An Analysis of Ensemble Methods for Deep Learning in the Diagnosis of Eye Diseases

Chia Wei Kit Samuel
University of Toronto
samuel.chia@mail.utoronto.ca

Jason Li
University of Toronto
lijason.li@mail.utoronto.ca

Ryoto Tamanoi
University of Toronto
r.tamanoi@mail.utoronto.ca

Abstract

Automated medical imaging is a increasingly important research area as manual diagnosis can be slow and expensive. However, in automated medical imaging it is often difficult to obtain large high-quality datasets. Ensemble methods have long given heuristic results in increasing model accuracy for when the options of more data and compute have been ruled out. There has also been evidence to suggest that sometimes increasing model width will improve performance while increasing model depth will not [9]. However, there has been little research in the best practices in applying ensemble techniques. In this paper we provide a rigorous analysis of different ensemble methods applied to different pre-trained models. Our main benchmark will be against the classification task of assigning disease labels to different eye fundi. We show that with stacking, there is a slight improvement in test performance compared to the benchmark.

1 Introduction

Supervised machine learning assigns labels to data points. This technology has applications in automated medical image analysis as explored in Shen et al. [13]. Often in this task, there does not exist datasets healthy in both size and quality. Furthermore, increasing model size or complexity does not seem to help with prediction performance [9]. Ensemble learning is a techniques used to improve test accuracy where dataset quality is the limiting factor [10]. We choose to apply a multitude of different ensemble techniques to various deep image models. We use ocular disease recognition to benchmark the performance of the newly introduced learning techniques.

2 Related Work

2.1 Applying deep learning to medical diagnosis

There has been rapid progress in the applications of deep learning to medical imaging [5]. Because of the breadth of medical imaging and the novelty of recent deep learning advancements, there are still many applications that have not yet been fully explored. Recent studies have shown promising results regarding deep learning for diagnosis of eye diseases [7, 2]. However, to the best of our knowledge, applying ensemble methods to the task of eyeball imaging has yet to be studied.

Preprint. Under review.

Table 1: Label count of images in training data

Diagnosis (label)	Count	Fraction of total data
Normal	1944	0.529
Diabetes	985	0.268
Glaucoma	149	0.041
Cataracts	191	0.052
Age related macular degeneration	174	0.047
Hypertension	75	0.020
Pathological myopia	158	0.043

2.1.1 Analysis of ensemble methods in deep learning

Ensemble methods have been a popular and effective way to improve the performance of machine learning models. Both theoretical research and empirical data strongly support the general effectiveness of ensemble learning improving the predictive power of deep learning models [12]. Despite its effectiveness, application has often been left as an afterthought compared to other parts of the machine learning model. Ensemble learning is guided by an overarching principle; combine a diverse set of models that make uncorrelated errors to harness what each different model has learnt. While there are a myriad of different ensemble methods, which ones are the best and how they should be employed is not obvious.

3 Data

The dataset by Ning Li et al. [9] is a collection of images of the eye labeled with 8 different eye conditions. In actual implementation of our models, we download the same dataset preprocessed to 512x512 colour images from [8]. The 6392 images in the dataset are labeled as either normal, labeled with the most common ailments, or labeled with "other ailments". For the purpose of this paper and to simplify the task of classification, we remove any multi-labeled images in the dataset and any eyes labeled as "other". This leaves us a total of 3676 images. The final label counts can be seen in table 1.

The left and right eyes are also labeled individually and no relationship is maintained between eyes from the same patient. During training we can split up left and right eyes helping prevent overfitting.

Data augmentation. With eyes classified as "normal" accounting for more than 50% of all samples and eyes with hypertension only making up 2% of all samples, the dataset is high unbalanced. To increase the breadth of our training data and reduce this imbalance, we applied some data augmentation on minority classes. The following transformations were applied to balance the classes in our training dataset: random resized crop, random rotation, color jitter, random horizontal and vertical flip, random clipped zoom, and random brightness enhance.

Finally, we applied an L_p normalization to all the images.

4 Models

4.1 Evaluation

If we combine all the not normal categories into one, we can get a binary classification for the dataset. Using this we can count true positives and false negatives. These counts are extremely important in automated medical diagnosis where usually a false negative is much worse than a false positive.

Let TP, TN, FP, FN be true positive, true negative, false positive, and false negative respectively. We define the following metrics for model evaluation [10].

$$\text{accuracy} = \frac{TP + TN}{TP + FN + TN + FN} \tag{1}$$

$$\text{F1} = \frac{2TP}{2TP + FP + FN} \tag{2}$$

Table 2: Pre-trained models after 10 epochs of fine-tuning without ensemble methods

Model	Test accuracy	F1 Score	Sensitivity
ResNet-50	0.686	0.562	0.495
ResNeXt_32x4d	0.683	0.503	0.426
DenseNet121	0.681	0.505	0.339
EfficientNet-B1	0.665	0.485	0.448
GoogleNet	0.654	0.442	0.354

Table 3: Performance of ensemble models on test set

Ensemble method	Test accuracy
Hard majority vote	0.690
Stacking	0.707
Bagging	0.673

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3}$$

We choose these metrics based on their ability to measure both type 1 and type 2 error.

4.2 Base models

In our analysis we will primarily use pre-trained models. We do this because of their performance and so we can benchmark our results against [9].

We apply our ensemble methods on the ResNet50 and ResNeX_32x4d [3]. We will also train on DenseNet121, EfficientNet-B1, and GoogleNet [4, 17, 16].

We fine tune these models on our training set for 10 epochs as suggested by [18]. The resulting training accuracies can be seen in table 2. For all models, we use cross entropy loss for training and we use Adam optimizer for gradient descent optimization [6].

These experiments reproduce the results in [9].

4.3 Ensemble Methods

We will focus on the following methods in our paper:

Hard majority vote. This ensemble method takes n models and trains them all on the dataset individually. Then, when it comes to test time, all models are predict the class of a test image independently. The final predicted label is the most frequently predicted label.

Bagging. In bagging, we draw subsets of the training data to train the models individually. Once training is done, we use hard majority vote to choose a final prediction. This should, in theory, reduce overfitting in the model because we are training on subsets of the entire dataset.

Stacking. Stacking trains n different models independently on the entire dataset. Then with the weights frozen. The n models have their results concatenated and put through a feed forward neural network. The weights on the network are learnt. Unlike hard majority vote and bagging, how the result is “voted” on is learnt instead of hard coded.

5 Results

Our ensemble experiment results are summarized in table 3.

Comparing the test accuracy with the base models, we see that there is an improvement of test accuracy of 2% for hard majority vote and a 3% improvement for stacking.

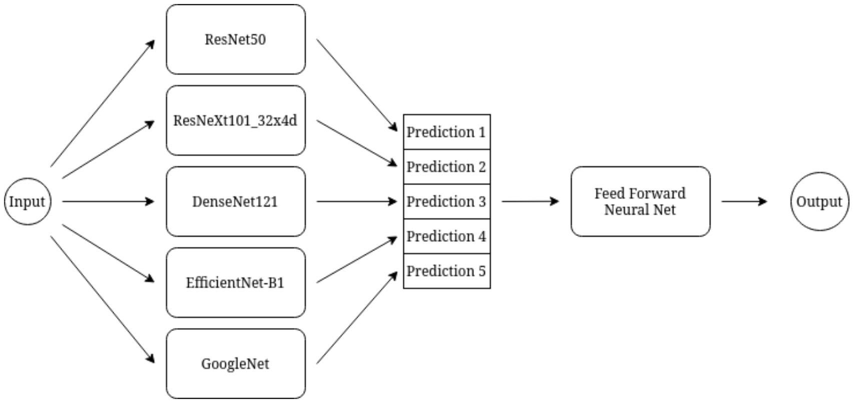

Figure 1: Stacking ensemble method diagram

(a) False positive

(b) False negative

Figure 2: Example of misclassified fundus under stacking model

The effect of data scarcity can be most clearly seen in our test accuracy for normal fundi and cataractic fundi. Where we have around 10% misclassification for both but we have an around 50% misclassification rate for the other labels. This suggests that the extra complexity of an ensemble model does not result in overfitting leading to decreased testing performance. Further ensemble method test results can be found in the appendix.

6 Discussion

Challenges. For many of the dataset labels, there were only a few training images. So augmenting and transforming the data was needed to allow for model training convergence. It was difficult to find the best base models to build the ensemble methods on top of because of the number of choices available and how long even fine-tuning a model takes. Also because the training took so long, it was difficult to optimize hyperparameters.

Extension. We hope to extend this project by analyzing more ensemble methods and to do more fine tuning on the hyperparameters. We would also like to extend our work to successfully train and analyze models for multi-class classification. Finally, we would like to incorporate both patient fundi in diagnosis to more closely match real world clinical practices in ophthalmology.

7 Conclusion

We have shown that shown a more complex base model does not necessarily imply better test performance with our study of the base models. Under the exploration of base models, we validated the results of Ning et al. [9]. We found slight improvement of test accuracy compared to the benchmark study in single class classification with hard majority vote and stacking. And stacking had the highest test accuracy out of any model we tested. By integrating multiple machine learning models together, we can surpass the performance of single base models.

References

[1] Xiao-Yan Gao et al. "Improving the accuracy for analyzing heart diseases prediction based on the ensemble method". In: *Complexity* 2021 (2021), pp. 1–10.

[2] Parampal S Grewal et al. "Deep learning in ophthalmology: a review". In: *Canadian Journal of Ophthalmology* 53.4 (2018), pp. 309–313.

[3] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385 (2015). arXiv: `1512.03385`. URL: `http://arxiv.org/abs/1512.03385`.

[4] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. "Densely Connected Convolutional Networks". In: *CoRR* abs/1608.06993 (2016). arXiv: `1608.06993`. URL: `http://arxiv.org/abs/1608.06993`.

[5] Jonghoon Kim, Jisu Hong, and Hyunjin Park. "Prospects of deep learning for medical imaging". In: *Precision and Future Medicine* 2.2 (2018), pp. 37–52.

[6] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: `1412.6980 [cs.LG]`.

[7] Yogesh Kumar and Surbhi Gupta. "Deep transfer learning approaches to predict glaucoma, cataract, choroidal neovascularization, diabetic macular edema, drusen and healthy eyes: an experimental review". In: *Archives of Computational Methods in Engineering* 30.1 (2023), pp. 521–541.

[8] Larxel. *Ocular Disease Recognition*. https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k.

[9] Ning Li et al. "A benchmark of ocular disease intelligent recognition: One shot for multi-disease detection". In: *Benchmarking, Measuring, and Optimizing: Third BenchCouncil International Symposium, Bench 2020, Virtual Event, November 15–16, 2020, Revised Selected Papers 3*. Springer. 2021, pp. 177–193.

[10] Dominik Müller, Iñaki Soto-Rey, and Frank Kramer. "An analysis on ensemble learning optimized medical image classification with deep convolutional neural networks". In: *Ieee Access* 10 (2022), pp. 66467–66480.

[11] Samiksha Pachade et al. "Retinal Fundus Multi-Disease Image Dataset (RFMiD): A Dataset for Multi-Disease Detection Research". In: *Data* 6.2 (2021). ISSN: 2306-5729. DOI: `10.3390/data6020014`. URL: `https://www.mdpi.com/2306-5729/6/2/14`.

[12] Omer Sagi and Lior Rokach. "Ensemble learning: A survey". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8.4 (2018), e1249.

[13] Dinggang Shen, Guorong Wu, and Heung-Il Suk. "Deep learning in medical image analysis". In: *Annual review of biomedical engineering* 19 (2017), pp. 221–248.

[14] Bin Sheng et al. "An overview of artificial intelligence in diabetic retinopathy and other ocular diseases". In: *Frontiers in Public Health* 10 (2022).

[15] Jayanthi Sivaswamy et al. "A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis". In: *JSM Biomedical Imaging Data Papers* 2.1 (2015), p. 1004.

[16] Christian Szegedy et al. "Going Deeper with Convolutions". In: *CoRR* abs/1409.4842 (2014). arXiv: `1409.4842`. URL: `http://arxiv.org/abs/1409.4842`.

[17] Mingxing Tan and Quoc V. Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". In: *CoRR* abs/1905.11946 (2019). arXiv: `1905.11946`. URL: `http://arxiv.org/abs/1905.11946`.

[18] Edna Chebet Too et al. "A comparative study of fine-tuning deep learning models for plant disease identification". In: *Computers and Electronics in Agriculture* 161 (2019), pp. 272–279.

A Appendix

A.1 Contributions

Samuel Chia

- Suggested and analyzed ensemble methods
- Researched and augmented the dataset

Jason Li

- Suggested and analyzed the eye disease classification
- Drafted and edited the paper
- Loaded, processed, and analyzed the model results

Ryota Tamanoi

- Fine-tuned the base models
- compared and analyzed which pre-trained model to decide which one to use
- Wrote code implementing the ensemble methods

A.2 Further results

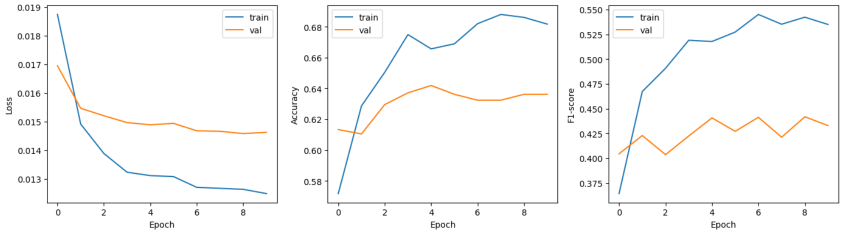

Figure 3: ResNet model training

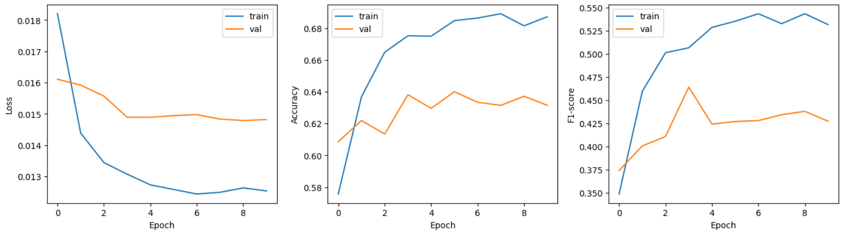

Figure 4: ResNeXt_32x4d model training

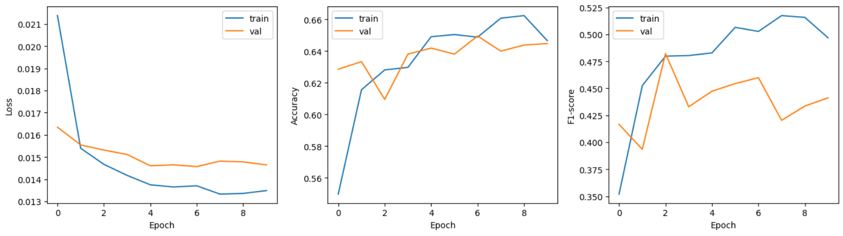

Figure 5: DenseNet121 model training

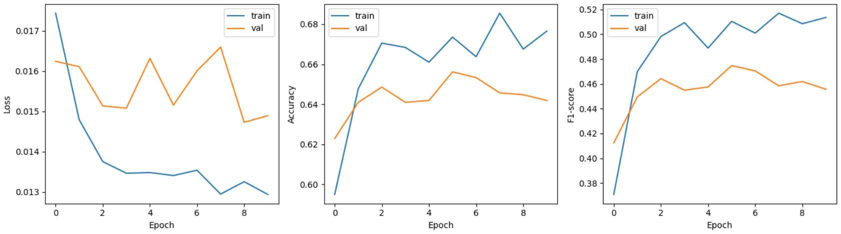

Figure 6: EfficientNet-B1 model training

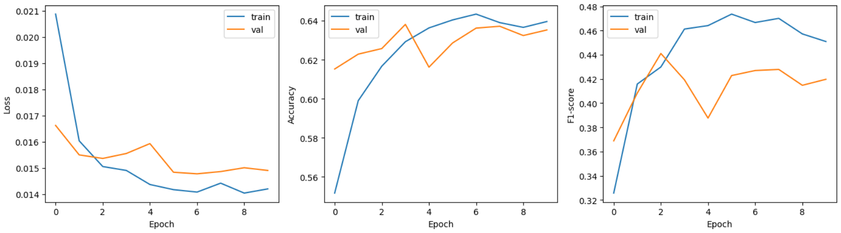

Figure 7: GoogleNet model training

```
------Reporting classification metrics for Ensemble Stacking model------
Sensitivity is 0.46255506607929514
Normal (N)
True Negatives: 267, 91.75%
False Positives: 24, 8.25%

Diabetes (D)
True Positives: 48, 36.36%
False Negatives: 82, 62.12%
Misclassified: 2, 1.52%

Glaucoma (G)
True Positives: 6, 20.00%
False Negatives: 24, 80.00%
Misclassified: 0, 0.00%

Cataract (C)
True Positives: 23, 95.83%
False Negatives: 1, 4.17%
Misclassified: 0, 0.00%

Age related Macular Degeneration (A)
True Positives: 10, 50.00%
False Negatives: 8, 40.00%
Misclassified: 2, 10.00%

Hypertension (H)
True Positives: 0, 0.00%
False Negatives: 5, 62.50%
Misclassified: 3, 37.50%

Pathological Myopia (M)
True Positives: 18, 85.71%
False Negatives: 2, 9.52%
Misclassified: 1, 4.76%
```

```
------Reporting classification metrics for ResNet50 model------
Sensitivity is 0.4590909090909091
Normal (N)
True Negatives: 261, 89.69%
False Positives: 30, 10.31%

Diabetes (D)
True Positives: 48, 36.36%
False Negatives: 83, 62.88%
Misclassified: 1, 0.76%

Glaucoma (G)
True Positives: 6, 20.00%
False Negatives: 22, 73.33%
Misclassified: 2, 6.67%

Cataract (C)
True Positives: 24, 100.00%
False Negatives: 0, 0.00%
Misclassified: 0, 0.00%

Age related Macular Degeneration (A)
True Positives: 4, 20.00%
False Negatives: 8, 40.00%
Misclassified: 8, 40.00%

Hypertension (H)
True Positives: 1, 12.50%
False Negatives: 4, 50.00%
Misclassified: 3, 37.50%

Pathological Myopia (M)
True Positives: 18, 85.71%
False Negatives: 2, 9.52%
Misclassified: 1, 4.76%

------Reporting classification metrics for ResNext101_32x4d model------
Sensitivity is 0.4260089686098655
Normal (N)
True Negatives: 267, 91.75%
False Positives: 24, 8.25%

Diabetes (D)
True Positives: 46, 34.85%
False Negatives: 84, 63.64%
Misclassified: 2, 1.52%

Glaucoma (G)
True Positives: 3, 10.00%
False Negatives: 26, 86.67%
Misclassified: 1, 3.33%

Cataract (C)
True Positives: 23, 95.83%
False Negatives: 1, 4.17%
Misclassified: 0, 0.00%

Age related Macular Degeneration (A)
True Positives: 4, 20.00%
```

```
False Negatives: 11, 55.00%
Misclassified: 5, 25.00%

Hypertension (H)
True Positives: 0, 0.00%
False Negatives: 5, 62.50%
Misclassified: 3, 37.50%

Pathological Myopia (M)
True Positives: 19, 90.48%
False Negatives: 1, 4.76%
Misclassified: 1, 4.76%

------Reporting classification metrics for DenseNet121 model------
Sensitivity is 0.3391304347826087
Normal (N)
True Negatives: 275, 94.50%
False Positives: 16, 5.50%

Diabetes (D)
True Positives: 31, 23.48%
False Negatives: 101, 76.52%
Misclassified: 0, 0.00%

Glaucoma (G)
True Positives: 2, 6.67%
False Negatives: 27, 90.00%
Misclassified: 1, 3.33%

Cataract (C)
True Positives: 23, 95.83%
False Negatives: 1, 4.17%
Misclassified: 0, 0.00%

Age related Macular Degeneration (A)
True Positives: 3, 15.00%
False Negatives: 16, 80.00%
Misclassified: 1, 5.00%

Hypertension (H)
True Positives: 1, 12.50%
False Negatives: 5, 62.50%
Misclassified: 2, 25.00%

Pathological Myopia (M)
True Positives: 18, 85.71%
False Negatives: 2, 9.52%
Misclassified: 1, 4.76%

------Reporting classification metrics for EfficientNet-B1 model------
Sensitivity is 0.4479638009049774
Normal (N)
True Negatives: 253, 86.94%
False Positives: 38, 13.06%

Diabetes (D)
True Positives: 50, 37.88%
False Negatives: 79, 59.85%
Misclassified: 3, 2.27%
```

```
Glaucoma (G)
True Positives: 1, 3.33%
False Negatives: 27, 90.00%
Misclassified: 2, 6.67%

Cataract (C)
True Positives: 24, 100.00%
False Negatives: 0, 0.00%
Misclassified: 0, 0.00%

Age related Macular Degeneration (A)
True Positives: 5, 25.00%
False Negatives: 9, 45.00%
Misclassified: 6, 30.00%

Hypertension (H)
True Positives: 0, 0.00%
False Negatives: 5, 62.50%
Misclassified: 3, 37.50%

Pathological Myopia (M)
True Positives: 19, 90.48%
False Negatives: 2, 9.52%
Misclassified: 0, 0.00%

------Reporting classification metrics for GoogleNet model------
Sensitivity is 0.35398230088495575
Normal (N)
True Negatives: 271, 93.13%
False Positives: 20, 6.87%

Diabetes (D)
True Positives: 37, 28.03%
False Negatives: 93, 70.45%
Misclassified: 2, 1.52%

Glaucoma (G)
True Positives: 4, 13.33%
False Negatives: 26, 86.67%
Misclassified: 0, 0.00%

Cataract (C)
True Positives: 23, 95.83%
False Negatives: 1, 4.17%
Misclassified: 0, 0.00%

Age related Macular Degeneration (A)
True Positives: 0, 0.00%
False Negatives: 16, 80.00%
Misclassified: 4, 20.00%

Hypertension (H)
True Positives: 0, 0.00%
False Negatives: 6, 75.00%
Misclassified: 2, 25.00%

Pathological Myopia (M)
True Positives: 16, 76.19%
```

```
False Negatives: 4, 19.05%
Misclassified: 1, 4.76%
```