Transfer Learning-based One Versus Rest Classifier for Multiclass Multi-Label Ophthalmological Disease Prediction

Akanksha Bali, Vibhakar Mansotra
Department of Computer Science and IT
University of Jammu
Jammu, India

***Abstract*—The main objective of this paper is to propose transfer learning technique for multiclass multilabel opthalmological diseases prediction in fundus images by using the one versus rest strategy. The proposed transfer learning-based techniques to detect eight categories (seven diseases and one normal class) are Normal, Diabetic retinopathy, Cataract, Glaucoma, Age-related macular degeneration, Myopia, Hypertension and Other abnormalities in fundus images collected and augmented from Ocular Disease Intelligent Recognition (ODIR) dataset. To increase the data set, no differentiation between left and right eye images has been done and these images were used on VGG-16 CNN network to binary classify each disease separately and trained 8 separate models using one versus rest strategy to identify these 7 diseases plus normal eyes. In this paper, various results has been showcased such as accuracy of each organ and accuracy of the overall model compared to benchmark papers. Base line accuracy have increased from 89% to almost 91% and also proposed model has improved the performance of identifying disease drastically prediction of glaucoma has increased from 54% to 91%, normal images prediction has increased from 40% to 85.28% and other diseases prediction has increased from 44% to 88%. Out of 8 categories prediction, proposed model prediction rate has improved in 6 diseases by using proposed transfer learning technique vgg16 and eight different one versus classifier classification algorithms.**

***Keywords*—Fundus images; one versus rest strategy; transfer learning; VGG-16; augmentation**

I. Introduction

The motivation of this article comes from the fact that According to WHO at present, there are 2.2 billion people around the globe suffering with visual disability out of 2.2 billion at least a billion people could have been treated from visual impairment [1]. In the 21st century, eye blindness became normal because of high exposure towards electronic gadgets such as television, laptops etc. from early ages, though most of the eye diseases could be cured if detected in starting stages of the disease. Eyes are organs of the visual system, they capture light rays and regulate intensity using diaphragm and form an image using a lens. Eyes forward the captured image to the brain via optic nerve by converting them to electro-chemical impulses, any disturbances in the above process creates visual impairment or eye disorder. The study of eye diseases and disorders to diagnose is called Ophthalmology.

The primary sources for the cause of eye blindness are 1.exposure towards electronic gadgets 2.Lack of accessibility for medical facilities especially in developing and undeveloped countries 3. People in rural areas have a higher rate compared to their counterparts living in city 4. Aging people [2] and Indigenous people (tribes) 5. Accidents like facial fractures [3] e.t.c. The most common eye diseases occurring in day to day life are due to Diabetic Retinopathy, Glaucoma, Cataract, AMD, Hypertension, and Myopia. Some of these won't result in vision impairment but cannot be neglected from detection and treatment.

Diabetic retinopathy, a common disease for blindness in the age group of 20 to 70, initial days of this disease, diabetes patients of type -1 and 60 percent of type-2 diabetes suffered from retinopathy [4]. Glaucoma is a type of eye diseases having a common feature in cupping and atrophy of optic nerve head, visual field loss; often increase in intraocular pressure [5]. In 1990, 37 million people were estimated to be blind and 40 percent of them suffered from cataract [6] and it can be corrected with surgery but lack of facilities for treating cataract is a rising concern in developing countries and undeveloped countries.

Age related macular and degeneration (AMD) [7] is a natural thing but the numbers of these cases are increasing day by day due to the sharp rise in mortality rate because of development of medical facilities [8] and stable governments. Hypertension affects the retina significantly and the study of retaining provides valuable information to treat hypertension [9]. Myopia (short sightedness) considered as disorder, it can be corrected with glasses, contact lens and surgery (Lasik treatment), though it is less threatening but number of people suffering with this disorder has taken a step curve especially in children [10].

The state of the art solutions for retinopathy based on the classifications of these diseases are tedious tasks with the advancement of computers and computing techniques; various methods are proposed for the classification of objects. The most common types of classification are 1. Single labelled classification generates yes or no situations [11], such as person is suffering from eye disease enough to understand the

person is suffering but not to understand the reasons for suffering. 2. Multi labelled classification though it is computationally expensive, but it provides better intuitions and further depth analysis and it identifies more objects.

Machine learning algorithms have significantly done well in the field of image classifications, segmentation and enhancement techniques. Machine learning algorithms have generated better results initially but failed to learn more features with the increase of the dataset. Neural networks, a part of machine learning algorithms, have done significantly well with the data and deep learning algorithms that came into picture generated more features with increase of data. Deep learning algorithms became a common norm for any image classification, segmentation tasks.

The two common ways to represent eyes in the form of images are Fundus Photography, capturing images at the fundus. The main areas covered at fundus photography are the central and peripheral retina, optic disc and macula. The second way is Optical Coherence tomography (OCT), a technique to capture 2d, 3d and micrometer resolution images with low coherence light.

The paper findings as discussed in the upcoming section:

1) This paper focuses on Fundus images to classifyNormal, Diabetic Retinopathy, Glaucoma, Cataract, AMD, Hypertension, Myopia and other diseases using multi labelled dataset ODIR 2019 consisting of 5000 fundus images of patients of both eyes.

2) The primary reason to work on these datasets is its multi-disease data sets as most of the datasets encountered are mainly focused on one ophthalmic disease.

3) The paper carefully analyzes the machine learning, data variance and generative n probabilistic aspects of retinopathy and tries to devise a cross entropy modelling system for multi label classification.

4) It also formulates the utility of transfer learning, into the diverse and difficult field of retinopathy.

The rest of the paper, structured as: section 2, literature review briefly describes various research proposed in eye organ segmentation using both machine learning and deep learning algorithms, section 3 explains the methodology part which comprises subsection 1) data collection. 2) preprocessing and augmentation 3) Training, Section 4 explains the results and analysis which comprise subsection 1) performance metrics, section 5 describes the discussion and section 6 concludes the paper.

II. Literature Review

Deep learning and Machine Learning became solutions for computer vision problems such as image enhancement, segmentation and classification, especially in biomedical imaging. Many researchers have proposed various approaches for the classification of ophthalmological diseases.

J. Liu et al. [12], proposed an SVM based classification approach to classify myopia with an accuracy of 87% on test data from the Singapore Eye Research Institute. T.V. Phan et al. [13], proposed an SVM and random forest based method to classify AMD. V. Gulshan et al. [14], uses a deep convolutional neural network to identify diabetic retinopathy and DME in fundus images. H. Pratt et al. [15], proposed a CNN and data augmentation technique for the classification of micro-aneurysms, exudate and haemorrhages on the retina. The proposed architecture achieves sensitivity and accuracy around 95% and 75% on 5000 validation images.

J. Y. Choi et al. [16], used matConvNet for automatic detection of multiple retinal diseases on the STARE database, consisting of nine eye diseases. Optimal results were obtained by random forest transfer learning based VGG19. P. M. Burlina et al. [17], proposed a DCNN for the classification of AMD, this model compared with a pre trained DCNN by performing transfer learning. Y. Chai et al. [18], proposed a method to combine deep learning models with domain knowledge for automatic glaucoma detection on fundus images. The proposed model outperformed AlexNet, VGG16, InceptionV3 in accuracy, sensitivity and specificity. F. Grassmann et al. [19], utilized various convolutional neural networks to classify nine types of eye disease due to age, three types of AMD and one ungradable image, to classify these thirteen classes, ensembling has been done over six different neural network architectures.

M. N. Bajwa et al. [25], proposed a framework containing two stages, the first stage uses a CNN to localize and extract the optical disc from the retinal fundus image and the other one uses deep convolutional neural network for classifying disc extracted in the first stage. Due to the lack of original ground truth images, they proposed rules for generation of semi automatic ground truth images. They achieved a 2.7% improvement to the previously produced results on the ORIGA dataset. S. Keel et al. [26],proposed Inception V3 architecture for classification and severity possibility threshold on neovascular age-related macular degeneration. V. Das et al. [27], proposed CNN based classification detection techniques for DME and AMD up to two stages. The evaluation of the proposed method is performed on the OCT dataset and achieves a decent score of sensitivity, specificity and accuracy around 99.6%, 99.87% and 99.6% on test data. T. Li et al. [28], provided a new dataset of 13673 fundus images from 9598 patients for diabetic retinopathy and these images were classified into 6 types based on quality and DR level.

Y. Peng et al. [29], proposed DeepSeeNet architecture to measure the score of severity in the range 0 to 5 for the age related eye diseases study. T. Pratap et al. [30], used a pre trained CNN architecture for transfer learning to extract features for classifying levels of cataract and these features were classified using SVM. In this paper [31], M. S. Alabshihy et al. used direct technique such as problem transformation, multilabel cad system, segmentation, MSVM on dataset named as DiaretDB having two classes named as DR and hypertension and achieved an overall accuracy of 96.1%. Md. T. Islam et al. [32], proposed a classification model for eight ocular diseases using contrast limited adoptive histogram equalization as a pre-procesing step and CNN has used for feature extraction.

T. Nazir et al. [33], proposed a deep learning approach for segmentation of diabetic retinopathy, diabetic macular edema

(DME) and glaucoma using a fast region based convolutional neural network to localize and fuzzy k means to segment. A multi task loss was used as a loss for CNN. Intersection over union, mean average precision and dice coefficient as evaluation metrics and achieved mean average precision of 0.94. X. Pan et al. [34], compared DenseNet, Resnet50 and VGG16 to automatic classification and detection the four kinds of lesions of diabetic retinopathy such as non-perfusion regions, microaneurysms, leakages, and laser scars in fundus fluorescein angiography images. Sensitivity, specificity and region of curve were used as evaluation metrics. M. Aamir et al. [35], proposed a two phased CNN based architecture for classification, one for glaucoma detection and other for rating glaucoma in different scales like advanced, moderate, early, on fundus images, and adaptive thresholding was done before applying CNN. Sensitivity, specificity, accuracy and precision were used as metrics for evaluation. C. G. Gonzalo et al. [36], proposed a CNN ensembling methods to identify AMD and diabetic retinopathy in color fundus images. Inputs for CNN are contrast enhanced image and RGB image derived from original color fundus images.

R. Sarki et al. [37], proposed CNN based architecture for multi classification of diabetic eye disease in two ways, a low level multi class diabetic eye disease and another one is a high level multi class eye diabetic disease. Maximum accuracy for mild multi-classification and multi-classification are 88.3% and 85.95% using VGG16. K. Shankar et al. [38], proposed a synergic deep learning model for classifying the levels of diabetic retinopathy. The proposed method outperforms AlexNet, ResNet, GoogleNet and VggNet-19 with respect to accuracy, sensitivity and specificity. A. Ram et al. [39], used a CNN for feature extraction for classifying normal, cataract, myopia and AMD. The objective of this paper is to correlate the relationship between the number of classes and number of fully connected layers.

J. Wang et al. [40] used feature extraction based efficientnet in the first part and custom neural network in the second part for multilabel classification of fundus images. N. Gour et al. [41], used transfer learning for classification on fundus images by two approaches. In the first approach, images of both eyes were individually given as input for CNN and the results were concatenated and in the second method, images of both eyes were concatenated and given as input to CNN. Various state level architectures have used instead of CNN to generate better results and VGG16 pretrained architecture performed significantly. N. Li et al. [42], created a database of 10,000 fundus images of both eyes from 5000 patients to classify 8 diseases and multi level classification of images has improved significantly with the increase of complexity in state of art deep neural networks like AlexNet, ResNet, GoogleNet. J. He et al. [43], proposed a dense correlation network (DCN) for classifying multi labelled diseases. DCN consists of three modules for features extraction, features correlation and calculating classification score; a multi label soft margin loss was used as a loss function and produced way better results than benchmark deep neural networks.

In this paper [45], D. Muller et al. suggested ensemble heterogenous DL models for multi eye disease prediction. They also used fivefold cross validation on RFMID multilabel data containing 3200 images (1920 training data, 640 testing data and validation data each). This dataset contains twenty nine multilabel classes. The techniques used are data augmentation, bagging and stacking, transfer learning and stacked logistic regression. They achieved 0.95 AUROC for multilabel disease risk prediction by using ensemble DL models. In this paper [46], A. C. Garcia et al. used Resnet, Resnest, EfficientNet, ViT, Deit, NasNet, HRnet, CycleGAN on RFMID-2021 mainly focussed on the ERM category and achieved f1score of 86.82. In this paper [47], L. P. Cen et al. used DCNN on 249,620 images and 275,543 labels collected from different sources and achieved f1score, sensitivity, specificity, AUC of 0.923, 0.978, 0.996, 0.9984 resp. for multi-label classification dataset (Table I).

TABLE I. LITERATURE SURVEY OF RECENT STUDIES WITH SHORTCOMINGS

Reference No.	Author and year	Techniques used	Results	Limitations
[20]	M. Mateen et al., 2018	Gaussian mixture model (GMM), visual geometry group network (VGGNet), singular value decomposition (SVD) and principle component analysis (PCA)	Accuracy = 98.34	Does not consider the overfitting problem.
[21]	Q. Meng et al., 2019	Deep CNN with attention map mechanism	Accuracy = 94.5	Accurate classification accuracy can be checked on more data.
[22]	H. Chen et al., 2019	Deep hierarchical multi-label classification	AUC = 88.7	Results need to be evaluated using other performance parameters also.
[23]	L. Faes et al., 2019	Automated deep learning model	Sensitivity = 73·3–97·0 Specificity = 67–100% AUPRC = 0·87–1·00	Study should compare several state-of-art models with proposed one.
[24]	C. C. Jordi et al., 2019	VGG16 and InceptionV3	AUC = 88.71 F1-score = 88.76	Results can be improved with additional pre-processing steps.
[32]	Md. T. Islam et al., 2019	Shallow CNN architecture	F1-score = 85 Kappa score = 31 AUC = 80.5	Recent neural models can be implemented to evaluate results.
[44]	E. S. Kumar et al., 2021	Multi-Disease Classification Framework (MDCF) using stacking	AUC = 97.42 F1-score = 94.32	Different ensemble methods can be used.

A. Limitation in existing Architectures

The current state of the art methodologies has primarily 5 main limitations to highlight, as is discussed in the literature review section.

- Lack of a complete autonomous system to provide multi label classification at a medical acceptable rate. Majority of the research is done using one or multiclass eye disease. Applicability and deployment of deep learning techniques in classifying multilabel data is still in infancy stage.
- Absence of transfer running bio fueled ventures in retinopathy.
- The absence of high-volume dataset, diminishing the choice of deep learning.
- Use of traditional Machine learning techniques to measure the correlation of dimensional simulation, although feature extraction is difficult and eluding.
- Multiclass labelling pertains to quasi dimensional binary loss problem, general in multiclass labelling but deadly for retinopathy. To overcome this the paper hypotheses the use of cross-entropy modelling strategies.

III. Proposed Methodology

In this section, a pipelined architecture has been proposed based on deep learning using transfer learning techniques from Imagenet dataset to multi labelled classification of eye diseases. ODIR Dataset contains supervised data of 8 eye categories. They are Normal (N), Glaucoma (G), Diabetic Retinopathy (D), AMD (A), Hypertension (H), Cataract(C), Myopia (M) and other abnormalities (O) on fundus images as shown in Fig. 1. In this section, proposed architecture and database used has been explained.

A. Data Collection

The paper used Ocular Disease Intelligent Recognition (ODIR) dataset consisting 5000 images of ophthalmic patients of left and right eye, age and diagnostic key words of Doctor collected by Shanggong Medical Technology Co., Ltd. from different hospitals in China to classify diabetic retinopathy, glaucoma, cataract, age-related macular degeneration, hypertension, pathological myopia and other abnormalities. These multi labelled fundus images are captured by various cameras to create different resolutions. The Table II describes the count of each disease in the dataset and the augmentation details.

B. Pre-Processing and Augmentation

Images are cropped towards the centre to avoid the area which does not generate much information and various augmentation techniques were applied on the data sets labelled as Hypertension (H), Glaucoma (G), Cataract (C), AMD (A), Myopia (M) to create the balance among the datasets as these 5 diseases are largely outnumbered by other diseases. The augmented data set has increased from 7473 images to 14072 images as shown in Table II.

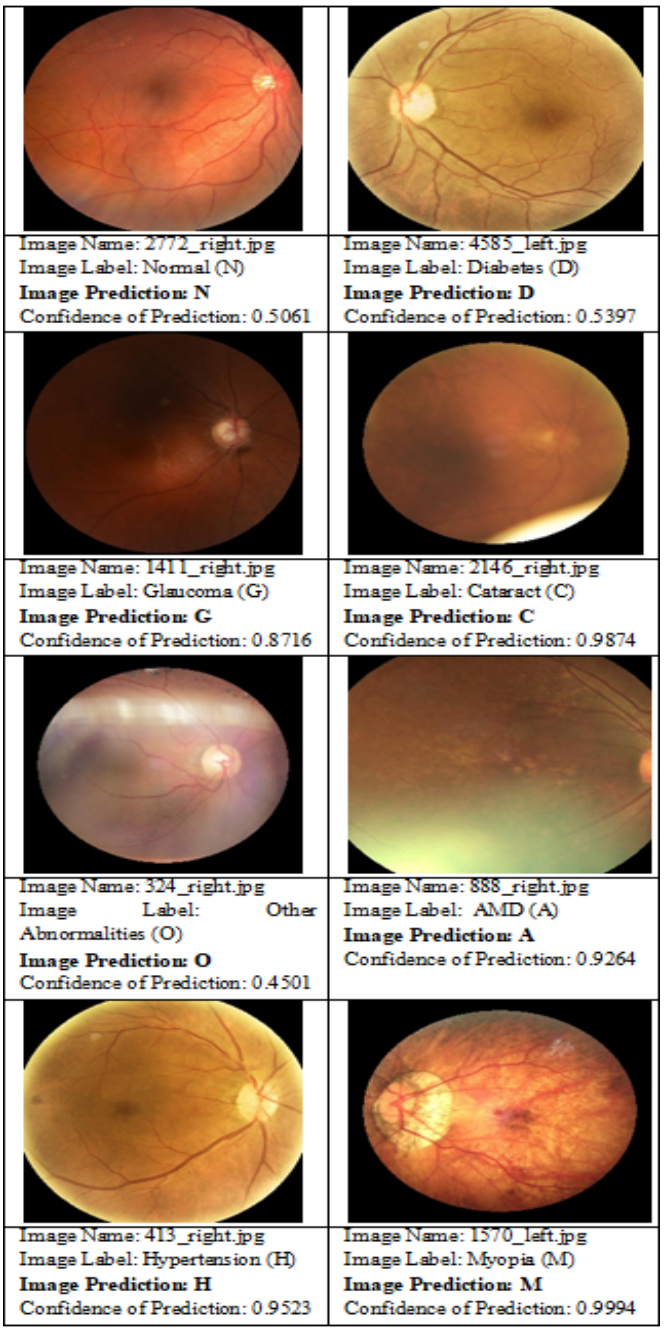

Fig. 1. Sample Images of 8 Eye Categories and their Confidence Levels of Prediction.

Each image of these 5 diseases were used to generate 8 more images using these 8 augmentation techniques, they are 1) Vertical flip, 2) Horizontal flip, 3) Both horizontal and vertical flip, 4) Clipped center of image and zoom of the original image, 5) Clipped center of image and zoom of the vertical flipped image, 6) Image rotation plus brightness enhancement, 7) Image rotation of the original image and 8) Image rotation of the vertical flipped image. Since it is a multi labelled dataset, augmentation of image containing two diseases creates 16 images in this augmentation approach, though no diabetic retinopathy images are augmented, its images are augmented due to multi labelled dataset.

TABLE II. TABULAR DESCRIPTION OF DATA SAMPLES AND AUGMENTATION DETAILS OF EACH DISEASE

Class	Total Samp-les	Training Split (70%)	Testing Split (30%)	Augmented	Augmented Sample (on training dataset)	Total Sampl-es (used for training)
N	3098	2169	929	No	0	2169
D	1801	1261	540	No	792	2053
G	326	229	97	Yes	1832	2061
C	313	218	95	Yes	1744	1962
A	277	193	84	Yes	1544	1737
H	193	135	58	Yes	1080	1215
M	268	188	80	Yes	1504	1692
O	1197	847	350	No	336	1183
Total	7473	5240	2233	Yes	8832	14072

C. Training

In proposed method, one versus rest strategy has used to classify each disease against all other 8 possibilities. The workflow of proposed methodology is shown in Fig. 2 and Fig. 3. Fig. 2 and Fig. 3 represent the flowchart and algorithm of our proposed methodology respectively. Eight different one versus rest classifier classification algorithms based on vgg16 were trained to detect N, D, G, C, A, H, M, O categories. Each eye image was taken separately (didn't consider left and right images of eyes as the same image) to double the size of the data set and implemented preprocessing and various augmentation methods to increase the data size to avoid over-fitting and to add versatility among the data. Images were cropped to the centre of fundus images and reshaped the size of image to 224×224 to avoid computational problems and these reshaped images are suitable to use VGG16 transferred weights.

1) Transfer learning: LeCun et al [48], proposed a convolutional neural network for extracting features in images, speeches and time series data. The basic layers in any convolutional neural network include convolution, pooling, batch normalization, fully connected layers as shown in Fig. 4. Various architectures are proposed by tweaking these layers by repeating more layers of one type or by changing the order of layers using different activation functions. The CNN's are great at localization and extracting features. CNNS are utilized in various fields like object detection, CNN's generated state of art results in segmentation, classification and enhancement on biomedical images and it requires a lot of data and computational power to perform matrix operations.

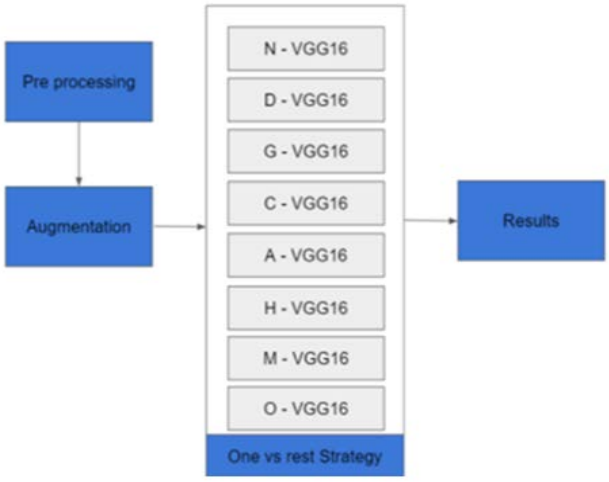

Fig. 2. Flowchart of Proposed Methodology.

For each batch on Train data do:

1. Resize image to 224 *224 Pixel.
2. Crop image to center.
3. If Image Labeled as G, C, H, A, M do:
   i. Vertical Flip.
   ii. Horizontal Flip.
   iii. Both Horizontal and Vertical Flip.
   iv. Clipped Center of image and zoom the original image.
   v. Clipped Center of image and zoom the vertical flipped image.
   vi. Image rotation plus Brightness enhancement.
   vii. Image rotation of original image.
   viii. Image rotation of vertical flipped image.
4. Else: No Augmentation.
5. For each label in data do:
   i. Train the VGG16 Model.
6. END

END

Fig. 3. Algorithm of Proposed Methodology.

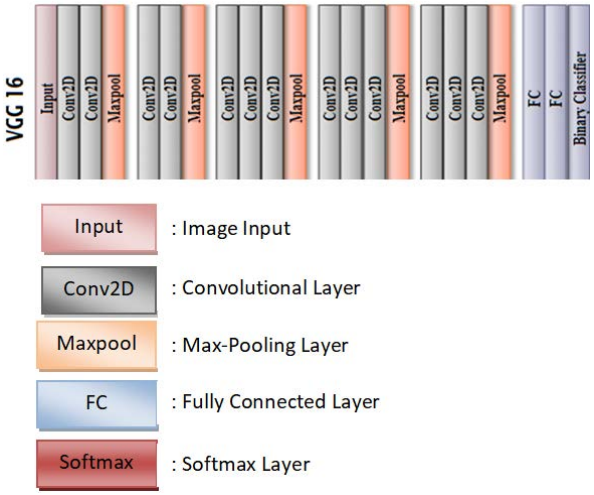

Fig. 4. Vgg16 Architecture. [48].

TABLE III. FILTERS IN CNN LAYERS

BLOCK ID	Number of layers	Number of CNN layers	Filters dimension	Number of Pooling layers
1	3	2	(3× 3 × 64)	1
2	3	2	(3× 3 × 128)	1
3	4	3	(3× 3 × 256)	1
4 & 5	4	3	(3× 3 × 512)	1

CNNs model weights are stored in open domains which are trained on large databases like ImageNet. Various CNN architectures have done well on ImageNet databases to detect and classify objects, the recent architectures are AlexNet[49], GoogleNet[50], VGGNet[51], MobileNet[52] and ResNet[53]. Collecting a large volume of multi labelled data in the medical field is a tedious task and very time consuming. Recently researchers moved to transfer learning where they use pre trained models on standard datasets and they use these pre-trained models on their datasets. There are various benefits in transfer learning like 1) it saves a lot of time for training the model and flexible enough to adjust trainable layers and non trainable layers. 2) It was trained on large dataset and has more parameters that are useful for learning, on a low sized datasets if we apply these standard models they mostly end up overfitting the data.

Proposed model uses VGG16 where its first 10 layers are not trainable and the rest are fine tuned and the last layer was adjusted to binary classification. The VGG16 contains 5 blocks, details about filters, number of convolutional layers and max pooling layers were given in Table III. In pooling layers kernel size is of 2×2 with a stride movement of 2 and in CNN layers ReLU used as activation function. Final parameters after block5 is 102764544 and these are flattened and forwarded to three sequentially fully connected neural networks. The last one uses softmax as activation for classification that outputs a vector of probability R as shown in equation (1) and the rest uses ReLU as activation function P that introduces nonlinearity to the network as shown in equation (2).

The softmax equation is defined by:

$$R = \begin{pmatrix} R1 \\ R2 \end{pmatrix} \text{ Where } R_i = \frac{e^{x_i}}{\sum_{k=1}^{n} e^{x_k}} \quad (1)$$

The Relu equation is defined by:

$$P(z) = \max(0, z) \quad (2)$$

Detailed implementation of these architectures is shown in Fig. 4, and these architectures were trained 8 times, as shown in Fig. 2 and Fig. 3 i.e. one versus rest strategy to classify 8 diseases individually and their results are amalgamated for multi diseases classification. The model in Fig. 4 was trained for 16 epochs and utilized a batch size of 32 to train this model with a validation split of 0.2. Stochastic gradient descent [54] as shown in equation (3) was used as an optimization algorithm and hyper parameters such as learning rate, momentum, decay are adjusted as 0.0001, 0.9, 0.000006 and Nesterov Accelerated gradient as shown in equation (4) uses parameters θ calculated from momentum term γw_{t-1} gives approximation term for next parameter value through $\theta - \gamma w_{t-1}$ that improves the generalisation performance has applied with stochastic gradient to increase the speed of convergence. Binary cross entropy as shown in equation (5) was used for loss function and accuracy used as the measure of metrics. To train this model, hardware specifications of cpu with 4 crores of ram size 32 gb and 11 gb GPU of Tesla K80 Architecture have utilized.

The SGD equation for each training example $t^{(i)}$ and label $u^{(i)}$

$$\theta = \theta - \eta . \nabla_\theta \boldsymbol{J}(\theta; t^{(i)}; u^{(i)}) \quad (3)$$

The nesterov accelerated gradient equation is

$$w_t = \gamma w_{t-1} + \eta \nabla_\theta \mathbf{J}(\theta - \gamma w_{t-1})$$
$$\theta = \theta - w_t \quad (4)$$

The Binary cross entropy function is given by

$$BCE = -\frac{1}{N} \sum_{i=0}^{N} y_i . \log(\acute{y}) + (1 - y_i) . \log(1 - \acute{y}) \quad (5)$$

IV. RESULT

A comparison table was provided on various techniques used on ODIR dataset along with proposed method to classify these diseases in Table IV. These proposed models were evaluated on 2233 testing images of ODIR database as mentioned in section III (Table II) and results have been compared on the basis of accuracy of each disease with proposed method and with the base paper proposed by N. Gour et al. [41] and these results were shown in Fig. 5, contains details of training loss, validation loss of each disease and overall accuracy and individual disease accuracy. Proposed model detected myopia with more accuracy as its false positive rate and false negative rate are very less compared to other diseases as a result it showed significant improvement in accuracy, f1-score, and precision and recall whereas classification error became quite negligible. Identifying normal images became a quite tricky and its results are mediocre compared to identifying other diseases.

A. Performance Metrics

Various metrics such as Accuracy, Specificity, Precision, Sensitivity, and Classification error, F1-score, False Positive Rate, Negative Predictive Value and False Negative Rate of each disease on testing data have been shown in Table V and the barcharts for each disease and for the overall model shown in Fig. 6 and Fig. 7. The equations for performance metrics is shown in the equations 6 to 14.

1) Accuracy: It is defined as the ratio of sum of true positives and true negatives to the total number of samples.

$$Accuracy = (\text{True Positives} + \text{True Negatives}) \div \text{Total Number of Samples} \quad (6)$$

TABLE IV. COMPARISON OF EXPERIMENTS CONDUCTED ON ODIR DATASETS AND THEIR RESULTS

Reference No.	Method used	Eyes merging	Results
[32]	Proposed cnn architecture	No fusion techniques were used	F1-score: 0.85, kappa score: 0.31, AUC value :0.85
[39]	Model based on two cases 1)featured based efficient net and 2) custom based neural network for multilabel classification	No fusion techniques were used	Overall Validation Accuracy: 0.90, F1-score: 0.85 (for image size 299*299) Overall Validation Accuracy:0.92,F1-score: 0.89 (for image size 448*448)
[40]	Transfer learning applied on two cases 1) transfer learning on concatenated left and right eye images 2) transfer learning individually on left images and right images and they were merged before classification	Uses concatenation as fusion technique	Overall Validation Accuracy 0.89
[41]	Transfer learning	Both left and right eyes are merged using sum, product and concatenation	The mean of kappa, AUC and F1 score. Better results achieved for inception -v4 (0.7516) for product as fusion technique
[42]	Transfer learning with spatial correlation module	Uses concatenation as fusion technique	Uses average of kappa, AUC and F1-score. Resnet -101 produces better results around 0.827
proposed method	**Transfer learning using VGG-16**	**No fusion (considered left and right images as separate images)**	**Validation Accuracy :90.85 F1-score :0.91**

2) Specificity: Measures the accurate identification of true negative values, it is also called Selectivity and True Negativity.

$$Specificty = (\text{True Negatives}) \div (\text{False Positives} + \text{True Negatives}) \quad (7)$$

3) Precision: Measures the number of true positive values obtained over the total number of positive values, it is also called as positive Predictive value.

$$Precision = (\text{True Positives})\_ \div \_(\text{False Positives} + \text{True Positives}) \quad (8)$$

4) Sensitivity: The ratio of true positives to true positives and false negatives is called sensitivity; it is alsoreferred as True Positive Rate and Recall.

$$Sensitivity = (\text{True Positives}) \div (\text{True Positives} + \text{False Negatives}) \quad (9)$$

5) Classification error (C.E): It is defined as the ratio of sum of false positives and false negatives to the total number of samples.

$$C.E = (\text{True Positives} + \text{True Negatives}) \div \text{Total Number of Samples} \quad (10)$$

6) F1- score: It is defined as harmonic mean between precision and recall.

$$F1 - Score = 2 * (\text{Precision} \times \text{Recall}) \div (\text{Precision} + \text{Recall}) \quad (11)$$

7) False positive rate (FPR): It is measured as the ratio between false positive to false positive and true negative.

$$FPR = \text{False Positive} \div (\text{False Positive} + \text{True Negative}) \quad (12)$$

8) Negative predictive value (NPV): It is measured as the ratio between true negative to true negative and false negative.

$$NPV = \text{True Negative} \div (\text{True Negative} + \text{False Negative}) \quad (13)$$

9) False negative rate (FNR): It is measured as the ratio between false negative to false negative and true positive.

$$FNR = \text{False Negative} \div (\text{False Negative} + \text{True Positive}) \quad (14)$$

		Baseline (VGG16 - Multiclass)	Proposed Approach (VGG16 - One vs Rest, Binary model for each disease class)						
Eye Disorder	Disease Label	Baseline (Validation Accuracy)	Total Training Samples	Total Validation Samples	Training Time (per Epoch)	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy
Normal	N	40.00%	9887	2472	282s - 29ms/step	0.2735	87.09%	0.3135	85.28%
Diabetic Retinopathy	D	89.00%	9887	2472	271s - 27ms/step	0.3302	85.16%	0.4536	82.56%
Glaucome	G	54.00%	9887	2472	270s - 27ms/step	0.0696	97.60%	0.2891	91.34%
Cataract	C	97.00%	9887	2472	272s - 27ms/step	0.0409	98.60%	0.1111	97.05%
AMD	A	88.00%	9887	2472	270s - 27ms/step	0.0505	98.29%	0.3808	90.45%
Hypertension	H	95.00%	9887	2472	272s - 27ms/step	0.1021	96.03%	0.214	93.08%
Myopia	M	90.00%	9887	2472	282s - 29ms/step	0.0384	98.68%	0.0592	98.18%
Others	O	44.00%	9887	2472	271s - 27ms/step	0.2458	91.44%	0.3911	88.88%
Overall Accuracy		**89.06%**							**90.85%**

Fig. 5. Comparison between Proposed Approach and Approach given by N. Gour et al. [41].

TABLE V. METRICS ON TEST DATA

Class	Confusion Matrix	FPR	NPV	FNR	Specificity	PPV	Sensitivity/ Recall	Classification Error	F1-Score	Accuracy
N	[[776, 386], [290, 639]]	0.3766	0.6879	0.2720	0.6234	0.6678	0.7279	0.3233	0.6966	0.6767
D	[[1545, 6], [524, 16]]	0.2727	0.0296	0.2533	0.7273	0.9961	0.7467	0.2535	0.8536	0.7465
G	[[1886, 108], [36, 61]]	0.6391	0.6289	0.0187	0.3609	0.9458	0.9813	0.0689	0.9632	0.9311
C	[[1955, 41], [11, 84]]	0.3280	0.8842	0.0056	0.6720	0.9795	0.9944	0.0249	0.9869	0.9751
A	[[1921, 86], [43, 41]]	0.6772	0.4881	0.2189	0.3228	0.9572	0.9781	0.0617	0.9676	0.9383
H	[[1887, 146], [36, 22]]	0.8690	0.3793	0.0187	0.1309	0.9282	0.9813	0.0871	0.9539	0.9129
M	[[1991, 20], [8, 72]]	0.2174	0.9000	0.0040	0.7826	0.9901	0.9959	0.0134	0.9930	0.9866
O	[[1741, 0], [350, 0]]	NaN	0.0000	0.1674	NaN	1.0000	0.8326	0.1674	0.9087	0.8326
Overall		0.41	0.49	0.09	0.45	0.93	0.90	0.12	0.91	0.87

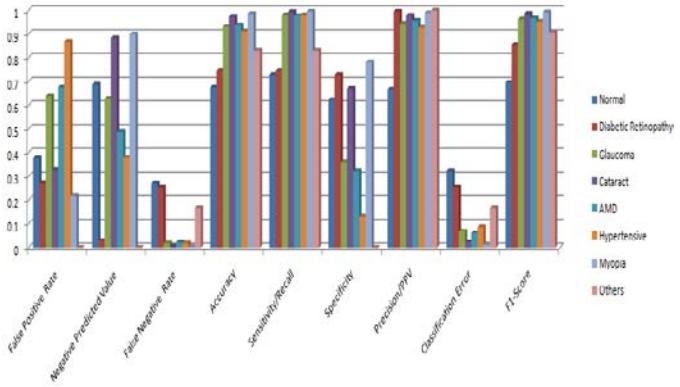

Fig. 6. Metrics in the Form of Barcharts for each and Every Disease.

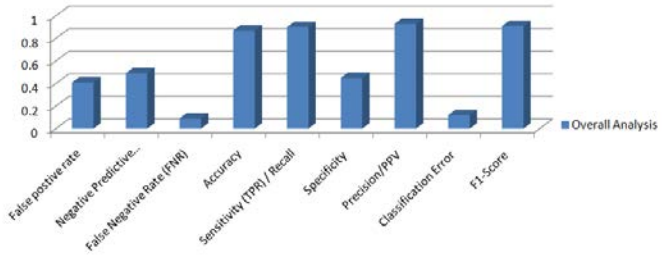

Fig. 7. Overall Analysis in the Form of Barcharts.

V. DISCUSSION

The paper demonstrates the lack of embedded machine learning techniques to solve the multi-class classification of retinopathy images. An important aspect of understanding the problem of the binary loss function is that because of the persistent 8 different classes, the exponential loss function is not able to provide substantially valid results as per the medical requirement. Because of which most of the retinopathy algorithms tend to create only theoretical steps to the earthquake in an autonomous system for multiclass classification in this field. The paper also takes into account various transfer learning accomplishments in the depth of image classification and computer vision. Furthermore, the paper can showcase the genesis idea of combining transfer learning into multiclass labelling techniques by using cross-entropy loss modelling ideas. A holistic discussion on various heuristics used is provided in the previous section and would be concluded in section 6. In the final analysis as per the literature review, we tend to understand the significance of machine learning ideas that have only been bubbled up and not implemented thoroughly. The paper ideation and implementation successfully answered where the research question is posted. While analyzing the results, it can be observed that transfer learning (Vgg16) with one verses rest classifier technique have produced exceptionally better results than baseline paper [41].

From the literature review, it is observed that no prediction is done on odir-19 testing dataset and most of the research is mainly focused on multiclass or one eye disease prediction. From the results, it is analyzed that training has produced reasonably better results (sensitivity, Accuracy, positive predictive value, F1-score) on overall testing data. Normal class shows average performance in all cases due to similarity in the features.

VI. CONCLUSION

The paper proposed a pipeline to identify multiple diseases on ODIR datasets where the aim was to increase baseline accuracy from 89% to almost 91% and also proposed model

has improved the performance of identifying disease drastically prediction of glaucoma has increased from 54% to 91%, Normal images prediction has increased from 40% to 85.28% and Other diseases prediction has increased from 44% to 88%. Out of 8 categories (seven diseases plus normal class) prediction proposed model prediction rate has improved in 6 categories, except in diabetic retinopathy and hypertension where proposed model accuracy has decreased by 6.44% and 2% respectively. The reason for achieving high accuracy in other categories is due to augmentation techniques where more balanced data is created as possible. That has clearly shown in less annotated diseases like glaucoma. The further research will be on working to create more data using other augmentation techniques like generating artificial images and working various transfer learning algorithms to improve the accuracy of each disease in multi-labeled classification problems. As for future work, other DL algorithms need to be explored for training the model and a study regarding hyperparameter optimization should be done to find the optimal model configuration. Moreover, other multi-label datasets with the latest eye diseases should be explored and predicted.

REFERENCES

[1] WHO: World report on vision. World Health Organisation (2019), https://www.who. int/publications-detail/world- report-on-vision

[2] C. C. W. Klaver, R. C. W. Wolfs, J. R. Vingerling, A. Hofman, and P. T. V. M. D. Jong, "Age-specific prevalence and causes of blindness and visual impairment in an older population: the Rotterdam Study." Archives of ophthalmology, Vol. 116, No. 5, 1998, pp. 653-658.

[3] M. H. Ansari, " Blindness after facial fractures: a 19-year retrospective study," Journal of oral and max- illofacial surgery, Vol. 63, No. 2, 2005, pp. 229-237.

[4] D. S. Fong, L. Aiello, T. W. Gardner, G.L. King, G. Blankenship, J. D. Cavallerano, F. L. Ferris, F. L., 3rd, R. Klein, & American Diabetes Association, "Retinopathy in diabetes," Diabetes care, Vol. 27, 2004, pp. S84–S87. https://doi.org/10.2337/diacare.27.2007.s84

[5] B. Thylefors, and A. D. Negrel. "The global impact of glaucoma." Bulletin of the World Health Organization, Vol. 72, No. 3, 1994, pp. 323-326.

[6] D. Allen, and A. Vasavada, "Cataract and surgery for cataract," Bmj, Vol. 333, No. 7559, 2006, pp. 128-132.

[7] AREDS2 Research Group, E. Y. Chew, T. Clemons, J. P. SanGiovanni, R. Danis, A. Domalpally, W. McBee, R. Sperduto, and F. L. Ferris, "The Age-Related Eye Disease Study 2 (AREDS2): study design and baseline characteristics (AREDS2 report number 1)," Ophthalmology, Vol. 119, No. 11, 2012, pp. 2282–2289. https://doi.org/10.1016/j.ophtha.2012.05.027

[8] E. Nolte, R. Scholz, V. Shkolnikov, and M. McKee, "The contribution of medical care to changing life expectancy in Germany and Poland," Social science & medicine (1982), Vol. 5, No. 11, 2002, pp. 1905–1921. https://doi.org/10.1016/s0277-9536(01)00320-3

[9] L. Konstantinidis, and Y. G. Crosier, "Hypertension and the eye." Current opinion in ophthalmology, Vol. 27, No. 6,2016, pp. 514-521.

[10] I. G. Morgan, K. O. Matsui, and S. M. Saw, "Myopia." The Lancet, Vol. 379, No. 9827, 2012, pp. 1739-1748.

[11] D. T. Munroe, and M. G. Madden, "Multi-class and single-class classification approaches to vehicle model recognition from images," proc. AICS, 2005, pp. 1-11.

[12] J. Liu, D. W. K. Wong, J. H. Lim, N. M. Tan, Z. Zhang, H. Li, F. Yin, B. Lee, S. M. Saw, L. Tong, and T. Y. Wong, "Detection of pathological myopia by PAMELA with texture-based features through an SVM ap- proach." Journal of Healthcare Engineering, 2010, pp. 1-11.

[13] T. V. Phan, L. Seoud, H. Chakor, and F. Cheriet, "Automatic screening and grading of age-related macular degeneration from texture analysis of fundus images." Journal of ophthalmology, 2016.

[14] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," JAMA, 2016, pp. E1-E9.

[15] H. Pratt, F. Coenen, D. M. Broadbent, , S. P. Harding, and Y. Zheng, "Convolutional Neural Networks for Diabetic Retinopathy," in International Conference On Medical Imaging Understanding and Analysis (MIUA), Procedia Computer Science, ELSEVIER, Vol. 90, 2016, pp. 200-205.

[16] J. Y. Choi, T. K. Yoo, J. G Seo, , J. Kwak, T. T. Um, and T. H. Rim, "Multi-categorical deep learning neural network to classify retinal images: A pilot study employing small database," PLoS ONE, Vol. 12, No. 11, 2017, pp. 1-17.

[17] P. M. Burlina, N. Joshi, K. D. Pacheco, T. Y. A. Liu, and N. M. Bressler, "Assessment of Deep Generative Models for High-Resolution Synthetic Retinal Image Generation of Age-Related Macular Degeneration," JAMA Ophthalmology, 2018, doi:10.1001/jamaophthalmol.2018.6156.

[18] Y. Chai, H. Liu, and J. Xu, "Glaucoma diagnosis based on both hidden features and domain knowledge through deep learning models." Knowledge-Based Systems, Vol. 161, 2018, pp. 147-156.

[19] F. Grassmann, J. Mengelkamp, C. Brand, S. Harsch, M. E. Zimmermann, B. Linkohr, A. Peters, I. M. Heid, C. Palm, and B. H.F. Weber, "A Deep Learning Algorithm for Prediction of Age-Related Eye Disease Study Severity Scale for Age-Related Macular Degeneration from Color Fundus Photography," AMERICAN ACADEMY OF OPHTHALMOLOGY, 2018, pp. 1-11.

[20] M. Mateen, J. Wen, S. Song, & Z. Huang, "Fundus image classification using VGG-19 architecture with PCA and SVD," Symmetry, Vol. 11, No. 1, 2019.

[21] Q. Meng, Y. Hashimoto, & S. I. Satoh, " Fundus image classification and retinal disease localization with limited supervision," In Asian conference on pattern recognition, Springer, 2019, pp. 469-482.

[22] H. Chen, S. Miao, D. Xu, G. D. Hager, & A. P. Harrison, "Deep hierarchical multi-label classification of chest X-ray images. In International Conference on Medical Imaging with Deep Learning," arxiv, in Proceedings of Machine Learning Research (PMLR), 2019, pp. 109-120.

[23] L. Faes, S. K. WagnerD. J. Fu, X. Liu, E. Korot, J. R. Ledsam, & P. A. Keane, "Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study," The Lancet Digital Health, Vol. 1, No. 5, pp. e232-e242.

[24] C.C. Jordi, N.D.R. Joan Manuel, V.R. Carles, "Ocular Disease Intelligent Recognition Through Deep Learning Architectures", Universitat Oberta de Catalunya, 2019, pp. 1–114.

[25] M. N. Bajwa, M. I. Malik, S. A. Siddiqui, A. Dengel, F. Shafait,W. Neumeier, and S. Ahmed, "Two-stage framework for optic disc localization and glaucoma classification in retinal fundus images using deep learning." BMC medical informatics and decision making, Vol. 19, No. 136, 2019, pp. 1-16.

[26] S. Keel, Z. Li, J. Scheetz, L. Robman, J. Phung, G. Makeyeva, K. Aung, C. Liu, X. Yan, W. Meng, R. Guymer, R. Chang, and M. He, "Development and validation of a deep-learning algorithm for the detection of neovascular age-related macular degeneration from colour fundus photographs," Clin. Exp. Ophthalmol., Vol. 47, No. 8, pp. 1009–1018, 2019, doi: 10.1111/ceo.13575.

[27] V. Das, S. Dandapat, P. K. Bora, "Multi-scale deep feature fusion for automated classification of macular pathologies from OCT images," in Biomedical Signal Processing and Control, Vol. 54, 2019, pp. 1-10. doi: 10.1016/j.bspc.2019.101605.

[28] 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, 2019, pp. 511-522. doi:10.1016/j.ins.2019.06.011.

[29] Y. Peng, S. Dharssi, Q. Chen, T. D. Keenan, E. Agrón, W. T. Wong, E. Y. Chew, and Z. Lu, "DeepSeeNet: a deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs," Ophthalmology, Vol. 126, No. 4, 2019, pp. 565-575.

[30] T.Pratap, P. Kokil, "Computer-aided diagnosis of cataract using deep transfer learning," Biomedical Signal Processing and Control, Vol. 53, 2019.

[31] M. S. Alabshihy, A. A. Maksoud, M. Elmogy, S. Barakat, and F. A. Badria, "Diagnosis of Diverse Retinal Disorders Using a Multi-Label Computer-Aided System, "Trends in Ophthalmology Open Access Journal, Vol. 2, No. 3, 2019, pp. 140-157.

[32] Md.T. Islam, S. A. Imran,A. Arefeen, M. Hasan, and C. Shahnaz, "Source andCamera Independent Ophthalmic Disease Recognition fromFundus Image Using Neural Network," in IEEE International Conference on Signal Processing, Information, Communication & Systems(SPICSCON), 2019, pp. 59-63.

[33] T. Nazir, A. Irtaza, A. Javed, H. Malik, D. Hussain, and R. A. Naqvi, "Retinal Image Analysis for Diabetes-Based Eye Disease Detection Using Deep Learning." Applied Sciences, Vol. 10, No. 18, 2020.

[34] X. Pan, K. Jin, J. Cao, Z. Liu, J. Wu, K. You, Y. Lu, Y. Xu, Z. Su, J. Jiang, K. Yao, and J. Ye, "Multi-label classification of retinal lesions in diabetic retinopathy for automatic analysis of fundus fluorescein angiography based on deep learning." Graefe's Archive for Clinical and Experimental Ophthalmology, Vol. 258, No. 4, 2020, pp. 779-785.

[35] M. Aamir, M. Irfan, T. Ali, G. Ali, A. Shaf, A. S. S, A. Al-Beshri, T. Alasbali, and M. H. Mahnashi, "An Adoptive Threshold-Based Multi-Level Deep Convolutional Neural Network for Glaucoma Eye Disease Detection and Classification," Diagnostics, Vol. 10, No. 8, 2020, doi: 10.3390/diagnostics10080602.

[36] C. G. Gonzalo, V. S. Gutierrez, P. H. Martınez, I. Contreras, Y. T. Lechanteur,A. Domanian,B. V. Ginneken, and C. I. Sanchez, "Evaluation of a deep learning system for the joint automated detection of diabetic retinopathy and age-related macular degeneration ," Acta Ophthalmologica, 2020 , pp. 368-377.

[37] R. Sarki, K. Ahmed, H. Wang, and Y. Zhang, "Automated detection of mild and multiclass diabetic eye diseases using deep learning," Heal. Inf. Sci. Syst., Vol. 8, No. 1, 2020, pp. 1–9, doi: 10.1007/s13755-020-00125-5.

[38] K. Shankar, , A. R. W. Sait, , D. Gupta, , S.K. Lakshmanaprabu, A. Khanna, and H. M. Pandey, "Automated detection and classification of fundus diabetic retinopathy images using synergic deep learning model," Pattern Recognition Letters, Vol. 133, 2020, pp. 210-216.

[39] A. Ram, and C. C. Reyes-Aldasoro, "The relationship between Fully Connected Layers and number of classes for the analysis of retinal images," arxiv, 2020, [Online]. Available: http://arxiv.org/abs/2004.03624.

[40] J. Wang, L. Yang, Z. Huo, W. He, and J. Luo, "Multi-Label Classification of Fundus Images with EfficientNet," IEEE Access, Vol. 8, 2020, pp. 212499–212508, doi: 10.1109/ACCESS.2020.3040275.

[41] N. Gour and P. Khanna, "Multi-class multi-label ophthalmological disease detection using transfer learning based convolutional neural network," Biomed. Signal Process. Control, Vol. 66, 2021, doi: 10.1016/j.bspc.2020.102329.

[42] N. Li, T. Li, C. Hu, K. Wang, and H. Kang, "A Benchmark of Ocular Disease Intelligent Recognition: One Shot for Multi-disease Detection," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), Vol. 12614 LNCS, 2021, pp. 177–193, doi: 10.1007/978-3-030-71058-3_11.

[43] J. He, C. Li, J. Ye, Y. Qiao, and L. Gu, "Multi-label ocular disease classification with a dense correlation deep neural network," Biomed. Signal Process. Control, Vol. 63, 2021, doi: 10.1016/j.bspc.2020.102167.

[44] E. S. Kumar & C. S. Bindu, "MDCF: Multi-Disease Classification Framework On Fundus Image Using Ensemble Cnn Models," Journal of Jilin University, Vol. 40, No. 09, pp.35-45, 2021

[45] D. Müller, I. Soto-Rey, and F. Kramer, "Multi-Disease Detection in Retinal Imaging based on Ensembling Heterogeneous Deep Learning Models", arxiv, 2021, Available : http://arxiv.org/abs/ 2103.14660.

[46] A.C. Garcia, M.G.Dominguez, J. Heras, A. Ines, D.Royo, and M. A. Zapala, "Prediction of Epiretinal membrane from Retinal fundus images using deep learning, in CAEPIA 2021, LNAE 12882 19th ed. Springer, 2021, pp. 2-13.

[47] L. P. Cen, J. Ji, J. W. Lin, S. T. Ju, H. J. Lin, T. P. Li, Y. Wang, J. F. Yang, Y. F. Liu, S. Tan, L. Tan, D. Li, Y. Wang, D. Zheng, Y. Xiong, H. Wu, J. Jiang, Z. Wu, D. Huang, T. Shi, B. Chen, J. Yang, X. Zhang, L. Luo, C. Huang, G. Zhang, Y. Huang, T. K. Ng, H. Chen, Weiqi Chen, C. P. Pang, and M. Zhang, "Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks," NATURE COMMUNICATIONS, Vol. 12, 2021.

[48] Y. LeCun, and Y. Bengio. "Convolutional networks for images, speech, and time series." The handbook of brain theory and neural networks, Vol. 3361, No. 10, 1998, pp. 255-258.

[49] A. Krizhevsky, I. Sutskever, G.E. Hinton, "Imagenet classification with deep convolutional neural networks," in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[50] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, "Going deeper with convolutions," in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[51] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv, 2015. arXiv preprint arXiv:1409.1556.

[52] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv, 2017, arXiv preprint arXiv:1704.04861.

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

[54] L. Bottou, "Stochastic gradient descent tricks," Neural networks: Tricks of the trade. Springer, Berlin, Heidelberg, 2012, pp. 421-436.