# Multi_Disease Detection From Retinal fundus images Based on Machine Learning Models

**Your Name**

Omid Jadidi

Mohammad Abedi

Mohammad Koosheshi

Shervin Issakhani

# ABSTRACT

Preventable or undiagnosed impairment and blindness have a huge impact on billions of people worldwide.  The World Health Organization (WHO) estimates the prevalence of blindness and visual impairment to 2.2 billion people worldwide, of whom at least 1 billion affections could have been prevented. Automated multi-disease detection models offer great potential to address this problem via clinical decision support in diagnosis. This study proposed a multi-class classification pipeline for retinal imaging which utilizes ensemble learning to combine the predictive capabilities of several heterogeneous deep convolutional neural network models like VGG 19, DenseNet 201, Inception V3, ResNet 50, and a CNN-based defined model. Our pipeline includes strategies like transfer learning with frozen weights trained on the ImageNet dataset and fine-tuning method. Furthermore,  we integrated ensemble learning techniques. Through evaluation, our results demonstrated that our proposed model can detect a variety of lesions in the color images of the fundus, which lays a foundation for assisting doctors in diagnosis and makes it possible to carry out rapid and efficient large-scale screening of fundus lesions.

# INTRODUCTION

Proper diagnosis and detection of retinal diseases among which Diabetic Retinopathy (DR) is the most common and severe microvascular complication have been an extensive challenge in recent years. Although medical facilities have progressed to a great extent in the last 30 years and enabled experts to cure the majority of visual impairment diseases, population growth and aging made fast and accurate diagnosis an unresolved challenge[cite 1]. Patients can be diagnosed quickly with fundus photography of the retina, but its accuracy is dependent largely on the physician's experience. Based on the World Health Organization's estimates, a world report on vision 2019, there is a total number of 2.2 billion people worldwide who suffer from visual impairment. If a correct and inexpensive diagnosis existed, at least 1 out of 2.2 billion affections could have been avoided [cite 2]. In extreme cases, an in-time diagnosis can protect a person from blindness.

In the past few years, the use of machine learning and even deep learning models to classify images, specifically medical images, has been on the rise [cite]. Multiple reasons account for this dramatic increase such as accurate classification, early diagnosis, and inexpensive try. The debate, now, is more related to just increasing the model accuracy and reducing the computational cost and effort.

Multiple studies have classified retinal images based on various diseases by implementing deep learning models.

Müller et al. proposed an ensemble learning approach to multi-disease detection for retinal imaging. Their approach included multiple strategies like transfer learning, class weighting, real-time image augmentation, and focal loss utilization. They used the Retinal Fundus Multi-Disease Image Dataset (RFMiD) which consists of 3200 retinal images that were annotated with 46 conditions.

Cheng et Al. proposed a two-part model to construct a multi-label classification model to be accurate on fundus images. The first part was image feature extraction and the second part was graphic convolutional network. Their dataset consisted of 7500 fundus images and 8 diseases. With the help of these two parts, they reached an average accuracy of 93 percent.

In this paper, we intend to move toward an accurate and reliable multi-disease detection model based on ensemble, transfer, and deep learning techniques.

In this study, we implement the Retinal Fundus Images (RFI) containing 11 conditions and 21746 images to validate our model's effectiveness in detecting rare diseases.

## 2. Methods

The implemented medical image classification pipeline to construct a multi-class classification model which would be suitable for fundus images can be summarized in the following core parts and is illustrated in Fig.1 :

- Resize images and defining the batch size
- Four defined Convolutional Neural Network (CNN) blocks inspired by the Inception model (block a,b,c,d)
- Individual training for multi-disease labels and disease risk detection utilizing transfer learning
- Fine-tuning method
- Multiple deep learning model architectures
- Ensemble learning strategy

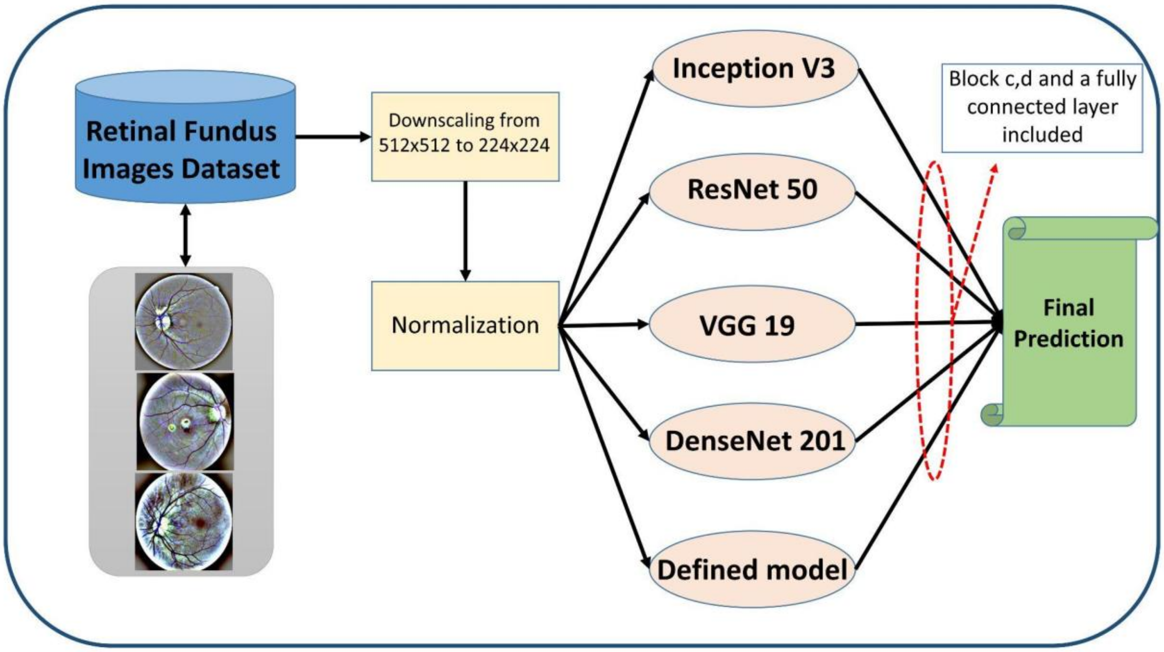

Retinal Fundus Images Dataset
Downscaling from 512x512 to 224x224
Normalization
Inception V3
ResNet 50
VGG 19
DenseNet 201
Defined model
Block c,d and a fully connected layer included
Final Prediction

*Figure 1- Flowchart diagram of the implemented medical image analysis pipeline for multi-disease detection in retinal imaging. The workflow is starting with the retinal imaging dataset (RFMiD) and ends with computed predictions for novel images.*

## 2.1 Dataset

The RFI dataset is an open-source dataset from Kaggle that consists of 21746 retinal images for which 20077 images were used as training dataset, 433 and 1236 images were used as test and validation dataset respectively. Each of these images has a width and height of  512, and 3 channels (RGB channels).  Four of these eleven classes belong to DR diseases which indicated the severity of this condition categorized as mild, moderate, severe, and proliferate DR. Wet and dry AMD, glaucoma, cataract, Pathological Myopia(PM), and Hypertensive Retinopathy (HR) account for six other condition classes. Ultimately, the last class belongs to normal fundus images with no diagnosed conditions. These classes are listed in Tab.1 with their frequency.

| Disease | Training samples | Test samples | Validation samples |
|---|---|---|---|
| Dry AMD | 1276 | 54 | 30 |
| Wet AMD | 545 | 23 | 19 |
| Mild DR | 2294 | 102 | 42 |
| Moderate DR | 4982 | 216 | 90 |
| Severe DR | 1635 | 107 | 49 |
| Proliferate DR | 1295 | 91 | 30 |
| Cataract | 1369 | 112 | 24 |
| HR | 1220 | 94 | 30 |
| PM | 1142 | 102 | 21 |
| Glaucoma | 1678 | 156 | 44 |
| Normal Fundus | 2641 | 179 | 54 |

*Table 1-   Annotation frequency for each class in the dataset.*

Figure 2 depicts a sample of each disease.

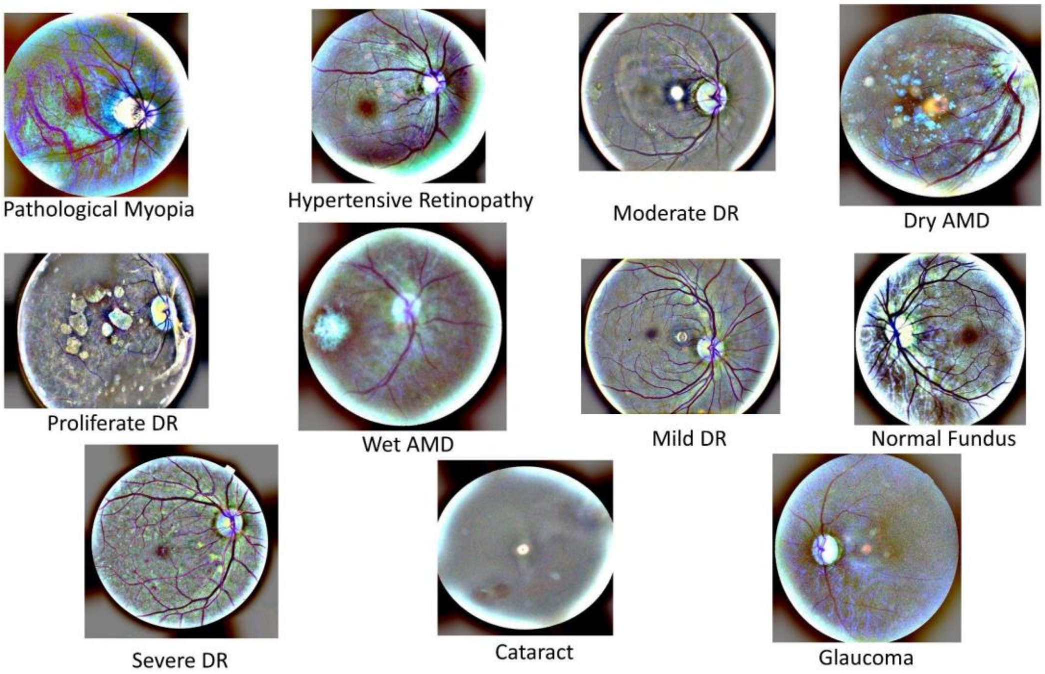


*Figure 2- A sample of all of the classes - Each of these images has a width and height of 512, and 3 channels (RGB channels).*

## 2.2 Preprocessing

In order to simplify the pattern-finding process of the deep learning model, as well as to decrease computational cost, several preprocessing methods were applied to the dataset. The common input image size for well-known deep learning models that is applicable in image classification is normally 224 by 224. Thus, each image resolution was first converted from 512x512x3 to 224x224x3. This would reduce the runtime as well.

Afterward, each pixel of each image which entails a number between 0 to 255 was divided by 255 to convert to a range between 0 to 1. This is called normalization, The reason lies within the fact that the gradient descent would converge better, and also the computation of high numeric values may become more complex in deep neural networks.

We also wanted to implement image augmentation to balance class distribution, but image augmentation was already applied to the raw dataset.

## 2.3 Deep Learning Models

Medical image classification is dominated by the unmatched deep convolutional neural network model. Nevertheless, the hyper parameter configuration and architecture selection are highly dependent on the required computer vision task, as well as the key difference between pipelines. Thus, our pipeline combines five different types of image classification models: 1) Inception, 2) ResNet, 3) VGG, 4) DenseNet, and 5) Our CNN model inspired by inception. The first four model types were pre-trained on the ImageNet dataset, and the top fully connected layers are not included. The ImageNet is a large visual database designed for use in visual object recognition software research. More than 14 million images have been hand-annotated by the project to indicate what objects are pictured and it contains more than 20,000 classes. For the fitting process, we applied transfer learning training, with frozen architecture layers, and a fine-tuning strategy with unfrozen layers. The transfer learning fitting with frozen weights was performed for 15 epochs using the Adam optimizer with an initial learning rate of 0.01. Whereas the fine-tuning had a maximal training time of 10 epochs utilizing adam optimizer with a learning rate of  1 -E05.

The model loss function is categorical cross-entropy as we have a multi-class classification problem. It is formulated as below:

$$Loss = -\sum_{i=1}^{output\ size} y_i \,.\, log\ \hat{y}_i$$

In which $y_i$ is the true value of the input, whereas $\hat{y}_i$ is the predicted value based on our model, and $i$

indicated the class number. It will calculate the average difference between the actual and predicted probability distributions for all classes in the problem. The score is minimized and a perfect cross-entropy value is 0.

### 2.3.1 Inception Version 3

Inception-v3 is a convolutional neural network deep architecture from the Inception family that makes several improvements including using Label Smoothing, Factorized 7 x 7 convolutions, and the use of an auxiliary classifier to propagate label information lower down the network and enhance vanishing gradient problem. Figure 3 illustrates its overall architecture. The reduction blocks shown in the figure have two main purposes: 1) They reduce the probability of overfitting, and 2) reduce the number of parameters with the help of asymmetric convolution layers, consequently, reducing the computational effort.

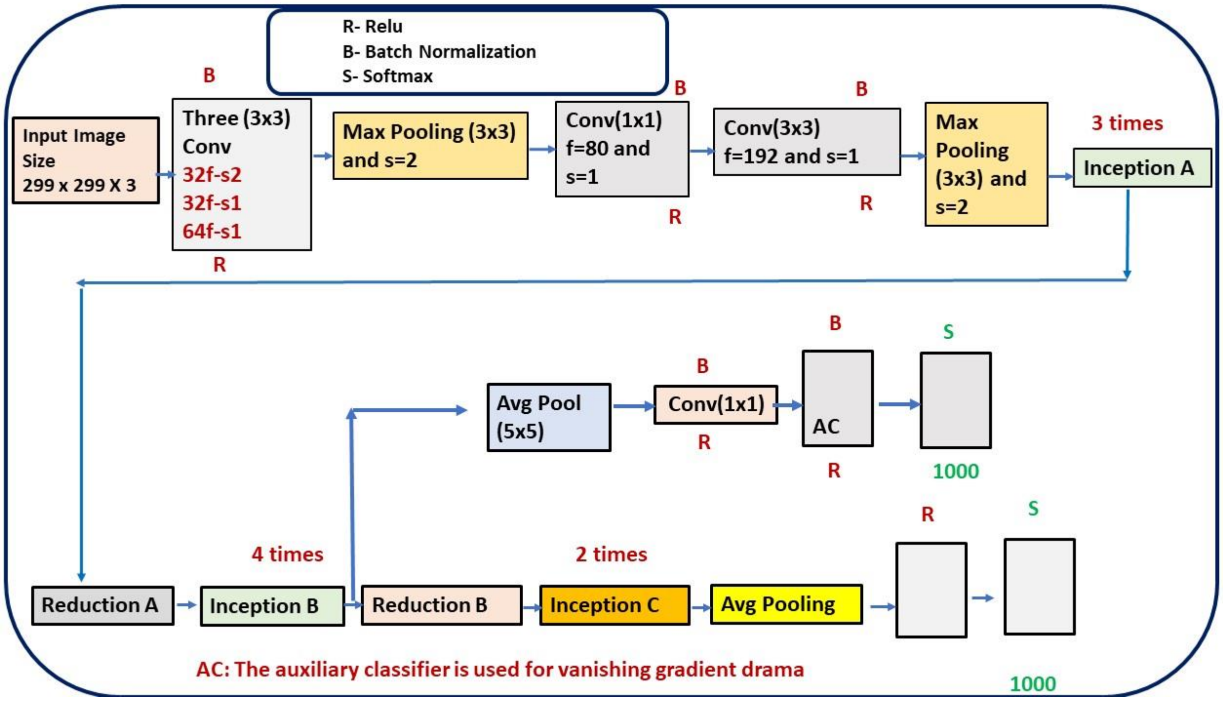


*Figure 3- Inception V3 architecture - It is made up of 42 layers. There are 3 inception and 2 reduction blocks.*

We took advantage of this brilliant model in one of five final models.

### 2.3.2 ResNet 50 Model

Residual Network (ResNet) is one of the famous deep learning models that was introduced by Shaoqing Ren, Kaiming He, Jian Sun, and Xiangyu Zhang in their paper. ResNet50 is a variant of the ResNet model which has 48 Convolution layers along with 1 MaxPool and 1 Average Pool layer.

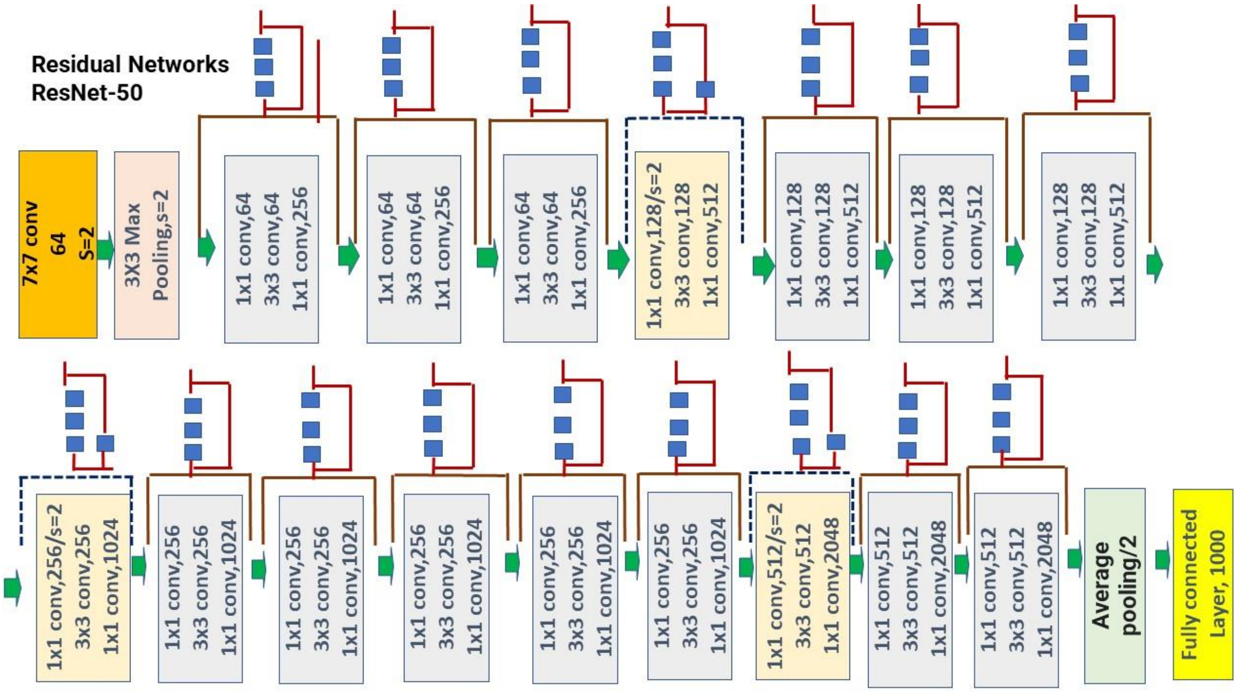

Residual Networks
ResNet-50
7x7 conv 64 S=2
3X3 Max Pooling,s=2
1x1 conv,64 3x3 conv,64 1x1 conv,256
1x1 conv,64 3x3 conv,64 1x1 conv,256
1x1 conv,64 3x3 conv,64 1x1 conv,256
1x1 conv,128/s=2 3x3 conv,128 1x1 conv,512
1x1 conv,128 3x3 conv,128 1x1 conv,512
1x1 conv,128 3x3 conv,128 1x1 conv,512
1x1 conv,128 3x3 conv,128 1x1 conv,512
1x1 conv,256/s=2 3x3 conv,256 1x1 conv,1024
1x1 conv,256 3x3 conv,256 1x1 conv,1024
1x1 conv,256 3x3 conv,256 1x1 conv,1024
1x1 conv,256 3x3 conv,256 1x1 conv,1024
1x1 conv,256 3x3 conv,256 1x1 conv,1024
1x1 conv,256 3x3 conv,256 1x1 conv,1024
1x1 conv,512/s=2 3x3 conv,512 1x1 conv,2048
1x1 conv,512 3x3 conv,512 1x1 conv,2048
1x1 conv,512 3x3 conv,512 1x1 conv,2048
Average pooling/2
Fully connected Layer, 1000

*Figure 4- ResNet 50 architecture - ResNet50 is a variant of the ResNet model which has 48 Convolution layers along with 1 MaxPool and 1 Average Pool layer.*

Figure 4 shows a summary of ResNet 50 architecture. The skip connections in ResNet solve the problem of vanishing gradient in deep CNNs by allowing alternate shortcut paths for the gradient to flow through. Also, the skip connection helps if any layer hurts the performance of architecture, then it will be skipped by regularization.

### 2.3.3 VGG 19

VGG is a classical convolutional neural network architecture. It was based on an analysis of how to increase the depth of such networks. The network utilizes small 3 x 3 filters. Otherwise, the network is characterized by its simplicity: the only other components being pooling layers and a fully connected layer. VGG19 is a variant of the VGG model which in short consists of 19 layers (16 convolution layers, 3 fully connected layers, 5 MaxPool layers, and 1 SoftMax layer).

### 2.3.4 DenseNet 201 Model

In a DenseNet architecture, each layer is connected to every other layer, hence the name Densely Connected Convolutional Network. For L layers, there are L(L+1)/2 direct connections. For each layer, the feature maps of all the preceding layers are used as inputs, and their own feature maps are used as input for each subsequent layer. DenseNet-201 is a convolutional neural network that is 201 layers deep. We loaded a pre-trained version of the network trained on more than a million images from the ImageNet database

### 2.3.5 Defined CNN model

In the last part, we designed a CNN architecture that is inspired by the inception model. All of the parameters in this model are trainable. It contains four blocks, and these blocks are stacked consecutively. Block a and b architecture are virtually the same, with the difference that in block a max-pooling layer performs the operation on input before concatenation occurs, however, in block b max-pooling layer is placed after the concatenation (figure 5). The same trend happens in blocks c and d respectively. Figure 6 depicts a summary of this defined CNN model.

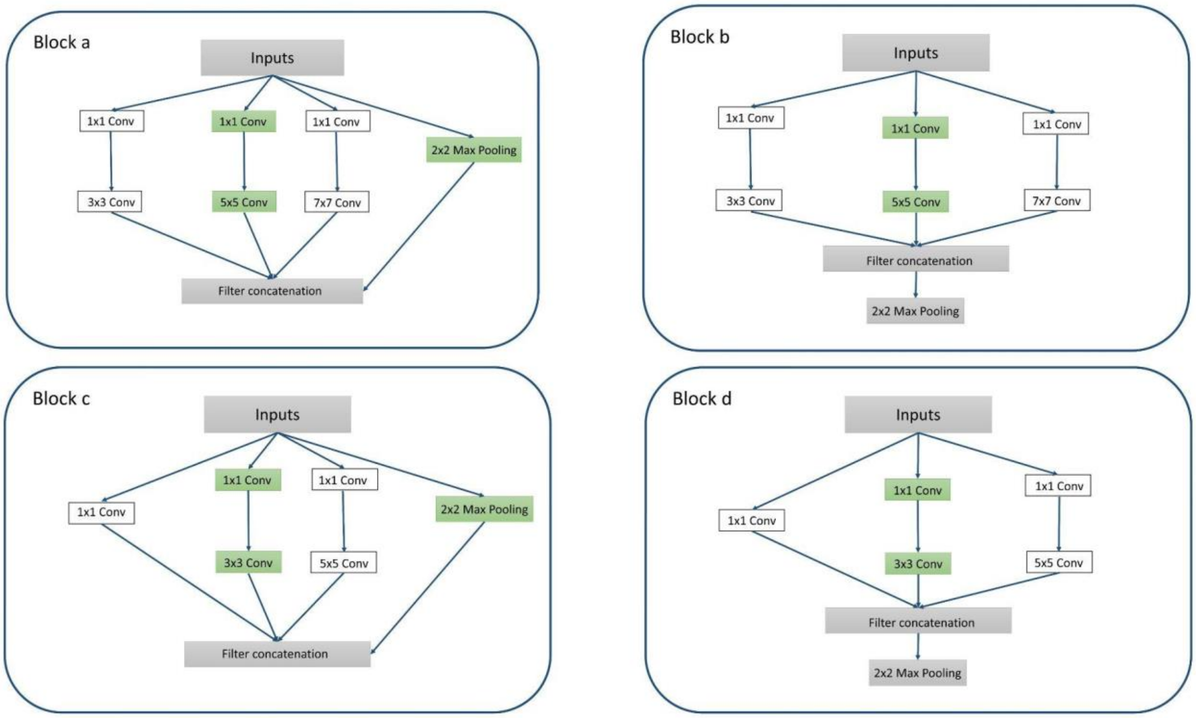


*Figure 6- a,b,c, and d blocks architecture- All of them are composed of simple CNN layers like simple convolution and max-pooling layers.*

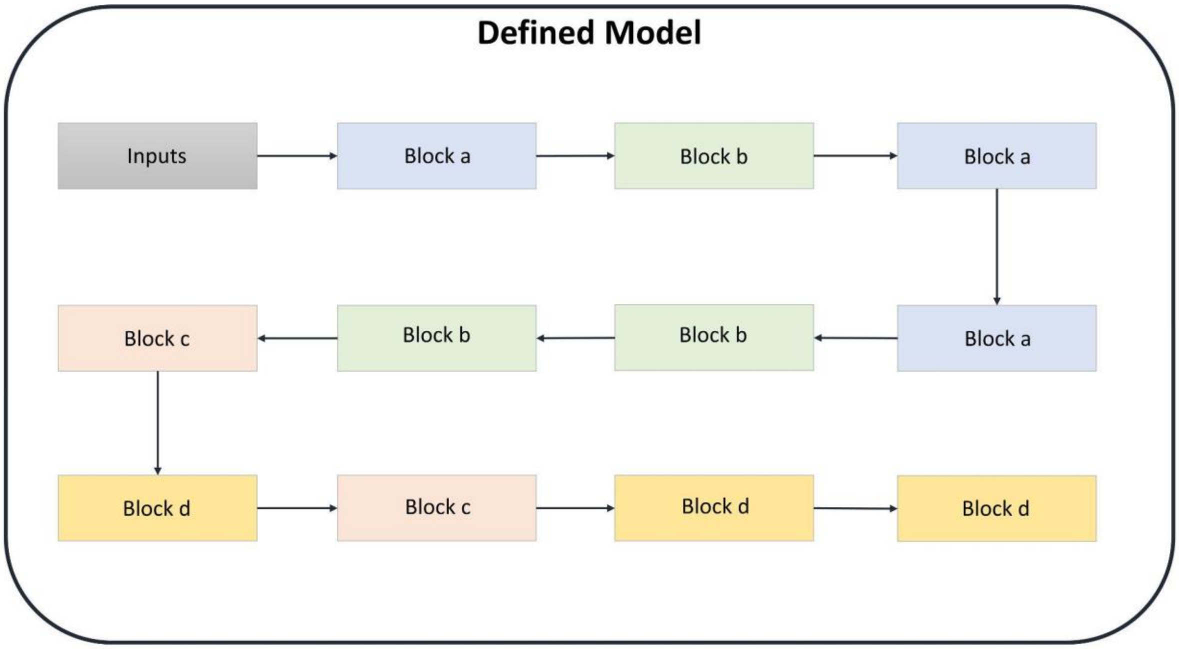


*Figure 6- Defined model architecture - This is inspired by inception. It consists of 11 CNN blocks followed by each other in the shown order.*

### 2.4 Ensemble Learning Strategy

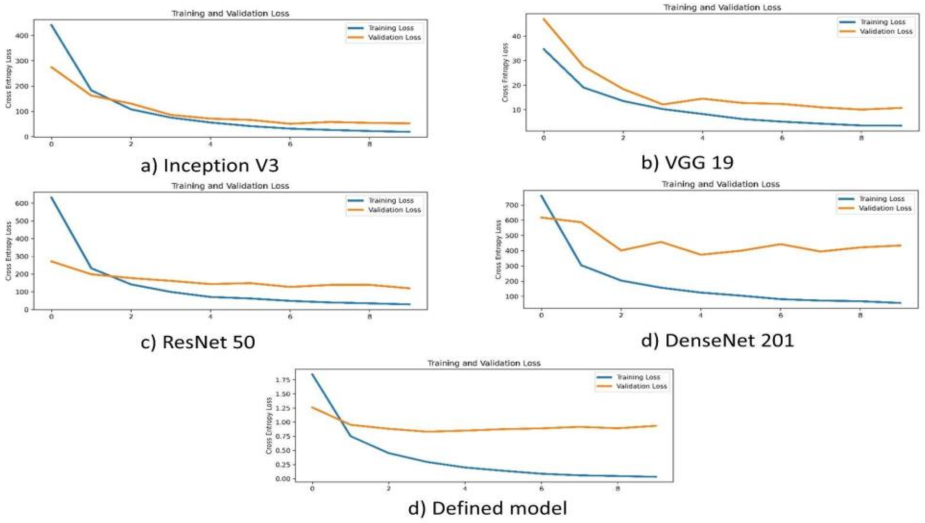
Next to the utilization of various deep learning models that we discussed in the previous section, we also applied ensemble learning.  Ensemble learning is a general meta approach to machine learning that seeks better predictive performance by combining the predictions from multiple models.  Our aim was to create a large variety of models which were trained to construct a better decision-making structure. This approach not only allowed a more efficient usage of the available training data but also increased the reliability of a prediction.

# Results and Discussion

The sequential training of a complete cross-validation for one architecture alone on Google Colab GPU which is NVIDIA TESLA K80 GPU took approximately 2 hours with 25 epochs on average for each deep convolutional neural network model. In the first place we fitted each of the five mentioned models followed by block c,d, and a fully connected layer with 11 outputs which was the classes on the dataset and saved the trainable parameter weights in the memory. We also tried to train each model with two fully-connected layers at the top, but by doing this, the number of trainable parameters increased and the chance of overfitting rose dramatically.

## 3-1 Performance Evaluation

Figure 7 illustrates the categorical cross-entropy loss function values for each trained model separately. No signs of overfitting were observed for the VGG 19, ResNet 50, and Inception v3, however, DenseNet 201 and our defined CNN model showed a trend of overfitting in the fine-tuning phase (Figure 7). It is evident from figure 7 that inception v3 worked really well on our dataset since the training and validation loss function values are virtually the same and we did not encounter any overfitting issues, although our defined model loss is the least one among them, the validation loss has not changed significantly.


a) Inception V3

b) VGG 19

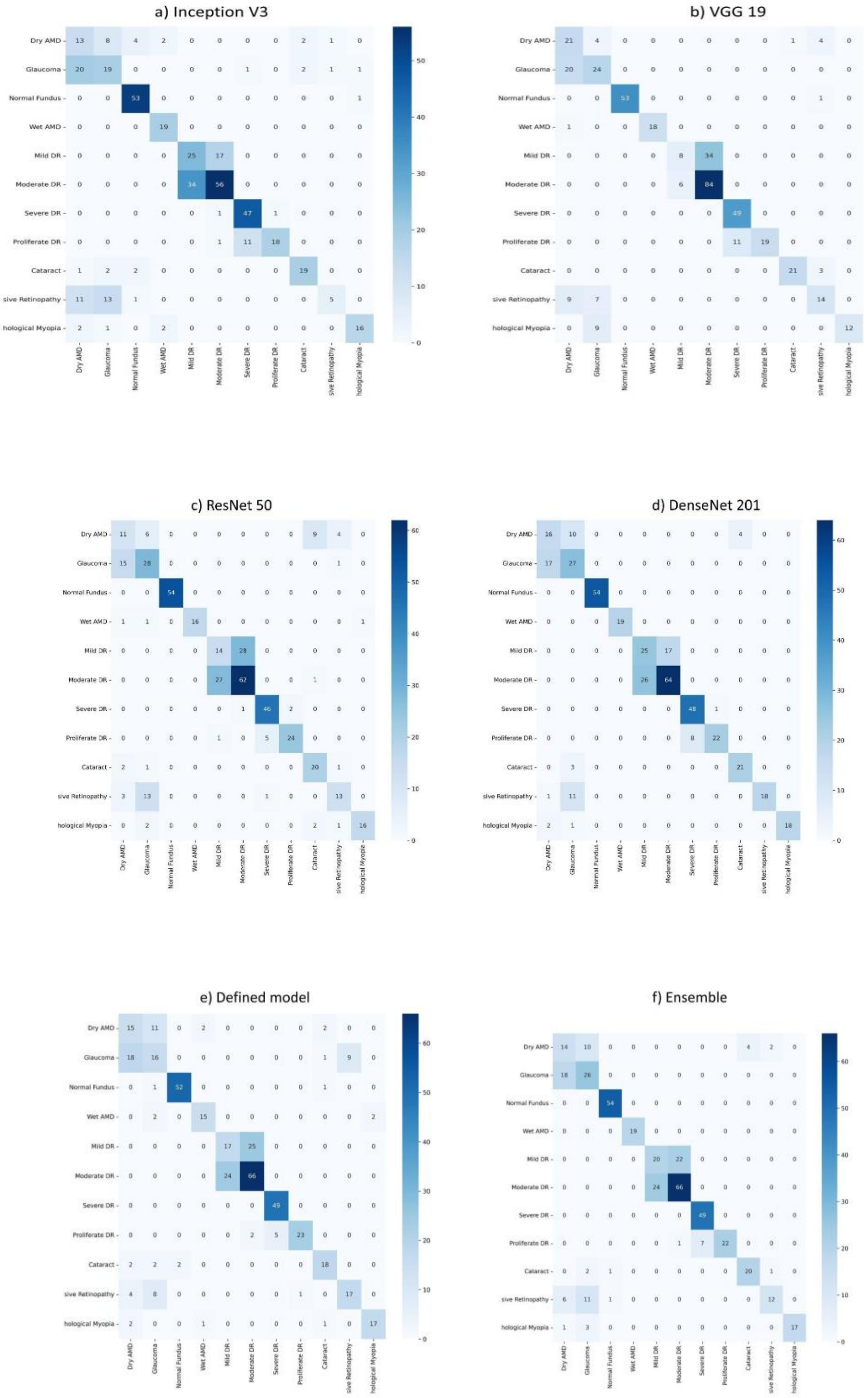
c) ResNet 50

d) DenseNet 201

d) Defined model

*Figure 7 - Loss course during the training process for training and validation data. The lines were computed via locally estimated scatterplot smoothing.*

Figure 8 depicts the confusion matrix for each of the trained models individually. The confusion matrix is a metric that shows the effectiveness of the trained model where output can be two or more classes. It is a table that, in our case, displays a combination of the predicted and actual values. By comparing the confusion matrix of different models, it can be seen that the Glaucoma disease class has the most number of true negatives and false positives which is not a desirable result. We speculated that the reason is that Glaucoma looked like just hypertensive retinopathy. Both of them are diagnosed with the thinning of the neuroretinal rim such that the optic disc appears excavated. Moreover, it might be because of a lack of data. Also, the reason might lie within the fact that it was difficult for the model to distinguish the thinned neuroretinal rims from the background of the image. The 2017 report of Tan et al. also verified our opinions. Tan et al. proposed a ten layers CNN architecture for DR lesion segmentation and achieved a sensitivity of 0.46 for segmentation of microaneurysms. Tan's job showed that microaneurysms were very difficult to distinguish from the surrounding background pixels. For soft exudates and hard exudates, we found that these two kinds of lesions often accompanied multiple other fundus lesions at the same time, which made the model difficult to extract the features of all fundus lesions. Moreover distinguishing Mild from Moderate DR was also a challenge for the proposed models. These two classes belong to the same condition with just a difference in severity, so it is comprehendible why this task is this much difficult for the trained model. Despite these difficulties, it can be seen that the Ensemble confusion matrix has ameliorated a lot.


a) Inception V3

b) VGG 19

c) ResNet 50

d) DenseNet 201

e) Defined model

f) Ensemble

*Figure 8 - Confusion matrix for each model after the training process - It is a good representation of each model's efficiency*

As one of the metrics of the represented model, the F-1 score for each class in each trained model is shown in Tab 2 and Fig 9. F1 score is defined as the harmonic mean between precision and recall. It is used as a statistical measure to rate performance. In other words, an F1 score (from 0 to 1, 0 being the lowest and 1 being the highest) is a mean of an individual's performance, based on two factors i.e. precision and recall.

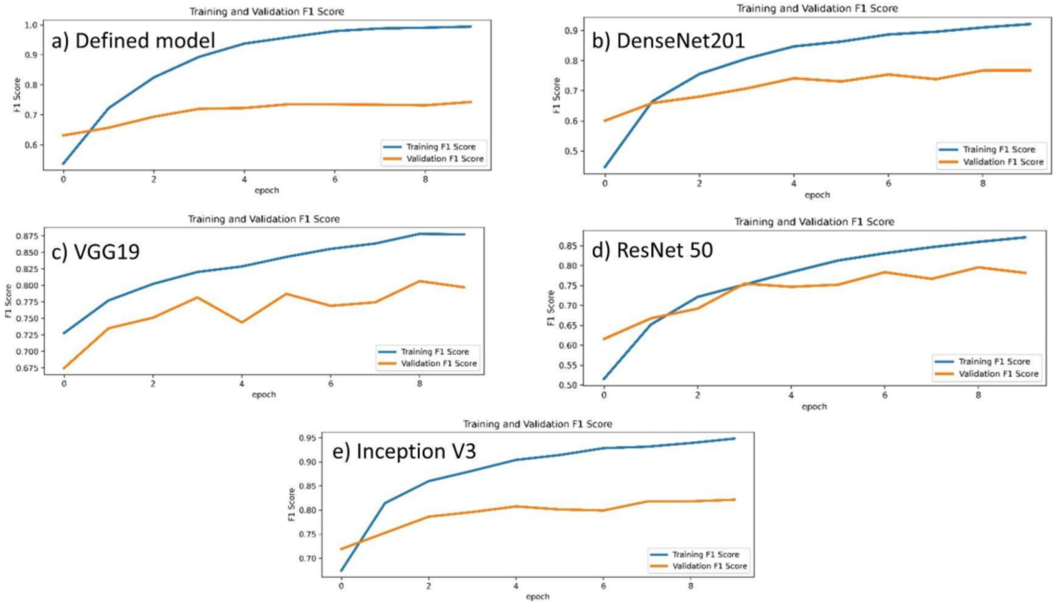


*Figure 9 - F-1 score change during the training process for training and validation data. The lines were computed via locally estimated scatterplot smoothing.*

| Disease | Inception V3 | Defined model | VGG 19 | DenseNet | ResNet | Ensemble |
|---|---|---|---|---|---|---|
| Dry AMD | 0.34 | 0.42 | 0.51 | 0.48 | 0.35 | 0.40 |
| Glaucoma | 0.44 | 0.38 | 0.54 | 0.56 | 0.59 | 0.54 |
| Normal Fundus | 0.93 | 0.96 | 0.88 | 1 | 1 | 0.98 |
| Wet AMD | 0.90 | 0.81 | 0.97 | 1 | 0.91 | 1 |
| Mild DR | 0.50 | 0.40 | 0.28 | 0.53 | 0.33 | 0.46 |
| Moderate DR | 0.68 | 0.72 | 0.80 | 0.74 | 0.68 | 0.73 |
| Severe DR | 0.87 | 0.95 | 0.89 | 0.91 | 0.91 | 0.84 |
| Proliferate DR | 0.73 | 0.85 | 0.77 | 0.83 | 0.85 | 0.83 |
| Cataract | 0.81 | 0.76 | 0.91 | 0.85 | 0.71 | 0.53 |
| HR | 0.27 | 0.60 | 0.53 | 0.75 | 0.52 | 0.89 |
| PM | 0.82 | 0.85 | 0.72 | 0.92 | 0.84 | 0.73 |
| Average | 0.66 | 0.70 | 0.72 | 0.78 | 0.70 | 0.75 |

*Table 2- F-1 score for each trained model for each diseases - It can be seen that the more data that we have - the higher the F-1 score is.*

Tables below depicts the classification report for each model.

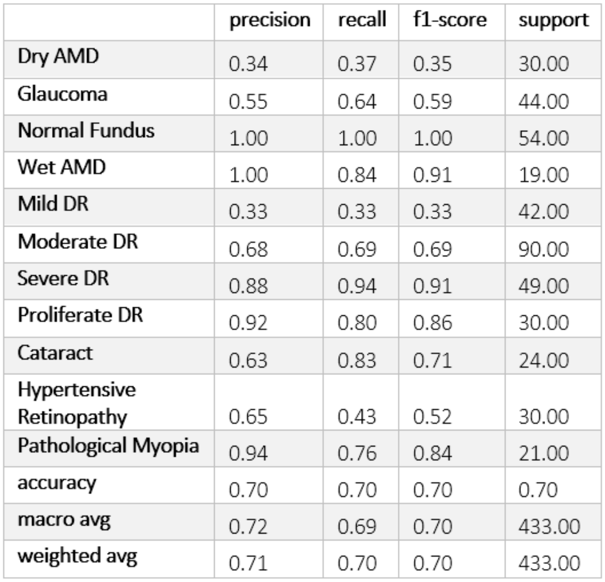
*Table 1 ensemble*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Dry AMD | 0.36 | 0.47 | 0.41 | 30.00 |
| Glaucoma | 0.50 | 0.59 | 0.54 | 44.00 |
| Normal Fundus | 0.96 | 1.00 | 0.98 | 54.00 |
| Wet AMD | 1.00 | 1.00 | 1.00 | 19.00 |
| Mild DR | 0.45 | 0.48 | 0.47 | 42.00 |
| Moderate DR | 0.74 | 0.73 | 0.74 | 90.00 |
| Severe DR | 0.88 | 1.00 | 0.93 | 49.00 |
| Proliferate DR | 1.00 | 0.73 | 0.85 | 30.00 |
| Cataract | 0.83 | 0.83 | 0.83 | 24.00 |
| Hypertensive Retinopathy | 0.80 | 0.40 | 0.53 | 30.00 |
| Pathological Myopia | 1.00 | 0.81 | 0.89 | 21.00 |
| accuracy | 0.74 | 0.74 | 0.74 | 0.74 |
| macro avg | 0.78 | 0.73 | 0.74 | 433.00 |
| weighted avg | 0.76 | 0.74 | 0.74 | 433.00 |

*Table 2 Defined Model*

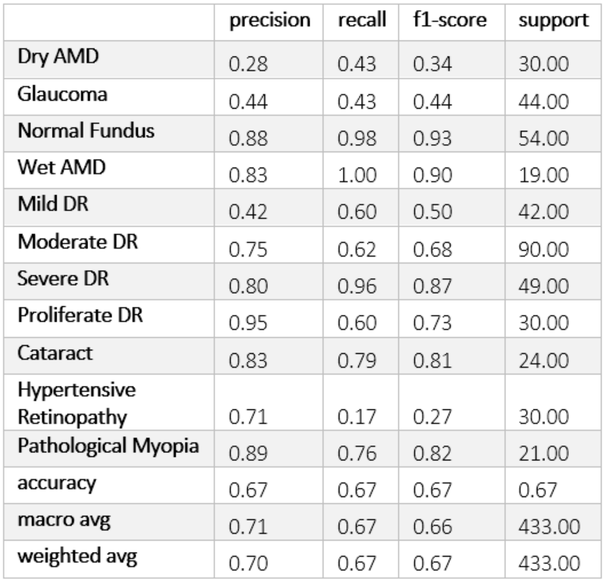
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Dry AMD | 0.37 | 0.50 | 0.42 | 30.00 |
| Glaucoma | 0.40 | 0.36 | 0.38 | 44.00 |
| Normal Fundus | 0.96 | 0.96 | 0.96 | 54.00 |
| Wet AMD | 0.83 | 0.79 | 0.81 | 19.00 |
| Mild DR | 0.41 | 0.40 | 0.41 | 42.00 |
| Moderate DR | 0.71 | 0.73 | 0.72 | 90.00 |
| Severe DR | 0.91 | 1.00 | 0.95 | 49.00 |
| Proliferate DR | 0.96 | 0.77 | 0.85 | 30.00 |
| Cataract | 0.78 | 0.75 | 0.77 | 24.00 |
| Hypertensive Retinopathy | 0.65 | 0.57 | 0.61 | 30.00 |
| Pathological Myopia | 0.89 | 0.81 | 0.85 | 21.00 |
| accuracy | 0.70 | 0.70 | 0.70 | 0.70 |
| macro avg | 0.72 | 0.70 | 0.70 | 433.00 |
| weighted avg | 0.71 | 0.70 | 0.71 | 433.00 |

*Table 3 Resnet 50*



| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Dry AMD | 0.34 | 0.37 | 0.35 | 30.00 |
| Glaucoma | 0.55 | 0.64 | 0.59 | 44.00 |
| Normal Fundus | 1.00 | 1.00 | 1.00 | 54.00 |
| Wet AMD | 1.00 | 0.84 | 0.91 | 19.00 |
| Mild DR | 0.33 | 0.33 | 0.33 | 42.00 |
| Moderate DR | 0.68 | 0.69 | 0.69 | 90.00 |
| Severe DR | 0.88 | 0.94 | 0.91 | 49.00 |
| Proliferate DR | 0.92 | 0.80 | 0.86 | 30.00 |
| Cataract | 0.63 | 0.83 | 0.71 | 24.00 |
| Hypertensive Retinopathy | 0.65 | 0.43 | 0.52 | 30.00 |
| Pathological Myopia | 0.94 | 0.76 | 0.84 | 21.00 |
| accuracy | 0.70 | 0.70 | 0.70 | 0.70 |
| macro avg | 0.72 | 0.69 | 0.70 | 433.00 |
| weighted avg | 0.71 | 0.70 | 0.70 | 433.00 |

*Table 4 cr_inc_c_d_11D*



| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Dry AMD | 0.28 | 0.43 | 0.34 | 30.00 |
| Glaucoma | 0.44 | 0.43 | 0.44 | 44.00 |
| Normal Fundus | 0.88 | 0.98 | 0.93 | 54.00 |
| Wet AMD | 0.83 | 1.00 | 0.90 | 19.00 |
| Mild DR | 0.42 | 0.60 | 0.50 | 42.00 |
| Moderate DR | 0.75 | 0.62 | 0.68 | 90.00 |
| Severe DR | 0.80 | 0.96 | 0.87 | 49.00 |
| Proliferate DR | 0.95 | 0.60 | 0.73 | 30.00 |
| Cataract | 0.83 | 0.79 | 0.81 | 24.00 |
| Hypertensive Retinopathy | 0.71 | 0.17 | 0.27 | 30.00 |
| Pathological Myopia | 0.89 | 0.76 | 0.82 | 21.00 |
| accuracy | 0.67 | 0.67 | 0.67 | 0.67 |
| macro avg | 0.71 | 0.67 | 0.66 | 433.00 |
| weighted avg | 0.70 | 0.67 | 0.67 | 433.00 |

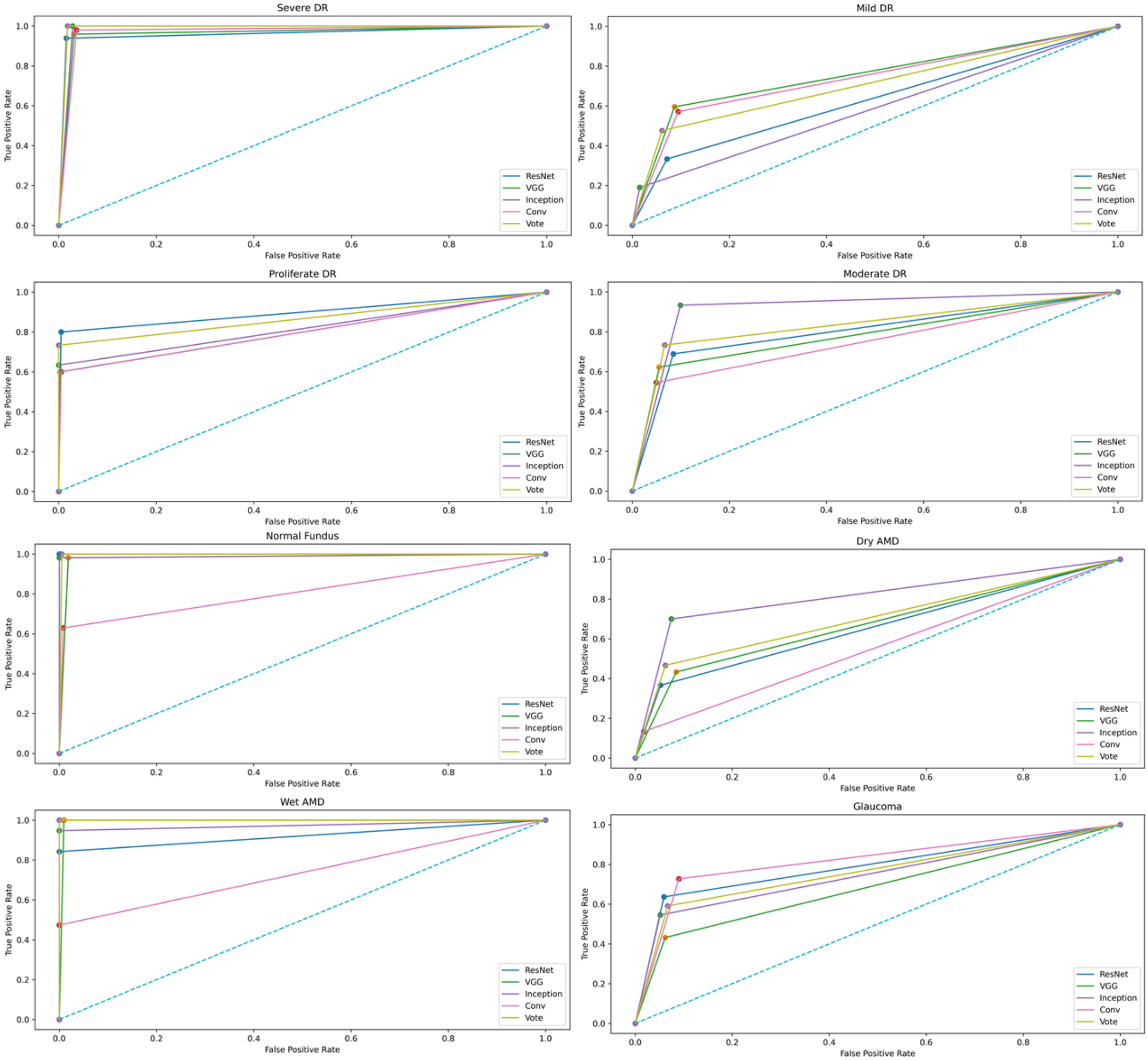
*Table 5 Densnet*

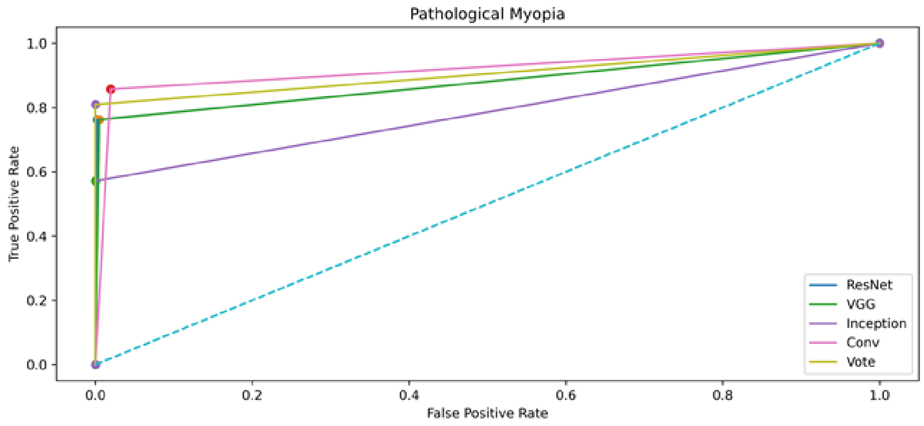
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Dry AMD | 0.44 | 0.53 | 0.48 | 30.00 |
| Glaucoma | 0.52 | 0.61 | 0.56 | 44.00 |
| Normal Fundus | 1.00 | 1.00 | 1.00 | 54.00 |
| Wet AMD | 1.00 | 1.00 | 1.00 | 19.00 |
| Mild DR | 0.49 | 0.60 | 0.54 | 42.00 |
| Moderate DR | 0.79 | 0.71 | 0.75 | 90.00 |
| Severe DR | 0.86 | 0.98 | 0.91 | 49.00 |
| Proliferate DR | 0.96 | 0.73 | 0.83 | 30.00 |
| Cataract | 0.84 | 0.88 | 0.86 | 24.00 |
| Hypertensive Retinopathy | 1.00 | 0.60 | 0.75 | 30.00 |
| Pathological Myopia | 1.00 | 0.86 | 0.92 | 21.00 |
| accuracy | 0.77 | 0.77 | 0.77 | 0.77 |
| macro avg | 0.81 | 0.77 | 0.78 | 433.00 |
| weighted avg | 0.79 | 0.77 | 0.77 | 433.00 |

*Table 6 vgg*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Dry AMD | 0.41 | 0.70 | 0.52 | 30.00 |
| Glaucoma | 0.55 | 0.55 | 0.55 | 44.00 |
| Normal Fundus | 1.00 | 0.98 | 0.99 | 54.00 |
| Wet AMD | 1.00 | 0.95 | 0.97 | 19.00 |
| Mild DR | 0.57 | 0.19 | 0.29 | 42.00 |
| Moderate DR | 0.71 | 0.93 | 0.81 | 90.00 |
| Severe DR | 0.82 | 1.00 | 0.90 | 49.00 |
| Proliferate DR | 1.00 | 0.63 | 0.78 | 30.00 |
| Cataract | 0.95 | 0.88 | 0.91 | 24.00 |
| Hypertensive Retinopathy | 0.64 | 0.47 | 0.54 | 30.00 |
| Pathological Myopia | 1.00 | 0.57 | 0.73 | 21.00 |
| accuracy | 0.75 | 0.75 | 0.75 | 0.75 |
| macro avg | 0.79 | 0.71 | 0.72 | 433.00 |
| weighted avg | 0.76 | 0.75 | 0.73 | 433.00 |

As the last evaluation metric, for the complex multi-label evaluation, we computed the popular area under the receiver operating characteristic (AUROC) curve (Fig 9). ROC curves are widely used in laboratory medicine to assess the diagnostic accuracy of a test, choose the optimal cut-off of a test, and compare the diagnostic accuracy of several tests. ROC curves also proved useful for the evaluation of machine learning techniques. It can be seen that in conditions with less dataset the ROC curve has a gap from the ideal form.


Severe DR
Mild DR
Proliferate DR
Moderate DR
Normal Fundus
Dry AMD
Wet AMD
Glaucoma
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ResNet
VGG
Inception
Conv
Vote


Pathological Myopia
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ResNet
VGG
Inception
Conv
Vote

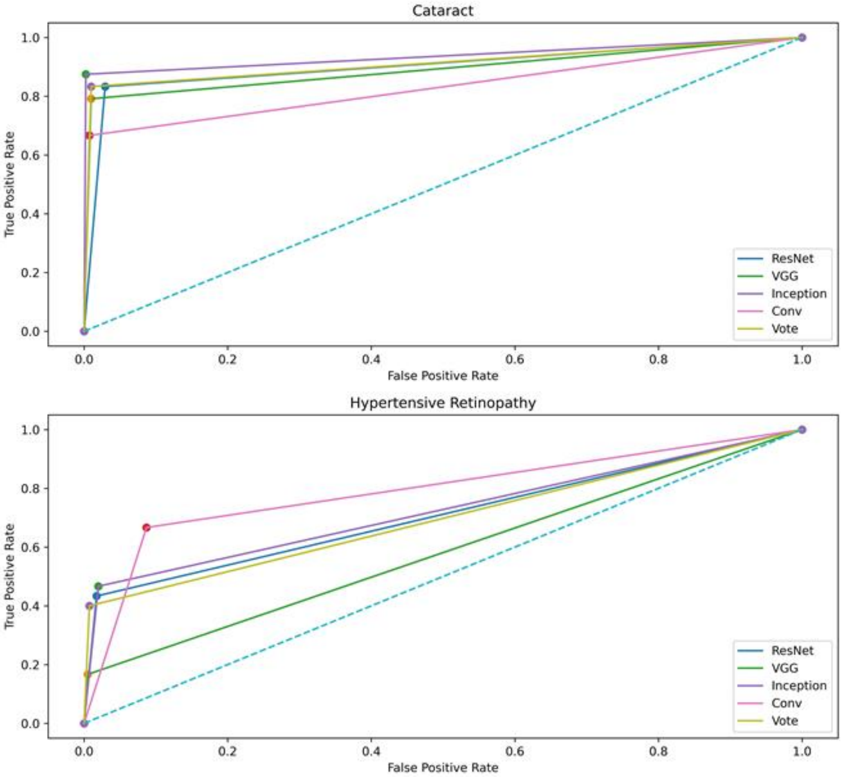


*Figure 10 - Receiver operating characteristic (ROC) curves for each model type applied in our pipeline for each disease. The ROC curves showing the individual model performance measured by the true positive and false-positive rates.*

Our multi-label classification model based on Ensemble and Transfer Learning can detect eleven types of fundus lesions at the same time.

## 3-2 Experiments and Improvements

Additionally, we experimented with using another fully-connected layer for training all model types. This resulted in inferior models for disease label classification, and, the extra fully-connected layer fitted disease risk detector models showed more overfitting with uneven performance. Further experimentation with loss functions for the disease risk detector models could provide the solution to avoid overfitting.

An important note would be the utilization of more data, especially for the classes like Wet and Dry AMD which form the lowest percentage of dataset images. To overcome this challenge, other publicly available datasets like Kaggle DR, RFMiD, IDRiD, Messidor, or APTOS are allowed to be used as additional training data. Our pipeline which was exclusively trained on the RFI dataset could be further improved with more retinal images of very rare conditions. Further research in retinal disease detection would include incorporating image cropping strategies, employing more architectures (especially with different input resolutions) to increase model ensembles, and utilizing specific retinal filters or retinal vessel segmentation as additional information to utilize for prediction.

## CONCLUSION

In this study, we introduced a powerful multi-disease detection pipeline for retinal imaging which exploits ensemble learning techniques to combine the predictions of various deep convolutional neural network models. Next to state-of-the-art strategies, such as transfer learning, Inception V3, ResNet 50, DenseNet 201, and VGG 19 we defined and used multiple convolutional neural network blocks to create an ensemble of models. With a stacking approach of defined blocks, we combined the knowledge of all neural network models to compute highly accurate and reliable retinal condition predictions. Next to an internal performance evaluation, we also proved the precision and comparability of our pipeline through a confusion matrix.