Journal of Electrical Engineering, Vol. 74, No. 4, 2023, pp. 328-335

sciendo
PAPERS

Detecting glaucoma from fundus images using ensemble learning

Veronika Kurilová[1,2], Szabolcs Rajcsányi[1], Zuzana Rábeková[1], Jarmila Pavlovičová[1], Miloš Oravec[1], Nora Majtánová[2,3]*

Glaucomatous changes of the optic nerve head could be detected from fundus images. Focusing on optic nerve head appearance, and its difference from healthy images, altogether with the availability of plenty of such images in public fundus image databases, these images are ideal sources for artificial intelligence methods applications. In this work, we used ensemble learning methods and compared them with various single CNN models (VGG-16, ResNet-50, and MobileNet). The models were trained on images from REFUGE public dataset. The average voting ensemble method outperformed all mentioned models with 0.98 accuracy. In the AUC metric, the average voting ensemble method outperformed VGG-16 and MobileNet models, which had significantly weaker performance when used alone. The best results were observed using the ResNet-50 model. These results confirmed the significant potential of ensemble learning in enhancing the overall predictive performance in glaucomatous changes detection, but the overall performance could be negatively affected when models with weaker prediction performance are included.

Keywords: ensemble learning, neural networks, deep learning, glaucoma, optic disc

1 Introduction

Glaucoma is one of the leading causes of irreversible blindness in the world. If detected and treated early, the progression could be slowed down or even stopped [1]. There is a lack of ophthalmologists, even more in developing countries [2]. Therefore, developing new screening possibilities, which could be used without an available ophthalmologist, are favourable. In this article, we will focus on detecting glaucoma using ensemble learning methods consisting of deep learning models, on fundus images.

Glaucomatous changes of the optic disc are visible on fundus examination [3]. Such fundus examination can be performed on a standard fundus camera, or also by mobile phone using special fundus attachment with magnifying lens [4]. Glaucoma screening using a smartphone could be a new screening option in remote areas [5]. An output, digitalized fundus images, could be further analysed using artificial intelligence methods. Artificial intelligence methods offer fast and precise image detection possibilities, are frequently used on medical data [6], and are successfully applied to glaucomatous optic disc images [7]. To obtain more precise results, ensemble learning methods could be applied, also in the field of deep learning image classification.

Ensemble learning is the method of combining multiple predictions to make a decision. The goal is that the accuracy of aggregated prediction models will outperform the best individual predictor [8]. In our work, we used the hard and average voting ensemble method. For hard voting, the class with the most votes from the individual models was selected as the final prediction. Average voting, on the other hand, means calculating the average of the probabilities predicted by each model for each class, with the class with the highest average probability being the chosen prediction. Through these methods, we aimed to leverage the strengths of each individual model to make a collective, more robust prediction.

2 Related work

In this part, we will summarize related works in the field of glaucoma screening or diagnosis using ensemble methods (Table 1). Some works use ensemble methods consisting of machine learning classifiers only [9], other ones are combining deep learning models with machine learning classifiers [10, 11], and finally, some are using deep learning models only [12-16]. Another principle

[1] Faculty of Electrical Engineering and Information Technology, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia
[2] Department of Ophthalmology of the Slovak Medical University and University Hospital in Bratislava, Antolská, 11, 85107 Bratislava, Slovakia
[3] Faculty of Medicine, Slovak Medical University, Limbová 12, 833 03 Bratislava, Slovakia
* nora.majtanova@gmail.com

https://doi.org/10.2478/jee-2023-0040, Print (till 2015) ISSN 1335-3632, On-line ISSN 1339-309X

Table 1 Related works in the field of glaucoma screening or diagnosis using ensemble methods

Authors	Used classifiers and neural networks	Dataset	Number of images	Evaluation
Cho et al.**[12]**	56 CNN models based on InceptionNetV3, InceptionResNet-V2	private	3460	88.1% (Acc.) 0.975 (AUC-ROC)
Deepa et al. **[13]**	VGG-16, ResNet-50, GoogLeNet	private DRISHTI-GS DRIONS-DB HRF	1150 101 110 30	91.13% (Acc.) 88.96% (Acc.)
Elangovan et al. **[10]**	Alexnet, GoogLeNet, VGG-16, VGG-19, Squeezenet, ResNet-18, ResNet-50, ResNet-101, EfficientNet-b0, MobileNet-v2, Densenet-201, Inception-V3, Xception, SVM classifier	DRISHTI-GS1-R ORIGA-R RIM-ONE2-R LAG-R ACRIMA-R	101 650 455 5504 705	93.4% (Acc.) 79.6% (Acc.) 91.3% (Acc.) 99.6% (Acc.) 99.5% (Acc.)
Fu et al. **[14]**	ResNet-50, U-net	ORIGA Singapore Chinese Eye Study (SCES) Singapore Indian Eye Study (SINDI)	650 1676 5783	 84.29% (Acc.) 0.9183 (AUC) 74.95% (Acc.) 0.8173 (AUC)
Li et al. **[11]**	U-net, Random Forest classifier	ORIGA Singapore Chinese Eye Study (SCES)	650 1676	0.96 (AUC) 86% (Acc.) 0.98 (AUC)
Mahdi et al. **[15]**	Inception, DenseNet	private	-	93.84% (Acc.)
Patra et al.**[17]**	Support Vector Machine, Random Forest, Multilayer Perceptron	REFUGE ORIGA	400 650	99.86% (Acc.) 88.48% (Acc.)
Tekuoabou Koumetio et al.[9]	AlexNet, AdaBoosting, Bagging Random Forest, Gradient Boosting classifiers	private	401	 81.44% (Acc.) 87.55% (Acc.) 90.3% (Acc.) 88.61%
Wang et al.**[16]**	SeResNext-50, VGG-16, DenseNet-161, EfficientNetB5, EfficientNetB7, InceptionNetV3	Rotterdam EyePACS AIROGS	101 442 (training)	0.9654 (pAUC)

is using only one classifier with an ensemble of inputs – more variants or types of images. An ensemble of machine learning classifiers was used by Tekuabou Koumetio et al. [9] after extracting the features using Alexnet. They used the Bagging classifier, AdaBoosting classifier, Random Forest classifier, and Gradient Boosting Classifier, with best accuracy of 90.3% using the Random Forest classifier. The machine learning model "Majority-Voting-Ensemble" consisting of the Support Vector Machine, Random Forest, and Multilayer Perceptron classifiers was proposed by Patra et al. [17] with accuracy of 99.86% and 88.48%, on REFUGE and ORIGA datasets, respectively. 13 deep learning models altogether with Support Vector Machine classifier were presented in the ensemble approach in [10]. Deep learning segmentation of the optic disc and optic cup, and ensemble Random Forest classifier were tested elsewhere [11]. Mahdi et al. [15] proposed an ensemble of Inception and DenseNet for glaucoma detection on private dataset. VGGnet-16, ResNet-50, and GoogLeNet were used by Deepa et al. [13] to classify glaucomatous fundus images. They used one private dataset and three publicly available datasets – DRISHTI-GS, DRIONS-DB and HRF with 88.96% accuracy on a mixture of public tested datasets. Wang et al. [16] classified glaucoma images from the Rotterdam EyePACS AIROGS dataset to glaucoma referrable or non-referrable using three variants of an ensemble of convolutional neural network classifiers – SeResNext-50, VGG-16, DenseNet-161, EfficientNetB5, EfficientNetB7, and InceptionNetV3, with best pAUC 0.9654. Cho et al. [12] proposed 56 types of the CNN models trained on colour and red-free fundus images, and seven types of image filters to classify images into

three classes – non-glaucoma, early-stage glaucoma, and late-stage glaucoma. Fu et al. [14] presented DE-Net: Disc-aware Ensemble Network that analyses, together as an ensemble, the whole fundus image and optic disc region of the fundus image.

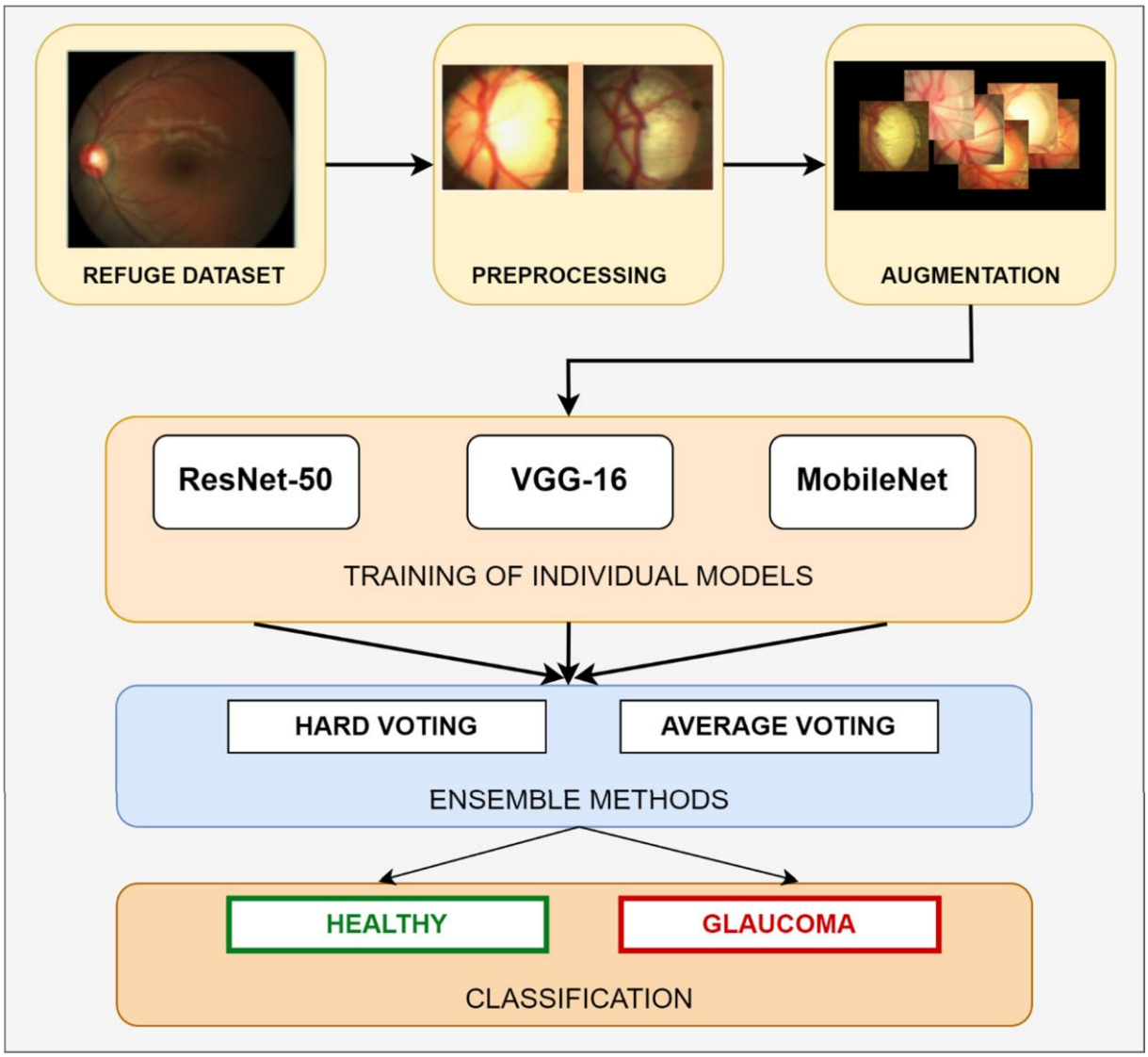

Fig. 1 The scheme of the proposed method

3 Methods

The scheme of our method is listed in Fig. 1.

3.1 REFUGE dataset

The Retinal Fundus Glaucoma Challenge (REFUGE) database [18] was created as part of a competition to stimulate progress in glaucoma detection. This diverse and quality database contains 1200 colour fundus images, divided into a training set of 400, a validation set of 400, and a test set of 400 images. The test set has not been provided with classification labels, so we decided not to use these unlabelled images. Each of 800 images in training and validation set has detailed segmentation annotations of the optic disc and cup. We divided these labelled images into 80%, 10%, and 10% for training, validation, and test set, respectively.

3.2 Preprocessing and data augmentation

Images were resized into 224×224 pixels, Contrast limited adaptive histogram equalization (CLAHE) [19] as a preprocessing was used. After consolidating the dataset into the dataframe we could see that the dataset was imbalanced. We used various augmentation techniques – horizontal and vertical flipping, random rotations up to 45 degrees, and their combination to obtain more images for training. These augmentations were applied to 40% of randomly chosen samples of the dataset. After the data augmentation, we obtained 1280 augmented images, 123 glaucomatous, and 1157 healthy images, additionally.

3.3 Ensemble of deep learning models

We have used a transfer learning approach for our models, the models were pretrained on the ImageNet database [20]. Our ensemble method consisted of ResNet-50 [21], VGG-16 [22], and MobileNet [23] deep learning models.

The output from the base models was first flattened, changing the multi-dimensional feature maps into a format that could be inputted into a traditional dense layer. Following that, we included a dense layer with a 'ReLu' activation function. This layer functions as a hidden layer, giving the model additional computational complexity. Next, we added a dropout layer. This method adds randomness to the model, helping its ability to generalize from the training data. Another dense layer was added, this one with just a single unit and a 'sigmoid' activation function. This final layer effectively reduces the model's predictions to a binary outcome - indicating the presence or absence of glaucoma. The model was compiled with the Adam optimizer [24] selected for its computational efficiency and low requirements. A specific learning rate was set to control how quickly the model learns during the training process. The initial layers of the model were set to non-trainable to preserve the valuable, pre-trained features of the base model while allowing our added layers to learn from the new data. For the loss function, we selected binary cross-entropy, which is often used in binary classification problems like ours.

Several metrics were chosen to monitor during training – binary accuracy, Area Under the Curve (AUC), precision, recall, true positives, true negatives, false positives, and false negatives. These metrics give us a comprehensive view of the model's performance and will also be useful in calculating sensitivity and specificity, two key indicators of a model's diagnostic ability in medical applications.

3.4 Hyperparameters setting

We evaluated various hyperparameters settings using the Keras library, with the best five settings listed in Table 2, with the best results marked in bold. In all three individual models, 128 dense units were included. With optimal hyperparameters, we observed an increase of the AUC score per epoch throughout the fine-tuning process.

We implemented the early stopping to avoid overfitting. We set an early stopping patience of 6. Each of the convolutional neural network models was fine-tuned individually using the RandomSearch function across 50 epochs, where an epoch is one full pass through the entire training dataset. The metric we aimed to optimize during this tuning was the validation Area Under the Curve (AUC).

Table 2 Hyperparameters settings of ResNet- 50, VGG-16, and MobileNet models. The best five models are listed descendingly according to validation AUC, with the best model marked in bold.

Hyperparameters			**AUC**	
dense unit	dropout rate	learning rate	training	validation
		ResNet-50 [21]		
128	**0.3**	**0.01**	**0.987**	**0.948**
64	0.5	0.001	0.965	0.941
64	0.5	0.01	0.939	0.939
128	0.2	0.01	0.996	0.914
128	0.5	0.01	0.977	0.912
		VGG-16 [22]		
128	**0.5**	**0.001**	**0.986**	**0.937**
64	0.2	0.001	0.982	0.936
128	0.5	0.01	0.959	0.924
128	0.2	0.01	0.991	0.883
256	0.5	0.001	0.996	0.880
		MobileNet [23]		
128	**0.2**	**0.001**	**0.999**	**0.951**
256	0.2	0.001	1.000	0.940
64	0.3	0.001	0.993	0.913
64	0.2	0.001	0.959	0.908
128	0.5	0.001	0.995	0.908

3.5 Training of the models

Each model was trained for 200 epochs, each epoch represented a full cycle through the training data, and we divided this data into batches of 32 samples for more efficient processing. To retain only the highest-performing model from the training process, we used a Keras tool known as ModelCheckpoint. Throughout the training process, we kept our focus on the validation AUC as our primary metric. By keeping our focus on validation AUC, we aimed to optimize the model's ability to accurately detect true cases of glaucoma while minimizing false diagnoses. This focus continued to serve us well in the training stage, guiding our models toward better performance in glaucoma classification.

3.6 Ensemble method

In the phase of combining the outputs of our individual models, often referred to as the "ensembling stage", we made use of two distinct voting schemes that fall under the fusion method of ensemble learning. We have used hard voting and average voting ensemble method.

Our work was programmed in Python and TensorFlow framework.

4 Results

The performance of individual models and ensemble methods on the testing dataset is listed in Table 3. Training of each model over the 200 epochs and corresponding AUC is depicted in Fig. 2.

Confusion matrices of all models are depicted in Fig. 3. The ResNet-50 model outperformed two other used models with significantly better results. Weak sensitivity results were observed using VGG- 16 and MobileNet models. We have saved the best AUC models to perform on the testing dataset. Since we haven't employed Early Stopping in our experiments and all the models were trained for a given number of training epochs, overtraining can be observed on the charts in Figure 2. However, since the best result was saved based on the validation set AUC, the results presented should be indicative.

If we compare the results of individual models to ensemble methods, the best accuracy, 0.98, was achieved by Average voting method. On the other hand, the best AUC was achieved by the individual model – ResNet-50.

Table 3 Results of tested models and ensemble methods on the testing dataset

Model	Sensitivity	Specificity	Accuracy	AUC
ResNet-50 [21]	0.85	0.98	0.97	0.92
VGG-16 [22]	0.5	1.0	0.95	0.75
MobileNet [23]	0.6	1.0	0.96	0.80
Hard Voting	0.65	1.0	0.97	0.83
Average Voting	0.75	1.0	0.98	0.88

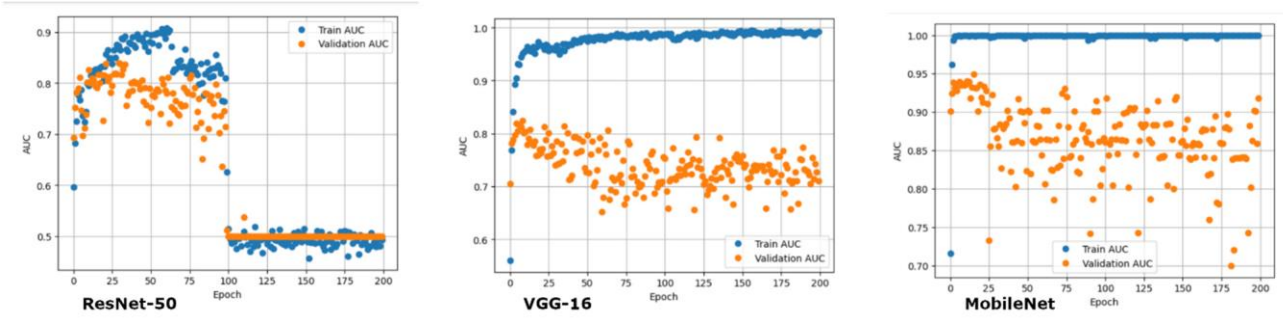

Fig. 2 Training of each model over the 200 epochs and corresponding AUC

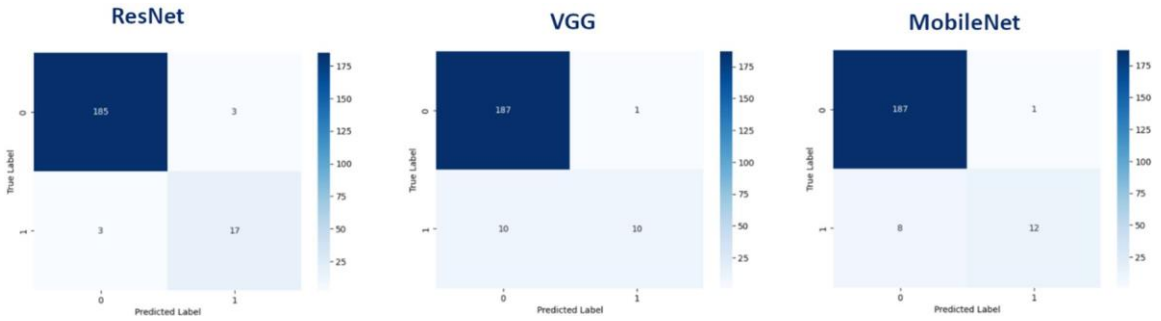

Fig. 3 Confusion matrices of individual models on the testing dataset

Examples of classified optic disc images by the average ensemble method are shown in Figure 4. In the figure, first two optic discs were classified correctly as glaucoma and healthy, the third and fourth disc are classified incorrectly, as healthy but they should be glaucomatous.

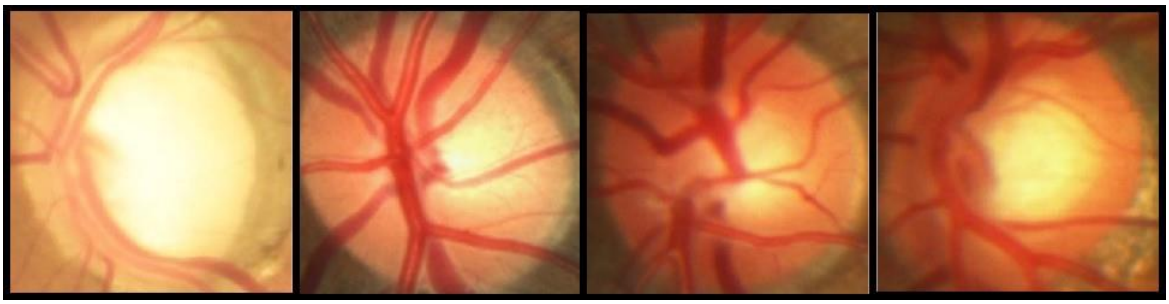

Fig. 4 Examples of classified optic disc images. True positive, true negative, and two images of false negative classification (from left to right).

5 Discussion

We observed that ensemble methods could provide better accuracy than individual models itself. Ensemble methods have better generalization abilities when compared to some individual models and have potential in clinical-screening application. This potential could be enhanced from the clinical point of view when smartphone-suitable deep learning models, such as MobileNet, are used. On the other hand, in our case, one individual model, the ResNet-50, performed better in the AUC metric alone than the ensemble method. This could be due to the weak performance of two other used models, which could affect the final ensemble results.

During our training process, we observed overtraining of the models which could be explained by incorrectly chosen learning rate during training. However, the used learning rate was chosen carefully after hyperparameters testing. We used the best-performing model to test the testing dataset, before the AUC drop because of overtraining.

The results are also influenced by the quality of the used dataset classification labels, as shown in the third and fourth image in the Figure 4. If we compare these false negative findings, we can conclude only the fourth image as falsely classified, based on the opththalmologist opinion. The third image from the left could be also classified as healthy one, and it is not markedly sure in which class it belongs, as evaluated by ophthalmologist.

If we compare our work with other authors, who used ensemble methods for glaucoma screening or diagnosis, only one work evaluated their results on the same dataset [17] as we did. Four of the nine works used a combination of deep learning with machine learning classifiers [9-11, 17], mostly with feature extraction using deep learning models. The other five works were using only deep learning models, from which two of the works were combining different types of images as input for the ensemble method [12, 14]. We can compare our final results with the most recent work by Patra et al. from

2023 [17], because of evaluated results on the same database, REFUGE. Their majority voting ensemble method consisted of individual classifiers with accuracies of 0.9986, 0.9916, and 0.9791 for the Support Vector Machine, Random Forest, and Multilayer Perceptron classifier, respectively. We can deduce, that machine learning classifiers performed much better on this training dataset containing 400 images, compared to MultiLayer Perceptron and deep learning ResNet-50 model which had similar accuracies (0.9791 vs. 0.97). In future work, we want to enlarge our dataset with other public datasets, and also try other deep learning models to improve the accuracy of our ensemble method. Our goal is to test also smartphone-suitable deep learning models which could enhance glaucoma screening possibilities in remote areas.

6 Conclusion

We applied hard and average voting ensemble methods to three deep learning models to classify optic discs from fundus images whether the presence of glaucoma. The best accuracy was achieved using the average voting method, in contrast to the best AUC which was obtained by the ResNet-50 model alone. VGG-16 and MobileNet models performed weakly, and that might negatively affect the overall ensemble method performance.

Acknowledgements

APVV-22-606
This research was supported by APVV grant number APVV-22-606.

VEGA 1/0202/23
This research was supported by grant VEGA 1/0202/23 AIDabiomeDIA – AI in Development of Advanced Biometrics and Medicine Diagnostics.

Internal FEI STU grant
This research was supported by Internal FEI STU grant 2022-23-04 to support young excellent research teams.

References

[1] European Glaucoma Society Terminology and Guidelines for Glaucoma, 4th Edition - Chapter 3: Treatment principles and options Supported by the EGS Foundation: Part 1: Foreword; Introduction; Glossary; Chapter 3 Treatment principles and options. Br J Ophthalmol 2017 Jun;101(6): 130-195. doi: 10.1136/bjophthalmol-2016-EGSguideline.003. PMID: 28559477; PMCID: PMC5583689. https://pubmed.ncbi.nlm.nih.gov/28559477/ (accessed July 25, 2023)

[2] S. Resnikoff, W. Felch, T.-M. Gauthier, and B. Spivey, "The number of ophthalmologists in practice and training worldwide: a growing gap despite more than 200 000 practitioners," *British Journal of Ophthalmology*, vol. 96, no. 6, pp. 783–787, Jun. 2012, doi: 10.1136/bjophthalmol-2011-301378.

[3] J. J. Kanski and B. Bowling, *Clinical Ophthalmology A Systemic Approach*. in 7. Elsevier, 2011.

[4] B. Raju, N. S. D. Raju, J. D. Akkara, and A. Pathengay, "Do it yourself smartphone fundus camera – DIYretCAM," *Indian J Ophthalmol*, vol. 64, no. 9, pp. 663–667, Sep. 2016, doi: 10.4103/0301-4738.194325.

[5] A. Neto, J. Camara, and A. Cunha, "Evaluations of Deep Learning Approaches for Glaucoma Screening Using Retinal Images from Mobile Device," *Sensors (Basel)*, vol. 22, no. 4, p. 1449, Feb. 2022, doi: 10.3390/s22041449.

[6] Y.-K. Chan, Y.-F. Chen, T. Pham, W. Chang, and M.-Y. Hsieh, "Artificial Intelligence in Medical Applications," *Journal of Healthcare Engineering*, vol. 2018, p. e4827875, Jul. 2018, doi: 10.1155/2018/4827875.

[7] C. Zheng, T. V. Johnson, A. Garg, and M. V. Boland, "Artificial intelligence in glaucoma," *Current Opinion in Ophthalmology*, vol. 30, no. 2, p. 97, Mar. 2019, doi: 10.1097/ICU.0000000000000552.

[8] "Ensemble learning: A survey - Sagi - 2018 - WIREs Data Mining and Knowledge Discovery - Wiley Online Library." https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1249 (accessed Jul. 25, 2023).

[9] S. C. Tekouabou Koumetio, E. A. Abdellaoui Alaoui, I. Chabbar, W. Cherif, and H. Silkan, "Using Deep Features Extraction and Ensemble Classifiers to Detect Glaucoma from Fundus Images," in *Emerging Trends in ICT for Sustainable Development*, M. Ben Ahmed, S. Mellouli, L. Braganca, B. Anouar Abdelhakim, and K. A. Bernadetta, Eds., in Advances in Science, Technology & Innovation. Cham: Springer International Publishing, 2021, pp. 63–70. doi: 10.1007/978-3-030-53440-0_8.

[10] P. Elangovan and M. K. Nath, "En-ConvNet: A novel approach for glaucoma detection from color fundus images using ensemble of deep convolutional neural networks," *International Journal of Imaging Systems and Technology*, vol. 32, no. 6, pp. 2034–2048, 2022, doi: 10.1002/ima.22761.

[11] S. Li, Z. Li, L. Guo, and G.-B. Bian, "Glaucoma Detection: Joint Segmentation and Classification Framework via Deep Ensemble Network," in *2020 5th International Conference on Advanced Robotics and Mechatronics (ICARM)*, Dec. 2020, pp. 678–685. doi: 10.1109/ICARM49381.2020.9195312.

[12] H. Cho *et al.*, "Deep Learning Ensemble Method for Classifying Glaucoma Stages Using Fundus Photographs and Convolutional Neural Networks," *Current Eye Research*, vol. 46, no. 10, pp. 1516–1524, Oct. 2021, doi: 10.1080/02713683.2021.1900268.

[13] N. Deepa, S. Esakkirajan, B. Keerthiveena, and S. B. Dhanalakshmi, "Automatic Diagnosis of Glaucoma using Ensemble based Deep Learning Model," in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Mar. 2021, pp. 536–541. doi: 10.1109/ICACCS51430.2021.9441817.

[14] H. Fu *et al.*, "Disc-Aware Ensemble Network for Glaucoma Screening From Fundus Image," *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2493–2501, Nov. 2018, doi: 10.1109/TMI.2018.2837012.

[15] M. M. Mahdi, M. A. Mohammed, H. Al-Chalibi, B. S. Bashar, H. A. Sadeq, and T. M. J. Abbas, "An Ensemble Learning Approach for Glaucoma Detection in Retinal Images," *Majlesi Journal of Electrical Engineering*, vol. 16, no. 4, pp. 117–122, 2022.

[16] E. Wang, A. Durvasula, D. Deng, A. Sivajohan, E. Ho, and K. Lane, "Ensemble Network for Glaucoma Screening in AIROGS Challenge".

[17] A. Patra, A. Nandi, M. Z. Lazarus, and S. Lenka, "An Ensemble Framework for Glaucoma Classification Using Fundus Images," in *Soft Computing: Theories and Applications*, R. Kumar, A. K. Verma, T. K. Sharma, O. P. Verma, and S. Sharma, Eds., in Lecture Notes in Networks and Systems. Singapore: Springer Nature, 2023, pp. 573–588. doi: 10.1007/978-981-19-9858-4_49.

[18] J. I. Orlando *et al.*, "REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs," *Medical Image Analysis*, vol. 59, p. 101570, Jan. 2020, doi: 10.1016/j.media.2019.101570.

[19] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics gems IV*, USA: Academic Press Professional, Inc., 1994, pp. 474–485.

[20] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *arXiv:1409.0575 [cs]*, Jan. 2015, Accessed: Jun. 15, 2021. [Online]. Available: http://arxiv.org/abs/1409.0575

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385 [cs]*, Dec. 2015, Accessed: May 24, 2021. [Online]. Available: http://arxiv.org/abs/1512.03385

[22] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556 [cs]*, Apr. 2015, Accessed: May 24, 2021. [Online]. Available: http://arxiv.org/abs/1409.1556 A. G. Howard *et al.*, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications." arXiv, Apr. 16, 2017. doi: 10.48550/arXiv.1704.04861.

[23] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization." arXiv, Jan. 29, 2017. doi: 10.48550/arXiv.1412.6980.

Received 8 August 2023