a decision tree as a surrogate system. In the case of contour images, the decision tree provides better average R-squared values for all architectures, which seems logical since it is able to generate more general nonlinear approximations. However, for the models trained with the original images, the differences on average between the two types of models are very small for all architectures, even with an advantage in some cases for linear regression.

There is no clear criteria as to when a surrogate model should be considered adequate in order to be confident that it sufficiently approximates the black box model. It is not clear what the best cut-off point for R-square is, although some publications speak of 80% of the explained variance [16]. In any case, from the experiments described above it seems clear that in the problem addressed, the most valid option is the one represented by the surrogate models based on decision trees that approximate the predictions of the CNNs trained with the disc and cup contours. For this reason, the experiments that we will describe in the remainder have been raised in this scenario. In addition, to carry out a better evaluation, we have used the neural network models trained with the extended set (Dataset XC) as explained in section 4.1. Figure 7 shows the average values of R-squared obtained for these models. It is observed how they are even higher than those shown in figure 5. Figure 8 displays a scatter plot akin to that seen in figure 6, allowing us to compare the

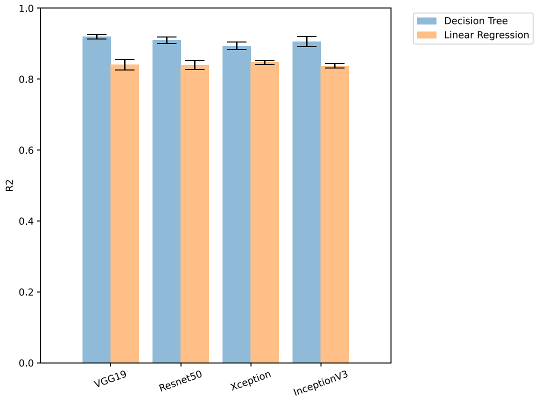


Figure 7. Average R-squared values obtained for each of the architectures with the Dataset XC described in section 3.1, with two different types of surrogate models, linear regression and decision trees.

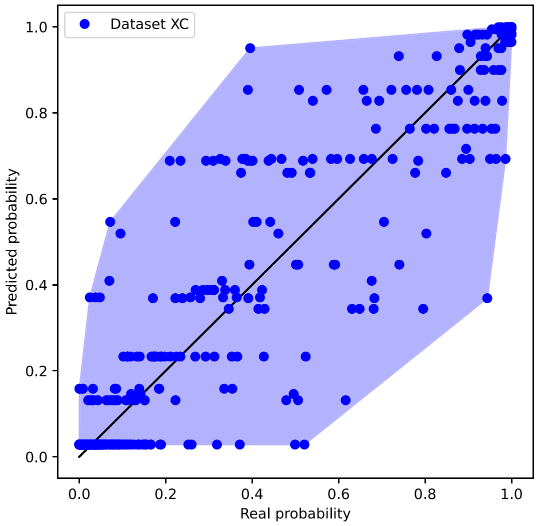


Figure 8. Scatter plot comparing the output probabilities of a VGG19 model (Real probability) against those of its corresponding surrogate decision tree (Predicted probability), for the Dataset XC.

predictions from one of the folds for the VGG19 architecture with the corresponding surrogate decision tree for Dataset XC.

4.3. Surrogate model interpretation

As already mentioned, decision trees are often used as surrogate models because of their good properties of simplicity, clarity and intuitiveness. Feature importance in decision trees refers to the calculation of the usefulness of each feature in predicting the target variable. Decision trees make splits in the data based on the values of the input features, and the importance of each feature is determined by how much these splits improve the purity of the resulting groups of data. In our case, the importance of each feature is determined by its contribution to the reduction of the mean squared error (MSE).

When an observation from the dataset is evaluated by the splitting rule, it either satisfies the rule and goes to the left of the node, or it does not satisfy the rule and goes to the right. This process continues until a terminal node is reached, at which point the prediction for the observation is made based on the values of the target variable in that group.

Let us denote each node as $N_m, m \in 1,..,k$ where k is the number of nodes of the tree. Every node in the decision tree, except for the leaf nodes at the final depth, has two child nodes: a left child N_m^l and a right child N_m^r.

Figures 9–12, show decision trees fitted to the predictions made by each architecture with the extended dataset (XC). Specifically, the trees corresponding to the fold for which the highest R-squared value has been obtained for each architecture are shown. The color of each node represents the predicted value. Lower

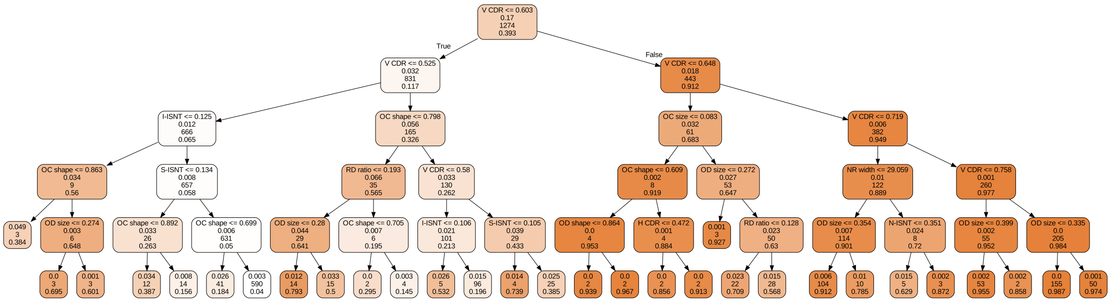

V CDR <= 0.603 0.17 1274 0.393
True
False
V CDR <= 0.525 0.032 831 0.117
V CDR <= 0.648 0.018 443 0.912
I-ISNT <= 0.125 0.012 666 0.065
OC shape <= 0.798 0.056 165 0.326
OC size <= 0.083 0.032 61 0.683
V CDR <= 0.719 0.006 382 0.949
OC shape <= 0.863 0.034 9 0.56
S-ISNT <= 0.134 0.008 657 0.058
RD ratio <= 0.193 0.066 35 0.565
V CDR <= 0.58 0.033 130 0.262
OC shape <= 0.609 0.002 8 0.919
OD size <= 0.272 0.027 53 0.647
NR width <= 29.059 0.01 122 0.889
V CDR <= 0.758 0.001 260 0.977
0.049 3 0.384
OD size <= 0.274 0.003 6 0.648
OC shape <= 0.892 0.033 26 0.263
OC shape <= 0.699 0.006 631 0.05
OD size <= 0.28 0.044 29 0.641
OC shape <= 0.705 0.007 6 0.195
I-ISNT <= 0.106 0.021 101 0.213
S-ISNT <= 0.105 0.039 29 0.433
OD shape <= 0.864 0.0 4 0.953
H CDR <= 0.472 0.001 4 0.884
0.001 3 0.927
RD ratio <= 0.128 0.023 50 0.63
OD size <= 0.354 0.007 114 0.901
N-ISNT <= 0.351 0.024 8 0.72
OD size <= 0.399 0.002 55 0.952
OD size <= 0.335 0.0 205 0.984
0.0 3 0.695
0.001 3 0.601
0.034 12 0.387
0.008 14 0.156
0.026 41 0.184
0.003 590 0.04
0.012 14 0.793
0.033 15 0.5
0.0 2 0.295
0.003 4 0.145
0.026 5 0.532
0.015 96 0.196
0.014 4 0.739
0.025 25 0.385
0.0 2 0.939
0.0 2 0.967
0.0 2 0.856
0.0 2 0.913
0.023 22 0.709
0.015 28 0.568
0.006 104 0.912
0.01 10 0.785
0.015 5 0.629
0.002 3 0.872
0.002 53 0.955
0.002 2 0.858
0.0 155 0.987
0.001 50 0.974

Figure 9. Decision Tree for the InceptionV3 arquitecture using the Dataset XC as explained in section 4.3.

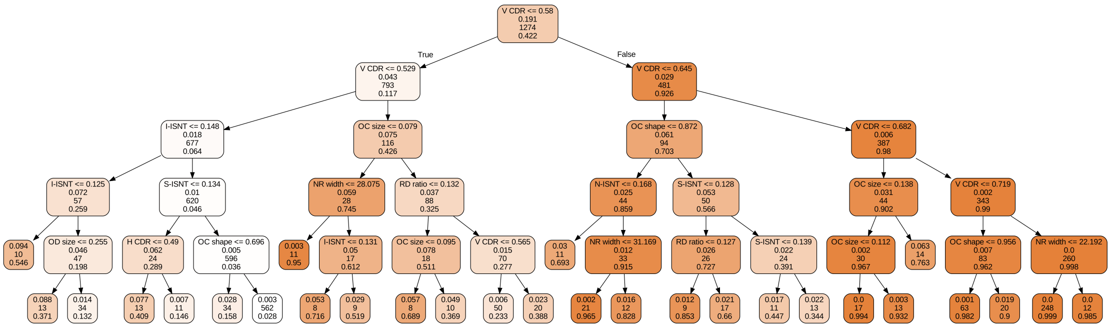

V CDR <= 0.58 0.191 1274 0.422
True
False
V CDR <= 0.529 0.043 793 0.117
V CDR <= 0.645 0.029 481 0.926
I-ISNT <= 0.148 0.018 677 0.064
OC size <= 0.079 0.075 116 0.426
OC shape <= 0.872 0.061 94 0.703
V CDR <= 0.682 0.006 387 0.98
I-ISNT <= 0.125 0.072 57 0.259
S-ISNT <= 0.134 0.01 620 0.046
NR width <= 28.075 0.059 28 0.745
RD ratio <= 0.132 0.037 88 0.325
N-ISNT <= 0.168 0.025 44 0.859
S-ISNT <= 0.128 0.053 50 0.566
OC size <= 0.138 0.031 44 0.902
V CDR <= 0.719 0.002 343 0.99
0.094 10 0.546
OD size <= 0.255 0.046 47 0.198
H CDR <= 0.49 0.062 24 0.289
OC shape <= 0.696 0.005 596 0.036
0.003 11 0.95
I-ISNT <= 0.131 0.05 17 0.612
OC size <= 0.095 0.078 18 0.511
V CDR <= 0.565 0.015 70 0.277
0.03 11 0.693
NR width <= 31.169 0.012 33 0.915
RD ratio <= 0.127 0.026 26 0.727
S-ISNT <= 0.139 0.022 24 0.391
OC size <= 0.112 0.002 30 0.967
0.063 14 0.763
OC shape <= 0.956 0.007 83 0.962
NR width <= 22.192 0.0 260 0.998
0.088 13 0.371
0.014 34 0.132
0.077 13 0.409
0.007 11 0.146
0.028 34 0.158
0.003 562 0.028
0.053 8 0.716
0.029 9 0.519
0.057 8 0.689
0.049 10 0.369
0.006 50 0.233
0.023 20 0.388
0.002 21 0.965
0.016 12 0.828
0.012 9 0.853
0.021 17 0.66
0.017 11 0.447
0.022 13 0.344
0.0 17 0.994
0.003 13 0.932
0.001 63 0.982
0.019 20 0.9
0.0 248 0.999
0.0 12 0.985

Figure 10. Decision Tree for the VGG19 arquitecture using the Dataset XC as explained in section 4.3.

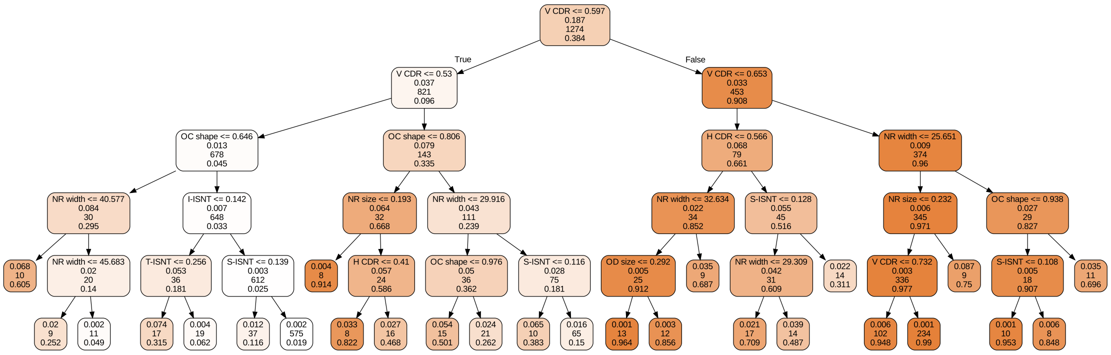

V CDR <= 0.597 0.187 1274 0.384
True
False
V CDR <= 0.53 0.037 821 0.096
V CDR <= 0.653 0.033 453 0.908
OC shape <= 0.646 0.013 678 0.045
OC shape <= 0.806 0.079 143 0.335
H CDR <= 0.566 0.068 79 0.661
NR width <= 25.651 0.009 374 0.96
NR width <= 40.577 0.084 30 0.295
I-ISNT <= 0.142 0.007 648 0.033
NR size <= 0.193 0.064 32 0.668
NR width <= 29.916 0.043 111 0.239
NR width <= 32.634 0.022 34 0.852
S-ISNT <= 0.128 0.055 45 0.516
NR size <= 0.232 0.006 345 0.971
OC shape <= 0.938 0.027 29 0.827
0.068 10 0.605
NR width <= 45.683 0.02 20 0.14
T-ISNT <= 0.256 0.053 36 0.181
S-ISNT <= 0.139 0.003 612 0.025
0.004 8 0.914
H CDR <= 0.41 0.057 24 0.586
OC shape <= 0.976 0.05 36 0.362
S-ISNT <= 0.116 0.028 75 0.181
OD size <= 0.292 0.005 25 0.912
0.035 9 0.687
NR width <= 29.309 0.042 31 0.609
0.022 14 0.311
V CDR <= 0.732 0.003 336 0.977
0.087 9 0.75
S-ISNT <= 0.108 0.005 18 0.907
0.035 11 0.696
0.02 9 0.252
0.002 11 0.049
0.074 17 0.315
0.004 19 0.062
0.012 37 0.116
0.002 575 0.019
0.033 8 0.822
0.027 16 0.468
0.054 15 0.501
0.024 21 0.262
0.065 10 0.383
0.016 65 0.15
0.001 13 0.964
0.003 12 0.856
0.021 17 0.709
0.039 14 0.487
0.006 102 0.948
0.001 234 0.99
0.001 10 0.953
0.006 8 0.848

Figure 11. Decision Tree for the ResNet50 arquitecture using Dataset XC as explained in section 4.3.

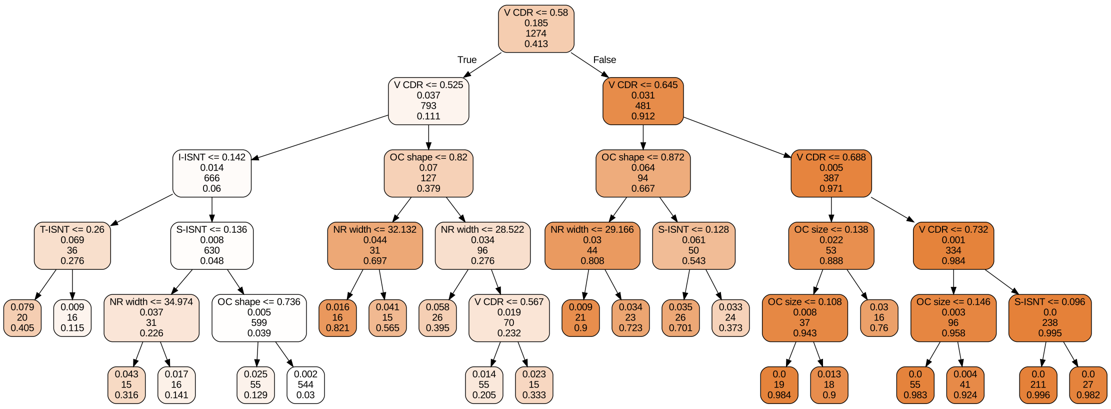

V CDR <= 0.58 0.185 1274 0.413
True
False
V CDR <= 0.525 0.037 793 0.111
V CDR <= 0.645 0.031 481 0.912
I-ISNT <= 0.142 0.014 666 0.06
OC shape <= 0.82 0.07 127 0.379
OC shape <= 0.872 0.064 94 0.667
V CDR <= 0.688 0.005 387 0.971
T-ISNT <= 0.26 0.069 36 0.276
S-ISNT <= 0.136 0.008 630 0.048
NR width <= 32.132 0.044 31 0.697
NR width <= 28.522 0.034 96 0.276
NR width <= 29.166 0.03 44 0.808
S-ISNT <= 0.128 0.061 50 0.543
OC size <= 0.138 0.022 53 0.888
V CDR <= 0.732 0.001 334 0.984
0.079 20 0.405
0.009 16 0.115
NR width <= 34.974 0.037 31 0.226
OC shape <= 0.736 0.005 599 0.039
0.016 16 0.821
0.041 15 0.565
0.058 26 0.395
V CDR <= 0.567 0.019 70 0.232
0.009 21 0.9
0.034 23 0.723
0.035 26 0.701
0.033 24 0.373
OC size <= 0.108 0.008 37 0.943
0.03 16 0.76
OC size <= 0.146 0.003 96 0.958
S-ISNT <= 0.096 0.0 238 0.995
0.043 15 0.316
0.017 16 0.141
0.025 55 0.129
0.002 544 0.03
0.014 55 0.205
0.023 15 0.333
0.0 19 0.984
0.013 18 0.9
0.0 55 0.983
0.004 41 0.924
0.0 211 0.996
0.0 27 0.982

Figure 12. Decision Tree for the Xception arquitecture using the Dataset XC as explained in section 4.3.

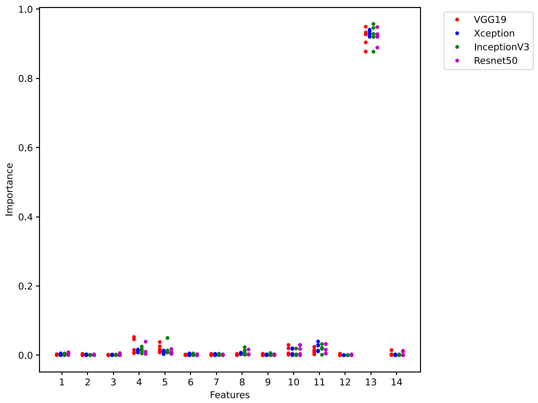


Figure 13. Feature importance for the surrogate decision trees for each of the architectures with the Dataset XC as explained in section 4.3.

values are represented by lighter colors, while higher values are represented by darker tones of orange. In these trees, each node contains some information in the following order (from top to bottom):

1. A splitting rule that uses the feature f and a threshold value of t. If an observation's f value is less than or equal to t, it follows the path to the left of the node, and if it is greater than t, it follows the path to the right. Leaf nodes do not have a splitting rule.
2. The mean squared error. The statistic that is used as the splitting criteria. It is calculated with the following equations:

$$\bar{y}_m = \frac{1}{n_m}\sum_{y\in Q_m}^{n} y \tag{7}$$

$$H(N_m) = \frac{1}{n_m}\sum_{y\in Q_m}^{n} (y-\bar{y}_m)^2 \tag{8}$$

   where Q_m is the data with n_m samples at node m.
3. The samples value represents the number of observations in the training dataset that follows the path to this node.
4. The predicted value for the node $\bar{y}_m$ equation (7). If an observation follows the path that leads to this node, that would be the predicted value for this obervation.

An essential property of a node with child nodes is its node importance equation (9). The node importance is calculated using a specific formula, where the basic idea is that if the mean squared error in the child nodes is low, then the importance of the node, and particularly the feature used in its splitting rule, is high. In other words, the lower the MSE in the children, the greater the importance of the node and the feature used to create the split.

$$U(N_m) = W(N_m)H(N_m) - W\left(N_m^l\right)H\left(N_m^l\right) - W(N_m^r)H(N_m^r) \tag{9}$$

where the weight $W(N_m) \in (0,1]$ is determined for every node in the tree and is found by dividing the number of observations in that node by the total number of observations in the dataset.

The importance of each feature is calculated by summing the importance of each node that uses that feature as a splitting criterion. Figure 13 shows the feature importances obtained with the trained decision trees for each architecture and fold, for the Dataset XC. Thus, what we observe is the relevance of each feature in the trained decision tree model to mimic the predictions of the corresponding CNN, i.e. the probability that the input image corresponds to an eye with glaucoma. In this regard, it is quite clear that there is one feature that stands out above the others and that is the vCDR. We have already referred to it in section 3.2, as it is a very popular measure for the diagnosis of glaucoma. The importance value assigned to it in the four network architectures considered seems to corroborate this popularity. It does not mean that the neural network is calculating the cup-to-disc ratio as such, but rather that it plays a very prominent role in the decisions made by the surrogate model, perhaps because the network learns to calculate features that may be

Table 6. The five most important features, on average, for the surrogate decision trees for each architecture, with the Dataset XC as explained in section 4.3

Architecture	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
VGG19	V CDR	I-ISNT	S-ISNT	OC shape	OC size
Xception	V CDR	OC shape	I-ISNT	S-ISNT	OC size
InceptionV3	V CDR	OC shape	S-ISNT	I-ISNT	NR width
ResNet50	V CDR	OC shape	OC size	I-ISNT	S-ISNT

strongly correlated with it. However, this is a somewhat controversial measure, as some authors point it out as not too reliable for the diagnosis of this pathology [44, 45, 54]. Jonas *et al* [45] note that the interindividual variability of cup-to-disc ratios and their dependence on optic disc size should also be taken into account. The cup-to-disc ratio is a relative measure independent of cup and disc size, and, also, of the magnification of the camera used but the optic disc size is an absolute measure. In fact, although to a much lesser extent, figure 13 shows other features that contribute to model prediction which are absolute measures such as the inferior and superior rim widths given by I-ISNT and S-ISNT, and the size of the optic cup, OC size. Some other features such as OC shape and NR width also contribute globally to the predictions but still to a lesser extent. Table 6 shows the five most important features for the surrogate decision trees per architecture, on average. Kumar *et al* [54] conclude in their article that the rim-to-disc ratio outperforms cup-to-disc ratio for glaucoma prescreening. However, according to the results found, in none of the 20 surrogate models considered does this characteristic seem to be of relevant importance.

Apart from this global analysis of the importance of the features, it is also interesting to do a more local analysis based on the representation of the specific decision trees shown.

There are common patterns that are repeated in the trees of all architectures and are used in most of the training set samples. One common pattern that can be found in all trees is the use of a split at the root node which implies that the vCDR is used around a value of 0.6. This generates a left sub-tree in which the lower values are predicted and a right sub-tree in which the higher values are predicted, in general.

Within the left subtrees of all architectures, a large number of samples in the training set with low glaucoma probability values follow a pathway in which the I-ISNT, S-ISNT, and OC shape features are used. The order in which these features are used varies by architecture, but their use appears consistent across all architectures for most of the clearly negative glaucoma cases.

In the subtrees on the right, the vast majority of samples with a high probability value for glaucoma follow a trajectory marked by the use of V CDR.

There are also specific patterns with which each architecture defines paths in the tree to process a smaller set of samples, for which the probability of glaucoma has intermediate values.

The tree corresponding to the InceptionV3 architecture (figure 9) uses OC Size, OD Size and NR Width to decrease the predicted value of samples traversing the right subtree, and OC Shape and RD Ratio to slightly increase the predicted value of samples traversing its left subtree.

The tree corresponding to the VGG19 architecture (figure 10) uses the OC Shape and OC Size features in the opposite way, i.e. it uses OC Shape to decrease the predicted value in the right subtree and OC Size to increase it in the left subtree.

The tree corresponding to the ResNet50 architecture (figure 11) prefers to use NR Size to increase the predicted value of samples traversing its left subtree instead of RD Ratio, and uses H CDR instead of OC Size to decrease the predicted value of samples traversing its right subtree.

Finally, the tree corresponding to the Xception architecture (figure 12) mainly uses OC Shape in both subtrees to adjust the predicted value. In the case of the left subtree, a threshold of 0.82 is used in this feature to increase the predicted value for samples that are below it. For the right subtree, a threshold of 0.87 is used to reduce the predicted value of the samples that exceed it.

4.4. Comparison with saliency maps

As mentioned in the introduction, despite their popularity, some authors have expressed doubts regarding the reliability and robustness of the saliency maps for interpreting CNN models in the field of medicine [11, 14]. One of the issues they have highlighted is the poor correlation between saliency maps corresponding to models with different architectures.

To go a little deeper into this question, we have performed some experiments with our own disc and cup contour images (Dataset C), and with the 20 models trained with them described in section 3. In particular, we have calculated the degree of correlation between saliency maps corresponding to the same images but obtained from CNN models with different architectures, following the same methodology as in [55]. The

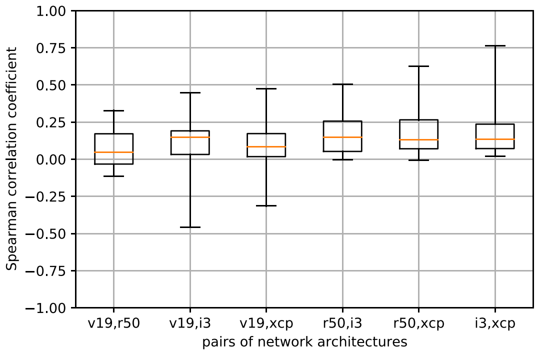
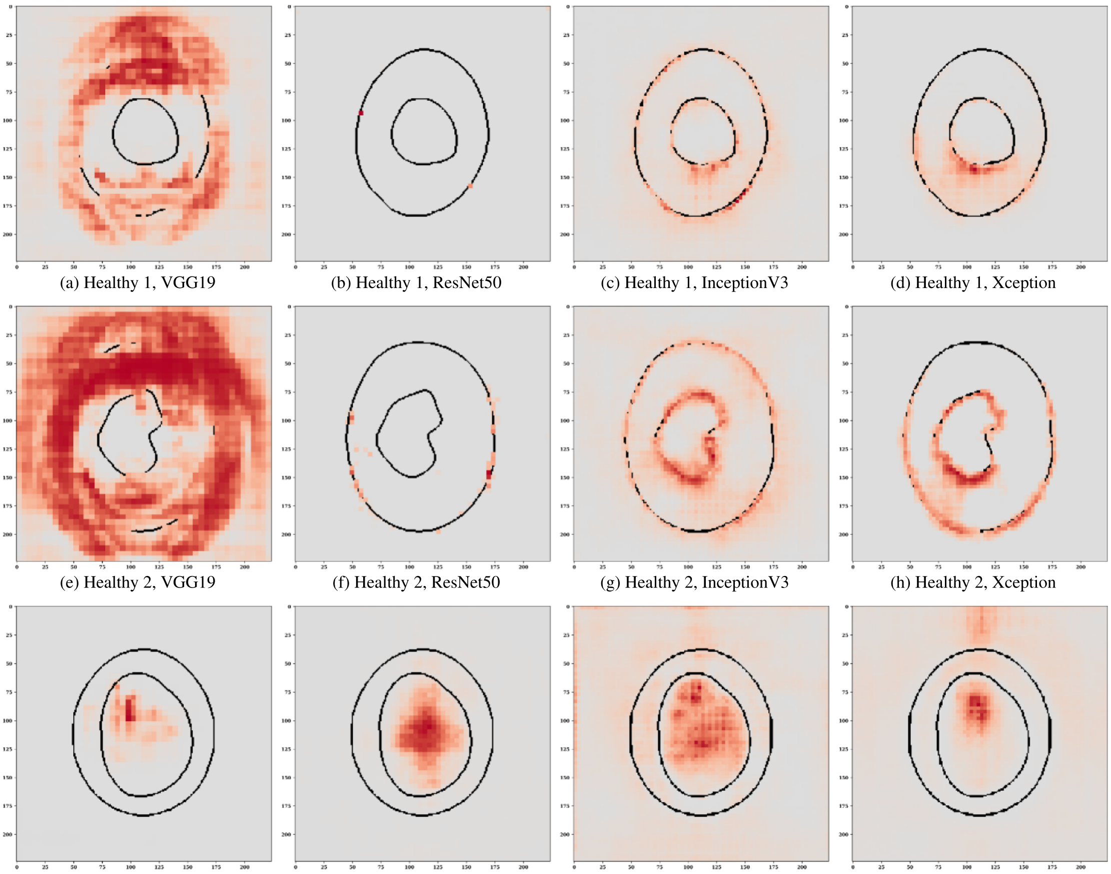


Figure 14. Boxplots showing the correlation coefficients between saliency maps for each pair of CNN architectures as explained in section 4.4.


(a) Healthy 1, VGG19 (b) Healthy 1, ResNet50 (c) Healthy 1, InceptionV3 (d) Healthy 1, Xception

(e) Healthy 2, VGG19 (f) Healthy 2, ResNet50 (g) Healthy 2, InceptionV3 (h) Healthy 2, Xception

(i) Glaucoma, VGG19 (j) Glaucoma, ResNet50 (k) Glaucoma, InceptionV3 (l) Glaucoma, Xception

Figure 15. Examples of saliency maps computed with the occlusion method for two healthy images (1st and 2nd rows) and one glaucoma image (3rd row) for the different convolutional neural network architectures trained with the Dataset XC: VGG19 (1st column), ResNet50 (2nd column), InceptionV3 (3rd column), and Xception (4th column).

saliency maps used were generated with well-known gradient backpropagation, integrated gradient, occlusion and CAM-based methods [56]. Figure 14 shows the results obtained where each boxplot has been calculated from the considered saliency methods between a pair of network architectures for the same fold. It is observed that on average the correlation is quite low in all possible combinations of architectures.

On the other hand, figure 15 shows some saliency maps computed with the occlusion method [57] corresponding to three example images for the models of different network architectures trained with the Dataset XC referred to in the previous section. Apart from the extreme difficulty of interpreting what the networks may be doing with the contour information to make the decision, it is possible to see the important differences, in some cases, between the saliency maps.

Therefore, in view of these results, it is worth considering the usefulness of this interpretability technique in a problem such as the diagnosis of glaucoma. Unlike other medical problems in which there may be clear

Table 7. Balanced accuracy results (column B. Acc.) achieved by the surrogate decision tree models per network architecture and fold on the Dataset XC. The last row of the table also shows the balanced accuracy obtained for a decision tree classifier, trained with the real labels of the data set, i.e. without surrogation. The surrogate models that performed better than the decision tree classfier are highlighted in bold.

Network / Model	Fold	B. Acc.
VGG19	1	**0.8989**
VGG19	2	0.8927
VGG19	3	**0.9009**
VGG19	4	0.8920
VGG19	5	**0.8945**
ResNet50	1	0.8773
ResNet50	2	0.8771
ResNet50	3	0.8846
ResNet50	4	0.8862
ResNet50	5	**0.8992**
InceptionV3	1	**0.9022**
InceptionV3	2	0.8886
InceptionV3	3	**0.9031**
InceptionV3	4	**0.9039**
InceptionV3	5	0.8936
Xception	1	**0.8969**
Xception	2	**0.8940**
Xception	3	0.8921
Xception	4	**0.8945**
Xception	5	**0.8944**
Decision Tree Classifier		0.8937

and well-located biomarkers of the pathology under consideration, in the case of glaucoma we find ourselves in a quite different scenario. In accordance with what was seen in the previous section, even assuming that the neural networks may be calculating features correlated with indicators commonly used by specialists such as the vCDR, it is practically impossible to deduce this fact from the simple observation of the saliency maps. As already mentioned in section 1, the usefulness of trying to explain at the pixel level what the network may be doing to generate its predictions is questionable. In this sense, the explored alternative of using a surrogate system as a means of interpretation seems to be more useful in a problem of this nature, also offering a more coherent result between the different architectures analyzed.

4.5. Knowledge distillation

So far we have seen how decision trees are able to reproduce the predictions of CNNs providing a certain interpretability mechanism for their decisions. However, as Rudin points out [6], rather than producing explanations faithful to the original model, these surrogate models show trends in the way predictions are related to features. Thus, while they may provide useful and more interpretable information than techniques such as saliency methods, as discussed in the previous section, they are not an entirely satisfactory solution.

For this reason, we decided to go a step further and evaluate on the real outcome, i.e. the original class test labels, the performance of the decision tree acting as a surrogate model that has learned to make predictions in a similar way as CNN does. Note that the surrogate model itself never sees the real outcome but the predictions of the neural networks. We would therefore be talking about a form of knowledge distillation in which we transfer knowledge from the network to the decision tree. Since the surrogate models have been trained for regression to mimic the probability value of being a glaucomatous eye, to do the classification we have simply thresholded this probability value at 0.5 to determine whether or not it is glaucoma. With this in mind, table 7 shows the balanced accuracy values found, on the Dataset XC, for the surrogate decision trees corresponding to the 20 network models considered, along with the value obtained for a decision tree trained without surrogation using the real outcome. Figure 16 visualizes the decision tree for this case in which a pattern similar to those we had already discussed is observed, with the vCDR playing a major role as the discriminant feature.

Table 7 shows that in some cases the knowledge distillation process produces decision tree models with better results. This finding is very interesting not only because of the improvement in itself, but also because we have managed to obtain intrinsically interpretable models with performance equivalent to CNN models, thus overcoming the limitations in the explanations referred to earlier in this section.

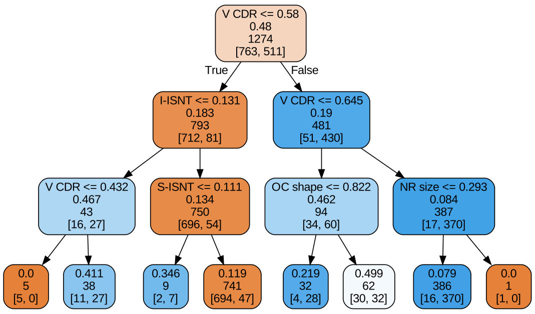


Figure 16. Decision Tree trained with the real class labels using the Dataset XC as explained in section 4.5.

5. Conclusions

The main conclusions we can draw from the experiments performed with surrogate interpretable models are as follows.

First, for this glaucoma diagnostic problem using CNNs trained with simplified images of disc and cup outlines, surrogate decision trees can offer a more intuitive and interpretable alternative to saliency methods when trying to explain their decisions despite the existing limitations of not dealing with the original model and data.

Second, we have shown that surrogate decision trees can closely approximate the predictions of neural networks and achieve similar performance when evaluated on the actual outcome, i.e. the class labels, and, in some cases, outperform the decision tree without surrogation, thanks to the use of knowledge distillation. Thus, we could dispense with CNNs and instead use an intrinsically explainable model with similar performance. From a practical point of view, we think this would be the most recommendable approach for specialists as a diagnostic tool.

Thirdly, in all cases the enormous difference found between the vCDR and the rest of the features in terms of their importance in the decision trees considered highlights the already known role of this feature as an indicator to distinguish between a healthy and a glaucomatous eye.

Finally, as future work, we propose to look at the possibility of increasing the number of features extracted from the fundus images to include not only geometric features but also other features extracted from structures such as vessels, peripapillary atrophy, the pallor, etc. Perhaps in this way the surrogate model could approximate the predictions of neural networks trained with the original images without simplification. On the other hand, the influence that the segmentation of the relevant structures of the optic nerve has on the results obtained could be addressed. In this work we have used manual segmentation of the disc and the cup carried out by an expert but automatic methods could be used to segment these structures and others.

Data availability statement

The data cannot be made publicly available upon publication because they contain sensitive personal information. The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

Funding from the University of La Laguna and the Spanish Ministry of Science, Innovation and Universities is acknowledged.

ORCID iDs

Francisco Fumero https://orcid.org/0000-0001-9806-2477
José Estévez https://orcid.org/0000-0002-7452-2958

References

[1] Bourne R R A *et al* 2013 Causes of vision loss worldwide, 1990–2010: a systematic analysis *Lancet Glob. Health* **1** e339–49
[2] Tham Y-C, Li X, Wong T Y, Quigley H A, Aung T and Cheng C-Y 2014 Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis *Ophthalmology* **121** 2081–90

[3] van der Velden B H M, Kuijf H J, Gilhuijs K G A and Viergever M A 2022 Explainable artificial intelligence (XAI) in deep learning-based medical image analysis *Med. Image Anal.* **79** 102470

[4] Tjoa E and Guan C 2021 A survey on explainable artificial intelligence (XAI): toward medical XAI *IEEE Trans. Neural Netw. Learn. Sys.* **32** 4793—4813

[5] van der Veer S N *et al* 2021 Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries *J. Am. Med. Inf. Assoc.* **28** 2128–38

[6] Rudin C 2019 Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead *Nat. Mach. Intell.* **1** 206–15

[7] Ordish J and Hall A 2020 *Black Box Medicine and Transparency: Interpretable Machine Learning* (PHG Foundation)

[8] Hanif A M, Beqiri S, Keane P A and Campbell J 2021 Applications of interpretability in deep learning models for ophthalmology *Curr. Opin. Ophthalmol.* **32** 452–8

[9] Petch J, Di S and Nelson W 2022 Opening the black box: the promise and limitations of explainable machine learning in cardiology *Can. J. Cardiol.* **38** 204–13

[10] Simonyan K, Vedaldi A and Zisserman A 2014 Deep inside convolutional networks: visualising image classification models and saliency maps *Workshop at Int. Conf. on Learning Representations*

[11] Arun N T *et al* 2020 Assessing the (un)trustworthiness of saliency maps for localizing abnormalities in medical imaging *Radiol.: Artif. Intell.* **3** e200267

[12] Saporta A *et al* 2022 Benchmarking saliency methods for chest x-ray interpretation *Nat. Mach. Intell.* **4** 867–78

[13] Singh A, Jothi Balaji J, Rasheed M A, Jayakumar V, Raman R and Lakshminarayanan V 2021 Evaluation of explainable deep learning methods for ophthalmic diagnosis *Clin. Ophthalmol.* **15** 2573–81

[14] Van Craenendonck T, Elen B, Gerrits N and De Boever P 2020 Systematic comparison of heatmapping techniques in deep learning in the context of diabetic retinopathy lesion detection *Transl. Vis. Sci. Tech.* **9** 64

[15] Ayhan M S, Kümmerle L, Kühlewein L, Inhoffen W, Aliyeva G, Ziemssen F and Berens P 2022 Clinical validation of saliency maps for understanding deep neural networks in ophthalmology *Med. Image Anal.* **77** 102364

[16] Molnar C 2022 *Interpretable Machine Learning* 2nd edn

[17] Ribeiro M T, Singh S and Guestrin C 2016 Why should i trust you?': explaining the predictions of any classifier *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, KDD '16* (Association for Computing Machinery) pp 1135–44

[18] Chan-Lau J A, Hu R, Ivanyna M, Qu R and Zhong C 2023 Surrogate data models: interpreting large-scale machine learning crisis prediction models *IMF Working Papers* **2023**

[19] Cowan N 2010 The magical mystery four: how is working memory capacity limited and why? *Curr. Dir. Psychol. Sci.* **19** 51–57

[20] Escamez C S F, Giral E M Martinez S P and Fernandez N T 2021 High interpretable machine learning classifier for early glaucoma diagnosis *Int. J. Ophthalmol.* **14** 393–8

[21] Kooner K S *et al* 2022 Glaucoma diagnosis through the integration of optical coherence tomography/angiography and machine learning diagnostic models *Clin. Ophthalmol.* **16** 2685–97

[22] Chai Y, Liu H and Xu J 2018 Glaucoma diagnosis based on both hidden features and domain knowledge through deep learning models *Knowl. Based Syst.* **161** 147–56

[23] Mehta P *et al* 2021 Automated detection of glaucoma with interpretable machine learning using clinical data and multimodal retinal images *Am. J. Ophthalmol.* **231** 154–69

[24] Oh S, Park Y, Cho K J and Kim S J 2021 Explainable machine learning model for glaucoma diagnosis and its interpretation *Diagnostics* **11** 510

[25] Xu Y *et al* 2021 A hierarchical deep learning approach with transparency and interpretability based on small samples for glaucoma diagnosis *npj Digit. Med.* **4** 1–11

[26] Krishna Adithya V, Williams B M, Czanner S, Kavitha S, Friedman D S, Willoughby C E, Venkatesh R and Czanner G 2021 EffUnet-SpaGen: an efficient and spatial generative approach to glaucoma detection *J. Imaging* **7** 92

[27] MacCormick I J C *et al* 2019 Accurate, fast, data efficient and interpretable glaucoma diagnosis with automated spatial analysis of the whole cup to disc profile *PLoS One* **14** e0209409

[28] Singh L K, Pooja P, Garg H, Khanna M and Bhadoria R S 2021 An enhanced deep image model for glaucoma diagnosis using feature-based detection in retinal fundus *Med. Biol. Eng. Comput.* **59** 333–53

[29] Kinger S, Kulkarni V, Nayak J, Behera H, Naik B, Vimal S and Pelusi D 2022 Explainability of deep learning–based system in health care *Computational Intelligence in Data Mining* (*Smart Innovation, Systems and Technologies*) (Springer) pp 619–33

[30] Gheisari S, Shariflou S, Phu J, Kennedy P J, Agar A, Kalloniatis M and Golzan S M 2021 A combined convolutional and recurrent neural network for enhanced glaucoma detection *Sci. Rep.* **11** 1945

[31] Li X, Xiong H, Li X, Zhang X, Liu J, Jiang H, Chen Z and Dou D 2023 G- LIME: statistical learning for local interpretations of deep neural networks using global priors *Artif. Intell.* **314** 103823

[32] Visani G, Bagli E and Chesani F 2022 OptiLIME: optimized lime explanations for diagnostic computer algorithms (arXiv:2006.05714)

[33] Palatnik de Sousa I, Maria Bernardes Rebuzzi Vellasco M and Costa da Silva E 2019 Local interpretable model-agnostic explanations for classification of lymph node metastases *Sensors* **19** 2969

[34] van der Linden I, Haned H and Kanoulas E 2019 Global aggregations of local explanations for black box models (arXiv:1907.03039)

[35] Ahern I, Noack A, Guzman-Nateras L, Dou D, Li B and Huan J 2019 NormLime: a new feature importance metric for explaining deep neural networks (arXiv:1909.04200)

[36] Elshawi R, Al-Mallah M H and Sakr S 2019 On the interpretability of machine learning-based model for predicting hypertension *BMC Med. Inform. Decis. Mak.* **19** 146

[37] Karatza P, Dalakleidi K, Athanasiou M and Nikita K 2021 Interpretability methods of machine learning algorithms with applications in breast cancer diagnosis, *2021 43rd Annual Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBC)* pp 2310–3

[38] Dua D and Graff C 2019 Machine learning repository (available at: https://archive.ics.uci.edu)

[39] Batista F, Diaz–Aleman T, Sigut J, Alayon S, Arnay R and Angel–Pereira D 2020 Rim–one dl: a unified retinal image database for assessing glaucoma using deep learning *Image Anal. Stereol.* **39** 161–7

[40] Fumero F, Alayon S, Sanchez J L, Sigut J and Gonzalez-Hernandez M 2011 Rim–one: an open retinal image Stereology *2011 24th Int. Symp. on Computer–Based Medical Systems (CBMS)* pp 1–6

[41] Fumero F, Sigut J F, Alayó S, Gonzalez–Hernandez M and González de la Rosa M 2015 Interactive tool and database for optic disc and cup segmentation of stereo and monocular retinal fundus images *Short Papers Proc. 23rd Int. Conf. in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG 2015)* ed V Skala pp 91–98

[42] Orlando J I *et al* 2020 Refuge challenge: a unified framework for evaluating automated methods for glaucoma assessment from fundus photographs *Med. Image Anal.* **59** 101570

[43] Sivaswamy J, Chakravarty A, Joshi Ujjwal G D and Syed T A 2015 A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis *JSM Biomed. Imaging Data Pap.* **2** 1004

[44] European Glaucoma Society 2021 European glaucoma society terminology and guidelines for glaucoma, 5th edition *Br. J. Ophthalmol.* **105** 1–169

[45] Jonas J B, Budde W M and Panda-Jonas S 1999 Ophthalmoscopic evaluation of the optic nerve head *Surv. Ophthalmol.* **43** 293–320

[46] Jonas J B, Gusek G C and Naumann G O H 1988 Optic disc, cup and neuroretinal rim size, configuration and correlations in normal eyes *Investigative Ophthalmol. Vis. Sci.* **297** 1151–8

[47] Spaeth G L, Hednerer J, Liu C S, Kesen M R, Altangerel U, Bayer A, Katz L J, Myers J S, Rhee D and Steinmann W 2002 The disc damage likelihood scale: reproducibility of a new method of estimating the amount of optic nerve damage caused by glaucoma *Trans. Am. Ophthalmol. Soc.* **100** 181–5

[48] Armaly M F 1969 The cup/disc ratio: the findings of tonometry and tonography in the normal eye *Arch. Ophthalmol.* **82** 191–6

[49] Simonyan K, Zisserman A, Bengio Y and LeCun Y 2015 Very deep convolutional networks for large-scale image recognition *3rd Int. Conf. on Learning Representations, ICLR 2015 Conf. Track Proc.*

[50] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* pp 770–8

[51] Szegedy C, Vanhoucke V, Ioffe S, Shlens J and Wojna Z 2016 Rethinking the inception architecture for computer vision *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* pp 2818–26

[52] Chollet F 2017 Xception: Deep learning with depthwise separable convolutions *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* pp 1800–7

[53] Brodersen K H, Ong C S, Stephan K E and J M Buh-mann 2010 The balanced accuracy and its posterior distribution *2010 20th Int. Conf. on Pattern Recognition* pp 3121–4

[54] Kumar J R H, Seelamantula C S, Kamath Y S and Jampala R 2019 Rim-to-disc ratio outperforms cup-to-disc ratio for glaucoma prescreening *Sci. Rep.* **9** 7099

[55] Fumero F, Sigut J, Estévez J and Diaz-Aleman T 2023 Systematic application of saliency maps to explain the decisions of convolutional neural networks for glaucoma diagnosis based on disc and cup geometry (available at: https://papers.ssrn. com/abstract=4327677) (Accessed 1 March 2023)

[56] Teng Q, Liu Z, Song Y, Han K and Lu Y 2022 A survey on the interpretability of deep learning in medical diagnosis *Multimedia Syst.* **28** 2335–55

[57] Zeiler M D and Fergus R 2014 Visualizing and understanding convolutional networks *Computer Vision – Eccv 2014* ed D Fleet, T Pajdla, B Schiele and T Tuytelaars (Springer) pp 818–33