Image Anal Stereol 2020;39:161-167 doi: 10.5566/ias.2346
Original Research Paper

RIM-ONE DL: A UNIFIED RETINAL IMAGE DATABASE FOR ASSESSING GLAUCOMA USING DEEP LEARNING

FRANCISCO FUMERO✉,[1], TINGUARO DIAZ-ALEMAN[2], JOSE SIGUT[1], SILVIA ALAYON[1], RAFAEL ARNAY[1] AND DENISSE ANGEL-PEREIRA[2]

[1]Department of Computer Engineering and Systems, University of La Laguna, Campus de Anchieta, 38200, Tenerife, Spain, [2]Servicio de Oftalmología, Hospital Universitario de Canarias, 38320, Tenerife, Spain
e-mail: ffumerob@ull.edu.es, vtdac@hotmail.com, jfsigut@ull.edu.es, salayon@ull.edu.es, rarnay@ull.edu.es, dangelp30@hotmail.com
(Received February 11, 2020; revised October 5, 2020; accepted October 5, 2020)

ABSTRACT

The first version of the **R**etinal **IM**age database for **O**ptic **N**erve **E**valuation (RIM-ONE) was published in 2011. This was followed by two more, turning it into one of the most cited public retinography databases for evaluating glaucoma. Although it was initially intended to be a database with reference images for segmenting the optic disc, in recent years we have observed that its use has been more oriented toward training and testing deep learning models. The recent REFUGE challenge laid out some criteria that a set of images of these characteristics must satisfy to be used as a standard reference for validating deep learning methods that rely on the use of these data. This, combined with the certain confusion and even improper use observed in some cases of the three versions published, led us to consider revising and combining them into a new, publicly available version called RIM-ONE DL (RIM-ONE for **D**eep **L**earning). This paper describes this set of images, consisting of 313 retinographies from normal subjects and 172 retinographies from patients with glaucoma. All of these images have been assessed by two experts and include a manual segmentation of the disc and cup. It also describes an evaluation benchmark with different models of well-known convolutional neural networks.

Keywords: Convolutional Neural Networks, Deep Learning, Glaucoma Assessment, RIM-ONE.

INTRODUCTION

The term glaucoma refers to a group of pathologies that affect the optic nerve and involve the loss of retinal ganglion cells, which is frequently associated with an increase in intraocular pressure. It is one of the leading causes of blindness in the world (Tham *et al.*, 2014) and, since the progression of the disease is typically asymptomatic, early detection is quite difficult. There are different techniques for viewing the retina that help in making this diagnosis. One of them is a retinography, which is a color photograph of the fundus of the eye. Fig. 1 shows an example of a retinography, where the most relevant parts for diagnosing glaucoma have been highlighted: the optic disc, the cup and the neuroretinal rim. In fact, since this is the most important part of the retinography, it is typical to cut around it and discard the rest. As in other fields, automated learning-based diagnoses, and more specifically the technique known as deep learning, have taken on great significance. The key to the proper functioning of these methods is the availability of a sufficient amount of data with which to train and test the system. In addition, validating these methods requires a reference standard that can be used for comparison. In the case at hand, this involves having public retinography databases that satisfy a series of requirements, which must also be clearly defined. In the very recent paper on the Retinal Fundus Glaucoma Challenge (REFUGE) (Orlando *et al.*, 2020), written by researchers from 20 institutions, a very important step is taken in this direction by suggesting certain criteria that can be used to compare these methods in terms of both classifying the glaucoma and segmenting the disc and cup:

1. Availability of publicly-accessible sets of images, labelled by several experts, sufficient enough in number that they can be used in deep-learning methods.
2. Clear separation between training and test sets. As noted in (Trucco *et al.*, 2013), a comparison of the results may be unreliable without said separation.
3. Presence of diversity in the set of images, with diversity meaning having images captured by various devices involving different patient ethnicities, and images taken in different lighting, contrast, noise and other conditions.
4. In addition to having a preliminary diagnosis, include also manual reference segmentations of the disc and cup.

5. Provide an evaluation framework that includes the metrics used and the format for presenting the results.

The first open version of the set of fundus images called **R**etinal **IM**age database for **O**ptic **N**erve **E**valuation (RIM-ONE) was published in 2011 (Fumero *et al.*, 2011). Two other versions followed, in 2014 and 2015. They will be referred to in this paper as RIM-ONE v1, v2 and v3, respectively. Since its publication, it has been cited 179 times, becoming, based on our data, the most cited public set of retinographies for evaluating glaucoma. Although it was initially intended to be used as a database of reference images for segmenting the optic disc, in recent years we have observed that its use has been more oriented toward training and testing deep learning models. This, combined with the certain confusion and even improper use observed in some cases of the three versions published, led us to consider revising and combining them into a new version called RIM-ONE DL (RIM-ONE for **D**eep **L**earning), optimized for a deep-learning context in keeping with the specifications explained earlier. Furthermore, for benchmarking purposes, we studied the performance of various convolutional neural network models that are very popular, due to their widespread use for the classification and semantic segmentation of natural images. As a result, we hope to lay the foundations to have RIM-ONE DL become a reference for evaluating glaucoma, like the previous versions were.

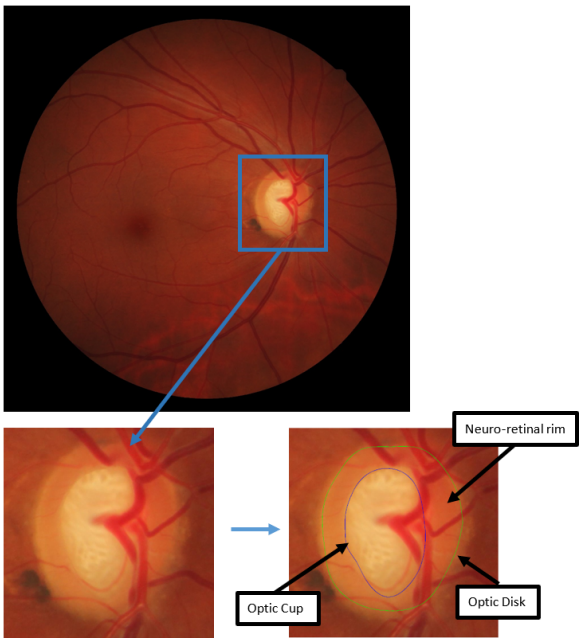


Fig. 1: Sample retinography with the most relevant regions for diagnosing glaucoma.

The rest of the paper is structured as follows. The section on "Related Work" describes other sets of public images for evaluating glaucoma. The "RIM-ONE database" section presents the different versions of this imaging database and analyzes the feedback received over its nearly ten years of use. The "Materials and Methods" section describes RIM-ONE DL, underscoring the changes made with respect to previous versions and explaining the neural network models and the metrics used in the benchmark. We conclude with the "Results and Discussion" section.

RELATED WORK

There are not too many public sets of images of the fundus of the eye for evaluating glaucoma, whether for classification or for segmenting the optic disc and cup. ORIGA (Zhang *et al.*, 2010) is composed of 168 images from patients with glaucoma and 482 from healthy patients. The data also include the disc and cup segmentation. The problem with this database is that even though it appears to have been public at one point, as far as we can tell, it stopped being public quite some time ago. DRISHTI-GS (Sivaswamy *et al.*, 2014) consists of 70 images of glaucoma and 31 normal images. It also includes the disc and cup segmentation. DR HAGIS (Holm *et al.*, 2017), HRF (Odstrcilik *et al.*, 2013), and LES-AV (Orlando *et al.*, 2018) are small sets with 39, 45 and 22 images, respectively, with no disc and cup segmentation. The ACRIMA set (Diaz-Pinto *et al.*, 2019) contains 396 images from patients with glaucoma and 309 images from healthy patients. It also does not include segmentation of the disc or cup. Finally, the recent REFUGE database (Orlando *et al.*, 2020) contains 120 images from patients with glaucoma and 1080 images from healthy patients with disc and cup segmentation.

It is important to note that of all the sets mentioned, only REFUGE satisfies the additional requirements of offering images from different cameras, as well as a clear division of the training and test data. There seems to be a clear need, then, to expand the number of public image databases available that comply with these requirements.

THE RIM-ONE DATABASE

The images in the three versions of RIM-ONE include healthy and glaucomatous eyes, and were taken in various Spanish hospitals: Hospital Universitario de Canarias (HUC), in Tenerife, Hospital Universitario Miguel Servet (HUMS), in Zaragoza,

and Hospital Clínico Universitario San Carlos (HCSC), in Madrid. The repository of retinographies is accessible through the website of the research group at `https://medimrg.webs.ull.es`.

CRITICAL REVISION OF THE THREE VERSIONS OF RIM-ONE

RIM-ONE v1 was presented in 2011 at the 24th International Symposium on Computer-Based Medical Systems (CBMS) (Fumero *et al.*, 2011). The main goal of this work was to provide an open database of retinographies from 118 healthy subjects and 51 patients with various stages of glaucoma. In addition to a diagnosis, it included a manual segmentation of the optic disc, carried out by five experts in the field. This was the main goal of its publication, to provide a reference for assessing methods to segment the disc. The images were taken at the three hospitals mentioned above using a Nidek AFC-210 non-mydriatic fundus camera with a 21.1-megapixel Canon EOS 5D Mark II body, with a vertical and horizontal field of view of 45°.

RIM-ONE v2 was published in 2014. It contains 255 images from healthy subjects and 200 images from patients with glaucoma that are manually segmented by a medical specialist. In this case, the images were taken at HUC and the Hospital Universitario Miguel Servet using the same camera as in v1. It is important to note that this version was designed as an extension of the first; as a result, some images are duplicated. It also includes some test-retest cases that give rise to images that are practically identical.

RIM-ONE v3 was published in 2015 and contains 85 images of healthy subjects and 74 images of patients with glaucoma. The main difference between this version and the two previous ones is that the images were captured only in the HUC with a non-mydriatic Kowa WX 3D stereo fundus camera. The images are centered on the ONH using a field angle of 34°, giving a final stereo image with a horizontal field of view of 20° and a vertical field of view of 27°, with a total resolution of 2144 x 1424 pixels (1072 x 1424 pixels per image in the stereo pair). Having stereo images available resulted in a manual segmentation of not only the optic disc, but of the cup as well. Two specialists carried out this task with help from the freely distributed DCSeg tool. More details on this version and on the tool are available at (Fumero *et al.*, 2015). It should be noted that some of the subjects whose retinographies are contained in v2 are also present in v3.

As was stated in the introduction, a critical review of these three versions is necessary in order to determine how well they comply with the criteria in place for using them in methods based on deep learning:

1. Although it may be tempting to combine the three versions for use in deep-learning problems, as indicated earlier, indiscriminately combining the images could result in their inappropriate use. It should also be noted that there is no consistency between the three versions in terms of the labelling of the images, as this was not always done by the same experts.
2. Since RIM-ONE was not originally designed for deep learning, a clear division between training and test images was never established.
3. The RIM-ONE images were taken in different hospitals with different cameras, but only one camera was used in each version.
4. In this area there is also little consistency between versions, since in the first two only the disc is segmented, while the cup is also segmented in v3. As happened with the diagnosis, the specialists involved in the manual segmentation were not the same in every case.
5. Although version 1 does propose a criterion for evaluating the quality of the disc segmentation, no details are given on the metrics for evaluating the diagnosis, since that was not the initial idea.

The “Materials and Methods” section details the process used to attempt to resolve these problems with RIM-ONE DL.

FEEDBACK ON THE EXPERIENCE WITH RIM-ONE

Using a procedure similar to that presented in (Decencière *et al.*, 2014) for the well-known Messidor database for segmenting the optic disc in diabetic retinography images, in this section we will focus on the feedback received over the nearly 10 years that RIM-ONE has been publicly available. To provide a quantitative idea of the impact that its publication has had, we used as a reference the total number of citations, which is 179. Moreover, Fig. 2 shows the recorded trend in terms of the primary purpose for which RIM-ONE has been used. In its early years, a large majority of the uses were centered on segmentation tasks. However, the last two years have seen a very significant increase in the number of publications in which its use was associated with deep-learning problems, a use that in 2019 even outpaced the number of works involving segmentation. This reinforces the need to have a revised and updated version of RIM-ONE that can satisfy this new trend.

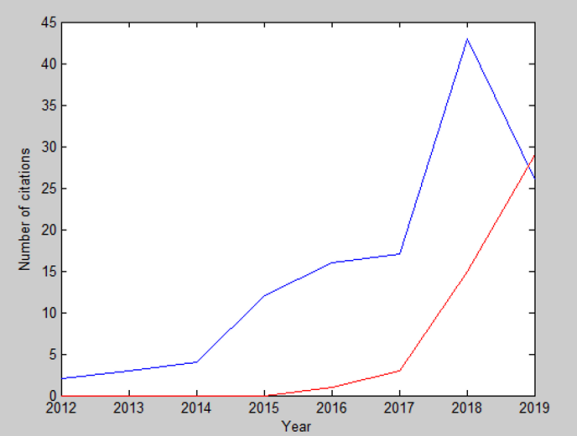


Fig. 2: Number of citations of RIM-ONE. Blue: RIM-ONE for segmentation, Red: RIM-ONE for Deep Learning.

MATERIALS AND METHODS

RIM-ONE DL

In this section, we describe RIM-ONE DL, which results from combining the three previous versions. This new version does away with the duplicate images contained in v1 and v2, and it also eliminates the test-retest images in v2. The images of the same patient in v2 and v3 were also deleted, with only the left image from v3 being retained. The result is a single image per patient and eye. Moreover, all the images were cropped squarely around the head of the optic nerve using the same proportionality criterion, something that was not done in the previous versions. Table 1 shows the minimum and maximum size of these cropped images per version and the hospital where the images were taken. Furthermore, the format of all the images in this new version is PNG, and the file names are prefixed with "r1", "r2" or "r3" according to the RIM-ONE version they were extracted from.

Table 1: Minimum and maximum size of the cropped images of RIM-ONE DL per version, indicating the hospital where the images were taken.

Version	Hospital	Min. Size	Max. Size
v1	HCSC, HUMS	316	708
v2	HUC, HUMS	274	793
v3	HUC	318	626

As concerns its use in deep-learning problems, and as discussed in previous sections, we note the following:

1. The final set of images consists of 313 images from healthy subjects and 172 images from patients with glaucoma. In order to standardize the criterion of experts for classifying glaucoma, two experts again reviewed all the images and re-labelled them after a visual inspection. In the event of a disagreement between them, a third specialist with 20 years of experience was consulted, who made the final decision.
2. A clear division is established between the training and test sets, with two variants. In one, the test set is built randomly, while in the other, the samples taken in the HUC are used for training and the samples taken in the two other hospitals (in Madrid and Zaragoza) are used for testing.
3. This combined version exhibits great diversity in terms of the cameras and hospitals.
4. In addition to the ground truth for classification, the set includes the manual segmentation of the disc and cup performed by one of the specialists.
5. The sub-section below details the evaluation framework for classifying the glaucoma disease.

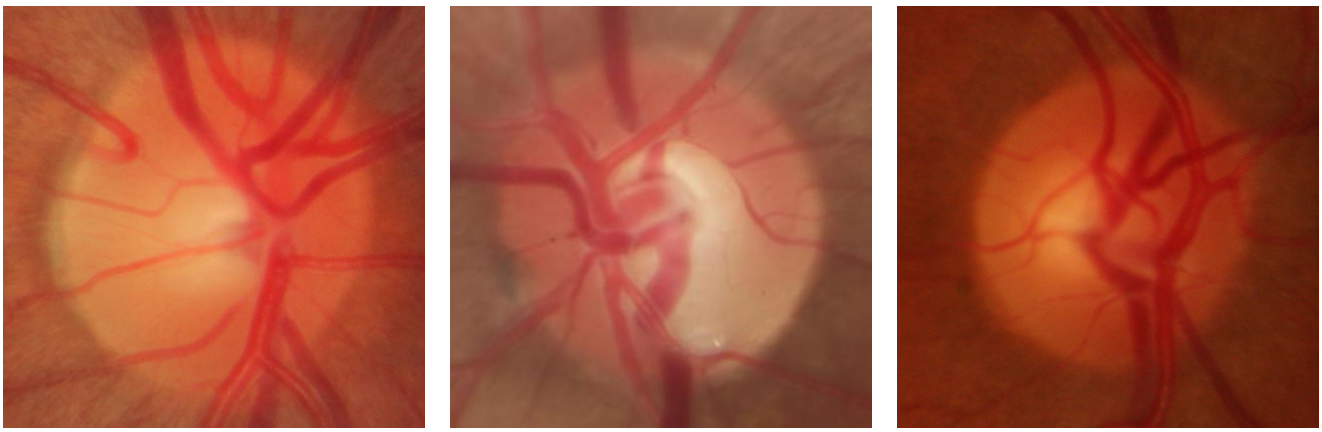
Fig. 3 shows some examples of the images contained in RIM-ONE DL, indicating the hospital they were taken in. This database is publicly available at the following location: `https://github.com/miag-ull/rim-one-dl`

EVALUATION FRAMEWORK

The evaluation framework proposed contains four main elements: definition of the training and test sets, neural network models used, training and testing strategy employed, and the metrics considered in the evaluation.

As concerns the training and test sets used, as noted in the preceding sub-section, two variants are considered. In the first variant, the set of images was divided at random into training and testing images using a 70:30 ratio, respectively. In the second variant, the images taken at the HUC were used for training (195 normal and 116 glaucoma), and the images taken at the two other hospitals were used for testing (118 normal and 56 glaucoma). The only processing done to the images involved re-scaling them in intensity in the 0-1 range and resizing them to 224x224x3.

In terms of the neural network models used, most of the architectures contained in the Keras Deep Learning Framework were tested: Xception, VGG16, VGG19, ResNet50, InceptionV3, InceptionResNetV2, MobileNet, DenseNet121, NASNetMobile and MobileNetV2. In every case, the size of the input layer was set to 224x224x3, and a GlobalAveragePooling2D layer was added to the convolutional base, followed by a fully-connected output layer with two outputs,



(a) Image of a healthy subject taken at Hospital Universitario de Canarias.

(b) Glaucoma image taken at Hospital Universitario Miguel Servet.

(c) Image of a healthy subject taken at Hosp. Clínico Universitario San Carlos.

Fig. 3: Examples of images included in RIM-ONE DL, indicating in which hospital they were taken.

using SoftMax to distinguish between the Normal and Glaucoma classes.

The training strategy was the same with both variants of the data sets. We started with the pre-trained networks using the weight values of ImageNet provided by Keras and fine tuned all the layers. Diaz-Pinto *et al.* (2019) found that this yielded the best results. To avoid overfitting, data augmentation consisting of random rotations (-30°, 30°), vertical and horizontal flip, and zoom (0.8, 1.2) was used.

The steps carried out to fine tune the networks were as follows:

1. Freeze the convolutional base network.
2. Train the part that was added.
3. Unfreeze all the layers in the base network.
4. Jointly train all the layers in the network.

To increase the reliability of the experiments, a 5-fold cross-validation was applied, with a proportion of 80% for the training set and 20% for the validation set in each fold. For each fold, the steps listed were followed in order to determine the most suitable number of epochs in 2 and 4. For the training in step 2, a batch size of 32 was used, along with an RMSprop optimizer with a learning rate of 2e-5, and categorical cross-entropy as a loss function. For the training in step 4, the learning rate was set to 1e-5. Once the validation phase was complete, the final model for each network was trained using the whole training data (no folds) and following the same four steps as before for the number of epochs that maximized the average validation accuracy across folds in steps 2 and 4. This final model was used to evaluate each network in the test set.

As concerns the metrics used, the outline proposed in (Orlando *et al.*, 2020) was used, in which the area under the curve (AUC) is used as a reference evaluation measure. This measure was complemented with the sensitivity value ($Se = Tp/(Tp+Fn)$) at a specificity of 0.85 ($Sp = Tn/(Tn+Fp)$), where Tp, Fp, Tn and Fn are the number of true positives, false positives, true negatives and false negatives, respectively. This allows for an assessment of the performance of the various networks when a low rate of false positives is imposed. The third measure included is accuracy, which is fairly standard in this type of problem, although it is well-known that it can exhibit some bias in data sets whose classes are not properly balanced.

RESULTS AND DISCUSSION

Tables 2 and 3 and figures 4 and 5 show the results of the glaucoma classifications obtained under the conditions described in the preceding section. In the case of the random test sample, the results are highly satisfactory. The VGG19 network model not only provided the highest AUC, but its sensitivity also equalled 1, the highest possible. The other network model with similar characteristics, VGG16, also yielded good results. Although a direct comparison with the results of the REFUGE challenge is not possible, it is interesting to note that the winning team (Son *et al.*, 2018) attained an AUC of 0.9885 with a sensitivity of 0.9752 for a test sample consisting of 360 images from healthy subjects and 40 images from patients with glaucoma. In the case of the test sample from the hospitals in Madrid and Zaragoza, there is a significant drop in all the metrics, with the best

response again being obtained by networks VGG19 and VGG16. This drop could be explained by the fact that a set of test images was used whose visual appearance was rather different from that of the images used during training. It is important to keep in mind that the images were captured in different hospitals under different circumstances, which seems to have affected the networks. The lack of robustness of this type of system to distortions that can affect the images, such as noise, contrast or lighting, has been analyzed by various authors (Borkar and Karam, 2019). Again, it is difficult to cite any work to compare against in this regard. The closest would be (Diaz-Pinto *et al.*, 2019), in which the Xception network was trained using images from some public databases, and it was trained with different databases whose images were taken under varying conditions. Its results are in keeping with those stemming from our experiments, with the exception that in our case, the experts who performed the reference diagnoses were the same for the training and test groups, unlike in the aforementioned work. This leads us to think that even though this factor could have some influence, the fact that the images were taken in different conditions is likely to be more relevant.

Table 2: Evaluation of the different networks using the random test set.

Network	AUC	Se	Acc.
VGG19	0.9867	1.0000	0.9315
VGG16	0.9834	0.9615	0.9247
Xception	0.9771	0.9808	0.9178
ResNet50	0.9755	0.9808	0.9110
MobileNetV2	0.9738	0.9423	0.9041
DenseNet	0.9726	0.9615	0.9041
MobileNet	0.9712	0.9615	0.9315
InceptionResNetV2	0.9685	0.9808	0.9110
InceptionV3	0.9597	0.9423	0.8904
NASNetMobile	0.9290	0.9231	0.7534

Table 3: Evaluation of the different networks using the test set from Madrid and Zaragoza.

Network	AUC	Se	Acc.
VGG19	0.9272	0.8750	0.8563
VGG16	0.9177	0.8214	0.8506
InceptionV3	0.9015	0.7500	0.8046
Xception	0.8982	0.7500	0.7989
DenseNet	0.8919	0.7143	0.7816
MobileNet	0.8912	0.7500	0.8276
ResNet50	0.8855	0.7321	0.8333
InceptionResNetV2	0.8396	0.625	0.7644
NASNetMobile	0.7969	0.6071	0.7989
MobileNetV2	0.7765	0.4464	0.5287

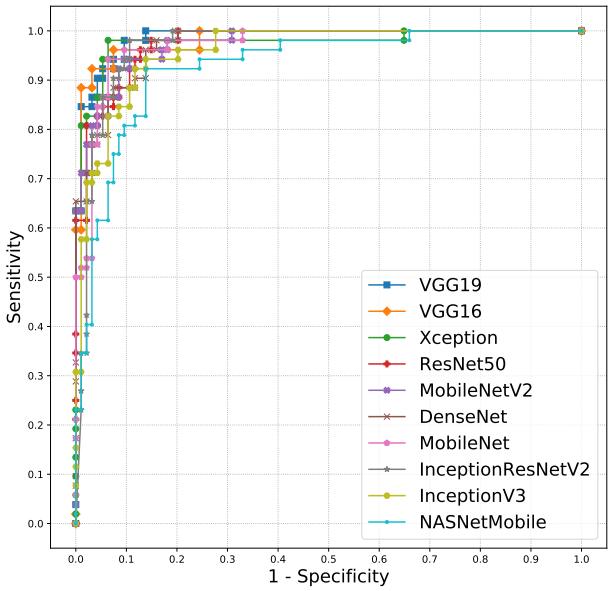


Fig. 4: ROC Curves for all the networks using the random test set.

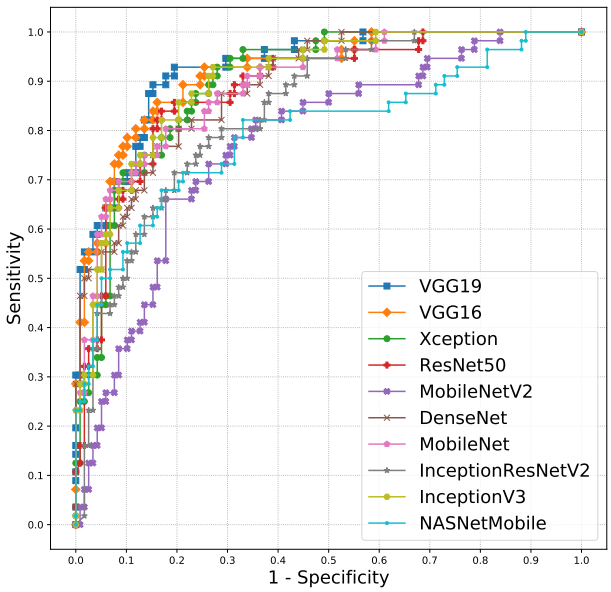


Fig. 5: ROC Curves for all the networks using the test set from Madrid and Zaragoza.

REFERENCES

Borkar TS, Karam LJ (2019). DeepCorrect: Correcting DNN Models Against Image Distortions. IEEE T Image Process 28:6022–34.

Decencière E, Zhang X, Cazuguel G, Lay B, Cochener B, Trone C, Gain P, Ordonez R, Massin P, Erginay A, Charton B, Klein JC (2014). Feedback

on a publicly distributed database: the Messidor database. Image Anal Stereol 33:231–4.

Diaz-Pinto A, Morales S, Naranjo V, Köhler T, Mossi JM, Navea A (2019). CNNs for automatic glaucoma assessment using fundus images: an extensive validation. Biomed Eng Online 18:1–19.

Fumero F, Alayon S, Sanchez JL, Sigut J, Gonzalez-Hernandez M (2011). RIM-ONE: An open retinal image database for optic nerve evaluation. In: 2011 24th International Symposium on Computer-Based Medical Systems (CBMS).

Fumero F, Sigut J, Alayon S, Gonzalez-Hernandez M, Gonzalez de la Rosa M (2015). Interactive tool and database for optic disc and cup segmentation of stereo and monocular retinal fundus images. In: Short communications proceedings. Plzen, Czech Republic: Václav Skala - UNION Agency.

Holm S, Russell G, Nourrit V, McLoughlin N (2017). DR HAGIS - a fundus image database for the automatic extraction of retinal surface vessels from diabetic patients. J Med Imaging Bellingham 4:1–11.

Odstrcilik J, Kolar R, Budai A, Hornegger J, Jan J, Gazarek J, Kubena T, Cernosek P, Svoboda O, Angelopoulou E (2013). Retinal vessel segmentation by improved matched filtering: evaluation on a new high-resolution fundus image database. IET Image Process 7:373–83.

Orlando JI, Fu H, Barbosa Breda J, van Keer K, Bathula DR, Diaz-Pinto A, Fang R, Heng PA, Kim J, Lee J, Lee J, Li X, Liu P, Lu S, Murugesan B, Naranjo V, Phaye SSR, Shankaranarayana SM, Sikka A, Son J, van den Hengel A, Wang S, Wu J, Wu Z, Xu G, Xu Y, Yin P, Li F, Zhang X, Xu Y, Bogunović H (2020). REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Med Image Anal 59:101570.

Orlando JI, van Keer K, Barbosa Breda J, Blaschko MB, Blanco P, Bulant CA (2018). Towards a glaucoma risk index based on simulated hemodynamics from fundus images. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-Lopez C, Fichtinger G, eds., Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Lecture Notes in Computer Science. Springer International Publishing, 65–73.

Sivaswamy J, Krishnadas S, Datt Joshi G, Jain M, Syed Tabish A (2014). Drishti-GS: Retinal image dataset for optic nerve head (ONH) segmentation. In: 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI).

Son J, Bae W, Kim S, Park SJ, Jung KH (2018). Classification of Findings with Localized Lesions in Fundoscopic Images Using a Regionally Guided CNN. In: Stoyanov D, Taylor Z, Ciompi F, Xu Y, Martel A, Maier-Hein L, Rajpoot N, van der Laak J, Veta M, McKenna S, Snead D, Trucco E, Garvin MK, Chen XJ, Bogunovic H, eds., Computational Pathology and Ophthalmic Medical Image Analysis, Lecture Notes in Computer Science. Springer International Publishing, 176–84.

Tham YC, Li X, Wong TY, Quigley HA, Aung T, Cheng CY (2014). Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. Ophthalmology 121:2081–90.

Trucco E, Ruggeri A, Karnowski T, Giancardo L, Chaum E, Hubschman JP, Al-Diri B, Cheung CY, Wong D, Abràmoff M, Lim G, Kumar D, Burlina P, Bressler NM, Jelinek HF, Meriaudeau F, Quellec G, Macgillivray T, Dhillon B (2013). Validating retinal fundus image analysis algorithms: issues and a proposal. Invest Ophth Vis Sci 54:3546–59.

Zhang Z, Yin FS, Liu J, Wong WK, Tan NM, Lee BH, Cheng J, Wong TY (2010). ORIGA-light: An online retinal fundus image database for glaucoma analysis and research. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology.