A Comparative Analysis of Glaucoma Feature Extraction and Classification Techniques in Fundus Images

Debora F. de Assis, Paulo César Cortez

***Abstract*—Glaucoma is an asymptomatic chronic eye disease that, if not treated in the early stages, can lead to blindness. Therefore, detection in the early stages is essential to preserve the patient's quality of life. Thus, it is crucial to have a non-invasive method capable of detecting this disease through images in the fundus examination. In the literature, datasets are available with fundus images; however, only a few have glaucoma images and labels. Learning from an imbalanced dataset challenges machine learning, which limits supervised learning algorithms. We compared approaches to extract and classify three public datasets with 2390 images: ACRIMA, REFUGE, and RIM-ONE DL. First, we evaluated extracted features non-structural from HOG, LBP, Zernike, and Gabor filters and features obtained from transfer learning. Then, we classified them with Multilayer Perceptron (MLP), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGB). Finally, each classifier was evaluated individually and in a voting classifier (VOT). We extracted and classified features from transfer learning models in the same process. Also, they were classified using traditional machine learning. Due to class imbalance, we undersampled the majority class normal by applying the following methods: random choice, near miss, and cluster centroid. We also evaluated our model using a cross-dataset approach. Therefore, we efficiently identified glaucoma in different fundus images using network VGG19 and a voting classifier. In addition, balancing classes reduced false negatives and improved model quality. Our approach achieved an average F1-score equals to 94.69%, accuracy rate of 94.77%, precision of 96.10%, recall of 93.45%, and specificity of 96.08%.**

***Index Terms*—Glaucoma, Classification, Transfer Learning, Cluster Centroid, RIM-ONE DL.**

I. Introduction

GLAUCOMA is a chronic eye disorder caused by a particular pattern of progressive optic nerve damage, usually related to increased intraocular pressure. This disease can cause irreversible damage to the optic disc that progressively atrophies the visual field, causing blindness in more advanced cases [1]. Glaucoma can be considered a disease group that presents optic neuropathy, caused both by alterations (optical disc aspect) and by functional deficit (visual field alteration) [2].

According to [3], glaucoma can be treated but it has no cure. This disease tends to be asymptomatic in the early stages, with notable symptoms in the advanced stages. The patient vision is maintained when appropriately treated, usually with eye drops that reduce or stabilize intraocular pressure.

Periodic evaluations with an ophthalmologist are the best way to detect this disease. Fundoscopy is commonly recommended to obtain information about the fundus image due to its economy and straightforwardness. In a fundoscopy exam, the specialist visualizes the eye structures, with particular attention to the optic nerve, the retinal vessels, and the retina itself. These regions are essential in identifying glaucoma, as they evaluate the pupillary reflex response and visual acuity. However, this exam is not confirmatory because human vision incorporates subjective factors, limiting the analysis of a specialist [4].

Different devices that capture fundus images can lead to heterogeneity in the acquisition, with blurred images captured or different angles. Therefore, developing methods to diagnose glaucoma that encompasses image recognition and adapts to image qualities from varied cameras is essential for patient treatment.

Several works have used computer vision techniques because they effectively identify many diseases. If applied properly, these techniques can transmit relevant information and support a doctor in providing an accurate diagnosis.

Manifold techniques have been developed and applied in the literature to aid in diagnosing glaucoma. However, some nuances related to different datasets, classifiers, and imbalance classes make the detection less generalized. In this work, we applied undersampling techniques to balance normal and pathological samples to enhance model generalization and reduce false negatives. For glaucoma, false negatives are more harmful to the model than false positives, as it is more important to diagnose the patient with glaucoma so the doctor can perform the necessary tests. In addition, assuming that the test is a false negative, the patient will be diagnosed as healthy and will not be referred to a specialist, which can aggravate the stage of the disease and delay the treatment.

In addition, we used three datasets to consider image variety, including the enhanced current RIM-ONE-DL dataset combined with three previous versions by removing duplicate images.

Features non-structural extraction methods, such as texture and entropy information, are widely applied to describe an image. However, Convolutional Neural Networks (CNN) have stood out for their ease and diversity of application. In this work, we compared these two feature extraction methods,

The authors are with Department of Teleinformatics Engineering (DETI), Federal University of Ceará (UFC), Campus do Pici, Ceará, Brazil e-mails: ({debora.ferreira, cortez}@lesc.ufc.br).

The authors would like to thank the Brazilian National Council for Scientific and Technological Development (CNPq) under Grant No. 313599/2019-0 and Coordination for the Improvement of Higher Education Personnel (CAPES)-Finance Code 001 for the support to this research.

Digital Object Identifier: 10.14209/jcis.2023.6

evaluated traditional Machine Learning (ML) models and transfer learning networks to proceed with the classification.

Our main contribution was a pipeline combining VGG19 and a voting classifier for glaucoma detection, using a dataset balanced with cluster centroid. In addition, we investigated and compared different methods to feature extraction and image classification for glaucoma detection. We provided an extensive analysis of traditional Machine Learning and Deep Learning (DL). We explored balanced classes to improve performance. In addition, we evaluated the following techniques to generalize:

- An analysis of class balancing for model stability, using a random, near miss, and cluster centroid undersampling;
- Non-parametric statistical tests for comparison between models with imbalanced and balanced classes;
- Model generalization with 10-fold cross-validation;
- Cross-dataset analysis to check model generalization with images not seen in the training stage.

We organized the content of this work into five sections. In Section 2, we discussed results related to the literature. Section 3 described the datasets, feature extraction, transfer learning, evaluated methods, and the Mann-Whitney test. In Section 4, we produced the analysis and discussed the experimental design results. Finally, we indicated the conclusions and future work in Section 5.

II. Related Works

Due to the improvement of computer vision techniques, several studies have been developed to detect glaucoma from fundus images. Table I presents a methods summary used in related works and in our best model. Some authors described the fundus image with features non-structural extractors or structural features calculated from the optic disc and cup segmentation. Recently, studies employed deep learning with pre-trained network architectures to extract and classify images in a single process. In addition to these techniques, there are also combinations of different methods to improve the model and increase accuracy.

In the related works, authors that chose to extract features non-structural to describe images used statistical information, texture, and entropy, among others. For example, in work developed by Talaat *et al.* [5], the authors applied features extraction using statistical and texture information. After that, they reduced the amount of data to balance the classes, and performed the classification using SVM.

Parashar *et al.* [6] and Khan *et al.* [7] also applied non-structural extractors to describe the images and classified them using traditional ML methods. Parashar *et al.* [6] proposed a 2-D compact variational mode decomposition (2-DC-VMD). First, they decomposed images into several variational modes (VMs) and extracted the features such as Kapur Entropy (KE), Renyi Entropy (RE), Yager Entropy (YE), Shannon entropy (SE), and energy (En). Then, they used the least squares SVM multiclass classifier (MC-LS-SVM) to detect glaucoma stages (healthy, early, and advanced). On the other hand, [7] applied discrete wavelets transform (DWT) based on the bivariate shrinkage method. They involved a non-Gaussian bivariate probability distribution function to model the statistics of wavelet coefficient images.

Another medical image classification approach is extracting CNNs features and merging them with other information, such as texture or geometric features. Thakur *et al.* [8] combined structural features such as the Gray Level Cooccurrence Matrix (GLCM), Gray Level Run Length Matrix (GLRM), Higher Order Spectra (HOS), the First Order Statistical (FoS), Higher-Order Cumulative (HOC) and Discrete Wavelet Transform (DWT). Furthermore, features non-structural were also merged, such as Cup-to-Disc Ratio (CDR) and Disc Damage Likelihood Scale (DDLS).

Claro *et al.* [9] combined features non-structural (LBP, GLCM, HOG, Tamura, GLRLM, morphology) and seven CNN architectures. The classification was performed with Randon Forest (RF). Another study that combined non-structural information with features from CNNs was Benzebouchi *et al.* [10], in which the authors proposed a multimodal representation based on extracting features from different CNNs with features non-structural from GLCM, Hu Moments, and Central Moments.

Raghavendra *et al.* [11] proposed a convolutional network to extract features and classified them with Linear Discriminant Analysis (LDA). Some authors used CNNS architectures to extract features and classify images in a single process, as [12] did. In this case, the network starts randomization of the weights and performs the extraction step from the beginning with the dataset used. This method differs from the pre-trained models, which can be used to classify new images using weights learned. It is optional to train the network with random initial weights, thus reducing the computational cost of training the model. The authors [13], [14], [15], [16], [17] and [18] used pre-trained models to identify glaucoma.

Norouzifard *et al.* [13] used the VGG19 and ResNetInceptionV2 networks to classify glaucoma. Gómez-Valverde *et al.* [14] explored the application of different CNN architectures to demonstrate the network's performance. Similarly, Batista *et al.* [15], who proposed an improvement in the RIM-ONE dataset, separated the dataset randomly and by the hospital. They performed the extraction and classification with different transfer learning methods in the two proposed separation forms. Juneja *et al.* [16] proposed a CNN for glaucoma classification based on GC-NET.

Elangovan *et al.* [17] used a CNN architecture to extract image information and compared classification traditional methods with the network using Softmax. Another work that also compared different classification methods was from Singh *et al.* [19], in which they proposed a multimodality-based approach for the detection of glaucoma. They classified the features with six traditional ML methods and two deep learning approaches. Finally, Ubaidah *et al.* [18] extracted and classified the images with the MobileNet network to multiclass dataset. However, they used data augmentation in all the dataset, including validation and testing, which leads to optimistic results.

Although many studies presented performances with high accuracy rates in detecting glaucoma, some authors need to improve their evaluation of the results. The difference in the

dataset and results evaluation makes it difficult to compare the methods. Some studies use repetition, k-fold, or separation of training and test without validation with external data. Thus, using only accuracy rates as a comparison does not imply a fair comparison.

To our best knowledge, the most used method to validate the results is k-fold cross-validation, as [5] used 10-fold for the training and validation step. Other works that used 10-fold to validate are [6], [9], [10], [14] and [18]. [8], applied 5-fold to validate the classification of features structural and non-structural. [19] used 5-fold for validation. Another form of validation was with fifty repetitions in [11], twenty repetitions in [12], and five repetitions in [17]. [15] did not use repetition due to the random and specific separation by the hospital. In [13], repetition was not used. In [7] and [16], validation techniques were not presented.

The authors that did not use any validation method could obtain optimistic results. For example, [8] presented classification with a separation of 70% for training and 30% for testing and obtained higher results than with a 5-fold cross-validation.

The studies may use different model validation techniques as repetition or cross-validation. Results with these techniques are more generalized than those with only one run, as the repetition methods are generalized and less biased. However, involving only one run can generate optimistic model results. Therefore, as the majority, we used 10-fold cross-validation to generalize the model.

Tests on new images are necessary to evaluate the models generalization. However, most related works used only the test set of the dataset applied in the training stage. In the literature, few studies evaluated the performance of models with a dataset not included in the training stage. To our best knowledge, only [13] used external validation with the HRF dataset. The results showed lower rates than the studies that separated the data set for training and testing.

Although those works applied different datasets and combinations of them, there must be no replicas of images in the training and test data, which can generate optimistic results since the algorithm has already learned the features of that specific image. [6], [9], and [14] used dataset merge from previous versions of RIM-ONE, which had image duplication. Therefore, it can return favorable results since identical images could be present in the training and testing data. With dataset improvement in [15], we used the RIM-ONE DL dataset, which contains unduplicated images from the three previous versions of the RIM-ONE dataset.

Most public datasets with glaucoma image labels are imbalanced, compromising model training. In our review, only [5] applied undersampling. They used 40 images for each class and 10-fold cross-validation to obtain results, that is, fewer test images.

To our best knowledge, most studies comparing evaluated models only used descriptive information. Few authors applied statistical tests. For example, [6] and [7] used tests to compare their models. [7] applied tests to define the best features.

Thus, some related works presented limitations in generalization, such as the lack of tests using the external dataset, insufficient validation with few test images, or not using repetition to generalize the results. In addition, it was necessary to improve the model without significantly reducing the data set, especially in works that evaluated less than ten images in the test. Other points to be enhanced are datasets with replicated images, such as studies that used the combination of versions 1, 2, or 3 of RIM-ONE and works that used data augmentation in the training, validation, and test stages. Furthermore, the inferential analysis must be accomplished for comparative works to prove the significant difference in models.

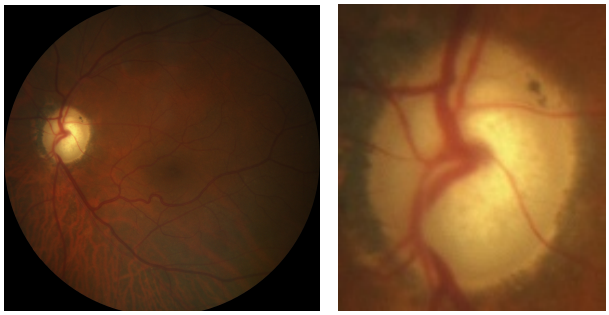
To overcome these limitations, we explored an efficient methodology to identify glaucoma in fundus images using existing techniques. Therefore, we performed an extensive analysis comparing features non-structural extractions, five transfer learning algorithms, four classifiers, and six evaluation metrics, including a confusion matrix and statistical tests to compare models.

III. Materials and Methods

We presented an experimental design model in the workflow illustrated in Fig. 2. We compared three models: extracted features non-structural from images and machine learning traditional models classification; features extracted from transfer learning networks and classification with traditional machine learning; and transfer learning networks to extracted features and classification from images. Moreover, we applied class balancing and compared the results. In addition, we evaluated the best model based on statistical tests. Each method applied was described in the following sections.

We used fundus images from three public datasets: ACRIMA, REFUGE, and RIM-ONE DL. Since the REFUGE dataset images were obtained from the entire retina, we cropped the optic disc region based on the publicly available area in the dataset. The total images are 2390: 1702 normal and 688 glaucoma classes. Fig. 1 (a) illustrates the original image, and 1 (b) shows the cropped image.

Fig. 1: REFUGE example image



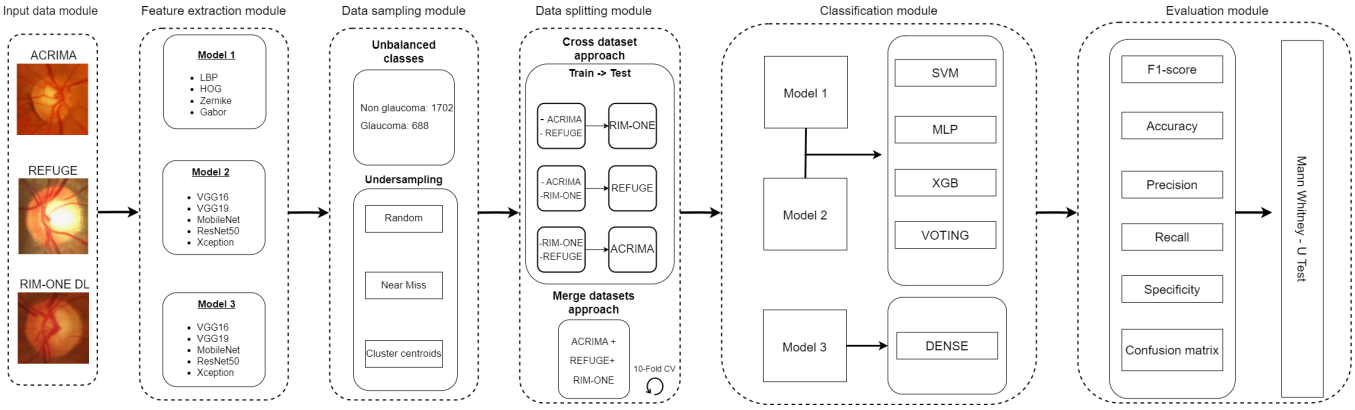
(a) Original (b) Cropped

We performed a comparison of different extraction methods and classifications to identify glaucoma. The first method is a traditional/baseline model, in which descriptors obtain feature vectors. Next, we resized the images to 256 $\times$ 256 pixels and converted them to gray scale. After that, we extracted the features of the images from Local Binary Patterns (LBP),

TABLE I: A summary of related works

Work	Dataset	Pre-process	Features	Classify	Validate	Class Balanced	Statistical Test	Results
Our method	**ACRIMA, REFUGE, RIM-ONE DL**	**Resize (224 × 224)**	**VGG19**	**Voting**	**10-fold**	**Random, NearMiss, Centroid Cluster**	**Mann- -Whitney**	**F1: 94.69%, Acc: 94.77%, Prec: 96.10%, Rec: 93.45%, Spe: 96.08%**
[5]	REFUGE	Cropped, Resize (224 × 224), RGB, HSV, CIELAB, Median filter	Statistics informations, GLCM, RLM	SVM	10-fold	Random	None	Acc: 92.50%, Rec: 95.00%, Spe: 90.00%
[6]	RIM-ONE v1, RIM-ONE v2	Resize (240 × 240), Green color, CLAHE	2-DC-VMD, KE, RE, SE, YE, FD, Energy	MC-LS-SVM	10-fold	None	F test, Kruskall- -Wallis	Acc: 96.23%, Rec: 98.00%, Spec: 94.50%, AUC: 96.25%
[7]	RIM-ONE v2	2D-WTD RGB plane	Contrast, Homogeneity, Energy, Entropy, RMS, Variance	LV-SVM	None	None	T test, Wilcoxon	Acc: 91.22%, Rec: 85.51%, Spec: 98.50%
[8]	DRISHTI-GS, RIM-ONE v3	Cropped, Resize (512 × 512)	DWT, GLRM, GLCM, HOS, FoS, DDLS, CDR	k-NN, SVM, NN, RF, NB	5-fold	None	None	Acc: 93.20%, Prec: 92.00%, Rec:91.00%, Spe: 92.00%
[9]	DRISHTI-GS, RIM-ONE (v1, v2 e v3), JSIEC, HRF, ACRIMA	Cropped disc	LBP, GLCM, HOG, Tamura, GLRLM, CNN	RF e CNN	10-fold	None	None	Acc: 92.78%, Prec: 92.80%, AUC: 96.60%, Kappa: 81.85%
[10]	RIM-ONE V2	Otsu, Resize (100 × 100)	CNN, GLCM, Hu moments, Central Moments	BWWV CNN SVM	10-fold	None	None	Acc: 99.78%, Rec: 99.50%, Esp: 100%
[11]	Kasturba Medical	Resize (64 × 64)	CNN	LDA	50 repts	None	None	Acc: 98.13%, Rec: 98.00%, Esp: 98.30%
[12]	DRISHTI-GS, ORIGA, HRF	HE, CLAHE, Resize (256 × 300)	CNN	CNN	20 Repts	None	None	F1: 95.70%, Acc: 93.50%, Rec: 97.70%, Spe: 92.60%, Prec: 93.80%, AUC: 95.10%
[13]	UCLA, HRF	Resize (299 × 299)	VGG19, ResNetInceptionV2	VGG19, ResNetInceptionV2	None	None	None	Acc: 95.00%, Rec: 90.10%, Spec: 100.00%
[14]	ESPERANZA, RIM-ONE v1, RIM-ONE v2, RIM-ONE v3, DRISHTI-GS	Data augment, Resize (256 × 256 and 224 × 224)	VGG19, GoogLeNet, ResNet50, DENet	VGG19, GoogLeNet, ResNet50, DENet	10-fold	None	None	Prec: 88.05%, Rec: 87.01%, Spec: 89.01%, AUC: 0.94
[15]	RIM-ONE DL	Resize (224 × 224)	VGG16, DenseNet, MobileNet, VGG19, Xception, ResNet50, MobileNetV2, InceptionResNetV2, InceptionV3, NasNetMobile	VGG16,DenseNet, MobileNet, VGG19, Xception, ResNet50, MobileNetV2, InceptionResNetV2, InceptionV3, NasNetMobile	None	None	None	Acc: 93.15%, Rec: 100.00%, AUC: 98.67%
[16]	RIM-ONE V1, DRISHTI-GS	Cropped, Denoising, Data augment	GC-NET, VGG16, InceptionV3, Xception, ResNet50, DenseNet121	GC-NET, VGG16, InceptionV3, Xception, ResNet50, DenseNet121	None	None	None	Acc: 97.51%, Rec: 98.78%, Spe: 96.20%
[17]	DRISHTI-GS	Resize (224 × 224), Data augmentat, Contrast	DenseNet201	SVM k-NN, NB, Softmax	5 reps	None	None	Acc: 96.48%, Prec: 95.82%, Rec:98.88%, Spe:92.10%, F1: 97.28%
[19]	REFUGE, ORIGA, ACRIMA	RGB, Tozero, Gray scaler	CDR, GLCM, GLRM, SRE, GLU, RPC, DCGAN, VGGCapsNet	RF, k-NN, NB, SVM, XGboost, DeepNet, DCGAN, VGG-CapsNet	5-fold	None	None	Acc: 95.56%, Prec: 95.00%, Rec: 93.00%, Spe: 97.00%, kappa: 94.00%
[18]	RIM-ONE v1	Resize (224 × 224), Data augmentat	MobileNet	MobileNet	10-folds	None	None	Acc: 99.00%, Rec:99.00%, F1: 99.00%

Fig. 2: Experiments Workflow



Histogram Oriented Gradient (HOG) and Zernike moments. Statistical information was obtained after applying the Gabor filter. As a result, we extracted 135 features from the images: a vector of 32 features obtained from the HOG descriptor, 66 from the LBP, 25 from the Zernike moments, and 12 statistical information from the image filtered with Gabor.

We chose these extractors after exhaustive tests with several descriptors applied in the literature, as in studies from [9] and [10]. In addition, we adjusted their parameters for proper performance in identifying glaucoma in fundus images.

In the second method, we applied CNNs as feature extractors. We used the transfer learning architectures: VGG16, VGG19, ResNet50, MobileNet, and Xception. As preprocessing, we resized the images to the standard size of each network. We resized the VGG16, VGG19, MobileNet, and Resnet50 networks to 224 $\times$ 224, and Xception networks to 299 $\times$ 299. We removed the last fully connected layers and sent the resulting vector to traditional ML classifiers.

The features of each model were classified using the Multilayer Perceptron (MLP), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGB). To obtain a better predictive performance, instead of using only the classifiers individually, we applied a voting classifier that consists of a classification committee using SVM, MLP, and XGB. This model aggregated the findings of each classifier to predict the output class based on the majority votes.

We chose these classifiers for presenting satisfactory results in the studies from [5], [8] and [19]. Also, in comparison with other classifiers by grid search techniques and exhaustive comparison tests, the chosen classifiers got the best results.

In the third method, we used the VGG16, VGG19, InceptionV3, ResNet50, MobileNet, and Xception architectures to extract and classify the images. We removed the fully connected layer and added dense layers to train the features. We used the weights of the architectures trained with ImageNet. We employed the sigmoid activation function, the cross-entropy loss function, and Adam optimizer.

We applied 10-fold cross-validation, ensuring a generalization and avoiding favorable results. The final results referred to the average of each metric: accuracy, precision, recall, specificity, and F1-score. Also, we evaluated the average amount of true positives, true negatives, false positives, and false negatives.

Public datasets fundus images vary in size, quality, and also depends on the camera that obtains the image. In addition, the number of images in the class is imbalanced, with many images of the non-glaucoma classes and few glaucoma images. To solve this problem and improve the model, we applied class balancing, which reduced the number of observations to match the smallest category. In addition, we used different techniques for undersampling: random, near miss, and cluster centroid.

Commonly used in the literature to select observations, random separation does not present a technique for choosing images. Instead, the Near Miss technique uses the borderline idea among the category to choose the most representative samples. Finally, the Cluster centroid undersampling uses the selection of samples based on each class centroid, seeking a grouping among the samples selected [20].

To verify the difference in the model performance when adding the balancing approach, we applied the Mann-Whitney test to compare the results. In addition, we performed the analysis with the cross-dataset to verify the model behavior when evaluating a dataset not seen in the training stage.

Our analyses were performed on an Intel(R) Core (TM) CPU i7-1165G7 2.80GHz, with 16GB of RAM and a 64-bit operating system. For computational analysis, we used Python and R as programming languages. The programming code implementations are public on https://github.com/DeboraFA/Glaucoma_binary_classification_undersampling.

A. *Datasets*

ACRIMA dataset was created from a project (TIN2013-46751-R) founded by the Ministerio de Economía y Competitividad of Spain, which develops automatic algorithms for assessing retinal diseases. The ACRIMA dataset comprises 705 fundus images, 396 images from the glaucomatous class, and 309 from the non-glaucoma label. Most of the fundus images in this dataset were taken from the previously dilated eyes and centered on the optic disc [21].

According to [22], the Retinal Fundus Glaucoma Challenge (REFUGE) was made public in 2018 in partnership with the MICCAI Workshop on Ophthalmological Image Analysis

(OMIA). This dataset provides segmented fundus images and clinical glaucoma labels: 1080 images from the non-glaucoma and 120 from the glaucoma classes.

The Open Retinal Image dataset for Optic Nerve Evaluation Deep Learning (RIM-ONE DL) results from the combined of the three previous versions of RIM-ONE (v1, v2 e v3). This new version eliminated the duplicate images contained in v1 and v2. In addition, they excluded the images of the same patient in v2 and v3, leaving only the left image of dataset version v3. The images in the three versions of RIM-ONE included healthy and glaucomatous eyes from Spanish hospitals: Hospital Universitario de Canarias (HUC) in Tenerife, Hospital Universitario Miguel Servet (HUMS) in Zaragoza, and Hospital Clınico Universitario San Carlos (HCSC), in Madrid. The final set consisted of 485 images, 313 healthy and 172 glaucoma fundus images [15].

B. Features descriptors

Obtaining image features is an exhaustive and non-intuitive process, requiring prior knowledge of images and descriptors. Furthermore, verifying the correlation among these features is essential, as adding some variables can cause multicollinearity problems.

Therefore, selecting suitable extractors with relevant features is crucial for distinguishing image classes. We defined feature extractors: LBP, HOG, Zernike moments, and statistical information of the image after the Gabor filter.

1) Histogram of Oriented Gradients (HOG): is a global feature descriptor that describes an image with a locally Histogram Oriented Gradient. The HOG is calculated by three step-sequence: gradient computation, orientation, and histogram generation. These histograms represent occurrences of specific gradient orientation in a local part of images [23, 24].

We divided the image into cells of 128 $\times$ 128 pixels and blocks of 1 $\times$ 1 pixels. To calculate HOG, we used eight orientations per histogram for each cell. Subsequently, we calculated result normalization using the square root method in 8 orientations, introducing an invariance to lighting, shading, contrast, and edges. Finally, the HOG descriptors from all blocks accumulated from a dense superimposed grid of blocks covering a detection window, into a combined feature vector. The total features descriptor is equal to 32.

2) Local Binary Pattern (LBP): described by Ojala in 1996, is a particular case of the spectral texture model, defined by Wang [25]. LBP assumes that texture information is divided into textural units.

After applying LBP, the resulting features are equal to $P+2$, where P is the number of gray levels in the image. So, to this work, the value of P is equal to 64. Furthermore, the added LBP input tabulates all patterns that are not uniform, adding an extra rotation level and grayscale invariance. Thus, we obtained the feature vector with a dimension of 66, representing the input images textures.

3) Zernike Moments: map an image onto a set of complex Zernike polynomials. As these polynomials are orthogonal to each other, Zernike moments can represent the image properties without redundancy or overlapping information between moments. Their magnitudes are independent of the object rotation angle. The Zernike moments' calculation of an image consists of three steps: radial polynomials analysis, calculation of base Zernike functions, and measuring Zernike moments by projecting the image onto base functions [26].

We used a radius equal to 128, which referred to the input image radius of dimension 256 $\times$ 256. Therefore, the quantity of features for this descriptor was equal to 25.

4) Gabor Filter: was developed in 1964 by Dennis Gabor and it is a linear filter used for texture analysis. This filter analyzes image frequency content in specific region directions. This filters are appropriate for texture representation and discrimination [27].

First, we applied Gabor filters to the images. After that, we calculated statistical information in filtered images: mean, asymmetry, and kurtosis. Then, three pieces of information are calculated for each resulting image, totaling 15 features for this descriptor.

C. Transfer Learning

Convolutional Neural Networks (CNNs) were introduced by Yann Lecun *et al.* in the 1990s. This designed network receives input data matrices, a color image composed of three 2D matrices containing pixel intensities channels color. The main convolutional network properties are local connections, shared weights, pooling, and the use of multiple layers [28].

Transfer learning networks have a layered architecture that uses a pre-trained network. They can be used, without their final layer, as an image feature extractor. We applied these architectures in our study: VGG16, VGG19, MobileNet, ResNet50, and Xception.

1) VGG: The VGG model concept improves overall network performance by increasing the layer depth. Its strategy is transforming the convolution kernel layer into small multivolume laminated kernels, reducing the model parameters number, and making the network more discriminative [29].

In the VGG16 architecture, the number 16 refers to the 16 layers with weights. The difference between VGG16 and VGG19 architectures comes down to three additional layers existing in VGG19, with an extra convolution layer in the fourth, one in the fifth, and one in the sixth block.

2) MobileNet: The Mobilenet kernel layer presents one of the essential network properties: the separable convolutions in depth. The principal depth convolutions property is to split the standard convolution into two nucleus, a depth convolution and a point convolution 1 $\times$ 1. Deep-separable convolution uses 8-9 times less computational cost than standard convolution [30].

3) ResNet50: Researchers at Microsoft Research developed residual neural networks (ResNet). The central aspect of differentiating from other CNN is the residual block concept. That uses shortcuts between layers, adding the layer's input initial values to the ReLU output function $y = F(x, W_i)) + x$, where the function $F(x, W_i)$ represents the residual map learned by the block and x the initial image [31].

The layers in a traditional network are learning the actual $(H(x))$ output while the residual network layers are learning

the residual ($R(x)$). These shortcut connections skip over layers and are arranged in residual blocks. These blocks have three convolution layers with k filters 1 $\times$ 1, 3 $\times$ 3, and 1 $\times$1, respectively, for each layer. Given an input x, the residual mapping $F(x)$ is denoted by $F(x)+x$. This mapping adds the input of a residual block (x identity) to the output resulting from that same block [32].

4) Xception: François Chollet proposed in 2017 the work "Xception: Deep Learning with Depthwise Separable Convolutions". This architecture is a linear stack of separable deep convolutional layers with residual connections. The architecture is based on separable layers in depth. That network is a more robust version of the Inception. The Xception architecture has 36 convolutional layers forming the basis for extracting features from the network. The 36 convolutional layers are structured in 14 modules, all with residual linear connections around them, except for the first and last modules [33].

D. Evaluating Metrics

To evaluate the model quality and compare the performance of the classifiers, we used the metrics: F1-score, accuracy, precision, recall, and specificity.

The False Positive (FP) means that the test result indicates the presence of the disease when the individual is healthy. On the other hand, the True Positive (TP) suggests that the test correctly predicted those belonging to a condition. The True Negative (TN) indicates that the test correctly predicted healthy owners. Another indicator, the False Negative (FN) test result, shows that the individual is healthy when he has the disease.

The accuracy (A) was used to assess the proportion of correct predictions, where $N = TP + TN + FN + FP$. Precision (P) indicates the measure of patients that the model correctly identifies as having a disease among all patients who do have it. The performed calculation is below,

$$A = \frac{TP + TN}{N}, \quad P = \frac{TP}{TP + FP}$$

The recall (R) is the model capacity to correctly predict the cases in which the individual has the disease. Specificity (S) is the model ability to predict healthy individuals correctly. Moreover, the F1-score (F1) is the weighted harmonic mean of precision and recall, reaching its best value of 1.

$$R = \frac{TP}{TP + FN}, \quad S = \frac{TN}{TN + FP}, \quad F1 = 2\left(\frac{P * R}{P + R}\right)$$

E. Mann-Whitney Test

The Mann-Whitney U Test null hypothesis (H_0) indicates that the two independent groups are homogeneous and have the same distribution. The two variables corresponding to the two groups, represented by two continuous cumulative distributions, are stochastically equal. The alternative hypothesis (H_1) establishes that the first group data distribution differs from the second group data distribution [34].

According to [34], the Mann-Whitney U test initially calculated a W statistic for each group. Then, distribution was assigned to each one of the two sample values to build ranking:

$$W_x = n_x n_y + \frac{n_x(n_x + 1)}{2} - R_x$$

$$W_y = n_x n_y + \frac{n_y(n_y + 1)}{2} - R_y,$$

where n_x and n_y are the sizes of each sample, R_x and R_y are the sum of the rows of the observations of samples x and y, respectively. The W statistic is defined as the minimum of U_x and U_y. The approximation of the z normal, when there are large enough samples, is given by the expression $z = \frac{W-\mu}{\sigma}$ where μ and σ are the means and standard deviation of W.

We chose a significance level equals to 0.05, which is the value to reject the null hypothesis. We supposed the W statistic p-value results in a p-value equals to or greater than 0.05. In that case, there was no evidence of the null hypothesis rejection, that is, there was no difference between the medians of the distributions. On the other hand, if the value is less than 0.05, there is evidence that the classifiers medians are different.

IV. Results

Table II contains the method 1 and 2 results for each classifier and the method 3 results with the classification adding the dense layer. We compared the accuracy, F1-score, precision, recall, specificity, true positive, true negative, false positive, and false negative. In addition, we presented the average 10-fold for each metric $\pm$ standard deviation. The separation dataset resulted in 2151 observations for training and 239 tests to each fold.

The F1-score metric is the harmonic mean between precision and recall. This metric is suitable for evaluating the overall quality index of the model in an imbalanced/ disproportionate class. Therefore, we used this metric to evaluate the model and verify the division of classes.

We used accuracy to measure overall model performance. A high rate indicated an adequate average between classes. While non-glaucoma/normal classes had high accuracy rates, the glaucoma class had more false negatives. The results for precision consider false positives more harmful than false negatives, so the higher the precision of the model, the lower the false positive rate.

The high specificity indicated that the model could avoid false positives, which makes it more probable to classify individuals as non-glaucomatous. Considering the recall rate, the model correctly identified individuals with glaucoma. Although this rate is considered adequate, false negatives are higher than false positives, indicating that the model is more susceptible to errors in glaucoma classification. In addition, we also evaluated the average number of observations from TP, TN, FP, and FN.

The traditional model rates remained close among the classifiers. In general, voting classifiers achieve the best results. The traditional model showed higher specificity rates, while recall rates were lower than specificity. It was also possible to verify that the average of false negative errors was more

TABLE II: Average 10-fold cross-validation results using an imbalanced dataset combination.

Classify	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	TP	TN	FP	FN
Traditional									
SVM	84.24 ± 2.06	91.05 ± 1.19	85.43 ± 2.64	83.14 ± 2.46	94.24 ± 1.17	57	160	10	12
XBG	80.52 ± 1.81	89.12 ± 1.11	83.30 ± 3.31	78.05 ± 2.77	93.59 ± 1.59	54	159	11	15
MLP	83.85 ± 2.26	90.71 ± 1.43	84.24 ± 3.79	83.58 ± 2.41	93.59 ± 1.83	58	159	11	11
VOT	**85.27 ± 1.52**	**91.67 ± 0.87**	**86.95 ± 2.25**	**83.72 ± 2.49**	**94.89 ± 1.06**	58	162	9	11
VGG16									
SVM	86.75 ± 2.28	92.51 ± 1.23	88.32 ± 2.35	85.32 ± 3.40	95.42 ± 1.04	59	162	8	10
XBG	82.52 ± 2.81	90.54 ± 1.49	88.17 ± 3.01	77.62 ± 3.45	95.77 ± 1.17	54	163	7	15
MLP	84.95 ± 2.13	91.42 ± 1.31	86.11 ± 3.92	84.01 ± 3.32	94.42 ± 1.90	58	161	9	11
VOT	**86.39 ± 3.26**	**92.47 ± 1.54**	**89.48 ± 1.92**	**83.71 ± 5.62**	**96.00 ± 0.87**	58	163	7	11
Dense	84.89 ± 3.38	91.38 ± 1.74	85.50 ± 3.96	84.71 ± 6.76	94.06 ± 2.09	58	160	10	11
VGG19									
SVM	85.24 ± 2.16	91.72 ± 1.14	87.41 ± 2.36	83.28 ± 3.44	95.12 ± 1.09	57	162	8	12
XBG	81.49 ± 2.79	89.92 ± 1.60	86.71 ± 4.59	77.04 ± 3.27	95.12 ± 1.93	53	162	8	16
MLP	82.59 ± 3.03	90.21 ± 1.72	84.81 ± 4.05	80.67 ± 4.16	94.06 ± 2.05	56	160	10	13
VOT	**86.32 ± 2.26**	**92.30 ± 1.27**	**88.40 ± 3.08**	**84.45 ± 3.25**	**95.48 ± 1.34**	58	162	8	11
Dense	83.41 ± 2.25	90.63 ± 1.58	85.82 ± 6.18	81.54 ± 3.76	94.30 ± 2.93	56	160	10	13
MobileNet									
SVM	86.38 ± 3.18	92.38 ± 1.71	88.94 ± 3.81	84.15 ± 4.63	95.71 ± 1.64	58	163	7	11
XBG	82.13 ± 2.45	90.25 ± 1.18	86.77 ± 2.17	78.04 ± 3.53	95.18 ± 0.86	54	162	8	15
MLP	85.91 ± 3.45	92.13 ± 1.85	88.44 ± 2.79	83.57 ± 4.41	95.59 ± 1.06	58	163	7	11
VOT	**87.21 ± 2.29**	**92.93 ± 1.18**	**90.78 ± 2.19**	**84.01 ± 3.57**	**96.53 ± 0.93**	58	164	6	11
Dense	86.01 ± 2.49	92.18 ± 1.46	88.97 ± 4.70	83.42 ± 3.33	95.71 ± 1.94	58	163	7	11
ResNet50									
SVM	82.85 ± 2.16	90.67 ± 1.17	88.02 ± 2.90	78.33 ± 2.87	95.65 ± 1.21	54	163	7	15
XBG	81.68 ± 3.20	90.04 ± 1.53	86.54 ± 2.29	77.47 ± 5.03	95.12 ± 0.87	53	162	8	16
MLP	83.54 ± 2.29	90.71 ± 1.12	85.22 ± 2.79	82.12 ± 4.45	94.18 ± 1.38	57	160	10	12
VOT	**85.01 ± 2.54**	**91.67 ± 1.30**	**88.08 ± 2.35**	**82.27 ± 4.28**	**95.48 ± 1.02**	57	162	8	12
Dense	83.63 ± 2.51	90.79 ± 1.26	85.51 ± 5.17	82.29 ± 5.60	94.33 ± 2.11	57	160	10	12
Xception									
SVM	82.28 ± 4.82	90.25 ± 2.30	85.79 ± 3.89	79.37 ± 7.22	94.65 ± 1.65	55	161	9	14
XBG	77.10 ± 2.91	87.49 ± 1.68	81.77 ± 4.66	73.11 ± 3.55	93.30 ± 2.04	50	159	11	19
MLP	81.91 ± 3.09	89.75 ± 1.63	83.20 ± 3.14	80.82 ± 4.77	93.36 ± 1.49	56	159	11	13
VOT	83.11 ± 2.51	90.75 ± 1.22	87.53 ± 3.14	79.35 ± 4.52	95.36 ± 1.38	55	162	8	14
Dense	**83.99 ± 2.11**	**91.09 ± 1.17**	**87.11 ± 3.65**	**81.25 ± 3.58**	**95.07 ± 1.59**	56	162	8	13

significant than the average of false positive errors. For this pathology, the false negative is more harmful, in which the patient is diagnosed as healthy and, in fact, has glaucoma, which can delay treatment due to not being referred to a specialist, aggravating the disease.

The VGG16 and VGG19 models presented similar results and they were close to the results from the traditional model considering the voting classifiers. However, the specificity rates are higher than the recall rates. The dense layer classification does not outperform the classifiers average rates. For VGG16 and VGG19, the accuracy and recall rates increased compared to the traditional model, and the mean number of false negatives remained lower than the false positives'. Thus, the VGG16 and VGG19 models using the voting classifier proved suitable for evaluating optic disc images and distinguishing glaucoma and non-glaucoma class.

The MobileNet model demonstrated satisfactory results. The results using traditional ML classifiers were better than adding a dense classification layer. Therefore, the MobileNet network presented the best rates with voting classification, with high rates above 85%.

The ResNet50 and Xceptions models' results were satisfactory. However, they are lower than other CNN architectures and the traditional model. The voting classifier obtained the best results for ResNet50, and the dense layer using the sigmoid classifier was best for the Xception network.

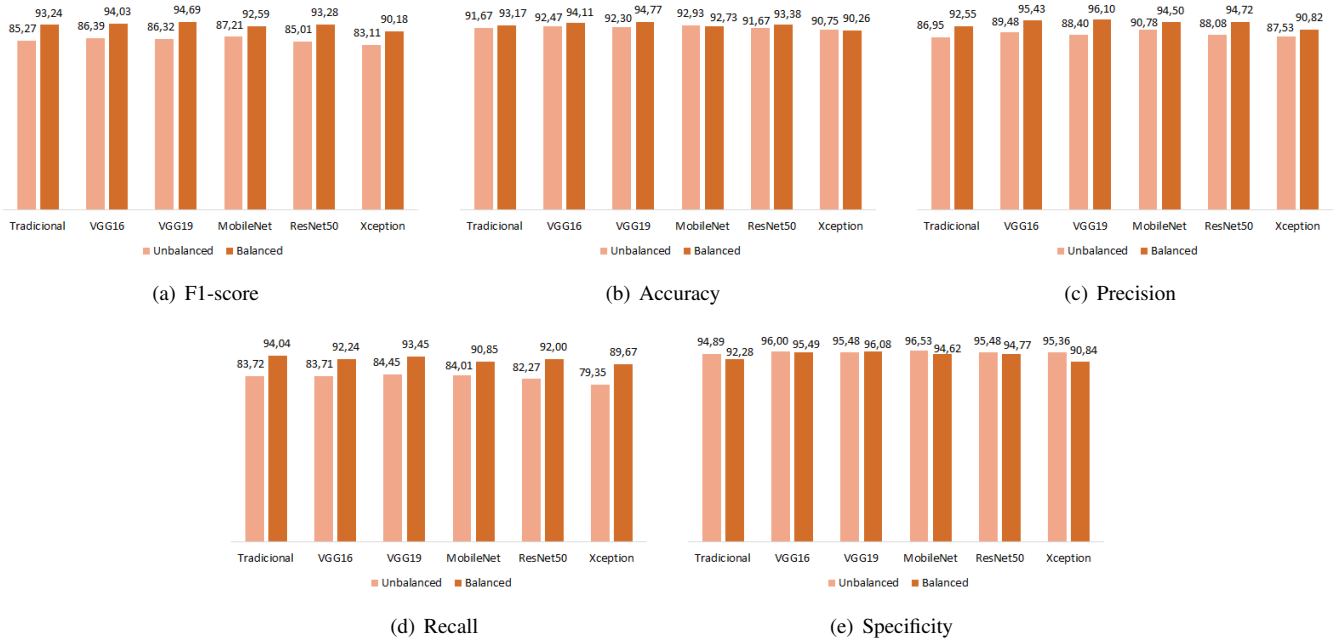
The analyzed models presented satisfactory results using different classifiers. The results presented average above 80% for voting and dense classifier, and the variation in each fold is low for the models. The standard deviation reinforces this observation.

The SVM classifier did perform well in the evaluated models, especially in VGG16. The XGB was a good classifier for distinguishing glaucoma and non-glaucoma classes and presented satisfactory results in all models, with average accuracy rates above 87% and precision above 81%. The MLP classifier performed excellently, with high accuracy rates above 90% in all models.

The voting classifier proved to be an excellent alternative to enhance the model, improving the classification rate in all evaluated models. In addition, it presented better rates than the individual classifier. Although the dense classification layers showed reasonable rates, the voting classifier had better results.

The classification of the evaluated models had average recall

Fig. 3: Compare imbalanced and balanced results using voting classifier.



(a) F1-score (b) Accuracy (c) Precision

(d) Recall (e) Specificity

rates lower than specificity, indicating that the algorithm has difficulty identifying the image with glaucoma. In addition, the models have many false negatives, in which the individual is classified as non-glaucoma when he has the disease. So, the model needed to learn adequate features to distinguish the class.

The combined datasets have imbalanced classes. To solve this problem and improve glaucoma classification, we used balanced classes. We expect the false negative rates to reduce and consequently increase the model rates, especially the F1-score that evaluates the model quality. We reduced non-glaucoma class to an equal amount of glaucoma observation. We compared three methods for undersampling: Random, Near Miss, and Cluster Centroid.

We organized three datasets to form a single dataset without considering the training and test directory. We used all glaucoma images and selected non-glaucoma images to the balanced dataset, a total of 1376 images. The best average rates were achieved using cluster centroid.

Fig. 3 presents imbalanced and balanced data results with cluster centroid for the voting classifier. Graphs are in the interval of 0 to 100. The model improved overall rates, mainly by reducing false negatives. The rates were higher with balanced data, mainly for recall and F1-score. On the other hand, specificity was the metric that showed a reduction in rates with balanced data. Descriptively, the class balancing showed improvement for the evaluated models. More details can be observed in GitHub.

Classification using traditional ML increased the average rates of all models, mainly for the F1-score, precision, and recall metrics. Accuracy increased in almost all models for the voting classifier, especially with VGG19, which achieved the highest rates. The voting classifier showed higher rates than the sigmoid classification using a dense layer.

Descriptively, we obtained the best results with balanced models. Nonetheless, we require inferential evidence to substantiate this assertion. Therefore, we performed the Mann-Whitney test to compare the best imbalanced and balanced models, verifying if they differed in data balancing and imbalanced.

The models presented better results with the voting classifier. Therefore, the comparison was performed with the voting classifier for imbalanced and balanced models.

Considering the F1-score metric, all models showed a significant difference between the imbalanced and balanced results, they are the traditional model (W = 0, p-value $<$ 0.01), VGG16 (W = 0, p-value $<$ 0.01), VGG19 (W = 98, p-value $<$ 0.01) MobileNet (W = 5, p-value $<$ 0.01), ResNet50 (W = 0, p-value $<$ 0.01), and Xception (W = 2, p-value $<$ 0.01).

For accuracy metric, the traditional model (W = 30, p-value = 0.13), VGG16 (W = 26, p-value = 0.07), MobileNet (W = 46, p-value = 0.76), and Xception (W = 65, p-value = 0.27) did not present a significant difference between their imbalanced and balanced accuracy results. On the other hand, the models VGG19 (W = 87, p-value = 0.01) and ResNet50 (W = 21, p-value = 0.03) showed a p-value less than 0.01 indicating that there is evidence that balanced and imbalanced results are different.

For precision metric, all models showed a significant difference between imbalanced and balanced, except Xception (W = 26, p-value = 0.08). The models were: the traditional model (W = 4, p-value $<$ 0.01), VGG16 (W = 4.5, p-value $<$ 0.01), VGG19 (W = 92, p-value $<$ 0.01), MobileNet (W = 12.5, p-value $<$ 0.01) and ResNet50 (W = 3, p-value $<$ 0.01).

As for the recall metric, all models showed significant differences for the imbalanced and balanced results: traditional (W = 1, p-value $<$ 0.01), VGG16 (W = 3.5, p-value $<$ 0.01),

Fig. 4: False Positive.

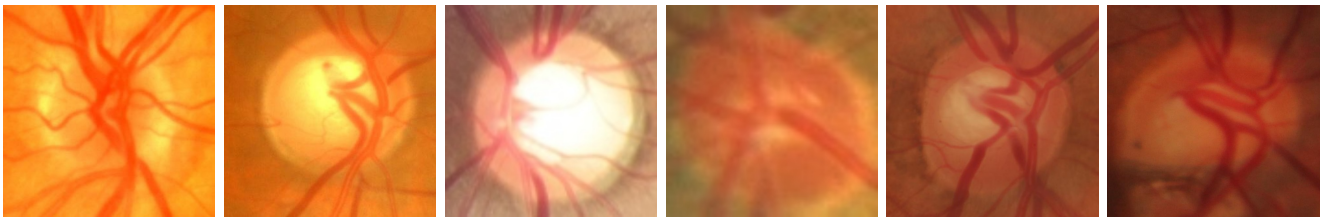
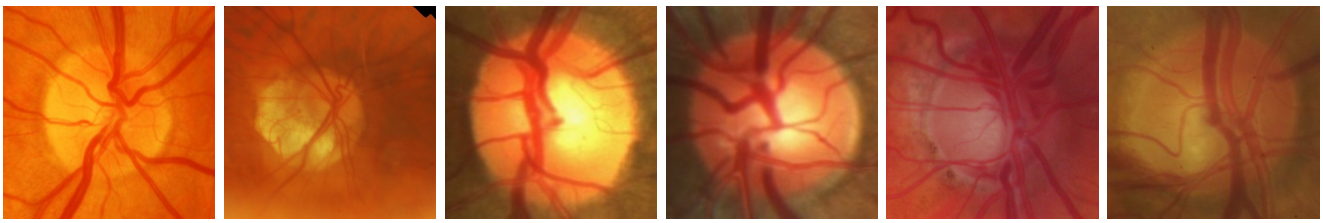


Fig. 5: False Negative.



VGG19 (W = 95.5, p-value $<$ 0.01), MobileNet (W = 8, p-value $<$ 0.01), ResNet50 (W = 5, p-value $<$ 0.01), and Xception (W = 3, p-value $<$ 0.01).

Finally, the specificity metric, the traditional model (W = 69, p-value = 0.15), VGG16 (W = 68, p-value =0.17), VGG19 (W = 66, p-value = 0.22), MobileNet (W = 75.5, p-value=0.05) and ResNet50 (W = 62, p-value = 0.36) showed no significant difference. Only the Xception model (W = 85, p-value = 0.01) showed a significant difference between the imbalanced and balanced data results.

Considering the F1-score and recall metrics, all models showed a significant difference between balanced and unbalanced results. As for accuracy, the models that showed a significant difference were VGG19 and ResNet50. For precision, the models showed significant differences, except with the Xception model. Finally, only Xception showed a significant difference for the specificity metric.

Thus, we considered VGG19 the best model, with data balancing using cluster centroid, for presenting the highest rates, and having a difference in all metrics between the results with balanced data, except specificity.

The images have different qualities and contrasts, which makes it difficult to distinguish and generalize the model to identify the pathology. Fig. 4 and Fig. 5 show examples of incorrectly classified images. False positives indicated that the image was classified as glaucoma when the image belonged to the non-glaucoma class. False negatives, on the other hand, suggest that the image was classified as non-glaucomatous when it actually belongs to the glaucoma class.

Fig. 4 shows the false positives. In the first image, it is possible to visualize the blood vessels and the cup disc, the lighter region in the optic disc center. However, if the image is very bright it is hard to find the boundary between the optic disc and the cup. In the second, third, and fifth images, it is observed a large cup disc and the entry of blood vessels are further away from each other, which could be easily confused with the glaucoma fundus image. The fourth and last images are blurred, which makes it difficult to identify parameters that distinguish class.

Fig. 5 exemplifies the false negatives. The first image cannot distinguish the optic disc from the cup. The second, fifth, and sixth images are blurred, and it is not easy to distinguish their features. The third and fourth images show a small cup disc, which could be mistaken for a healthy optic disc.

We applied the cross-dataset method to evaluate performance with new images. We performed three analyzes; in the first one, we used a RIM-ONE DL and ACRIMA junction for training and tested with REFUGE. In the second one, we used REFUGE and RIM-ONE DL for training and tested with ACRIMA. Finally, we trained with REFUGE and ACRIMA and tested with RIM-ONE DL.

Table III presents the cross-dataset results. The first column indicates datasets used in training, the second indicates the test dataset, and the third indicates balanced classes. The remaining columns indicate the metrics obtained.

To introduce a dataset not previously seen by the model, we observed that the rates significantly decreased. In addition, images from the REFUGE dataset showed more similarities between classes, which made it difficult to identify glaucoma and non-glaucoma types. In contrast, images from the ACRIMA dataset showed a more significant distinction between categories and better rates of accuracy.

The traditional model results for the REFUGE and ACRIMA test datasets are better when using training data balancing. The F1-score rate for ACRIMA was 46.99% and 74.53% for the imbalanced and balanced data, respectively. For the REFUGE dataset, F1-score was 30.47% and 37.01%. For the RIM-ONE DL, the F1-score rate decreased from 64.07% to 60.54%.

We showed the VGG16 network an improvement in the classification of new data with class balancing, especially for the ACRIMA dataset, achieving F1-score of 73.23%. For imbalanced data, testing with the ACRIMA dataset got F1-score equals to 51.92%. The F1-score was 30.28% with the REFUGE test, while with class balanced, the test data reached a rate of 34.95%. F1-score with imbalanced classes with RIM-

TABLE III: Cross-dataset for the voting classifier

Train	Test	Method	F1-score	Accuracy	Precision	Recall	Specificity	TP	TN	FP	FN
Traditional											
RIM-ONE + REFEUGE	ACRIMA	Imbalanced	46.99%	58.72%	84.31%	32.58%	92.23%	129	285	24	267
		Cluster Centroid	**74.53%**	**70.92%**	**73.35%**	**75.76%**	**64.72%**	**300**	**200**	**109**	**96**
RIM-ONE + ACRIMA	REFUGE	Imbalanced	30.47%	61.58%	18.60%	84.17%	59.07%	101	638	442	19
		Cluster Centroid	37.01%	71.92%	23.86%	82.50%	70.74%	99	764	316	21
ACRIMA + REFUGE	RIM-ONE	Imbalanced	64.07%	63.92%	49.52%	90.70%	49.20%	156	154	159	16
		Cluster Centroid	60.54%	61.03%	47.23%	84.30%	48.24%	145	151	162	27
VGG16											
RIM-ONE + REFEUGE	ACRIMA	Imbalanced	51.92%	60.85%	83.71%	37.63%	90.61%	149	280	29	247
		Cluster Centroid	**73.23%**	**62.98%**	**61.66%**	**90.15%**	**28.16%**	**357**	**87**	**222**	**39**
RIM-ONE + ACRIMA	REFUGE	Imbalanced	30.28%	87.33%	33.67%	27.50%	93.98%	33	1015	65	87
		Cluster Centroid	34.95%	88.83%	41.86%	30.00%	95.37%	36	1030	50	84
ACRIMA + REFUGE	RIM-ONE	Imbalanced	58.82%	53.81%	43.01%	93.02%	32.27%	160	101	212	12
		Cluster Centroid	58.51%	51.75%	42.09%	95.93%	27.48%	165	86	227	7
VGG19											
RIM-ONE + REFEUGE	ACRIMA	Imbalanced	46.50%	55.60%	71.96%	34.34%	82.85%	136	256	53	260
		Cluster Centroid	**71.16%**	**67.80%**	**71.61%**	**70.71%**	**64.08%**	**280**	**198**	**111**	**116**
RIM-ONE + ACRIMA	REFUGE	Imbalanced	37.94%	80.92%	28.11%	58.33%	83.43%	70	901	179	50
		Cluster Centroid	35.92%	78.00%	25.34%	61.67%	79.81%	74	862	218	46
ACRIMA + REFUGE	RIM-ONE	Imbalanced	58.95%	59.79%	46.20%	81.40%	47.92%	140	150	163	32
		Cluster Centroid	59.18%	55.05%	43.65%	91.86%	34.82%	158	109	204	14
MobileNet											
RIM-ONE + REFEUGE	ACRIMA	Imbalanced	52.01%	59.43%	77.50%	39.14%	85.44%	155	264	45	241
		Cluster Centroid	**75.67%**	**70.35%**	**70.19%**	**82.07%**	**55.34%**	**325**	**171**	**138**	**71**
RIM-ONE + ACRIMA	REFUGE	Imbalanced	34.10%	71.33%	22.14%	74.17%	71.02%	89	767	313	31
		Cluster Centroid	35.80%	74.00%	23.77%	72.50%	74.17%	87	801	279	33
ACRIMA + REFUGE	RIM-ONE	Imbalanced	61.78%	65.57%	50.94%	78.49%	58.47%	135	183	130	37
		Cluster Centroid	60.30%	61.44%	47.49%	82.56%	49.84%	142	156	157	30
ResNet50											
RIM-ONE + REFEUGE	ACRIMA	Imbalanced	49.10%	55.89%	69.77%	37.88%	78.96%	150	244	65	246
		Cluster Centroid	**70.98%**	**63.69%**	**64.40%**	**79.04%**	**44.01%**	**313**	**136**	**173**	**83**
RIM-ONE + ACRIMA	REFUGE	Imbalanced	43.88%	90.83%	56.58%	35.83%	96.94%	43	1047	33	77
		Cluster Centroid	53.10%	91.17%	56.60%	50.00%	95.74%	60	1034	46	60
ACRIMA + REFUGE	RIM-ONE	Imbalanced	59.02%	53.61%	42.97%	94.19%	31.31%	162	98	215	10
		Cluster Centroid	54.17%	41.03%	37.39%	98.26%	9.58%	169	30	283	3
Xception											
RIM-ONE + REFEUGE	ACRIMA	Imbalanced	37.31%	52.34%	71.43%	25.25%	87.06%	100	269	40	296
		Cluster Centroid	**64.67%**	**54.75%**	**57.59%**	**73.74%**	**30.42%**	**292**	**94**	**215**	**104**
RIM-ONE + ACRIMA	REFUGE	Imbalanced	34.47%	75.92%	23.68%	63.33%	77.31%	76	835	245	44
		Cluster Centroid	34.05%	74.50%	22.97%	65.83%	75.46%	79	815	265	41
ACRIMA + REFUGE	RIM-ONE	Imbalanced	55.97%	51.34%	41.21%	87.21%	31.63%	150	99	214	22
		Cluster Centroid	54.25%	42.27%	37.73%	96.51%	12.46%	166	39	274	6

ONE DL came to 58.82%, while with balanced classes, this index dropped to 58.51%, but recall increased to 95.93%.

The VGG19 network results with imbalanced data for the REFUGE and RIM-ONE DL tests were better than the ones with balanced data. The ACRIMA testing showed improvement with balanced data, reaching F1-score equals to 71.16%. Considering VGG19, the imbalanced model with the REFUGE test obtained F1-score equals to 37.94%, while for the balanced model, the F1-score was equal to 35.92%. For the RIM-ONE DL, the F1-score obtained was 58.95% and 58.51% for training with imbalanced and balanced data, respectively. For VGG19, imbalanced results are better for specificity rates, while results with balanced training have increased precision and reduced recall.

The MobileNet network results were similar to the VGG16's. The results for ACRIMA and REFUGE test data were better for balanced training. On the other hand, the RIM-ONE DL results were better with the imbalanced data, with the F1-score equals to 61.78%. For balanced data, there was an improvement in recall rates.

The ResNet50 model improved the rates with balanced data for the ACRIMA and REFUGE test, increasing all the evaluated metrics. For ACRIMA, the F1-score model reached 49.10% for the imbalanced training data and 70.98% for the balanced training data. For the REFUGE results with balanced and imbalanced data, the F1-score reached 43.88% and 53.10%, respectively. Results with balanced training data for testing with RIM-ONE DL were worse than those with imbalanced training data. And finally, the results with the Xception model improved only with the ACRIMA test dataset.

Accuracy is a metric for comparing results. However, we had to evaluate more than this metric since analyzing the false positives and negatives are crucial to verify each model's performance. For example, the VGG16 model presented an

accuracy rate of 88.83% with the REFUGE base for testing. However, the recall rate was low; the model identified only 36 of the 120 images with glaucoma. As for the RIM-ONE DL dataset, the same VGG16 network correctly classified more glaucoma images than non-glaucoma images, with a recall rate of 95.93%, while its specificity was only 27.48%.

We can obtain different model evaluations if we compare only one metric. Thus, analyzing many metrics and checking false positive and false negative rates is essential. The best results were from VGG19, VGG16 and traditional models, mainly with balanced data and using ACRIMA as a test dataset.

According to the results, some CNN architectures can get essential and sufficient information to distinguish a normal and glaucomatous fundus image. Even using different image datasets, our models correctly identified the classes. Thus, we chose the VGG19 network with a dense layer to classify the features, as it had the best rates with balanced data and presented a good performance in the cross-dataset. Furthermore, we improved the general model rates with the data balancing, especially with the VGG19 architecture, which achieved the best results.

We compared our proposed model with other related works, according to Table I. We obtained competitive results using various datasets, image qualities, and cameras. Our results are promising, with average rates higher than the ones founded in several studies, achieving satisfactory rates for identifying glaucoma in medical images.

[5], [6] and [7] applied a method of extraction of features non-structural in which the descriptors were chosen to obtain information from the images; [5], and [7] applied statistical information, [6] used information such as energy and entropy. We achieved competitive results even with our traditional model. We emphasize that the comparison is affected by the form of validation of the data and the bases used, mainly in works that use private datasets and without the application of evaluation techniques, which makes it difficult to replicate the techniques. Still, we got satisfactory results with merging datasets and 10-fold cross-validation.

Some authors merged features from different methodologies and added other descriptors manually to improve image class distinction, such as [8], which used structural and non-structural descriptors to represent images. [9], [10], and [19] used Convolutional Neural Networks with manually chosen feature extractors. [19] applied geometric features and CNN models to evaluate the results. The combination of these techniques can generate higher computational costs. In addition, some methods presented excellent results without the need for combining two or more methodologies.

Most related works applied cross-validation or repetition to obtain model results. Although [18] presented 10-folds validation, the authors opted for a methodology that presents optimistic results since the data increase was used before separating training, validation, and testing. Despite the changes in the image, variations of the same image can occur during training and testing, generating optimistic results once the data augmentation images come from the original image.

We compared different extraction methods and used 10-fold cross-validation to generalize our results. Our proposed methodology achieved promising results for joining the ACRIMA, REFUGE, and RIM-ONE DL datasets. The metrics rates were similar and even better than some works in the literature. Our results were satisfactory due to data validation since we used a 10-folds average. We emphasized that our study used the recent RIM-ONE DL dataset, which is few analyzed in the literature and it is composed of images from the three datasets of the RIM-ONE version, RIM-ONE v1, v2, and v3.

Some studies used the old versions of RIM-ONE, which could have duplicate images if the datasets are combined, as in the works by [6], [9] and [14].

The RIM-ONE DL dataset showed promising results when evaluated individually using transfer learning, as seen in [15] and with other datasets, such as our analysis. However, if used only for test analyzes, the values obtained are lower than those obtained with images of this dataset used in training, as shown in the cross-dataset analysis in Table III. Few authors dealed with this approach. Just [13] compared the model with an independent dataset, demonstrating a 12% reduction in retest accuracy. Furthermore, images from different datasets are labeled differently. It increases the noise and makes it challenging to generalize automatic classification.

Few related works approached dataset reduction to balance classes. In our work, we evaluated techniques that proved efficient for improving model performance. [5] also reduced the data to obtain better results. However, they used only random undersampling to define the images of the majority class. They had just 80 images and used 10-fold cross-validation, just eight images to test for each folder. Although we merged different datasets in addition to the REFUGE dataset, we got better results than [5].

In general, our method achieved promising results and showed high rates of reduced false negatives and false positives. In this way, the proposed method with balanced classes improved the model and increased the average rates obtained in test results. Therefore, the cluster centroid balancing method proved efficient in improving the model in learning the image features. In addition, we compared our results through non-parametric tests to verify the improvement in applying a balance to the dataset. Unfortunately, to our best knowledge, only some authors compared these results using statistical tests. For example, only [6], and [7] used statistical methods to validate their improvements.

Our methodology has many strengths, including analysis with different datasets, model evaluation using 10-fold cross-validation and cross-dataset methods. Moreover, the cluster centroid undersampling method improved the evaluation of the metric. However, some limitations can be improved. As in works involving medical images, the generalization with other datasets still needs to be enhanced due to image quality differences. A common way to deal with this issue is the application of pre-processing techniques such as CLAHE to improve image contrast. However, it is necessary to conduct more analysis to improve the model accuracy and generalization. Also, the acquisition angle changes according to

the dataset, and our method does not have an automatic localization technique of disc optics.

V. Conclusions

This paper exploited different feature extractions from fundus images and evaluated classification from traditional ML and DL methods. We presented an overview of descriptors and classifiers to identify glaucoma. Furthermore, we showed that the VGG19 network with the voting classifier is an efficient method to detect glaucoma in optic disc images. Our methodology has been adequate for encompassing different public datasets: ACRIMA, REFUGE, and RIM-ONE DL. We applied the recently available RIM-ONE DL dataset, which combines previous versions and removes duplicate images. In addition, we evaluated our models with 10-fold cross-validation and cross-dataset, allowing the generalization of the result.

The balanced method proved to be an alternative for the reduction of false positives and false negatives. The cluster centroid stood out from the others in the undersampling step and it has increased rates. The Mann-Whitney test confirmed this information, for most metrics had a significant difference to the results with imbalanced and balanced classes.

The VGG19 network achieved the best results with the combined dataset and class balancing. Regarding F1-score, the results achieved average rates of 94.69%, accuracy of 94.77%, precision of 96.10%, a recall of 93.45%, and a specificity of 96.08%. For the cross-dataset, the best result was training using the REFUGE and RIM-ONE DL datasets. The test dataset ACRIMA achieved F1-score equals to 75.67%, accuracy of 70.35%, precision of 70.19%, recall of 82.07% and specificity of 55.34%.

Our work achieved excellent results. However, we suggest implementing improvements for future works to overcome the limitations mentioned, such as real-time recognition of glaucoma. Furthermore, we recommend applying automatic detection algorithms such as YOLO to identify the optic disc. In addition, we also suggest including optic disc and cup segmentation to calculate structural features such as CDR in fundus images. Finally, adding other public datasets and evaluating other network CNN architectures for extraction and classification is also recommended.

References

[1] M. Ulieru, A. Grabelkovsky, Telehealth approach for glaucoma progression monitoring (2003).

[2] S. P. Syaefullah, Epidemiology approach for glaucoma (2021).

[3] W. H. Organization, et al., Blindness and vision impairment prevention. priority eye diseases, 2018.

[4] M. D. Abràmoff, M. K. Garvin, M. Sonka, Retinal imaging and image analysis, IEEE Reviews in Biomedical Engineering 3 (2010) 169–208.

[5] M.-A. Talaat, N. Raed, A. Medhat, R. Ashraf, M. Essam, R. Y. ElKashlan, L. Abdel-Hamid, Glaucoma detection from retinal images using generic features: Analysis & results, in: Proceedings of the 2019 2nd International Conference on Watermarking and Image Processing, 2019, pp. 10–15. doi:10.1145/3369973.3369976.

[6] D. Parashar, D. Agrawal, 2-d compact variational mode decomposition- based automatic classification of glaucoma stages from fundus images, IEEE Transactions on Instrumentation and Measurement 70 (2021) 1–10.

[7] S. I. Khan, S. B. Choubey, A. Choubey, A. Bhatt, P. V. Naishadhkumar, M. M. Basha, Automated glaucoma detection from fundus images using wavelet-based denoising and machine learning, Concurrent Engineering (2021) 1063293X211026620.

[8] N. Thakur, M. Juneja, Classification of glaucoma using hybrid features with machine learning approaches, Biomedical Signal Processing and Control 62 (2020) 102137.

[9] M. Claro, R. Veras, A. Santana, F. Araujo, R. Silva, J. Almeida, D. Leite, An hybrid feature space from texture information and transfer learning for glaucoma classification, Journal of Visual Communication and Image Representation 64 (2019) 102597.

[10] N. E. Benzebouchi, N. Azizi, A. S. Ashour, N. Dey, R. S. Sherratt, Multi-modal classifier fusion with feature cooperation for glaucoma diagnosis, Journal of Experimental & Theoretical Artificial Intelligence 31 (2019) 841–874.

[11] U. Raghavendra, H. Fujita, S. V. Bhandary, A. Gudigar, J. H. Tan, U. R. Acharya, Deep convolution neural network for accurate diagnosis of glaucoma using digital fundus images, Information Sciences 441 (2018) 41–49.

[12] O. Deperlioglu, U. Kose, D. Gupta, A. Khanna, F. Giampaolo, G. Fortino, Explainable framework for glaucoma diagnosis by image processing and convolutional neural network synergy: Analysis with doctor evaluation, Future Generation Computer Systems 129 (2022) 152–169.

[13] M. Norouzifard, A. Nemati, H. GholamHosseini, R. Klette, K. Nouri-Mahdavi, S. Yousefi, Automated glaucoma diagnosis using deep and transfer learning: Proposal of a system for clinical testing, in: 2018 International Conference on Image and Vision Computing New Zealand (IVCNZ), IEEE, 2018, pp. 1–6.

[14] J. J. Gómez-Valverde, A. Antón, G. Fatti, B. Liefers, A. Herranz, A. Santos, C. I. Sánchez, M. J. Ledesma-Carbayo, Automatic glaucoma classification using color fundus images based on convolutional neural networks and transfer learning, Biomedical optics express 10 (2019) 892–913.

[15] F. J. F. Batista, T. Diaz-Aleman, J. Sigut, S. Alayon, R. Arnay, D. Angel-Pereira, Rim-one dl: A unified retinal image database for assessing glaucoma using deep learning, Image Analysis & Stereology 39 (2020) 161–167.

[16] M. Juneja, N. Thakur, S. Thakur, A. Uniyal, A. Wani, P. Jindal, Gc-net for classification of glaucoma in the retinal fundus image, Machine Vision and Applications 31 (2020) 1–18.

[17] P. Elangovan, M. K. Nath, et al., Detection of glaucoma from fundus image using pre-trained densenet201 model (2021).

[18] I. D. S. Ubaidah, Y. Fu'Adah, S. Sa'Idah, R. Magdalena, A. B. Wiratama, R. B. J. Simanjuntak, Classification of glaucoma in fundus images using convolutional neural network with mobilenet architecture, in: 2022 1st International Conference on Information System Information Technology (ICISIT), 2022, pp. 198–203. doi:10.1109/ICISIT54091.2022.9872945.

[19] L. K. Singh, M. Khanna, et al., A novel multimodality based dual fusion integrated approach for efficient and early prediction of glaucoma, Biomedical Signal Processing and Control 73 (2022) 103468.

[20] B. R. Silva, R. J. Silveira, M. G. da Silva Neto, P. C. Cortez, D. G. Gomes, A comparative analysis of undersampling techniques for network intrusion detection systems design, Journal of Communication and Information Systems 36 (2021) 31–43.

[21] A. Diaz-Pinto, S. Morales, V. Naranjo, T. Köhler, J. M. Mossi, A. Navea, CNNs for automatic glaucoma assessment using fundus images: an extensive validation, Biomedical engineering online 18 (2019) 1–19.

[22] J. I. Orlando, H. Fu, J. B. Breda, K. van Keer, D. R. Bathula, A. Diaz-Pinto, R. Fang, P.-A. Heng, J. Kim, J. Lee, et al., Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs, Medical image analysis 59 (2020) 101570.

[23] M.-K. Cheon, W.-J. Lee, C.-H. Hyun, M. Park, Rotation invariant histogram of oriented gradients, International Journal of Fuzzy Logic and Intelligent Systems 11 (2011) 293–298.

[24] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 1, IEEE, 2005, pp. 886–893.

[25] Y. Wang, J. See, R. C.-W. Phan, Y.-H. Oh, Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition, PloS one 10 (2015) e0124674.

[26] S.-K. Hwang, W.-Y. Kim, A novel approach to the fast computation of Zernike moments, Pattern Recognition 39 (2006) 2065–2076.

[27] K. G. Derpanis, Gabor filters (2007).

[28] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, nature 521 (2015) 436–444.

[29] Y. Wu, X. Qin, Y. Pan, C. Yuan, Convolution neural network based transfer learning for classification of flowers, in: 2018 IEEE 3rd International Conference on Signal and Image Processing (ICSIP), IEEE, 2018, pp. 562–566.

[30] H.-Y. Chen, C.-Y. Su, An enhanced hybrid MobileNet, in: 2018 9th International Conference on Awareness Science and Technology (iCAST), IEEE, 2018, pp. 308–312.

[31] T. Pomari, E. Rezende, G. Ruppert, F. Balieiro, T. Carvalho, Associando redes convolucionais e características de iluminação para detectar falsificações em imagens, Conference on graphics, pattern and images, 31 (SIBGRAPI) (2018) 31.

[32] C. S. Bezerra, Uma abordagem de segmentação de íris para fins biométricos usando aprendizagem profunda, 2019.

[33] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.

[34] N. Nachar, et al., The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution, Tutorials in quantitative Methods for Psychology 4 (2008) 13–20.

Débora Ferreira de Assis is a PhD candidate in Teleinformatics Engineering at Federal University of Ceará (UFC), Brazil. She received a Master degree in Teleinformatics Engineering from Federal University of Ceará, Brazil, in 2020. Her research interests are image processing, video recognition, statistics and artificial intelligence. Currently, she participates in one research projects and she works as a researcher at the Computer Systems Engineering Laboratory (LESC). ORCID: 0000-0003-0750-5784. IDLattes: 5346773401665003.

Paulo Cesar Cortez received the B.Sc.degree in electrical engineering from the Federal University of Ceará, Brazil, in 1982, and the M.Sc.and Ph.D. degrees in electrical engineering from the Federal University of Paraíba, in 1992 and 1996, respectively. He is currently a Full Professor at the Department of Teleinformatics Engineering, Federal University of Ceará. His fields of interest include image and signal analysis, computer vision, biomedical signal processing, and biomedical systems. ORCID: 0000-0002-4020-3019; IDLattes: 5024602152304064.