Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2024.00121

Research Article

The evaluation of the effect of data balancing over the classification performances of ensemble of networks for the diabetic retinopathy

Mothna Mezher AL-RUBAYE[1,*], Hamza Osman İLHAN[1]

[1]*Department of Computer Engineering, Faculty of Electrical & Electronics, Yildiz Technical University, İstanbul, 34220, Türkiye*

ARTICLE INFO

Article history
Received: 29 April 2024
Revised: 07 August 2023
Accepted: 09 October 2023

Keywords:
Data Balancing; Diabetic Retinopathy; Ensemble of CNNs; Soft Voting

ABSTRACT

Diabetic retinopathy (DR) is a retinal condition that occurs due to diabetes mellitus and might lead to blindness. Early identification and treatment are crucial to slow down or prevent vision loss and degeneration. However, categorizing DR into several levels of severity remains a challenging problem due to the complexity of the disease. The Diabetic Retinopathy Grading System divides retinal pictures into five severity categories: No DR, Mild Non-Proliferative Diabetic Retinopathy (NPDR), Moderate NPDR, Severe NPDR, and Proliferative Diabetic Retinopathy. In this study, three deep learning models, namely ResNet50, Densenet201, and InceptionV3, were utilized for the classification of the APTOS 2019 diabetic retinopathy image dataset. For the individual experiments of the models, transfer learning with fine-tuning and layer freezing was applied. Additionally, a decision-level fusion idea using soft voting was implemented across the three pre-trained models. The maximum accuracy achieved for the classification of the original imbalanced dataset was 85% with the fusion idea. To further improve the classification performance, a balancing technique based on oversampling with augmentation operations was applied to the original APTOS 2019 dataset. The proposed approach, which involves the idea of soft voting-based fusion across models along with data balancing, improved the classification performance and achieved an accuracy of 90%.

Cite this article as: Al-Rubaye MM, İlhan HO. The evaluation of the effect of data balancing over the classification performances of ensemble of networks for the diabetic retinopathy. Sigma J Eng Nat Sci 2024;42(5):1563–1574.

INTRODUCTION

Diabetic Retinopathy (DR) is one of the critical diabetes complications that cause blindness for the diabetic patients. In order to prevent blindness's effect on further diabetic stages, eye screening tests should be performed repeatedly and periodically. But the diagnosis of DR over the color fundus images obtained by retinal cameras, requires experienced clinicians to identify the presence and also, the severity level of the disease [1]. Additionally, after the determination of DR, the regional analysis of each fundus image, which is a highly time-consuming process for clinicians, should be performed for the classification of the severity level. In this regard, computerized analyze systems can be used to alleviate these concerns and provide much more effective and accurate DR diagnose systems [2]. Most of the

***Corresponding author.**
*E-mail address: moshnamothna@gmail.com

This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkilic

computer's analyzes based on autonomous systems employ deep learning-based techniques which achieve comprehensive and high-performance outputs such as in the diagnosis of COVID19 [3]. morphological analysis of sperm [4], classification of the melanoma [5] etc.

Deep Learning (DL) is a field of machine learning that considers methods especially for analyzing the images using deep convolutional neural networks [2]. The "deep" term refers to the depth of the model which indicates the number of hidden layers in the network. The key factor in the performance of deep networks is the data. Deep learning achieves higher accuracy compared to the traditional machine learning techniques when the network is fed with enough data. The core of DL architecture consists of artificial neural networks (ANNs). In contrast to the traditional ANN, which remains quite limited, DL utilizes the Convolutional Neural Networks (CNNs) which are the special type of ANN with convolution layers designed to process pixel data (images) [6]. One of the most effective benefits of using DL is to use the raw data without the need of applying an extra feature extraction step [7].

The important features which summarize the image can be automatically extracted by the model convolution filters during the learning process in DL [6]. In the medical applications, due to this automatic feature extraction benefit of DL, pre-information about the disease and the significant effects of the disease over the images is no longer a requirement for developing an analyzing system. The feature extractor component in deep learning is a combination of convolution and pooling. A convolution is a multiplication of an image matrix (the actual matrix) and a kernel/filter (another smaller matrix). Convolution kernels are determined automatically during training to minimize the error between predicted and actual values. The pooling layer helps to reduce the spatial representation of the image to reduce the number of parameters and the amount of computation in the network.

In literature, various architectures have been developed, including Alex Net, Google Net, Residual Network (ResNet), VGG16 etc. [8]. Each architecture has a different number of parameters and combination of convolution, pooling, normalization, SoftMax, ReLU layers. Various studies have been carried out for the detection and severity classification of DR. Gayathri et al. employed feature extraction and selection techniques to derive the appropriate characteristics from retinal fundus images [9]. They used the MR-MR (maximum relevance-minimum redundancy) feature selection and ranking approach for the selection of top-ranked features. Then, a variety of machine learning classifiers including SVM, Naive Bayes, Random Forest, and multilayer perception (MLP), were utilized on three datasets (IDRiD, MESSIDOR, and DIARETDB0). They reported that MLP outperformed all other classifiers by using their suggested feature extraction and selection strategy for all datasets in terms of binary classification. In another study [10]. Gayathri et al. utilized the Wavelet Transform and Haralick feature extraction techniques. The directional characteristics in fundus images are consistently extracted by the Haralick features, which are based on second order statistics. They focused on both binary (DR and No DR) classification and multi-class structure including the severity levels of DR. For both scenarios (binary and multiclass), Random Forest was reported as the most accurate network when compared to SVM and Decision Tree over the classification of MESSIDOR, KAGGLE, and DIARETDB0 datasets.

With the growing popularity and effectiveness of deep learning-based networks, several customized networks have been designed especially for the classification of DR images. Pratta et al. proposed a custom CNN architecture to recognize the complex characteristics of micro-aneurysms, exudate, and retinal hemorrhages [11]. Their proposed model resulted in 75% accuracy over 5000 validation data. To extract deep features from the retinal fundus, Gayathri et al. used a lightweight CNN model [12]. The features from the CNN output were utilized in the several machine learning algorithms (SVM, AdaBoost, Naive Bayes, and Random Forest). They used the IDRiD, MESSIDOR, and KAGGLE datasets to test the models. Their findings demonstrated that combining CNN feature extraction with the J48 classifier resulted in an accuracy of 99.89% for binary classification and 99.59% for multi-class classification.

Macsik et al. proposed a local binary convolutional neural network (LBCNN) with deterministic filter generation mode which can approximate the performance of the conventional convolutional neural network (CNN) with fewer learnable parameters and with less memory utilization [13]. LBCNN architectures produced effective results when used with retinal fundus image datasets for binary classification. An average accuracy 89.71 was obtained for the EyePACS dataset.

For multiclass DR classification, Sarki et al. created an automated classification system using a CNN model [14] with image processing and optimization techniques. They reached a maximum accuracy of 86% on the original APTOS2019 dataset using the ResNet50[15]. According to their experiments performed over Eye PACS fundus image dataset, ResNet50 was the largest overfitting model. The most ideal, effective, and trustworthy DL algorithm for DR detection was reported as EfficientNetB4, followed by InceptionV3, NasNet Large, and DenseNet169. The maximum validation accuracy was obtained by EfficientNetB4 with an accuracy of 79.11%. DenseNet201 achieved a 76.80% accuracy score as the second place.

Addition to single usage of transfer learning-based models for the classification, hybrid approaches including regular CNNs with pre-trained networks such as DenseNet, ResNet etc. were utilized for the DR specific problems. Raja et al. designed two hybrid models, which contain a custom CNN with DenseNet and ResNet pre-trained architectures [16]. The characteristics of the eye are extracted using the suggested deep learning architectures. The accuracy of the

single CNN model was measured as 75%, whereas hybrid CNN with DenseNet resulted in 93% accuracy score. Patel and Chaware used the transfer learning approach with the MobileNetv2 network model customized for the classification of APTOS2019 dataset [17]. With the fine-tuning operations (freezing the first 169 layers), the validation accuracy had increased from 50% to 81%. This accuracy was achieved without any preprocessing augmentation or balancing processes. Alyoubi et al. implemented a deep learning-based model (CNN512) achieving an accuracy of 84.1% on the APTOS2019 dataset after the preprocessing stage [18]. The applied processing methods included image enhancement, noise removing, cropping, color normalization, and data augmentation. The CNN512 was designed using 32 layers including 6 convolution layers. Bodapati et al. designed a deep CNN using a transfer learning approach that integrates different representation forms of DR images acquired from the Xception and VGG16 models [19]. Despite the application of several structures, the recorded accuracy was measured as 82.54% on the APTOS2019 dataset. Agus et al. had investigated 3 different image preprocessing techniques with augmentation operations for the classification of the APTOS2109 dataset. The used model was EfficientNet-B7 with hyperparameters tuning. However, the best testing accuracy was 84% [20].

In this study, the performance evaluation of well-known pre-trained deep learning networks for the multi-level classification problem of DR has been investigated in terms of individual and combined usage of networks. A publicly available unbalanced dataset, APTOS2019, is used in the experiments. To enhance the classification performance, a data balancing approach, namely oversampling, has been applied to the dataset using data augmentation operations. The remainder of the research is organized into the following sections. Section II introduces the materials and the methods used in the experiments. The results of the experiments obtained by the individual and combined networks with data balancing pre-processing are presented in Section III. Lastly, the discussion and conclusions of the research are presented in Section IV.

MATERIALS AND METHODS

Initially, three pre-trained deep learning models were individually employed in the classification of the original APTOS 2019 dataset to evaluate the solo model performances. Additionally, each experiment was also repeated over the balanced version of the dataset created using the data augmentation techniques for oversampling. In the final decision step, a decision level fusion technique was implemented over the model predictions. The flowchart of the presented study is given in Figure 1. The details of the presented study will be given in subsections.

Dataset Information

Dataset for the Asia Pacific Tele-Ophthalmology Society 2019 Screening for Blindness (APTOS 2019), which was organized by Aravind Eye Hospital in India, was used in this study. APTOS 2019 dataset includes 3662 fundus images that were gathered from several rural Indian subjects [21]. Fundus photography was used by the hospital staff to collect retinal image samples from rural parts of India. These RGB fundus images were taken over an extended period. Later, a team of skilled physicians evaluated and classified the collected samples into five classes using the International Clinical Diabetic Retinopathy

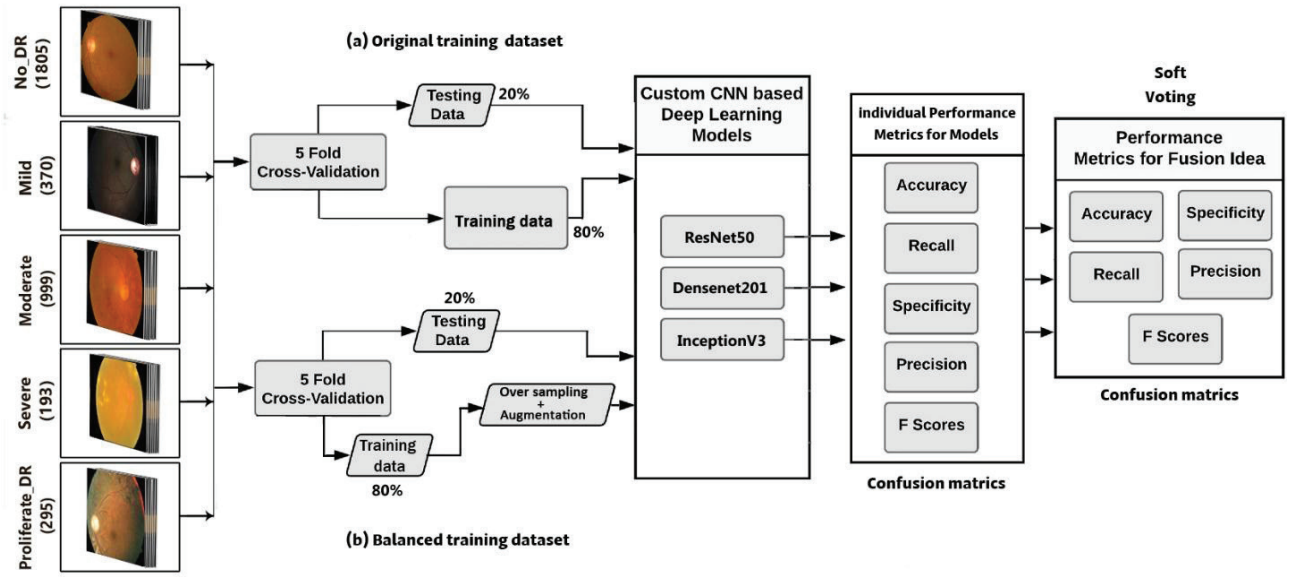
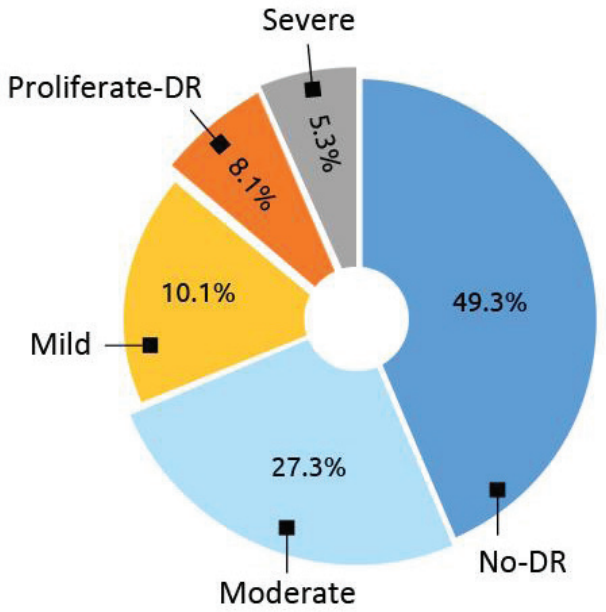


Figure 1. Flow chart of the proposed approach for classification technique.



Class Names	# of images
No-DR	1805
Mild	370
Moderate	999
Severe	193
Proliferate-DR	295
Total	3662

Figure 2. Class Names and Sample distributions of APTOS2019 dataset

Disease Severity Scale (ICDRSS). APTOS 2019 is an imbalanced dataset in which the sample distributions are not equal for each class. The distribution of the samples across the classes is given in Figure 2.

APTOS 2019 includes five classes as no DR, moderate DR, mild DR, proliferative DR, and severe DR. The healthy retinal samples were represented in the first group (no DR). The rest of classes represent the severity levels of the diabetic retinopathy disease. The most severe status is indicated by the Proliferative DR class, which indicates the images having retinal hemorrhage. Mild DR class is the initial stage of the DR. The resolutions of the images vary from 474 × 358 to 3388 × 2588 pixels in width and height, respectively. This difference might cause problems for further processing. Therefore, an image resizing procedure has been implemented before the network's initial layers as in [22].

Methodology

The flowchart of the study is given in Figure 1. As an initial part of the study, the dataset was organized according to a 5-fold cross validation schema to perform the experiments more objectively by using the same data for training and testing sets of each network evaluation. Addition to original training set based experiments, a data balancing idea based on oversampling was implemented over the original training set to evaluate the effects of data balancing approach to model performances. In the oversampling, data augmentation techniques were employed to augment the number of images in classes with limited samples to match the sample count of the most abundant class. The augmentation method was applied only for the training data. Otherwise, applying augmentation techniques to the test set will give higher accuracies in the classification due to biasing effects, but it will not represent the real classification accuracy for the APTOS dataset. Three pre-trained models as ResNet-50, Densenet201, and InceptionV3 were utilized for the experiments in terms of Transfer Learning idea. In the first stage of the study, each network was separately trained and evaluated using the original and balanced version of APTOS dataset. Then, decision-level fusion was performed over the predictions of individual models. The performances were evaluated using confusion matrix-based metrics. The details will be given in subsections.

K-Fold cross validation

K-Fold cross-validation is a widely used technique in machine learning and model evaluation. It is employed to assess the performance of a predictive model when dealing with a limited amount of data [23]. The process involves dividing the available dataset into K subsets of roughly equal size, where K is a positive integer usually chosen as 5 or 10. Each subset is referred to as a fold. In the implementation of K-Fold Cross validation, the original dataset is randomly divided into K folds. Each fold contains an equal number of samples. The evaluation is performed K times. During each evaluation, one-fold is selected as the test set, while the remaining (K-1) folds are combined to create the training set. The model is trained on the training set and then evaluated on the test set. This process is repeated K times, with each fold being used as the test set once. The performance of the model is measured in each evaluation using metrics based on the confusion matrix, such as accuracy. Additionally, the confusion matrix of each evaluation is stored. Once all K evaluations are completed, the confusion matrices of each evaluation are summed to obtain the confusion matrix for the whole dataset. Furthermore, the performance metrics from each iteration are averaged to derive a single performance score for the model. This average score represents the model's overall performance on the dataset [24].

In this study, K was selected as 5, in which 20% of the dataset is organized for the test set and the remaining 80% is reserved for training of the model for each fold. The evaluation is performed 5 times with a different test set to measure the model performances over the entire dataset. The final confusion matrix was obtained by summing up each fold confusion matrix.

Data balancing

The size of datasets has a critical impact on the model performances in DL. Small datasets cause overfitting in the training of the models while large datasets are scarce because of the exhaustive and time-consuming data acquisition process. Mostly, pretrained models can alleviate some of these concerns by extracting more informative features with the few data because of the enormous previous training procedure. Still, models might fail to generalize the performance for imbalanced datasets. APTOS 2019 dataset is also an imbalanced dataset, which has a variation in number of images for each class as shown in Figure 2. Therefore, an alternative comparison to original dataset-based performance evaluation, dataset balancing has been applied to original data in terms of over-sampling with data augmentation techniques in the presented study. The total number of images of the less sampled classes were increased to the total number of images in the highest sampled class with the data augmentation techniques as demonstrated in Figure 3. But this increment was performed only for the classes in the training set of each fold to avoid the biasing effect for evaluation of the model.

In the data augmentation step, spatial and pixel value effects such as scaling, reflection, adjusting the brightness, rotation, and blurring were performed over the original images to generate augmented versions of images. In terms of rotation, images were turned on both the vertical and horizontal axes or in a randomly chosen direction in a range between -30 and 30. In the same way, images were also scaled on either axis with a range of 0.8 - 1.2. For brightness adjustments, a Gaussian variance function was used (between 0.1 and 0.9) to both brighten and blur the images in the dataset randomly. Lastly, reflection for both X and Y directions were applied. The implemented effects were demonstrated over an example image in Figure 4.

Transfer learning

Classical learning is primarily based on the training of discrete isolated models for a particular tasks and datasets. Model cannot be used in another task because there is no transferable knowledge in the classical learning process. However, Transfer Learning (TL) is an optimization technique for machine learning where a model created for one job can be the basis for a model in another task. It permits the transfer of information from big to small data sets. TL

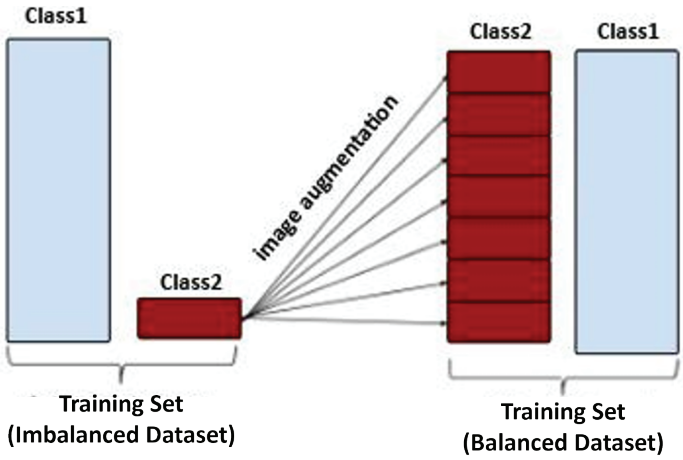


Figure 3. Demonstration of oversampling in training set.

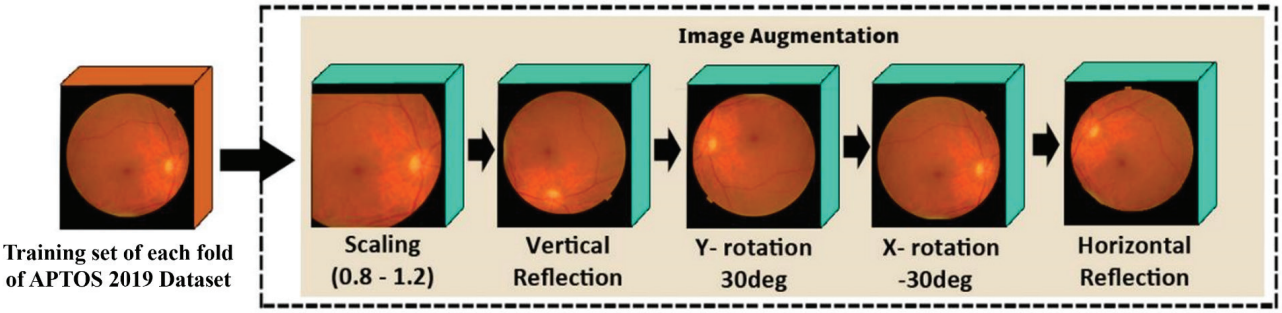


Figure 4. The effect of the implemented data augmentation techniques over an example image.

enables the use of information (features, weights, etc.) from previously learned models to train new models, which significantly speeds up the training process for huge datasets. It is a common method in deep learning to employ pre-trained models as a starting point in computer vision and similar applications since developing neural network models requires enormous computing and time resources. In this study, the output layers of models that have previously been trained were replaced with the class count of APTOS 2019 dataset. Then, each model was re-trained

with several hyper-parameters as part of a transfer learning technique known as fine-tuning. Although transfer learning requires the fewest number of parameters when compared to the model training from the scratch, still takes long training times to obtain the most accurate results [25].

In the implementation of TL, freezing some layers is feasible to speed up the training, improve model performance, and prevent the first few layers' weights from increasing excessively. When a model is trained without the layers being frozen at a predetermined length, the layers freeze one at a time, which is both ineffective and impractical. A precise layer freezing implementation can be applied so that several different models can be used quickly and extremely efficiently. However, freezing layers can frequently lead to overfitting problems. In this study, freezing of the different number of layers have been tested and freezing of the first 10 layers have been observed for a more efficient way in model training.

Decision-level fusion

The classification results of the individual models are different due to having different architectures. The combination of these results could increase the performance. This combination process for the resulting predictions is called decision-level fusion [26]. One of the fusion approaches is soft voting, which is based on the argmax of the sums of the predicted probabilities obtained from the outputs of each model. The soft voting depends on the obtained class assignment probabilities for each image to reduce the misclassification effect of incorrectly classified models. The soft voting prediction output can be written as in Equation 1.

$$\overline{y} = \arg\max_{l} \sum_{j=l}^{m} w_j \; p_{l,j} \tag{1}$$

where w_j is a weight that can be given to determine the contribution of each classifier and $p_{l,j}$ represents the predicted probability of the class label l and the classifier j. The idea of the soft voting is also demonstrated over an example image in Figure 5. When there are conflicts in the class assignments, a soft-voting approach can improve its performance in accordance with other network predictions that more precisely identify the sample [25].

The limitation of the soft voting technique is that the results will be incorrect if at least one of the three classification models does not provide a true result with a high probability value relative to the other two models. Furthermore, another critical limitation of the fusion approach emerges when all constituent models within the fusion structure collectively misclassify an image by assigning it to an incorrect class label. In such circumstances, the fusion technique encounters a significant challenge, as it lacks the capability to accurately determine the correct class assignment for the image. Consequently, this scenario undermines the effectiveness of the fusion idea, as the aggregated predictions from the individual models fail to yield an accurate consensus, thereby impeding the successful classification of the image.

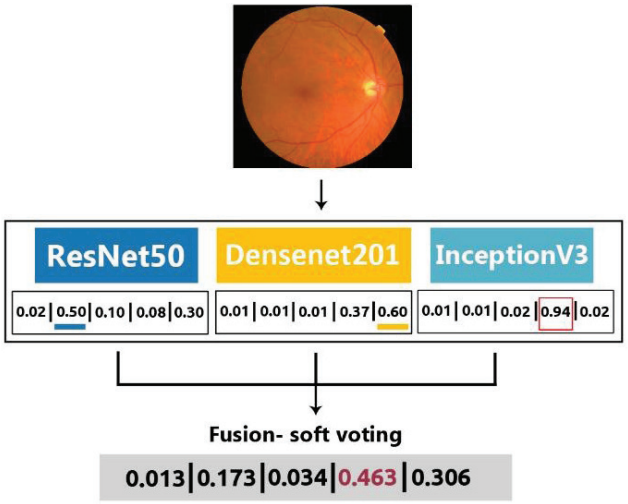


Figure 5. Demonstration of soft-voting technique.

Performance metrics

In the performance analysis of the experiments, five metrics such as Accuracy, Precision, Recall, Specificity and F1 Score were used to compare the models in terms of both individual and fusion performances. Metrics were calculated from the confusion matrices according to true positives, true negatives, false positives, and false negatives. In the equations, TP, and TN stand for true positive and true negative which represent the correctly identified class centered positive and negative samples, respectively. False positive (FP) and false negative (FN) indicate the number of mis-classified samples that are indeed negative and positive, respectively.

Specificity reflects the conditional probability of a true negative which has been given a secondary class. As such, it estimates the likelihood of a negative labeling. It is denoted by Equation 2.

$$Specificity = \frac{TN}{TN+FP} \tag{2}$$

Accuracy represents the most common stat for measuring the performances of models which directly informs that

how many samples were properly identified. It was calculated as in Equation 3.

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

The most significant performance metric used for classifying issues for balanced data sets is often accuracy scores. Accuracy, however, is not useful for specific tasks like analyzing the system consistency for a certain class. It is necessary to compute the Precision and Recall metrics. The Precision score represents how well class-based, accurately predicted samples performed. It was calculated as in Equation 4.

$$Precision \frac{TP}{TP + FP} \quad (4)$$

How percent of the data the system should correctly predict is expressed by the recall metric, which is calculated by Equation 5.

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

F-score offers more informative comparison due to including both precision and recall scores. The harmonic average of the Recall and Precision values is used to calculate the F-score rather than the arithmetic mean as shown in Equation 6. The harmonic average is used because it lessens the impact of outliers on the average.

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (6)$$

RESULTS AND DISCUSSION

Three pre-trained deep network architectures as ResNet50, DenseNet201 and Inception-v3 were utilized in the experiments. As an initial step, APTOS 2019 dataset was re-organized and stored according to a 5-fold cross validation schema to compare the models more objectively by using the identical training and testing sets for each fold in the experiments. In the training of the models, the processes of fine tuning for hyper-parameter determination and layer freezing to use the pretrained model weights were employed. After getting individual classification results of each model, the soft voting-based class fusion technique was performed. The same procedure was repeated over the balanced dataset after oversampling the training sets of the original dataset with the data augmentation techniques. In the data balancing step, data augmentation techniques were not applied to the images in test sets of each fold to avoid biasing effects.

The experiments were performed using the MATLAB 2021a over GPU-accelerated hardware. GeForce 970x (6 GB) was employed in the training phase of the networks to accelerate the process. In addition to the GPU module, the testing platform had 16 GB of local ram and Intel i7 CPU with a 3.2 GHz processing speed. The determination of the hyper-parameter has a critical impact for each deep learning-based study. Therefore, different parameters for epoch size, minibatch size and learning rate have been tested as preliminary study. After these hyperparameter tests, the final parameters for the training of the models were determined as 20, 32 and 10^{-4} for epoch size, batch size and learning rate, respectively.

The individual and ensemble (decision level fusion) classification performances of the models over the original dataset is given in Table 1. In the individual experiments, similar performance scores were obtained. The highest classification accuracy for the individual usage of the models was measured as 80% for the ResNet50 over the original imbalanced dataset.

In addition to general accuracy scores, the confusion matrices of the individual classification performances were given in Figure 6. High classification result (above 95%) is achieved for the first class (no DR). Misclassification is mostly observed between the first two stages of disease (Between the mild and moderate stages of the disease). The images of third class (moderate) are more distinctive than other stages of DR. Models failed in the classification of the images for second, fourth and fifth classes (Mild, Severe and Proliferate stages of disease). Therefore, precision and recall scores were measured much lower.

In the ensemble of the individual models, the highest accurate predictions of the individual models were aimed

Table 1. Individual and Ensemble model performances over the original imbalanced dataset

	Individual Model Performances			Ensemble (Decision Level Fusion)
	ResNet50	**Densenet201**	**InceptionV3**	**Soft Voting**
Precision	64%	68%	63%	**77%**
Recall	62%	65%	61%	64%
Accuracy	80%	78%	79%	**85%**
Specificity	94%	95%	94%	**96%**
F1-Score	63%	66%	62%	**70%**

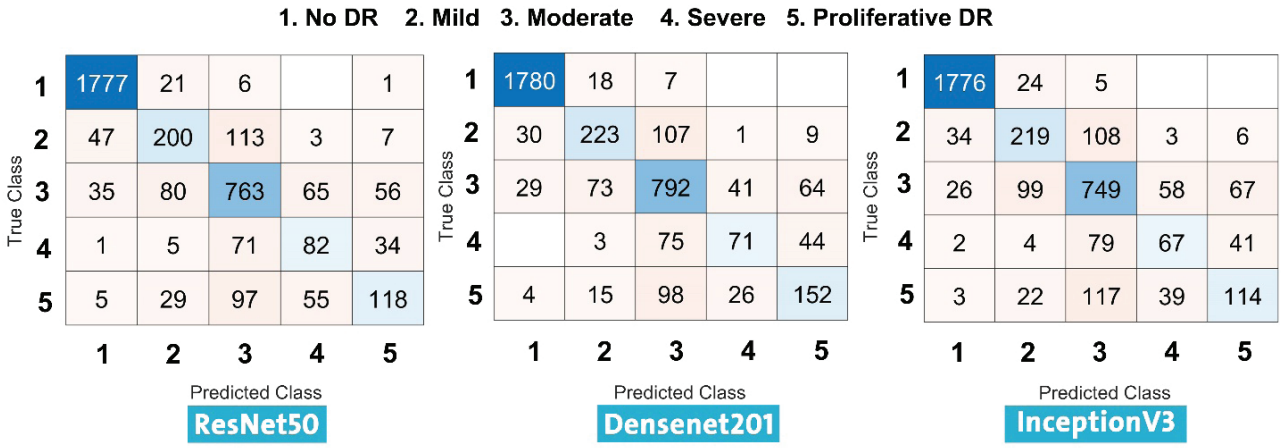


Figure 6. Confusion matrices for individual usage of the models over imbalanced dataset.

to select in the classification step by using the decision-level fusion idea including soft voting implementation over the prediction results of each model. The probabilities for every class for every image classification prediction were obtained to apply the voting process. The ensemble results over the original imbalanced dataset are presented in the right column of Table 1. The maximum accuracy is achieved as 85% with the ensemble idea compared to 80% obtained by the individual usage of ResNet50. Precision was also significantly increased. The confusion matrix of the soft voting-based ensemble idea is shown in Figure 7. The dramatic effect of the model fusion was observed over the third class (moderate DR) when compared to individual model outputs. Alternatively, precision scores increased by almost 10% with the decision level fusion idea.

APTOS 2019 dataset is an imbalanced dataset in which the distributions of the number of images per class is different. In this study, a dataset balancing approach, oversampling, was performed to training sets of the models to evaluate the balancing effect over the model classification performances. In the oversampling process, data augmentation techniques were used to generate the new images. Test sets were reserved with the original images due to avoiding the bias effects. According to the presented results

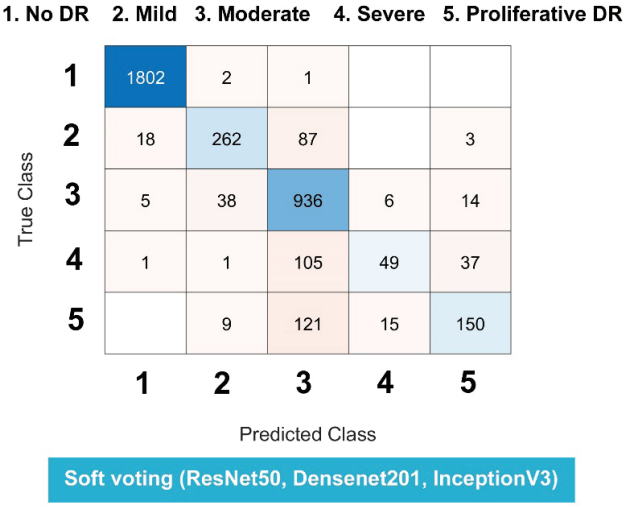


Figure 7. Confusion matrix of model ensemble approach over the imbalanced dataset.

in Table 2, the individual performance of DenseNet201 was improved by 4% compared to the network trained by an imbalanced dataset. Additionally, the performances of ResNet50 and InceptionV3 were also increased by around

Table 2. Individual and Ensemble model performances over the balanced dataset by oversampling

	Individual Model Performances			Ensemble (Decision Level Fusion)
	ResNet50	**Densenet201**	**InceptionV3**	**Soft Voting**
Precision	68%	69%	65%	**77%**
Recall	65%	67%	63%	64%
Accuracy	82%	82%	80%	**90%**
Specificity	95%	95%	94%	**96%**
F1-Score	66%	68%	64%	**70%**

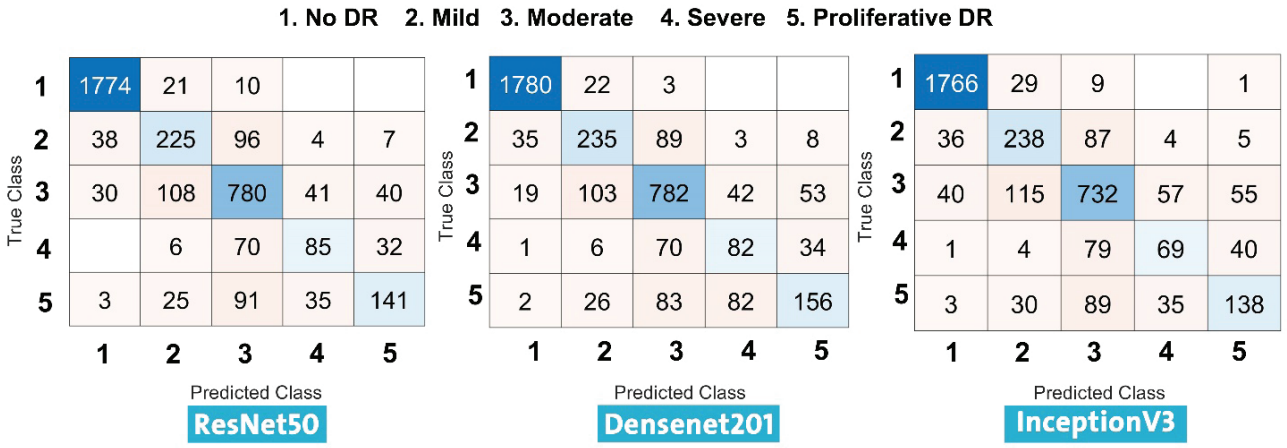


Figure 8. Confusion matrices for individual usage of the models over balanced dataset.

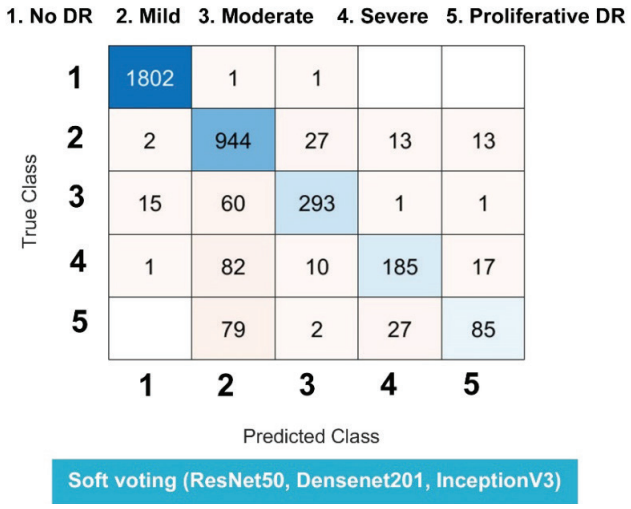


Figure 9. Confusion matrix of model ensemble approach over the balanced dataset by oversampling.

2%. 82%, 82%, and 80% accuracies have been measured for the individual usage of ResNet50, Densenet201, and InceptionV3 by using the balanced dataset, respectively.

The confusion matrices of the networks are shown in Figure 8. The number of correctly classified images (matrix diagonal elements) increased for classes 2, 4, and 5 compared to the experiments conducted with the imbalanced dataset. In other words, balancing the dataset resulted in more accurate class assignments for the mild, moderate, and proliferative stages of DR.

Applying the soft voting over the individual model results, an accuracy of 90% was achieved with the balancing dataset, as shown in the right column of Table 2. Similar to the results obtained with the unbalanced dataset, the ensemble ideas of the models significantly improved the precision scores of individual networks in the experiments conducted with the balanced dataset. The confusion matrix of the fusion approach is demonstrated in Figure 9. The number of correctly classified images had increased for all classes, especially the 4th (Severe DR) and 5th (Proliferate DR) classes.

The accuracy of the classification is considered the most important criteria. However, the training time and the complexity of the methods are also important factors. These factors are greatly related to the size of the used architecture in the model. There are different transfer deep learning architectures. In Table 3, the architectural structure complexities of the utilized models are given with the classification performances and the training times for APTOS2019 dataset.

The depth of a network is directly related to the number of layers it contains, and each layer also contains parameters. Each layer of a network learns features by processing data from previous layers, which may require adjusting more parameters. However, increasing depth does not

Table 3. Complexity and Training times of utilized pretrained models

Model	Size	Parameters	Depth	Imbalanced Dataset		Balanced Dataset	
				Accuracy	Training Time	Accuracy	Training Time
ResNet50	98MB	25M	50	80%	39 min	82%	335 min
InceptionV3	92MB	23M	159	79%	45 min	80%	315 min
DenseNet201	80MB	20M	201	78%	73 min	82%	326 min

Table 4. Performance Result Summary of the Related Works

Paper	Method	Accuracy
Sarki et al. 2019 [14]	ResNet50 with fine-tuning, data augmentation, and volume increase on original APTOS	86%
Wang L. and Schaefer A. 2020 [27]	Transfer learning to a pretrained MobileNetV2 and used a weighted loss function	77%
Pak et al. 2020 [28]	DenseNet,, ResNet and Inception with an initial preprocessing stage	79%
Li et al. 2020 [29]	InceptionV3, ResNet-50, DenseNet with data preprocessing (Noise removing, normalization and augmentation) .	80%
Lazuardi et al. 2020 [30]	EfficientNet-B4 and EfficientNet-B5 with CLAHE (contrast limited adaptive histogram equalization) and image central cropping.	83.87% 83.89%
Patel R. and Chaware A 2021 [17]	Implementation of MobileNetV2 using Transfer Learning and fine-tuning operations.	81%
Alyoubi et al. 2021 [18]	deep learning-based model (CNN512) with image enhancement, noise removing, cropping, color normalization and data augmentation.	84.1%
Bodapati et al. 2021 [19]	Combination of a custom deep CNN model with the integration of Xception and VGG16 architectures.	82.54%
Oulhadj et al. 2022 [31]	Densenet-121, Xception, Inception-v3, Resnet-50 with an initial preprocessing stage including the elimination the effect of the background.	85.28%
Agus et al. 2022 [20]	EfficientNet-B7 with hyperparameters tuning after 3 different image preprocessing techniques and augmentation operations.	84%
Proposed	**Soft voting over three pretrained models (ResNet, DesNet, Inception) with data balancing**	**90%**

always guarantee better results. Deeper networks can have more learning capacity, but they might also require more data and have longer training times. Additionally, they can become prone to overfitting. When using the deep learning networks, it's important to carefully balance both depth and the number of parameters to enhance the model's performance. This balance can vary depending on the specific tasks or datasets. For the APTOS 2019 dataset, deeper networks such as Inception and DenseNet cannot surpass the less layered network, ResNet, in classifying the original imbalanced dataset due to data scarcity. When the dataset was balanced with data augmentation, all the networks resulted in nearly identical classification accuracies. However, the training times of the models were extended. In the classification of the balanced version, three networks completed the training period in similar times while ResNet, being a lightweight network in comparison to the others in terms of depth, completed the training in a shorter time in the original dataset classification.

In Table 4, the results of the proposed approach were compared to the novel studies which had only covered APTOS 2019 dataset and used the similar deep learning methods. The proposed model outperforms similar studies by achieving a classification accuracy of 90% on a dataset with 5 classes. This improvement can be attributed to the implementation of data balancing and decision-level fusion of three pre-trained architectures as ResNet, DenseNet and Inception.

CONCLUSION

Addition to the fusion idea of the deep networks, a specific novel approach was implemented to enhance the accuracy of classifying the APTOS2019 dataset in this study. The challenge with this dataset lies in its imbalance and the limited number of images for the 4th and 5th classes. Balancing the dataset holds significance in ensuring equal representation of various data categories. This balance plays a crucial role in machine learning algorithms, as they may exhibit bias when data is imbalanced. Numerous techniques are available to balance the dataset, one of which is oversampling. This method can be utilized to improve the performance of machine learning models, especially when addressing small minority groups. In this study, data balancing was applied using the idea of oversampling and the augmentation techniques. Over the balanced dataset, the performance of the individual models and the fusion idea were also evaluated. As a result, the classification performance of the APTOS2019 diabetic retinopathy image data set increased to 90% from 85% by data balancing. Results proved that the proposed fusion idea over the balanced dataset was able to classify retinal images with high accuracy. Therefore, an automated examination system based on the proposed approach helps non-experts' ophthalmologists in the initial examination process of the diabetic retinopathy images.

In the future study, the classes with low samples of this data set could be increased by adding new real samples. These samples could be taken from certain hospital records to recreate more informative and balanced data. Thus, the accuracy of the models will greatly improve. Additionally, different deep learning models including the fundus image centered designs will be tested in the fusion manner.

AUTHORSHIP CONTRIBUTIONS

Authors equally contributed to this work.

DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

REFERENCES

[1] Mohamed Q, Gillies MC, Wong TY. Management of diabetic retinopathy: a systematic review. JAMA 2007;298:902–916. [CrossRef]

[2] Akram MU, Khalid S, Tariq A, Khan SA, Azam F. Detection and classification of retinal lesions for grading of diabetic retinopathy. Comput Biol Med 2007;45:161–171. [CrossRef]

[3] Shorten C, Khoshgoftaar TM, Furht B. Deep learning applications for COVID-19. J Big Data 2021;8:1–54. [CrossRef]

[4] Bodapati JD, Veeranjaneyulu N. Feature extraction and classification using deep convolutional neural networks. J Cyber Secur Mobil 2019;261–276. [CrossRef]

[5] Li Y, Shen L. Skin lesion analysis towards melanoma detection using deep learning network. Sensors (Basel) 2018;18:556. [CrossRef]

[6] Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J. Deep learning-based text classification: a comprehensive review. ACM Comput Surv 2021;54:1–40. [CrossRef]

[7] Kussul N, Lavreniuk M, Skakun S, Shelestov A. Deep learning classification of land cover and crop types using remote sensing data. IEEE Geosci Remote Sens Lett 2017;14:778–782. [CrossRef]

[8] Adnan MM, Rahim MSM, Rehman A, Mehmood Z, Saba T, Naqvi RA. Automatic image annotation based on deep learning models: a systematic review and future challenges. IEEE Access 2021;9:50255–50256. [CrossRef]

[9] Amin J, Sharif M, Anjum MA, Raza M, Bukhari SAC. Convolutional neural network with batch normalization for glioma and stroke lesion detection using MRI. Cogn Syst Res 2020;59:304–311. [CrossRef]

[10] Gayathri S, Gopi VP, Palanisamy P. Automated classification of diabetic retinopathy through reliable feature selection. Phys Eng Sci Med. 2020;43:927–945. [CrossRef]

[11] Gayathri S, Krishna AK, Gopi VP, Palanisamy P. Automated binary and multiclass classification of diabetic retinopathy using Haralick and multiresolution features. IEEE Access 2020;8:57497–57504. [CrossRef]

[12] Pratt H, Coenen F, Broadbent DM, Harding SP, Zheng Y. Convolutional neural networks for diabetic retinopathy. Procedia Comput Sci 2016;90:200–205. [CrossRef]

[13] Macsik P, Pavlovicova J, Goga J, Kajan S. Local binary CNN for diabetic retinopathy classification on fundus images. Acta Polytech Hung 2022;19:29–31. [CrossRef]

[14] Sarki R, Michalska S, Ahmed K, Wang H, Zhang Y. Automatic detection of diabetic eye disease through deep learning using fundus images. bioRxiv 2020;8:763136. [CrossRef]

[15] Das D, Biswas SK, Bandyopadhyay S. Detection of diabetic retinopathy using convolutional neural networks for feature extraction and classification (DRFEC). Multimed Tools Appl 2022;82:29943–30001. [CrossRef]

[16] Raja SMV, Panjanathan R. Diabetic retinopathy classification using CNN and hybrid deep convolutional neural networks. Symmetry (Basel) 2020;14:1932. [CrossRef]

[17] Patel R, Chaware A. Transfer learning with fine-tuned MobileNetV2 for diabetic retinopathy. 2020 Int Conf Emerg Technol (INCET). 2020;1–4. [CrossRef]

[18] Alyoubi WL, Abulkhair MF, Shalash WM. Diabetic retinopathy fundus image classification and lesions localization system using deep learning. Sensors (Basel) 2021;21:3704. [CrossRef]

[19] Bodapati JD, Shaik NS, Naralasetti V. Composite deep neural network with gated-attention mechanism for diabetic retinopathy severity classification. J Ambient Intell Humaniz Comput 2021;12:9825–9839. [CrossRef]

[20] Agus EM, Mochammad HCM, Yufis A, Fitri B, Hanung AN, Zaidah I. Classification of diabetic retinopathy disease using convolutional neural network. Int J Inform Vis 2022;6:12–18. [CrossRef]

[21] Khalifa NEM, Loey M, Taha MHN, Mohamed HNET. Deep transfer learning models for medical diabetic retinopathy detection. Acta Inform Med 2019;27:327. [CrossRef]

[22] Sikder N, Masud M, Bairagi AK, Arif ASM, Nahid AA, Alhumyani HA. Severity classification of diabetic retinopathy using an ensemble learning algorithm through analyzing retinal images. Symmetry (Basel) 2021;13:670. [CrossRef]

[23] Rodriguez JD, Perez A, Lozano JA. Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE Trans Pattern Anal Mach Intell 2009;32:569–575. [CrossRef]

[24] Perez L, Wang J. The effectiveness of data augmentation in image classification using deep learning. arXiv 2017;1712.04621.

[25] Yüzkat M, Ilhan HO, Aydin N. Multi-model CNN fusion for sperm morphology analysis. Comput Biol Med 2021;137:104790. [CrossRef]

[26] Kieu LM, Ou Y, Truong LT, Cai CA. Class-specific soft voting framework for customer booking prediction in on-demand transport. Transp Res C Emerg Technol 2020;114:377–390. [CrossRef]

[27] Wang L, Schaefer A. Diagnosing diabetic retinopathy from images of the eye fundus. CS230. Stanford. Edu.

[28] Pak A, Ziyaden A, Tukeshev K, Jaxylykova A, Abdullina D. Comparative analysis of deep learning methods of detection of diabetic retinopathy. Cogent Eng 2020;7:1805144. [CrossRef]

[29] Li Y, Hsu JS, Bari N, Qiu X, Viswanathan M, Shi W, et al. Interpretable evaluation of diabetic retinopathy grade regarding eye color fundus images. 2022 IEEE Int Conf Biomed Eng Informat (BIBE) 2022;11–16. [CrossRef]

[30] Lazuardi RN, Abiwinanda N, Suryawan TH, Hanif M, Handayani A. Automatic diabetic retinopathy classification with EfficientNet. 2020 IEEE Reg 10 Conf (TENCON). 2020;10:756–760. [CrossRef]

[31] Oulhadj M, Riffi J, Chaimae K, Mahraz AM, Ahmed B, Yahyaouy A, et al. Diabetic retinopathy prediction based on deep learning and deformable registration. Multimed Tools Appl 2022;81:28709–28727. [CrossRef]