Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.0320/ACCESS.2020.DOI

Diabetic Retinopathy Diagnosis from Fundus Images Using Stacked Generalization of Deep Models

HARSHIT KAUSHIK[1], DILBAG SINGH[2], MANJIT KAUR[2], HAMMAM ALSHAZLY[3], ATEF ZAGUIA[4], AND HABIB HAMAM[5]

[1]School of Computing and Information Technology, Manipal University, Jaipur, Rajasthan, India
[2]School of Engineering and Applied Sciences, Bennett University, Greater Noida, India
[3]Department of Computer Science, Faculty of Computers and Information, South Valley University, Qena 83523, Egypt
[4]Department of Computer Science, College of Computers and Information Technology, Taif University, Taif, 21944, Saudi Arabia
[5]Faculty of Engineering, Moncton University, NB, E1A3E9, Canada

Corresponding author: Manjit Kaur (e-mail: manjitbhinder8@gmail.com)

This work was supported by Taif university Researchers Supporting Project Number (TURSP-2020/114), Taif University, Taif, Saudi Arabia.

ABSTRACT Diabetic retinopathy (DR) is a diabetes complication that affects the eye and can cause damage from mild vision problems to complete blindness. It has been observed that the eye fundus images show various kinds of color aberrations and irrelevant illuminations, which degrade the diagnostic analysis and may hinder the results. In this research, we present a methodology to eliminate these unnecessary reflectance properties of the images using a novel image processing schema and a stacked deep learning technique for the diagnosis. For the luminosity normalization of the image, the gray world color constancy algorithm is implemented which does image desaturation and improves the overall image quality. The effectiveness of the proposed image enhancement technique is evaluated based on the peak signal to noise ratio (PSNR) and mean squared error (MSE) of the normalized image. To develop a deep learning based computer-aided diagnostic system, we present a novel methodology of stacked generalization of convolution neural networks (CNN). Three custom CNN model weights are fed on the top of a single meta-learner classifier, which combines the most optimum weights of the three sub-neural networks to obtain superior metrics of evaluation and robust prediction results. The proposed stacked model reports an overall test accuracy of 97.92% (binary classification) and 87.45% (multi-class classification). Extensive experimental results in terms of accuracy, F-measure, sensitivity, specificity, recall and precision reveal that the proposed methodology of illumination normalization greatly facilitated the deep learning model and yields better results than various state-of-art techniques.

INDEX TERMS Convolutional neural networks, diabetic retinopathy, early diagnosis, fundus images, gray world algorithm, ensemble learning

I. INTRODUCTION

Diabetic retinopathy (DR) is a medical condition that is caused by the damage to the blood vessels of the light-sensitive tissue at the back of the eye (retina), which can eventually cause complete blindness and various other eye problems depending on the severity of the disease. Though the treatment is available, it is estimated that numerous people go blind every day because of this disease [1]. It is observed that $40\% - 45\%$ of diabetic patients are likely to have DR in their life, but due to lack of knowledge and delayed diagnosis, the condition escalates quickly [2].

The Early Treatment DR Study Research Group (ETDRS) has shown that if DR is correctly diagnosed on time, it may reduce the chances of vision loss by 50% [3]. The prevalence of DR is maximum i.e., 25.04% in the people who fall in the age bracket of 61-80 [4]. Till now retinal images are manually assessed by ophthalmologists and clinicians for predicting DR after the eye fundoscopic exam and to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3101142, IEEE Access

analyze signs such as cotton wool spots, retinal swellings, and hemorrhages [2]. However, it is usually observed that during image the acquisition process, the fundus images show various kinds of irrelevant illuminations, non-uniform light distribution, blurred or darkened candidate regions, which subsequently affect the diagnostic process and result in biased predictions[5]. To detect DR, it is essential to obtain results with high precision irrespective of any bias to avoid a wrong judgment that may lead to a serious problem or in some cases, even permanent blindness. During the fundoscopic test, if the obtained image is highly saturated, it becomes difficult to carry out a proper visual assessment test even by a trained ophthalmologist or a clinician and hence, the presence of non-uniform illuminations can impede correct predictions [6]. Therefore, luminosity normalization becomes a significant pre-processing aspect for a diverse set of retinal images. Some of the previous work have considered normalizing the luminosity of the retinal images using various statistical, mathematical, and particularly HSV color space based models to desaturate the image [5]–[7].

The previous diagnostic studies of DR can be classified into two types: 1) Automatic detection of the disease (binary), and b) Classification of different stages of the disease (multi-class). In our study, our focus is to automate the diagnostic process and to combine the luminosity normalization pre-processing pipeline with an advanced artificial intelligence technique. Till now various image processing techniques have been presented to detect DR by considering the definitive candidate regions such as cotton wool spots, exudate, hemorrhages, and blood vessels, as reported in [8], [9]. These methods rely on manual feature extraction but, since most of the retinal images depict non-uniform features, thus generalizing feature set for all images may give inappropriate diagnostic results when a large database is considered.

Various architectures of the multi-layered perceptron, convolutional neural networks (CNN), and machine learning algorithms have also been implemented for automatic disease detection [10]–[12]. However, none of these studies has addressed the problem of non-uniform reflectance and over-saturation of a fundus image surface for developing an unbiased DR diagnostic tool. Therefore, to alleviate this issue we have presented a novel color constancy technique to reduce irrelevant reflectance in fundus images and for the feature extraction of the pre-processed images, a stacked generalization of deep CNNs is developed, which can also be considered as a superior cross-validation technique for neural networks models.

The main contributions of this paper can be summarized as follows:

- We solve the non-ideal illumination and color degradation problems by using the gray world color constancy schema to desaturate the retinal images. This will enable ophthalmologists to use color of the images as a reliable cue for recognizing the DR signs and avoid the various distortions related to light distribution and color, which may hinder the diagnostic results.
- Scaling factor is an important step in color correction technique such as gray world in our case, therefore, the color channel with minimum mean is considered as a reference to calculate the gray world illuminant.
- To automate the diagnostic process and to make predictions using the desaturated images, a stacked generalization of three custom CNNs is developed, which is fed into a single meta-learner to extract the most optimum weights from the sub-networks to achieve better performance. This method differs from a usual voting classifier because the evaluation metrics (e.g. accuracy and mean squared error) are not averaged or voted, but rather the meta-learner model gets multiple prediction probabilities as input, which are combined to generate better features and thus achieve accurate results.
- We consider the Exponential Linear Unit (ELU) activation function for each sub-model due to its fast convergence and more accurate results.
- To monitor the generalization of error and avoid conditions such as overfitting and bias-variance trade-off during training, techniques such as exponential learning rate decay and early-stopping are also applied to give an overall regularization effect on the proposed model.
- Extensive experiments and comparisons between the proposed model with the existing works in the diagnosis of DR have been drawn to validate our model and findings.

The rest of the paper is organized as follows. Section II discusses the literature review of DR diagnosis. In Section III we state our motivations. Image normalization and the stacked generalization of deep CNNs model are discussed in Section IV. Section V presents the experimental setup. Section VI presents the quantitative analysis. The discussion is presented in Section VII. Finally, our conclusions and possible future work are presented in Section VIII.

II. LITERATURE REVIEW

Different techniques have been presented by researchers to deal with retinal image normalization, balancing luminosity distribution, contrast normalization, and computer-aided diagnostic systems, which have proved to be of great importance in the field of retinal imaging. The literature survey of this study covers two major categories of DR works to ensure that an overall view is given for better understanding. The works of each category were evaluated based on different performance metrics and design attributes based on the data pattern and proposed experimental design.

The first category comprises of works, which solely focused on an image processing based methodology for DR detection. Zhou et al. [5] presented a luminosity adjustment technique in which a luminance matrix is obtained by the gamma correction of value channel in HSV color space to improve the quality of individual RGB channels. For improving the contrast of images, contrast limited adaptive histogram equalization (CLAHE) technique was used that involves a kernel based iterative process to normalize the histogram of image pixels to avoid congestion of the pixels in a particular range, thus improving the image quality. In [7], the authors proposed the histogram equalization-based image processing technique for fundus image enhancement and developed a CNN model for classification. They used a small dataset of 400 images and achieved a sensitivity of 96.67% and specificity of 93.33%. Bhaskar et al. [13] proposed a technique to normalize the contrast and luminosity of the fundus images by assuming that all the neighborhood pixels are independent and identical to each other. In [14], a retinal enhancement technique based on Speeded up Adaptive Contrast Enhancement (SUACE) algorithm integrated with the Tyler-coy algorithm was proposed. The SAUCE algorithm uses a gray-scale image obtained by Principal Component Analysis (PCS), which was then fed into the Tyler-coy algorithm to remove the discontinuities of blood-vessel for better prediction results. In 2015, [15] presented an illumination correction technique using a low-pass filter and Gaussian filter. By using the low pass filter, the background of the image is normalized and then superimposed with the results of the Gaussian filter, thus removing any sort of foreground noise that existed earlier. Singh et al. [16] used the usual histogram equalization technique for low-radiance images to clip away the pixel-values based on the threshold, which was calculated by taking the average median value of the image to enhance the normalization results. They used structural similarity index measure and the Euclidean distance to validate their prediction results. Although numerous techniques proposed methodologies for image contrast enhancement, but none of them focused on image desaturation for developing a DR system.

The second category depicts the various deep/ machine learning methodologies, which have been presented for early DR detection. Most of the previous work was focused on the development of traditional machine learning and ensemble deep learning techniques. Recently, Zhou et al. [10] proposed a multiple instance learning technique, which is a weakly supervised technique to detect DR in fundus images. Initial image processing steps such as resizing, Gaussian smoothing were implemented before feature extraction. Their detection model was divided into two parts. First, they created a bag of image patches for detecting lesions. Second, a pre-trained Alexnet [17] model was utilized for automatic feature extraction. The model achieved an AUC score of 92.5%. In [18], an ensemble approach using deep transfer learning models to detect DR was proposed. The models including ResNet [19], Densenet [20], Inception [21], and Xception [22] for extracting features and performed extensive hyperparameter tuning to achieve better results. They gave per-class metrics where the highest AUC of the imbalanced class was 97%, but they did not consider any image pre-processing technique to normalize the images. The authors used a dataset of images that contains spatial noise and distortions such as blurring and darkened corners, which require a more advanced technique to get reliable results.

A stacking technique of machine learning algorithms was presented in [1] to prepare a DR screening tool. Lesions and microaneurysms are extracted and then classified using an ensemble classifier. The model's performance was evaluated using accuracy, sensitivity, and specificity, and achieved 90%, 91%, 90%, respectively. In 2017, another improved ensemble technique was presented by Somasundaram and Alli [23]. Machine learning bagging ensemble classifier (ML-BEC) was considered for the prediction of DR. They implemented the t-distributed Stochastic neighbor embedding (t-SNE) algorithm to separate the images into similar and dissimilar pairs. Saleh et al. [24] presented an ensemble technique for DR risk assessment, which justifies the presence or absence of the disease. They prepared a dominance-based rough set balanced rule ensemble (DRSA-BRE) and compared their works with the random forest classifier. The best sensitivity score achieved was near 80%. Similarly, various DR detection methods have been presented in this field [25]–[27]. However, none of these solve the problem of non-uniform illuminations, which can play a major role in detection of proliferative and Non-proliferative DR.

Table 1 summarizes the most relevant work from two major categories for DR detection along with the used performance evaluation metrics and their limitations. From the presented research literature in Table 1, we can infer that most of the techniques focused on retinal contrast enhancement and machine/deep learning models for classification without addressing the non-uniform reflectance of fundus data during image acquisition. Therefore, to alleviate these issues we developed a pipeline for image illumination normalization and a novel feature extraction model for early DR detection.

III. MOTIVATION

Since most of the proposed approaches focused mostly on machine learning, deep learning, and image processing techniques to extract candidate features such as lesions, hemorrhages, exudates and cotton-wool spots but they ignored solving the variance in scene illumination and light degradation, which affects the performance and may result in biased prediction results. In our proposed method, we have used a dataset that has multi-sourced images. Therefore, various types of noise and distortions are encountered in the images. To overcome such issues, we aim to explore the research area of combining artificial intelligence and image processing to develop a complete illumination proof diagnostic tool for DR. The methodology that has been applied is discussed in the following sections.

IV. METHODOLOGY

Figure 1 demonstrates the different stages of the proposed methodology in the form of a model pipeline. After the data acquisition, the image luminosity is normalized by the color constancy based gray world algorithm. The image processing pipeline is shown in the figure in which the illuminant K' from the images is used to normalize the image. The data is split into training and test sets for the stacking convolutional

TABLE 1: A summary of the related works presented in this study.

References	**Methodology**	**Performance evaluation metrics**	**Limitations**
Mei et al. [5]	Image processing methodology for image enhancement	Average standard deviation	Image enhancement is discussed, however, its advantage in DR detection is missing
ÃŰmer et al. 2018 [7]		Sensitivity and Specificity	The experimental dataset consisted of only 400 retinal images
A. M. et al. 2017 [14]		True positive rate, false positive rate, Accuracy	The obtained experimental accuracy is less than the previous state-of-art methods
Mustafa et al. 2015 [15]		Signal noise ratio (SNR)	Utilized Gaussian filter for noise correction, but did not discuss any disease detection module
Navdeep et al.2019 [16]		Structural similarity index (SSI), Euclidean distance	Used histogram equalization for image enhancement, but preprocessing steps for luminosity normalization were not taken
BÃąlint et al. 2014 [1]	Machine Learning and Deep Learning for feature extraction	Accuracy, Sensitivity, Specificity	Used ensemble machine learning algorithms, however, obtained less accuracy on the dataset
Lei et al. 2017 [10]		Area under curve (AUC)	Transfer learning used for feature extraction but did not present extensive comparative analysis in the study
Sehrish et al. 2019 [17]		AUC	Ensemble of transfer learning models was proposed; however, image luminosity normalization was not done. AUC value is also less
S.K et al. 2017 [18]		Accuracy, Classification time, DR detection rate	A bagging classifier was used but no image preprocessing steps for irrelevant illuminations were taken.
Emran et al. 2018 [19]		Sensitivity	Proposed a machine learning ensemble model for feature extraction, but model sensitivity was only 80%.
Proposed Methodology	Stacked CNNS	F-measure, precision, recall, accuracy, AUC, sensitivity, specificity, mean squared error (MSE), Peak signal to noise ratio (PSNR), Confusion Matrix analysis.	A diagnostic system is proposed based on a novel gray world assumption for luminosity normalization and a stacked integrated deep learning algorithm for feature extraction. Data augmentation is also applied. One limitation we observed is that our model neglected the fainting lesions even after the image normalization was done and thus, it might be difficult to predict the mild stage of DR. However, some advanced candidate region-based segmentation techniques could be used for candidate feature extraction to achieve better results.

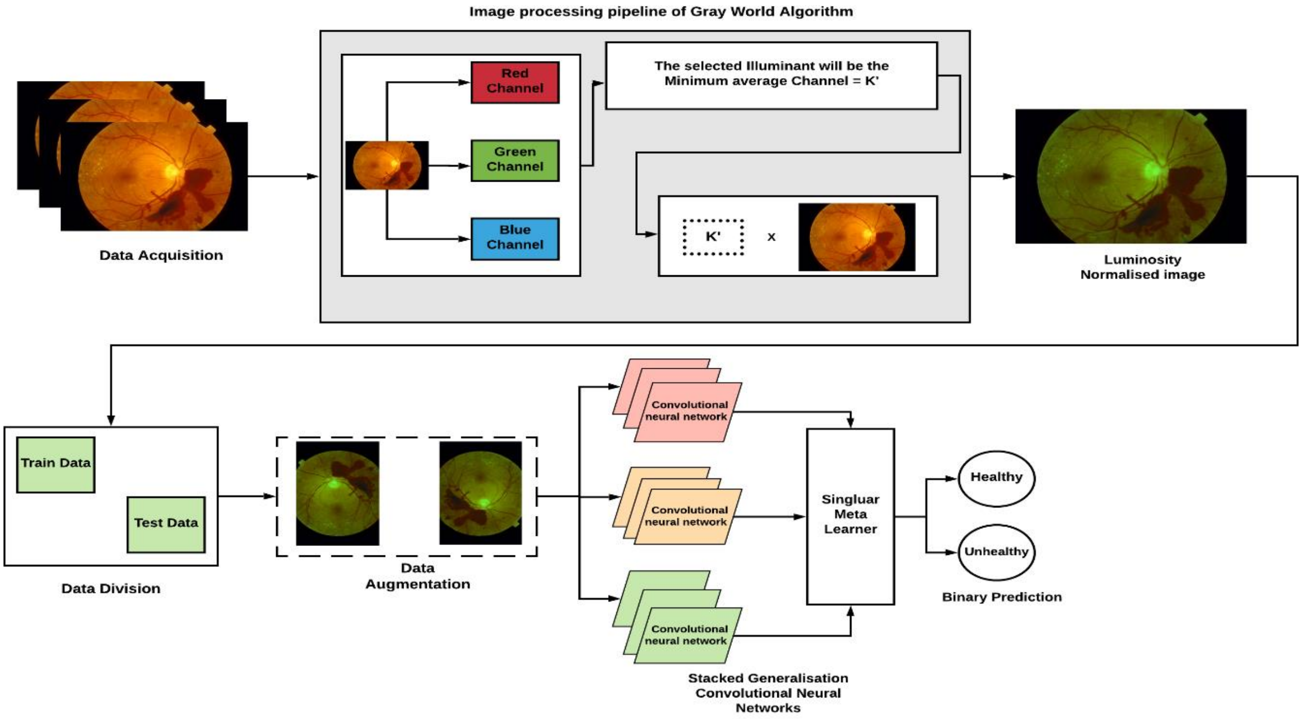


Image processing pipeline of Gray World Algorithm
Data Acquisition
Red Channel
Green Channel
Blue Channel
The selected Illuminant will be the Minimum average Channel = K'
K'
x
Luminosity Normalised image
Train Data
Test Data
Data Division
Data Augmentation
Convolutional neural network
Convolutional neural network
Convolutional neural network
Singluar Meta Learner
Healthy
Unhealthy
Binary Prediction
Stacked Generalisation Convolutional Neural Networks

FIGURE 1: A diagrammatic flow of the proposed methodology and the training process.

model. Three different sub-models of CNNs are fed into a single meta-learner classifier for feature extraction. The fusion strategy of the stacking model is based on the weighted majority from each of the sub-model for generating better features for classification. Data augmentation technique is also applied to improve the diversity of images in the dataset. Finally, the meta-learner classifier produces the diagnostic result as healthy (No DR) or unhealthy (DR).

A. LUMINOSITY NORMALIZATION USING GRAY WORLD ALGORITHM

We use a color constancy algorithm for image normalization as a pre-processing step. In our experiments, we dealt with a dataset that contains images from multiple sources having color variations, varying illuminations, and a non-uniform light distribution, which resulted in a large amount of heterogeneity among images. In case of retinal images, heterogeneity among images can cause a major difference in appearance, for example, some part of the image gets highlighted near the center, but boundaries get blurred, and these non-uniformities can seriously affect the diagnostic results. Therefore, it is necessary to propose a color calibration methodology for these images [28]. We have implemented the color constancy algorithm to remove the unnecessary surface reflectance and to make the color of the image invariant to such illuminations and other color-related aberrations. Generally, there are four common color constancy algorithms: Gray World, Shades of Gray, General Gray World, and Max-RGB. In this paper, we have consider the Gray World algorithm.

Gray world algorithm assumes that the average surface reflectance of the image is achromatic and therefore variations could be done by including the average pixel values and scaling them by a scaling factor, which is computationally inexpensive to calculate. Gray world is also a statistical algorithm, which uses less computation power. According to [29] most of the existing algorithms are based on assumptions. For instance, the Max-RGB color constancy algorithm assumes the presence of white patch in the image to calculate the illuminant, whereas the gray world algorithm uses the average reflectance, and encourages a data-driven approach in color constancy [30].

As explained above, it is assumed in the gray world algorithm that the (R,G,B) color channels have linear values, which means that the average reflection in standard light is gray. But it is not what we see in real life. It is based on the hypothesis that the average of each channel (R,G,B) in an image I is always equal, i.e., gray [31]. However, the average is not constant and is either greater or less than the gray value. This deviation from the original gray value gives us the illumination change. The illuminant of the image is then estimated in the RGB mode, which is then used to normalize each channel of the image to transform the image under a canonical light resource.

The stepwise algorithm of this popular luminosity normalization schema is explained below.

STEP 1: Pixel Level Normalization

Initially, to get the color of the light source, pixel-level normalization is carried out by calculating the average pixel value of each sensor channel. Consider an image as:

$$I = [R_{wh}, G_{wh}, B_{wh}] \tag{1}$$

R_{wh}, G_{wh}, and B_{wh} represent the sensor channels, whereas w and h depict the image width and height, respectively. The mean pixel value can be calculated as:

$$\mu_j = \frac{1}{j} \sum_j I_j \tag{2}$$

Here, $j = R, G, B$.

STEP 2: Gray World Illuminant Calculation

In the gray world color correction, one of the color channels is selected as a reference to calculate the illuminant but the intensity of the resultant normalized image degrades and may hinder the diagnostic results. Therefore, in the proposed method, a compressed color channel technique of the Gray world algorithm is used in which the color channel with minimum average magnitude is selected as proposed in [32]. The scaling factors obtained based on this magnitude can be expressed as:

$$\beta_r = \frac{\bar{X}_{min}}{\bar{R}_{wh}} \tag{3}$$

$$\beta_g = \frac{\bar{X}_{min}}{\bar{G}_{wh}} \tag{4}$$

$$\beta_b = \frac{\bar{X}_{min}}{\bar{B}_{wh}} \tag{5}$$

Thus, the resultant illuminant belongs to each sensor channel of the image. For example, if the image is I, then the component of illuminant is e_c, where $c \in [R, G, B]$.

STEP 3: Scaling Individual Image Channel

The normalized image is obtained by scaling each individual color channel by multiplying it with the scaling factor as:

$$R' = R_{wh} \times \beta_r \tag{6}$$

$$G' = G_{wh} \times \beta_g \tag{7}$$

$$B' = B_{wh} \times \beta_b \tag{8}$$

R', G', and B' represent the normalized channels of the resultant color normalised image.

Figure 2 shows the normalised images obtained after applying the gray world color constancy algorithm.

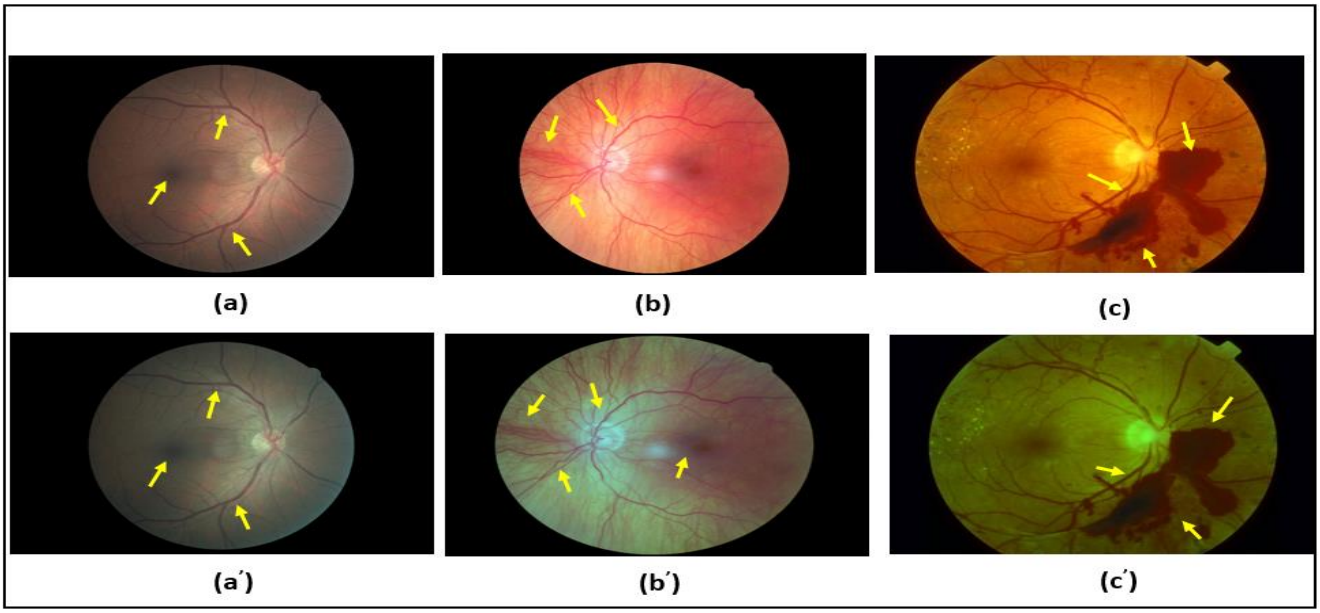


FIGURE 2: Results of image normalization using colour constancy algorithm. The first row shows three original images, whereas the second row shows their corresponding colour normalized images after applying the gray world algorithm. The yellow arrow points out that features visible in original image such as blood vessels, macula, haemorrhages are not affected after luminosity normalization.

B. ARTIFICIAL DATA GENERATION

To add more diversity to the dataset, data augmentation technique is used in which artificial data is generated from the pre-existing images. The data is generated for each mini batch in an iterative process during model training. The applied augmentation steps include horizontal flip, width shift, height shift, fill mode, and zoom range. Table 2 illustrates the data augmentation parameters. Figure 3 shows nine generated images in an iterative process during model training with a rotation angle set to 74 degrees.

TABLE 2: Data augmentation parameters considered for retinal image generation.

Parameters	Values
Width shift	0.1
Height shift	0.1
Rotation range	15
Zoom range	0.1
Horizontal flip	True
Fill mode	Nearest

C. CONVOLUTIONAL NEURAL NETWORK

The CNN models are based on the principle of layer-wise abstraction for feature learning. The complexity as well as the number of features increase with the model depth. CNNs follow an analogous feed-forward architecture just like an artificial neural network, but they are much better in the generalization for computer vision-related problems. They are commonly known as ConvNets and usually consist of an input layer, hidden layers, and an output layers. The hidden layers have some activation functions, fully connected layers, and pooling layers. The top layers of a CNN model tend to learn low-level features such as edges, color, and shapes, whereas the deeper layer focuses on learning high-level features.

Typically, CNN models are a stack of alternating convolutions with various sizes of filters, pooling, and fully connected layers. The difference between a fully connected layer and a convolution layer is that the convolution layer is partially connected and receives inputs from a sub-area of the previous layer, whereas in a fully connected layer all the previous neurons are related to the next neurons for feature transmission [33]. A kernel or commonly known as the filter is a sliding window over the image, which is an array of numbers, where these numbers are the weights that are updated continuously. The area over which it slides is called the receptive field. In our model, we apply 3×3 filters with a depth of 3 since we have colored images of size $96 \times 96 \times 3$. The filter convolution over an image results in an element-wise multiplication with pixel values represented as:

$$L[m,n] = (f \times h)[m,n] = \sum_{j} \sum_{k} h[j,k] \times f[m-j,n-k] \tag{9}$$

Here, the input image is f, the kernel is denoted by h, and the indices of rows and columns are represented by m and n, respectively. After completion of the first convolutional layer, a feature map is generated which is the input for the

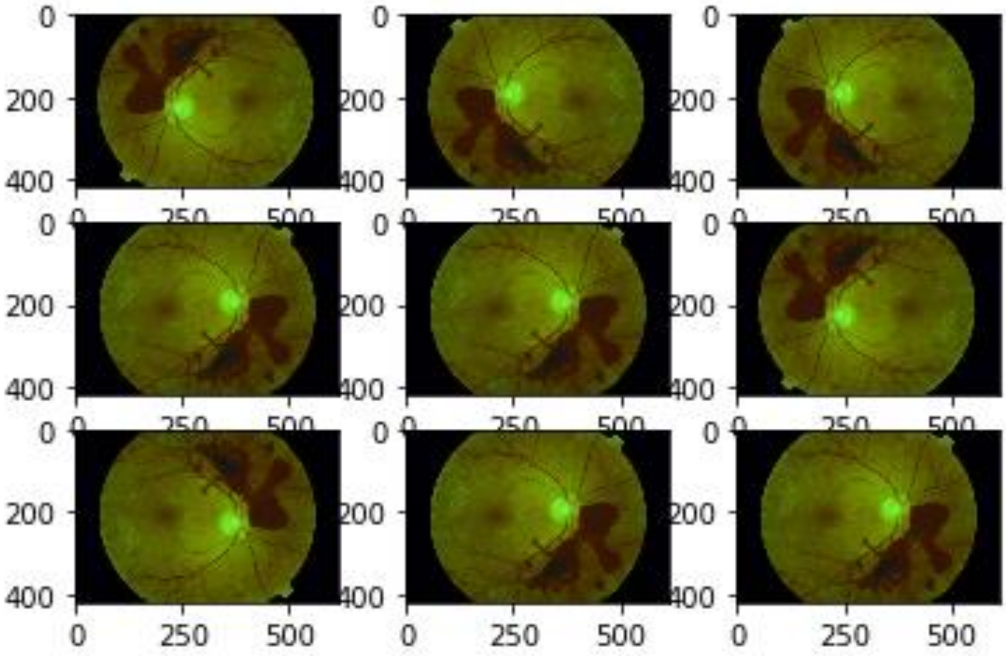


FIGURE 3: An illustration of data augmentation in retinal images.

second layer. We consider max pool as the pooling layer, which down sample the resulting feature maps and increases the receptive field on the filters [33]. To induce non-linearity to the feature maps, activation functions are applied. In our case, we utilize the Exponential Linear Unit (ELU) activation function, which is defined as:

$$ELU(x) = \begin{cases} x, & if \ \ x > 0 \\ \alpha(e^x - 1), & if \ \ x \leq 0 \end{cases} \tag{10}$$

Although it is computationally expensive but it converges faster as compared to other activation functions. According to [34] the presence of the extra parameter α controls the saturation point for negative values and thereby it is computationally inexpensive as compared to other functions. This is what differentiates it from the commonly used Rectified Linear Unit (Relu) activation function. In a neural network, forward and backward propagation are one of the most important factors in determining the convergence performance of a CNN. The forward propagation has two steps. First, the calculation of Z, which is determined as:

$$Z = W^{[l]} \times A^{[l-1]} \times B^{[l]} \tag{11}$$

Here W is a tensor which has a filter and B is the bias term. The second step, applying the activation function as follow:

$$A^{[l]} = K^{[l]}(Z^{[l]}) \tag{12}$$

Here, k denotes the activation function. This process is followed by a backward propagation in which partial derivatives are calculated to update W and B for improving gradient descent convergence, and reducing the error. The partial derivative can be calculated as:

$$DX^{[l]} = \frac{\partial L}{\partial A} \tag{13}$$

Here, X denotes $A[L]$, $W[L]$, $B[L]$, which are activation, weight, and bias, respectively. The Weights are updated as:

$$W = W_i - \eta \frac{\partial l}{\partial w} \tag{14}$$

Here, η is the learning rate, W_i is initial weight. For the CNN model, the images are resized into $96 \times 96 \times 3$ before feeding into the neural network.

D. STACKED GENERALIZATION OF CNNS

The methodology of stacking multiple sub-models into a single meta-learner classifier to combine the prediction probabilities to reduce the generalization error by deducing individual biases of the sub-models is called the stacked generalization [35]. It is different from the usual model averaging methodology in terms of the fusion strategy because the classification results are not averaged, but the final output is decided by the weighted majority of the sub-models. In the case of deep learning, multiple CNNs with different architectures are merged before giving the final output. Different ensemble techniques have been applied in similar work [1], [24], however these were machine learning algorithms-based techniques, such as ensemble models of K-nearest neighbor, NaÃŕve Bayes and Decision tree. In contrary, we have used CNNs for stacked generalization.

In order to reduce the bias in machine learning, usually, the crude cross-validation techniques such as 10-fold cross-validation and Leave-one-out cross-validation are applied.

However, if those techniques are applied in CNNs, the complexity and computational time of the model increase tremendously unlike most of the machine learning algorithms. A CNN deals with millions of parameters during the forward and backward propagation [24]. Thus, different fusion-based ensemble methods have been presented by researchers every year to combine multiple predictions in the most optimized way. There are various fusion strategies for ensemble models including model averaging, where we combine the predictions from several independently trained models as adopted in [36]. In our case, we use the weighted majority based fusion technique for the stacked model.

Consider an ensemble of M independent classifiers D_1,...,D_M, with individual accuracies $p_1, p_2..., p_M$. Each classifier D_i produces $c - dimensional$ vector $[d_{i,1}, ..., d_{i,c}]^T \in 0, 1^c, i = 1, ...M$, where $d_{i,j}$ =1 if D_i labels x in ω_j, and 0 otherwise. The majority vote will result in an ensemble decision for class ω_k if

$$\sum_{i=1}^{M} d_{i,k} = max_{j=1}^{c} \sum_{i=1}^{M} d_{i,j} \qquad (15)$$

When introducing weights or coefficients of importance b_i,$i = 1, 2, ..., M$, and rewriting Eq.(15 as: choose class label ω_k if

$$\sum_{i=1}^{M} b_i d_{i,k} = max_{j=1}^{c} \sum_{i=1}^{M} b_i d_{i,j} \qquad (16)$$

The outputs are combined by the maximum weighted majority as shown in Eq. (16) [37]. This results in an improved overall performance because the models that perform well individually, contribute more to the final metrics compared to less performing models. A stacked generalization is a multi-level learning model because at each level it aims to select the most appropriate bias for minimizing the overall generalization error.

Figure 4 shows the stacked model of CNNs prepared for our experiments. Three different CNN architectures are prepared to be fed into the meta-learner classifier. Therefore, three different copies of the input data for each of the network is made, and after the concatenation a 3-element vector of prediction probabilities is created, which can be seen in the concatenate_7 layer of the stacked model from 3 different sub-models and 2-class labels after applying the sigmoid function to produce result as either 0 (No DR) or 1 (DR).

V. PERFORMANCE ANALYSIS

A. DATASET DESCRIPTION

The dataset used in our experiment was acquired from a Kaggle competition and is a benchmark dataset for DR diagnosis provided by EyePACS [38]. EyePACS is a web-based system designed to remotely help patients deal with DR issues without the need of a doctor. It is a platform where clinicians could collaborate and share their work, which could be further used for research purposes. EyePACS shared their data with Google and Kaggle to host a competition for tackling DR where people could contribute to open further research areas through using their open-source retinal database. The dataset is highly imbalanced with the number of healthy images overshadowing the number of severe and advanced stage of DR images.

Figure 5 shows a corpus of sample retinal images obtained from the EyePACS dataset. Since we are detecting DR and not classifying the stages therefore, we used a subset of balanced data and the images were divided and put into two different folders of healthy and unhealthy retinal images. Figure 6 explains about the feature descriptions of the retinal fundus images used in our experiments. We have used 2471 images and divided them as 20% for validation and 80% for training. All images are colored and kept in the original .JPEG format. Since images were not acquired by the same camera and were taken in different lighting conditions with visible illumination variance, color combination, camera angles, therefore, an image normalization technique was implemented before feature extraction.

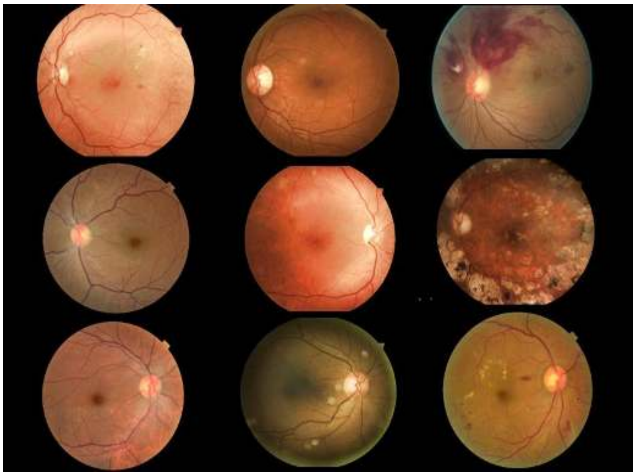


FIGURE 5: Sample fundus images from the EyePACS dataset.

B. MODEL BUILDING

The goal of our experiment is to make an accurate validation tool for doctors to detect DR by avoiding unnecessary reflectance properties of fundus images and making light a reliable factor other than unnecessary distortions. A stacked generalization of three different CNNs was prepared and fed into a meta-learner as shown in Figure 4. Out of 2471 images, 495 images are kept for testing the model and 1976 images are used for training. During the stacking process, it is important to train the meta-learner on a separate dataset other than the data on which individual sub-networks are trained to avoid any sort of overfitting and bias in the results. That is why each of the sub-model is trained on the same training set, however the test results by the meta-learner were tested on a separate test set.

Table 3 gives the information about the class-wise dis-

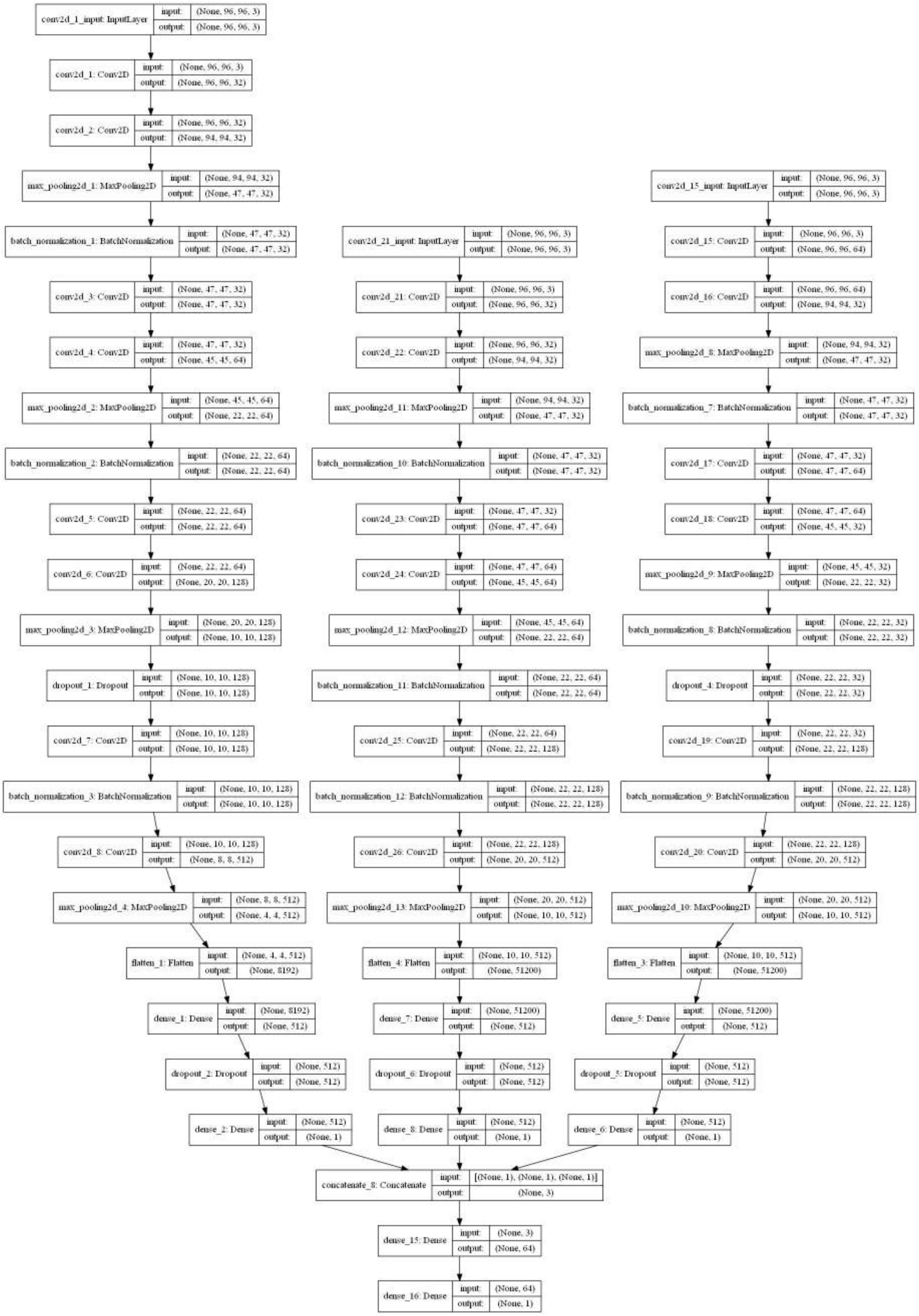


FIGURE 4: An illustration of the stacked CNNs concatenated on top of the meta-learner classifier.

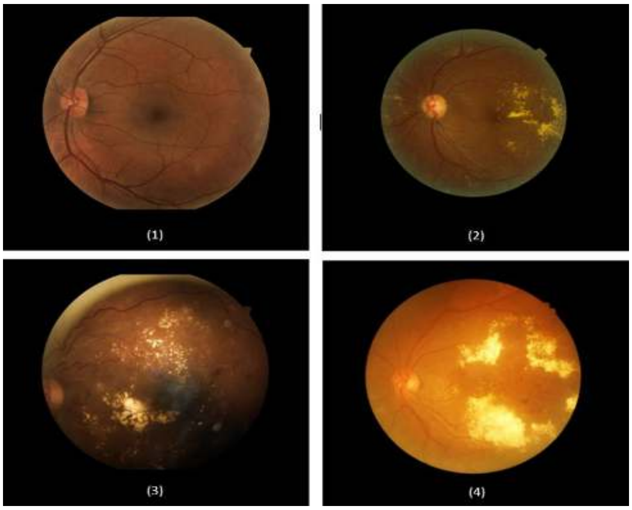


FIGURE 6: DR dataset for fundus image diagnosis. Figure 6 (1) shows the healthy retinal image having no signs of any type of lesions and hemorrhage. Figure 6 (2) shows the unhealthy retinal image having mild stage DR because it has some lesion presence. Figure 6 (3) shows the presence of yellowish irregular edges known as the hard exudates in the unhealthy retinal image. Figure 6 (4) shows the unhealthy retinal image having severe stage of DR due to prominent presence of cotton wool spots, which are caused due to the accumulations of the axoplasmic material in the retina.

tribution of the dataset used for our experiments. Image processing and normalization are applied using libraries such as OpenCV, NumPy, and PIL and for the model development, Keras and TensorFlow libraries are used.

TABLE 3: Label wise depiction of dataset division used for training and testing of the stacked CNN and other transfer learning models.

Class Labels	Training Dataset	Testing Dataset
Healthy (No DR)	900	235
Unhealthy (Having DR)	1076	260
Total	**1976**	**495**

The different CNNs which are used as sub-models are comprised of many successive convolution layers, pooling layers, and batch normalization layers. The architecture of the three sub-models is not similar in terms of the number of layers and the combination of pooling and batch normalization. The reason for stacking different architectures is to achieve different results for better generalization. However, hyperparameters such as learning rate, batch size, epochs are identical for each sub-model. The final fully connected layers are apply sigmoid function to give the binary diagnosis. Table 4 illustrates the layer-wise ConvNet configuration of the first CNN which is the part of the stacking model.

The models are trained for 18 epochs with batch size of 16. Accuracy was considered as the initial metric of evaluation. Stochastic gradient descends (SGD) was used as the model optimizer. To prevent the model from overfitting, hyperparameter tuning was also applied but the problem of overfitting persisted. To address this issue two important techniques were utilized. First, we applied a learning rate decay. The initial learning rate is set to 0.00009 and a rate decay equivalent to learning rate/100 is applied. Therefore, a gradual decrease in the learning rate with a factor of 100 improved the gradient descend convergence. Second, two regularization techniques are also implemented which are discussed in the next subsection.

C. REGULARIZATION BY CALL-BACK FUNCTIONS

Call-back is a set of functions that are applied to induce a regularization effect to generalize the deep learning model and stabilize the estimates to combate overfitting. Usually, the regularization techniques increase the bias and reduce the variance of the model [39]. First, we applied a strategy known as early stopping in which the training is stopped prematurely as the validation loss tends to increase resulting in a steep increase in the loss curve and decrease in the model performance, and thus giving us an optimal stopping point. For early stopping, the hyper-parameter patience was set to 2. In the deep learning models, it is commonly seen that the validation loss vs. training loss graph gets stuck at an inflection point, where the loss does not decrease and no improvement in performance is detected. That inflection point looks like a plateau formed in the graph. Thus, the second function we used for regularization was to reduce the learning rate by a factor of 100 when such plateau is reached during the training process.

D. MODIFIED DEEP TRANSFER LEARNING MODEL

To fairly compare our findings, we implemented two different deep transfer learning models, which are ResNet50 [19] and VGG-16 [40]. Transfer learning models can be used in different ways to transfer the learned features from pretrained models [41]. However, the most prominent work follows two scenarios, which are known as feature extraction and fine-tuning. In our experiments, we utilize the pretrained models as feature extractors.

We used the same dataset and divided it in the similar ratio as done for the stacked generalization CNN model. Data augmentation was also applied for improving the diversity of the data. Both ReseNet50 and VGG-16 models were fine-tuned as it is necessary to improve the performance. In our case we did the layer-wise fine tuning similar to [42], [43], as it is more effective and less time consuming. We added the fully connected layer head to ResNet50 and VGG-16, which consists of a pooling layer, fully connected layer and the final layer having sigmoid function to give us the binary output. The weights of both VGG-16 and ResNet50 were frozen so that only the fully connected layers were adjusted. Similarly, we trained the networks for 18 epochs with a batch-size of 16.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3101142, IEEE Access

TABLE 4: Layerwise configuration of a single CNN architecture which is fed into the meta-learner classifier.

Operation Layer	Filter Size	Padding	Strides	Output Image Size	Dropout	Activation
Convolution Layer	$3 \times 3 \times 32$	same	1	$96 \times 96 \times 32$	-	Elu
Convolution Layer	$3 \times 3 \times 32$	-	1	$94 \times 94 \times 32$	-	-
Max-pool2D	2×2	valid	-	$47 \times 47 \times 32$	-	-
Batch-Normalization	-	-	-	$47 \times 47 \times 32$	-	-
Convolution Layer	$3 \times 3 \times 32$	same	1	$47 \times 47 \times 32$	-	Elu
Convolution Layer	$3 \times 3 \times 64$	valid	1	$45 \times 45 \times 64$	-	-
Max-pool2D	2×2	valid	-	$22 \times 22 \times 64$	-	-
Batch-Normalization	-	-	-	$22 \times 22 \times 64$	-	-
Convolution Layer	$3 \times 3 \times 64$	same	1	$22 \times 22 \times 64$	-	Elu
Convolution Layer	$3 \times 3 \times 128$	valid	1	$20 \times 20 \times 128$	-	-
Max-pool2D	2×2	valid	1	$10 \times 10 \times 128$	-	-
Dropout	-	-	-	$10 \times 10 \times 128$	10%	-
Convolution Layer	$3 \times 3 \times 128$	same	1	$10 \times 10 \times 128$	-	Elu
Batch-Normalization	-	-	-	$10 \times 10 \times 128$	-	-
Convolution Layer	$3 \times 3 \times 512$	same	valid	$8 \times 8 \times 512$	-	Elu
Max-pool2D	2×2	valid	1	$4 \times 4 \times 512$	-	-
Flattening Layer	-	-	-	8192	-	-
Fully Connected Layer	-	-	-	512	-	Elu
Dropout	-	-	-	512	50%	-
Output	-	-	-	1	-	-

VI. QUANTITATIVE ANALYSIS

Three major checkpoints are cleared in our experiments. First, solving the multi-sourced dataset problem of normalizing non-uniform luminosity by desaturating images using their statistical features such as mean pixel values and an optimum scaling factor. Second, developing an automated detection system for the normalized fundus images using an advanced artificial intelligence technique known as stacked generalization of CNNs, which uses the principle of weighted majority of sub-models. Third, to support our experimental results with proof, various comparisons are drawn with the benchmark deep transfer learning models.

A. LUMINOSITY NORMALIZATION ANALYSIS

The results of the proposed image illumination normalization technique are shown in Figure 2. It is visible that the images are color calibrated using the gray world algorithm. After applying this algorithm, the images are desaturated using the illuminant, which is taken as the minimum magnitude color channel. This aids in the transformation of a uniformly luminous image, and thus removing the presence of the unnecessary reflectance. The saturation loss of these images helped to reduce unnecessary hindrance like noise, non-uniform light distribution, and non-ideal illuminations. Therefore these images are reliable and could be used for further analysis. In medical imaging, degradation of image features is a major issue while implementing normalization techniques. The yellow arrows in Figure 2 clearly show the presence of features such as blood vessels, hemorrhages, retinal macula, which play a major role in decision-making for DR diagnosis.

To support our arguments and provide more concrete proof for the proposed luminosity normalization technique we have calculated the Peak signal-to-noise ratio (PSNR) and mean squared error (MSE) of the transformed image. PSNR can be defined as the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the quality of its representation [44]. PSNR is measured in Decibels (dB) and in most cases, a higher value of PSNR indicates that the enhanced or reconstructed image is of superior quality. On the other hand, MSE tells us about the difference in the images by computing the average of the squared errors between two images. The lesser the value of MSE, the better the image enhancement technique. Mathematically, PSNR and MSE can be defined as follows:

$$PSNR = 10\, log_{10}(\frac{(2^n - 1)^2}{MSE}) \tag{17}$$

$$MSE = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (k(i,j) - l(i,j))^2 \tag{18}$$

Here, M and N define the number of rows and columns in the image, respectively. $k(i,j)$ represents the original referenced image and $l(i,j)$ represents the luminosity normalised image. n stands for the max value of a pixel in the image. Figure 7 shows the difference between statistical values of PSNR and MSE between normal gray image and gray world normalised image.

B. STACKED CNN MODEL ANALYSIS

The images of our model are available in the .JPEG format. Since it is a lossy compression, the images lose a significant amount of information that makes the feature extraction imprecise and difficult as explained in [45]. Therefore, a superior feature extraction technique known as stacked generalization of CNNs has been implemented. Since it is a binary classification task, high values of accuracy and other evaluation metrics are expected. To prepare a robust model, hyperparameter tuning is performed, which shows the potential fluctuation and improvement in accuracy and loss. The experimental results reveal that the proposed stacked CNN model achieves an accuracy of 97.92% on the training set with a training loss of 0.066. On the test set the model achieves an accuracy of 97.77% with a test loss of 0.078.

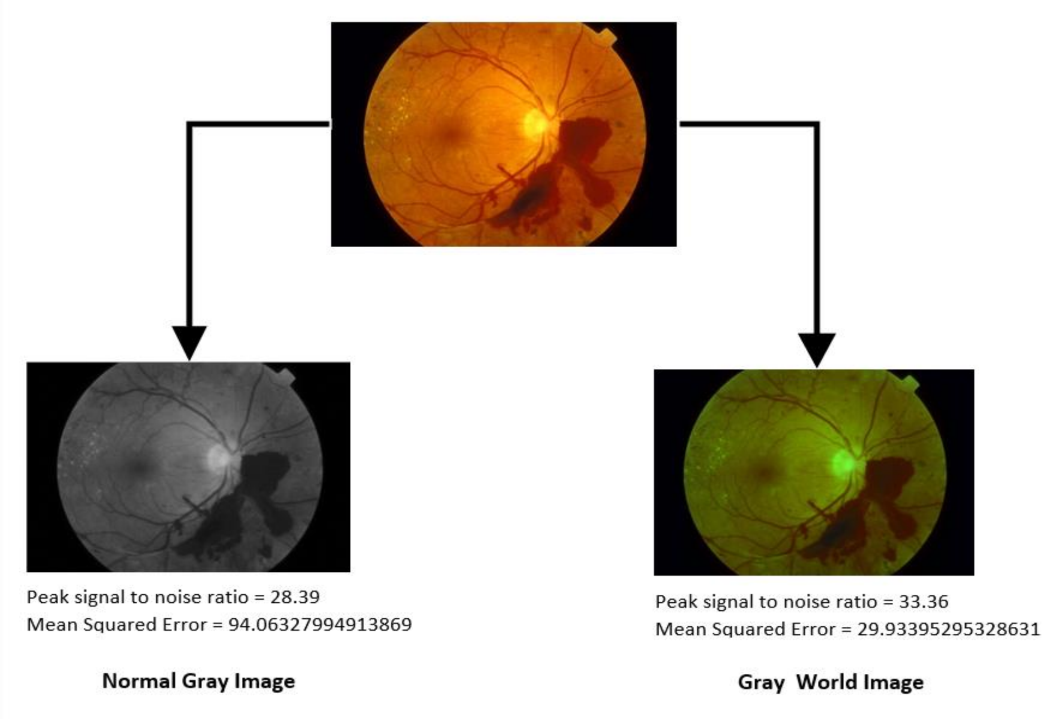

Peak signal to noise ratio = 28.39
Mean Squared Error = 94.06327994913869
Normal Gray Image
Peak signal to noise ratio = 33.36
Mean Squared Error = 29.93395295328631
Gray World Image

FIGURE 7: Statistical comparative analysis between normal gray image and gray world normalized image based on PSNR and MSE values.

Table 5 shows the evaluation metrics obtained after using various activation functions with and without data augmentation. It can be clearly observed that the ELU activation function works best with data augmentation and gave better results in terms of accuracy, sensitivity, and specificity.

Table 6 depicts the performance metrics such as train loss, test loss, train accuracy, test accuracy, and area under curve (AUC) values in comparison with the other deep transfer learning models. Experimental results reveal that the proposed model achieves an AUC value of 0.9979. AUC is an important performance measure that proves the model's reliability over other solutions.

Table 7 shows a report containing precision, recall, and F-measure scores of all the competent models. These are important metrics in the evaluation of a computer-aided diagnostic system. Precision score depicts the exactness and tells how often the predicted value is correct, whereas the F1-measure, which is the harmonic mean of recall and precision, reveals the test accuracy. From Table 7, we observe that the proposed stacked CNN model outperforms all other competitive models.

Figure 8 shows the confusion matrix for each model, which summarizes the predicted results and the type of errors compactly over the test set. Although VGG-16 has a lesser number of false negatives (FN) as compared to the stacked CNN model but the model obtains greater number of false positives (FP) and lesser number of true negatives (TN). However, the proposed stacking ensemble model has a greater number of TN and zero FP, which reveals its accuracy for both healthy (No DR) and unhealthy (Having DR) images. From Figure 8 (d) we observe that our proposed model has got only 11 FN, which means that only 11 out of 495 patients are falsely predicted as not having DR. Since we are dealing with a real-life problem in the medical domain, reducing false negatives as well as achieving considerably higher values for true positives are important. The medical domain is a field of precision, it is important to consider such metrics of evaluation that directly deal with the correct and incorrect predictions. So, we have considered sensitivity and specificity, which are discussed in detail below.

TABLE 7: Comparative analysis between the proposed model and other deep transfer learning models on similar dataset.

Architecture	Precision	Recall	F1-Score
VGG-16	0.92	0.97	0.94
ResNet50	1.00	0.94	0.96
CNN	1.00	0.95	0.974
Stacked CNNs	**1.00**	**0.96**	**0.979**

C. SENSITIVITY AND SPECIFICITY ANALYSIS

Sensitivity and specificity play a crucial role in the medical domain. Higher values of sensitivity and specificity prove the reliability of a diagnostic model. Sensitivity is the ability of the model to successfully predict the actual positive value [46], which in our case, to correctly predict the unhealthy fundus image of a patient as having DR. Mathematically, it

TABLE 5: Performance of different activation functions in the proposed model with data augmentation.

Learning Rate	Evaluation Metrics	Without Data Augmentation			With Data Augmentation		
		Relu	LeakyRelu	Elu	Relu	LeakyRelu	Elu
0.00001	Accuracy	94.45%	96.77%	96.32%	95.67%	95.72%	94.23%
	Sensitivity	95.32%	94.31%	92.23%	96.12%	93.13%	91.63%
	specificity	91.09%	93.06%	90.42%	93.33%	91.49%	90.47%
0.00003	Accuracy	95.21%	89.73%	96.99%	94.33%	96.72%	97.12%
	Sensitivity	96.89%	90.08%	91.74%	92.24%	96.00%	95.21%
	specificity	91.08%	88.32%	90.99%	90.52%	93.34%	94.11%
0.00009	Accuracy	95.72%	85.92%	96.59%	92.21%	93.24%	97.77%
	Sensitivity	94.88%	91.82%	90.12%	94.23%	89.42%	96.86%
	specificity	93.91%	90.21%	89.74%	95.21%	90.21%	100%

TABLE 6: An evaluation metric report of the proposed models.

Model	Training Loss	Test Loss	Train Accuracy	Test Accuracy	AUC
VGG-16	0.115	0.132	95.85%	94.34%	0.9910
ResNet50	0.082	0.095	97.36%	96.96%	0.9972
CNN	0.083	0.098	97.46%	97.37%	0.9980
Stacked CNNs without data augmentation	0.160	0.188	94.92%	92.99%	0.9776
Stacked CNNs with data augmentation	**0.066**	**0.078**	97.92**%**	97.77**%**	**0.9979**

can be measured in terms of percentage as:

$$Senstivity = \frac{TP}{TP + FN} \times 100 \qquad (19)$$

Specificity, on the other hand, shows how accurately is the model in detecting those people who do not have DR. In other words, it correctly predicts the healthy fundus image. Achieving high values of specificity may also have a business impact as it can save time for an ophthalmologist to carry out further tests if an earlier report is correctly predicted negative. Mathematically, specificity can be measured in terms of percentage as:

$$Specificity = \frac{TN}{TN + FP} \times 100 \qquad (20)$$

The sensitivity, specificity, and accuracy of our proposed models are plotted in Figure 9, where we can see that our proposed stacking ensemble CNN model achieves higher scores than other competitive models. However, the sensitivity of VGG-16 is a slightly higher than other models, but due to more false positives, its specificity is low.

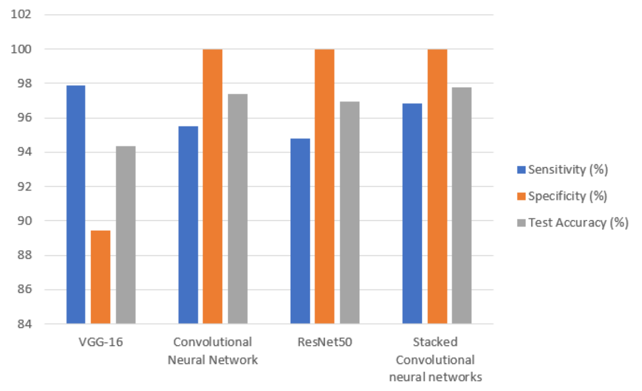


FIGURE 9: Bar plots for evaluation metrics of the proposed stacked CNNs model with VGG-16, CNN, and ResNet50.

D. PERFORMANCE DURING TEST PHASE

We closely monitored the performance of the proposed stacked CNN model with and other competent deep transfer learning models during training and testing stages. However monitoring at testing stage is very important to see the performance on unseen data to validate the generalizability and cross-check how well the model has learned during training. All the models are trained for the same number of epochs on the same dataset to minimize all the possible redundancies and discrepancies. Figure 10 shows the convergence of the loss curve during the testing phase. It is visible that the proposed model outperforms all other models till the end of all the iterations as its loss curve goes to the global optimum point of 0.078.

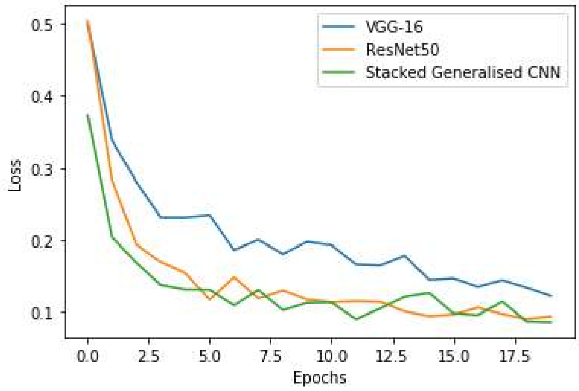
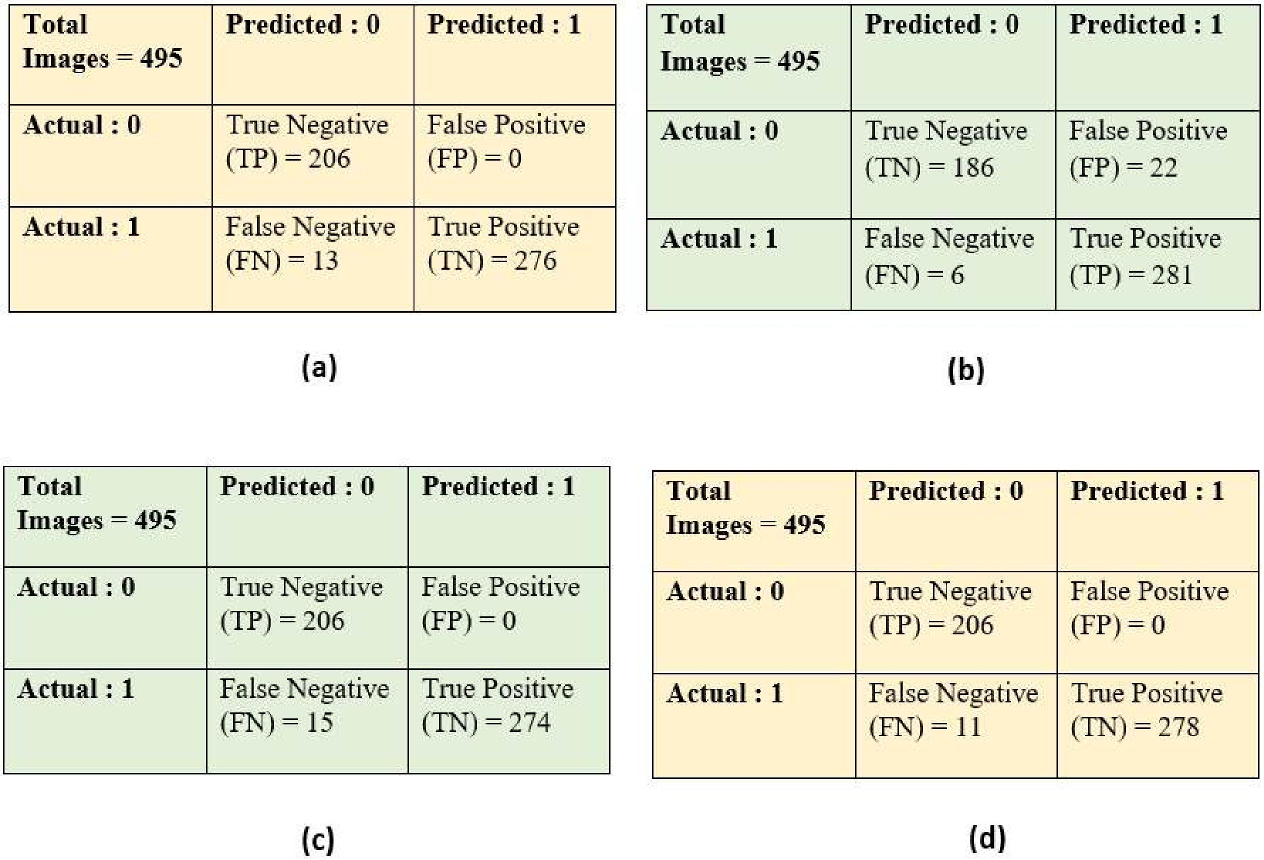


FIGURE 10: Test stage performance analysis of the proposed stacked CNN model with competitive transfer learning models.

VII. DISCUSSION

The contribution of our experiments includes the use of a publicly available EyePACS dataset from Kaggle. The image data is multi-sourced with various discrepancies due to



Total Images = 495	Predicted : 0	Predicted : 1
Actual : 0	True Negative (TP) = 206	False Positive (FP) = 0
Actual : 1	False Negative (FN) = 13	True Positive (TN) = 276

(a)

Total Images = 495	Predicted : 0	Predicted : 1
Actual : 0	True Negative (TN) = 186	False Positive (FP) = 22
Actual : 1	False Negative (FN) = 6	True Positive (TP) = 281

(b)

Total Images = 495	Predicted : 0	Predicted : 1
Actual : 0	True Negative (TP) = 206	False Positive (FP) = 0
Actual : 1	False Negative (FN) = 15	True Positive (TN) = 274

(c)

Total Images = 495	Predicted : 0	Predicted : 1
Actual : 0	True Negative (TP) = 206	False Positive (FP) = 0
Actual : 1	False Negative (FN) = 11	True Positive (TN) = 278

(d)

FIGURE 8: Confusion matrices of the proposed model and other compared models. (a) CNN, (b) VGG-16, (c) ResNet50, and (d) stacked CNN model.

various reasons such as different cameras and lighting conditions. Therefore, image normalization is very important. The images are pre-processed for luminosity normalization using the gray world color constancy algorithm to enhance the candidate regions by reducing the unnecessary lighting and reflectance. To confirm and support the results of our normalization step, we analyzed the enhanced images based on PSNR and MSE measures. The PSNR value was improved as shown in Figure 7, which proves the importance and effect of our color correction schema.

Researchers have presented similar work related to color constancy and retinal image enhancement using various techniques as described in the literature review. However, an automated tool using these techniques has not been presented. Most of the algorithms focused on extracting features such as cotton wool spots, exudates, lesion presence, hemorrhage detection for disease diagnosis, but did not discuss luminosity normalization as a pre-processing step. The diagnostic decision-making stage was handled by stacked generalization of CNNs, which proved to be better than other competitive models including VGG-16, ResNet50 and CNNs. Comparisons are also drawn between the proposed model and other models in terms of accuracy, sensitivity, specificity, precision, recall, and F-measure.

There are two main theories behind developing an automated validation tool that could remove the non-ideal illuminations from retinal fundus images using deep learning. The first was to reduce the human effort in extracting manual features for Diagnosis and let the power of artificial intelligence and image processing techniques extract and enhance features automatically. The second was the adaptability of deep learning models to solve a variety of problems and the availability of optimization methodologies such as various regularization techniques for better performance. Our major focus was also to reduce the number of false negatives and the experimental results on unseen test data showed that only 2.2% false negatives, which proves the reliability of our model. Our method is also economically viable to implement as it does not require expensive equipment/gadgets with high graphical processing unit (GPU) power. According to [47] sensitivity values achieved in detecting DR greater than 60% proves to be cost-effective. Since our model was trained with a dataset having a lot of variances, it also proves the high adaptability and robustness of our model to perform accu-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3101142, IEEE Access

rately with fundus images having non-ideal illuminations.

TABLE 8: Performance analysis of related works on binary fundus dataset.

Study	Accuracy	Sensitivity	Specificity
Zago et al. [12]	91.2%	94.0%	NA
Tahira et al.[27]	95.2%	96.1%	96.5%
Moazam et al.[48]	87.72%	92.4%	81.25%
Hemanth et al.[49]	97.00%	94.00%	98.00%
Gadekallu et al.[50]	97.30%	91%	97%
Stacked CNN model	97.92%	97.77%	100%

Table 8 compares the performance of the proposed model on binary classification with previous work conducted on DR detection using similar multi-sourced datasets. To verify the results, our proposed model has been tested on several binary and multi-class datasets. It can be observed that [50] obtained an accuracy of 97.30%, however, the sensitivity of their model is lower compared to our proposed stacked model. Therefore, our model is superior in detecting true positives accurately.

Table 9 compares the performance of our model with previous studies for multi class classification on the Kaggle dataset [51]. The dataset has five satges of DR including: healthy, mild, moderate, severe, advanced as summarized in Table 10. The proposed model achieved the highest sensitivity and specificity values and outperformed all other models with a final test accuracy of 87.45%. This accuracy score is inferior to the binary classification results due to the imbalanced data as depicted in the given dataset. Table 11 presents the performance of the proposed model on various binary and multi-class datasets in terms of accuracy and precision. Considering all metrics of Tables 8, 9, and 10 together, it can be concluded that the proposed model outperforms state-of-art models and is successful in both binary and multi-class classification of DR.

Figure 11 depicts the sensitivity and specificity of the proposed models compared with the different machine/deep learning methods carried out in the literature [2], [23].

TABLE 9: Performance analysis of related works on multi-class fundus dataset.

Study	Accuracy	Sensitivity	Specificity
Sehrish et al. [18]	80.8%	NA	86.7%
Alexander et al.[52]	NA	92%	72%
Carson et al.[53]	74.5%	95%	96%
Chetoui et al.[54]	NA	95.8%	97.1%
Stacked CNN model	87.45%	96.30%	97.28%

TABLE 10: Class distribution of the Kaggle multi-class dataset [51].

Healthy	Mild	Moderate	Severe	Advanced	Total
25810	2443	5292	873	708	35126

Figure 12 shows the ROC curve of the proposed model for binary classification task where it obtains an AUC value of 0.99. The results provided in Table 8 and Figure 12 prove the potential of the proposed stacking deep learning technique. Our model is able to outperform the conventional methods for diagnosis. Finally, our stacked generalization of CNNs achieve accuracy of 97.92% on the train set and 97.77% on the test set, a sensitivity of 96.86%, and a specificity of 100% in binary classification. The Proposed model also outperforms ResNet50 in terms of accuracy and F-measure. For multi-class classification, the model achieves train and test accuracy of 96.45%, 96.30% respectively as reported in Table 9.

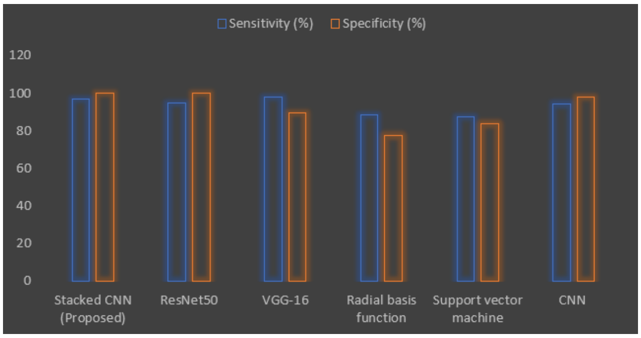


FIGURE 11: Sensitivity and specificity based graphical analysis of the proposed model with the work of [2], [23].

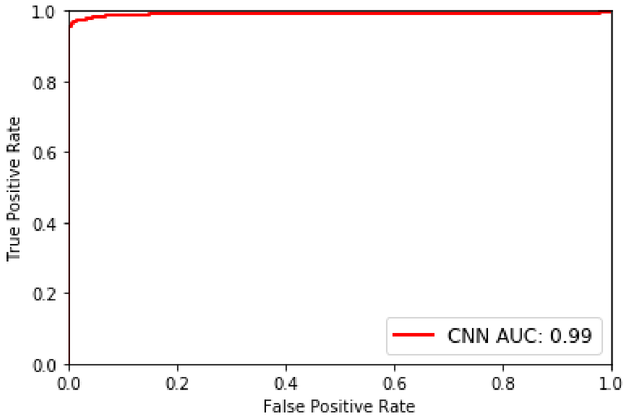


FIGURE 12: ROC curve of the proposed stacked CNN model with AUC value of 0.99.

VIII. CONCLUSION AND FUTURE WORK

We proposed to solve the problem of non-ideal illuminations in the retinal fundus images using the gray world algorithm and to develop an automated DR prediction system. A stack generalization-based ensemble model is prepared using three different CNNs. The performance of image normalization is measured using statistical metrics such as the PSNR and MSE of the original and enhanced images. The stacked ensemble model is an advanced technique of stacking different neural networks whose combined results are produced based on a fusion strategy that combines the best weights of the individual neural networks. Machine learning models

TABLE 11: Performance evaluation of the proposed model on the publicly available fundus datasets.

Dataset	Number of samples	Classification	Accuracy		Precision
			Train	Test	
DIAREDBI [55]	89	Binary	97.84%	96.21%	0.96
Messidor [56]	1200	Multi-class	97.42%	95.21%	0.95
DDR [57]	13673	Multi-class	94.25%	92.13%	0.92
IDRID [58]	516	Multi-class	97.42%	95.23%	0.95
STARE [59]	20	Binary	98.57%	98.11%	0.98
E-Optha [60]	381	Binary	98.48%	97.31%	0.97

are extensively utilized to classify and detect DR in fundus images. However, these techniques require suitable preprocessing and feature extraction methods to improve the results especially when the images are from different sources. DR images are generally taken from different cameras under different lighting conditions and to mitigate these effects we adopted an efficient color constancy technique. Extensive experiments are conducted to evaluate the performance of the proposed model in binary as well as multi-class DR classification tasks. Considering the obtained results using various evaluation metrics, we validate our model, which outperforms state-of-art models in binary and multi-class classification tasks.

For future work, we think of diversifying and increasing the images in the dataset for improving the feature extraction capabilities. Metaheuristic techniques can be used for hyperparameter optimization to achieve more competitive results. The patient's family medical history, daily diet, and nutrition intake can be included in the dataset to provide insightful information for the disease.

REFERENCES

[1] B. Antal and A. Hajdu, “An ensemble-based system for automatic screening of diabetic retinopathy,” Knowledge-Based Systems, vol. 60, pp. 20–27, 2014.

[2] R. Gargeya and T. Leng, “Automated identification of diabetic retinopathy using deep learning,” Ophthalmology, vol. 124, no. 7, pp. 962–969, 2017.

[3] E. T. D. R. S. R. Group et al., “Early photocoagulation for diabetic retinopathy: Etdrs report number 9,” Ophthalmology, vol. 98, no. 5, pp. 766–785, 1991.

[4] S. S. Gadkari, Q. B. Maskati, and B. K. Nayak, “Prevalence of diabetic retinopathy in india: The all india ophthalmological society diabetic retinopathy eye screening study 2014,” Indian journal of ophthalmology, vol. 64, no. 1, p. 38, 2016.

[5] M. Zhou, K. Jin, S. Wang, J. Ye, and D. Qian, “Color retinal image enhancement based on luminosity and contrast adjustment,” IEEE Transactions on Biomedical Engineering, vol. 65, no. 3, pp. 521–527, 2017.

[6] W. A. Mustafa, H. Yazid, and M. M. M. Abdul Kader, “Luminosity correction using statistical features on retinal images,” vol. 37, pp. 74–84, 2018.

[7] Ö. Deperlıoğlu and U. Köse, “Diagnosis of diabetic retinopathy by using image processing and convolutional neural network,” in Proceedings of 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2018, pp. 1–5.

[8] J. Wang, Y. Bai, and B. Xia, “Simultaneous diagnosis of severity and features of diabetic retinopathy in fundus photography using deep learning,” IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 12, pp. 3397–3407, 2020.

[9] V. M. G. S. Gupta, S. Gupta, P. Sengar et al., “Extraction of blood veins from the fundus image to detect diabetic retinopathy,” in Proceedings of the 1st IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES), 2016, pp. 1–3.

[10] L. Zhou, Y. Zhao, J. Yang, Q. Yu, and X. Xu, “Deep multiple instance learning for automatic detection of diabetic retinopathy in retinal images,” IET Image Processing, vol. 12, no. 4, pp. 563–571, 2018.

[11] M. Masood, T. Nazir, M. Nawaz, A. Mehmood, J. Rashid, H.-Y. Kwon, T. Mahmood, and A. Hussain, “A novel deep learning method for recognition and classification of brain tumors from mri images,” Diagnostics, vol. 11, no. 5, p. 744, 2021.

[12] G. T. Zago, R. V. Andreão, B. Dorizzi, and E. O. T. Salles, “Diabetic retinopathy detection using red lesion localization and convolutional neural networks,” Computers in Biology and Medicine, vol. 116, p. 103537, 2020.

[13] K. U. Bhaskar and E. P. Kumar, “Extraction of hard exudates using functional link artificial neural networks,” in IEEE International Advance Computing Conference (IACC), 2015, pp. 420–424.

[14] A. Bandara and P. Giragama, “A retinal image enhancement technique for blood vessel segmentation algorithm,” in Proceedings of the IEEE international conference on industrial and information systems (ICIIS), 2017, pp. 1–5.

[15] W. A. Mustafa, H. Yazid, and S. B. Yaacob, “Illumination correction of retinal images using superimpose low pass and gaussian filtering,” in Proceedings of the 2nd International Conference on Biomedical Engineering (ICoBE), 2015, pp. 1–4.

[16] N. Singh, L. Kaur, and K. Singh, “Histogram equalization techniques for enhancement of low radiance retinal images for early detection of diabetic retinopathy,” Engineering Science and Technology, an International Journal, vol. 22, no. 3, pp. 736–745, 2019.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proceedings of the 25th International Conference on Neural Information Processing Systems, 2012, pp. 1097–1105.

[18] S. Qummar, F. G. Khan, S. Shah, A. Khan, S. Shamshirband, Z. U. Rehman, I. A. Khan, and W. Jadoon, “A deep learning ensemble approach for diabetic retinopathy detection,” IEEE Access, vol. 7, pp. 150 530–150 539, 2019.

[19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[20] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700–4708.

[21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818–2826.

[22] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1251–1258.

[23] S. Somasundaram and P. Alli, “A machine learning ensemble classifier for early prediction of diabetic retinopathy,” Journal of Medical Systems, vol. 41, no. 12, pp. 1–12, 2017.

[24] E. Saleh, J. Błaszczyński, A. Moreno, A. Valls, P. Romero-Aroca, S. de la Riva-Fernandez, and R. Słowiński, “Learning ensemble classifiers for diabetic retinopathy assessment,” Artificial Intelligence in Medicine, vol. 85, pp. 50–63, 2018.

[25] K. Oh, H. M. Kang, D. Leem, H. Lee, K. Y. Seo, and S. Yoon, “Early detection of diabetic retinopathy based on deep learning and ultra-wide-field fundus images,” Scientific Reports, vol. 11, no. 1, pp. 1–9, 2021.

[26] H. Khalid, R. Schwartz, L. Nicholson, J. Huemer, M. H. El-Bradey, D. A. Sim, P. J. Patel, K. Balaskas, R. D. Hamilton, P. A. Keane et al., “Wide-field optical coherence tomography angiography for early detection and objective evaluation of proliferative diabetic retinopathy,” British Journal of Ophthalmology, vol. 105, no. 1, pp. 118–123, 2021.

[27] T. Nazir, A. Irtaza, J. Rashid, M. Nawaz, and T. Mehmood, “Diabetic retinopathy lesions detection using faster-rcnn from retinal images,” in

Proceedings of the 1st International Conference of Smart Systems and Emerging Technologies (SMARTTECH), 2020, pp. 38–42.

[28] K. A. Goatman, A. D. Whitwam, A. Manivannan, J. A. Olson, and P. F. Sharp, "Colour normalisation of retinal images," in Proceedings of Medical Image Understanding and Analysis, 2003, pp. 49–52.

[29] D. Liu, "Comparison analysis of color constancy algorithms," 2013.

[30] V. Agarwal, B. R. Abidi, A. Koschan, and M. A. Abidi, "An overview of color constancy algorithms," Journal of Pattern Recognition Research, vol. 1, no. 1, pp. 42–54, 2006.

[31] G. Chen and X. Zhang, "A method to improve robustness of the gray world algorithm," in Proceedings of the 4th International Conference on Computer, Mechatronics, Control and Electronic Engineering, 2015, pp. 243–248.

[32] N. M. Kwok, D. Wang, X. Jia, S. Chen, G. Fang, and Q. P. Ha, "Gray world based color correction and intensity preservation for image enhancement," in Proceedings of the 4th International Congress on Image and Signal Processing, vol. 2, 2011, pp. 994–998.

[33] S. Indolia, A. K. Goswami, S. Mishra, and P. Asopa, "Conceptual understanding of convolutional neural network-a deep learning approach," Procedia Computer Science, vol. 132, pp. 679–688, 2018.

[34] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," arXiv preprint arXiv:1811.03378, 2018.

[35] D. H. Wolpert, "Stacked generalization," Neural networks, vol. 5, no. 2, pp. 241–259, 1992.

[36] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Ensembles of deep learning models and transfer learning for ear recognition," Sensors, vol. 19, no. 19, p. 4139, 2019.

[37] L. I. Kuncheva, Combining pattern classifiers: methods and algorithms, 2014.

[38] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," Jama, vol. 316, no. 22, pp. 2402–2410, 2016.

[39] Y. Gal, "Uncertainty in deep learning," University of Cambridge, vol. 1, no. 3, p. 4, 2016.

[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proceedings of the International Conference on Learning Representations (ICLR), 2015, pp. 1–14.

[41] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Handcrafted versus CNN features for ear recognition," Symmetry, vol. 11, no. 12, p. 1493, 2019.

[42] X. Li, T. Pang, B. Xiong, W. Liu, P. Liang, and T. Wang, "Convolutional neural networks based transfer learning for diabetic retinopathy fundus image classification," in Proceedings of the 10th international congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI), 2017, pp. 1–11.

[43] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Deep convolutional neural networks for unconstrained ear recognition," IEEE Access, vol. 8, pp. 170 295–170 310, 2020.

[44] N. M. W. A. Mustafa, H. Yazid, M. Jaafar, M. Zainal, and A. S. Abdul-Nasir, "A review of image quality assessment (iqa): Snr, gcf, ad, nae, psnr, me," Journal of Advanced Research in Computing and Applications, vol. 7, no. 1, pp. 1–7, 2017.

[45] T. Nazir, A. Irtaza, Z. Shabbir, A. Javed, U. Akram, and M. T. Mahmood, "Diabetic retinopathy detection through novel tetragonal local octa patterns and extreme learning machines," Artificial Intelligence in Medicine, vol. 99, p. 101695, 2019.

[46] R. Trevethan, "Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice," Frontiers in Public Health, vol. 5, p. 307, 2017.

[47] J. C. Javitt, J. K. Canner, R. G. Frank, D. M. Steinwachs, and A. Sommer, "Detecting and treating retinopathy in patients with type i diabetes mellitus: a health policy model," Ophthalmology, vol. 97, no. 4, pp. 483–495, 1990.

[48] M. M. Fraz, W. Jahangir, S. Zahid, M. M. Hamayun, and S. A. Barman, "Multiscale segmentation of exudates in retinal images using contextual cues and ensemble classification," Biomedical Signal Processing and Control, vol. 35, pp. 50–62, 2017.

[49] D. J. Hemanth, O. Deperlioglu, and U. Kose, "An enhanced diabetic retinopathy detection and classification approach using deep convolutional neural network," Neural Computing and Applications, vol. 32, no. 3, pp. 707–721, 2020.

[50] T. R. Gadekallu, N. Khare, S. Bhattacharya, S. Singh, P. K. R. Maddikunta, and G. Srivastava, "Deep neural networks to predict diabetic retinopathy," Journal Of Ambient Intelligence and Humanized Computing, pp. 1–14, 2020.

[51] "Diabetic retinopathy detection," https://www.kaggle.com/c/diabetic-retinopathy-detection, accessed: 2021-06-15.

[52] A. Rakhlin, "Diabetic retinopathy detection through integration of deep learning classification framework," BioRxiv, p. 225508, 2018.

[53] C. Lam, D. Yi, M. Guo, and T. Lindsey, "Automated detection of diabetic retinopathy using deep learning," AMIA summits on translational science proceedings, vol. 2018, p. 147, 2018.

[54] M. Chetoui and M. A. Akhloufi, "Explainable end-to-end deep learning for diabetic retinopathy detection across multiple datasets," Journal of Medical Imaging, vol. 7, no. 4, p. 044503, 2020.

[55] T. Kauppi, V. Kalesnykiene, J.-K. Kamarainen, L. Lensu, I. Sorri, A. Raninen, R. Voutilainen, H. Uusitalo, H. Kälviäinen, and J. Pietilä, "The diaretdb1 diabetic retinopathy database and evaluation protocol." in BMVC, vol. 1, 2007, pp. 1–10.

[56] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay et al., "Feedback on a publicly distributed image database: the messidor database," Image Analysis & Stereology, vol. 33, no. 3, pp. 231–234, 2014.

[57] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, and H. Kang, "Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening," Information Sciences, vol. 501, pp. 511–522, 2019.

[58] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, and F. Meriaudeau, "Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research," Data, vol. 3, no. 3, p. 25, 2018.

[59] A. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," IEEE Transactions on Medical Imaging, vol. 19, no. 3, pp. 203–210, 2000.

[60] E. Decenciere, G. Cazuguel, X. Zhang, G. Thibault, J.-C. Klein, F. Meyer, B. Marcotegui, G. Quellec, M. Lamard, R. Danno et al., "Teleophta: Machine learning and image processing methods for teleophthalmology," Irbm, vol. 34, no. 2, pp. 196–203, 2013.

• • •