Deep Ensemble Model for Retinal Diseases Detection and Classification

Amogh Jayant Dabholkar
ECE Department
Georgia Institute of Technology
Atlanta, USA
adabholkar6@gatech.edu

Aaryan Shah
ECE Department
Georgia Institute of Technology
Atlanta, USA
aaryan.s@gatech.edu

Parima Mehta
ECE Department
Georgia Institute of Technology
Atlanta, USA
parimam@gatech.edu

***Abstract*—Early diagnosis of retinal diseases is imperative to preempt severe vision impairment and blindness. This paper aimed to achieve high accuracy and generalization performance in predicting retinal diseases. Deep Learning models have proven to be extremely effective in solving convoluted problems in the area of Image Processing. Moreover, ensemble learning yields high generalization performance by reducing variance. Thus, a synthesis of transfer, ensemble, and deep learning was used in this paper to build an accurate and dependable model for the retinal disease detection and classification task. An in-depth analysis of the performance of widely used deep neural network architectures was made to build the disease-classifier set. Ensemble learning strategies like *bagging* via *k-fold cross-validation* and *stacking* of logistic regression models were adopted to create a collection of models to be used to make ultimate dependable predictions. Finally, the entire model was evaluated with the help of *Retinal Fundus Multi-Disease Image Dataset (RFMiD)*. The results clearly demonstrated the power of exploiting the synergy between deep learning and ensemble learning models and their useful application in retinal disease detection and classification.**

***Index Terms*—Deep Learning, Ensemble learning, Retinal Image Analysis, multi-Disease classification, transfer learning**

I. Introduction

The retina is the innermost, light-sensitive layer of our eyes, which is extremely delicate and responsible for translating the image of the visual world into electrical neural impulses to the brain to create visual perception[1]. Any damage to the retina has the potential to cause serious implications, diseases, and disabilities such as vision impairment and temporary or permanent loss of vision. According to the World Health Organization (WHO), there are at least 2.2 billion people in the world who have vision impairment and further in almost half of these cases, vision impairment is yet to be addressed or could have been avoided[2]. Hence, early detection and diagnosis of ocular pathologies have become of paramount importance to prevent retinal diseases, visual impairment, and blindness.

Over the past years, computerized clinical decision support systems (CDSS) have seen a rapid growth in their implementation to help clinicians in their complex decision-making[3]. Additionally, in recent years, the machine learning community has grown exponentially and several sophisticated deep learning frameworks such as convolutional neural networks (CNN) have been applied to detect a vast array of ophthalmological diseases using image classification[4][5]. However, several of these models are application-specific and lack the capabilities to detect rare retinal diseases with reasonable accuracy.

The screening of the eye through visualization of the retina, using the color fundus photos, presents a very unique opportunity to examine the systemic microcirculation in the retina in a non-invasive way. Detailed clinical observations of the retinal fundus features not only provides insightful information about the eye disease but also have led to the identification of early symptoms of diverse long-term diseases such as diabetes, stroke, and hypertension[6].

In this project, we have developed a machine learning model for automatic ocular disease classification of frequent diseases and rare pathologies using the Retinal Fundus Multi-disease Image Dataset (RFMiD) consisting of a total of 3200 fundus images captured using three different fundus cameras with 28 conditions annotated through adjudicated consensus of two senior retinal experts. There are several machine learning pipelines which are currently used for image classification. With the RFMiD dataset, we have studied, built, and developed an ensemble model which implements different deep learning models thus resulting in an ensemble that yields better results than either of those state-of-the-art deep models could individually.

The main aim was to expand on the work done by Dominic et al., 2021 [7]. We have explored and built on top of their existing machine learning pipeline which involves transfer learning followed by ensemble learning (for multiple deep learning models) and strived to improve upon it. Throughout the course of this project, we got exposure to studying and building upon the following machine learning principles:

- k-fold cross validation
- image augmentation
- multiple deep learning architectures such as ResNet152, InceptionV3, DenseNet201, and EfficientNetB4
- ensemble learning principles such as bagging, stacking, and boosting
- binary logistic regression for classification

Our ultimate goal was to experiment with different models, and try to improve the existing model by fine-tuning different parameters, and thus creating a novel machine learning model to detect and classify ocular pathologies.

Section II presents the methodology we adapted to complete the project successfully and Section III presents the results we were able to produce. Next, Section IV discusses a few of the key obstacles we faced and a brief overview on how we overcame them, and Section V concludes our work and discusses the future scope for this project. Lastly, Section VI highlights the individual contributions of the team members towards this project which is then followed by the References section.

II. Methodology

Figure 1 shows the methodology implemented for the entire project. The first step included procuring and cleaning the RFMiD dataset. Next, as part of the data preprocessing step, several techniques were implemented to create a clean and workable dataset. Techniques used during this step include image augmentation, up-sampling, cropping and padding, and normalization. Next in the pipeline included building and training twenty classification models and ten detection models. Further, ensemble learning techniques such as bagging and stacking were used in conjunction with stratified k-fold cross-validation and logistic regression to create a total of 29 models. Finally, the fully trained model was evaluated on the test dataset using performance metrics such as AUROC (Area Under Receiver Operating Characteristics) curves and mAP (Mean Average Precision) values.

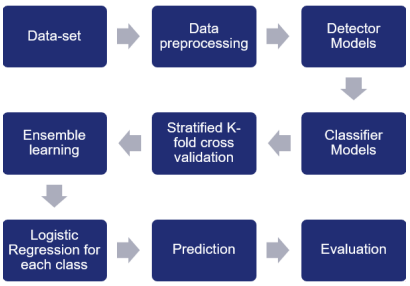

Fig. 1. Methodology used for the project

A. RFMiD Dataset

The RFMiD (Retinal Fundus Multi-disease Dataset) dataset, which is publicly available data, was used for the training and building the model for this project. The dataset consists of 3200 images out of which 1920 images were used as the training dataset. The images obtained from the RFMiD dataset consisted of 46 classes each representing rare and challenging diseases, which were thoroughly adjudicated by two senior retinal experts. To train the model, 27 of the most significant classes were used while the classes representing extremely rare diseases were clubbed in the class - 'OTHER'. Hence, the resulting training dataset consisted of 28 classes. Finally, the model was evaluated using a test dataset which consisted of 1280 images.

B. Data Preprocessing

To build a robust model and in general increase the data variability, the project involved implementing several preprocessing techniques.

First, image augmentation was applied to balance class distribution and real-time augmentation to obtain unique images in each epoch. Techniques such as flipping, rotation, and altering brightness, contrast, saturation, and hue were implemented. Using these techniques the dataset was prepared for image up-sampling.

Next, via image up-sampling, we ensured that each class/label occurred for a minimum threshold value. In the project, the threshold value was set to 100. Doing so, the training images increased from 1920 to 3354.

Further, post image up-sampling, we also applied square padding to avoid aspect ratio loss which might occur during posterior sizing and cropping to ensure that the fundus was at the center of the image. Both these techniques were applied to all the images individually, and the resulting images from these processes were further resized as per the requirements of the model where they were used as inputs.

Finally, we implemented value intensity normalization on the images before they were fed to the deep learning architecture. The intensities were zero-centered via the Z-Score normalization approach and to avoid data snooping, the training and test datasets were separately normalized.

C. Deep Learning Architecture

In today's world, in the domain of medical image classification, deep convolutional neural network models are unequivocally state-of-the-art. Our end-to-end pipeline combines two types of image classification models. All these models are pre-trained on the ImageNet dataset, followed by transfer learning with most frozen layers except for the classification head, combined with a fine-tuning method for unfrozen layers. The models are as follows:

- **Detector Models:** The disease risk detectors for binary classification into normal or abnormal images.
  - **DenseNet201:** Connects each layer to every other layer in a feed-forward fashion. Ensures strong gradient flow in forward and backward propagation.
  - **EfficientNetB4:** Constructed using Neural Architecture Search and Auto ML to optimize accuracy and efficiency (FLOPS). State of the art Top-1 and Top-5 accuracy on CIFAR-100 and others.
- **Classifier Models:** The disease label classifiers for multi-label annotation of abnormal images.
  - **ResNet152:** Made up of residual blocks which are made up of skip connections. This variant is used as a standard baseline for transfer learning.
  - **InceptionV3:** Made up of inception modules that consist of filter banks with all shapes. V3 optimizes the filter bank to factorize the larger sized filters
  - **DenseNet201:** Connects each layer to every other layer in a feed-forward fashion. Ensures strong gradient flow in forward and backward propagation.

– **EfficientNetB4:** Constructed using Neural Architecture Search and Auto ML to optimize accuracy and efficiency (FLOPS). State of the art Top-1 and Top-5 accuracy on CIFAR-100 and others.

D. Ensemble Learning Principles

1) Bagging: To improve the performance and accuracy of the machine learning model, bagging is generally applied. It is proven to be useful for the bias-variance tradeoff and helps reduce the variance of the predicted model. Further, in our case, it also helps us deal with high-dimensional data quite efficiently.

As a bagging approach, stratified 5-fold cross-validation was applied. By doing so, a large variety of models was created by training them on different subsets of the training data. It should be noted that the word stratified should be emphasized as it ensured that each fold is representative of all the strata of the data. Hence, the stratified approach ensures that each of the 28 classes is represented in each fold in the same proportion.

Through this approach, we not only achieved efficient use of the training data but also avoided overfitting and increased the reliability of the prediction. As a result of this approach, 20 disease label classifier models and 10 disease risk detector models were created.

2) Stacking: Stacking is generally used as it helps us choose the best model by combining the predictions from multiple machine learning models obtained on the same training dataset. Further, it helps us in using the usefulness of multiple models which are useful on the dataset in their unique ways. Given the creation of 30 models and a comparatively small dataset, stacking finds huge importance in this project.

For this project, at the end of the deep learning architecture, a binary logistic regression algorithm was applied to each class individually. Hence, the predictions obtained from all the 30 models were used in calculating the classification of each of the 28 classes. As a result, 29 distinct models were built, 28 for each of the classes and one for the detection of any disease in the image. In the end, individual class probabilities obtained from the binary logistic regression models were concatenated to build the final prediction.

It should be noted that all the logistic regression models were trained using stratified 5-fold cross-validation to avoid overfitting and training the models on the same images as seen in the deep learning architecture.

III. Results

A. Loss Curves

No signs of overfitting were observed for the classifiers as well as the detector models as can be seen in Fig. 2. The gray areas surrounding the fitting curves show the confidence intervals. To reduce the complexity, the loss curves have been averaged across all folds. During the training process, the strategy of choosing the model with the best validation set performance was used and resulted in powerful classifiers and detectors.

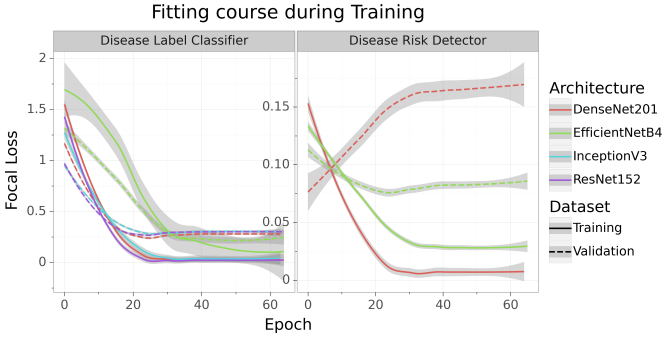

Fig. 2. Loss Curve

B. ROC Curves

Figure 3 shows the ROC (Receiver Operating Characteristics) curves for each of the model architectures which have been macro-averaged. The TPR (True Positive Rate) which is on the Y-Axis should preferably be more than FPR (False Positive Rate) on the X-Axis, which is the case in all our models. Ensembler makes the entire system near perfect based on the curve and this will also be reflected by the evaluation metrics in the following section.

C. Evaluation Metrics

We used AUROC (Area under ROC curve) and mAP (Mean Average Precision) to evaluate our models and ensemble performance. Both metrics were macro-averaged across cross-validation folds and classes. The results are as follows:

Model Type	Architecture	AUROC	mAP
Classifier	DenseNet 201	0.971515	0.914628
Classifier	EfficientNet B4	0.966678	0.908115
Classifier	ResNet 152	0.969704	0.911869
Classifier	Inception V3	0.921541	0.521429
Detector	DenseNet 201	0.968519	0.991890
Detector	EfficientNet B4	0.982201	0.996926
Ensembler	Logistic Regression	0.999507	0.997512

TABLE I
EVALUATION METRICS TABLE

The pipeline for multiple disease detection showed a very robust detection as well as classification performance. It also displayed the ability to detect rare retinal image conditions. While the classifier models individually were only able to achieve an AUROC of approximately 0.97 and a mAP of 0.91, the detectors showed extremely high predictive power of 0.98 AUROC and 0.99 mAP. Somehow, Inception V3 shows a poor mAP of 0.52 and a slightly reduced 0.92 AUROC.

In general, it is a complex task to train a multi-label classifier and detector especially with the main class imbalance issue between the conditions revealed a hard challenge for building a reliable model. The upsampling and usage of focal loss for countering the heavy imbalance in the dataset made a crucial contribution As expected from a robust model, all of the labels were accurately detected including the 'OTHER' class. Overall, the applied ensemble learning system and its

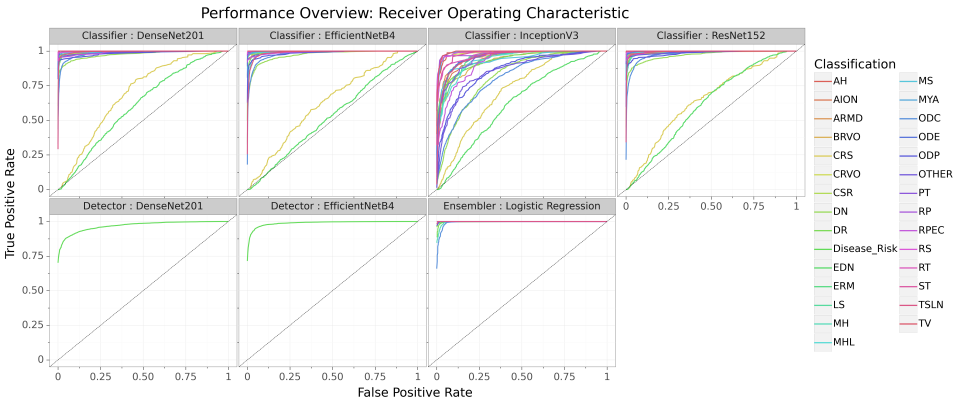

Fig. 3. ROC

strategies concluded in a crucial performance improvement as compared to the individual deep CNN models.

IV. Obstacles

During the course of the project, we faced several obstacles. This section sheds some light on a few of the key obstacles and the approaches taken to overcome them.

A. Data Procurement & Pre-processing

The original problem statement was first defined in a private competition on Kaggle, and the training and test dataset were not publicly available. Hence, we instead worked on the publicly available RFMiD dataset.

B. Dimensionality Reduction

The publicly available RFMiD dataset was huge and it contained 46 classes. Hence, we performed data cleaning and dimensionality reduction, thereby reducing the number of classes to 28 classes which represented 27 most significant classes and one class representing sparsely represented extremely rare diseases.

C. Time Complexity

Given the nature of the deep learning architecture and models involved in it, the entire training process was cumbersome as it required a huge amount of time given the low computing resources we possessed. As a result, we moved to GPU-based training on Google Colab, which oftentimes resulted in exceeding the daily limit of resource utilization.

V. Conclusions

In this project, we primarily created an automatic machine learning classification and detection model for ocular pathologies and rare diseases. We implemented a deep learning-based architecture to detect and classify these diseases and extensively applied several data pre-processing techniques such as image augmentation, up-sampling, and normalization among others to clean and prepare data for the deep learning pipeline. Further, we explored the use of ensemble learning principles of bagging and stacking. Via bagging, we applied stratified 5-fold cross-validation to facilitate learning by representing each class in equal proportion, and via stacking and the subsequent binary logistic regression model approach, we were able to create a final trained model by using the predictions obtained from several models thereby increasing the accuracy and reliability of the trained model. This entire approach led to a creation of a well-trained model which yielded good results on the test dataset while preventing overfitting and data snooping.

As a future scope, we believe that other sophisticated models could be used instead of the ones currently implemented in the pipeline. During the project timeline, as a first step towards that direction, we replaced the Inception V3 model with the MobileNet V3 model [15] as the mAP value for Inception V3 is 0.52, which can be observed from Table 1, and has some scope of improvement. Built by Google AI using Auto ML and Neural Architecture Search (NAS), it helps in efficient hyperparameter tuning, thereby reducing the human bias in the entire process. But due to the extensive computing resources required in the process, we were unable to train the model satisfactorily, under multiple hyperparameter settings, and as a result we obtained even more sub-optimal results when compared to the results obtained from the Inception V3 model.

Next, in the future we believe using more state-of-the-art deep learning models, and data augmentation and up-sampling techniques would yield better results. Lastly, we believe training the model on more publicly available datasets like Kaggle DR, IDRiD, Messidor or APTOS would make the model more reliable, accurate, and robust.

VI. Member Contributions

1) Literature review and survey - Aaryan
2) Study of proposed framework - Amogh and Parima
3) Implementation of individual algorithms - Amogh
4) Definition of performance metrics - Parima
5) Data Pre-processing - Amogh and Aaryan
6) Ensemble Learning principles - Parima
7) Model training and implementation - Aaryan
8) Improvisation and fine-tuning - Amogh and Aaryan
9) Presentation and project report - Aaryan, Amogh, Parima

REFERENCES

[1] Wikipedia, "Retina" https://en.wikipedia.org/wiki/Retina (accessed Oct. 19, 2021)

[2] World Health Organization, "Blindness and vision impairment." https://www.who.int/news-room/fact-sheets/detail/blindness-andvisual-impairment (accessed Oct. 19, 2021).

[3] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R.N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: benefits, risks, and strategies for success," npj Digital Medicine, vol. 3, no. 1. Nature Research, pp. 1–10, Dec. 01, 2020, doi: 10.1038/s41746-020-0221-y.

[4] G. Litjens et al., "A survey on deep learning in medical image analysis," Med. Image Anal., vol. 42, no. December 2012, pp. 60–88, 2017, doi: 10.1016/j.media.2017.07.005.

[5] A. Das, R. Giri, G. Chourasia and A. A. Bala, "Classification of Retinal Diseases Using Transfer Learning Approach," 2019 International Conference on Communication and Electronics Systems (ICCES), 2019, pp. 2080-2084, doi: 10.1109/ICCES45898.2019.9002415.

[6] S. Pachade et al., "Retinal Fundus Multi-Disease Image Dataset (RFMiD): A Dataset for Multi Disease Detection Research," Data, vol. 6, no. 2, p. 14, Feb. 2021, doi: 10.3390/data6020014.

[7] Dominik Müller, Iñaki Soto-Rey and Frank Kramer. (2021) Multi-Disease Detection in Retinal Imaging based on Ensembling Heterogeneous Deep Learning Models. arXiv e-print: https://arxiv.org/abs/2103.14660

[8] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, vol. 2017-January, pp. 2261–2269, Aug. 2016, Accessed: Feb. 27, 2021.[Online]. Available: http://arxiv.org/abs/1608.06993.

[9] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," 36th Int. Conf. Mach. Learn. ICML 2019, vol. 2019-June, pp. 10691–10700, May 2019,Accessed: Feb. 27, 2021

[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Dec. 2016, vol. 2016 - December, pp. 2818–2826, doi: 10.1109/CVPR.2016.308

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Dec. 2016, vol. 2016-December, pp. 770 778, doi:10.1109/CVPR.2016.90.

[12] G. Quellec, M. Lamard, P. H. Conze, P. Massin, and B. Cochener, "Automatic detection of rare pathologies in fundus photographs using few-shot learning," Med. Image Anal., vol. 61, p. 101660, Apr. 2020, doi: 10.1016/j.media.2020.101660

[13] 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, p. e0187336, Nov. 2017, doi: 10.1371/journal.pone.0187336.

[14] "Home - RIADD (ISBI-2021) - Grand Challenge." https://riadd.grand-challenge.org/Home/ (accessed Oct. 19, 2021).

[15] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for MobileNetV3. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019.

[16] Kaggle DR. Diabetic Retinopathy Detection

[17] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, Fabrice Meriaudeau, April 24, 2018, "Indian Diabetic Retinopathy Image Dataset (IDRiD)", IEEE Dataport, doi: https://dx.doi.org/10.21227/H25W98.

[18] Decencière et al.. Feedback on a publicly distributed database: the Messidor database. Image Analysis & Stereology, v. 33, n. 3, p. 231-234, aug. 2014. ISSN 1854-5165. Available at: http://www.ias-iss.org/ojs/IAS/article/view/1155 or http://dx.doi.org/10.5566/ias.1155.

[19] APTOS(2019). APTOS 2019 blindness detection.