Transfer Learning-Based Stack Ensemble Deep Learning Approach to Predict the Severity of Diabetic Retinopathy

Deva Kumar S[1,*], Venkatramaphanikumar S[1], Venkata Krishna Kishore Kolli[1]

[1]Department of Computer Science & Engineering, Vignan's Founation for Science, Technology & Research, Vadlamudi, Guntur, Andhra Pradesh, India

Corresponding Author: sdevakumar31@gmail.com

ABSTRACT

Diabetic retinopathy (DR) is the most commonly occurring eye disorder and the main reason underlying blindness in diabetics all around the world. Many technologies have emerged today for the accurate diagnosis of DR at an early stage. Of these, deep learning (DL) is one of the most effective methods. This research focuses on the prediction of DR severity into five classes, Normal, Mild, Moderate, Severe, and Proliferative DR (PDR), using pre-trained models. Transfer learning using models, such as EfficientNetB0, MobileNet, and Xception, were implemented with customization. Further, the Stack Ensemble model was applied to combine the predictions of all these pre-trained models using meta classifiers, such as Random Forest and Extra Trees Classifier to grade the DR severity. The performance the proposed model was evaluated on the KAGGLE and the Asia Pacific Tele-Ophthalmology Society (APTOS) retina datasets. The final outcome revealed that the proposed model outperformed state-of-the-art pre-trained models, with an accuracy of 0.96 and 0.97 on the KAGGLE and APTOS datasets, respectively.

Keywords: Diabetic Retinopathy, EfficientNetB0, MobileNet, Xception, Stack Ensemble, Random Forest Classifier, Extra Trees Classifier

DOI Number: 10.48047/NQ.2022.20.20.NQ109072 **NeuroQuantology2022;20(20): 711-725**

I. INTRODUCTION

Diabetes incidence is increasing rapidly, making it one of the most common metabolic disorders in adults over the last few decades. Most of the people are unaware about diabetes characteristics and related complications. Of its complications, diabetic retinopathy (DR) has gained a lot of attention. The World Health Organization (WHO) states that diabetes, a polygenic disease, will be the seventh leading cause of death [1]. According to statistics, there are approximately 61.3 million diabetes patients between the ages of 20 and 79. By 2030, it is predicted that affected population will rise to 102 million [2]. About seventy-five percent of diabetics in the American, European, and Asian countries have been estimated to be suffering from a specific kind of DR [3]. Many studies introduced several digital screening methods to identify DR in its early stage using fundus images. The digital screening of fundus images helps ophthalmologists identify the severity of DR easily and rapidly. The emergence of digital screening methods has led to the development of automated techniques for DR detection. These automated techniques have met ophthalmologists' expectations and received a positive response.

DL models are capable of learning directly from retinal images and CNN architectures are meant to segment exudates. CNN models are more successful with respect to the prediction of DR severity in its early stages from the color fundus retinal images. Deep models lack generalizability and they require large-scale annotated data. Pre-trained models, such as AlexNet, VGGNet, ResNet, etc., are outperforming with higher accuracy in retinal image classification.

Many diagnostic algorithms have been developed in recent years that can automatically identify DR and categorize fundus images. It takes a long time to extract handcrafted features. Thus, a new technology which is able to learn features by itself grabbed the researcher's attention. Artificial intelligence (AI) helps in mass screening for DR and telemedicine. It helps monitor the disease's progression while reducing the inter-person variability during DR categorization. Based on severity, there are five classes of DR: No DR, Mild DR, Moderate DR, Severe DR, and Proliferative DR (PDR). Their features are listed in Table 1.

Table 1: Diabetic Retinopathy Types and their Features

DR Types	Standard features of fundus images
No DR	No abnormalities
Mild	Signs of Microaneurysms
Moderate	Numerous signs of Microaneurysms, Exudates, and Hemorrhages
Severe	Four quadrants of the retina have irregular characteristics.
Proliferative DR	Vitreous hemorrhage, Severe retinal proliferative

DL models [4] are capable of processing data and feature extraction from images. For Data imaging, DL can be used along with CNNs. CNNs learn how to perform their tasks [5] by correcting itself and undergo multiple repetitions. They guide themselves by analyzing labeled data and provide relevant output. CNN adjusts its hyper parameters to reduce error rate.In DL we can implement pre-trained models for classification of severity of DR images. Images are pre-processed to resize them and make them uniform. Further, the Stack Ensemble model is implemented on pre-trained models to increase their accuracy and to classify the fundus images of KAGGLE and APTOS datasets. The proposed model's primary goal is to detect, and fundus images can be used to predict DR automatically. To classify the severity of DR images, the proposed Stack Ensemble technique is used to optimize the performance of pre-trained models.
The major objectives of this study were as follows:

- Extraction of deep feature representations from retinal images using pre-trained neural networks
- Introducing transfer learning-based Stack Ensemble approach to extract lesion regions and to improve the accuracy in severity grade classification
- Investigating the proposed model performance on datasets namely APTOS and Kaggle with respect to detection of DR severity using minimum number of trainable parameters.

Section 2 provides a comprehensive examination of the cutting-edge DL methods and DR methodologies. The proposed Stack Ensemble technique for DR is presented in Section 3. Sections 4 and 5 give the performance evaluation of the proposed technique on benchmark datasets, along with the conclusion.

II.RELATED WORK

Authors presented a systematic study on usage of various practices, including traditional and DL models, for the detection of DR. In earlier days, researchers mainly concentrated on hand-crafted procedures to demonstrate the severity of the disease. Generally, DR is categorized into two types, non-PDR (NPDR) and PDR [6]. Based on the severity, DR can be divided into five types. It is quiet easy to treat DR in its earlier stage. Microaneurysms (MAs) refer to the small rounded swellings in tiny blood vessels. MA is the only lesion that indicates the initial stage of DR [7][8]. Hence, MA detection is vital for the early diagnosis of DR.

a) Traditional practices

eISSN 1303-5150

Traditional image classification process comprises three stages: pre-processing of images, feature extraction, and classification of images. S. S. Rubini et al. [9] proposed an automated DR detection using hessian-based candidate selection algorithm (AHCS) for feature extraction and SVM for classification. These methods were tested on real-world images and were found to be significantly less effective, with a probability of $p < 0.005$. Mookiah et al. [10] proposed a system that uses hybrid features including exudates/vessel area, texture, and entropy for DR classification. G. Quellec et al. [11] proposed an adapted wavelet algorithm followed by Powell's direction set descent. In this method, the images are divided into three types: color photographs, green filtered photographs, and angiographs. An algorithm for MA detection from image analysis was developed. They attained 83.62% sensitivity for three images with 39 injuries. The detection of red lesions method was proposed by [12] in which the lesions were labeled as candidates. The structures of blood vessels from the candidates were subtracted to reduce the false positives. They attained a sensitivity of 94% and specificity of 87% when tested on 89 images. The disadvantage was that it had a long computation time of 3 min per image. Pre-processing techniques, such as correlation, have been applied [13] for the identification of the bleeding images. In their study, they obtained an average of four false positives per image, indicating a sensitivity of 85%. The diagnosis of early DR [14] was carried out using PCA and Radon transform, and then the features were classified using a hierarchical system of classifiers. The performance of the proposed method was evaluated on diaretDB1 datasets, attaining the sensitivity and specificity of 92.32% and 88.06%, respectively.

P. Adarsh et al. [15] proposed an image processing methodology to extract features, like blood vessels, exudates, and MAs, and textured them to provide input to multi-class support vector machine. Further, model run on 219 DR images helped achieve a sensitivity and specificity of 90.6% and 93.6%, respectively. B. Antal et al. [16] proposed an ensemble method to extract various features from the different fundus images using traditional image processing techniques, such as lesion-specific, image level, and functional structures. The model was evaluated on Messidor dataset, achieving a sensitivity, specificity and accuracy of 90%, 91%, and 90%, respectively. L Seoud et al. [17] proposed a dynamic shape feature extraction algorithm for the detection of red lesion and then used these hand-crafted features for DR screening. D K Prasad et al. [18] proposed to detect the MAs and exudates usingusing morphological techniques on segmentation algorithms. They extracted the features using Haar wavelet and feature selection was done using PCA. Finally, diabetic and non-diabetic samples were classified using a single rule and BPNN classifiers. This model was evaluated on DIARETDB1 dataset, which contained 89 retinal fundus images. The BPNN classifier achieved 93.8%, 93.3%, and 95.23% of accuracy, sensitivity, and specificity, respectively. A. P. Bhatkar et al. [19] proposed a Multi-Layer Perception Neural Network (MLPNN) to classify DR images as normal and abnormal.

b) Approaches of Deep Convolutional Neural Network (Deep CNN)

Many of the researchers have applied the CNN to improve the performance of their models with respect to different computing objectives, including semantic segmentation, classification [20], and objects detection [21]. In medical image processing, authors have applied various DL techniques for different tasks, such as classifying, segmenting, detecting, and registration. M. Bakator et al. [22] and Y. Guo et al. [23] used CNNs to classify the severity levels of DR and achieved a higher accuracy. In general, computer-aided diagnosis is meant for early diagnosis of the disease along with its severity [24]. Further, ophthalmologists make use of this information to validate their decision/opinion. Q. H. Nguyen *et al.* [25] have proposed a model to compute DR severity using varying illumination and field of view. A novel cross-disease network was devised

eISSN 1303-5150

for grading diabetic macular edema (DME) and DR, specifically to examine the relationship between them [26] [27]. In DL pre-processing stage is an important process for the removal of noise and to improve the quality [28] [29]; however, feature extraction method is not necessary. Authors of [30] [31] proposed a CNN architecture with data augmentation for the classification of MAs, exudates, and hemorrhages on the fundus image. They achieved a sensitivity and accuracy of 95% and 75%, respectively, on 5,000 validation images. D Lin et al. [32] and H Pratt et al. [33] proposed a ResNet architecture for classifying fundus images into normal and abnormal types and achieved an accuracy and sensitivity of 85% and 86%, respectively. The image segmentation related to the medical field has played a prominent role and grabbed a lot of attraction from medical scientists and researchers for image identification and analysis [34][35]. In comparison to the conventional segmentation approaches [36], DL techniques have yielded better performance. B. Harangi et al. [37] proposed an ensemble approach for detecting DR and DME. For five-class DR and three-class DME, they achieved accuracies of 90.07% and 96.85%, respectively. K Xu et al. devised a deep CNN [38] methodology with applied data augmentation for multiple heterogeneous sources and achieved 94.5% accuracy. Table 2 presents a systematic survey on DR with different models with its performance.

Table 2: Authors worked on DR with different models and their performance

Ref	Methodology	Dataset	Accuracy	Sensitivity	Specificity
[15]	Feature classification using multiclass SVM	DIARETDB1 DIARETDB0	95.3%	90.6%	93.65%
[16]	The image-level, lesion-specific and anatomical features were extracted and then applied classified with Deep Learning.	MESSIDOR	90%	90%	91%
[24]	A novel DR grading Computer-Aided Diagnosis system based on deep learning	KAGGLE	74%	***	***
[25]	Convolutional Neural Network	KAGGLE	83%	***	***
[26]	Cross Disease Attention Network	IDRiD MESSIDOR	92.6%	90.6%	92.0%
[41]	Combine multilevel features from several Exception architecture based convolutional layers	APTOS	83.09%	88.24%	87.00%
[42]	Convolutional neural network	APTOS	77%	***	***
[44]	Ensemble Classifier	KAGGLE	80.8%	51.5%	86.7%

A deep CNN architecture was previously proposed with standard machine learning (ML) methods [39]. They attained an AUC, sensitivity, and specificity of 0.94, 93%, and 87%, respectively, on the Messidor-2 dataset. Authors of [40] recommended a DL model to classify DR along with other eye disorders. RDR and STDR square measurements revealed a sensitivity and specificity of 90.5% and 91.6%, and 100% and 91.9% respectively. The fully convolutional Networks (FCN) outperform other DL models in terms of semantic medical picture segmentation. S. H. Kassani et al. [41] proposed four in-depth models, including InceptionV3, MobileNet, ResNet50, and Xception, and achieved higher accuracy of 83.09% on APTOS dataset. A DL model for DR diagnosis was presented [42], with an accuracy of 77% on APTOS. G. T. Zago et al. [43] designed a lesion localization model using deep network patch-based approach and achieved a

sensitivity of 94% and CI of 0.95. Some more studies [44] [45] employing ensemble for DR detection achieved an accuracy of more than 80% on the Kaggle dataset. It is clear that DR severity prediction can be done with low error rate using transfer learning and deep ensemble models with pre-trained models.

III. STACK ENSEMBLE METHODOLOGY FOR PREDICTION OF DR

To overcome the barrier of the existing methodologies, we proposed a novel solution using the Stack Ensemble technique for pre-trained models to classify DR images. The proposed model comprises three steps: a) Pre-processing, b) learning of pre-trained models, and c) construction of Stack Ensemble model.

A. Pre-processing

Two datasets were used to evaluate the proposed method's performance: KAGGLE and APTOS, which contained 35126 and 3662 images, respectively. Both these datasets have five different classes, as shown in Figure 1. Firstly, all of the images are resized from their high-resolution sizes to 224*224 as part of the pre-processing.

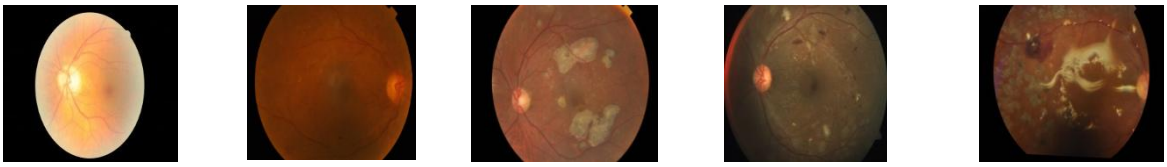


Figure 1. a) Normal, b) Mild, c) Moderate, d) Severe, d) Proliferative DR.

B. Models of pre-trained models

EfficientNetB0: This model heavily relies on baseline networks. They enhanced the model by improving accuracy and efficiency using a new baseline network. EfficientNetB0 architecture is depicted in Figure 2. It divides the original convolution into two stages to reduce the cost of calculation significantly, with a minimal loss of accuracy. These are depth- and point-wise convolutions, respectively. MobileNetV2 was the previous name for MBConv, an inverted bottleneck Conv. It is the most important component of EfficientNetB0. In comparison to old layers, it was paired with an in-depth dissociable convolution, which decreased the computation by roughly k^2. The peak and width of the 2-dimensional convolution window are specified by k, where k represents the kernel size. Larger networks with bigger breadth, depth or resolution tend to attain higher accuracy.

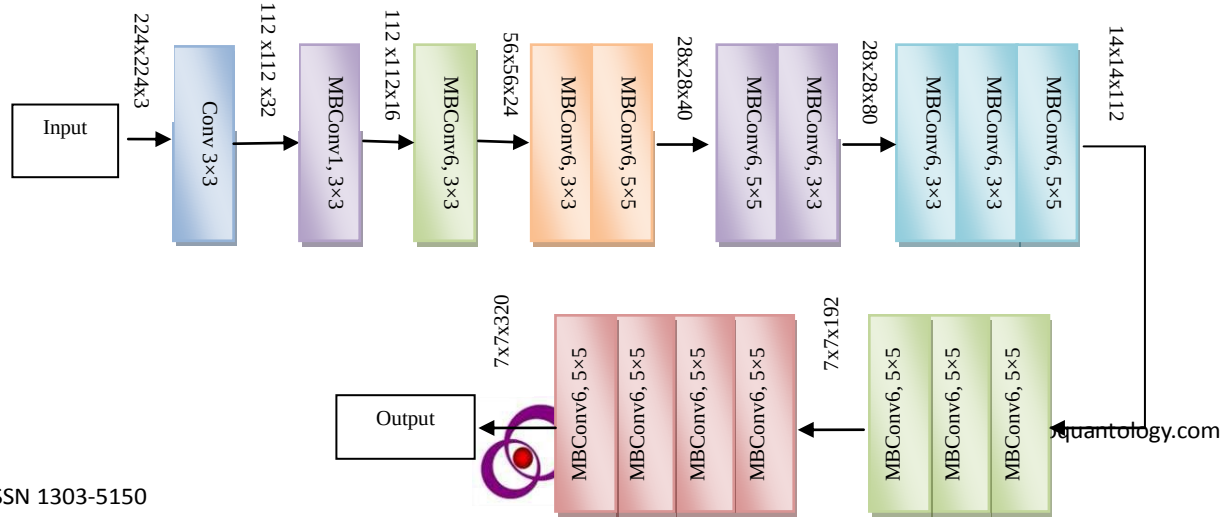


Figure 2: Architecture of the EfficientNetB0

First, let's define a ConvNet, as represented by equation (1):

$$N = I.\ F_i^{L_i}(X_{<H_i, W_{i,} C_i>}) \quad (1)$$

All layers are parameterized uniformly with a constant ratio to reduce the design space. The compounded scaling is done to increase the model's accuracy, as shown in the equation (2):

$$d1, w1, r1 max \qquad Accuracy\big(N(d1, w1, r1)\big) \quad (2)$$

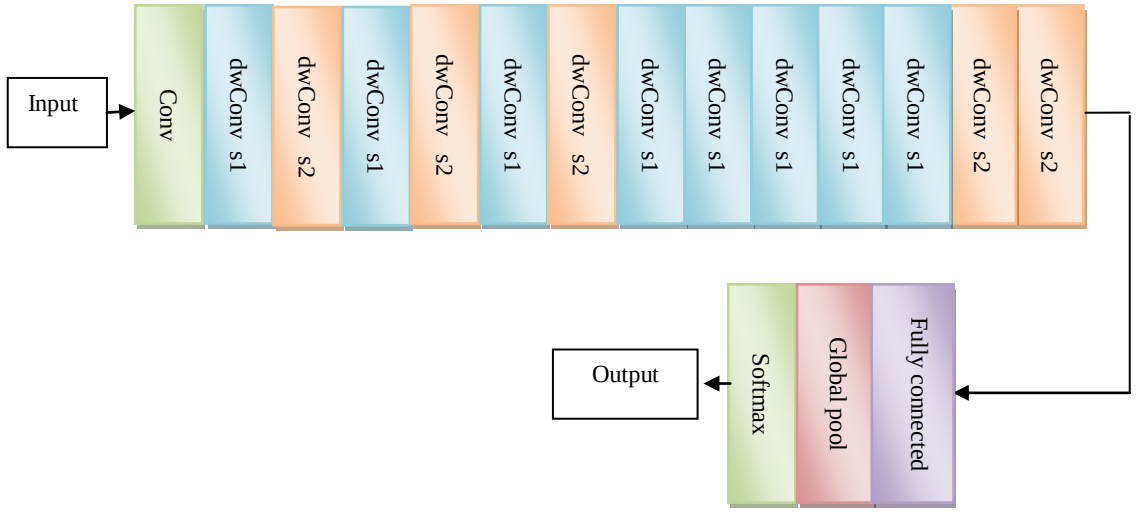
$$N(d1, w1, r1) = \hat{F}_i^{d.\hat{L}}(X_{<r.\hat{H}I, r.\hat{W}I, w.\hat{C}i>})$$

Where the parameters w1, d1, and r1 reflect network scales: width, depth, and resolution. The classification for the pre-defined parameters is as follows:

depth: $d1 = \alpha^{\phi}$, width: $w1 = \beta^{\phi}$, resolution: $r1 = \gamma^{\phi}$ s.t. $\alpha.\beta^2.\gamma^2 = 2$, $\alpha \geq 1$, $\beta \geq 1$, $\gamma \geq 1$ (3)

In equation (3), α, β, and γ are constants found by a grid searching. The user-defined coefficient (ϕ) monitors how the resources are assigned to the scaling of the model. The scaling method performs better and reduces the number of parameters.

MobileNetV3: MobileNetV3 is a CNN model with 30 layers. This model is used for image classification, detection, and segmentation. These models are small, low-latency power models that have many use cases. This model has comparatively less computational time than other models. To build a lightweight DNN, depth-wise separable convolutions are used, which is the core layer for MobileNet. Each input is subjected to a single filter. When the depth-wise and point-wise convolution outputs are combined, a 1×11×1 convolution is applied. Figure 3 shows the model with its architectural structure. The main principle behind MobileNet is that by using depth-wise separable convolution, the speed will be much faster than the regular convolution.



eISSN 1303-5150

Figure 3: Generic architecture of MobileNet

Xception: Xception is a more advanced version of Inception, with depth-wise separable convolution replacing Inception modules. Here, point-wise convolution is not present. Depth-wise convolutions are alternatives to classical convolution and are more efficient in computation time. It comprises 36 layers. Depth-wise separable convolutions are used that have approximately identical parameter counts as the Inception V1 model. For extracting and reducing the features, convolution and pooling are applied. Equation (4) represents the convolution performed on an image. Figure 4 shows the Xception architecture.

$$C(p,q) = (I*w)(p,q) = \sum_k \sum_l I(p-k, q-l)w(k,l) \quad (4)$$

Where w is the convolution kernel size of (k,l).

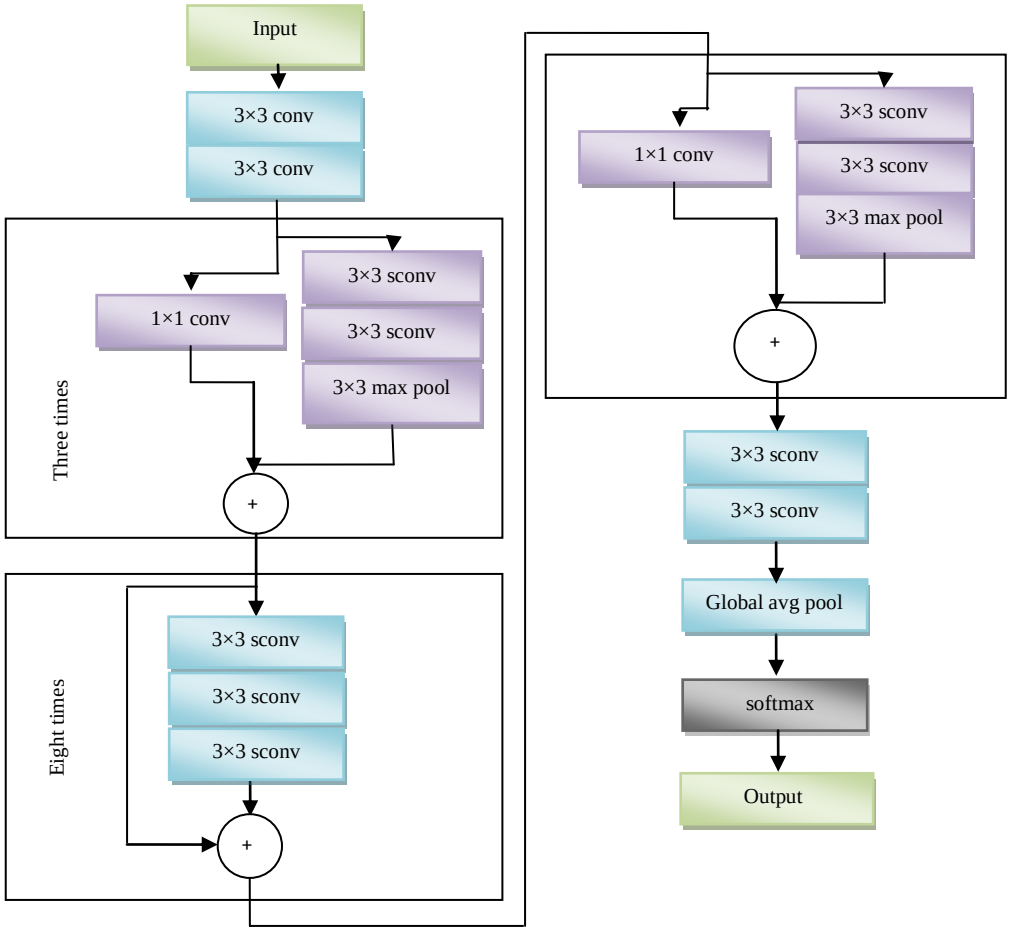


Figure 4: Architecture of the Xception module

C. **Construction of Stack Ensemble model**

Ensemble methods are meta-algorithms that combine all techniques into one predictive model. These techniques are of two types: i) sequential ensemble techniques (e.g., AdaBoost) and ii) ensemble techniques in parallel (e.g., random forest). In our proposed methodology, three different pre-trained models were used, namely EfficientNetB0, MobileNet, and Xception. The outputs of all these pre-trained models are combined further to construct the Stack Ensemble

model. Later, weights are loaded, and then, the layers are frozen and the Global Average Pooling is performed.

The architecture of the Stack Ensemble model is shown in Figure 5. It is an ensemble of various classifiers to serve as a single classifier. Features extracted by various classifiers will be passed to a meta classifier. For final prediction, Random Forest Classifier (RFC) and Extra Trees Classifier (ETC) were used. The algorithm below summarizes stacking.

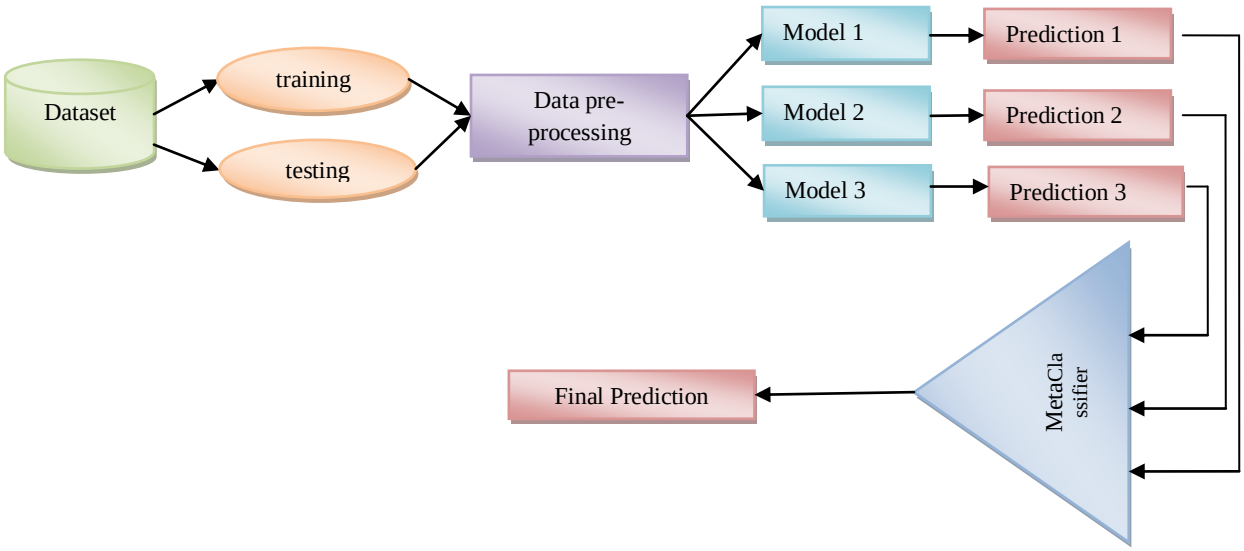


Figure 5: Architecture of the proposed Transfer learning-based Stack Ensemble approach

Algorithm: Stack Ensemble
Input: train data $D_{train}=\{xi, yi\}_{i=1}^{m}$, test data $D_{test}=\{xi, yi\}_{i=1}^{m}$, where x_i is the i^{th} input fundus image from the dataset and the label of the corresponding x_i is y_i.
Output: Ensemble Classifier E

Step 1: Load dataset
Step 2: Resize all pictures from dataset to 224 * 224
Step 3: Load pre-trained models
Model 1➀EfficientNetB0
Model 2➁MobileNet
Model 3➂Xception
Step 4: Learning from the lower-level classifier
for pr=1 to M do
learn h_p based on D_value
end for
Step 5: Construct a new dataset based on the predicted values
for i=1 to N do
$D_{h\_value}=\{\dot{x}_i, y_i\}$, where $\dot{x}_i=\{h_1(x_i),\ldots h_M(x_i)\}$
end for
Step 6: Learn the classifier
Learn E based on D_{h_value}

eISSN 1303-5150

return E

The random forest classifier and extra trees classifier meta-classifiers use features from various pre-trained models. In the process of training, random forest constructs individual decision trees where predictions are taken to make the final prediction. Gini is used to make a final choice after gathering numerous results. If each decision tree has just two children, equation (5) was considered:

$$ni_j = weight_j C_j - weight_{l(j)} C_{l(j)} - weight_{r(j)} C_{r(j)} \quad (5)$$

Where ni_j= indicates the importance of node j, $weight_j$= the weighted number of samples that arrive at node j, C_j = node j impurity value, $left_j$ = from left split node j child node, $right_j$ = from right split node j child node. Decision trees' each feature's importance is calculated as shown in equation (6):

$$fii = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ nodes} ni_k} \quad (6)$$

Where fi_i = denotes the signification of feature i and ni_j = node j importance. Normalize the values between 0 and 1 and divide them with the sum of all features, as shown in equation (7):

$$normfii = \frac{fi_i}{\sum_{j \in all\ features} fi_j} \quad (7)$$

The final feature importance of random forest level is calculated as shown in equation (8):

$$RFfii = \frac{\sum_{j \in all\ trees} normfi_{ij}}{T} \quad (8)$$

Where $RFfi_i$ denotes that the feature i is calculated using all trees in the RF algorithm, normfii = T (the total number of trees), and N = normalized feature importance for i in tree j.

Extremely Randomized Trees Classifier (Extra Trees Classifier) is a learning model that serves as an RF but differs from the construction decision tree. By using the original sample, decision tree is constructed. As the feature set provides k sample features, the decision tree selects the best feature. According to Gini, each feature is ordered in descending order. Five decision trees and k value is built for the above data. The entropy of the data is calculated as shown in equation (9):

$$Entropy(S) = \sum_{i=1}^{c} -p_i log_2(p_i) \quad (9)$$

The number of unique class labels is c, and the fraction of rows having the output label as 'I' is p_i. These predictions are used to train the meta classifier, which made the final prediction.

IV. RESULTS AND DISCUSSIONS

In this, Existing model results and the Proposed model results were compared. A high-end GPU machine is used which runs on 64GB with an Intel® Xenon® Platinum 8276 CPU@2.20GHz processor and a 4GB NVIDIA GRID V100D-4Q graphics card using the Keras framework and TensorFlow-GPU. Along with the DL package, the TensorFLow back-end library was used.

Performance Evaluation Metrics

Accuracy: It is the one metric used for evaluating the classification model. The accuracy test differentiates between the affected and the unaffected cases as shown in equation (10):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad 10$$

eISSN 1303-5150

Where TP = True positive which correctly identifies the positive class, TN = True negative that correctly identifies the negative class, FP = False positive that incorrectly identifies the positive class, and FN = False negative that incorrectly identifies the negative class.

Sensitivity (or) Recall: As shown in equation (11), the positive values are divided and the relevant positive samples are referred to as recall.

$$Sensitivity = \frac{TP}{TP + FN} \quad (11)$$

Precision: It is simply the division of the positive values returned by the classifier and the correct positive values, as shown in equation (12):

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

F1-Score: The test's accuracy is evaluated by precision and recall, as shown in equation (13):

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (13)$$

The two datasets, KAGGLE [46] and APTOS [47], were trained on three different pre-trained models. As already mentioned, the same methodology was implemented on all EfficientNetB0, MobileNet, and Xception models. For all these pre-trained models, the image input size is 224×224, 30 epochs, and a batch size of 8 is used for the classification of images. The RMSprop() optimizer is used to compile the model, with a learning rate of 0.01. Table 3 shows the models performance on which KAGGLE dataset is trained. Here, 7026 images were used for testing, of which 5164, 511, 1046, 173, and 132 images represented Normal, Mild, Moderate, Severe, and Proliferative DR.

Table 3: Pre-trained CNN models performance evaluation on KAGGLE dataset

Model Name	No. of Epochs	Batch Size	Optimizer	Learning Rate	Accuracy	Precision	Recall	F1-Score
EfficientNetB0	30	8	rmsprop( )	0.01	0.70	0.40	0.46	0.41
MobileNet					0.74	0.50	0.37	0.41
Xception					0.77	0.51	0.37	0.41

Similar to the KAGGLE dataset, the APTOS dataset is also trained on EfficientNetB0, MobileNet, and Xception models. The models trained on the APTOS dataset performance is shown in Table 4. Here, 733 images were used for testing, of which 361, 85, 190, 33, and 64 images represented Normal, Mild, Moderate, Severe, and Proliferative DR.

Table 4. Pre-trained CNN models performance evaluations on APTOS dataset

Model Name	No. of Epochs	Batch Size	Optimizer	Learning Rate	Accuracy	Precision	Recall	F1-Score
EfficientNetB0	30	8	rmsprop( )	0.01	0.82	0.73	0.61	0.65
MobileNet					0.78	0.67	0.62	0.61
Xception					0.79	0.69	0.60	0.61

When the Stack Ensemble model is applied to EfficientNetB0, MobileNet, and Xception models that are processed on the KAGGLE dataset, an accuracy of 96% was obtained. Stack Ensemble was applied on various pre-trained models, including EfficientNetB0, MobileNet, Xception, and

the proposed model performance is presented in Table 5. Here, 7026 images were used for testing, of which 5176, 454, 1088, 172, and 136 images represented Normal, Mild, Moderate, Severe, and Proliferative DR.

Table 5: Pre-trained models performance comparison with proposed model on Kaggle dataset

Model Name	Accuracy	Precision	Recall	F1-Score
EfficientNetB0	0.90	0.78	0.82	0.78
MobileNet	0.93	0.91	0.84	0.87
Xception	0.93	0.93	0.81	0.86
Proposed model	0.96	0.93	0.84	0.87

Stack Ensemble is processed on APTOS dataset, an accuracy of 97% was obtained. Stack Ensemble was applied on various pre-trained models, including EfficientNetB0, MobileNet, Xception, and the proposed model performance is shown in Table 6. Here, 733 images were used for testing, of which 341, 79, 208, 44, and 61 images represented Normal, Mild, Moderate, Severe, and Proliferative DR.

Table 6: Pre-trained models performance comparison with proposed model on APTOS dataset

Model Name	Accuracy	Precision	Recall	F1-Score
Efficient_Net_B0	0.94	0.94	0.90	0.92
MobileNet	0.95	0.95	0.92	0.93
Xception	0.94	0.92	0.90	0.90
Proposed model	0.97	0.95	0.92	0.93

The implementation of Stack Ensemble model on pre-trained models led to excellent results with contrast to the existing models. Table 7 shows the proposed modelperformane with other state-of-the-art models trained on KAGGLE and APTOS datasets.

Table 7: Performance Comparison of Existing models with proposed model

Authors	Corpus	# of Samples	Accuracy	Precision	Recall	F1-Score
[24]	Kaggle	7025	0.74	-	-	-
[25]	Kaggle	35126	0.83	-	-	-
[33]	Kaggle	80000	0.75	-	-	-
[34]	Kaggle	35000	0.85	-	-	-
[37]	Kaggle	22700	0.90	-	-	-
[38]	Kaggle	1000	0.94	-	-	-
[43]	Kaggle	15919	0.82	-	-	-
[44]	Kaggle	35126	0.80	0.63	0.65	0.53
[45]	Kaggle	4476	0.88	0.95	0.94	-
[33]	APTOS	3662	0.96	-	-	-
[34]	APTOS	3662	0.77	-	-	-
[41]	APTOS	3662	0.83	-	-	-
Proposed	Kaggle	35126	0.96	0.93	0.84	0.87
Model	APTOS	3662	0.97	0.95	0.92	0.93

eISSN 1303-5150

The proposed model exhibited an accuracy of 96% on KAGGLE and 97% on APTOS datasets. From Table 7, it is clearly evident that the proposed model outperformed other models employed in the previous studies for the prediction of disease severity using DR images.

V. CONCLUSION

DR is prevalent in a majority of diabetic patients; manually detecting it is a time-consuming operation. Many techniques have been used to detect DR more rapidly. Because ophthalmologists face a barrier in preventing DR, overcoming this disadvantage and detecting DR promptly is critical. Here, we proposed the Stack Ensemble model using DL approach. The different images of DR were classified based on disease severity. We used two datasets, i.e., KAGGLE and APTOS, to train and evaluate the model, with accuracies of 96% and 97% for KAGGLE and APTOS, respectively. The findings indicated that compared to the conventionally used model, our model could help make adequate predictions more rapidly and precisely. This model will aid ophthalmologists in diagnosing DR in a shorter time frame. The future studies would need to focus on further tuning and improving the model proposed in this study.

Conflicts of Interest:

The writers finalized that they don't have any known interests of financial personal relationships that may have been used such as to influence this work which was reported in this paper.

REFERENCES:

[1] S. Wan, Y. Liang, and Y. Zhang, "Deep convolutional neural networks for diabetic retinopathy detection by image classification," *Comput. Electr. Eng.*, vol. 72, pp. 274–282, 2018, doi: 10.1016/j.compeleceng.2018.07.042.

[2] K. Ogurtsova*et al.*, "IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040," *Diabetes Res. Clin. Pract.*, vol. 128, pp. 40–50, 2017, doi: 10.1016/j.diabres.2017.03.024.

[3] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, and H. Kang, "Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening," *Inf. Sci. (Ny).*, vol. 501, pp. 511–522, 2019, doi: 10.1016/j.ins.2019.06.011.

[4] R. A. P. Karthigaikumar, "Multi-retinal disease classification by reduced deep learning features," *Neural Comput. Appl.*, 2015, doi: 10.1007/s00521-015-2059-9.

[5] M. P. Bala and S. Vijayachitra, "Early detection and classification of microaneurysms in retinal fundus images using sequential learning methods," *Int. J. Biomed. Eng. Technol.*, vol. 15, no. 2, pp. 128–143, 2014, doi: 10.1504/IJBET.2014.062743.

[6] B. Zhang, X. Wu, J. You, Q. Li, and F. Karray, "Detection of microaneurysms using multi-scale correlation coefficients," *Pattern Recognit.*, vol. 43, no. 6, pp. 2237–2248, 2010, doi: 10.1016/j.patcog.2009.12.017.

[7] E. Treatment and D. Retinopathy, "Early Photocoagulation for Diabetic Retinopathy: ETDRS Report Number 9," *Ophthalmology*, vol. 98, no. 5, pp. 766–785, 1991, doi: 10.1016/S0161-6420(13)38011-7.

[8] B. Antal and A. Hajdu, "Improving microaneurysm detection using an optimally selected subset of candidate extractors and preprocessing methods," *Pattern Recognit.*, vol. 45, no. 1, pp. 264–270, 2012, doi: 10.1016/j.patcog.2011.06.010.

[9] S. S. Rubini and A. Kunthavai, "Diabetic Retinopathy Detection Based on Eigenvalues of the Hessian Matrix," *Procedia - Procedia Comput. Sci.*, vol. 47, pp. 311–318, 2015, doi: 10.1016/j.procs.2015.04.001.

[10] M. R. K. Mookiah*et al.*, "Evolutionary algorithm based classifier parameter tuning for

eISSN 1303-5150

automatic diabetic retinopathy grading: A hybrid feature extraction approach," *Knowledge-Based Syst.*, vol. 39, pp. 9–22, 2013, doi: 10.1016/j.knosys.2012.09.008.

[11] G. Quellec, M. Lamard, P. M. Josselin, G. Cazuguel, B. Cochener, and C. Roux, "Optimal Wavelet Transform for the Detection of Microaneurysms in Retina Photographs," *IEEE Trans. Med. Imaging*, vol. 27, no. 9, pp. 1230–1241, 2008, doi: 10.1109/TMI.2008.920619.

[12] M. Esmaeili, H. Rabbani, A. M. Dehnavi, and A. Dehghani, "A new curvelet transform based method for extraction of red lesions in digital color retinal images," *Proc. - Int. Conf. Image Process. ICIP*, pp. 4093–4096, 2010, doi: 10.1109/ICIP.2010.5652820.

[13] J. P. Bae, K. G. Kim, H. C. Kang, C. B. Jeong, K. H. Park, and J. M. Hwang, "A study on hemorrhage detection using hybrid method in fundus images," *J. Digit. Imaging*, vol. 24, no. 3, pp. 394–404, 2011, doi: 10.1007/s10278-010-9274-9.

[14] R. Rosas-Romero, J. Martínez-Carballido, J. Hernández-Capistrán, and L. J. Uribe-Valencia, "A method to assist in the diagnosis of early diabetic retinopathy: Image processing applied to detection of microaneurysms in fundus images," *Comput. Med. Imaging Graph.*, vol. 44, pp. 41–53, 2015, doi: 10.1016/j.compmedimag.2015.07.001.

[15] P. Adarsh and D. Jeyakumari, "Multiclass SVM-based automated diagnosis of diabetic retinopathy," *Int. Conf. Commun. Signal Process. ICCSP 2013 - Proc.*, pp. 206–210, 2013, doi: 10.1109/iccsp.2013.6577044.

[16] B. Antal and A. Hajdu, "An ensemble-based system for automatic screening of diabetic retinopathy," *Knowledge-Based Syst.*, vol. 60, pp. 20–27, 2014, doi: 10.1016/j.knosys.2013.12.023.

[17] L. Seoud, T. Hurtut, J. Chelbi, F. Cheriet, and J. M. P. Langlois, "Red Lesion Detection Using Dynamic Shape Features for Diabetic Retinopathy Screening," *IEEE Trans. Med. Imaging*, vol. 35, no. 4, pp. 1116–1126, 2016, doi: 10.1109/TMI.2015.2509785.

[18] D. K. Prasad, L. Vibha, and K. R. Venugopal, "Early detection of diabetic retinopathy from digital retinal fundus images," no. December, pp. 240–245, 2016, doi: 10.1109/raics.2015.7488421.

[19] A. P. Bhatkar and G. U. Kharat, "Detection of Diabetic Retinopathy in Retinal Images Using MLP Classifier," *Proc. - 2015 IEEE Int. Symp. Nanoelectron. Inf. Syst. iNIS 2015*, pp. 331–335, 2016, doi: 10.1109/iNIS.2015.30.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.

[21] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," *arXiv*, 2020.

[22] M. Bakator and D. Radosav, "Deep learning and medical diagnosis: A review of literature," *Multimodal Technol. Interact.*, vol. 2, no. 3, 2018, doi: 10.3390/mti2030047.

[23] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, 2016, doi: 10.1016/j.neucom.2015.09.116.

[24] T. Araújo *et al.*, "DR|GRADUATE: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images," *Med. Image Anal.*, vol. 63, 2020, doi: 10.1016/j.media.2020.101715.

[25] Q. H. Nguyen *et al.*, "Diabetic retinopathy detection using deep learning," *ACM Int. Conf. Proceeding Ser.*, pp. 103–107, 2020, doi: 10.1145/3380688.3380709.

[26] X. Li, X. Hu, L. Yu, L. Zhu, C. W. Fu, and P. A. Heng, "CANet: Cross-Disease Attention Network for Joint Diabetic Retinopathy and Diabetic Macular Edema Grading," *IEEE*

Trans. Med. Imaging, vol. 39, no. 5, pp. 1483–1493, 2020, doi: 10.1109/TMI.2019.2951844.

[27] J. Ahmad, H. Farman, and Z. Jan, *Deep Learning Methods and Applications BT - Deep Learning: Convergence to Big Data Analytics*. Springer Singapore, 2019.

[28] S. Dutta, B. C. S. Manideep, S. M. Basha, R. D. Caytiles, and N. C. S. N. Iyengar, "Classification of diabetic retinopathy images by using deep learning models," *Int. J. Grid Distrib. Comput.*, vol. 11, no. 1, pp. 89–106, 2018, doi: 10.14257/ijgdc.2018.11.1.09.

[29] V. Gulshan *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA - J. Am. Med. Assoc.*, vol. 316, no. 22, pp. 2402–2410, 2016, doi: 10.1001/jama.2016.17216.

[30] G. Quellec, K. Charrière, Y. Boudi, B. Cochener, and M. Lamard, "Deep image mining for diabetic retinopathy screening," *Med. Image Anal.*, vol. 39, pp. 178–193, 2017, doi: 10.1016/j.media.2017.04.012.

[31] L. Deng, "A tutorial survey of architectures, algorithms, and applications for deep learning," *APSIPA Trans. Signal Inf. Process.*, vol. 3, no. 2014, 2014, doi: 10.1017/ATSIP.2013.99.

[32] D. lin, A. V. Vasilakos, Y. Tang, and Y. Yao, "Neural networks for computer-aided diagnosis in medicine: A review," *Neurocomputing*, vol. 216, pp. 700–708, 2016, doi: 10.1016/j.neucom.2016.08.039.

[33] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional Neural Networks for Diabetic Retinopathy," *Procedia Comput. Sci.*, vol. 90, no. July, pp. 200–205, 2016, doi: 10.1016/j.procs.2016.07.014.

[34] M. T. Esfahani, M. Ghaderi, and R. Kafiyeh, "Classification of diabetic and normal fundus images using new deep learning method," *Leonardo Electron. J. Pract. Technol.*, vol. 17, no. 32, pp. 233–248, 2018.

[35] A. Srinivasan and S. Sundaram, "Applications of deformable models for in-dopth analysis and feature extraction from medical images-A review," *Pattern Recognit. Image Anal.*, vol. 23, no. 2, pp. 296–318, 2013, doi: 10.1134/S1054661813020132.

[36] G. Litjens*et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, no. December 2012, pp. 60–88, 2017, doi: 10.1016/j.media.2017.07.005.

[37] B. Harangi, J. Toth, A. Baran, and A. Hajdu, "Automatic screening of fundus images using a combination of convolutional neural network and hand-crafted features," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 2699–2702, 2019, doi: 10.1109/EMBC.2019.8857073.

[38] K. Xu, D. Feng, and H. Mi, "Deep convolutional neural network-based early automated detection of diabetic retinopathy using fundus image," *Molecules*, vol. 22, no. 12, 2017, doi: 10.3390/molecules22122054.

[39] R. Gargeya and T. Leng, "Automated Identi fi cation of Diabetic Retinopathy Using Deep Learning," *Ophthalmology*, pp. 1–8, 2017, doi: 10.1016/j.ophtha.2017.02.008.

[40] D. S. W. Ting *et al.*, "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," *JAMA - J. Am. Med. Assoc.*, vol. 318, no. 22, pp. 2211–2223, 2017, doi: 10.1001/jama.2017.18152.

[41] S. H. Kassani, P. H. Kassani, R. Khazaeinezhad, M. J. Wesolowski, K. A. Schneider, and R. Deters, "Diabetic Retinopathy Classification Using a Modified Xception Architecture," *2019 IEEE 19th Int. Symp. Signal Process. Inf. Technol. ISSPIT 2019*, pp. 0–5, 2019, doi: 10.1109/ISSPIT47144.2019.9001846.

[42] O. Dekhil, A. Naglah, M. Shaban, M. Ghazal, F. Taher, and A. Elbaz, "Deep Learning

Based Method for Computer Aided Diagnosis of Diabetic Retinopathy," *IST 2019 - IEEE Int. Conf. Imaging Syst. Tech. Proc.*, pp. 19–22, 2019, doi: 10.1109/IST48021.2019.9010333.

[43] G. T. Zago, R. V. Andreão, B. Dorizzi, and E. O. Teatini Salles, "Diabetic retinopathy detection using red lesion localization and convolutional neural networks," *Comput. Biol. Med.*, vol. 116, p. 103537, 2020, doi: 10.1016/j.compbiomed.2019.103537.

[44] S. Qummar*et al.*, "A Deep Learning Ensemble Approach for Diabetic Retinopathy Detection," *IEEE Access*, vol. 7, pp. 150530–150539, 2019, doi: 10.1109/ACCESS.2019.2947484.

[45] Z. Gao, J. Li, J. Guo, Y. Chen, Z. Yi, and J. Zhong, "Diagnosis of Diabetic Retinopathy Using Deep Neural Networks," *IEEE Access*, vol. 7, pp. 3360–3370, 2019, doi: 10.1109/ACCESS.2018.2888639.

[46] Kaggle dataset [Online]. Available, https://kaggle.com/c/diabetic-retinopathy-detection.

[47] APTOS Kaggle dataset [Online]. Available, https://www.kaggle.com/c/aptos2019-blindness-detection.

eISSN 1303-5150