Glaucoma Grading Using Multimodal Imaging and MultiLevel CNN

Marcos M. Ferreira, Geraldo Braz Junior, João D. S. Almeida, Anselmo C. Paiva

***Abstract*—Glaucoma is considered the leading cause of blindness. Because there is no cure for the disease, treatment must begin promptly to prevent disease progression, which leads to severe visual impairment and, in some cases, total blindness. In this scenario, diagnosis in the early stages is essential and could be accomplished through screening exams. Recent studies have combined fundus images and Optical Coherence Tomography (OCT) volumes as indicators of disease progression in computer vision methods. In this work, we develop a method based on deep learning, which uses fundus images and OCT volumes to aid in detecting glaucoma at early or progressive stages. In this way, it can have clinical use, being able to be used as a tool not only for the detection of the disease but also that it helps in searching for more severe cases of the disease. As a result, the proposed method achieves 0.886 kappa score.**

***Index Terms*—Glaucoma Diagnosis, Multimodal Imaging, Deep Learning.**

I. Introduction

Glaucoma is considered one of the leading causes of vision impairment and the main cause of irreversible blindness [1], [2], [3]. The number of people (aged 40 - 80 years) in Latin America and the Caribbean with glaucoma in 2013 was about 6.59 million, increasing to 8.11 million in 2020 and approximately 13 million in 2040 [2]. Glaucoma is a disorder in which excessive Intraocular Pressure (IOP) causes gradual damage to the optic nerve. This damage causes bilateral blindness, which can progress to total vision loss. In addition to high IOP, other risk factors such as older age, family history, and ethnicity are other factors that should be considered [4].

Despite being irreversible, vision loss can be avoided if medical procedures are performed in the early stages of the disease to prevent its progress. According to the World Health Organization (WHO) [3], only 11% of people who received timely diagnosis and treatment reported having moderate or severe visual impairment or blindness resulting from more severe forms of the condition. However, as glaucoma is asymptomatic in the early stages, early diagnosis is complex, and the disease is usually diagnosed when the effects of increased IOP have already caused damage to vision [4].

Because visual impairment due to glaucoma is irreversible, interventions are needed in the early stages of the disease to prevent complete vision loss [5]. Glaucoma screening can be made with color fundus photography images and Optical Coherence Tomography (OCT) [6]. Through fundus images and OCT scans, it is possible to analyze biomarkers that indicate glaucoma, such as the cup-to-disc ratio and retinal nerve fiber layer (RNFL) thicknesses [7]. However, some factors that make early diagnosis difficult, such as the absence of symptoms in the initial stages and the great number of patients potentially per specialists qualified to perform the diagnosis, which remained for a long time to analyze a large number of exams. In Brazil, the ratio of ophthalmologists per inhabitant is on average 1:10875, whereas in the northern region this ratio is 1:19512 [8]. Furthermore, human beings are susceptible to fatigue, emotional fluctuation, and other factors that can lead to biased results [9]. In this context, automatic methods that can help specialists detect glaucoma using imaging exams may have great potential for clinical use. Methods based on deep learning techniques have shown promising results in medical image classification tasks.

This present study proposes a method based on deep learning which can grade the stage of glaucoma using fundus and OCT images, not only to facilitate early diagnosis but also to allow the identification of advanced cases which require urgent intervention. In this work, we perform model optimization, seeking the best parameters to obtain the best classifiers for each type of medical image.. We also investigate the performance of a multilevel architecture and the use of ensembles strategies to classify two modalities of images as normal (no glaucoma), early-stage or more advanced (progressive) glaucoma, achieving a kappa score of 0.89, which implies a strong level of agreement with specialists, as well as related works.

The main contributions of this work are a) a CNN architecture optimized for grading the glaucoma stage based on multimodal medical image (fundus image, optic disc, OCT volumes) and ensemble strategies; b) input magnification, adding the optic disc region obtained from the fundus image in a separate input, and thus making its features gain more importance in the process; c) easily configurable method and expandable convolutional neural network architecture for glaucoma grading.

The remaining sections are organized as follows. Section II presents some related works. Section III presents the proposed method. Section IV presents the results and evaluation of our method. Section V concludes this paper.

II. Related Work

Several methods have been proposed to detect glaucoma, most applied in datasets formed by fundus images. Recently,

Marcos M. Ferreira is with Applied Computing Center, Federal University of Maranhão (UFMA) Brasil e-mail:marcos.ferreira@discente.ufma.br.

Geraldo Braz, João D. s. Almeida and Anselmo C. Pailva are with Applied Computing Center, Federal University of Maranhão (UFMA) Brasil e-mail:{geraldo, jdallyson, paiva}@nca.ufma.br

methods that use other types of images, such as OCT and ultrasonic biomicroscopy have been proposed. These images are widely used because it is possible to detect diseases such as macular degeneration, diabetic retinopathy, and glaucoma.

A multimodal model was used to perform binary glaucoma classification combining multiple convolutional networks to extract features from fundus and OCT images in [5]. Similarly, [6] obtain feature maps based on the RNFL and retinal ganglion cell (RGC) thicknesses using CNN and Random Forest classifier. Ma et al. [10] describe a method used to classify fundus images of patients in three classes: healthy, suspect, and early-stage glaucoma. The method uses a total of 15 clinical features, such as age, IOP, RNFL thickness and uses segmentation to calculate the optic disc ratio. Experiments were carried out with different classifiers, with the best AUC result achieved by the SVM classifier. Xiong et al. [11] proposed a model called FusionNet. This model combines OCT images with images obtained from visual field assessment reports. It uses a multilevel network, with different levels formed by convolutional layers, followed by a classifier which predicts the probabilities that the images are from a patient with glaucoma.

Fang et al. [12] describe an architecture that combines features extracted from fundus and OCT images. The target is to classify images into normal, early-stage glaucoma, and moderate or advanced-stage glaucoma. Feature extraction was performed using an encoder, ResNet34.The best result was obtained by combining the fundus images, the OCT volumes, and the optic disc region using ordinal regression. In the same way, Li et al. [13] present a method that uses a hierarchical concatenation of features performed by a network fed by fundus and OCT images. At each network branch, only variations of ResNet were evaluated. Using the proposed concatenation technique, it was possible to increase the total number of features used for classification, in addition to using features from different scales. To achieve a higher kappa, an ensemble strategy of the value predictions of the two best models was used.

Tian et al [14] present a network named GC-Net to perform glaucoma classification into normal, early-stage glaucoma and progressive glaucoma, using as input optic disc region. Architecture is formed by a pre-trained CNN used as feature extractor, and an attention module formed by a global attention block and a class attention block.

This work proposes a glaucoma grading method using 2D fundus images and 3D OCT volumes through deep learning, based on visual feature extraction and transfer learning. A hyperparameter optimization was carried out, evaluating different CNN models, handling each imaging modality as input. Using ensemble and feature combination strategies, it was possible to use features extracted by different CNNs to grade glaucoma stage. Through activation maps, it was possible to demonstrate the contribution of each imaging modality for predicting the stage of glaucoma.

III. Proposed Method

The proposed method uses convolutional neural networks combined in a multilevel architecture for predicting glaucoma stage in early or progressive (intermediate and advanced glaucoma). Our proposal combines fundus images and OCT volumes, thus providing a multimodal analysis. In this work, we used multilevel architectures and ensemble strategies, allowing designing models that use more than one medical imaging modality for grading glaucoma stage.

The method used for developing this research comprises five main steps: Image acquisition, image preprocessing, model construction, models ensemble and evaluation of the research method. The steps of the proposed method, along with some keywords that summarize each step, are shown in Fig. 1.

A. Image Acquisition

The dataset used in this work was made available to the participants of the GAMMA challenge (*Glaucoma Grading from Multi-Modality Images*) [7]. The multimodal dataset comprises two imaging modalities exams, fundus images and 3D OCT volumes, used for diagnosing Glaucoma.

The dataset is composed of pairs of images. Each pair consists of two imaging modalities, a fundus image and a volume scan of 256 3D OCT slices. In total, 200 pairs were made available, 100 that form the training set and 100 that form the test set. Both sets have pairs that belong to one of three possible classes: normal (class 0 - 50 pairs), early-stage glaucoma (class 1 - 26 pairs), and intermediate or advanced-stage glaucoma (class 2 - 24 pairs). Fig. 2 presents examples of image pairs, each with three OCT slices and a fundus image, belonging to one of three classes: normal (no glaucoma), early-stage glaucoma, and progressive glaucoma (moderate or advanced stage).

B. Image Preprocessing

In preprocessing, fundus images were resized to 128x128, and each slice of the OCT volumes was resized to 128x128. After resizing, a bilateral filter was applied for noise reduction present in OCT slices. We also downsample OCT depth to 64 using spline interpolation to minimize computational cost.

As optic disc is one of the eye regions most affected by the increase in intraocular pressure caused by glaucoma [7], we segmented this region by creating a third image with just the optical disc. We used this region of interest to increase the classification capacity of the models, given its relevance to the diagnosis and the fact that the available base for training is small. To capture the region, disc segmentation was performed using a UNET [15], pre-trained on another fundus image dataset, RIMONE [16]. From the results generated by the segmentation, the ROI was cropped, and new images were generated, containing only the optic disc region. Fig. 3 presents an example of a region of interest containing an optical disc extracted from a fundus image.

C. Models Construction

Convolutional Neural Networks are a class of deep neural networks that have been successfully used in image processing and classification applications. The main feature of this neural network is that in some layers, filters are used to perform

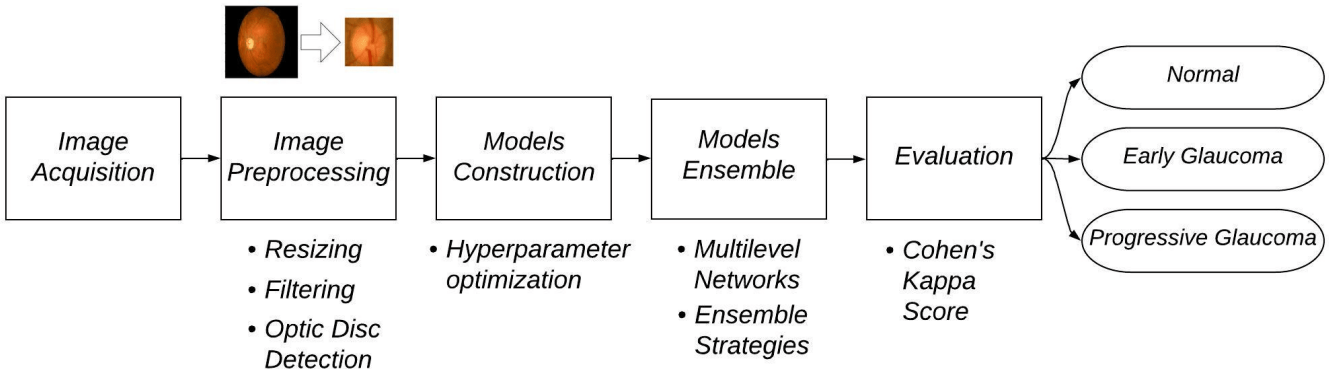

Fig. 1. Steps of the proposed method.

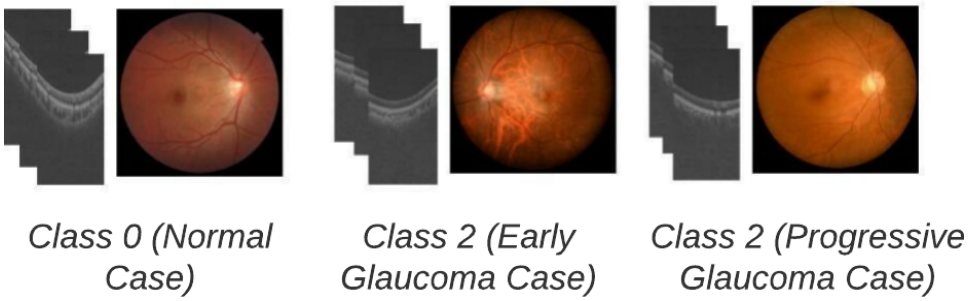

Fig. 2. Example of dataset images

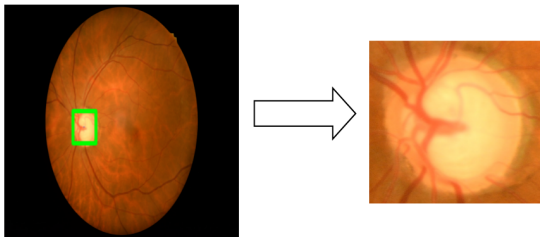

Fig. 3. (a) Original fundus image. (b) Region of interest obtained from segmentation.

the mathematical operation of convolution (Equation 1). The network adjusts the weights that form the filters during the training process, minimizing the need for manual feature extraction. As a result of the convolution operations, feature maps are obtained, which are used as input to a classifier, usually formed by fully connected layers.

$$s(t) = (x * w)(t) = \sum_{a=0}^{t} x(a)w(t-a) \tag{1}$$

in which x is the input, w is the kernel filter. The output s is the feature map at a index t. In machine learning applications, the input is usually a multidimensional array of data, and the kernel is usually an array of parameters adapted by the learning algorithm [17].

In most applications, CNNs receive 2D images as input, performing feature extraction from each image for classification. However, there are also 3D CNNs, which receive input volumes formed by several images named slices. In this work, we use multimodal models that receive 2D fundus images and 3D OCT volumes as input, combining the features extracted from each modality for classification. As 3D volumes were used in this work, the 3D feature extractor is formed by pretrained 3D CNN models, proposed by [18]. These 3D models were obtained from state-of-the-art 2D CNNs. Both CNN 2D and CNN 3D were loaded with Imagenet weights [19]. Imagenet is a large manually annotated image database, a benchmark provided for researchers to evaluate their methods and algorithms in the annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC). The weights obtained in the challenge were expanded to be used in a network with a 3D volume scan as input.

We evaluate the following pretrained CNNs: VGG19 [20], Inception V3 [21], DenseNet121/169 [22] and ResNet50/152 [23], with 2D CNNs fed with fundus images and 3D CNNs fed with OCT volumes. The CNNs were imported without the original classifier and loaded with Imagenet weights to perform transfer learning.

In the training step, the search for the best CNN hyperparameters for the classification of each imaging modality was carried out, using a Sequential Model-Based Optimization - SMBO [24]. The SMBO differs from Grid Search and Random Search because it considers the past performance of hyperparameters in its search. In contrast, in the other two methods, the search is independent of past evaluations. SMBO methods work by looking for the next set of hyperparameters

to evaluate in the objective function, selecting the hyperparameters that stand out in a surrogate probabilistic function, which is less expensive to evaluate. If the evaluated values also show promising results in the objective function, they will be incorporated into the set of the best hyperparameters.

The following hyperparameters were included in the optimization process: feature extractor (CNNs), batch size, dropout rate, learning rate, number of dense layers, number of dense nodes of the first dense layer and dense nodes divisor, which was used to calculate the total of dense nodes of the second layer (if was necessary). Table I presents the search space used for optimization of each hyperparameter. Optimization processes were carried out in search of models that have as input each imaging modality: two 2D models for extracting features from fundus and optic disc images and a 3D model for extracting features from OCT volumes.

The optimization was performed using the 100 labeled pairs that form the training set. Ten pairs were randomly selected to be used as a validation set. The training strategy used to avoid overfitting was early stopping, having as variables monitored the training accuracy or the validation loss.

TABLE I
HYPERPARAMETER SEARCH SPACE

Parameter	Search Space	Distribuition
CNN	VVGG19, InceptionV3, ResNet50, ResNet152, DenseNet121, DenseNet169	Categorical
Dropout rate	[0.0, 0.5, step=0.1]	Discrete Uniform
Batch size	[1,2,3]	Categorical
Learning rate	[1E-5, 1E-4 ,1E-3]	Categorical
Numbers of Layers	[1, 2]	Categorical
Numbers of dense nodes	[64, 128, 256,512]	Categorical
Dense nodes divisor	[2, 4, 8]	Categorical

D. Models Ensemble

A multilevel architecture, shown in Fig. 4, was chosen for the multimodal experiments. Figure presents examples of images used as input for each level for CNNs evaluated as feature extractors and the layers that are the classifier. A multilevel architecture makes it possible to use different images as input. The architecture comprises 3 levels, the first having fundus image as input, the second receiving optic disc region and the last receiving OCT volumes. We used optic disc as second level since this is a region of interest (ROI) because it is possible to evaluate the cup-to-disc ratio, which is a biomarker for assessing the progress of glaucoma.

Features extracted from each level are used for classification. We used the optimized models obtained previously, with classifier removal, combined in a multilevel architecture. We evaluate two feature combination strategies, feature concatenation and addition. In the first strategy, we combine feature vectors by concatenating one feature vector after another to obtain a new feature map, as used in previous works [9], [11]. In the second strategy, we combine feature vectors by adding all vectors to obtain a new feature map. Finally, we add three layers, two fully connected, which use the Relu activation function, and the last, which uses softmax regression to predict the class into normal (C0), early glaucoma (C1) or progressive glaucoma (C2). Then, optimization of the classification layers was performed, with the feature extractors layers having their weights frozen during the process.

As the images that make up the dataset have different features, it was decided to adopt ensemble strategies. In this case, a combination of the results obtained by the three best models in the optimization step for each input image is performed. Two strategies were evaluated to obtain the final result: average ensemble and mode ensemble. In average ensemble, consider a task with N classes and M classifiers. Being zij the value of the jth model (j=1, ... M) of the ith node of the last layer (i=1, ...,N). The average value of all models for the ith node is given by Equation 2 [9]. In the second strategy, each individual classifier votes for a class, and the majority wins. In statistical terms, the predicted target label of the ensemble is the mode of the distribution of individually predicted labels [25]. As there are three possible classes and three classifiers, in cases of a tie in the voting, the result predicted by the model fed with fundus images that reached the highest kappa in evaluation was chosen as the final result. An advantage of the ensemble approach is that it was possible to increase the resolution of the fundus images to 224x224 pixels for further testing. The same was not possible with OCT volumes due to the computational cost.

$$v(i) = \frac{1}{M}\sum_{j=1}^{M} zji \tag{2}$$

in which M is the number of classifiers used.

E. Evaluation

After the training steps, the models were saved and evaluated using the set test, formed by 100 unlabeled pairs. The results were saved in a CSV file and sent for online evaluation, which returned the corresponding Cohen's kappa coefficient (3).

$$k = \frac{p_0 - p_e}{1 - p_e} \tag{3}$$

in which p_0 is accuracy and p_e is the sum of the products of the actual and predicted numbers corresponding to each category, divided by the square of the total number of samples.

IV. RESULTS

In this work, a method for grading glaucoma stage using multimodal models is proposed. In the model building stage, the hyperparameters optimization of 2D and 3D CNNs was carried out, which received as input fundus images and optic disc images, with sizes of 128x128 and 224x224, and OCT volumes with a size of 128x128x64. In the ensemble step, models that achieved the highest kappa scores were combined in a multilevel architecture, in which the features extracted by each model were combined and used for classification. In this step, ensemble strategies were also evaluated, in which

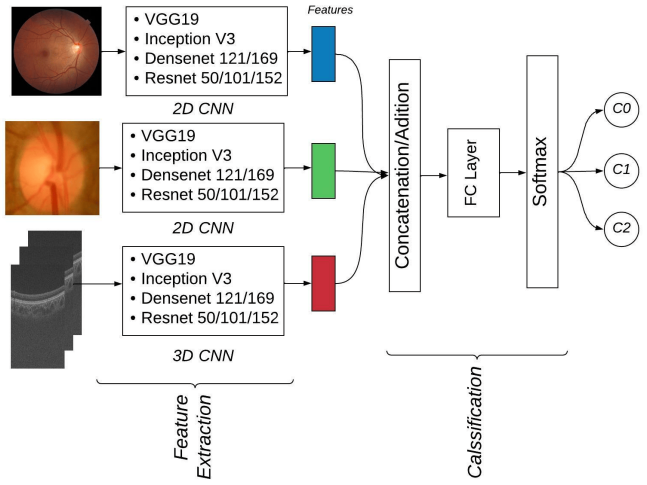

Fig. 4. Proposed multilevel architecture.

the predictions obtained by each model were combined to obtain the final result. The optuna [26] framework was used to carry out the optimization process. Since labels of the test set were not made available, it was not possible to calculate other metrics and the performance of the models is measured in terms only of Cohen's kappa coefficient, used to evaluate models in the GAMMA challenge. The bests results achieved in the model construction step are presented in Table II.

TABLE II
RESULTS OBTAINED IN MODELS CONSTRUCTION STEP.

Dataset	Resolution	CNN	Kappa
Fundus	128^2	VGG19	0.799
Fundus	224^2	Densenet169	0.855
Optic Disc	128^2	VGG19	0.731
Optic Disc	224^2	Densenet169	0.703
OCT	128^2x64	Densenet121	0.783

In the ensemble step, a multilevel architecture was evaluated, fed with fundus images, optic disc region and OCT volumes. The CNNs that achieved the highest scores had their classifier removed, and a new one was added. Two ensemble strategies were also evaluated in this step. For this approach, three models had their predictions values combined: 2D DenseNet169 fed with 224x224 fundus images; 2D VGG19 fed with 128x128 optic disc images; and 3D DenseNet121 fed with OCT volumes. The results are presented in the table III.

The results show that color fundus images outperform OCT volumes and optic disc images for grading the stage of glaucoma using a single modality. Results also suggest that ensembling predicted values outperform concatenation and adding features. It was possible to visualize which areas of the images were important for the prediction made by the models by creating activation maps, which helped understand why models that used fundus images performed better. Figs. 5, 6, 7, 8 present images and its respective activation maps,

TABLE III
RESULTS OBTAINED USING MULTILEVEL NETWORKS AND ENSEMBLE STRATEGIES. THE BEST RESULT IS SHOWN IN BOLD.

Images	Ensemble Strategy	Kappa
Fundus/Optic Disc/OCT	Features Concatenation	0.836
Fundus/Optic Disc/OCT	Features Addition	0.839
Fundus/Optic Disc/OCT	Average Ensemble	0.863
Fundus/Optic Disc/OCT	Model Ensemble	**0.886**

obtained using the features used for each model to perform classification. It is possible to visualize that the region of the optic disc was determined for the classification, which was already expected due to the cup-to-disc reason being a biomarker for the diagnosis.

In Fig. 6, it is possible to visualize a possible error committed by the model since the optic disc region was not used, which can be explained by this image has a bad quality near the disc region. Regarding OCT images, the primary biomarker for glaucoma is the retinal layer thicknesses, so to achieve better results, it would be necessary to perform a precise segmentation to estimate the thickness values, as carried out in work [6]. However, the annotation of the segmentation region of each layer was not available for training.

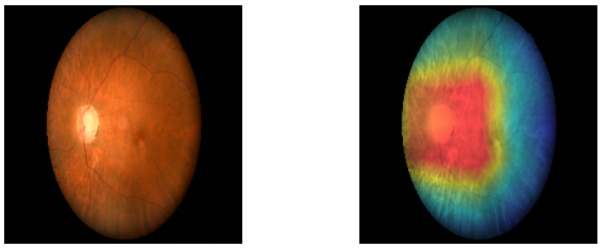

Fig. 5. Case study: success case. Optic disc was decisive for classification (class activation map image)

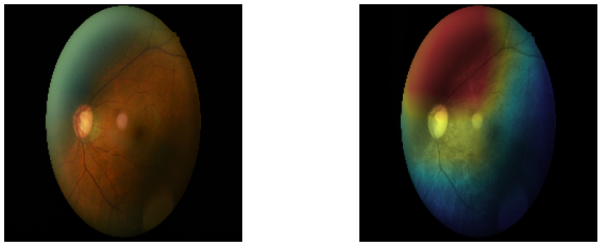

Fig. 6. Case study: error case. Optic disc was not decisive for classification (class activation map image)

Fig. 7 presents an activation map generated by a model that had as input images with the optic disc region. Through the activation maps, it is possible to find an explanation for the superior performance of the models that received the complete fundus images as input. The models that used images containing only the optical disc did not use the whole region, only a part, which could led to incorrect predictions.

Fig. 8 presents the activation map of one of the slices of an OCT volume used for grading the glaucoma stage. We observe

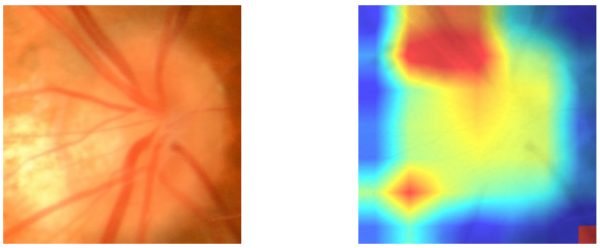

Fig. 7. Case study: activation map for a optic disc region.

that the model mainly used the central region of the image, close to the outer nuclear layer, and have not used the region next to the retinal nerve fiber layer, crucial for classifying the stage of glaucoma [6], [27], which can explain the lower results than those obtained by the models that used eye fundus images. Even so, the combination of results obtained by the models made it possible to achieve more accurate results, highlighting the importance of using more than one image modality to grade the stages of glaucoma.

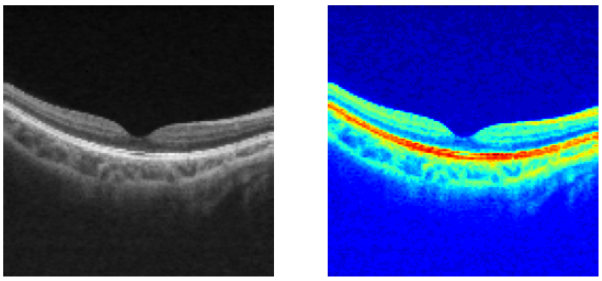

Fig. 8. Case study: activation map for OCT slice.

Table IV presents a comparison with related works that use the same dataset. The Kappa metric is used to measure interrater reliability. According to Cohen, a Kappa value between 0.81 - 1.00 indicates an almost perfect level of agreement. Another interpretation proposed by [28], suggests that values close to 0.90 indicate a strong or almost perfect level of agreement. In terms of comparison, the kappa values achieved with the proposed method are close to the best result achieved in the GAMMA challenge and other related works, indicating a strong level of agreement with classification made by specialists who used another exam, visual field reports [7].

The work [12] presents a baseline for the dataset. This work tests models with only one or two imaging modalities, evaluating a ResNet34 as a feature extractor. The work's main contribution was to present the dataset and analyze the feasibility of feature extraction methods for diagnosis. A multiresolution fusion features method was developed in [13] alongside an architecture with 3 branches, achieving a kappa score of 0.84 as the best result. The features were extracted by pre-trained ResNet models, like [12]. Only the best-selected model (manually selected using part of the training dataset) was evaluated with the test set. In [14], the authors locate and crop the optic disc as input to a CNN model that was modified using attention blocks. The method does not use OCT volume as input which we find in our experiments that refine the results as presented in Table III.

Our method explores a multilevel architecture using pre-trained and optimized CNN. We also add the optic disc image as input to emphasize regions important for glaucoma diagnosis. The method applies optimization to achieve the best model configuration. Also, each network is fine-tuned for the problem, which we can check by the presented applicability, where it is correctly defined concerning the expectation of manifestation of the disease. The final model, trained without sample selection, and evaluated in test set (available only with the GAMMA challenge organizers), reaches kappa=0.886, which is statistically classified with a high level of agreement.

TABLE IV
COMPARISON AMONG RELATED WORKS RESULTS.

Work	Dataset	Classes	Metric
Fang et al. [12]	Fundus/OCT	Normal/Early/Adv	0.86 (Kappa)
Li et al. [13]	Fundus/OCT	Normal/Early/Adv	0.89 (Kappa)
Tian et al. [14]	Fundus	Normal/Early/Adv	0.86 (Kappa)
Our Method	Fundus/OCT	Normal/Early/Adv	0.886 (Kappa)

V. CONCLUSIONS

In this work, we presented a method for glaucoma grading using two modalities of medical images: fundus and OCT volumes. The main objective of the work was to propose a method that uses multimodal medical imaging to detect glaucoma and identify the glaucoma stage.

The results show that the combination of visual features extracted from fundus image and OCT, it is possible to achieve good results not only in the detection of glaucoma but also in grading stages of the disease, allowing individuals with more severe cases of the disease can be identified more quickly. Using different images is possibly a promising approach to developing tools that can aid specialists in making diagnoses faster and more accurate and allow the analysis of different clinical parameters.

This work evaluates different CNN for different image modalities considering visual differences among them, carries out a search for optimization of hyperparameters for best models to classify each image and presents an explicability of which part of the images was more relevant for classification if there is a relationship with clinical parameters used by specialists

The results also suggest that ensemble strategies improve the model when compare to those obtained by multilevel networks where the concatenation and addition of resources were used, which can be explained by the fact that the imaging modalities are very distinct. The best results were achieved using the VGG19 and Densenet as a backbone, unlike the other related works, which used Resnet. However, none of the works mentioned used model optimization techniques.

As future works, it is intended to evaluate the use of learning transfer, through the training of the models in datasets with fundus images classified with the presence or absence of glaucoma, for later adjustment of the models using the images from the GAMMA dataset. It is also necessary to investigate approaches that use other techniques that are not based on convolutions, such as vision transformers networks, used by

[29], and Self-Organized Operational Neural Networks used by [30], who achieved good results in detecting the presence of glaucoma in fundus images. However, to use these networks, larger datasets are needed than those used for training convolutional neural networks. Another approach that deserves investigation is using a clinical parameter for classification, such as the optic disc cup ratio, used by [31]. This parameter, combined with the models' predictions, could increase the accuracy of the obtained results.

ACKNOWLEDGMENTS

The authors acknowledge the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil - Finance Code 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil, and Fundação de Amparo à Pesquisa Desenvolvimento Científico e Tecnológico do Maranhão (FAPEMA) (Brazil), Empresa Brasileira de Serviços Hospitalares (Ebserh) Brazil (Grant number 409593/2021-4) for the financial support.

REFERENCES

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

[2] Y.-C. Tham, X. Li, T. Y. Wong, H. A. Quigley, T. Aung, and C.-Y. Cheng, "Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis," *Ophthalmology*, vol. 121, no. 11, pp. 2081–2090, 2014.

[3] WHO, "World report on vision," 2019. https://www.who.int/publications/i/item/9789241516570.

[4] A. Sarhan, J. Rokne, and R. Alhajj, "Glaucoma detection using image processing techniques: A literature review," *Computerized Medical Imaging and Graphics*, p. 101657, 2019.

[5] P. Mehta, C. A. Petersen, J. C. Wen, M. R. Banitt, P. P. Chen, K. D. Bojikian, C. Egan, S.-I. Lee, M. Balazinska, A. Y. Lee, *et al.*, "Automated detection of glaucoma with interpretable machine learning using clinical data and multimodal retinal images," *American Journal of Ophthalmology*, vol. 231, pp. 154–169, 2021.

[6] G. An, K. Omodaka, K. Hashimoto, S. Tsuda, Y. Shiga, N. Takada, T. Kikawa, H. Yokota, M. Akiba, and T. Nakazawa, "Glaucoma diagnosis with machine learning based on optical coherence tomography and color fundus images," *Journal of healthcare engineering*, vol. 2019, 2019.

[7] J. Wu, H. Fang, F. Li, H. Fu, F. Lin, J. Li, L. Huang, Q. Yu, S. Song, X. Xu, *et al.*, "Gamma challenge: Glaucoma grading from multi-modality images," *arXiv preprint arXiv:2202.06511*, 2022.

[8] CBO, "Censo oftalmologico 2021," 2021. https://cbo.net.br/2020/admin/docs_upload/034327Censocbo2021.pdf.

[9] L. D. Nguyen, R. Gao, D. Lin, and Z. Lin, "Biomedical image classification based on a feature concatenation and ensemble of deep cnns," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–13, 2019.

[10] J. Ma, B. Lv, Y. Li, P. Fan, X. Zhao, H. Yuan, and Y. Zhang, "Multimodal primary open angle glaucoma early diagnosing program based on clinical process," 2021.

[11] J. Xiong, F. Li, D. Song, G. Tang, J. He, K. Gao, H. Zhang, W. Cheng, Y. Song, F. Lin, *et al.*, "Multimodal machine learning using visual fields and peripapillary circular oct scans in detection of glaucomatous optic neuropathy," *Ophthalmology*, 2021.

[12] H. Fang, F. Shang, H. Fu, F. Li, X. Zhang, and Y. Xu, "Multi-modality images analysis: A baseline for glaucoma grading via deep learning," in *International Workshop on Ophthalmic Medical Image Analysis*, pp. 139–147, Springer, 2021.

[13] Y. Li, M. El Habib Daho, P.-H. Conze, H. Al Hajj, S. Bonnin, H. Ren, N. Manivannan, S. Magazzeni, R. Tadayoni, B. Cochener, *et al.*, "Multimodal information fusion for glaucoma and diabetic retinopathy classification," in *Ophthalmic Medical Image Analysis: 9th International Workshop, OMIA 2022, Held in Conjunction with MICCAI 2022, Singapore, Singapore, September 22, 2022, Proceedings*, pp. 53–62, Springer, 2022.

[14] H. Tian, S. Lu, Y. Sun, and H. Li, "Gc-net: Global and class attention blocks for automated glaucoma classification," in *2022 IEEE 17th Conference on Industrial Electronics and Applications (ICIEA)*, pp. 498–503, 2022.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[16] F. J. F. Batista, T. Diaz-Aleman, J. Sigut, S. Alayon, R. Arnay, and D. Angel-Pereira, "Rim-one dl: A unified retinal image database for assessing glaucoma using deep learning," *Image Analysis & Stereology*, vol. 39, no. 3, pp. 161–167, 2020.

[17] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge: MIT Press, 2016. http://www.deeplearningbook.org.

[18] R. Solovyev, A. A. Kalinin, and T. Gabruseva, "3d convolutional neural networks for stalled brain capillary detection," *Computers in Biology and Medicine*, vol. 141, p. 105089, 2022.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[22] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *CoRR*, vol. abs/1608.06993, 2016.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Los Alamitos, CA, USA), pp. 770–778, IEEE Computer Society, jun 2016.

[24] F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Sequential model-based optimization for general algorithm configuration," in *Learning and Intelligent Optimization: 5th International Conference, LION 5, Rome, Italy, January 17-21, 2011. Selected Papers 5*, pp. 507–523, Springer, 2011.

[25] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.

[26] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[27] J.-C. Mwanza, J. D. Oakley, D. L. Budenz, R. T. Chang, O. J. Knight, and W. J. Feuer, "Macular ganglion cell–inner plexiform layer: Automated detection and thickness reproducibility with spectral domain–optical coherence tomography in glaucoma," *Investigative Ophthalmology and Visual Science*, vol. 52, pp. 8323–8329, 10 2011.

[28] M. McHugh, "Interrater reliability: The kappa statistic," *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, vol. 22, pp. 276–82, 10 2012.

[29] R. Fan, K. Alipour, C. Bowd, M. Christopher, N. Brye, J. A. Proudfoot, M. H. Goldbaum, A. Belghith, C. A. Girkin, M. A. Fazio, *et al.*, "Detecting glaucoma from fundus photographs using deep learning without convolutions: Transformer for improved generalization," *Ophthalmology Science*, vol. 3, no. 1, p. 100233, 2023.

[30] O. C. Devecioglu, J. Malik, T. Ince, S. Kiranyaz, E. Atalay, and M. Gabbouj, "Real-time glaucoma detection from digital fundus images using self-onns," *IEEE Access*, vol. 9, pp. 140031–140041, 2021.

[31] P. Shanmugam, J. Raja, and R. Pitchai, "An automatic recognition of glaucoma in fundus images using deep learning and random forest classifier," *Applied Soft Computing*, vol. 109, p. 107512, 2021.

Marcos Ferreira is a doctoral student at Federal University of Maranhão (UFMA) since 2021. He received his Master's Degree in Computer Science from Federal University of Maranhão (UFMA). His research interests include computer vision, deep learning and medical image processing.

Geraldo Braz Junior receive PhD in Electrical Engineering from Federal University of Maranhão. He is a professor at the Federal University of Maranhão. Has experience in computer vision, machine learning, deep learning and medical image processing.

João Almeida received a DSc. Degree in Electric Engineering from the Federal University of Maranhão (UFMA), Brazil, in 2013. Currently, he is a Professor at the Federal University of Maranhão (UFMA), where he teaches Intelligent Systems, Design, and Analysis of Algorithms and Topics in Image Processing. His research interests include medical image processing, pattern recognition, and machine learning.

Anselmo Paiva received a DSc. Degree in Computing from Pontifical Catholic University of Rio de Janeiro. He is an associate professor at the Federal University of Maranhão (UFMA). His research interests include virtual reality, augmented reality, graphic computing, medical image processing and volume rendering.