Cataract grading method based on deep convolutional neural networks and stacking ensemble learning

Yaroub Elloumi

Cataract Grading Method Based on Deep Convolutional Neural Networks and Stacking Ensemble Learning

Yaroub Elloumi
Mail : yaroub.elloumi@esiee.fr
ORCID : https://orcid.org/0000-0001-8878-7562

LIGM, Univ Gustave Eiffel, CNRS, ESIEE Paris, F-77454 Marne-la-Vallée, France
Medical Technology and Image Processing Laboratory, Faculty of medicine, University of Monastir, Tunisia
ISITCom Hammam-Sousse, University of Sousse, Tunisia

Abstract

Purpose The cataract is the most common cause of severe vision impairment or blindness worldwide. It is essential to periodically diagnose the retina in order to prevent cataract severity, and so to enhance the life quality of cataract-affected patients. Cataract grading through a fundus image is feasible with higher accuracy. However, a delay of early cataract screening is registered caused by deficiency of ophthalmologists and imaging devices. The challenge is to propose a CAD system to grade the cataract from retinal images.

Method In this paper, an ensemble learning framework for cataract grading is put forward, where three convolutional deep neural networks are stacked in order to provide higher performance grading. The main contributions of this work are given as follows: (1) Preprocessing and data augmentation of fundus images are performed to ensure the robustness of the cataract grading; (2) The well-known DL architectures (Inception-V3, MobileNet-V2 and NasNet-Mobile) are fine-tuned and learned as base classifiers; (3) A stacking method is propounded to combine the features of base classifiers.

Results The evaluation is conducted using a dataset of 590 fundus images selected from two public databases. The suggested framework achieves 93.97% accuracy, 95.59% sensitivity, 91.67% specificity, 94.20% precision and 94.89% F-measure for cataract grading.

Conclusion The proposed framework successfully grades fundus images into cataract severity. Moreover, stacking ensemble learning allows achieving a performance that significantly surpasses the ones realized by each DL architecture, applied separately.

Keywords: Fundus images, Cataract, Deep learning, Ensemble learning

1. Introduction

The eye lens is a circular and transparent component, located beyond the iris, which allows refracting the visual information into the retina [1]. The cataract is an ocular disease that affects the eye lens, hence becoming cloudy. As reported by the world health organization in 2020, the cataract caused 65.2 million people suffering from moderate or severe vision impairment or blindness, which exceeds widely the persons having the same severity caused by glaucoma, diabetic retinopathy, corneal opacities and trachoma all together. More than 50% of worldwide blindness cases are caused by the cataract where the number of blind people will exceed 40 million in 2025 [2, 3]. The cataract is involved due to smoking, genetics, alcohol, nutritional or metabolism disorders, medications or a long exposure to sunlight [5, 4, 6]. In addition, it is correlated with several other pathologies [1, 5].

The cataract can be diagnosed through the slit-lamp photography, which consists in reflecting light inside the lens structure and deducing the cataract disease from the non-uniform illumination of the refracted intensity. The Lens Opacities Classification System III, the American Cooperative Cataract Research Group protocol and the Oxford Clinical Cataract Classification are the well-known protocols for classifying the cataract disease, which are based on convoluted procedures and require well-experienced ophthalmologists [2, 3]. The ultrasound backscattering signal, the optical coherence tomography and the ultrasound biomicroscope are also used for cataract screening. However, their diagnosis processes are costly and based on complex operations. The cataract is easily diagnosed through the ophthalmoscopy where the blurriness of the retina components is similar to the quality of the visual acuity. Cataract screening and grading through a retinal fundus image is feasible with

higher accuracy, even with inexperienced graders [2, 3], which promotes accessing to eye diagnosis.

While the cataract develops slowly, it is hard to recognize it since its symptoms are similar to a poor visual acuity [3, 5]. Indeed, it is always diagnosed through a visual acuity test. This pathology imposes limiting activities and affects the life quality [1, 5]. Therefore, it is recommended to make a periodical diagnosis in order to enhance the life quality of cataract-affected patients[3]. However, the process of cataract diagnosis is time-consuming and leads to an important workload, especially with the higher rate of requested diagnosis. In addition, an alarming worldwide deficiency of ophthalmologists is registered [7]. The lack of ophthalmologists will be aggravated in future years and associated to an augmentation of elderly people. Consequently, a delay is registered to ensures early cataract diagnosis, and will persist in the years ahead.

Therefore, several studies have focused on proposing a Computer-Aided-Diagnosis (CAD) system for cataract detection and grading. The common challenge is to succeed in detecting and grading the cataract in terms of opacities. For this purpose, several methods have been suggested which have realized optimal performances of cataract detection. However, they have failed to achieve a higher performance of cataract grading, even resorting to Ensemble Learning (EL). Elsewhere, those methods have been always evaluated using private datasets where a lack a public database containing cataract-affected fundus images has been noticed, in contrast to the other ocular pathologies such as diabetic retinopathy and glaucoma. The evaluation processes have been always performed using few hundreds of fundus images, inadequately partitioned in terms of grading, which might not guarantee a reliable evaluation.

Therefore, our main objective is to provide a CAD system for cataract grading. The first challenge consists in ensuring higher accuracy when classifying images into cataract stages. In fact, several Deep Learning (DL) architectures are dedicated for the classification problem, which are varied in terms of processing principles, and so in terms of classification results. Our main idea consists in stacking features of DL architectures in order to provide higher performance grading. The second challenge is to ensure such grading even using a small dataset. To address this problem, we firstly select three DL architectures composed by lightweight convolutional blocks that allow converging the trained model without resorting to a large number of images. Secondly, we fine-tune those DL architectures initially trained using the ImageNet dataset. In addition, we perform a well-off data augmentation to diversify retina modelling.

The remainder of this paper is organized as follows. Section 2 presents a literature review about the cataract-affected retinal images and the existing methods for grading. Section 3 details the learning framework of cataract grading. Section 4 describes the experimentation and the discussion of the provided results. Section 5 presents the conclusion and some future work.

2. Literature review

The cataract occurs when proteins are accumulated in different locations of the lens [2, 6] which becomes cloudy. In terms of severity, vision becomes blurred with distressed colors and troubled with bright light [5, 6], which corresponds to vision impairment. If the lens is totally clouded, the light is disrupted when projected on the retina, which avoids reflecting real images, and hence leading to blindness [9, 5]. Respectively, fundus images are graded into non-cataract, mild, moderate or severe cataract, as shown in Fig.1. Before the cataract disease, all retinal components, even the micro-vascular structures, are explicitly illustrated. In the mild stage, only choroid and capillary vessels cannot be distinguished among the other components. The moderate stage is deduced when only the optic disk and the main vessels are visible. It becomes difficult to observe any retinal structures in the severe stage. Some work has been focused on proposing a CAD system for cataract detection and grading. Recently, the proposed methods have been based on machine learning, while varied in terms of types and principles.

Some methods have employed single-machine learning for grading. In the work suggested in [9], features were extracted, whose processing was based on the sketch method with discrete cosine transforms and discrete wavelet transforms. After that, a multiclass discriminant analysis algorithm was used for cataract classification. The dataset was composed by 445 fundus images where 199 are non-cataract and 148, 71 and 27 are in mild, moderate and severe cataract stages. The classification performance of cataract detection and grading was 90.9% and 77.1%, respectively. In [11], the suggested method consisted in performing a DL architecture to provide a feature map that was input to an SVM classifier so as to grade images. The DL architecture was composed successively by a convolution layer with a filter kernel size of 11, an activation layer with a ReLU function, a pooling layer, and a normalization layer. A dataset of 7,851 fundus images was used for training, where an accuracy rate of 90.82% was achieved. The method put forward in [12] extracted features reflecting visible structures, local standard deviation and contrast of vessels against background. The extracted features allowed to train a decision tree that ensured a four-class grading with 83.8% of accuracy. The

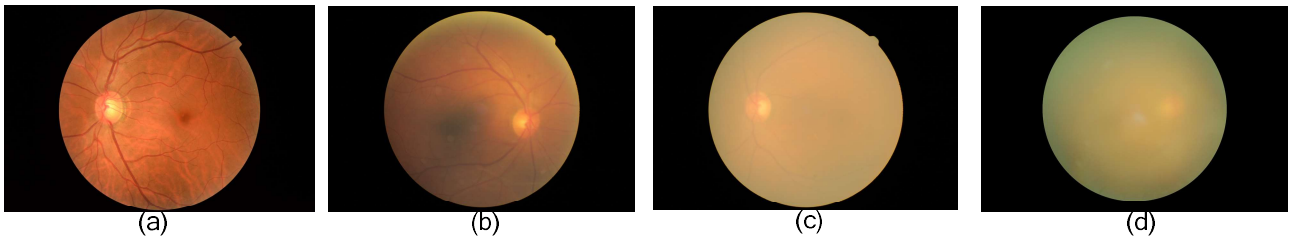

Figure 1. Retinal images of all cataract grades. (a) Non-cataract. (b) Mild cataract. (c) Moderate cataract. (d) Severe cataract

method suggested in [2] figured out the cataract grading using a deep neural network. Firstly, improved Haar wavelet features and retinal structure features were extracted from the retinal image. Then, the feature vector was input to a Multi-Layer Perceptron (MLP) classifier with an exponential discrete state transition. The accuracy of cataract detection and grading is respectively 92.85% and 89.23% using a dataset of 1,355 retinal images. In [13], green channels of retinal images were preprocessed by applying the histogram equalization and the top-bottom hat transformation. Then, the wavelet and texture features were extracted. The classification was done using Self Organizing Maps (SOM) and a Radial Basis Function (RBF) neural network to process clustering, where 91.7% of accuracy for cataract grading was achieved.

Several methods have exploited multiple machine learning algorithms to perform a higher performance of cataract detection and grading. In [4], The eigenvalues were calculated from the gray-level and gray-gradiant co-occurrence matrices to be used as texture features. In addition, the discrete wavelet transform and the discrete cosine transform were performed to provide wavelet and sketch features. Thereafter, The Support Vector Machine (SVM) and the Logistic Regression (LR) were used for classification. Each machine learning algorithm was performed several times as a binary classifier following the one-Vs-rest principle. Then, binary classifiers of each machine learning algorithm were combined to perform the four types of grading classification. A dataset of 2,000 retinal images was used, which was split equitably between training and testing steps. The ensemble classifier based on the SVM outperformed the LR one where the accuracy of the four-stage cataract grading was about 88.60%. The work described in [14] consisted in extracting independent wavelet, sketch and texture based features from preprocessed retinal images. The SVM and Back Propagation Neural Network (BPNN) were used as based classifiers, where their results were provided to EL to perform the final classification. The method was evaluated using 1,239 fundus images, where their numbers in terms of increased severity were 767, 246, 128, and 98. The experimentation showed that EL achieved a better performance than each single learning model, where the realized accuracy was 93.2% and 84.5 % for cataract detection and grading, respectively. In [3], the Haar wavelet decomposition was used to extract features that reflected how the retinal components were modelled in the fundus image. The optimal features were selected through a BPNN classifier. Thereafter, three types of adjacent two-class classification based on majority voting were performed to ensure the four-class classification problem of the cataract. The experimentation was performed using 1,355 retinal images, where the accuracy of the two-class and four-class classification was respectively 94.83% and 85.98%. The method exposed in [15] was dedicated to cataract detection. It started by extracting texture and sketch features from pre-processed images. Then, an EL experimentation was carried out utilizing the decision tree, BPNN and sequential minimal optimization classifiers and using a dataset of 374 fundus images. Each classifier was trained first by texture features, second by sketch features, and third by combining both of them. It was deduced that merging features would allow realizing better accuracy than using each kind of features separately. Thereafter, the classifiers trained using the same feature set were stacked. It was synthetized that the stacked learning method achieved the detection accuracy of 95.47%, which exceeded the one provided by each single classifier.

3. Ensemble learning framework for cataract severity grading

3.1. Preprocessing & data augmentation

The fundus images are captured through varied devices with different technologies, hence provided with multiple resolutions. For this purpose, all fundus images are resized in order to be normalized. The new sizes are determined with respect to the input of each employed DL architecture. In addition, the fundus images have varied and unbalanced contrasts caused either by the lamp used in the capture task or by the retina morphology, as presented in the first and second images in Fig.2. To avoid this limitation, the histogram equalization is applied to enhance the whole contrast ratio per images and to normalize the contrast of all images of the dataset [16], as shown in the third and fourth images in Fig.2.

The retinas are illustrated through different Field-Of-Views (FOVs) that depend on the capture devices. Moreover, the FOV gap leads to a similar difference of masks used to encompass the retinas. For this purpose, a data augmentation process is suggested, which consists in creating new images by modifying the initial ones. This step allows raising the dataset size to augment the method robustness. As regards the FOV and mask differences, each fundus image is zoomed, sheared to a random corner, and shifted and flipped in the horizontal and vertical directions, as respectively depicted in the fundus images in Fig.3. Furthermore, the fundus image is flipped respectively in horizontal, vertical and both horizontal and vertical directions.

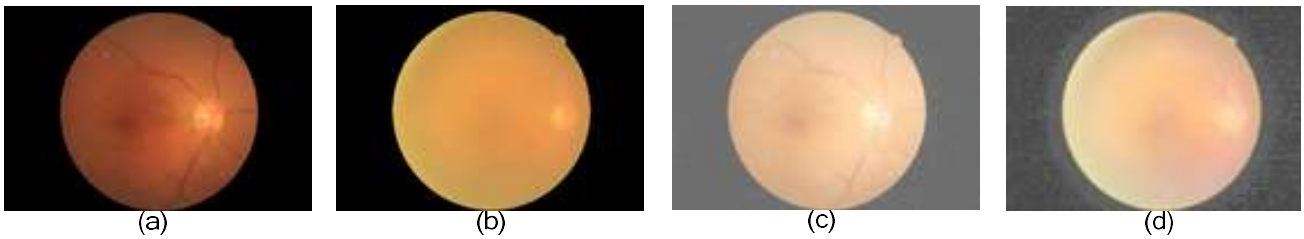

Figure 2. (a), (b): Original fundus images; (c), (d): Fundus images after histogram equalization.

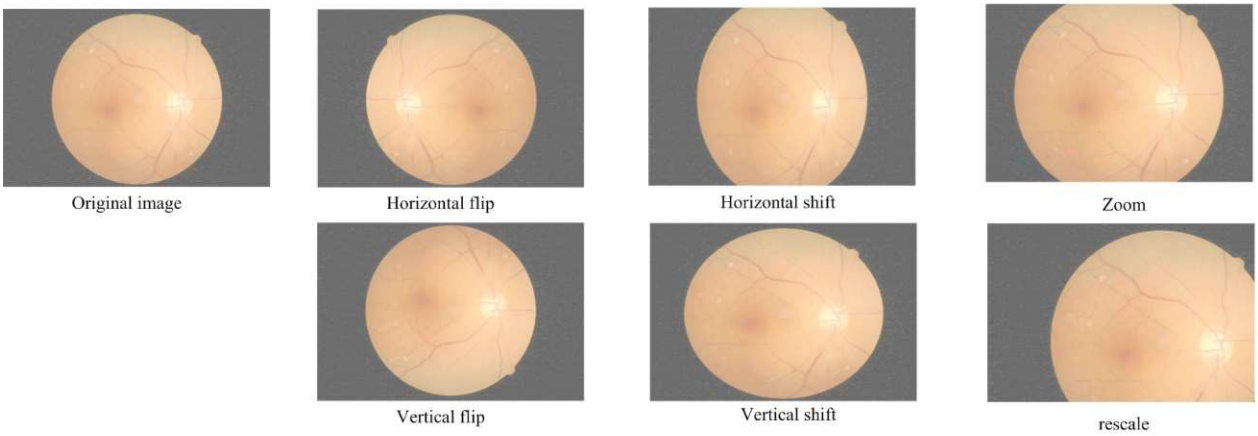

Figure 3. Different data augmentation processing.

3.2. DL model for cataract grading

3.2.1. Inception-V3

Inception-V3 is used for image classification, which is extended from the GoogLeNet network suggested by Google in 2014 [17]. Inception-V3 consists in performing several small convolutions on the same level instead of using large convolutions [18]. The first level of the Inception module is composed of three convolutions and one max-pooling. In the last level, the provided channels are non-linearly merged [19]. Therefore, the network parameters are reduced, which implies accelerating either the training or test steps. In addition, the features are efficiently extracted, which increases the classification accuracy [20]. Inception-V3 contains dense layers to increase the network depth, which decreases again the computational complexity. It is trained on the ImageNet dataset, to identify 1,000 classes [16]. Inception-V3 proves higher accuracy than the original GoogLeNet, Resnet-50 and AlexNet [20]. Due to its flexibility and accuracy, Inception-V3 achieves a higher performance when doing Transfer Learning (TL) even with small datasets [19].

3.2.2. MobileNet-V2

The MobileNet-V2 architecture contains convolutional layers, where each one is followed by a batch normalization and a ReLU6 nonlinear activation function, except the output layer [20, 21]. In MobileNet-V2, expensive convolutional layers are factorized into lightweight separable convolution blocks. In each block, the input is filtered through a (3x3) depthwise convolutional layer [23]. The provided output channels are projected via a (1x1) pointwise convolutional layer. The MobileNet-V2 blocks are built with residual connections that assist converging the network weights [24]. The use of (3x3) depthwise separable convolutions and the resolution modification through layers imply decreasing the computational complexity with respect to the standard convolutions despite a slight accuracy reduction [23, 25].

3.2.3. NasNet-Mobile

The Neural architecture search (Nas) framework is a research algorithm based on a neural architecture [27]. The objective is to provide an optimal Convolutional Neural Network (CNN) architecture for a given dataset, called Nas Network (NasNet). It automatically proposes a sequence of blocks that present a whole architecture. Thereupon, the provided NasNet architecture is unkown before the training process [28]. The NasNet research algorithm is based on a controller recurrent neural network that samples convolutional neural blocks and arranges them into different architectures. The controller uses reinforcement learning where the proposed architectures are trained and their accuracy is utilized to enhance the following ones [28, 29]. To classify whatever image size, the generated architectures are built using two types of convolutional cells [30]: (1) normal cells that provide a feature map having the same resolution than the input, and (2) reduction cells that provide a feature map where the height and the width are reduced by half. Following the same architecture building, the NasNet framework might provide two different architectures [29]. The first one is called NasNet-Large, which is composed of 84 million parameters. The second one is NASNet-Mobile made up of only four million parameters. This architecture is trained on ImageNet and performs acceptable classification. Despite the lower size, the TL of the NasNet-Mobile architecture maintains similar accuracy even with a smaller dataset [31].

3.3. Transfer learning and fine-tuning

TL and Fine-Tuning (FT) are well-known machine learning methods where the knowledge obtained through a trained model dedicated to a problem is used to resolve another one [32]. Generally, a great dataset is required to train a CNN, which is not guaranteed in several problems. For this purpose, both methods are used to avoid the problem of the dataset size, to enhance the performance, to reduce the redundancy and to accelerate the training [24]. The TL consists in training the model in order to adjust the objective of the related dataset to the target of the current one [32]. In the FT, the weight values are retrieved from the trained model and used as the initialization to be updated through the training process[24]. Due to the size of the dataset used in this study, our framework leads to extract the features of the pre-trained CNN model of Inception-V3, MobileNet-V2 and Nasnet-Mobile to be used for TL and FT. The selected CNN models are initially trained with the publicly "ImageNet" dataset that contains 1,000 categories [16].

The main goal of the targeted models is to classify the fundus image into: healthy, cataract mild stage, cataract moderate stage or cataract severe stage. Therefore, the last three layers are substituted to adapt the result into the aimed four classes. After the last global average pooling, a flatten layer is processed to merge the channel of the provided feature map. Thereafter, a first fully connected dense layer is applied with the ReLU activation function in order to reduce the feature map resolution. The last softmax fully connected dense layer is altered which is composed of four neurons to generate four probability distributions that correspond to the four output categories [32]. Furthermore, FT is preformed where the layer weights are unfreezing during training. Hence, their values are able to be updated based on the back propagation principle with a learning rate of 0.00001 [16]. A callback is activated to save the better weight after each epoch to improve the performance and accelerate model training. For the three DL architectures, the train process is performed using the "Adam" optimizer and the " categorical_crossentropy " loss function through 150 epochs.

3.4. Ensemble learning framework for cataract grading

An EL system is a learning model that clusters several classifiers, called base classifiers, to merge their decisions into a single one, called a meta classifier [33]. The main purpose is to provide a higher accurate decision than the ones separately given by each base classifier [33, 34]. These base classifiers are called homogeneous if they belong to the same algorithm trained to various datasets. In contrast, heterogeneous base classifiers are different methods learned from the same dataset. Two categories of EL systems are extensively used in the work appeared in the past few years [36]. The first one was the data level EL, where single classifiers were resampled to select the base classifier ones, such as the bagging method. The combination level was a second category of EL where individual classifiers were combined, including voting, ensemble selection and stacking. In the voting method, each base classifier would generate a label for each dataset instance, which were retrieved by the meta model that decided a unique label through a voting principle. Knowing that some base classifiers on an EL system might decrease the performance, the ensemble selection was a heuristic algorithm that allowed selecting an optimal subset of base classifiers that would guarantee a better performance. The base classifiers in the stacking technique provided their result as an input of the meta model. For the stacking EL, the training principles of classifiers were considerably different, which led to varied expertise. Stacking consists in combining the predictions of all base classifiers into another learning model [31, 36, 37].

Our main objective is to ensure cataract grading with a higher performance. The MobileNet-V2, Inception-V3 and Nasnet-Mobile models are selected as heterogeneous classifiers since they have different architectures and hence multiple prediction principles. Those models are known by their higher performance, even though their lightweight criteria prevent them from achieving optimal classification. Consequently, we design a stacked ensemble classifier where the three CNN models are considered as base-classifiers in order to combine the predictions. The classification capability and the greater difference of the base classifier leads to a better classification performance of the whole EL. The meta-classifier rectifies the prediction errors, reduces the bias carried out by the base classifiers and prevents the over-fitting effect[32, 35].Moreover, their difference allows benefiting from the advantages of each model to provide a complementary prediction. As a result, the meta-classifier ensures achieving high prediction accuracy. The choice of the meta-classifier is a critical issue in a way that the performance of the whole framework must be better than each base classifier, tested separately. For this purpose, several classifiers are experimentally evaluated, as detailed in section 4.2, where the SVM with the RBF

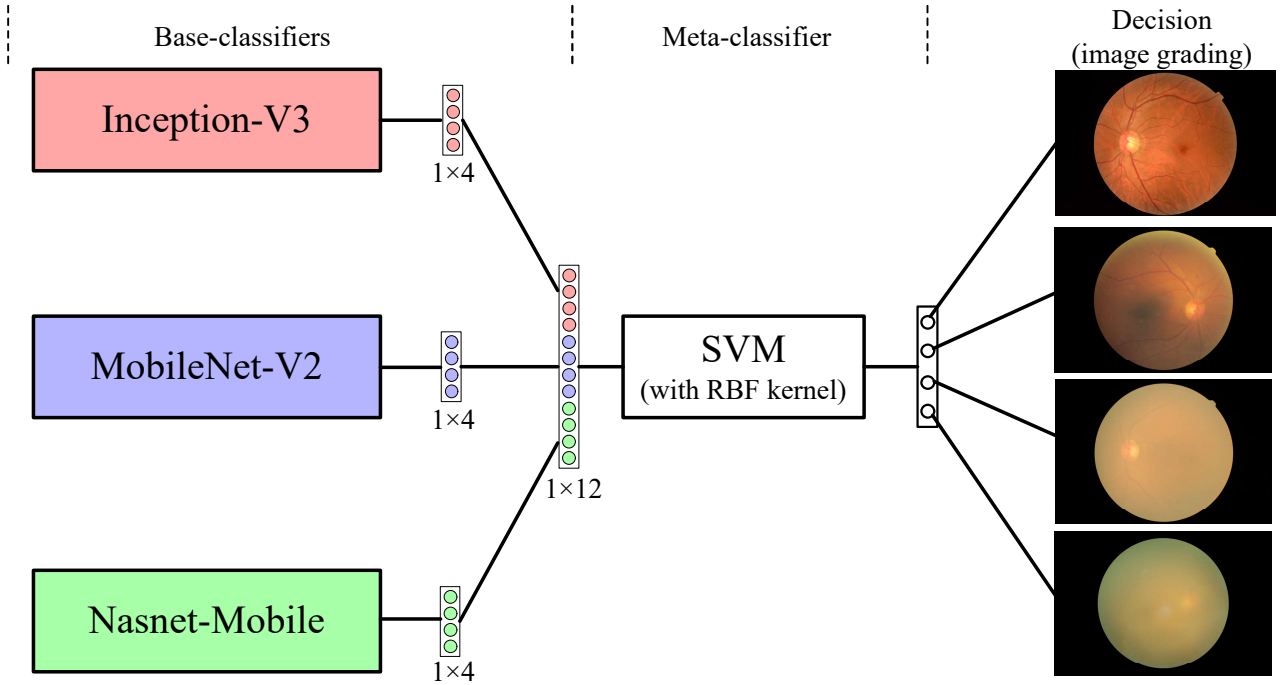

Figure 4. Processing pipeline of ensemble learning framework for cataract grading

kernel provides the optimal grading accuracy. The main steps of the EL framework, illustrated in Fig. 4, are as follows:

1) Extract the feature from each base classifier separately into a vector having a size of (1*4);
2) Concatenate the feature probability provided into three vectors provided from the base classifiers into a single meta-feature vector with a size of (1*12);
3) Provide the meta-feature vector to the meta classifier to classify the input image into cataract stages.

4. Experimental results

4.1. Dataset

Based on the literature review detailed in section 2, the evaluation of the cataract detection and grading methods is always conducted using private datasets. In our case, we construct a dataset of retinal images selected from two public databases recently published in the Kaggle platform. The "Cataract Dataset" [42] is composed of 600 fundus images having a size of 2592*1728, where 100 are affected by the cataract disease. The second database is called "Ocular Disease Recognition (ODiR)" [43] containing 8,000 fundus images with different image sizes, where 293 are reached by the cataract. The cataract-affected images of both databases are labeled through the intervention of two ophthalmologists belonging to different hospital centers.

The used dataset contains 590 fundus images, where 220 are confirmed as non-cataract, while 65, 145 and 160 images are diagnosed as mild, moderate and severe cataracts, respectively. Thereafter, each grading set is randomly partitioned into five subsets. three subsets, representing 60% of the whole dataset, are dedicated for training; each of the last two subsets are respectively used for validation and testing. To ensure a reliable evaluation, we conduct the evaluation through a 5-fold cross validation approach, where the five subsets are affected differently to the training, validation and testing process, as depicted in Fig.5.

4.2. Software environment and evaluation metrics

The framework is coded with python language and using the "Keras" API dedicated for Deep learning processing. The training and testing phase are run on the "google Colab" cloud service. To evaluate the performance of our method, five metrics are computed, which are Sensitivity (Sens), Specificity (Spec), Accuracy (Acc), Precision (Prec) and F-Measure (F-M) given in equations (1) – (5), respectively.

$$\text{Sens} = \frac{\text{TP}}{\text{TP}} \quad (1)$$

$$\text{Spec} = \frac{\text{TN}}{\text{TN}} \quad (2)$$

$$\text{Acc} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (3)$$

$$\text{Prec} = \frac{\text{TP}}{\text{TP+FP}} \quad (4)$$

$$\text{F-M} = \frac{2\times\text{TP}}{2\times\text{TP+FP+FN}} \quad (5)$$

where TP, TN, FP and FN are respectively True Positive, True Negative, False Positive and False Negative detected images.

4.3. Meta classifier algorithm choice

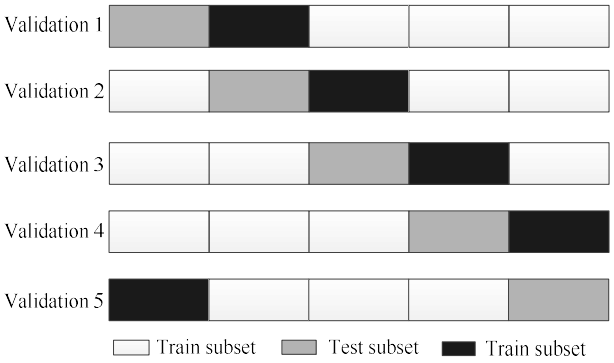

Figure 5. Dispatching of subsets for 5-fold cross validation

The aim of this sub-section is to identify the machine learning algorithm and its configuration, to be used as a meta classifier that allows achieving the highest performance of cataract grading. Within this objective, different machine learning algorithms are used having as an input the extracted features of all the three base classifiers. For instance, the Multi-Layer Perceptron (MLP) is employed, where the hidden layer size is varied among 8, 10 and 12, and where their results are shown in the second column of Table 1. For the Random Forest (RF), the tree number is fixed to 200, while the tree depth is varied following the above values {8, 10, 12}, where the performance metrics are illustrated in the third column. The SVMs are trained with the linear, RBF and

Table 1. Performance comparison between SVM and RF classifiers

Classifier		MLP				RF				SVM		
Parameters	Hidden layer size	8	10	12	Max depth	8	10	12	Kernel	Linear	Poly	RBF
Sensitivity		88.57	94.29	94.29		94.29	92.86	94.29		94.29	94.37	**95.59**
Specificity		91.67	87.50	85.42		85.42	87.50	87.50		85.42	89.36	**91.67**
Accuracy		89.83	91.53	90.68		90.68	90.68	91.53		90.68	92.37	**93.97**
Precision		93.94	91.67	90.41		90.41	91.55	91.67		90.41	93.06	**94.20**
F-Measure		91.18	92.96	92.31		92.31	92.20	92.96		92.31	93.71	**94.89**

polynomial kernels separately, where the achieved performance metrics are depicted in the fourth column. We deduce that using the SVM classifier with the RBF kernel allows achieving the higher performance of cataract grading, thus chosen as a meta classifier of the proposed framework.

4.4. Ensemble learning framework evaluation

4.4.1. Five-cross validation of the proposed framework

In this sub-section, we evaluate the cataract grading using the 5-fold cross validation approach based on the five metrics as shown in Table 2. For a better analysis of the achieved results, the box plots illustration is propounded in Fig.6 to present the performance of these validations.

Table 2. Cataract grading performance for 5-flod cross validation

Validations	Metrics				
	Sens	Spec	Acc	Prec	F-M
1	90,00%	91,67%	90,68%	94,03%	91,97%
2	94,29%	89,58%	92,37%	92,96%	93,62%
3	94,20%	89,80%	92,37%	92,86%	93,53%
4	94,12%	89,80%	92,31%	92,75%	93,43%
5	95,59%	91,67%	93,97%	94,20%	94,89%

We deduce that all validations achieved high performances in terms of the five metrics. The accuracy, sensitivity and F-measure values are very close to their average. Moreover, reduced gaps are deduced between values for specificity and precision with variation values of ±1.04% and ±0.72%, respectively. Accordingly, a higher performance is guaranteed whatever the fundus image dataset used for testing or training. This result approves of the robustness for the suggested framework when used in clinical context with new fundus images.

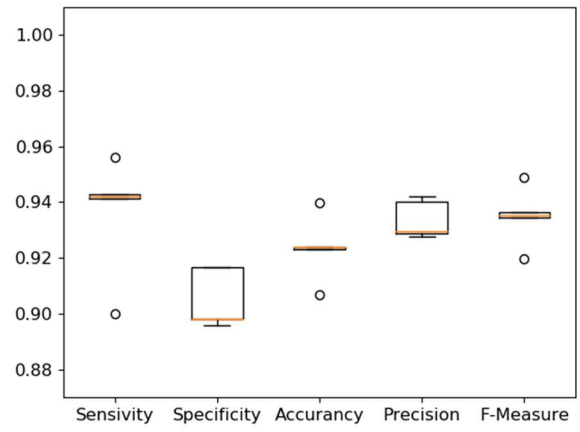

Figure 6. Performance visualization of cataract grading using box plots

4.4.2. Ensemble learning evaluation with respect to base classifiers

The four-grading performance of the suggested EL framework is evaluated and compared to the ones realized by each base classifier, separately. For this purpose, the experimentation consists in extracting features from each base classifier and given to an algorithm classifier to label images. To ensure a credible evaluation, the same SVM classifier with the RBF kernel is associated in the output of each DL architecture. In addition, the same subset of fundus images is used either for base classifiers or a meta-classifier, where their cataract-grading results is detailed in the confusion matrices of Fig.7.

The EL framework succeeds in properly classifying a larger number of fundus images, as deduced when adding and comparing blue cells for each matrix. Moreover, we deduce that it registers less numbers either of over-estimated or under-estimated fundus images, as noticed by half-matrices limited by the diagonal lines. The performance metrics are illustrated in Fig.8, where the EL framework significantly surpasses the base classifiers in all the five metrics.

Subsequently, we analyse the achieved result for each grade separately. For the non-cataract grade, the Inception-V3 architecture falls into four fundus images, in contrast to the two other DL architectures. Thereby, the EL framework settles for the features extracted by the Mobilenet-V2 and Nasnet-Mobile architectures. For the mild grade, all single classifiers fall into more than 33% of images. The proposed framework takes benefit from the diversity of all extracted features to enhance the final classification. In addition, the EL classification ensures the maximal labelling of moderate cataract affected images, like the Mobilenet-V2 architecture even with the lower accuracy achieved by Inception-V3 and Nasnet-Mobile. Finally, the framework takes the average results of the three base classifiers for the severe grade images, where it is overpassed by only one wrong image grading with respect to two base classifiers.

Figure 7. Confusion matrices of : (a) Inception-V3, (b) MobileNet-V2, (c) NasNet-Mobile, and (d) Ensemble learning

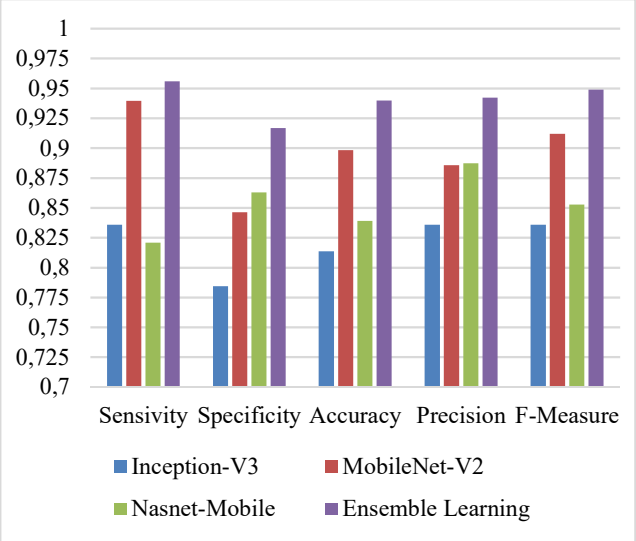

Figure 8. Performance metrics in terms of base classifiers and meta classifier

4.4.3. Cataract grading evaluation with respect to existing methods

We compare the grading performance of our method to the existing ones. The study is performed in terms of overall accuracy, where values are represented in the last column of Table 3. We deduce that our framework outperforms the methods based on a single CNN architecture [2, 11, 13]. By the way, those methods achieve accuracies similar the ones realized by any base-classifier, separately. Moreover, our framework surpasses the methods that employed an EL principles [3, 14]. The better grading result is due to the adequate choice of base classifiers. In addition, it is caused by the ensemble learning principle with consists at take benefit from all base classifiers rather than chosen the best ones.

Table 3. Performance comparison between existing methods of cataract grading

Methods of cataract grading	Classification algorithm	Acc (%)
Guo et al. [9] (2015)	Discriminant analysis algorithm	77.1
Xiong et al. [40] (2017)	Decision tree	83.84
Yang et al. [14] (2016)	Stacking EL of SVM and BPNN	84.5
Cao et al. [3] (2020)	Voting EL of two-class BPNN	85.98
Song et al. [4] (2019)	Eigenvalue computing + SVM	88.6
Zhou et al. [2] (2019)	CNN	89.23
Dong et al. [11] (2017)	CNN + SVM	90.82
Imran et al. [13] (2019)	SOM-RBF NN	91.7
Our method	Stacking EL of 3 CNN	**93.97**

5. Conclusion

The cataract is the first worldwide cause of vision impairment and blindness. The challenge is to propose a CAD system to grade the cataract from retinal images. For such a need, we have put forward in this paper an ensemble learning framework of stacking three DL architectures, where higher-performance cataract grading has been achieved. The realized performance has outperformed the ones realized by each single DL architecture. Accordingly, this framework will promote accessing to eye diagnosis, hence avoiding an advanced cataract state.

In our future work, we aim to extend our framework to consider other ocular pathologies, such as diabetic retinopathy, glaucoma or aged macular degeneration. Moreover, the employed lightweight learning machines, either as base-classifiers or as meta-classifiers, have low computational complexity. Consequently, the framework may be implemented in smartphone devices to be provided as a mobile-aided screening system for cataract grading, which will widely increase the accessibility to eye diagnosis.

Funding this work was supported by the PHC-UTIQUE 19G1408 Research program.

Conflict of interest The authors declare that they have no confict of interest to this work.

Acknowledgements The author would like to acknowledge the help of Dr. Aymen Daldoul and Dr. Nesrine Abroug, ophthalmologists respectively in the hospital center of Nevers agglomeration (France) and in Department of Ophthalmology of Fattouma Bourguiba University Hospital (Tunisia), to provide the groundtruth of retinal fundus images.

References

[1] S. Hu *et al.*, 'Unified Diagnosis Framework for Automated Nuclear Cataract Grading Based on Smartphone Slit-Lamp Images', *IEEE Access*, vol. 8, pp. 174169–174178, 2020, doi: 10.1109/ACCESS.2020.3025346.

[2] Y. Zhou, G. Li, and H. Li, 'Automatic Cataract Classification Using Deep Neural Network With Discrete State Transition', *IEEE Transactions on Medical Imaging*, vol. 39, no. 2, pp. 436–446, Feb. 2020, doi: 10.1109/TMI.2019.2928229.

[3] L. Cao, H. Li, Y. Zhang, L. Xu, and L. Zhang, 'Hierarchical method for cataract grading based on retinal images using improved Haar wavelet', Apr. 2019, *arXiv:1904.01261*.

[4] W. Song, Y. Cao, Z. Qiao, Q. Wang, and J. Yang, 'An Improved Semi-Supervised Learning Method on Cataract Fundus Image Classification', in *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, Jul. 2019, vol. 2, pp. 362–367, doi: 10.1109/COMPSAC.2019.10233.

[5] V. Agarwal, V. Gupta, V. M. Vashisht, K. Sharma, and N. Sharma, 'Mobile Application Based Cataract Detection System', in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, Apr. 2019, pp. 780–787, doi: 10.1109/ICOEI.2019.8862774.

[6] R. Sigit, M. Kom, M. B. Satmoko, D. K. Basuki, S. Si, and M. Kom, 'Classification of Cataract Slit-Lamp Image Based on Machine Learning', in *2018 International Seminar on Application for Technology of Information and Communication*, Sep. 2018, pp. 597–602, doi: 10.1109/ISEMANTIC.2018.8549701.

[7] W. L. Wong *et al.*, 'Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis', *Lancet Glob Health*, vol. 2, no. 2, pp. e106-116, Feb. 2014, doi: 10.1016/S2214-109X(13)70145-1.

[8] M. Akil and Y. Elloumi, 'Detection of retinal abnormalities using smartphone-captured fundus images: a survey', in *Real-Time Image Processing and Deep Learning 2019*, Baltimore, United States, May 2019, p. 21, doi: 10.1117/12.2519094.

[9] L. Guo, J.-J. Yang, L. Peng, J. Li, and Q. Liang, 'A computer-aided healthcare system for cataract classification and grading based on fundus image analysis', *Computers in Industry*, vol. 69, pp. 72–80, May 2015, doi: 10.1016/j.compind.2014.09.005.

[10] X. Qian, E. W. Patton, J. Swaney, Q. Xing and T. Zeng, "Machine Learning on Cataracts Classification Using SqueezeNet," *2018 4th International Conference on Universal Village (UV)*, Boston, MA, USA, 2018, pp. 1-3, doi: 10.1109/UV.2018.8642133.

[11] Y. Dong, Q. Zhang, Z. Qiao and J. Yang, "Classification of cataract fundus image based on deep learning," 2017 IEEE International Conference on Imaging Systems and Techniques (IST), Beijing, 2017, pp. 1-5, doi: 10.1109/IST.2017.8261463.

[12] Li Xiong, Huiqi Li, Liang Xu, "An Approach to Evaluate Blurriness in Retinal Images with Vitreous Opacity for Cataract Diagnosis", Journal of Healthcare Engineering, vol. 2017, Article ID 5645498, 16 pages, 2017. doi: https://doi.org/10.1155/2017/5645498

[13] A. Imran, J. Li, Y. Pei, F. Akhtar, J. Yang, and Q. Wang, 'Cataract Detection and Grading with Retinal Images Using SOM-RBF Neural Network', in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, Dec. 2019, pp. 2626–2632, doi: 10.1109/SSCI44817.2019.9002864.

[14] J.-J. Yang *et al.*, 'Exploiting ensemble learning for automatic cataract detection and grading', *Computer Methods and Programs in Biomedicine*, vol. 124, pp. 45–57, Feb. 2016, doi: 10.1016/j.cmpb.2015.10.007.

[15] N. Hnoohom and A. Jitpattanakul, 'Comparison of Ensemble Learning Algorithms for Cataract Detection from Fundus Images', in *2017 21st International Computer Science and Engineering Conference (ICSEC)*, Nov. 2017, pp. 1–5, doi: 10.1109/ICSEC.2017.8443900.

[16] F. Li, Z. Liu, H. Chen, M. Jiang, X. Zhang, and Z. Wu, 'Automatic Detection of Diabetic Retinopathy in Retinal Fundus Photographs Based on Deep Learning Algorithm', *Transl Vis Sci Technol*, vol. 8, no. 6, p. 4, Nov. 2019, doi: 10.1167/tvst.8.6.4.

[17] E. Kristiani, C. Yang, and C. Huang, 'iSEC: An Optimized Deep Learning Model for Image Classification on Edge Computing', *IEEE Access*, vol. 8, pp. 27267–27276, 2020, doi: 10.1109/ACCESS.2020.2971566.

[18] M. T. Hagos and S. Kant, 'Transfer Learning based Detection of Diabetic Retinopathy from Small Dataset', May 2019, *arXiv:1905.07203.*

[19] N. Dong, L. Zhao, C. H. Wu, and J. F. Chang, 'Inception v3 based cervical cell classification combined with artificially extracted features', *Applied Soft Computing*, vol. 93, p. 106311, Aug. 2020, doi: 10.1016/j.asoc.2020.106311.

[20] Y. Zhao, K. Xie, Z. Zou, and J.-B. He, 'Intelligent Recognition of Fatigue and Sleepiness Based on InceptionV3-LSTM via Multi-Feature Fusion', IEEE Access, vol. 8, pp. 144205–144217, 2020, doi: 10.1109/ACCESS.2020.3014508.

[21] Y. Wang, J. Yan, Q. Sun, J. Li, and Z. Yang, 'A MobileNets Convolutional Neural Network for GIS Partial Discharge Pattern Recognition in the Ubiquitous Power Internet of Things Context: Optimization, Comparison, and Application', *IEEE Access*, vol. 7, pp. 150226–150236, 2019, doi: 10.1109/ACCESS.2019.2946662.

[22] N. R. Gavai, Y. A. Jakhade, S. A. Tribhuvan, and R. Bhattad, 'MobileNets for flower classification using TensorFlow', in *2017 International Conference on Big Data, IoT and Data Science (BID)*, Dec. 2017, pp. 154–158, doi: 10.1109/BID.2017.8336590.

[23] A. S. Aguiar, F. N. D. Santos, A. J. M. D. Sousa, P. M. Oliveira, and L. C. Santos, 'Visual Trunk Detection Using Transfer Learning and a Deep Learning-Based Coprocessor', *IEEE Access*, vol. 8, pp. 77308–77320, 2020, doi: 10.1109/ACCESS.2020.2989052.

[24] A. Michele, V. Colin, and D. D. Santika, 'MobileNet Convolutional Neural Networks and Support Vector Machines for Palmprint Recognition', Procedia Computer Science, vol. 157, pp. 110–117, Jan. 2019, doi: 10.1016/j.procs.2019.08.147.

[25] M. Toğaçar, B. Ergen, and Z. Cömert, 'COVID-19 detection using deep learning models to exploit Social Mimic Optimization and structured chest X-ray images using fuzzy color and stacking approaches', *Computers in Biology and Medicine*, vol. 121, p. 103805, Jun. 2020, doi: 10.1016/j.compbiomed.2020.103805.

[26] H. Pan, Z. Pang, Y. Wang, Y. Wang, and L. Chen, 'A New Image Recognition and Classification Method Combining Transfer Learning Algorithm and MobileNet Model for Welding Defects', *IEEE Access*, vol. 8, pp. 119951–119960, 2020, doi: 10.1109/ACCESS.2020.3005450.

[27] A. Gupta, Anjum, S. Gupta, and R. Katarya, 'InstaCovNet-19: A deep learning classification model for the detection of COVID-19 patients using Chest X-ray', *Appl Soft Comput*, vol. 99, p. 106859, Feb. 2021, doi: 10.1016/j.asoc.2020.106859.

[28] T. Cogan, M. Cogan, and L. Tamil, 'MAPGI: Accurate identification of anatomical landmarks and diseased tissue in gastrointestinal tract using deep learning', *Computers in Biology and Medicine*, vol. 111, p. 103351, Aug. 2019, doi: 10.1016/j.compbiomed.2019.103351.

[29] A. Bahri, S. G. Majelan, S. Mohammadi, M. Noori, and K. Mohammadi, 'Remote Sensing Image Classification via Improved Cross-Entropy Loss and Transfer Learning Strategy Based on Deep Convolutional Neural Networks', *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 6, pp. 1087–1091, Jun. 2020, doi: 10.1109/LGRS.2019.2937872.

[30] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, 'Learning Transferable Architectures for Scalable Image Recognition', Apr. 2018, *arXiv:1707.07012.*

[31] A. S. Winoto, M. Kristianus, and C. Premachandra, 'Small and Slim Deep Convolutional Neural Network for Mobile Device', *IEEE Access*, vol. 8, pp. 125210–125222, 2020, doi: 10.1109/ACCESS.2020.3005161.

[32] M. Mateen, J. Wen, N. Nasrullah, S. Sun, and S. Hayat, 'Exudate Detection for Diabetic Retinopathy Using Pretrained Convolutional Neural Networks', Complexity, vol. 2020, Apr. 10, 2020, doi: https://doi.org/10.1155/2020/5801870.

[33] W. Książek, M. Hammad, P. Pławiak, U. R. Acharya, and R. Tadeusiewicz, 'Development of novel ensemble model using stacking learning and evolutionary computation techniques for automated hepatocellular carcinoma detection', *Biocybernetics and Biomedical Engineering*, vol. 40, no. 4, pp. 1512–1524, Oct. 2020, doi: 10.1016/j.bbe.2020.08.007.

[34] T. Zhou, H. Lu, Z. Yang, S. Qiu, B. Huo, and Y. Dong, 'The ensemble deep learning model for novel COVID-19 on CT images', *Applied Soft Computing*, vol. 98, p. 106885, Jan. 2021, doi: 10.1016/j.asoc.2020.106885.

[35] S. Agarwal and C. R. Chowdary, 'A-Stacking and A-Bagging: Adaptive versions of ensemble learning algorithms for spoof fingerprint detection', *Expert Systems with Applications*, vol. 146, p. 113160, May 2020, doi: 10.1016/j.eswa.2019.113160.

[36] D. Gupta and R. Rani, 'Improving malware detection using big data and ensemble learning', *Computers & Electrical Engineering*, vol. 86, p. 106729, Sep. 2020, doi: 10.1016/j.compeleceng.2020.106729.

[37] N. An, H. Ding, J. Yang, R. Au, and T. F. A. Ang, 'Deep ensemble learning for Alzheimer's disease classification', *Journal of Biomedical Informatics*, vol. 105, p. 103411, May 2020, doi: 10.1016/j.jbi.2020.103411.

[38] X.-N. Fan and S.-W. Zhang, 'LPI-BLS: Predicting lncRNA–protein interactions with a broad learning system-based stacked ensemble classifier', *Neurocomputing*, vol. 370, pp. 88–93, Dec. 2019, doi: 10.1016/j.neucom.2019.08.084.

[39] D. Han and L. Wang, 'Notice of Retraction: DIEN Network: Detailed Information Extracting Network for Detecting Continuous Circular Capsulorhexis Boundaries of Cataracts', *IEEE Access*, vol. 8, pp. 161571–161579, 2020, doi: 10.1109/ACCESS.2020.3021490.

[40] L. Xiong, H. Li, and L. Xu, 'An Approach to Evaluate Blurriness in Retinal Images with Vitreous Opacity for Cataract Diagnosis', Journal of Healthcare Engineering, vol. 2017, Apr. 26, 2017, doi: https://doi.org/10.1155/2017/5645498.

[41] A. Sharma, 'Emerging Simplified Retinal Imaging', *Dev Ophthalmol*, vol. 60, pp. 56–62, 2017, doi: 10.1159/000459690.

[42] Cataract Dataset, 2020. https://www.kaggle.com/andrewmvd/ocular-disease-recognition-odir5k

[43] Ocular Disease Recognition (ODiR), 2019. Retrieved from https://www.kaggle.com/jr2ngb/cataractdataset