Fine-tuning Convolutional Neural Networks: a comprehensive guide and benchmark analysis for Glaucoma Screening

Amed Mvoulana, Rostom Kachouri, Mohamed Akil

Fine-tuning Convolutional Neural Networks: a comprehensive guide and benchmark analysis for Glaucoma Screening

Amed Mvoulana
Gaspard-Monge Computer Science Laboratory
Université Gustave Eiffel
CNRS, ESIEE Paris
F-77454 Marne-la-Vallée
Email: amed.mvoulana@esiee.fr

Rostom Kachouri
Gaspard-Monge Computer Science Laboratory
Université Gustave Eiffel
CNRS, ESIEE Paris
F-77454 Marne-la-Vallée
Email: rostom.kachouri@esiee.fr

Mohamed Akil
Gaspard-Monge Computer Science Laboratory
Université Gustave Eiffel
CNRS, ESIEE Paris
F-77454 Marne-la-Vallée
Email: mohamed.akil@esiee.fr

***Abstract*—This work aimed at giving a comprehensive, in-detailed and benchmark guide on the route to fine-tuning Convolutional Neural Networks (CNNs) for glaucoma screening. Transfer learning consists in a promising alternative to train CNNs from scratch, to avoid the huge data and resources requirements. After a thorough study of five state-of-the-art CNNs architectures, a complete and well-explained strategy for fine-tuning these networks is proposed, using hyperparameter grid-searching and two-phase training approach. Excellent performance is reached on model evaluation, with a 0.9772 AUROC validation rate, giving arise to reliable glaucoma diagnosis-help systems. Also, a baseline benchmark analysis is conducted, studying the models according to performance indices such as model complexity and size, AUROC density and inference time. This in-depth analysis allows a rigorous comparison between model characteristics, and is useful for giving practioners important trademarks for prospective applications and deployments.**

I. Introduction

Glaucoma is a neurodegenerative eye disease, causing gradual vision loss and ending up to complete blindness [1]. Glaucoma is known as one of the most prevalent ocular diseases worldwide, as up-to-date projections estimate glaucoma burden to about 112 million people worldwide by 2040 [2]. Dispensing and ensuring early screening of the pathology remains essential, to inhibit the development of the spreading disease and avoid irreversible vision damages with in-time treatment. Glaucomatous optic neuropathy is mainly featured by structural changes within the optic nerve head (ONH), a yellowish and bright circular region within the retina where arteries and veins converge toward the brain. As the disease develops, gradual alteration of the ONH and surroundings occurs: prominence of the optic cup (OC) inside the optic disc (OD), gradual narrowing of the neuro-retinal rim (NRR), retinal nerve fiber layer (RNFL) loss, hemorrhages on the retinal layer (see Figure 1). Hence, to effectively diagnose the presence of the disease at the earlier stage, ophthalmologists explore the retina via dedicated imaging tools [3], and analyse the presence of such glaucomatous patterns.

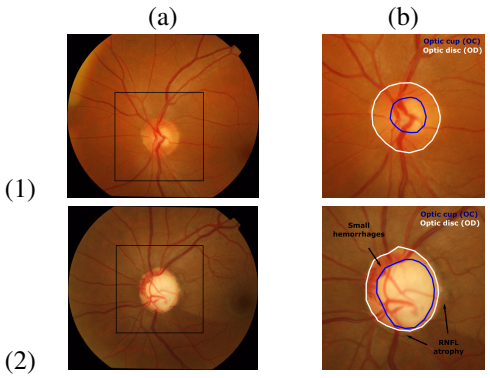

Fig. 1. Example of healthy (1) and glaucomatous (2) retinal images: (a) retinal image with framed ONH region, (b) ONH sub-image.

Computer-aided diagnosis (CAD), in combination with high-resolution digital imaging, has a great potential for assisting clinicians in their work, bringing more effectiveness, affordability and convenience in the task of early screening and diagnosis of glaucoma [4]. The deployment of such CAD systems constitutes a step forward in offering upgraded glaucoma screening strategies, for the development of ocular disease screening programs, widening the access to eye health.

In this work, we exploit powerful deep learning (DL) algorithms to provide effective glaucoma assessment from retinal fundus images. Here, a transfer learning strategy is leveraged andto assess the capacity of DL models on generalizing glaucoma detection. Based on such thorough study, a baseline benchmark analysis is conducted for comprehensive comparison between the exploited models.

The remainder of the paper is organized as follows. Section II introduces the existing works on glaucoma screening from retinal images. Novel approach for fine-tuning CNNs for glaucoma screening is detailed in Section III. Results are presented in Section IV, followed by a benchmark analysis in Section V. Discussion and conclusions are given in Section VI and VII.

II. Related work

Glaucoma gives arise to structural changes within the ONH, mainly characterized by a gradual increase in optic cup (OC) size inside the optic disc (OD), and conversely, a gradual narrowing of the neuro-retinal rim (NRR). Hence, most of the existing approaches for glaucoma assessment rely on the extraction of clinical measurements such as the cup-to-disc ratio (CDR) [5], the NRR area [6] or the ISNT sectors [7], to evaluate the morphological alterations occurring in the ONH. Extracting these clinical measurements requires a preceding segmentation of the OC and the OD areas. In one of the most relevant works for early glaucoma screening from retinal fundus images, Cheng et al. [8] proposed an energy-based superpixel classification method to segment both OC and OD areas. Diameter-based CDR calculation is then performed to lead to glaucoma assessment. In the work by Mvoulana et al. [9], authors exploited a non-supervised clustering method, in combination with a model-based operator, to detect the regions of interest then automatically compute area-based CDR. Both approaches obtain excellent results on final glaucoma assessment. However, reliability on glaucoma screening directly depends on an accurate segmentation of the ONH, which is known to be a challenging task in this biomedical imaging context.

Early advancements on deep learning, a derivated field of artificial intelligence, are at the core of a stunning revolution in the domain of computer vision. Convolutional Neural Networks (CNNs) have improved traditional algorithms in many tasks, including object detection, segmentation and pattern recognition, among various areas ranging from autonomous driving, natural language processing, speech recognition or medical image analysis [10]–[12]. CNNs mainly consist in an ensemble of non-linear modules, capable of extracting features from images at different levels of representation. The key advantage of deep learning is that the filters are automatically learned from data, using a general-purpose learning procedure [13]. Thereupon, recent studies have suggested the usefulness of exploiting CNNs to automatically learn glaucomatous patterns from retinal images, leading to the assessment of glaucoma disease. As a forrunner study, Chen et al. [14] developed a novel six-layers CNN architecture for glaucoma assessment, trained with private ORIGA and SCES datasets. This work has improved traditional state-of-the-art approaches for glaucoma screening, paving the way to develop CNNs to screen the disease. Also, Fu et al. [15] introduced a DENet architecture, consisting of four modules retrieving different hierarchical aspects of retinal fundus images (disc localization, ONH contextual information, etc.), finally aiming to screen glaucomatous neuropathy with high sensitivity. Nevertheless, specifying such novel DL algorithms and effectively trained them from scratch tends to be a complicated task, requiring great amount of data, with consistent and trust-worthy labelling, and substantial hardware resources. Instead, transfer learning has been validated as a valuable alternative to fully training CNNs, especially for designing intelligent systems for the screening of pathologies when data requirements are often deficient [16]. A promising path is to fine-tune CNN that have been pre-trained on large general-purpose dataset (ImageNet), i.e. restoring weights from a pre-trained model, adapting the network for the new classes of interest, and incrementally re-train its layers for better handling of the new classes. Several methods have exploited fine-tuning for the assessment of glaucoma, including the works in [17], [18], each fine-tuning ResNet50 for the detection of glaucoma. Distinguished results are obtained in these studies, however, the lack on in-depth explanation on CNN training strategy, and the use of private datasets make these medical-oriented works hardly reproducible. As a baseline study, Diaz-Pinto et al. [19] proposed an extensive validation of different DL models, including VGG-16, ResNet50 or Inception-v3, each fine-tuned for early screening of glaucoma. Explanation on fine-tuning strategy, specification of hyperparameters are given to further lead to glaucoma screening. Several publicly-available datasets are used for testing and comparing the ability of each model to screen the disease. However, arbitrary specifications of hyperparameters, being the same for all studied models, can jeopardize reliable convergence and narrow global performance for glaucoma screening. Also, a few explanation on computed architectures, and discussions about pros and cons were given to discuss their global generalization capacity, and other model characteristics such as model complexity.

In this work, we aim to make the following contributions for benchmarking well-known CNNs architectures to the task of early glaucoma screening:

- we describe the different used pre-trained CNNs, in a disseminated manner, and expose their advantages and limitations;
- we give a well-explained strategy for fine-tuning different ImageNet-trained CNNs, including optimal specification of hyperparameters, explanation of a two-pass training strategy for effective fine-tuning, and full exhibition of implementation skills for reliable convergence;
- and propose an extensive analysis and discussion of these architectures in terms of accuracy, accuracy density, memory size, inference time, to give a comprehensive report for further applications and deployments.

III. Novel approach for fine-tuning CNNs for glaucoma screening

A. Datasets

A thorough research of publicly-available datasets has been conducted to further build, train and evaluate our models for early glaucoma screening. Among many retinal images datasets dedicated to glaucoma disease and exploited by research works in the field, only a few are publicly available: DRISHTI-GS1 [21], RIM-ONE [22], HRF [20], ACRIMA

Dataset	Healthy	Glaucomatous	Total
HRF [20]	18	27	45
DRISHTI-GS1 [21]	31	70	101
RIM-ONE [22]	194	261	455
ACRIMA [19]	309	396	705
KIM-EYE [23]	786	758	1544

TABLE I
RETINAL IMAGES DATASETS FOR GLAUCOMA SCREENING.

[19] and KIM-EYE[1] [23]. A description of each dataset is drawn in Table I. Among these datasets, KIM-EYE appears as the most suitable choice to conduct DL training for glaucoma assessment, as it contains substantial amount of images to train and validate DL models, especially with the agreement of two well-trained specialists for assessing referable glaucomatous optic neuropathy. The dataset consists of 1544 retinal images, including 786 healthy (H) images and 758 glaucomatous (G) images divided into 289 glaucoma-early and 467 glaucoma-advanced cases. In our undergoing work, no distinction between early and advanced cases is done and all glaucoma cases are gathered in a same class.

B. Data preprocessing and augmentation

Before implementing DL architectures, prior preprocessing of retinal images is required. First, since glaucoma disease mainly manifests itself within and around the ONH, all images are cropped around the ONH. We exploit here the method proposed in [9] to effectively detect the ONH center, and crop the retinal image around the detected center using a (224x224) or a (299x299) window (depending on the default input size required by the models). Relevant prior studies have assessed the utility to operate this cropping, improving the ability of the algorithm to feature the presence of the disease [24]. Second, image normalization is computed to scale pixel intensities from $[0, 255]$ to $[-1; 1]$, as each pixel value X_i in the image X is rescaled following Eq. (1):

$$N(X_i) = \frac{X_i - \overline{X}}{\sigma(X)} \tag{1}$$

with $\overline{X}$ image mean, $\sigma(X)$ image standard deviation, and $N(X_i)$ the output normalized pixel.

Image normalization is crucial for optimal training: it allows to maintain each learned feature in a specific range, preventing from gradients going out of control when multiplying weights with initial inputs. Third, one-hot encoding is performed on the labels, passing from 1D vectors to 2D-representation of the labels. Fourth, to conduct both training and validation of the DL models, whole data is splitted into training (90%) and validation (10%) folders. During training phase, training split will thereafter be splitted for a 10-fold cross validation. Finally, to enlarge training dataset for improving accuracy and proper convergence, data augmentation is appplied: random geometric transformations such as rotation (range: 0-40 degrees), zooming (range: 0-20%), shear (range: 0-20%), and both horizontal and vertical flips.

[1]Kim-EYE refers to the name of the hospital where the dataset has been elaborated. See reference for more information.

C. ImageNet-trained CNNs

In our study, five of the well-known state-of-the-art CNNs architectures were selected for fine-tuning and benchmark analysis. These networks have been selected according to their differences across layer agencement or building block, giving a wide range of network configurations to explore.

1) VGG16: introduced by Simonyan et al. [25] at the ILSVR2014, VGG16 is known as one of the first CNNs proposed in the literature. The architecture is composed of 13 convolutional layers, interspersed with 5 pooling layers and ending with 3 dense layers (see Figure 2). Because of its intuitive sequential architecture, while achieving excellent accuracy on a wide range of computer vision domains, VGG16 has been extensively exploited by the community and remains a well-suggested CNN for benchmarking on a particular task.

2) ResNet50: developed by [26], ResNet50 is a 50-layer CNN architecture mainly featured by residual blocks. A typical residual block consists of 3 convolutional layers, mainstreamed by skip connections to "jump" over some layers. These shortcuts remedy the problem of vanishing gradient, i.e. when the loss function shrinks to zero after several iterations on deep networks, resulting in accuracy degradation. With ResNets, the gradients can flow directly through the skip connections backwards from later layers to initial filters.

3) Inception-v3: proposed by Szegedy et al. [27] in ILSVRC 2015, 48-layers Inception-v3 is the extension of GoogLeNet. Inception module is a multi-level feature extractor, as convolutions of size 1×1, 3×3 and 5×5 are computed within the same module. These modules allow to solve the problem of overfitting, as well as computational expense through dimensionality reduction with stacked 1 × 1 convolutions. The weights for Inception-v3 are smaller than both VGG and ResNets.

4) DenseNet121: proposed by Huang et al. [28], DenseNet121 is an extension of the previously-introduced ResNets. DenseNets blocks are featured as having connections to all following layers in the network. 121-layer DenseNet121 mainly consists of 4 DenseNet blocks, interspersed by compression/transition to reduce the number of feature maps exploited by the subsequent block. A fully-connected layer at the end allows to achieve final classification. One of the main advantage of DenseNet121 is the fewer number of parameters compared to ResNet50, achieving comparable accuracy on ImageNet dataset.

5) MobileNet: introduced by Howard et al. [29], the authors aimed at developing efficient and lighter models for further implementation of DL models on mobile platforms. Consisting of 88 layers with depthwise convolutions, MobileNet is the lighter model among the well-known ImageNet-trained DL models across the state-of-the-art, with outstanding 89.5% top-5 accuracy on ImageNet.

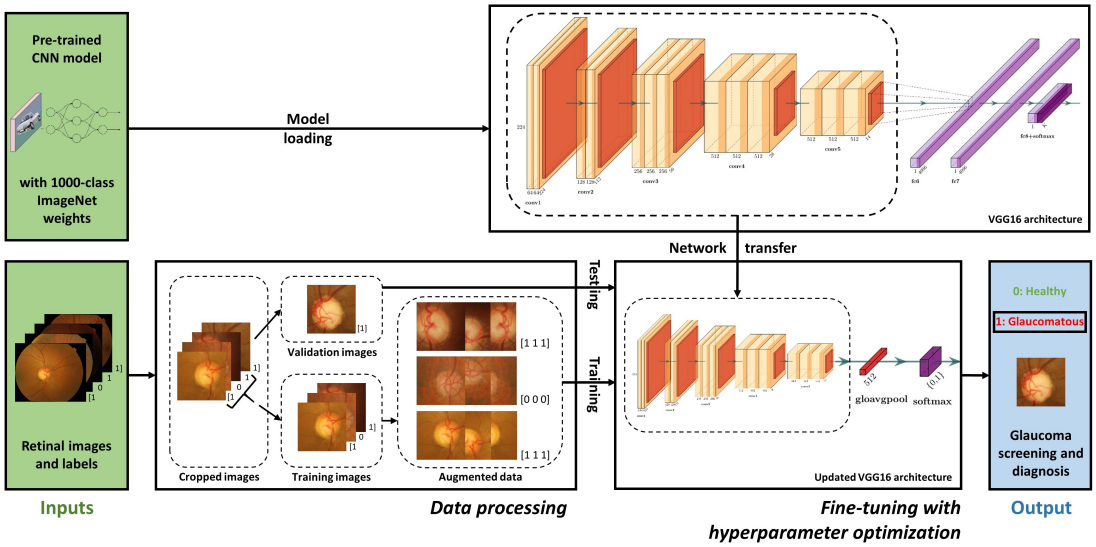

Fig. 2. Flowchart of the transfer learning process for glaucoma screening, with VGG16 architecture: data processing, pre-trained model loading, fine-tuning with a removal of the last fully-connected (FC) layers and softmax classifier, and a replacement with a global average pooling layer and the 2-output softmax classifier. [0] and [1] corresponds to the healthy and glaucomatous classes respectively.

D. Fine-tuning and grid search of hyperparameters

Fine-tuning firstly consists in loading the pre-trained models, and designing them for the targeted classification. Hence, we remove the last 1000-output fully-connected layer, corresponding to the 1000 classes of ImageNet, and replace it with a 2-output fully-connected layer, corresponding to our binary classification between healthy and glaucomatous images. A softmax classifier is computed to give the probabilities associated to each class. This last fully-connected layer is preceded by an average pooling operation, to minimize overfitting by reducing the total number of parameters in the model for better generalization. The computed "surgery" is illustrated in Figure 2, with VGG-16 network. From there, the new network can be fine-tuned. It starts with a "warm-up" phase, where we train only the new last layer of the model, and set all weights belonging to the previous layers as "non-trainable". This operation allows to preserve the powerful features learned from pre-training on ImageNet and contained in the body of the network, when backpropagating the gradient coming from the random values in the new layers. Once the new layer is pre-trained, fine-tuning can be operated by setting a few layers from the network as trainable. Hence, the model is trained a second time until reaching desired performance. In this context, a recent study [16] has suggested the usefulness of "deep tuning", referring to fine-tuning all layers in the model. Deep-tuning is recommended when it aims to exploit transfer learning for a target application far from the ImageNet source dataset, and setting all layers as trainable tends to improve results compared to fine-tuning a sub-part of the model.

To obtain the best performance on training-validation steps along both "warm-up" and "deep-tuning" phases, specifying the best combination of hyperparameters is crucial. In this direction, we perform a grid search of hyperparameters: it consists, for each hyperparameter, in specifying an interval of values to scan. Then, we automatically compute training step along all combinations of hyperparameters, according to the values in each interval, and select the combination of hyperparameters giving the best tradeoff between accuracy and global generalization. Such grid searching allows to find out the best hyperparameters for managing each architecture. In our work, we focused on grid-searching learning rate (LR) and batch size (BS) values, being the most impacting hyperparameters for proper convergence of fine-tuned models during training [16]. Grid search is operated along the sample of values $I_{LR} = \{1e^{-5}; 5e^{-5}; 1e^{-4}\}$ for learning rate, as low learning rate values allow to reliably follow loss landscape. Batch size is grid searched along the sample of values $I_{BS} = \{8, 12\}$, giving a good balance between memory cost and better generalization. Stochastic Gradient Descent (SGD) is chosen as the optimizer for all our study, offering reliable convergence (momentum = 0.9, decay = $1e^{-6}$). Number of epochs is specified as 120 for warm-up phase, and 80 for fine-tuning phase. Also, to enhance the reliable convergence of our networks, early stopping is computed along both stages, as the training phase is early interrupt after 20 epochs without decreasing on validation loss. A summary of all specified hyperparameters is given in Table II.

Learning rate	Optimizer	Momentum	Decay
$\{1e^{-5}; 5e^{-5}; 1e^{-4}\}$	SGD	0.9	$1e^{-6}$

Batch size	Epochs		Early stop.
	Warm-up	Fine-tuning	
$\{8; 12\}$	≤ 120	≤ 80	20

TABLE II
SPECIFIED HYPERPARAMETERS FOR OPTIMIZATION OF THE MODELS.

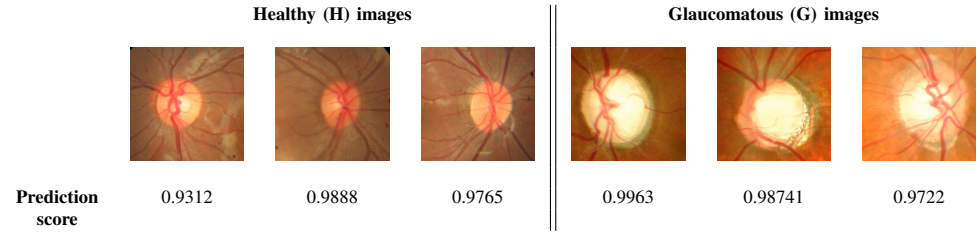

Fig. 3. Examples of correctly classified retinal images, using VGG16 network. Here, the prediction score on the ground-truth class (H or G) is given.

Model	AUROC	Acc	Sen	Spe	PPV	NPV	F1-score
VGG16	**0.9772**	**0.9290**	0.9130	**0.9419**	**0.9265**	0.9310	**0.9197**
ResNet50	**0.9703**	0.9161	0.9275	0.9070	0.8889	0.9398	0.9078
Inception-v3	0.9658	0.9189	0.9275	0.8837	0.8649	0.9383	0.8951
DenseNet121	0.9681	0.9161	0.9275	0.9070	0.8889	0.9398	0.9078
MobileNet	0.9626	0.9032	**0.9565**	0.8605	0.8462	**0.9610**	0.898

TABLE III
EVALUATION RESULTS OF THE FINE-TUNED DL MODELS ON KIM-EYE DATASET.

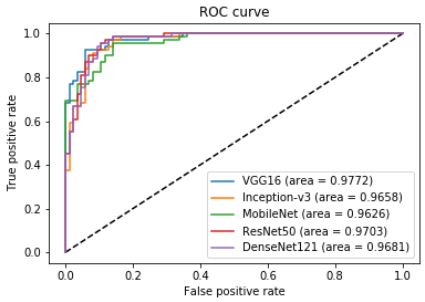

Fig. 4. ROC curve for fine-tuned DL models, with AUROC-integrated legend: VGG16, Inception-v3, MobileNet, ResNet50 and DenseNet121.

IV. RESULTS

A. Framework configuration

All experiments were conducted using Keras deep learning framework [30], including all implemented models with ImageNet weights. Models were trained using a NVIDIA 1080 Ti GPU, with a 11 Go RAM memory. Scripts were implemented on Jupyter notebooks.

B. Experimental results

Our algorithms were evaluated using the Receiver Operating Characteristic (ROC), illustrating the diagnostic quality of our binary glaucoma classifiers across true positive and false positive rates. The area under ROC (AUROC) curve is one of the most important metrics for evaluating classification performance, as it describes how much the model is capable of distinguishing the two healthy and glaucomatous classes. Also, accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and F1-score were calculated to feature the ability of the models to effectively classify the retinal images. Model evaluation were operated across the validation set (10% of total available data), consisting of 155 retinal images with 86 healthy and 69 glaucomatous images from KIM-EYE dataset.

In this direction, Figure 4 illustrates the ROC curve obtained for each DL model. Plotted ROC curves demonstrate the high capacity of implemented models to screen glaucoma, as they get closer to the left-hand border and then the top border along the ROC space. To emphasize model evaluation, Table III summarizes the obtained results when evaluating the fine-tuned models on KIM-EYE dataset. First, high-rate values are achieved on AUROC, testifying the excellent performance of DL models on glaucoma assessment across both classes. Results between 0.9626 and 0.9772 are obtained, with a top-value reached with VGG16 network. Second, global accuracy achieves excellent performance on all fine-tuned models, ranging from 0.9032 for MobileNet to about 0.93 for VGG16 and Inception-v3. According to sensitivity metric, associated to the classification among glaucomatous subjects, top-rate values across DL models are observed. A remarkable 0.9565 rate on MobileNet illlustrates the ability of the model to effectively detect glaucomatous patients. F1-score, which is a harmonic average of both sensitivity and PPV metric, also shows off great performance along the five models. In sum, fine-tuned models obtain excellent performance on testing set, according to performed metrics, which can attest their reliability on screening glaucoma.

As a qualitative outcome, a sample of retinal images from the testing set is presented, with the output prediction given by the DL models (see Figure 3). The output probability score comes from VGG16 network, and is associated to the prediction rate allocated to the ground-truth class. Hence, it allows to interpret the performed classification, where a prediction value superior to 0.5 indicates a correct diagnosis. Correctly classified examples, with high obtained prediction scores, testify the trustworthiness of the algorithm on screening the presence of the disease.

To enrich the evaluation of the fine-tuned models, a supplementary evaluation phase has been performed on other retinal image datasets. These datasets, which differ from the former training-validation dataset in terms of image acquisition settings, clinical cohort or expert labelling, allow to assess the ability of DL models on generalizing glaucoma screening from retinal images. In this direction, experimentation with ACRIMA and DRISHTI-GS1 datasets has been carried out to analyse each model's ability on assessing glaucoma from images with outlying conditions (for each, 50% for training and 50% for testing). We performed two experiments: (1) evaluation with the models trained on KIM-EYE train set, (2) evaluation with the models trained on KIM-EYE, ACRIMA and DRISHTI-GS1 train sets. In this direction, Figure 5

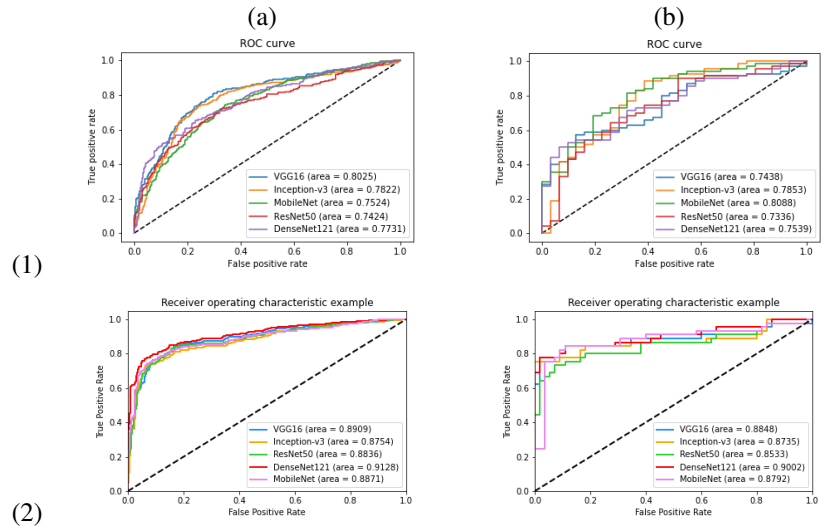

Fig. 5. ROC curve obtained from fine-tuned models on ACRIMA (a) and DRISHTI-GS1 (b) datasets, along experiments (1) and (2).

Dataset	Model	Experiment	AUROC	Acc	Sen	Spe	PPV	NPV	F1-score
(a) ACRIMA	VGG16	(1)	0.8025	0.6823	**0.8990**	0.4045	0.6139	**0.8362**	0.7607
		(2)	**0.8909**	**0.7801**	0.8738	**0.6602**	**0.7672**	0.8032	**0.8170**
	ResNet50	(1)	0.7424	0.6270	**0.8611**	0.3269	0.5319	**0.8253**	0.7217
		(2)	**0.8836**	**0.7730**	**0.8611**	**0.6602**	**0.7646**	0.7876	**0.8100**
	Inception-v3	(1)	0.7822	0.7078	**0.8662**	0.5049	0.5710	**0.8332**	0.7691
		(2)	**0.8754**	**0.7589**	0.8434	**0.6505**	**0.7557**	0.7643	**0.7971**
	DenseNet121	(1)	0.7731	0.6837	0.8485	0.4725	**0.6724**	0.8455	0.7508
		(2)	**0.9128**	**0.7901**	**0.8864**	**0.6667**	0.7731	**0.8207**	**0.8259**
	MobileNet	(1)	0.7524	0.6723	**0.8914**	0.3916	0.4928	**0.8741**	0.7535
		(2)	**0.8871**	**0.7702**	0.8561	**0.6602**	**0.7635**	0.7816	**0.8071**
(b) DRISHTI-GS1	VGG16	(1)	0.7438	0.7228	0.8143	0.5161	0.5801	0.7944	**0.8028**
		(2)	**0.8848**	**0.7500**	**0.8158**	**0.6905**	**0.7045**	**0.8056**	0.7561
	ResNet50	(1)	0.7336	**0.7723**	**0.90**	0.4839	0.5312	**0.8754**	**0.8456**
		(2)	**0.8533**	0.6750	0.7632	**0.5952**	**0.6304**	0.7353	0.6905
	Inception	(1)	0.7853	0.8020	**0.8857**	**0.6129**	**0.6915**	**0.8510**	**0.8611**
		(2)	**0.8735**	**0.7250**	0.8421	**0.6190**	0.6667	0.8125	0.7442
	DenseNet121	(1)	0.7539	0.6931	0.7286	0.6129	0.5904	0.7188	0.7469
		(2)	**0.9002**	**0.7500**	**0.8158**	**0.6905**	**0.7045**	**0.8056**	**0.7561**
	MobileNet	(1)	0.8088	0.7624	0.8286	0.6129	0.6430	0.8021	0.8286
		(2)	**0.8792**	**0.7850**	**0.8158**	**0.7429**	**0.7739**	**0.7941**	**0.8381**

TABLE IV
OBTAINED RESULTS ON MODEL EVALUATION, ALONG ACRIMA AND DRISHTI-GS1 DATASETS ACCORDING TO EVALUATION METRICS.

illustrates the obtained ROC curves on both ACRIMA (a) and DRISHTI-GS1 (b) datasets, across the two experiments on the five architectures. Also, Table IV summarizes the obtained results across the conducted experiments. These results globally demonstrate top-level performance on glaucoma assessment for both datasets. Especially, the main outcome relies on a higher capacity on glaucoma screening with the models trained on all datasets, allowing to enhance the global generalization when screening for glaucomatous patterns. This outcome is globally validated for all architectures, across ACRIMA and DRISHTI-GS1 datasets.

V. BENCHMARK ANALYSIS

The goal of this benchmark study is to provide a thorough analysis of each architecture into different implementation aspects. The fine-tuned architectures for glaucoma screening were compared using different performance indices:

1) Model depth: corresponds to the number of layers in the network, including average pooling and final softmax layer

2) Model complexity: corresponds to the number of parameters (in millions) in the network

3) Memory size: corresponds to the size of the model (in megabits) inside the disk

4) AUROC density: corresponds to the ratio between AUROC on testing phase and the number of parameters. AUROC density is important to measure the impact of the parameters on classification accuracy. AUROC density is calculated from the obtained AUROC on evaluation of the models trained with KIM-EYE (see Table III).

Model	Model depth	Model comp. (M)	Memory size (Mb)	AUROC	AUROC density	Inference time (ms)
VGG16	21	14.72	118	0.9772	0.0664	**7.60**
ResNet50	192	23.57	189	0.9703	0.0412	18.86
Inception-v3	313	21.81	175	0.9658	0.0443	31.88
DenseNet121	429	**7.04**	57.3	0.9681	0.1375	32.58
MobileNet	89	**3.23**	**26**	0.9626	**0.2979**	**10.64**

TABLE V
SUMMARY OF PERFORMANCE INDICES, FOR BENCHMARK ANALYSIS. THESE RESULTS ARE BASED ON THE RESULTS OUTLINED IN TABLE III.

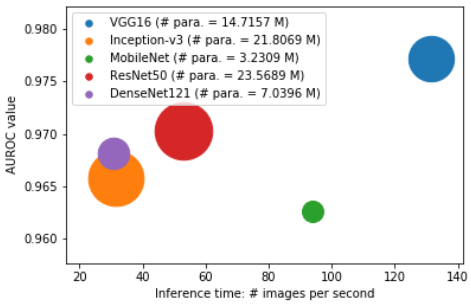

Fig. 6. Ball chart reporting inference time (images per second, along x-axis), AUROC value (along y-axis) and model complexity (# parameters, ball size).

5) Inference time: corresponds to the time (in milliseconds) for the architecture to give prediction on a sample image. Inference time is measured by computing the prediction on all test set, with batch size equal to 1, and divided total runtime by the number of images. This experiment was conducted 10 times and the average inference time was calculated.

Hence, Table V reports the benchmark performance indices from the five DL models. Also, Figure 6 illustrates reached performance indices across inference time, with the number of predicted images per second (x-axis), AUROC rate (along y-axis), and model complexity (ball size). The figure emphasizes the performance of each model on glaucoma screening with a top value for VGG16. Also, the chart suggests that increased model complexity do not necessarily induce model performance. Moreover, lighter and less complex models such as MobileNet and VGG16 tend to have a shorter inference time.

VI. DISCUSSION

In this study, we exploited transfer learning for glaucoma screening from retinal fundus images. Specifically, we chose to analyse five well-known state-of-the-art CNNs, each featured by its proper layer agencement, building block, and other architecture properties. We also proposed a benchmark analysis of such networks, according to specific parameters, offering important trademarks for applications and deployments.

The first goal was to provide a well-explained and in-detailed guide for fine-tuning deep convolutional networks for the purpose of glaucoma screening, using the strength of these networks at extracting low- to high-level features from digital images. The main challenge when exploiting transfer learning is to reliably transpose acquired knowledge to the targeted classification. It is even more the case when exploiting knowledge acquired from a far away classification task, as our study aimed at exploiting ImageNet-trained CNNs for medical imaging purpose. In this direction, performing the best strategy along data collection and preprocessing, hyperparameter specification and training routine is mandatory, on the pathway to generalizing the assessment of glaucoma. To do so, we firstly studied the different publicly available retinal images, dedicated to the screening of glaucomatous condition. Among a few available datasets, KIM-EYE has been chosen as a baseline training-validation dataset, consisting of enough retinal images to train our algorithms on, with trustworthy, consistant labelling by trained ophthalmologists. As the former specialist labelling has been conducted among healthy, moderate and severe glaucoma condition, the both moderate and severe classes have been gathered to form one class and equally balance the different classes before performing binary classification. Also, data preprocessing with image resizing and normalization, accompanied by data augmentation to remedy the lack of training data, allows to build a ready-to-use dataset for training the DL networks. According to training-validation sets, a 90-10 distribution has been considered to collect enough data for CNNs training. Second, hyperparameter specification is perhaps the most challenging and critical phase along fine-tuning route. To answer to this difficulty, a fully-automated grid search algorithm has been implemented, allowing to find the best parameters for proper convergence. Grid-searching has been conducted along two hyperparameters, learning rate and batch size, appearing as the ones having the biggest impact on reliable convergence. Grid-search of hyperparameters induces a longer training phase, in comparison with a traditional training phase, but worth the effort to find the most adapted parameters. Also, to find a balance between accurate choice of these hyperparameters and training consumption, we defined restricted intervals of values to grid-search, with a few but relevant values to scan. In our study, Stochastic Gradient Descent (SGD) has been chosen as the preferred optimizer compared to Adam, Nadam, Adagrad or RMSprop, offering satisfying convergence of all architectures. According to the obtained results, excellent performance has been reached when evaluating the performed models on KIM-EYE testing data. AUROC as well as global accuracy testify the ability of the models on detecting glaucomatous patterns from the images. When evaluating the models on different datasets, encouraging but dropping results were found on glaucomatous classification, as the evaluation in variant imaging conditions seems to disrupt the KIM-EYE-trained CNNs. However, when integrating samples from these abroad datasets, the models appeared to perform better and greater capacity in screening the pathology is observed.

The second goal of this study was to give a benchmark analysis of the five architectures, explore their different implementation characteristics and give a complete view for researchers on the pathway to developing practical applications or deploying such algorithms. Hence, different performance indices such as architecture depth, model complexity, memory

size, AUROC density, and inference time were computed. One of our main findings is that deeper and more complex networks do not necessarily transfer the better. For example, VGG16 architecture has been found as the most accurate model when designed for our task of early glaucoma screening. In terms of memory size and model complexity, this study suggests that MobileNet, followed by DenseNet121, are the preferred models to exploit when developing DL models for mobile deployment. As a metric measuring the impact of each parameter in system accuracy, AUROC density highlights the strength of MobileNet architecture in efficiently screen the disease with a few number of parameters. Inference time, which corresponds to the required time for giving prediction on a sample image, is reached within a few milliseconds of all architectures, opening the gate to real-time diagnosis-help systems. Among all architectures, both VGG16 and MobileNet predict glaucomatous neuropathy is around or less than 10 ms.

VII. Conclusion

In this paper, we proposed a novel approach for fine-tuning Convolutional Neural Networks (CNNs) for glaucoma screening from retinal fundus images. The study aimed at giving a precise, well-explained guide to fine-tuning five of the most known state-of-the-art CNNs architectures. In this work, a two-phase training strategy and a fully-automated hyperparameter grid-searching is operated, finally giving accurate and reliable models designed for screening glaucomatous subjects. Moreover, a benchmark analysis was conducted, highlighting the different features of implemented models in terms of model complexity, density or inference time. This thorough study aims at giving researchers standards on developing DL models for further applications and deployments.

References

[1] S. Kingman, "Glaucoma is second leading cause of blindness globally," *Bulletin of the World Health Organization*, vol. 82, pp. 887–888, 2004.

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

[3] M. D. Abràmoff, M. K. Garvin, and M. Sonka, "Retinal imaging and image analysis," *IEEE reviews in biomedical engineering*, vol. 3, pp. 169–208, 2010.

[4] P. S. Grewal, F. Oloumi, U. Rubin, and M. T. Tennant, "Deep learning in ophthalmology: a review," *Canadian Journal of Ophthalmology*, vol. 53, no. 4, pp. 309–313, 2018.

[5] M. F. Armaly and R. E. Sayegh, "The cup/disc ratio: The findings of tonometry and tonography in the normal eye," *Archives of Ophthalmology*, vol. 82, no. 2, pp. 191–196, 1969.

[6] P. J. Airaksinen, S. M. Drance, and M. Schulzer, "Neuroretinal rim area in early glaucoma," *American journal of ophthalmology*, vol. 99, no. 1, pp. 1–4, 1985.

[7] J. B. Jonas, G. C. Gusek, and G. O. Naumann, "Optic disc morphometry in chronic primary open-angle glaucoma," *Graefe's archive for clinical and experimental ophthalmology*, vol. 226, no. 6, pp. 522–530, 1988.

[8] J. Cheng, J. Liu, Y. Xu, F. Yin, D. W. K. Wong, N.-M. Tan, D. Tao, C.-Y. Cheng, T. Aung, and T. Y. Wong, "Superpixel classification based optic disc and optic cup segmentation for glaucoma screening," *IEEE transactions on medical imaging*, vol. 32, no. 6, pp. 1019–1032, 2013.

[9] A. Mvoulana, R. Kachouri, and M. Akil, "Fully automated method for glaucoma screening using robust optic nerve head detection and unsupervised segmentation based cup-to-disc ratio computation in retinal fundus images," *Computerized Medical Imaging and Graphics*, vol. 77, p. 101643, 2019.

[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[14] X. Chen, Y. Xu, D. W. K. Wong, T. Y. Wong, and J. Liu, "Glaucoma detection based on deep convolutional neural network," in *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2015, pp. 715–718.

[15] H. Fu, J. Cheng, Y. Xu, C. Zhang, D. W. K. Wong, J. Liu, and X. Cao, "Disc-aware ensemble network for glaucoma screening from fundus image," *IEEE transactions on medical imaging*, vol. 37, no. 11, pp. 2493–2501, 2018.

[16] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[17] S. Liu, S. L. Graham, A. Schulz, M. Kalloniatis, B. Zangerl, W. Cai, Y. Gao, B. Chua, H. Arvind, J. Grigg *et al.*, "A deep learning-based algorithm identifies glaucomatous discs using monoscopic fundus photographs," *Ophthalmology Glaucoma*, vol. 1, no. 1, pp. 15–22, 2018.

[18] Z. Li, Y. He, S. Keel, W. Meng, R. T. Chang, and M. He, "Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs," *Ophthalmology*, vol. 125, no. 8, pp. 1199–1206, 2018.

[19] A. Diaz-Pinto, S. Morales, V. Naranjo, T. Köhler, J. M. Mossi, and A. Navea, "Cnns for automatic glaucoma assessment using fundus images: an extensive validation," *Biomedical engineering online*, vol. 18, no. 1, p. 29, 2019.

[20] A. Budai, R. Bock, A. Maier, J. Hornegger, and G. Michelson, "Robust vessel segmentation in fundus images," *International journal of biomedical imaging*, vol. 2013, 2013.

[21] J. Sivaswamy, S. Krishnadas, A. Chakravarty, G. Joshi, A. S. Tabish *et al.*, "A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis," *JSM Biomedical Imaging Data Papers*, vol. 2, no. 1, p. 1004, 2015.

[22] F. Fumero, S. Alayón, J. L. Sanchez, J. Sigut, and M. Gonzalez-Hernandez, "Rim-one: An open retinal image database for optic nerve evaluation," in *2011 24th international symposium on computer-based medical systems (CBMS)*. IEEE, 2011, pp. 1–6.

[23] J. M. Ahn, S. Kim, K.-S. Ahn, S.-H. Cho, K. B. Lee, and U. S. Kim, "A deep learning model for the detection of both advanced and early glaucoma using fundus photography," *PloS one*, vol. 13, no. 11, 2018.

[24] J. I. Orlando, E. Prokofyeva, M. del Fresno, and M. B. Blaschko, "Convolutional neural network transfer for automated glaucoma identification," in *12th international symposium on medical information processing and analysis*, vol. 10160. International Society for Optics and Photonics, 2017, p. 101600U.

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

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[28] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[29] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[30] F. Chollet *et al.*, "Keras: The python deep learning library," *ascl*, pp. ascl–1806, 2018.