Intelligent Systems with Applications 16 (2022) 200140

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

Detection of glaucoma using three-stage training with EfficientNet

I. de Zarzà[a,b,c], J. de Curtò[*,a,b,c], Carlos T. Calafate[c]

[a] Centre for Intelligent Multidimensional Data Analysis, HK Science Park, Shatin, Hong Kong
[b] Studies of Computer Science, Multimedia and Telecommunications, Universitat Oberta de Catalunya, Barcelona
[c] Department of Computer Engineering, Universitat Politècnica de València, València

ARTICLE INFO

Keywords:
Glaucoma
Fundus images
EfficientNet

ABSTRACT

This paper sets forth a methodology that is based on three-stage-training of a state-of-the-art network architecture previously trained on Imagenet, and iteratively finetuned in three steps; freezing first all layers, then retraining a specific number of them and finally training all the architecture from scratch, to achieve a system with high accuracy and reliability. To determine the performance of our technique a dataset consisting of 17.070 color cropped samples of fundus images, and that includes two classes, normal and abnormal, is used. Extensive evaluations using baselines models (VGG16, InceptionV3 and Resnet50) are carried out, in addition to thorough experimentation with the proposed pipeline using variants of EfficientNet and EfficientNetV2. The training procedure is described accurately, putting emphasis on the number of parameters trained, the confusion matrices (with analysis of false positives and false negatives), accuracy, and F1-score obtained at each stage of the proposed methodology. The results achieved show that the intelligent system presented for the task at hand is reliable, presents high precision, its predictions are consistent and the number of parameters needed to train are low compared to other alternatives.

1. Introduction

Intelligent systems based on data-driven techniques have been proposed in recent years to solve a wide variety of tasks with unprecedented level of success. In the case of medical data, applications where methods that rely on computer vision and neural networks are used have proven to be very effective; (Behrad and Saniee Abadeh, 2022; de Curtò et al., 2022; Litjens et al., 2017; Ronneberger et al., 2015), showing performance levels that are similar or even better than human assessment.

In this article, we propose a three-stage training mechanism to design a reliable system to detect and assess glaucoma; (Diaz-Pinto et al., 2019a; 2019b), an irreversible neuro-generative eye disease that, according to the World Health Organization (WHO), affects more than 65 million people around the globe. Early detection and treatment are of utmost importance to prevent loss of visual capacity.

The system introduced achieves state-of-the-art performance on the application under consideration, and the methodology is general enough to be used in other clinical cases or widespread vision applications where learning highly descriptive features from raw pixel intensities is crucial. The design principles take into consideration performance, reliability, statistical significance, platform-aware latency and FLOPS needed to accomplish the task; with the ultimate goal to propose an expert system that could be seamlessly integrated with the clinical equipment (e.g. retinograph) for early diagnostic and treatment.

Our network achieves a mean average percentage F1-score across folds of 96.6 using EfficientNet-B0 (with standard deviation of 3.7) and EfficientNet-B4 (with standard deviation of 2.0), where the best F1 on a given fold is 99 on B0 and 98 on B4. For the case of EfficientNetV2, V2-B3 achieves a mean average F1-score across folds of 95.7 (with standard deviation of 2.3) and V2-S of 95.4 (with standard deviation of 1.6), where the best F1 on a given fold is 98 for both V2-B3 and V2-S. These results significantly outperform the baselines; VGG16 (83.2), InceptionV3 (91.1) and ResNet50 (88.9), and are also clearly better than the state-of-the-art reported results found in the literature, (Diaz-Pinto et al., 2019b).

Code and data used throughout the manuscript is released publicly under the badge initiative on reproducibility by Code Ocean[1]. A detailed notebook addressing all the stages of the methodology, as well as the dataset used, can be found in a runnable capsule environment.

The remainder of this paper is organized as follows: in the next

* Corresponding author.
E-mail addresses: dezarza@uoc.edu (I. de Zarzà), decurto@uoc.edu (J. de Curtò), calafate@disca.upv.es (C.T. Calafate).

[1] Permanent link to reproducible capsule: https://doi.org/10.24433/CO.8342269.v1

https://doi.org/10.1016/j.iswa.2022.200140
Received 5 August 2022; Received in revised form 20 September 2022; Accepted 6 October 2022
Available online 12 October 2022

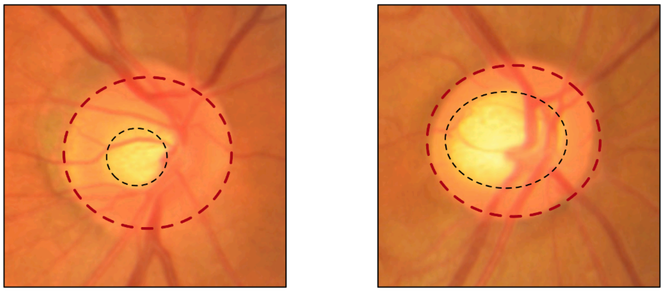


Fig. 1. Example of negative (normal) and positive (abnormal) samples. Highlighted inner circular region corresponds to Optic Cup and outer circular region to Optic Disc. Samples that are glaucomatous (right; with severe pathology) present abnormal size of the Optic Cup respect to normal samples (left). [Source: two random samples from the dataset under study].

section we present an overview and state of the art; then an exhaustive description of the data and the methods is provided, together with visualization, preliminary study, evaluation and selection of the best model with thorough experimentation. Finally, conclusions and further work are discussed.

2. Overview and state of the art

Several approaches to address retinal imaging problems are introduced, showing both traditional techniques based on hand-crafted features and also CNN based methodologies. A brief discussion of each procedure is provided as well as the type of data used in the experimentation. We broaden the analysis by mentioning the state-of-the-art techniques used in our work, that unlike previous publications take carefully into account accuracy and number of parameters.

In Chakravarty and Sivaswamy (2016) they propose a semi-supervised learning framework based on bag of words for early detection of glaucoma. In Geetha Ramani et al. (2012) they assess the pathology by the use of random tree classification, although the experiments are only reported on a dataset of 45 samples. Xiong et al. (2014) details the use of PCA and BAYES classifier. Mitra et al. (2018) proposes the use of CNN to predict bounding boxes with their corresponding class probability and confidence score, where initialization is done using k-means clustering. Wang et al. (2019) goes beyond this approach and applies a pathology-aware feature visualization approach for the diagnostic, where the method relies heavily on Generative Adversarial Networks (GANs). Guo et al. (2020) uses UNet++ in Zhou et al. (2018) to segment the Optic Disc and Optic Cup using feature extraction at several fields of view and then a gradient boosting decision tree to do the screening of glaucoma. Traditional methods have also shown to be effective to tackle related medical imaging problems, (Huang et al., 2021; Yang et al., 2020a; 2020b). For a meticulous analysis of several approaches see Barros et al. (2020), where both a description of hand-crafted methods and techniques based on deep convolutional networks (Maqsood et al., 2021; Rajinikanth et al., 2021; Wong et al., 2012) are presented. However, although most methods provide a well-thought effective methodology to address the problem, the majority of them have a shortfall on the used data, as are tested on very small datasets with limited statistical significance.

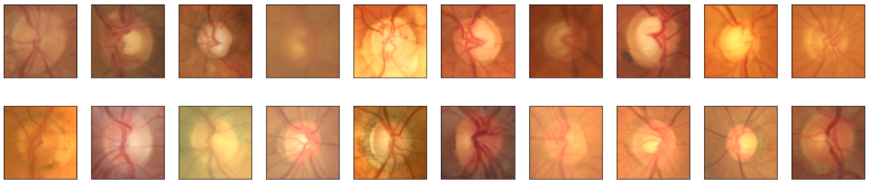


Fig. 2. Visual exploration of the samples of the pathology glaucoma; first row: positive, second row: negative.

Table 1
Statistics of the dataset consisting on 17.070 fundus images with positive (P: abnormal) and negative (N: normal) samples. The data is distributed into 10 folds (0 to 9) of 1.707 samples each with corresponding train, validation and test.

		Fold										
		0	1	2	3	4	5	6	7	8	9	Total
Train	(N)	754	740	739	743	746	758	754	642	748	733	7452
	(P)	625	639	640	636	633	621	625	737	631	646	6338
Validation	(N)	83	88	83	85	81	71	84	82	80	82	819
	(P)	71	66	71	69	73	83	70	72	74	72	721
Test	(N)	82	91	97	91	92	90	81	100	91	104	919
	(P)	92	83	77	83	82	84	93	74	83	70	821
Total		1707	1707	1707	1707	1707	1707	1707	1707	1707	1707	17070

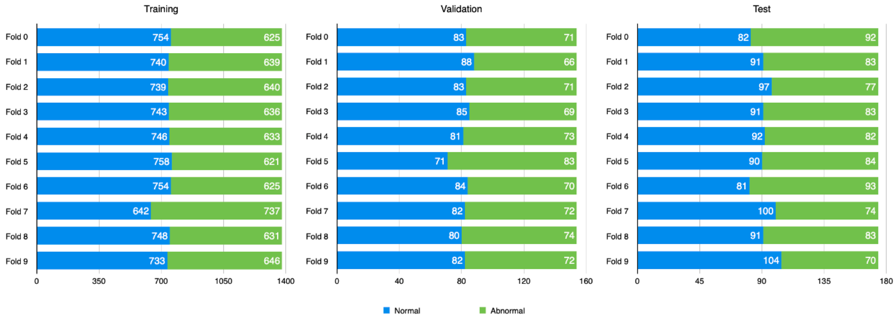


Fig. 3. Statistics of the dataset using a bar plot for sets of training, validation and testing.

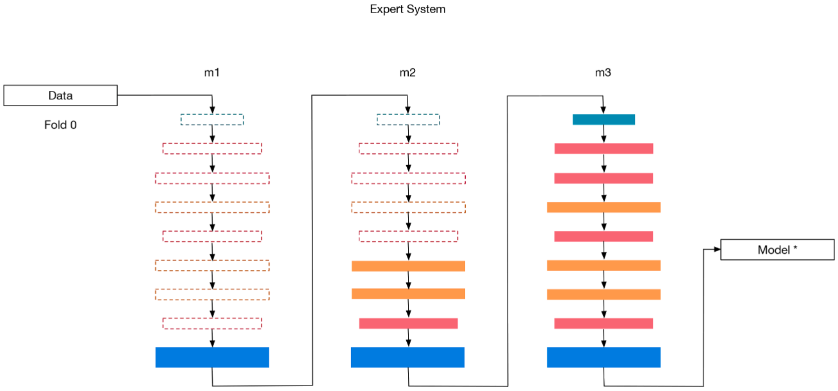


Fig. 4. Visual description of the proposed three-stage training system in one fold of the data. Transfer learning from ImageNet is put in place. Color layers are re-trained. In particular, weights from m1 are used to initialize the network when training m2, which unfreezes a given number of layers from the full model (in our application 20, keeping layers BatchNorm untrained). Afterwards, weights from m2 are used to initialize the network when training m3, which retrains the whole architecture. Finally, in the evaluation stage, the weights obtained (Model*) are then fed into 10-fold crossvalidation to retrain the network for each fold, and select the best model according to F1-score. The procedure is robust against hyperparameter choices.

Table 2
EfficientNet-B0, baseline network. Each row describes a stage c with $\widehat{L}_c$ layers, with input resolution $\langle \widehat{H}_c, \widehat{W}_c \rangle$ and output channels $\widehat{C}_c$.

Stage c	Operator $\widehat{\mathcal{F}}_c$	Resolution $\widehat{H}_c \times \widehat{W}_c$	**# Channels** $\widehat{C}_c$	**# Layers** $\widehat{L}_c$
1	Convn3x3	224 × 224	32	1
2	MBConvn1, k3x3	112 × 112	16	1
3	MBConvn6, k3x3	112 × 112	24	2
4	MBConvn6, k5x5	56 × 56	40	2
5	MBConvn6, k3x3	28 × 28	80	3
6	MBConvn6, k5x5	14 × 14	112	3
7	MBConvn6, k5x5	14 × 14	192	4
8	MBConvn6, k3x3	7 × 7	320	1
9	Convn1x1 & Pooling & FC	7 × 7	1280	1

Table 3
EfficientNetV2-S, example architecture. Extension to EfficientNet using both MB and Fused-MB Convolutions. Each row describes a stage c with $\widehat{L}_c$ layers, with given stride and output channels $\widehat{C}_c$.

Stage c	Operator $\widehat{\mathcal{F}}_c$	Stride	**# Channels** $\widehat{C}_c$	**# Layers** $\widehat{L}_c$
0	Convn3x3	2	24	1
1	Fused-MBConvn1, k3x3	1	24	2
2	Fused-MBConvn4, k3x3	2	48	4
3	Fused-MBConvn4, k3x3	2	64	4
4	MBConvn4, k3x3, SE0.25	2	128	6
5	MBConvn6, k3x3, SE0.25	1	160	9
6	MBConvn6, k3x3, SE0.25	2	256	15
7	Convn1x1 & Pooling & FC	-	1280	1

Table 4
Three-stage training system for several model baselines. Accuracy in Fold 0.

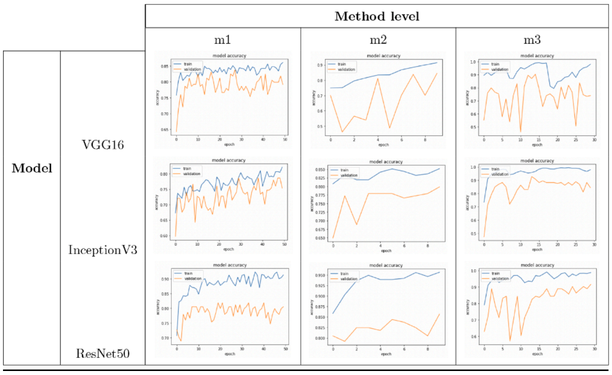
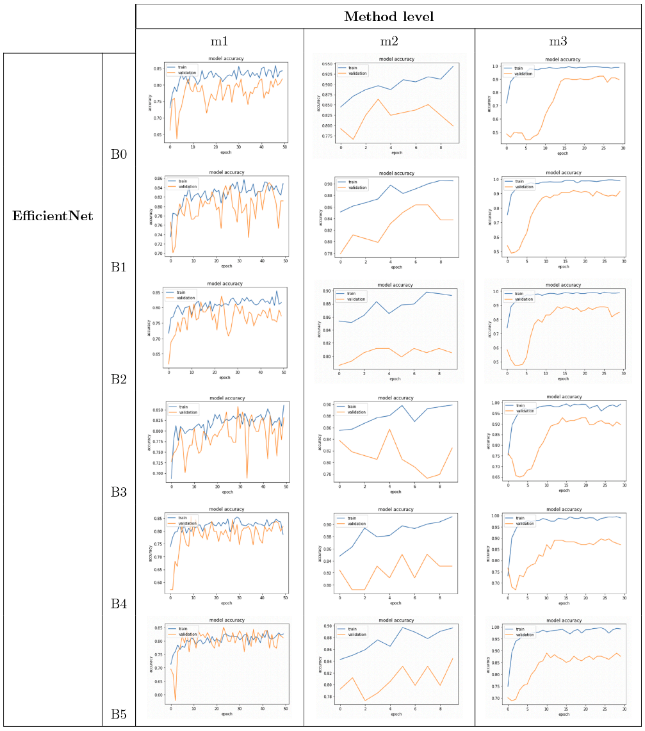


Table 5
Three-stage training system for several variants of EfficientNet. Accuracy in Fold 0.



The architecture proposed is built on EfficientNet in Tan and Le (2019) and EfficientNetV2 in Tan and Le (2021), using a three-stage training mechanism that broadens the finetuning steps proposed in Diaz-Pinto et al. (2019b). These architectures are built using Neural Architecture Search (NAS), (Pham et al., 2018; Zoph and Le, 2017; Zoph et al., 2018), in particular EfficientNet uses the AutoML MNAS framework presented in Tan et al. (2019) that optimizes the networks for accuracy and efficiency (FLOPS) and is based on previous work in Sandler et al. (2018) and Tan et al. (2019), but with a larger base model. Evaluation of the family of models is done from B0 to B5 in the case of EfficientNet, and from B0 to B3, S and M in EfficientNetV2. We use transfer learning from ImageNet to the particular application under study and see that the models achieve high accuracy with a reduced number of training parameters, compared to other state-of-the-art methodologies. Concomitant approaches in retinal image classification show the adequacy of the family of models EfficientNet for the given task, (Gupta et al., 2022a; 2022b; Islam et al., 2022; Jaiswal et al., 2021; Nawaz et al., 2022; Wang et al., 2020).

Table 6
Three-stage training system for several variants of EfficientNetV2. Accuracy in Fold 0.

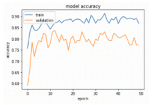
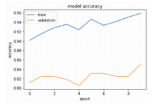
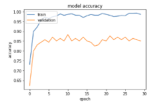
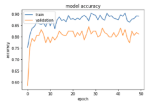
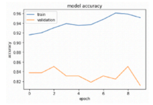
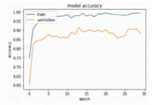
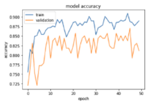
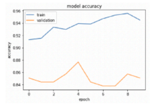
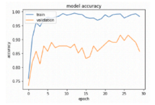
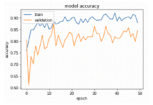
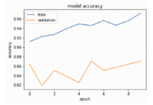
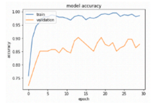
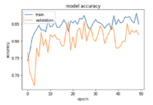
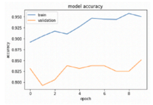
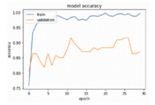
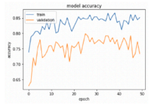
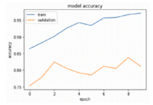
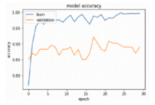
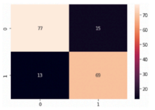
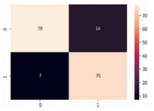
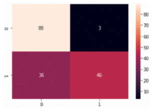
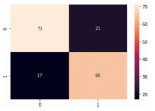
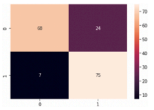
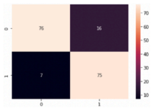
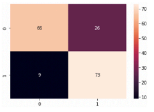
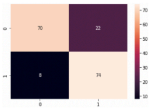
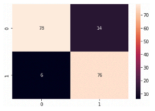
		Method level		
		m1	m2	m3
EfficientNetV2	B0			
	B1			
	B2			
	B3			
	S			
	M			

Table 7
Three-stage training system. Confusion Matrix Baseline Models in Fold 0. VGG16, InceptionV3 and ResNet50.

		Method level		
		m1	m2	m3
Model	VGG16			
	InceptionV3			
	ResNet50			

3. Data and methods

The dataset under consideration consists on 17.070 fundus images, which are digitalized photographs of the posterior part of the eye, with positive (abnormal) and negative (normal) samples of the pathology. The data is divided into 10 folds of 1.707 instances, each one with its corresponding sets of training, validation and testing. The sets are relatively balanced to reduce the number of false negatives. These samples are obtained using retinography, and thus their characteristics in terms of illumination and intensity are very particular and relatively homogeneous among all instances; such aspect is central for the correct detection of the samples. For this reason, using the raw pixels without normalization confers the network with significantly better generalization than when using min-max normalization, or normalization with standard deviation, as these types of preprocessing cause loss of information. This observation is very important as any type of non-linear transformation that affects or alters the brightness of the samples can severely degrade the performance of such a system; this can hold also when dealing with other medical data where image intensity is crucial.

Table 8
Three-stage training system. Confusion Matrix EfficientNet in Fold 0.

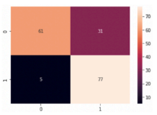
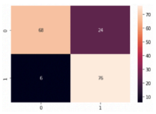
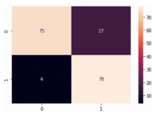
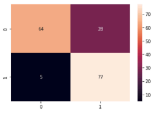
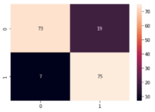
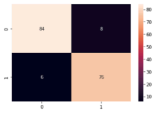
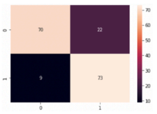
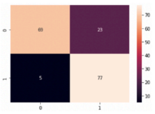
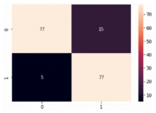
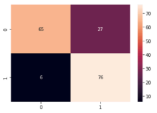
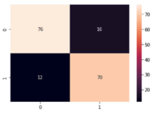
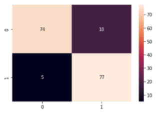
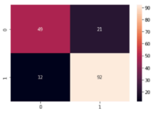
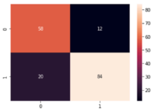
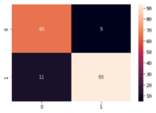
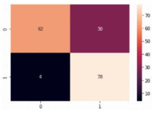
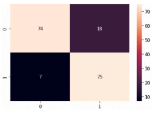
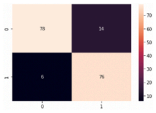
		Method level		
		m1	m2	m3
EfficientNet	B0			
	B1			
	B2			
	B3			
	B4			
	B5			

Table 9
Three-stage training system. Confusion Matrix EfficientNetV2 in Fold 0.

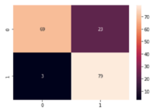
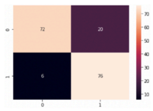
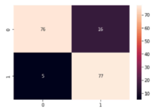
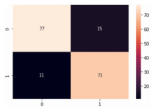
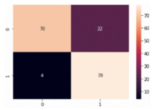
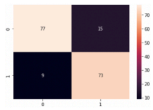
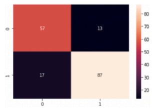
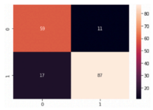
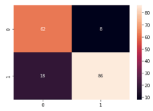
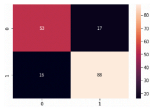
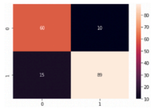
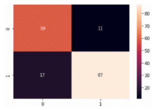
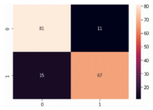
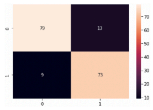
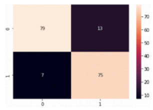
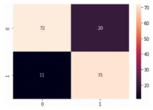
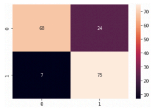
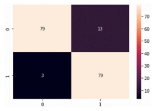
		Method level		
		m1	m2	m3
EfficientNetV2	B0			
	B1			
	B2			
	B3			
	S			
	M			

Table 10
Three-stage training system for several baseline models. F1-score, number of trainable parameters and number of non-trainable parameters in Fold 0.

			Method level		
			m1	m2	m3
Model	VGG16	F1-score (%)	84	88	76
		# trainable parameters	2.050	14.678.018	14.716.738
		# non-trainable parameters	14.715.712	39.744	1.024
	InceptionV3	F1-score (%)	78	82	87
		# trainable parameters	8.194	401.410	21.776.546
		# non-trainable parameters	21.806.880	21.413.664	38.528
	ResNet50	F1-score (%)	80	83	89
		# trainable parameters	8.194	5.518.338	23.542.786
		# non-trainable parameters	23.591.808	18.081.664	57.216

Table 11
Three-stage training system for several variants of EfficientNet. F1-score, number of trainable parameters and number of non-trainable parameters in Fold 0.

			Method level		
			m1	m2	m3
EfficientNet	B0	F1-score (%)	79	83	88
		# trainable parameters	5.122	1.126.706	4.012.670
		# non-trainable parameters	4.052.131	2.930.547	44.583
	B1	F1-score (%)	81	85	92
		# trainable parameters	5.122	1.355.602	6.518.306
		# non-trainable parameters	6.577.799	5.227.319	64.615
	B2	F1-score (%)	81	85	92
		# trainable parameters	5.634	1.637.594	7.706.628
		# non-trainable parameters	7.771.385	6.139.425	70.391
	B3	F1-score (%)	81	84	87
		# trainable parameters	6.146	1.946.210	10.702.378
		# non-trainable parameters	10.786.607	8.846.543	90.375
	B4	F1-score (%)	80	81	91
		# trainable parameters	7.170	2.643.314	17.555.786
		# non-trainable parameters	17.677.407	15.041.263	128.791
	B5	F1-score (%)	80	86	89
		# trainable parameters	8.194	3.446.914	28.348.978
		# non-trainable parameters	28.517.623	25.078.903	176.839

Table 12
Three-stage training system for several variants of EfficientNetV2. F1-score, number of trainable parameters and number of non-trainable parameters in Fold 0.

			Method level		
			m1	m2	m3
EfficientNetV2	B0	F1-score (%)	85	85	88
		# trainable parameters	5.122	594.226	6.865.174
		# non-trainable parameters	6.933.684	6.344.580	73.632
	B1	F1-score (%)	85	85	86
		# trainable parameters	5.122	594.226	6.865.174
		# non-trainable parameters	6.933.684	6.344.580	73.632
	B2	F1-score (%)	82	83	85
		# trainable parameters	5.634	700.406	8.692.720
		# non-trainable parameters	8.772.190	8.077.418	85.104
	B3	F1-score (%)	80	85	83
		# trainable parameters	6.146	860.892	12.827.552
		# non-trainable parameters	12.933.694	12.078.948	112.288
	S	F1-score (%)	85	87	88
		# trainable parameters	5.122	938.050	20.182.610
		# non-trainable parameters	20.333.920	19.400.992	156.432
	M	F1-score (%)	82	82	91
		# trainable parameters	5.122	3.050.626	52.863.478
		# non-trainable parameters	53.152.948	50.107.444	294.592

3.1. Fundus images

The samples are cropped to improve the sensitivity of the detector. The disease is characterized by an abnormal size of the Optic Cup, with respect to the Optic Disc, see Fig. 1. This is the reason why many earlier approaches were based on the Cup/Disc Ratio (CDR). As our approach is data driven, there is no need to use handcrafted intermediate features as feature selection. A random subset of the data is shown in Fig. 2, as well as, detailed statistics in Table 1 and Fig. 3.

3.2. Preliminary study and methodology

The preliminary study focuses on Fold 0 to set forth a design methodology that will serve as guiding principle of the manuscript.

The methodology under study proposes a three-stage training system (see Fig. 4 for a visual description) that consists on the following procedure in only one Fold of the data:

1. Start from a model trained on Imagenet, and only re-train the last added layers (GlobalAveragePooling2D, BatchNorm, Dropout and Fully Connected) of the system.

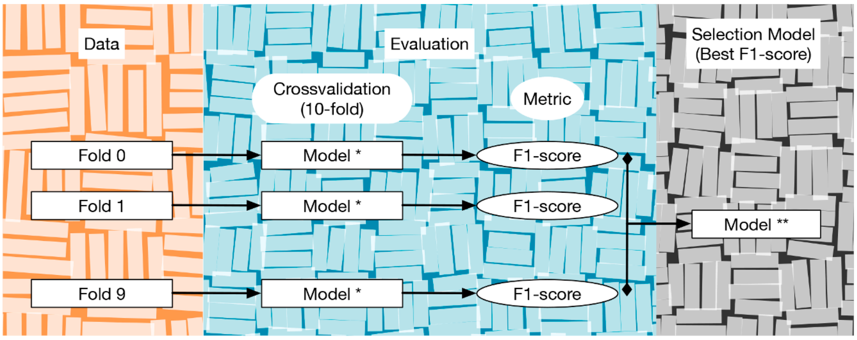


Fig. 5. Visual description of the evaluation. The weights obtained in the preliminary stage using Fold 0 (Model*: obtained from m3) are then used in 10-fold crossvalidation to retrain the network for each fold, and select the best weights of the model (Model **) according to F1-score.

Table 13
Evaluation of the F1-score: baseline models of the method consisting on VGG16, InceptionV3 and ResNet50. Thorough testing across folds with mean and standard deviation for the F1-score of all models under evaluation.

		Fold										Statistics	
		0	1	2	3	4	5	6	7	8	9	Mean	stdev
Baseline models	VGG16	83	88	89	83	95	93	60	90	86	65	83,2	1,1
	InceptionV3	91	94	86	92	93	90	91	94	89	92	91,1	2,4
	ResNet50	88	88	82	87	92	85	92	88	91	93	88,9	3,3

Table 14
Evaluation of the F1-score (%): methods based on EfficientNet. Thorough testing across folds with mean and standard deviation for the F1-score of all models under evaluation.

		Fold										Statistics	
		0	1	2	3	4	5	6	7	8	9	Mean	stdev
EfficientNet	B0	99	99	97	86	97	95	98	98	98	87	96,6	3,7
	B1	92	99	98	94	98	98	97	98	94	93	95,9	2,4
	B2	89	100	96	97	92	98	97	97	98	96	96,1	2,9
	B3	89	99	98	94	98	92	95	98	96	96	95,5	3,0
	B4	97	98	97	98	98	98	96	97	98	91	96,6	2,0
	B5	91	98	97	94	94	96	97	98	94	99	95,8	2,3

Table 15
Evaluation of the F1-score (%): methods based on EfficientNetV2. Thorough testing across folds with mean and standard deviation for the F1-score of all models under evaluation.

		Fold										Statistics	
		0	1	2	3	4	5	6	7	8	9	Mean	stdev
EfficientNetV2	B0	86	95	96	97	95	93	94	96	93	97	94,3	3,3
	B1	88	95	92	96	96	93	96	96	95	95	94,2	2,5
	B2	94	99	91	95	94	95	95	98	95	87	94,5	3,1
	B3	95	98	96	96	96	96	97	98	95	89	95,7	2,3
	S	93	97	95	97	95	97	97	98	93	93	95,4	1,6
	M	88	97	92	89	92	94	91	98	93	91	92,4	3,0

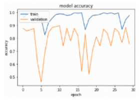
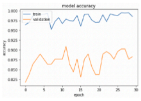
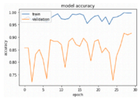
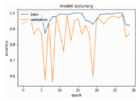
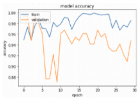
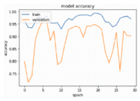
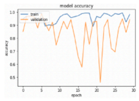
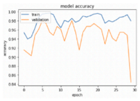
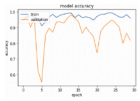
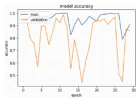
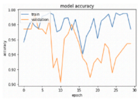
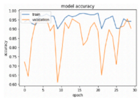
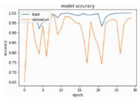
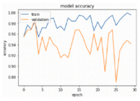
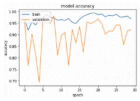
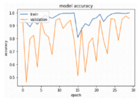
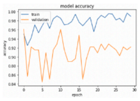
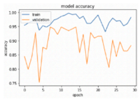
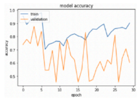
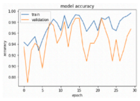
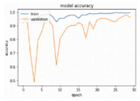
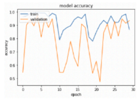
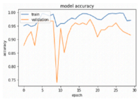
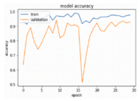
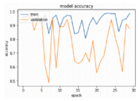
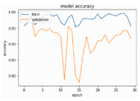
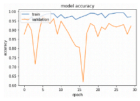
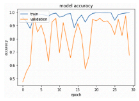
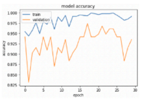
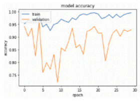
2. Use the weights from the previous stage to initialize a model that unfreezes a number of layers of the previous model (excluding BatchNorm), and retrain the system.
3. Use the weights from the previous iteration and retrain the whole network. Evaluate the classification report based on both the F1-score and confusion matrix to select the best hyperparameters.

Experimentation is based on baseline models (VGG16 in Simonyan and Zisserman (2015), InceptionV3 in Szegedy et al. (2016) and ResNet50 in He et al. (2016)) and then extended to variants of EfficientNet; Tan and Le (2019), and EfficientNetV2; Tan and Le (2021).

EfficientNet-B0 base model consists on the following layers, see Table 2. In this particular example, model m1 freezes all layers from stages 1 to 8, and retrains only layers corresponding to stage 9. Model m2 starts from the learned weights on m1 and retrains a subset of layers going backwards, in our case 20, while keeping BatchNorm layers untrained. Finally model m3 starts from the weights of m2 and retrains the whole network. The network proceeds in the same way with the case of variants of EfficientNetV2 (see Table 3 for a description of the architecture) and model baselines (VGG16, InceptionV3 and ResNet50).

The family of models EfficientNet and EfficientNetV2 is a compositional stack of modules MB and Fused-MB Convolutions (denoted MBConvn and Fused-MBConvn). These modules consist on the following inner operators:

Table 16
Evaluation on several model baselines (VGG16, InceptionV3 and ResNet50). Accuracy across folds (from 0 to 9).

	Baseline models		
Fold	VGG16	InceptionV3	ResNet50
0			
1			
2			
3			
4			
5			
6			
7			
8			
9			

- **MBConvn**: a 1 × 1 convolution, followed by a depthwise 3 ×3 convolution, a SE module in Hu et al. (2018), and finally another 1 × 1 convolution.
- **Fused-MBConvn**: a 3 × 3 convolution, followed by a SE module and finally a 1 × 1 convolution.

Worthy of mention is the fact that EfficientNet-B0 achieves state-of-the-art performance while keeping the number of parameters to train bounded to the same levels as ResNet50.

Tables 4, 5 and 6 show the accuracy in validation and training for the given three-stage training mechanism: m1 corresponding to the first stage where only the last layers are trained, m2 to the second stage where a number of layers are unfrozen, and m3 to the third stage where the whole network is retrained.

Confusion matrices of the corresponding models are shown in Tables 7, 8 and 9, where we can see that there is a clear performance increase due to the three-stage procedure, causing the number of false negatives to be drastically reduced. Although for this particular task we consider F1-score as the comparison metric, confusion matrices allow for the computation of other measures such as error-rate, accuracy, specificity, sensitivity, and precision.

Numerical progression of F1-score can be observed in Tables 10, 11 and 12, where the number of trainable and non-trainable parameters are reported for each method level under study.

Regarding the confusion matrices for the three-stage training system, we can observe that the expected behavior in terms of incorrect cases is having a higher number of false positives than false negatives. This is appropriate for designing a system to detect glaucoma, as the principle is to be able to always detect the disease if it is present, as the pathology is irreversible and early treatment can considerably improve the condition of the subject.

The manuscript uses the F1-score (higher is better), which is calculated as the harmonic mean between precision and recall, to choose among models. We can observe the increase in performance that the training in three steps confers to the design of the system, very much irrespective of the hyperparameters chosen (learning rate, number of epochs and optimizer).

In addition, transfer learning from Imagenet allows us to rapidly fine-tune the architecture in three stages, achieving high accuracy with limited training time.

The system is implemented using Keras with the following hyper-parameters:

- m1 (lr $= 1e-2$, dropout $= 0.2$, epochs $= 50$)
- m2 (lr $= 1e-4$, epochs $= 10$)
- m3 (lr $= 1e-4$, epochs $= 30$)

where 'adam' is the choice of optimizer and the GPU used in the experiments is a Tesla V100 SXM2 (16 GB).

Once the desired model is obtained in Fold 0, we pursue a thorough testing across folds (10-fold crossvalidation) to choose the weights that give better accuracy on a given test subset, see Fig. 5. In particular, we evaluate the mean and the standard deviation to determine statistical significance of the result.

For the evaluation, we perform 10-fold crossvalidation loading the weights from m3 and retraining on each fold with epochs $= 30$.

3.3. Evaluation and discussion

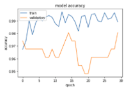
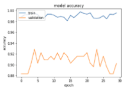
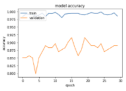
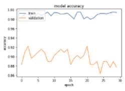
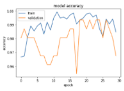
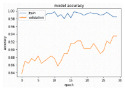
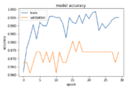
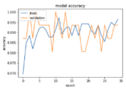
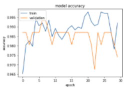
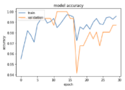
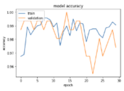
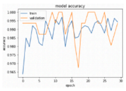
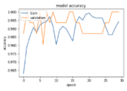
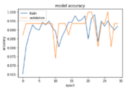
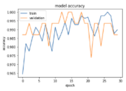
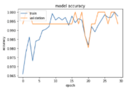
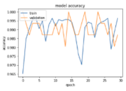
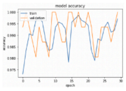
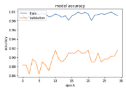
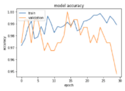
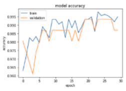
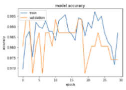
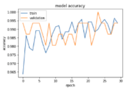
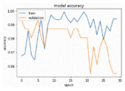
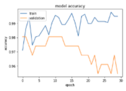
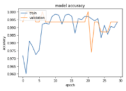
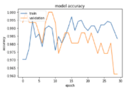
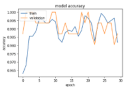
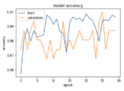
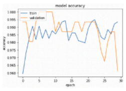
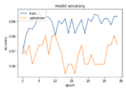
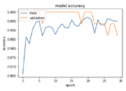
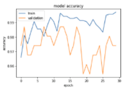
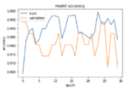
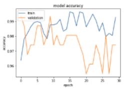
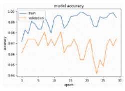
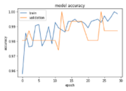
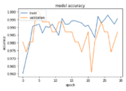
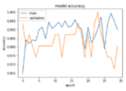
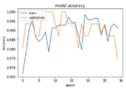
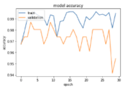
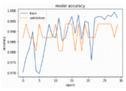
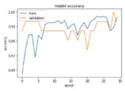
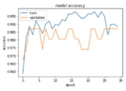
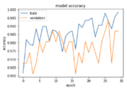
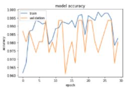
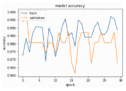
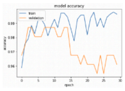
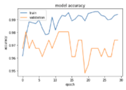
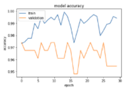
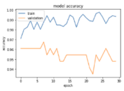
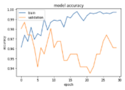
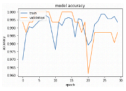
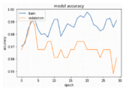
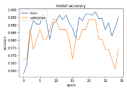
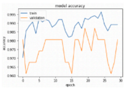
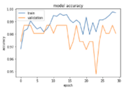
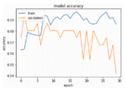
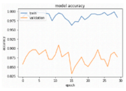
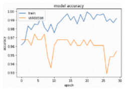
The manuscript builds on VGG16, InceptionV3 and ResNet50 as baseline models of the methodology, and then propose to use variants of EfficientNet to achieve state-of-the-art performance.

Extensive evaluation of every model across folds is performed. Tables 13, 14 and 15 show the F1-score across all folds of the dataset evaluating the method under consideration using as initial weights the corresponding weights of m3, that is, the result of the three-stage training in one fold, for each given model. Best results are highlighted, showing the statistical significance of the outputs by computing the mean and standard deviation along the folds.

Plots of accuracy of every model on the sets of training and validation for each fold are shown in Tables 16, 17 and 18 in order to visualize the level of generalization of the architecture.

The three-stage system presented, including variants of both EfficientNet and EfficientNetV2, considerably outperforms the given baselines (VGG16, InceptionV3 and ResNet50), which are similar in scope to the models reported in Diaz-Pinto et al. (2019b) but using the three-stage training introduced in the manuscript. In the case of the baseline models, InceptionV3 has clearly the highest mean F1-score (91.1) compared to VGG16 (83.2) and ResNet50 (88.9). Although InceptionV3 and ResNet50 show comparative performance in terms of number of trained parameters and overall accuracy achieved, the first is more effective with the problem at hand considering that we are dealing with a dataset in the order of the thousands. Should the data to train be increased, it would be expected that ResNet50 achieves slightly better performance due to its better handling of the gradient backpropagating through the layers.

Table 17
Evaluation on several variants of EfficientNet (B0-5). Accuracy across folds (from 0 to 9).

	EfficientNet					
Fold	B0	B1	B2	B3	B4	B5
0						
1						
2						
3						
4						
5						
6						
7						
8						
9						

EfficientNet models perform slightly better in terms of F1-score than EfficientNetV2, although variants of EfficientNetV2 show a better standard deviation across folds; being B2 the model that achieves higher accuracy on a given fold (100 in Fold 1), and B0 and B4 the ones that achieve best mean F1 across folds (96.6), where B4 has the lowest standard deviation (2.0); thus, a better generalization is achieved since the results are more consistent. In the case of EfficientNetV2, V2-B2 achieves the highest accuracy on a given fold (99 in Fold 1), while V2-B3 is the model that gets best mean F1 score across folds (95.7), and model S is the more consistent model according to the standard deviation of the F1-score (1.6). Error bars with mean and standard deviation are showed in Fig. 6.

Accuracy plots across folds show the considerably good ability to generalize of each model, showing the corresponding curves for the sets of training and validation.

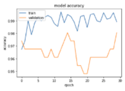
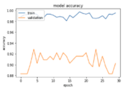
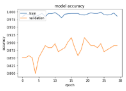
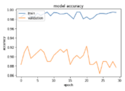
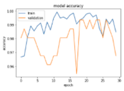
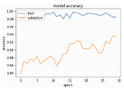
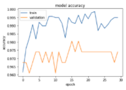
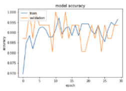
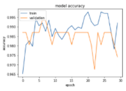
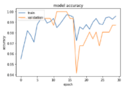
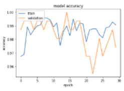
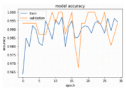
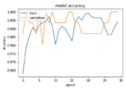
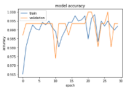
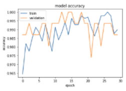
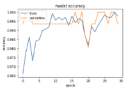
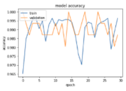
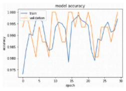
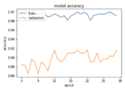
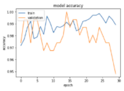
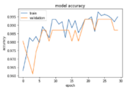
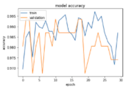
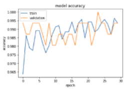
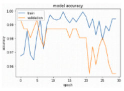
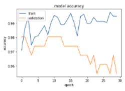
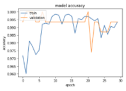
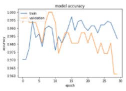
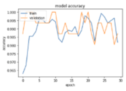
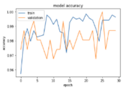
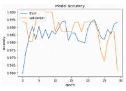
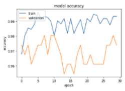
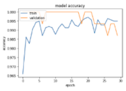
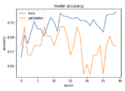
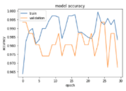
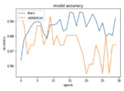
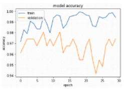
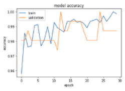
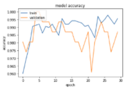
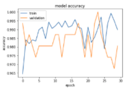
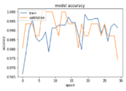
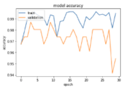
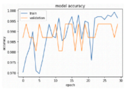
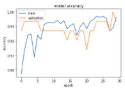
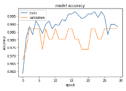
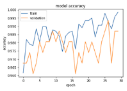
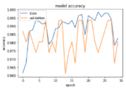
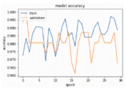
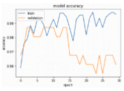
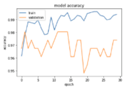
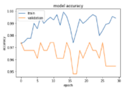
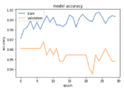
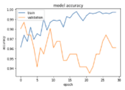
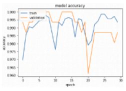
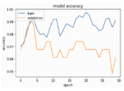
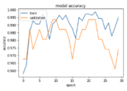
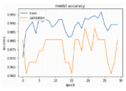
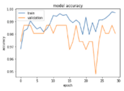
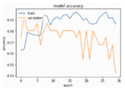
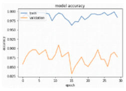
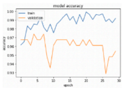
The methodology to tackle the problem in many train stages of the same architecture presents a robust behavior with state-of-the-art performance. Limitations of the technique mainly are due to EfficientNet and EfficientNetV2, where we inherit the necessity to train a large number of parameters, that is clearly less than other state-of-the-art CNN architectures, as the networks are found using NAS optimizing for overall FLOPS, but still very high compared to traditional hand-crafted methodologies where the number of parameters to learn is very low.

4. Conclusions and further work

In this work, an intelligent system to automatically detect glaucoma is presented. The methodology is based on a three-stage training procedure based on variants of EfficientNet, a recently proposed family of architectures found using NAS that achieves compelling accuracy on Imagenet, achieving consistent results that outperform the baseline methods. Transfer Learning from Imagenet to the given application under study is employed. The training mechanism applied bestows the system with robustness against hyperparameter choices. We use a dataset consisting of 17.070 fundus images, a considerable size compared to the number of samples used in other recent works, and where the sets used for training, validation and testing are well balanced; such fact confers the obtained models with a low number of false negatives, which is clearly desirable given the gravity and irreversibility of the pathology. Extensive evaluations are reported at each stage of the described procedure under study, as well as, visual interpretation of the results for the sets of training and validation. The F1-score in the test set is used as the target score metric to choose among models, along with a classification report and confusion matrix for each model in the preliminary stage. The proposed system is reliable, highly-accurate, consistent and resource-efficient.

The methodology achieves a mean average percentage F1-score across folds of 96.6 using EfficientNet-B0 (with standard deviation of 3.7) and EfficientNet-B4 (with standard deviation of 2.0), where the best F1 on a given fold is 99 on B0 and 98 on B4. For the case of

Table 18
Evaluation on several variants of EfficientNetV2 (B0-3, S and M). Accuracy across folds (from 0 to 9).

	EfficientNet					
Fold	B0	B1	B2	B3	B4	B5
0						
1						
2						
3						
4						
5						
6						
7						
8						
9						

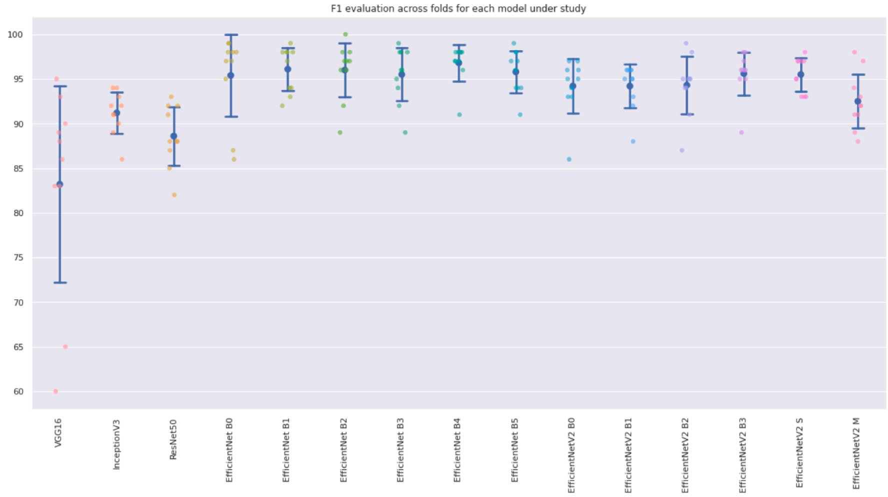


Fig. 6. F1 evaluation across folds for each model under study using the three-stage training procedure. Error bars with mean and standard deviation for each model are depicted. All architectures based on EfficientNet and EfficientNetV2 outperform the baseline methods (VGG16, InceptionV3 and ResNet50) being EfficientNet B4 and EfficientNetV2 S the best performing techniques.

EfficientNetV2, V2-B3 achieves a mean average F1-score across folds of 95.7 (with standard deviation of 2.3) and V2-S of 95.4 (with standard deviation of 1.6), where the best F1 on a given fold is 98 for both V2-B3 and V2-S. These results significantly outperform the baselines; VGG16 (83.2), InceptionV3 (91.1) and ResNet50 (88.9), and are also clearly better than the state-of-the-art reported results found in the literature, Diaz-Pinto et al. (2019b).

The three stage-training mechanism using variants of EfficientNet and EfficientNetV2 proposed, although targeted for the particular application of detecting the pathology of glaucoma, achieves a superior classification baseline to use in other clinical conditions, or in the more general case in any vision application where extracting features from raw pixel intensities can play an important role. Indeed, visual sensors are ubiquitous in many applications, such as self-driving cars or Unmanned Aerial Vehicles, where the use of transfer learning, and subsequent freeze, training and finetuning has proven to be very effective; therefore, the system proposed could be further integrated into the detection pipeline of such a system, for instance for lane detection in a self-driving vehicle, or for target recognition in drones.

CRediT authorship contribution statement

I. de Zarzà: Methodology, Validation, Investigation, Visualization, Software. **J. de Curtò:** Conceptualization, Visualization, Software, Writing – original draft. **Carlos T. Calafate:** Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the HK Innovation and Technology Commission (InnoHK Project CIMDA), the HK Research Grants Council (Project CityU 11204821) and City University of Hong Kong (Project 9610034). We acknowledge the support of Universitat Politècnica de València; R&D project PID2021-122580NB-I00, funded by MCIN/AEI/10.13039/501100011033 and ERDF.

References

Barros, D., Moura, J., Freire, C., Taleb, A., Valentim, R., & Morais, P. (2020). Machine learning applied to retinal image processing for glaucoma detection: Review and perspective. *BioMedical Engineering OnLine, 19*. https://doi.org/10.1186/s12938-020-00767-2

Behrad, F., & Saniee Abadeh, M. (2022). An overview of deep learning methods for multimodal medical data mining. *Expert Systems with Applications, 200*, 117006. https://doi.org/10.1016/j.eswa.2022.117006

Chakravarty, A., & Sivaswamy, J. (2016). Glaucoma classification with a fusion of segmentation and image-based features. *2016 IEEE 13th international symposium on biomedical imaging (ISBI)* (pp. 689–692).10.1109/ISBI.2016.7493360

de Curtò, J., de Zarzà, I., Yan, H., & Calafate, C. T. (2022). On the applicability of the hadamard as an input modulator for problems of classification. *Software Impacts, 13*, 100325. https://doi.org/10.1016/j.simpa.2022.100325

Diaz-Pinto, A., Colomer, A., Naranjo, V., Morales, S., Xu, Y., & Frangi, A. F. (2019a). Retinal image synthesis and semi-supervised learning for glaucoma assessment. *IEEE Transactions on Medical Imaging, 38*, 2211–2218. https://doi.org/10.1109/TMI.2019.2903434

Diaz-Pinto, A., Morales, S., Naranjo, V., Kohler, T., Mossi, J. M., & Navea, A. (2019b). CNNs for automatic glaucoma assessment using fundus images: An extensive validation. *BioMedical Engineering OnLine*.

Geetha Ramani, R., Balasubramanian, L., & Jacob, S. G. (2012). Automatic prediction of diabetic retinopathy and glaucoma through retinal image analysis and data mining techniques. *2012 international conference on machine vision and image processing (MVIP)* (pp. 149–152). https://doi.org/10.1109/MVIP.2012.6428782

Guo, F., Li, W., Tang, J., Zou, B., & Fan, Z. (2020). Automated glaucoma screening method based on image segmentation and feature extraction. *Medical and Biological Engineering and Computing, 58*. https://doi.org/10.1007/s11517-020-02237-2

Gupta, I. K., Choubey, A., & Choubey, S. (2022a). Mayfly optimization with deep learning enabled retinal fundus image classification model. *Computers and Electrical Engineering, 102*, 108176. https://doi.org/10.1016/j.compeleceng.2022.108176

Gupta, N., Garg, H., & Agarwal, R. (2022b). A robust framework for glaucoma detection using clahe and efficientnet. *The Visual computer, 38*, 2315–2328.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *CVPR*.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *CVPR*.

Huang, M., Feng, C., Li, W., & Zhao, D. (2021). Vessel enhancement using multi-scale space-intensity domain fusion adaptive filtering. *Biomedical Signal Processing and Control, 69*, 102799.

Islam, M. T., Mashfu, S. T., Faisal, A., Siam, S. C., Naheen, I. T., & Khan, R. (2022). Deep learning-based glaucoma detection with cropped optic cup and disc and blood vessel segmentation. *IEEE Access, 10*, 2828–2841. https://doi.org/10.1109/ACCESS.2021.3139160

Jaiswal, A. K., Tiwari, P., Kumar, S., Al-Rakhami, M. S., Alrashoud, M., & Ghoneim, A. (2021). Deep learning-based smart iot health system for blindness detection using retina images. *IEEE Access, 9*, 70606–70615. https://doi.org/10.1109/ACCESS.2021.3078241

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., … Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis, 42*, 60–88.

Maqsood, S., Damaševičius, R., & Maskeliūnas, R. (2021). Hemorrhage detection based on 3D CNN deep learning framework and feature fusion for evaluating retinal abnormality in diabetic patients. *Sensors (Basel, Switzerland)*.

Mitra, A., Banerjee, P. S., Roy, S., Roy, S., & Setua, S. K. (2018). The region of interest localization for glaucoma analysis from retinal fundus image using deep learning. *Computer Methods and Programs in Biomedicine, 165*, 25–35. https://doi.org/10.1016/j.cmpb.2018.08.003

Nawaz, M., Nazir, T., Javed, A., Tariq, U., Yong, H.-S., Khan, M. A., & Cha, J. (2022). An efficient deep learning approach to automatic glaucoma detection using optic disc and optic cup localization. *Sensors, 22*. https://doi.org/10.3390/s22020434

Pham, H., Guan, M., Zoph, B., Le, Q., & Dean, J. (2018). Efficient neural architecture search via parameters sharing. *ICML*.

Rajinikanth, V., Kadry, S., Damaševičius, R., Taniar, D., & Rauf, H. T. (2021). Machine-learning-scheme to detect choroidal-neovascularization in retinal oct image. *2021 seventh international conference on bio signals, images, and instrumentation (ICBSII)* (pp. 1–5).10.1109/ICBSII51839.2021.9445134

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *MICCAI*.

Sandler, M., Howard, A. G., Zhu, M., Zhmoginov, A., & Chen, L. (2018). Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *CVPR*.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *ICLR*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *CVPR*.

Tan, M., Chen, B., Pang, R., Vasudevan, V., & Le, Q. V. (2019). Mnasnet: Platform-aware neural architecture search for mobile. *CVPR*.

Tan, M., & Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *ICML*.

Tan, M., & Le, Q. V. (2021). Efficientnetv2: Smaller models and faster training. *ICML*.

Wang, J., Yang, L., Huo, Z., He, W., & Luo, J. (2020). Multi-label classification of fundus images with efficientnet. *IEEE Access, 8*, 212499–212508. https://doi.org/10.1109/ACCESS.2020.3040275

Wang, X., Xu, M., Li, L., Wang, Z., & Guan, Z. (2019). Pathology-aware deep network visualization and its application in glaucoma image synthesis. In D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, & A. Khan (Eds.), *Medical image computing and computer assisted intervention – MICCAI 2019* (pp. 423–431). Springer.

Wong, D. W. K., Liu, J., Tan, N.-M., Yin, F., Cheng, X., Cheng, C.-Y., Cheung, G. C. M., & Wong, T. Y. (2012). Automatic detection of the macula in retinal fundus images using seeded mode tracking approach. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*.

Xiong, L., Li, H., & Zheng, Y. (2014). Automatic detection of glaucoma in retinal images. *2014 9th IEEE conference on industrial electronics and applications* (pp. 1016–1019).10.1109/ICIEA.2014.6931312

Yang, J., Huang, M., Fu, J., Lou, C., & Feng, C. (2020a). Frangi based multi-scale level sets for retinal vascular segmentation. *Computer Methods and Programs in Biomedicine, 197*, 105752.

Yang, J., Lou, C., Fu, J., & Feng, C. (2020b). Vessel segmentation using multiscale vessel enhancement and a region based level set model. *Computerized Medical Imaging and Graphics, 85*, 101783.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation*, vol. 11045. 4th international workshop DLMIA* (pp. 3–11).10.1007/978-3-030-00889-5_1

Zoph, B., & Le, Q. (2017). Neural architecture search with reinforcement learning. *ICLR*.

Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. V. (2018). Learning transferable architectures for scalable image recognition. *CVPR*.