electronics

MDPI

Article

Multi-Stage Ensemble-Based System for Glaucomatous Optic Neuropathy Diagnosis in Fundus Images

Carlos A. Vásquez-Rochín , Miguel E. Martínez-Rosas *, Humberto Cervantes de Ávila, Gerardo Romo-Cárdenas , Priscy A. Luque-Morales and Manuel M. Miranda-Velasco

Autonomous University of Baja California (UABC), Faculty of Engineering, Architecture and Design (FIAD), Carretera Transpeninsular Ensenada-Tijuana 3917, Zona Playitas, Ensenada 22860, Baja California, Mexico
* Correspondence: emartine@uabc.edu.mx

Abstract: Recent developments in Computer-aided Diagnosis (CAD) systems as a countermeasure to the increasing number of untreated cases of eye diseases related to visual impairment (such as diabetic retinopathy or age-related macular degeneration) have the potential to yield in low-to-mid income countries a comfortable and accessible alternative to obtaining a general ophthalmological study necessary for follow-up medical attention. In this work, a multi-stage ensemble-based system for the diagnosis of glaucomatous optic neuropathy (GON) is proposed. GON diagnosis is based on a binary classification procedure working in conjunction with a multi-stage block based on image preprocessing and feature extraction. Our preliminary data show similar results compared to current studies considering metrics such as Accuracy, Sensitivity, Specificity, AUC (AUROC), F_1 score, and the use of Matthews Correlation Coefficient (MCC) as an additional performance metric is proposed.

Keywords: deep learning; computer-aided diagnosis; CNN; glaucoma

Citation: Vásquez-Rochín, C.A.; Martínez-Rosas, M.E.; de Ávila, H.C.; Romo-Cárdenas, G.; Luque-Morales, P.A.; Miranda-Velasco, M.M. Multi-Stage Ensemble-Based System for Glaucomatous Optic Neuropathy Diagnosis in Fundus Images. *Electronics* **2023**, *12*, 1046. https://doi.org/10.3390/electronics12041046

Academic Editors: Esteban Tlelo-Cuautle, Everardo Inzunza-González and Walter Leon-Salas

Received: 15 January 2023
Revised: 14 February 2023
Accepted: 17 February 2023
Published: 20 February 2023

1. Introduction

The sense of vision is fundamental in every aspect of human development, being one of the most dominant senses as an inseparable part of current social and interpersonal interaction. In 2020, the World Health Organization (WHO) published the *World Report on Vision*, in which it stated that at least 2.2 billion people around the globe suffer from some kind of visual deficiency (or impairment), of which less than 1 billion could have been prevented. These numbers tend to increase due to several factors related to population aging, reduced medical attention, and lack of educational opportunities [1,2].

In general, eye pathologies (such as diabetic retinopathy, cataract, and glaucoma) represent a major *quality-of-life* deterioration around the globe—especially for elderly people—which is why a better understanding of said pathologies is needed to develop better and more efficient methodologies for clinical environments [3]. For any ophthalmologist, an objective clinical measure is essential during an eye screening process. This kind of measurement can be acquired through a variety of techniques that are related to visual acuity, intraocular pressure, or fundus imaging [4]. In the latter—also known as *ophthalmoscopy*—there is a specific set of landmarks (also called *biomarkers*) that represent zones of interest for an ophthalmologist during the evaluation process of a *Color Fundus Photograph* (CFP), such as the *macula*, *fovea*, *optic disc* (OD), *optic cup* (OC), and vascular system (i.e., veins and arteries) [5,6]. Depending on the pathology, the type and number of lesions found in a CFP vary and could play a critical role in the process of diagnosis.

In the specific case of *glaucomatous optic neuropathy* (GON), the pathology is characterized by a variety of eye disorders that can present a set of almost imperceptible symptoms at early stages, all of which can lead to severe damage of the optic nerve and eventually, a certain level of peripheral vision loss, or even complete blindness [5,7] if left untreated (as seen in Figure 1). For these reasons, special attention to OD is required to perform an early

diagnosis of GON and adequate disease management. In recent years, there have been multiple approaches related to the diagnosis of retinal pathologies and secondary tasks related to it (i.e., semantic segmentation—the process of classifying a certain class within an image context and separating it from the rest of said image classes by overlaying it with a segmentation mask— or localization of biomarkers for deeper analysis) based on Artificial Intelligence (AI) and Machine Learning algorithms [8–14]. Most of these studies have yielded good results, related to evaluation metrics such as *Sensitivity* (SN) and *Specificity* (SP), compared with more classical approaches based on *Digital Image Processing* (DIP) of CFPs.

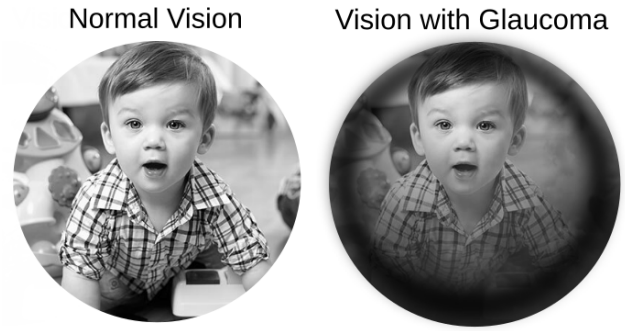

Figure 1. Examples of human vision without (**left side**) and with GON (**right side**).

In this work, a method for the development of a Multi-stage Ensemble-based *Computer-aided Diagnosis* (CAD) system for *referable* GON is proposed. It comprises a CNN-based classification system in conjunction with a subsystem that performs a morphological analysis of a set of particular fundus landmarks, in this case, OD and OC to imitate the general procedure, in which a specialist in the field of ophthalmology would perform an eye fundus examination. To further validate the proposed ensemble of classification models, external data were used to create a baseline for the their ability to generalize over *unseen* features. Moreover, a set of specific metrics have been proposed to compensate for the class imbalance found in each dataset. In the following sections, a general description of the current place of AI in the analysis of biomarkers and lesions related to guided diagnosis of retinal pathologies is presented. In addition to this, a comparison between the most recent DL-based systems performing GON diagnosis applying different methodologies is drawn.

2. CAD for Eye-Related Pathologies

2.1. AI for Image Analysis

AI is a term employed for describing the development of a computer program that models intelligent behavior with reduced human interaction [15]. As a ramification of this discipline, Machine Learning Classifiers (MLC) were developed to learn patterns from a dataset without input from a human through the implementation of a statistical model. Classical MLC, such as Logistic Regression (LR), Random Forest (RF), Support Vector Machines (SVM), etc., form part of what is called *supervised* Machine Learning (ML), which uses mathematical modeling to accomplish a given goal through the feeding of a large and structured training dataset (e.g., set of labeled CFPs). However, one known limitation of MLCs is that they require engineered (or handmade) filters to successfully recognize pattern within input data [16].

Additionally, networks based on artificial neural behavior, also called *Artificial Neural Networks* (ANN), implement what is known as *neurons*, which are a set of interconnected nodes capable of transforming a given input by assigning a *weight* to it before passing it to next one. These neurons undergo a *learning process* which adjusts said weights to the point of being capable of making a prediction. A more complex version of this is called *Deep Neural Networks* (DNN), containing multiple layers of processing (filters) between the

input and output layers and, as a branch of this category, *Convolutional Neural Networks* (CNN) are found [16,17]. CNN basically consists of a set of layers, each of which is a series of *convolutional layers* that applies a convolution to the input before passing it to the next layer of neurons. The major advantage of CNN is that it locally connects neurons to specific and smaller *receptive fields*, which means that the algorithm considers the *spatial structure* of a given input.

In the *state of the art* (SOTA), a plethora of methodologies for the detection or segmentation of these types of lesions can be found. Although methodologies of systems, such as the case of Dai et al. [18] and Sarhan et al. [19], resolve up to a certain point the problem caused by data imbalance, comparison between most CAD systems still remains *unfair* in terms of the reported performance metrics. This is mainly because of the performance shift of a given metric as the imbalance ratio (e.g., prevalence of minority class) varies from dataset to dataset, which could yield misleading (or even over-optimistic) results. In this regard, multiple research groups [20–23] have studied the effects of data imbalance and the effectivity of evaluation metrics implemented in binary classification tasks and they recommend a set of metrics depending on the nature of the pre-established goals of the study to be conducted.

2.2. AI for ONH Evaluation

Evaluation of OD, also known as *optic nerve head* (ONH), for GON assessment requires special attention to a set of key features. Some of these features in CFP correspond to *retinal nerve fiber layer* (RNFL) defects, *peripapillary atrophy* (PPA), neuroretinal rim notch, and vertical *cup-to-disk ratio* (CDR). In recent work from Phene et al. [24], a comparison between a DL classification system and a group of GON specialists was drawn, taking into account the relative importance of different features related to a referable GON patient. From this work, an Area Under the Receiver Operating Characteristic curve—also known as AUC (ROC)—that ranges from 0.881 to 0.945 across three datasets was achieved, and the authors concluded that the most relatively important features with regard to referable GON were: neuroretinal rim notching, RNFL defect, bared circumlinear vessels, and the presence of vertical CDR of 0.7 or more.

Vertical CDR is a specific feature from referable GON that has not been widely adopted but has been reported recently in multiple works [12,24–31], which does not share similar prevalence and bias in the validation results thanks to the diverse amount of CFP datasets employed between each work. Therefore, reports of different thresholds of vertical CDR to determine a referable case of GON are commonly found, which could be a product of the different methodologies applied to compute CDR and inherent characteristics of each dataset. From the literature, a threshold of $0.5 \leq \text{CDR} < 0.8$ is considered as a *possible* case of GON, and $\text{CDR} \geq 0.8$ as a *confirmed* case.

Table 1 presents current work related to classification systems for GON diagnosis, where the network architecture and CFP datasets employed in each proposed CAD system are detailed. In addition, data availability, capture technique of the acquisition device, and region of interest (ROI) are shown. All of the systems cited in this work differ from each other in the methodology used for classifying referable GON, the amount of data, and the clinical information used outside the classifier prediction to arrive to a satisfying decision. From a network architecture point of view, ResNet is found widely used either in conjunction to other architectures, or by itself, and that the region analyzed more frequently is the area comprised by the OD, although a mixture of both eye fundus as a whole and OD crops has been applied before with near-SOTA results [32].

Table 1. Results from recent AI systems for GON classification Non-Referable GON (NO-GON) vs. Referable GON (R-GON) based on CFP as input.

Author	Network Architecture	Database	Capture Technique	Region Analyzed
Ting et al., 2017 [12]	VGGNet	SiDRP 2010-2013, SIMES, SINDI, SCES, Singapure National Eye Center SiDRP 2014-2015: N/A	Mixed	Fundus
Chai et al., 2018 [25]	Multi-Branch Neural Network (MB-NN)	N/A	N/A	OD crop
Christopher et al., 2018 [33]	VGG16, Inceptionv3, ResNet-50	ADAGES: N/A, UCSD DIGS: N/A	Mydriatic	OD crop
Li et al., 2018 [26]	Inception-v3	LabelMe (subset): N/A	N/A	Fundus
Liu et al., 2018 [34]	ResNet-50	RIM-ONE: OA + private, HRF: OA	Non-mydriatic + N/A	N/A
Al-Aswad et al., 2019 [28]	ResNet-50	Clinic-based images: N/A, ORIGA: N/A	Mixed	OD crop
Hemelings et al., 2019 [35]	ResNet-50	University Hospitals Leuven: N/A	Non-mydriatic	Fundus
Kim et al., 2019 [32]	VGG-16, ResNet-152, Inception-v4	Glaucoma clinic, Samsung medical center: N/A	N/A	Mixed (OD cropped center 1:1 ratio)
Liu et al., 2019 [29]	ResNet	CGSA, Handan Eye Study, + additional hospital and clinic based images: N/A	Mydriatic	OD crop
Phene et al., 2019 [24]	Inception-v3	EYEPACS, Inoveon AREDS Aravind Eye Hospital, Sankara Nethralaya, Narayana Nethralaya, India	N/A	Fundus
Diaz-Pinto et al., 2019 [36]	VGG16, VGG19, InceptionV3, ResNet-50, Xception	HRF: OA, Drishti-GS1: OA, RIM-ONE: OA, sjchoi86-HRF: N/A, ACRIMA: OA	Mixed	OD crop
Li et al., 2020 [30]	ResNet-101	Shanghai Zhongshan Hospital Shanghai First People's hospital: NA	Mydriatic	OD crop
Sreng et al., 2020 [37]	AlexNet, GoogleNet, InceptionV3, XceptionNet, ResNet-50, SqueezeNet, ShuffleNet, MobileNet, DenseNet, InceptioN/AesNet, NasNet-Large	REFUGE: OA, ACRIMA: OA, ORIGA: OA, RIM-ONE: OA, DRISHTI-GS1: OA	N/A	OD crop
Civit-Masot et al., 2020 [31]	Generalized U-net + (VGG16, ResNet-50, Xception, MobileNetV2)	RIM-ONE: OA, DRISHTI: OA	N/A	OD crop

N/A = Not Available, OA= Open Access.

2.3. Classification of Imbalance Datasets

A dataset can be defined as *imbalanced* if its class distribution is *unequal*, meaning that a disproportion among the total number of samples per class is present in said dataset. For instance, even a difference of just one sample between classes could technically be considered an imbalanced dataset. Many tasks related to ML are affected by data imbalance, e.g., face recognition irregularity detection and natural language processing, among others. This phenomenon is particularly present in medical diagnosis, as prevalence of a disease varies between populations due to different factors, including socioeconomic status and ethnicity.

In the case of *binary classification* (BC), a discrimination between two classes is performed, in ML, this could be the output of a dedicated classification algorithm. As defined by Hicks et al. [22], a BC problem can be expressed as follows:

$$p(X, \alpha) = \alpha p_{\mathrm{P}}(X) + (1 - \alpha) p_{\mathrm{N}}(X) \, , \tag{1}$$

where data samples are represented by X, $p_{P/N}$ represent class distributions of positive and negative classes, and α is a *mixture* parameter of the positive class defined as $\alpha = \frac{\mathrm{N_P}}{\mathrm{N_P}+\mathrm{N_N}}$, with $\mathrm{N_{P/N}}$ as the total number of positive and negative data samples from the dataset of interest. The performance of an algorithm for BC can be summarized in the form of a *two-class* confusion matrix (CM):

$$\mathrm{CM} = \begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix} .$$

For the above, instances of a positive case being correctly identified as positive are denominated *True Positive* (TP), and in the case of correctly classified negative cases, *True Negative* (TN) denomination is used. On the other hand, *False Positive* (FP) and *False Negative* (FN) categories are used for instances related to positive and negative misclassifications, respectively. From these four base components, other elemental metrics could also be computed—metrics such as *True Positive Rate* (TPR) and *True Negative Rate* (TNR). In BC, performance results should—ideally—land only on the TP and TN (negative slope diagonal).

In general, positive (or minority) cases of a disease are less frequent than negative (or majority) cases in many medical applications. Therefore, maximization of TP instances and minimization of FN instances should be the main focus within this field, as, in general, FP instances would not be as detrimental as the other case.

A general sense of classification performance can be shown with the resultant CM. However, the need for more compact metrics has led scientists to develop evaluation metrics such as Accuracy, F_β-score, and Recall. Recent works have demonstrated that at least Accuracy, as a performance metric, does not have the same performance at different levels of imbalance ratio [21–23]. Instead, metrics such as *Matthews Correlation Coefficient* (MCC) should be considered for a *balanced* indication of performance; *Bookmaker Informedness* (BM) and *g-mean* of TRP and TNR (GBA) should be used for a *reasonable* comparison between different classifiers. In the literature [38], several approaches have been proposed to address this particular issue, which could be based on an *Algorithm*, *Data*, or *Ensemble-based* methodology.

A system based on the last category provides a suitable alternative to reach a satisfying decision in classification-related tasks. It is important to note that in order to achieve a desirable result, a level of *diversity* in the system should be established. Within this context, diversity is used as a term to relate to a situation where the output of the *base classifiers* (classifiers that compose the ensemble system) differ. This characteristic is visualized with statistical concepts such as *bias* and *variance*, which are related to *overfitting* and *underfitting* problems, respectively. The main objective in an ensemble system is the reduction of variance with the averaging procedure applied to a set of base classifiers. This work takes advantage of a combination technique called *Classifier fusion*, which aggregates all classifiers

involved to establish a final decision, assuming that every classifier is competent within the feature space and each one is expected to misclassify different examples.

3. Materials and Methods

During this section, a detailed explanation of the general work scheme for the proposed CAD system is given, including a DL-based classification and OD segmentation block. In addition to this, general aspects of the datasets employed during training, validation, and testing procedures are presented.

3.1. Ensemble-Based CAD System for GON

The current proposal for a CAD system is based on an *ensemble* of three CNNs in a classification block, with a total of two, which differ from each other in aspects of architecture to imitate the analysis from *multiple points of view* of a single CFP as input. Each classification block focuses on different regions of the input image, one of these is dedicated to the analysis of the image on a global aspect, and the other just to a *region of interest* (ROI) comprising the OD and OC. A classification structure based on an ensemble assimilates an inspection methodology similar to natural human behavior, in which a conclusion is reached based on the opinion of multiple experts in a field. Such a configuration was selected for the CAD system due to the performance improvement yield in comparison with classification systems based on just one CNN classification system [39]. Furthermore, a subsystem was added to the main workflow, using a U-shaped encoder–decoder network architecture (UNet) to generate a mask base on the semantic segmentation of two particular biomarkers of interest for the study of GON (OD and OC), and from which vertical CDR is computed based on the relationship of the vertical diameter of both biomarkers. Figure 2 establishes a general block diagram of the proposed CAD system.

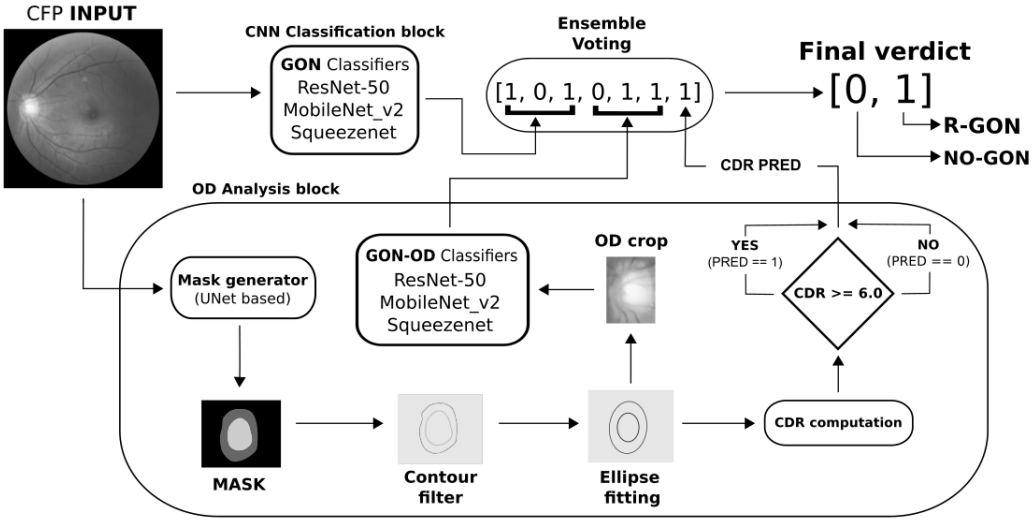

Figure 2. General scheme workflow for the proposed Multi-stage Ensemble-based CAD system.

3.2. Datasets

As stated before, the eye pathology of interest in this case is GON. Therefore, CFP datasets that provide clinical validation by a group of professionals in this discipline are required. For this study, a set of public CFP datasets were employed: LAG [40], iChallenge-GON [41], and RIM-ONE DL [42], whose labels correspond to a patient with a diagnosis of *referable GON* (R-GON) or not (NO-GON), which are needed for BC. Similarly to this, in a segmentation related task, a set of images with a manual segmentation of areas of interest (ground truth) by a specialist is essential. Fortunately, the iChallenge-GON dataset provides OD and OC ground truths for every image, which led to the respective training and

validation procedures of a segmentation network (U-Net in this case). An overview of the technical specifications of each dataset is shown in Table 2 for the required training process.

Table 2. CFP datasets specifications employed for Training, Validation, and Test procedures related to ensemble-based CAD system.

Datasets for Global Assessment	**Ethnicity**	**Class Distribution (R-GON/NO-GON)**	**Image Resolution (Mean H × W)**
LAG	Asian	1711/3143	500 × 500
iChallenge-GON	Asian	40/360	2056 × 2124
Datasets for OD-only Assessment			
LAG [a]	Asian	404/1564	418 × 391
RIM-ONE DL	Latin	172/313	503 × 503

[a] Custom subset of OD-cropping from OD detection algorithm.

3.3. Training process

As shown in Figure 2, a set of 3 different CNNs were employed for the analysis of GON. The first one, *ResNet-50*, is a CNN that employs a series of *skip connections* to assimilate information with fewer layers than a regular or *flat* CNN implementation. Since its introduction in 2015, it has remained as a SOTA architecture for tasks related to image classification, and this particular version of it manages only 50 layers of processing, which have been proven to yield acceptable results for pathologies such as GON and diabetic retinopathy [43].

In the case of *MobileNet_v2*, its architecture is based on a combination of *depthwise* and *pointwise* convolutions that significantly reduces the total amount of trainable parameters when compared to a network with the same depth with a flat implementation. The principal criteria for its selection were established based purely on the results gathered from current literature, which have presented near-SOTA results for classification of GON.

For the final network, *Squeezenet* was chosen. This network architecture has an *AlexNet performance level* with a reduction of 50× related to model size, using what is called a *fire module*, comprising a *squeeze convolution layer* (only 1×1 filters), which feeds into an expand layer that has a combination of 1×1 and 3×3 convolution filters.

For each training process, an additional training scheme based on *Scheduled Learning Rate* (SLR) was considered, which relies on a constant verification of a parameter (in this case, a *Validation score*) to determine whether the computed loss follows a correct trending (i.e., a reduction at each epoch) and if this condition is not fulfilled, it reduces α by a fixed factor in an attempt to improve the current behavior. The training configuration setup used in this training process is shown in Table 3.

Table 3. Parameters selected for Training process.

Model	**ResNet-50**	**MobileNet_v2**	**Squeezenet**
Batch size	16	32	
Optimizer	SGD (α = as stated below, momentum = 0.9), Adam (α = as stated below, $\beta_1 = 0.9$, $\beta_2 = 0.999$)		
Learning Rate (α)	1×10^{-3} (regular scheme), 1×10^{-3} to 1×10^{-5} (under SLR)		
Epochs	100		

In the context of an optimization algorithm, the function used to evaluate a candidate solution (i.e., a set of weights) is called *objective function*, and for the case of neural networks,

minimizing the error produced between a prediction and a target value is desired. At the same time, the objective function is often referred to as a cost function or a *loss function*, and the *value calculated* by the *loss function* is referred to simply as *Loss*. In this work, two of the best-performing optimization algorithms for training DL models were alternated: *Stochastic Gradient Descent* (SGD) and *Adaptive Moment Estimation* (Adam).

Nevertheless, in the current literature, it is found that there is no *standard combination* of parameters and hyperparameters that provides good results for a specific eye pathology. However, an $\alpha \approx 1 \times 10^{-3}$ and *momentum*-related parameter within the optimization algorithm stays close to 1. In the following section, the results from 6 out of a total of 24 training procedures are presented as the final selection of trained parameters for the implementation of the current iteration of the proposed ensemble-based CAD system.

4. Results

A set of CMs is shown in the following figures as a visualization of general performance on the best-performing model for GON classification. It is based on the data of the Train (TR) and Validation (VAL) sets. In addition to this, an external Validation dataset (TST) was selected as a measure of robustness (or *generalization*) for each model with information that differs from what is used to *see* as an input CFP.

From the models represented by Figures 3–5, it was observed that models with these trained parameters did not perform as expected, considering that the TST dataset was selected trying to match some of the characteristics (e.g., field of view, clean optic media, illumination) of the TR and VAL datasets. However, a general improvement in this regard can be observed in Figures 6–8, which represents the performance of the trained parameters for each model based on the classification of a ROI within a CFP, in this case, the OD and OC. This could indicate that even if a dataset has a higher number of samples, but lower quality overall, training cannot translate into better validation results. Based on these results, we suggest that further attention should be focused on a set of specific regions, and exclude non-valuable information within each CFP used as input for GON assessment, such as the typical FOV mask applied by commercially available acquisition systems.

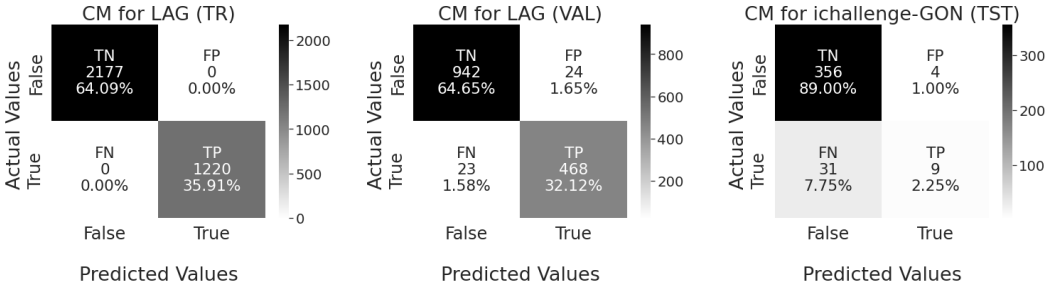

Figure 3. Confusion matrix of ResNet-50 trained parameters for GON.

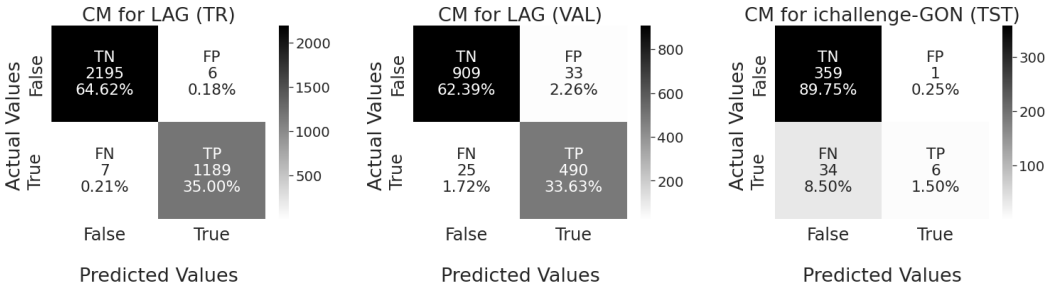

Figure 4. Confusion matrix of MobileNet_v2 trained parameters for GON.

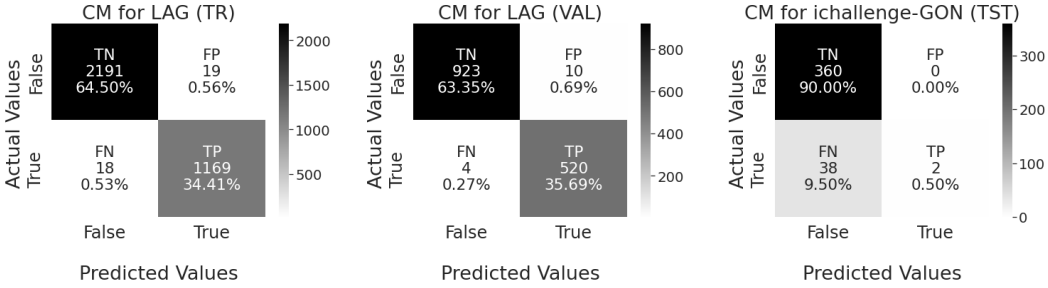

Figure 5. Confusion matrix of Squeezenet trained parameters for GON.

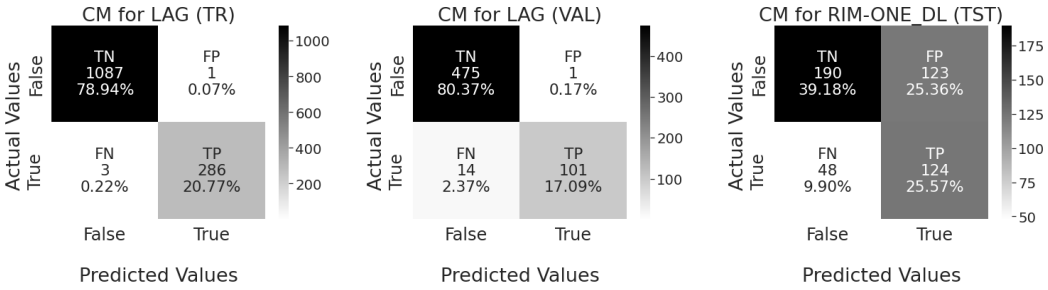

Figure 6. Confusion matrix of ResNet-50 trained parameters for GON-OD.

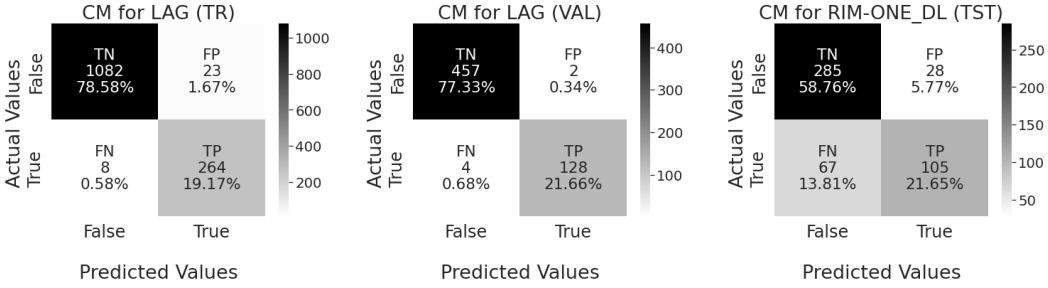

Figure 7. Confusion matrix of MobileNet_v2 trained parameters for GON-OD.

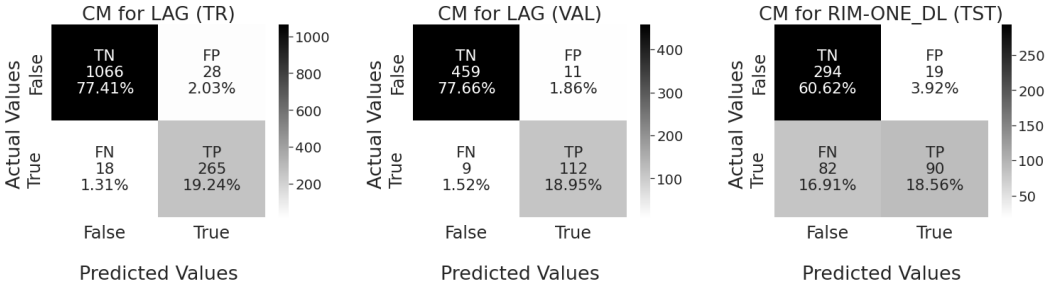

Figure 8. Confusion matrix of Squeezenet trained parameters for GON-OD.

In order to generate a mask of OD and OC (based on semantic segmentation) and compute the corresponding CDR value of a given OD from an input CFP, a model based on the U-Net architecture was trained on top of pretrained parameters from a VGG19 model. Said model was trained using a dataset containing 400 examples of a CFP paired with a png image (mask) where the *features of interest*, or classes, are highlighted in a different color. In this case, three different classes were taken into account, which considered the OC, OD, and *Background* (BG) or everything that is not related to the biomarkers of interest. Furthermore, to assign a certain weight to the prediction made to each class, a mean inverse relationship between the total amount of pixels in the image and the coverage area was calculated. In Table 4, results for OD and OC segmentation from a validation subset of iChallenge-GON dataset are presented.

Table 4. Results from validation of trained model (U-Net) for OD/OC segmentation.

Dataset	Precision			Recall			F1-Score			ACC	MCC
iChallenge-GON (Validation set)	BG	OD	OC	BG	OD	OC	BG	OD	OC	0.9965	0.905
	1.0	0.8693	0.6285	0.998	0.901	0.9878	0.999	0.8847	0.7682		

This provided enough data to perform CDR estimation via the computation of the vertical diameter of both OD and OC, from the bounding boxes of the respectively detected contours. Moreover, a specific CDR threshold for GON classification was established considering a point where the number of TP and TN cases were maximized, and the standard deviation from the computed CDR of validation data was at minimum. A set of examples in this regard are shown in Figures 9 and 10, respectively. Unfortunately, a comparison against an external validation set has not been possible due to the lack of data available with the same label quality. Regardless, this segmentation network was used in conjunction with multiple image manipulation techniques, based on OpenCV libraries, to generate a custom subset of OD-crops from the LAG dataset.

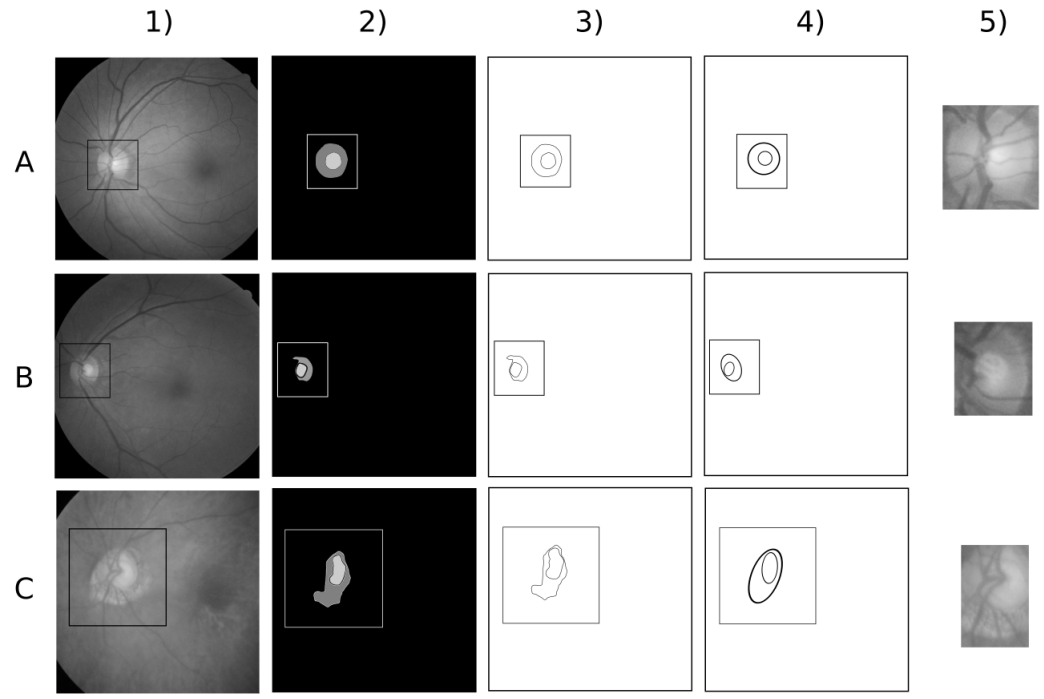

Figure 9. Performance of segmentation network (U-Net) on external dataset LAG. Columns A, B, and C represent examples of the best, mild, and worst scenarios, respectively. Each row, in this case, shows a step into the OD segmentation process, (1) input CFP, (2) output mask by U-Net, (3) output from contour filter algorithm, (4) ellipse-fitting output for detected contours, and (5) final OD crop.

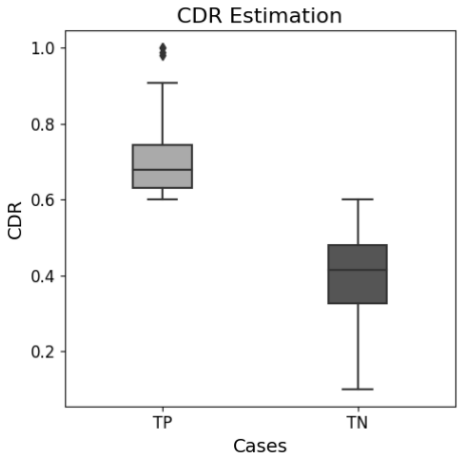

Figure 10. Boxplot related to CDR estimation based on TP and TN cases for LAG dataset.

In relation to GON classification performance, Table 5 shows results for *Global* and *OD-only* assessment from the proposed CAD system, taking into account different aspects of evaluation. First, an average value from all CNNs as a unit over TR and VAL sets is given. Finally, as a measure of robustness for the CAD system, the performance yield over the TST set in the overall average value was considered. In some cases, it had a negative effect due to the disparity of the image quality between TR/VAL and TST sets. This is mostly present in procedures related to the iChallenge-GON dataset and was expected up to a certain point. As a complement to the commonly reported metrics, BM, GBA, and MCC are included, considering their inherent properties discussed in previous sections.

Table 5. Results from Multi-stage ensemble-based CAD system proposed for GON classification.

Method [a]	Subsets for Evaluation	ACC (%)	SN (%)	SP (%)	AUC(ROC)	BM	GBA	MCC
Global Assessment	TR/VAL	98.39	97.93	98.64	0.9951	0.9752	0.9845	0.9647
	TR/VAL/TST	95.93	70.01	98.94	0.9231	0.6961	0.9307	0.7444
OD-only Assessment	TR/VAL	97.86	94.5	98.71	0.984	0.9445	0.9761	0.9345
	TR/VAL/TST	90.17	83.61	93.11	0.9119	0.7921	0.9297	0.7788

TR = Training set; VAL = Validation set; TST = Test set; [a] Average values from performance yield by ResNet-50, MobileNet_v2, and Squeezenet on each classification block.

A general comparison between the proposed CAD system and multiple studies found in the literature for GON classification is shown in Table 6, in which, for practical reasons, only those metrics reported in most cases were considered, and the results from the proposed CAD system were established based on the average value of both TR/VAL evaluation, as this was considered a less over-optimistic representation of the performance from the proposed system.

Table 6. Comparison between proposed CAD system for GON classification and currents studies.

Author	ACC (%)	SN (%)	SP (%)	AUC(ROC)
Ting et al., 2017 [12]	N/A	96.4	87.2	0.942
Chai et al., 2018 [25]	91.51	92.33	90.9	N/A
Christopher et al., 2018 [33]	N/A	84.0	83.0	0.91
Li et al., 2018 [26]	N/A	95.6	92.0	0.986
Liu et al., 2018 [34]	91.6	87.9	96.5	0.97
Al-Aswad et al., 2019 [28]	N/A	83.7	88.2	0.926
Hemelings et al., 2019 [35]	N/A	**99.0**	93.0	**0.996**
Kim et al., 2019 [32]	96.0	95.0	**100**	0.99
Liu et al., 2019 [29]	N/A	91.8286	90.043	0.9546
Phene et al., 2019 [24]	N/A	80.0	90.2	0.945
Diaz-Pinto et al., 2019 [36]	89.77	93.46	85.8	0.9605
Li et al., 2020 [30]	95.3	98.0	94.95	0.994
Sreng et al., 2020 [37]	93.308	N/A	N/A	0.924
Civit-Masot et al., 2020 [31]	88.0	91.0	86.0	0.96
Ours [a]	**98.125**	96.215	98.675	0.9895

N/A = Not Available; [a] Average values from performance of Global and OD-only classification blocks based on the subsets for TR/VAL from LAG dataset.

5. Discussion

From the comparison made in Table 6, it can be observed that specifically for GON, there are plenty of methodologies related to BC based on ML. However, there are a set of problems that have emerged during the development of the proposals presented in previous sections. First, there is an evident lack of reproducibility from the reported results, which might come from non-disclosure of CMs obtained by each dataset employed, and privacy concerns related to the use of private clinic/hospital data. Moreover, unlike other disciplines, the applications of DL techniques for tasks related to ophthalmology require strictly to undertake a labeling process that is performed by experts in the area to reduce the factor of human error during the training process. This is difficult in most cases due to a lack of resources to conduct mass clinical studies, or simply the low prevalence levels of some diseases (e.g., GON).

Second, low levels of generalization is a quite notorious problem within the datasets that are publicly available, since there are differences between key characteristics of every CFP, such as the model of the acquisition camera, resolution of CFPs, intensity of the light source, and other quality-related parameters. This causes some models within the SOTA to only obtain satisfactory results under a given group of datasets. A possible solution to this would be the use of *Domain Adaptation*, as suggested by Wang and Deng [44]. When testing their effectiveness in a real clinical environment, it denotes the extent to which work-related physicians are able to accept the results of a system that comes from a *black box process*, which is an inherent feature of DL. Some of the most frequently used procedures to reduce the negative impact of this type of system is the construction of *features maps* or *class activation maps* (heatmaps) that show which part of the CFPs analyzed by the DL system has greater relevance during the process of the final prediction.

6. Conclusions

In this work, a multi-stage ensemble-based CAD system is proposed. A mixture of deep neural network architectures was applied in conjunction with a dedicated U-shaped segmentation network in order to classify the information within a CFP as a referable case of GON.

The advent of high-performance results in DL-based systems for performing tasks related to the CAD of retinal pathologies using CPFs is promising. In some cases, they even deliver better performance compared with a medical specialist. However, there are still several aspects to improve in the current ensemble-based CAD system. First, the lack of

generalization of the less deep network—in this case, Squeezenet—adds a considerable amount of burden to the final classification assessment, which could be resolved with its replacement and experimentation of other networks such as DenseNet, or further improvements to the current training procedures for this specific model.

In terms of network structure, a tendency is observed in the selection of base structures used for classification tasks, such as ResNet and DenseNet, as well as a combination of other kinds of methodologies (e.g., transfer learning). In the case of segmentation tasks, a transition from the use of architectures such as CNNs (manually designed CNNs) to *Fully Connected Neural Networks* (FCN) is observed, as well as the use of other models such as *YOLOv7*, *Mask-RCNN*, and *DeepbV3+*. However, these results are not the product of the trivial use of different architectures applied to a particular classification and/or segmentation problem, but the product of a series of contributions in the form of processing blocks that manage, in some cases, to imitate the procedure by which a medical specialist or physician performs eye fundus examination.

In future work, the use of *multimodal* data for validation, or testing, CFPs could also be considered to produce more reliable predictions based on the complementary information obtained from in-depth anatomic studies such as OCT scans or IOP measurements.

Author Contributions: Conceptualization, C.A.V.-R., and M.E.M.-R.; Data curation, M.M.M.-V.; Formal analysis, M.E.M.-R.; Methodology, C.A.V.-R.; Resources, P.A.L.-M.; Software, C.A.V.-R.; Supervision, M.E.M.-R., H.C.d.Á., and P.A.L.-M.; Visualization, M.M.M.-V.; Writing—original draft, C.A.V.-R.; Writing—review and editing, G.R.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Council for Science and Technology (CONACyT), México, under Research Grant No. 903310.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. Publicly available datasets were analyzed in this study. These data can be found here: LAG: https://ieeexplore.ieee.org/document/8756196/ accessed on 12 May 2022, iChallenge-GON: http://refuge.grand-challenge.org accessed on 15 February 2022, RIM-ONE DL: https://www.ias-iss.org/ojs/IAS/article/view/2346 accessed on 18 November 2021.

Acknowledgments: The authors wish to thank the CONACyT for providing the funding to conduct this research, as well as to the people involved in this research at the Universidad Autónoma de Baja California (UABC), Programa para el Fortalecimiento y Calidad Educativa (PFCE) and the Programa para el Desarrollo Profesional Docente (PRODEP).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GON	Glaucomatous Optic Neuropathy
ML	Machine Learning
DL	Deep Learning
CAD	Computer-aided Diagnosis

References

1. Geneva: World Health Organization. World Report on Vision. 2019. Available online: https://www.who.int/publications/i/item/9789241516570 (accessed on 13 May 2022).
2. Bourne, R.R.; Steinmetz, J.D.; Flaxman, S.; Briant, P.S.; Taylor, H.R.; Resnikoff, S.; Casson, R.J.; Abdoli, A.; Abu-Gharbieh, E.; Afshin, A.; et al. Trends in prevalence of blindness and distance and near vision impairment over 30 years: An analysis for the Global Burden of Disease Study. *Lancet Glob. Health* **2021**, *9*, e130–e143. [CrossRef] [PubMed]
3. Zhang, J.; Tuo, J.; Wang, Z.; Zhu, A.; Machalińska, A.; Long, Q. Pathogenesis of Common Ocular Diseases. *J. Ophthalmol.* **2015**, *2015*, 734527. [CrossRef] [PubMed]
4. Assi, L.; Rosman, L.; Chamseddine, F.; Ibrahim, P.; Sabbagh, H.; Congdon, N.; Evans, J.; Ramke, J.; Kuper, H.; Burton, M.J.; et al. Eye health and quality of life: An umbrella review protocol. *BMJ Open* **2020**, *10*, e037648. [CrossRef] [PubMed]

5. Li, T.; Bo, W.; Hu, C.; Kang, H.; Liu, H.; Wang, K.; Fu, H. Applications of deep learning in fundus images: A review. *Med. Image Anal.* **2021**, *69*, 101971. [CrossRef] [PubMed]
6. Asiri, N.; Hussain, M.; Al Adel, F.; Alzaidi, N. Deep learning based computer-aided diagnosis systems for diabetic retinopathy: A survey. *Artif. Intell. Med.* **2019**, *99*, 101701. [CrossRef]
7. Tham, Y.C.; Li, X.; Wong, T.Y.; Quigley, H.A.; 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*, 2081–2090. [CrossRef]
8. Yan, Q.; Weeks, D.E.; Xin, H.; Swaroop, A.; Chew, E.Y.; Huang, H.; Ding, Y.; Chen, W. Deep-learning-based prediction of late age-related macular degeneration progression. *Nat. Mach. Intell.* **2020**, *2*, 141–150. [CrossRef]
9. Nawaz, M.; Nazir, T.; Javed, A.; Tariq, U.; Yong, H.S.; Khan, M.A.; Cha, J. An Efficient Deep Learning Approach to Automatic Glaucoma Detection Using Optic Disc and Optic Cup Localization. *Sensors* **2022**, *22*, 434. [CrossRef]
10. Peng, Y.; Dharssi, S.; Chen, Q.; Keenan, T.D.; Agrón, E.; Wong, W.T.; Chew, E.Y.; Lu, Z. DeepSeeNet: A Deep Learning Model for Automated Classification of Patient-based Age-related Macular Degeneration Severity from Color Fundus Photographs. *Ophthalmology* **2019**, *126*, 565–575. [CrossRef]
11. Mateen, M.; Malik, T.S.; Hayat, S.; Hameed, M.; Sun, S.; Wen, J. Deep Learning Approach for Automatic Microaneurysms Detection. *Sensors* **2022**, *22*, 542. [CrossRef]
12. Ting, D.S.W.; Cheung, C.Y.L.; Lim, G.; Tan, G.S.W.; Quang, N.D.; Gan, A.; Hamzah, H.; Garcia-Franco, R.; Yeo, I.Y.S.; Lee, S.Y.; et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA-J. Am. Med Assoc.* **2017**, *318*, 2211–2223. [CrossRef] [PubMed]
13. Pham, Q.T.; Ahn, S.; Song, S.J.; Shin, J. Automatic drusen segmentation for age-related macular degeneration in fundus images using deep learning. *Electronics* **2020**, *9*, 1617. [CrossRef]
14. Gondal, W.M.; Kohler, J.M.; Grzeszick, R.; Fink, G.A.; Hirsch, M. Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images. In Proceedings of the 2017 IEEE international conference on image processing (ICIP), Beijing, China, 17–20 September 2017; pp. 2069–2073. [CrossRef]
15. Lee, R.S.T. AI Fundamentals. In *Artificial Intelligence in Daily Life*; Springer: Singapore, 2020; pp. 19–37. [CrossRef]
16. Jammal, A.A.; Thompson, A.C.; Mariottoni, E.B.; Berchuck, S.I.; Urata, C.N.; Estrela, T.; Wakil, S.M.; Costa, V.P.; Medeiros, F.A. Human Versus Machine: Comparing a Deep Learning Algorithm to Human Gradings for Detecting Glaucoma on Fundus Photographs. *Am. J. Ophthalmol.* **2020**, *211*, 123–131. [CrossRef]
17. Ting, D.S.W.; Pasquale, L.R.; Peng, L.; Campbell, J.P.; Lee, A.Y.; Raman, R.; Tan, G.S.W.; Schmetterer, L.; Keane, P.A.; Wong, T.Y. Artificial intelligence and deep learning in ophthalmology. *Br. J. Ophthalmol.* **2019**, *103*, 167–175. [CrossRef]
18. Dai, L.; Fang, R.; Li, H.; Hou, X.; Sheng, B.; Wu, Q.; Jia, W. Clinical report guided retinal microaneurysm detection with multi-sieving deep learning. *IEEE Trans. Med Imaging* **2018**, *37*, 1149–1161. [CrossRef]
19. Sarhan, M.H.; Albarqouni, S.; Yigitsoy, M.; Navab, N.; Eslami, A. Multi-scale microaneurysms segmentation using embedding triplet loss. *Lect. Notes Comput. Sci.* **2019**, *11764 LNCS*, 174–182. [CrossRef]
20. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef]
21. Chicco, D.; Tötsch, N.; Jurman, G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Min.* **2021**, *14*, 13. [CrossRef]
22. Hicks, S.A.; Strümke, I.; Thambawita, V.; Hammou, M.; Riegler, M.A.; Halvorsen, P.; Parasa, S. On evaluation metrics for medical applications of artificial intelligence. *Sci. Rep.* **2022**, *12*, 5979. [CrossRef] [PubMed]
23. Halimu, C.; Kasem, A.; Newaz, S.H.S. *Empirical Comparison of Area under ROC Curve (AUC) and Mathew Correlation Coefficient (MCC) for Evaluating Machine Learning Algorithms on Imbalanced Datasets for Binary Classification*; ACM Press: New York, NY, USA, 2019; pp. 1–6. [CrossRef]
24. Phene, S.; Dunn, R.C.; Hammel, N.; Liu, Y.; Krause, J.; Kitade, N.; Schaekermann, M.; Sayres, R.; Wu, D.J.; Bora, A.; et al. Deep Learning and Glaucoma Specialists: The Relative Importance of Optic Disc Features to Predict Glaucoma Referral in Fundus Photographs. *Ophthalmology* **2019**, *126*, 1627–1639. [CrossRef] [PubMed]
25. Chai, Y.; Liu, H.; Xu, J. Glaucoma diagnosis based on both hidden features and domain knowledge through deep learning models. *Knowl.-Based Syst.* **2018**, *161*, 147–156. [CrossRef]
26. Li, Z.; He, Y.; Keel, S.; Meng, W.; Chang, R.T.; He, M. Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs. *Ophthalmology* **2018**, *125*, 1199–1206. [CrossRef] [PubMed]
27. Shibata, N.; Tanito, M.; Mitsuhashi, K.; Fujino, Y.; Matsuura, M.; Murata, H.; Asaoka, R. Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. *Sci. Rep.* **2018**, *8*, 14665. [CrossRef]
28. Al-Aswad, L.A.; Kapoor, R.; Chu, C.K.; Walters, S.; Gong, D.; Garg, A.; Gopal, K.; Patel, V.; Sameer, T.; Rogers, T.W.; et al. Evaluation of a Deep Learning System for Identifying Glaucomatous Optic Neuropathy Based on Color Fundus Photographs. *J. Glaucoma* **2019**, *28*, 1029–1034. [CrossRef] [PubMed]
29. Liu, H.; Li, L.; Wormstone, I.M.; Qiao, C.; Zhang, C.; Liu, P.; Li, S.; Wang, H.; Mou, D.; Pang, R.; et al. Development and Validation of a Deep Learning System to Detect Glaucomatous Optic Neuropathy Using Fundus Photographs. *JAMA Ophthalmol.* **2019**, *137*, 1353–1360. [CrossRef]
30. Li, F.; Yan, L.; Wang, Y.; Shi, J.; Chen, H.; Zhang, X.; Jiang, M.; Wu, Z.; Zhou, K. Deep learning-based automated detection of glaucomatous optic neuropathy on color fundus photographs. *Graefe's Arch. Clin. Exp. Ophthalmol.* **2020**, *258*, 851–867. [CrossRef]

31. Civit-Masot, J.; Dominguez-Morales, M.J.; Vicente-Diaz, S.; Civit, A. Dual Machine-Learning System to Aid Glaucoma Diagnosis Using Disc and Cup Feature Extraction. *IEEE Access* **2020**, *8*, 127519–127529. [CrossRef]
32. Kim, M.; Han, J.C.; Hyun, S.H.; Janssens, O.; Hoecke, S.V.; Kee, C.; Neve, W.D. Medinoid: Computer-aided diagnosis and localization of glaucoma using deep learning. *Appl. Sci.* **2019**, *9*, 3064. [CrossRef]
33. Christopher, M.; Belghith, A.; Bowd, C.; Proudfoot, J.A.; Goldbaum, M.H.; Weinreb, R.N.; Girkin, C.A.; Liebmann, J.M.; Zangwill, L.M. Performance of Deep Learning Architectures and Transfer Learning for Detecting Glaucomatous Optic Neuropathy in Fundus Photographs. *Sci. Rep.* **2018**, *8*, 16685. [CrossRef]
34. Liu, S.; Graham, S.L.; Schulz, A.; Kalloniatis, M.; Zangerl, B.; Cai, W.; Gao, Y.; Chua, B.; Arvind, H.; Grigg, J.; et al. A Deep Learning-Based Algorithm Identifies Glaucomatous Discs Using Monoscopic Fundus Photographs. *Ophthalmol. Glaucoma* **2018**, *1*, 15–22. [CrossRef] [PubMed]
35. Hemelings, R.; Elen, B.; Barbosa-Breda, J.; Lemmens, S.; Meire, M.; Pourjavan, S.; Vandewalle, E.; de Veire, S.V.; Blaschko, M.B.; Boever, P.D.; et al. Accurate prediction of glaucoma from colour fundus images with a convolutional neural network that relies on active and transfer learning. *Acta Ophthalmol.* **2019**, *98*, e94–e100. [CrossRef]
36. Diaz-Pinto, A.; Morales, S.; Naranjo, V.; Köhler, T.; Mossi, J.M.; Navea, A. CNNs for automatic glaucoma assessment using fundus images: An extensive validation. *BioMed. Eng. Online* **2019**, *18*, 1–9. [CrossRef] [PubMed]
37. Sreng, S.; Maneerat, N.; Hamamoto, K.; Win, K.Y. Deep learning for optic disc segmentation and glaucoma diagnosis on retinal images. *Appl. Sci.* **2020**, *10*, 4916. [CrossRef]
38. Fernández, A.; García, S.; Galar, M.; Prati, R.C.; Krawczyk, B.; Herrera, F. *Learning from Imbalanced Data Sets*; Springer International Publishing: Berlin/Heidelberg, Germany, 2018. [CrossRef]
39. Yip, M.Y.; Lim, G.; Lim, Z.W.; Nguyen, Q.D.; Chong, C.C.; Yu, M.; Bellemo, V.; Xie, Y.; Lee, X.Q.; Hamzah, H.; et al. Technical and imaging factors influencing performance of deep learning systems for diabetic retinopathy. *Npj Digit. Med.* **2020**, *3*, 31–34. [CrossRef] [PubMed]
40. Li, L.; Xu, M.; Liu, H.; Li, Y.; Wang, X.; Jiang, L.; Wang, Z.; Fan, X.; Wang, N. A Large-Scale Database and a CNN Model for Attention-Based Glaucoma Detection. *IEEE Trans. Med. Imaging* **2020**, *39*, 413–424. [CrossRef]
41. Baidu Inc. iChallenge-GON, 2020. Available online: http://refuge.grand-challenge.org (accessed on 15 February 2022).
42. Batista, F.J.F.; 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*, 161–167. [CrossRef]
43. Li, X.; Hu, X.; Yu, L.; Zhu, L.; Fu, C.W.; Heng, P.A. CANet: Cross-Disease Attention Network for Joint Diabetic Retinopathy and Diabetic Macular Edema Grading. *IEEE Trans. Med. Imaging* **2020**, *39*, 1483–1493. [CrossRef] [PubMed]
44. Wang, M.; Deng, W. Deep visual domain adaptation: A survey. *Neurocomputing* **2018**, *312*, 135–153. [CrossRef]