4

The Multivariate Regression based Neural Network Model Fundus Image Quality Assessment

In the previous chapter, a detailed discussion of the prepared FIQuA data-set is provided. For each image, the data-set contains a total of seven subjective quality scores taken from the ophthalmologists. The first *six* scores are numeric values in the range of [0,10] for six quality parameters of fundus images. The last score is the quality class of the fundus image. Through experiments, it has been validated that the six subjective scores are unique enough to classify the fundus images into quality classes (good, fair, and poor) efficiently. With a similar approach, a new neural network based fundus IQA method is proposed. This chapter includes a detailed description of the proposed fundus IQA method named: Multivariate Regression based neural network model. The structure of the rest of the chapter is as follows: Section 4.1 explains in detail the proposed model for fundus IQA; Section 4.2 contains the implementation details of the proposed model; In section 4.3, a detailed analysis of the experimental results are provided; and finally, section 4.4 discusses the conclusions

The proposed model is also a two-step process: *Block-1:* Multivariate linear regression-based model that extracts optimized features against training for the subjective scores of F1-F6, and *Block-2:* Fusion of the optimized features obtained from step Block-1 for the classification. The comparison between the previous fundus IQA work and the proposed model is illustrated in Fig. 4.1. CNNs have proved to give extraordinary results not only in case of image classification [Szegedy *et al.*, 2015; Krizhevsky *et al.*, 2017] and object detection tasks [Han *et al.*, 2015; Zhang *et al.*, 2017a, 2020] but also for quality assessment [Bosse *et al.*, 2016; Kim *et al.*, 2018; Kang *et al.*, 2014b; Hou *et al.*, 2015]. The motivation for using CNNs is the reported performance of CNN based IQA models [Kim and Lee, 2017; Kang *et al.*, 2014a; Bosse *et al.*, 2018] for natural images. These reported works proved that CNN models are very effective for IQA and outperform the state-of-the-art methods. The architecture of the proposed fundus IQA model is shown in Fig. 4.2. The subsequent subsections provide the description of the aforementioned steps.

4.1 MODEL DESCRIPTION

The proposed model is built leveraging on two popular concepts of learning based algorithms: (i) Transfer learning [Pan and Yang, 2010], and (ii) Ensemble learning [Liu and Yao, 1999]. As anticipated above and illustrated in Fig. 4.2, the model is divided into two blocks. A detailed description of each block is given below:

Block-1: The objective of this block is to derive the optimized features for the final classification. Transfer learning has been used to achieve the objective. Transfer learning is a popular machine learning strategy where weights obtained from popular pre-trained networks on ImageNet [Deng *et al.*, 2009] alike large data-sets, are used as initial parameters to train another network. These pre-trained CNN models, like AlexNet [Krizhevsky *et al.*, 2017], GoogLeNet [Szegedy *et al.*, 2015], ResNet [He *et al.*, 2016], DenseNet [Huang *et al.*, 2017], Xception [Chollet, 2017], etc., are used to solve other object detection and classification problems, not only in the domain of natural images but also for other image domains. The reason for adopting the transfer learning methodology is the limited number of fundus images available

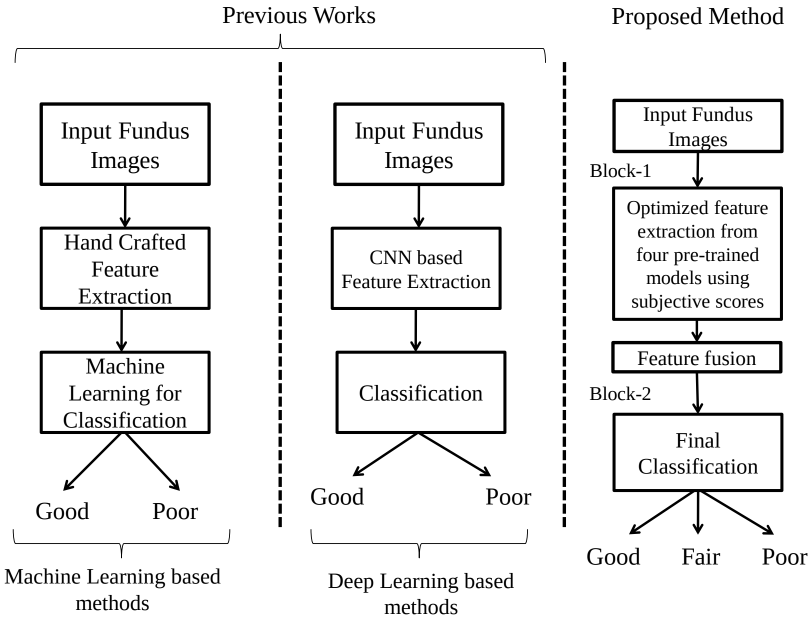

Previous Works
Proposed Method
Input Fundus Images
Hand Crafted Feature Extraction
Machine Learning for Classification
Good
Poor
Machine Learning based methods
Input Fundus Images
CNN based Feature Extraction
Classification
Good
Poor
Deep Learning based methods
Input Fundus Images
Block-1
Optimized feature extraction from four pre-trained models using subjective scores
Feature fusion
Block-2
Final Classification
Good
Fair
Poor

Figure 4.1 : Comparison Flow Chart of the state of the art fundus IQA methods and the proposed method.

for the training phase. Training a network from scratch requires a sufficiently large number of images to get the optimal values for the network weights. Recently, transfer learning methods are also used to address the challenges of fundus image quality assessment [Zago *et al.*, 2018; Chalakkal *et al.*, 2019; Fu *et al.*, 2019b].

As an initial setting for the training, we have used the weights of the following four pre-trained models: ResNet [He *et al.*, 2016], DenseNet [Huang *et al.*, 2017], Inception-V3 [Szegedy *et al.*, 2016], and Xception [Chollet, 2017]. ResNet is a deep residual learning based CNN architecture proposed by He *et al.* [He *et al.*, 2016]. ResNet50, ResNet101, and ResNet152 are its variants, where 50, 101, and 152 indicate the number of layers present in the architecture, respectively. DenseNet{121, 169, 210} was proposed by Huang *et al.* [Huang *et al.*, 2017] in 2017. The "dense" term indicates that each layer of this CNN model is connected to every layer of the architecture. Here 121, 169, and 201 indicate the depth of the model. Next, Inception-v3 is a successor version of GoogLeNet that also named Inception-v1. Each inception layer is built with six convolution layers, followed by one pooling layer. Finally, the Xception architecture is a linear stack of depthwise separable convolution layers with residual connections [Chollet, 2017]. Each model is trained individually on the subjective scores of F1-F6, by adding *five* fully connected (FC) layers at the end of the each network. The details of the FC layers are as follows: FC1: 1024×1, FC2: 512×1, FC3: 120×1, FC4: 24×1, FC5: 12×1. Here the first four FC layers are followed by the rectified linear unit (ReLu) [Nair and Hinton, 2010] activation function. The mathematical representation of the ReLu is given below:

$$y = max(0, x). \tag{4.1}$$

It produces the output y as x if the value of input x is positive and 0 otherwise. The ReLu activation is used because of its advantage over sigmoid and hyperbolic tangent activation functions as it avoids the vanishing gradient problem. The last FC5 layer, with the inclusion of sigmoid function, performs

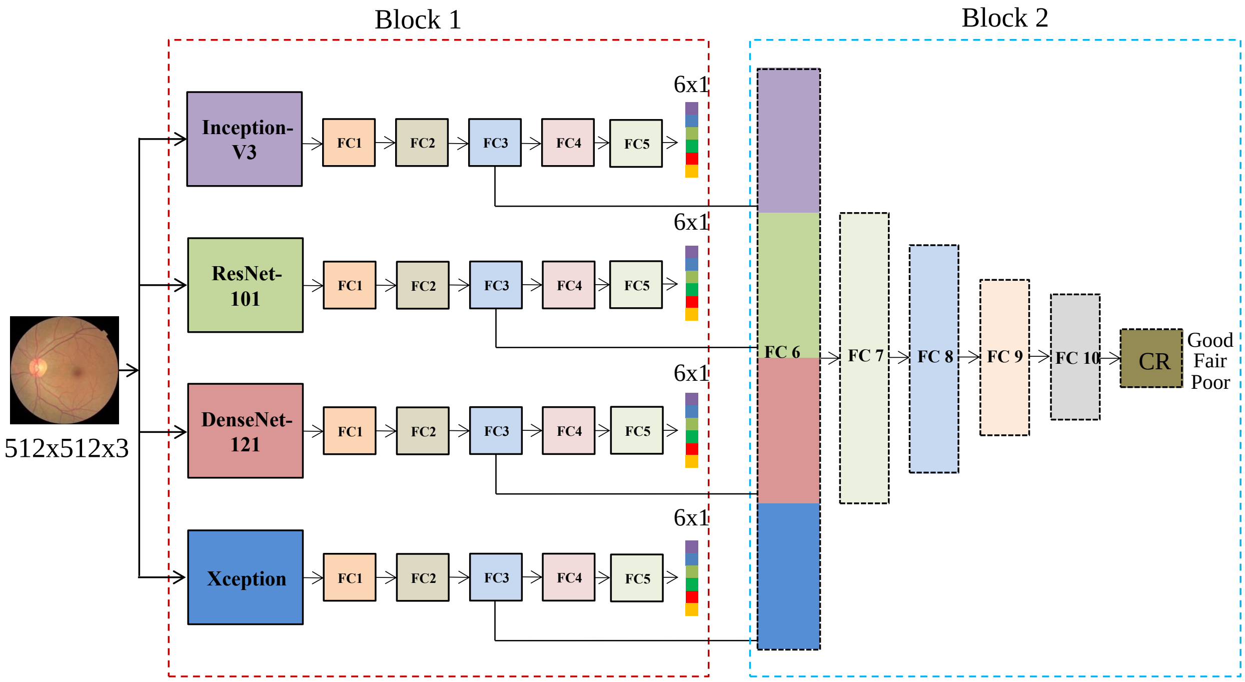


Figure 4.2 : Proposed Model. FC: Fully Connected Layer, FC1: 1024×1, FC2: 512×1, FC3: 120×1, FC4: 24×1, FC5: 12×1, FC6: 480×1, FC7: 120×1, FC8: 24×1, FC9: 12×1, FC10: 6×1, CR: Classification Result.

multivariate regression to derive the six numerical values corresponding to the F1-F6 quality parameters.

Fig. 4.2 shows that the CNN model takes the input image of size $512 \times 512 \times 3$ and in the fifth FC layer transforms it into a feature vector of size 12×1. In the last FC layer the model performs the multivariate linear regression onto the desired feature vector of size 6×1. Let $X_{(i)12\times 1}$ be the input feature vector obtained at the fourth FC layer and $Y_{(i)6\times 1}$ is the associated score vector for the i^{th} image. Then, the multivariate linear regression model can be represented as:

$$\hat{Y}_i = W_i X_i + E_i \tag{4.2}$$

where

- $\hat{Y}_i = [\hat{y}_{i1}, \hat{y}_{i2}, \hat{y}_{i3}, \hat{y}_{i4}, \hat{y}_{i5}, \hat{y}_{i6}]$ is the 6×1 predicted score vector for the i^{th} image.
- $X_i = [x_{i1}, x_{i2}, x_{i3}...., x_{i12}]$ is the 12×1 input feature vector for i^{th} image.
- $W_i = [W_{i1}, W_{i2}, W_{i3}...., W_{i6}]$ is the 6×12 weight matrix for the i^{th} image.
- $W_{ij} = [w_{ij_1}, w_{ij_2}, w_{ij_3}, ..., w_{ij_{12}}]$ is the 1×12 weight vector for j^{th} feature. Here, j = 1,2,3,...6.
- Finally, $E_i = [e_{i1}, e_{i2}, e_{i3}, e_{i4}, e_{i5}, e_{i6}]$ is the corresponding error matrix of size similar to Y.

It is important to mention that the batch normalization [Ioffe and Szegedy, 2015] method is used for the regularization of the model to avoid the over-fitting problem. Batch normalization is preferred over the dropout [Srivastava *et al.*, 2014] method as empirical results were better than in the case of batch normalization. All four models were trained to achieve the maximum correlation with the subjective scores of F1-F6. Furthermore, once each of the models was trained for the maximum correlation, the values of FC3 layers from each model were assembled and transferred to Block-2. The accuracy of the correlation results is discussed in Section 4.3.

Block-2: This block uses the concepts of both transfer learning and ensemble learning. The objective of ensemble learning is to collect the predictions from different models to conclude with better prediction results [Liu and Yao, 1999]. The optimized features of the FC3 (120×1) layer from each of the four models of Block-1 are combined to form the FC6: 480×1 layer and transferred to Block-2. Block-2 consists of 5 fully connected layers: FC6: 480×1, FC7: 120×1, FC8: 24×1, FC9: 12×1, FC10: 6×1 and finally the classification results. It is important to mention that the training of each block presented here is done individually. Block-1 was trained until the optimized features were derived. Afterwards, Block-2 was trained to get the optimized classification results. Similar to the previous block, the ReLu activation function follows each FC layer in Block-2 after the FC10 layer *softmax* function is applied to get the desired classification results.

4.2 IMPLEMENTATION DETAILS

- *Pre-processing:* Fundus images carry a large area of black background that might affect the training accuracy. Therefore, all the images were cropped to the boundary of the fundus area in order to reduce the area of black background. It is achieved by traversing the nearest pixel values that are close to zero to the center co-ordinates of the images. In addition, the fundus images provided on Kaggle are of high resolution. Hence, each image is further resized to the dimension of 512×512.

- *Loss Function:* In Block-1, the mean square error (MSE) function is used as the loss function, and can be represented as:

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N} ||(Y - \hat{Y})||^2 \tag{4.3}$$

  where L_{MSE} represents the loss computed for the Block-1, Y and $\hat{Y}$ represent the actual value and predicted value respectively, and N represents the number of samples. Moreover, in Block-2 the categorical cross entropy loss function is used. Its mathematical representation is as follows:

$$L_{CCE} = -\sum_{i=1}^{C} P_i log(\hat{P}_i) \tag{4.4}$$

  Here, L_{CCE} represents the loss computed for the Block-2, C represents the total number of classes, P and $\hat{P}$ represent the actual and predicted output respectively. It is important to mention that the *softmax* activation function should be applied to the target before computing the categorical loss.

- The back-propagation and adaptive moment estimation (ADAM) [Kingma and Ba, 2014] optimization methods are used for error minimization with learning rate of 10^{-4}. ADAM has been performed for *1000* epochs with the mentioned batch size of *8* images during the training process.

- Out of 1500 images, 1200 were used for the training and 300 for testing purpose. Here, 400 images were taken from each class for training and similarly 100 images from each category for testing.

- All the experiments were carried out on a computer system of 2.0 GHz CPU and GTX-1080 Ti GPU and the proposed model is implemented using the Python programming language with Keras library.

Table 4.1 : Correlation coefficients for the predicted values of F1-F6

	Feature	SROCC	PLCC	KCC	RMSE
Inception-V3	F1	0.9299	0.9544	0.8463	0.4163
	F2	0.9309	0.9392	0.8319	0.4294
	F3	0.9005	0.9346	0.7932	0.4453
	F4	0.9281	0.9413	0.8281	0.4187
	F5	0.9413	0.9512	0.8517	0.4014
	F6	0.9388	0.9477	0.8454	0.4004
ResNet-101	F1	0.8758	0.8906	0.7260	0.7646
	F2	0.8920	0.8644	0.7336	0.7731
	F3	0.8378	0.8588	0.6724	0.8531
	F4	0.8815	0.8916	0.7224	0.6887
	F5	0.8981	0.9032	0.7478	0.5871
	F6	0.9025	0.9066	0.7517	0.7571
DenseNet-121	F1	0.9011	0.9008	0.7477	0.7176
	F2	0.8906	0.8981	0.7404	0.7362
	F3	0.8706	0.8782	0.7072	0.8116
	F4	0.9032	0.9053	0.7463	0.5966
	F5	0.9133	0.9148	0.7647	0.4947
	F6	0.9079	0.9113	0.7581	0.6835
Xception	F1	0.9293	0.9469	0.8532	0.4262
	F2	0.9230	0.9348	0.8369	0.4294
	F3	0.9007	0.9363	0.7898	0.4406
	F4	0.9225	0.9324	0.8225	0.4220
	F5	0.9326	0.9398	0.8482	0.4106
	F6	0.9269	0.9333	0.8402	0.4262

4.3 RESULTS AND ANALYSIS

4.3.1 Evaluation methodology

Four commonly used standard measures recommended by the Video Quality Experts Group [Rohaly *et al.*, 2000] have been used to evaluate the performance of Block-1. These are the Spearman rank-order correlation coefficient (SROCC), the Kendall rank-order correlation coefficient (KCC), the Pearson Linear correlation coefficient (PLCC), and the root-mean-square error (RMSE). For the performance measurement of an IQA metric, SROCC and KCC evaluate the prediction monotonicity. The other two, PLCC and RMSE, measure the prediction accuracy. Higher values obtained in SROCC, KCC, and PLCC for an IQA metric indicate higher performance, whereas lower values of RMSE are associated with better performance. Furthermore, to evaluate the performance of Block-2 the following statistical parameters are used:

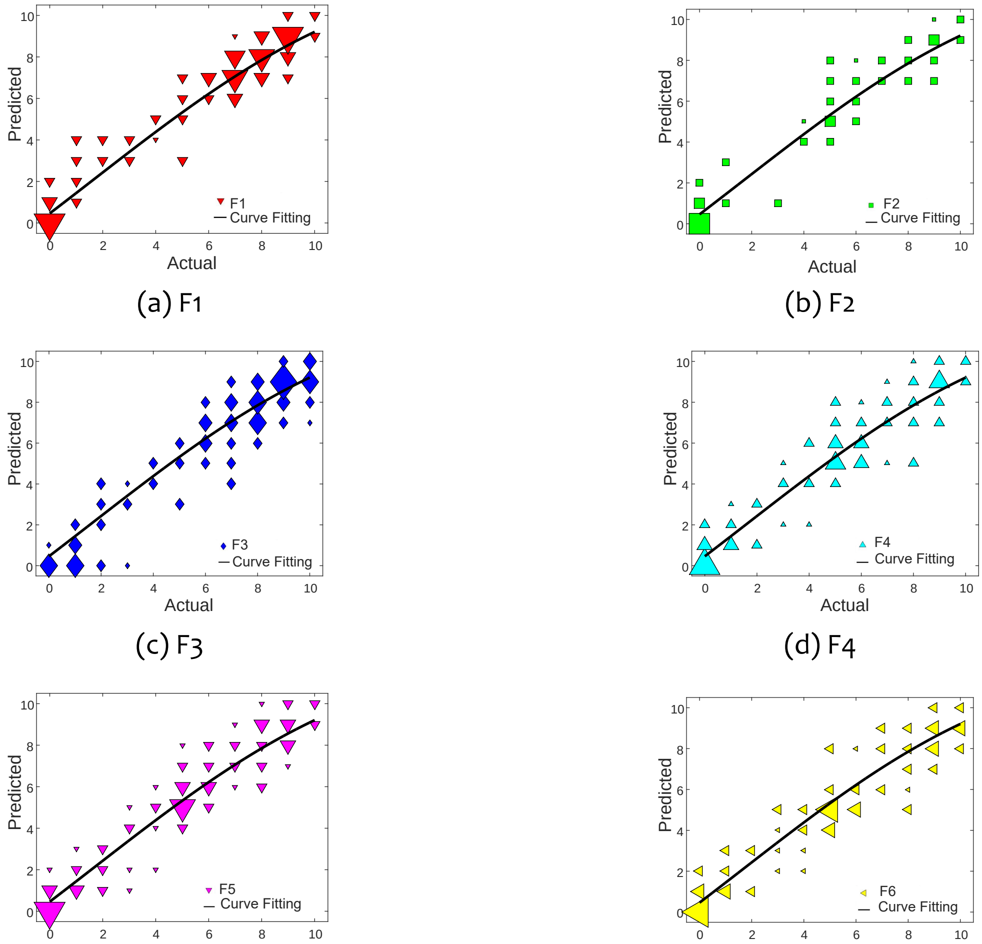
$$A = \frac{T}{N} * 100 \tag{4.5}$$

$$P = \frac{T_p}{T_p + F_p} \tag{4.6}$$

$$R = \frac{T_p}{T_p + F_n} \tag{4.7}$$

$$F_m = 2 * \left(\frac{PR}{P+R} \right). \tag{4.8}$$

Here A = Classification accuracy, P = Precision, R = Recall, T = Total number of correct classifications, N = Total number of samples, T_p = true positive, F_p = false positives, F_n = false negatives, and F_m = F-measure.



(a) F1 (b) F2

(c) F3 (d) F4

(e) F5 (f) F6

Figure 4.3 : Feature-wise plot of the predicted scores versus actual opinion scores.

4.3.2 Performance evaluation of Block-1

The feature-wise performance of each of the four models is shown in Table 4.1, reporting the correlation values calculated between the derived scores and the subjective score values for each quality parameter F1-F6. In addition, Table 4.1 shows that the highest results obtained for the SRCC, PLCC, and KCC are *0.94*, *0.95*, and *0.85* respectively and for RMSE the lowest result is *0.40*. It validates that the proposed model achieves a significantly high correlation between the subjective and derived scores. Furthermore, scatter plots with curve fitting of the mean of predicted values from each of the four models are shown in Fig. 4.3. These plots are obtained after performing logistic regression between predicted values and subjective OS values. These curves are obtained after non-linear fitting,

Table 4.2 : Performance evaluation of different models for classification results on FIQuA data-set..

Model	Accuracy (in%)	F-measure	Precision	Recall
ResNet50 [He *et al.*, 2016]	90.33	0.9032	0.9031	0.9033
ResNet101 [He *et al.*, 2016]	91.66	0.9165	0.9164	0.9166
ResNet152 [He *et al.*, 2016]	90.33	0.9032	0.9031	0.9033
DenseNet121 [Huang *et al.*, 2017]	92.33	0.9233	0.9233	0.9233
DenseNet169 [Huang *et al.*, 2017]	89.66	0.8963	0.8960	0.8966
DenseNet201 [Huang *et al.*, 2017]	90.66	0.9065	0.9064	0.9066
Inception-V3 [Szegedy *et al.*, 2016]	93.00	0.9300	0.9300	0.9300
Xception [Chollet, 2017]	93.33	0.9334	0.9335	0.9333
Proposed	**95.66**	**0.9566**	**0.9565**	**0.9566**

as suggested in [Larson and Chandler, 2008]. It can be observed from Fig. 4.3 that the consistency between the predicted and subjective values is very high. Here, the size of the object represents the frequency of the predicted values corresponding to the actual value, whereas larger size objects correspond to a higher frequency. It can also be observed that all the larger size objects lie in the close vicinity of the curve, indicating a high correlation between actual and predicted values. These high correlation results validate that the features obtained in previous FC layers are optimized. Now, the optimized features of FC-3 are ensembled together and transferred to Block-2 for the final classification of images.

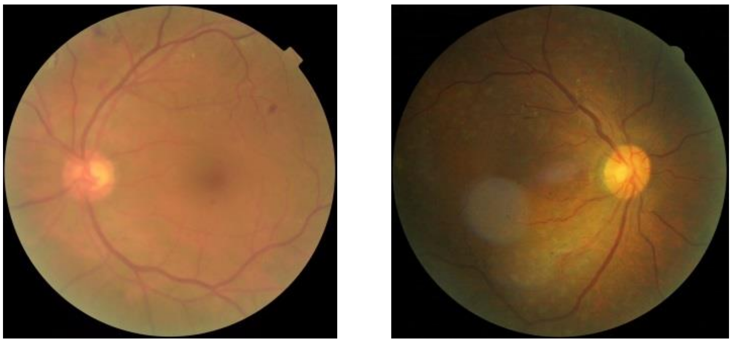
4.3.3 Performance evaluation of Block-2

Initially, the individual classification performance of different variants of each of the four models has been analyzed. Here, individual performance indicates that the 240×1 feature vector derived from Block-1 is used only to train Block-2 for final classification. Table 4.2 contains the performance results of Block-2 with three variants of both ResNet and DenseNet. It indicates that the Xception model achieves the highest individual accuracy (93.33%). However, the performance of the proposed ensemble model after the fusion of features got approximately 2% jump on overall accuracy with 95.66%. The confusion matrix of the prediction results of the proposed method is shown in Fig. 4.4. It can be observed from Fig. 3.4 (shown in the Chapter 3 page no. 30) and Fig. 4.4 that the accuracy of the proposed fundus IQA model is closely similar to the results of the classification using subjective scores. It indicates that the inclusion of subjective scores greatly helps to train the model to derive the optimized features for the classification. Also, for illustration purposes, two example images from the Fair category of the FIQuA data-set are shown in Fig. 4.5. Here, (a) and (b) are the sample images distorted with blur and uneven illumination distortions, respectively. It can be observed from the Fig. 4.5 that all the structural information is quite visible, yet due to the presence of a small proportion of distortions, ophthalmologists labeled them as a fair quality image. The proposed model also correctly classified these images as fair quality. It indicates the robustness of the model as it efficiently mimics the visual perception of ophthalmologists by detecting these distortions in the image.

Predicted Class

	Class	Good	Fair	Poor
Actual Class	Good	96	4	0
	Fair	4	93	3
	Poor	0	2	98

Figure 4.4 : Confusion matrix of the prediction results obtained on FIQuA data-set from the proposed fundus IQA model.



(a) (b)

Figure 4.5 : Sample images with different distortions from the Fair category of the FIQuA data-set that are correctly classified by the proposed model. Here (a) and (b) represent the images distorted with Blur and Uneven Illumination distortion, respectively.

4.3.4 Cross Data-set evaluation

The proposed fundus IQA model trained over the FIQuA data-set was also evaluated over two publicly available data-sets: DRIMDB [Sevik *et al.*, 2014] and EyeQ [Fu *et al.*, 2019b], specifically developed for fundus IQA. The DRIMDB [Sevik *et al.*, 2014] data-set was presented by U. Sevik. It contains 216 fundus images with three classes: Good (125), Poor (69), and Outlier (22). Next, Fu, *et al.* made a commendable effort and recently presented a large scale EyeQ data-set. The EyeQ data-set consists of 28,792 fundus images divided (with analogy to our approach) into three categories: Good, Usable, and Reject. Table 4.3 contains the classification results over the above mentioned data-sets. The results indicate that the proposed fundus IQA model achieves high classification accuracy over an unknown and large scale data-set given it was trained on a comparatively small data-set. Also, for comparison purposes, a performance summary of recent fundus IQA works that are developed and evaluated over DRIMDB and EyeQ data-set is presented in Table 4.4. It can be observed from both Table 4.3 and 4.4 that the performance of the proposed model outperforms the recent fundus IQA methods over the mentioned data-sets. It is essential to mention that despite being trained over a comparatively too small data-set (FIQuA), the performance of the proposed model is very close to the model proposed in [Fu *et al.*, 2019b]. It shows that the inclusion of adequate subjective inputs not only increases the performance of the model but also its generalizability over unknown image inputs. In our future work, we are planning to use reinforcement learning methods to achieve higher accuracy over the EyeQ data-set.

Table 4.3 : Performance evaluation of proposed method over DRIMDB and Eye-Quality (EyeQ) data-set

Data-set	Accuracy (in %)	Precision	Recall	F-measure
DRIMDB	98.96	0.9889	0.9859	0.9920
EyeQ	88.43	0.8697	0.8700	0.8694

Table 4.4 : Performance summary of recent fundus IQA works over DRIMDB and EyeQ data-set. Here (+) indicates that the work also includes fundus images from other proprietary data-sets.

Work	Year	Data-set	No. of Images	Accuracy (in %)
Wang *et al.* [2016]	2016	DRIMDB (+)	536	94.52
Shao *et al.* [2018]	2018	DRIMDB (+)	4372	92.39
Zago *et al.* [2018]	2018	DRIMDB (+)	1036	98.56
Chalakkal *et al.* [2019]	2019	DRIMDB (+)	7007	97.70
Fu *et al.* [2019b]	2019	EyeQ	28792	91.75

4.4 SUMMARY

- a new multivariate linear regression based neural network model for fundus image quality assessment is presented. The peculiarity of the model is that it derives the optimized features for classification, using the subjective inputs provided by the ophthalmologists.
- It consists of two blocks: Block-1 derives the optimized features from four pre-trained CNN models: Inception-V3, ResNet-151, DenseNet-121, and Xception that are trained through transfer learning against the six subjective scores provided by the ophthalmologists.
- Further, these optimized features are ensembled together and forwarded to Block-2 to classify the fundus images into three classes: Good, Fair, and Poor.
- The results show that the proposed model achieves a high correlation with subjective values. The correlation values obtained from Block-1 for SROCC, LCC, and KCC for each quality parameter (F1-F6) are approximately *0.941*, *0.954*, and *0.853* respectively, and for RMSE the result is *0.401*. It indicates that for each of the six features, the derived quality scores from the proposed model are closely similar to the subjective quality scores provided by the medical doctors.
- Further, using the derived ensembled features, the classification accuracy achieved by the Block-2 is 95.66%. It proves that the inclusion of the subjective scores helps achieving a high classification accuracy.

As mentioned in previous chapters, that the “Fair” category of fundus image quality provides two major advantages: (i) reducing the number of miss classifications between the good and poor categories, and (ii) it also indicates the need and requirement of enhancement in the image. In view of the second advantage, our next work is based on fundus image enhancement. The next chapter includes a detailed description of our proposed fundus image enhancement method namely: Residual Dense Connection (RDC) based UNet Model.

...