ENHANCING RETINAL DISEASE DETECTION USING RETINAL FUNDUS IMAGES BASED ON ENSEMBELING DEEP LEARNING HETEROGENEOUS MODELS

Dev Shrivastava **Shubhanshu Prasad**

Abstract

Automatic multi-disease detection models have showed promise in tackling the widespread issue of avoidable or undetected blindness and visual impairment. In this article, we propose a multi-disease detection pipeline for retinal disease identification, the pipeline was inspired from the work of Muller et. al. [3] did and the modifications were done on their proposed pipeline along with construction of a new data set for better training and testing of the models used. The pipeline uses ensemble techniques that combines the prediction power of many heterogeneous deep convolutional neural network models using ensemble learning. Modern techniques including transfer learning, class weighting, real-time picture augmentation, and the use of focal loss are all incorporated into the pipeline. Moreover, we use ensemble learning methods including stacked logistic regression models, bagging via 5-fold cross-validation, and heterogeneous deep learning models.We were able to validate and show excellent accuracy and dependability of our pipeline.

1 Introduction

In order to stop retinal abnormalities or ocular fundus diseases like diabetic retinopathy (DR), retinal detachment (RD), retinal artery occlusion (RAO), and many more from progressing and avert irreversible vision loss, early detection and precise diagnosis are essential. With the high prevalence of blindness and visual impairment worldwide, According to the World Health Organization (WHO), there are 2.2 billion blind and visually impaired persons in the globe. Out of these, at least 1 billion of these conditions could have been avoided or still need to be resolved, only if the detection methods were faster. This is especially prevalent in rural and remote areas, especially in developing countries, where there is insufficiency in ophthalmic service and a shortage of ophthalmologists, early detection and timely referral for treatment may not be available and hence the increasing number of cases.

Throughout the past ten years, the usage of clinical decision support (CDS) systems for diagnosis has increased. Modern deep learning models may become a robust tool to enable automatic and trustworthy medical image classification with astonishing accuracy equivalent to physicians. In retinal diseases, deep learning algorithms for AI-assisted diagnoses have been applied to screen for DR, AMD, retinopathy of prematurity, glaucoma, and papilledema.[natur paper]. These systems, however, focus on detection of one single disease at a time.

In this study we propose a pipeline that might be able to detect 37 different retinal disease using retinal fundus images. The ppipeline was inspired from the work of Muller et.al. The pipline utilizes power of ensemble learning to combine the strengths of deep learning models like convolutional neural networks, also applying state-of-the-art techniques like transfer learning, real-time augmentation for upsampling, and focal loss utilization.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

1.1 Previous work

In the work proposed by muller et. al.[3], the classification was done for 27 different disease. The pipline used consisted of creating ensemble of CNNs DenseNet201 and EfficientNetB4 for a disease detector system which classifys given image as either diseased or normal, with accuracy score of 98-99 percent. Furthermore, the disease classifier system which did the job of classifying the 27 disease was made by using four different architectures, ResNet152, InceptionV3, DenseNet201 and EfficientNetB4, each of which was stated to have an accuracy of above 95 percent. The ensembler was created by stacking binary logistic regression models, where the output of the CNNs was fed into the regression model, that was created for each seperate class. In total their pipeline had 6 CNN models, and 29 logistic regression models. The AUC score of the complete pipeline was over 0.95. The potential drawbacks in their study were the use of only one dataset that is RFMiD dataset, and the complexity of architecture which made the training of complete pipeline to take around 90 hours on a single NVIDIA TITAN RTX GPU.

In another study by Cen et.al.[2] A robust deep learning model was created which is able to classify 39 disease an conditions in total. The data set for the model is humongous and the validation through internal and external examination was very intensive. In total they had 249,620 images marked with 275,543 labels and were collected for training, validation and tests for the pipeline. The pipeline consisted of 4 groups of CNN and Mask R-CNN out of which 2 groups were custom made CNNs. In general they achieved a very high sensitivity and specificity of over 0.942 and 0.979 respectively. The study is different from ours in the sense that we have not considered condition like Silicon oil in eye, Laser spots, and Disc swelling and elevation.

2 Methodology

The implemented medical image classification pipeline can be summarized in the following core steps and is illustrate:

- Real-time image augmentation.
- Several deep learning model architectures.
- Class weighted Focal loss and up-sampling.
- Ensemble creation using Support Vector Regression and logistic regressions.

2.1 RFMiD 3.0

For this study we have custom build our own data set namely RFMiD3.0, which was created using Retinal Fundus Multi-Disease Image Dataset (RFMID)[4], RFMiD2.0[5], Indian Diabetic Retinopathy Image Dataset (IDRID)[6], Eye-Disease Data set[1], and some web scrapping.

After the collection and cleaning of data was done, a dataset of total 58 diseases was created and an image data of 11,372 images was formed (Table S1). Since the dataset is highly imbalanced and classification of all 58 diseases might not be possible. Hence the classification of 38 and one other category is made.

After data collection all the images were square padded so as to not lose the retinal image during resizing.

2.2 Pipeline

Our pipeline consists of first the data cleaning part where square padding was done, followed by upsampling by introducing flip, brightness variation, and variation in Hue. After this the images are first fed into a two CNN architecture consisting of DenseNet121, and ResNet152, thereafter a weighted average ensemble is created for disease risk detector which simply classifies the image into diseased or not. After This The image is fed into a series of CNNs which are two DensNet121, VGG-16, VGG-19, two DensNet201, ResNet152, EfficientNetB3, EfficientNetB7. All these CNNs are trained majorly on different subset of data depending upon the distribution. This step is crucial as the dataset is highly imbalanced getting high accuracy with just one classifier for multi-label problem will not be a fruitful attempt. Then these CNNs are stacked together by using their output to create

Table 1: Data Distribution

Disease	Train Set
DR	1098
CT	1038
GC	1007
MZ	343
ODC	303
TSLN	206
DM	400
DN	142
MYA	128
ARMD	105
BRVO	85
ODP	75
ODE	69
VS	7
Total	5006%

Table 2: Data Distribution

Disease	Train Set
LS	71
CRS	61
HR	50
RS	52
CSR	56
RT	38
CRVO	35
MS	33
CME	37
RPEC	34
CWS	22
AION	21
ERM	18
PTCR	8
Total	546

Table 3: Data Distribution

Disease	Train Set
AH	21
TV	20
EDN	16
MHL	19
TD	12
PT	18
RD	12
RP	10
ST	11
PRH	7
HTN	7
CF	8
MCA	9
Other	51
Total	263

a support vector regression ensemble. Another approach of using stacked logistic regression for individual classes is hypothesized for testing. The CNNs selected for the ensemble were first tested along with some other CNNs on part of training data as to compare accuracy of other networks and then selected, the training was done for 50 epochs for each CNN. The tried and tested networks with their accuracy is given in the following Tabel 4: CNN Architecture selection. After selecting the architectures, the disease risk detector was created, which is a binary classifier that classifies whether a given image is diseased or not. This classifier should be of high accuracy and for that purpose we selected the two highest scoring CNNs from tabel 4, i.e. DensNet121 and ResNet152. The classifiers were all initialized with weights of imagenet dataset, and last 8 layers were unfreezed for transfer learning application. Binnary crossentropy loss function was used along with sigmoid or tanh activation function. The Ensemble of these 2 classifiers was created using weighted average stacking, where DensNet121 was given weight 0.6 as it performed better during training, and ResNet152 was given weight 0.4.

Table 4: CNN Architecture selction

CNN	Train Accuracy	Test Accuracy
DensNet121	59.43%	76.74%
DensNet169	58.11%	71.23%
DensNet201	58.48%	75.40%
ResNet152	80.85%	91.18%
VGG-19	63.28%	79.82%

CNN	Train Accuracy	Test Accuracy
EfficientNetB7	61.11%	74.31%
EfficientNetB0	63.28%	79.82%
EfficientNetB3	59.55%	72.66%
InceptionV3	54.36%	59.16%
VGG-16	59.94%	81.82%

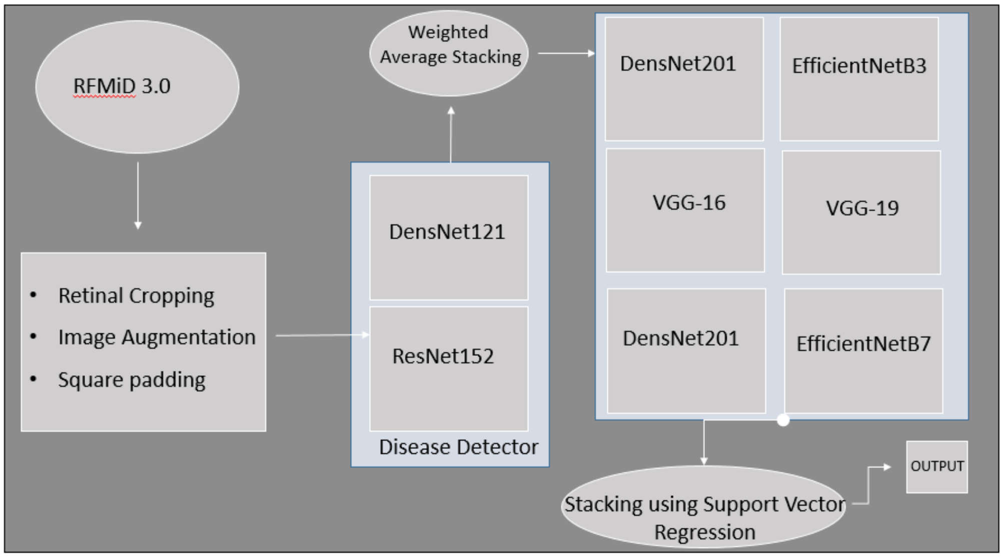

Figure 1: Pipeline

3 Results

In the proposed pipeline, the data collection and training and testing of individual neural networks is done, as well as the Disease Detector ensemble is also created.

Table 5: Disease Detector Ensemble

CNN	Train Accuracy	Test Accuracy
DensNet121	93.86%	96.17%
DensNet201	95.11%	91.13%
Ensemble	92.85%	95.63%

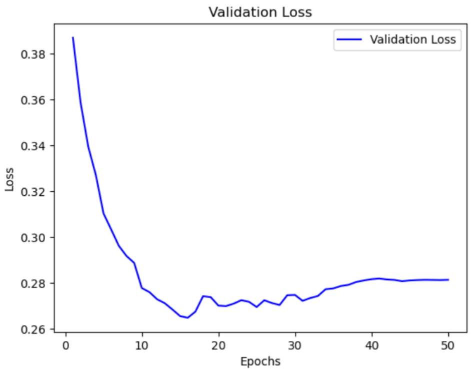

Figure 2: Val_Loss DensNet121

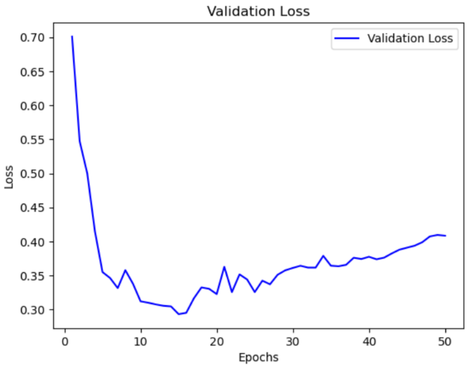

Figure 3: Val_Loss ResNet152

4 Furthermore

Other CNNs classifier were also trained but only on part of training dataset, training on whole dataset is not being possible in the provided lingraj server either due to space defficiency or due to server connection timed out or due to docker being restarted or NISER AP getting disconnected. The model history files were unfortunately removed during space management on my user because of which the training testing accuracy data is not available. In future we are going to first create the support vector regression stacking, and then running the entire dataset for the ensemble training and testing.

5

References

[1] 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, and et al., 2021.

[2] 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, and et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nature Communications*, 12, 2021.

[3] I. S.-R. Dominik Müller and F. Kramer. Multi-disease detection in retinal imaging based on ensembling heterogeneous deep learning models. 2021.

[4] S. Pachade, P. Porwal, D. Thulkar, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, L. Giancardo, G. Quellec, and F. Mériaudeau. Retinal fundus multi-disease image dataset (rfmid), 2020.

[5] S. Panchal, A. Naik, M. Kokare, S. Pachade, R. Naigaonkar, P. Phadnis, and A. Bhange. Retinal Fundus Multi-Disease Image Dataset (RFMiD) 2.0, Jan. 2023. RFMiD2.0 Dataset is classified into three sub- classes with disease distribution. The training set, validation set, and test set are three subfolders consisting of multi-labeled images. Each folder includes images and a .csv file containing labels. This distributed data can be utilized for the development of AI-based models.

[6] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, and F. Meriaudeau. Indian diabetic retinopathy image dataset (idrid), 2018.