Comparing RETFound with traditional computer vision models in eye disease classification

Amine Obeid
University of Texas at Austin
amine.obeid@utexas.edu

ABSTRACT

Foundation models have gained significant attention in the AI community in recent years because of their ability to handle large-scale data and adapt to a wide range of downstream tasks. In the medical domain, a notable example is RETFound, a foundational model introduced by Zhou et al. specifically for retinal images. According to its authors, RETFound demonstrates superior performance compared to existing models on a range of retinal imaging tasks. It has also shown in several studies its efficiency and good performance, pushing the edge of traditional models.

In this paper, we aim to further evaluate the effectiveness of RETFound by fine-tuning it on a publicly available retinal image dataset and evaluating the results compared to those of three well-known traditional computer vision models: ResNet50, DenseNet121, and EfficientNet_b0. To do so, we train all the models on the same training and validation dataset and using similar hyper-parameters and computer architecture. Then, we test all four models on a test dataset using three standard metrics: Accuracy, F1 score, and ROC-AUC. Our results show that finetuned RETFound didn't outperform the traditional models but achieved similar results to them. These findings could be explained as a limitation of the foundation model or as the impact we had while conducting this study due to the limited time constraint such as limited variety of hyper-parameters tested.

ACM Reference Format:
Amine Obeid. 2025. Comparing RETFound with traditional computer vision models in eye disease classification. In . ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Foundation models have gained significant attention in the AI community in recent years due to their ability of handling large-scale data and adapt to a wide range of downstream tasks. These models, often built on vast datasets, are typically trained using self-supervised learning techniques. They show remarkable versatility, and performance across diverse applications, as their primary use, as the name suggests, is acting as a foundation which can be fine-tuned and adapted for a variety of downstream tasks. [1].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference'17, July 2017, Washington, DC, USA
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

In the medical domain, a notable example is RETFound, a foundational model introduced by Zhou et al. specifically trained on a large dataset consisted of retinal images. Retinal images are beneficial for the diagnosis, prognosis, and systemic diseases of a wide range of diseases ranging from glaucoma to systemic conditions such as heart failure. RETFound was designed by learning rich representations from unlabeled retinal images. This allows it to then be fine-tuned for various tasks, such as classification or segmentation using smaller labeled datasets of retinal images. According to its authors, RETFound demonstrates superior performance compared to existing models on a range of retinal imaging tasks. However, while RETFound has shown promising results; we thought that testing RETFound's performance on a different dataset and comparing it to well-known models would be interesting. [5]

In this paper, we aim to further evaluate the effectiveness of RETFound by fine-tuning it on a publicly available retinal image dataset sourced from Kaggle consisting of approximately 4000 images. Our downstream task is image classification into one of three disease categories or as no disease. Then, to assess RETFound's performance, we compare it to three of the most famous computer vision models, all trained on the same dataset and under identical conditions, and evaluated using several standard evaluation metrics.

The rest of this paper will be divided as follows: Section 2 will discuss some related work on comparing RETFound to existing models that we have found. Then, Section 3 will discuss the methodology we have used and which is also reflected in Figure 1. Moreover, Section 4 will discuss our experimental results and an analysis of these results and lastly, we will conclude our findings in Section 5 and discuss the limitations of our work.

2 RELATED WORK

Several studies have explored the validation of RETFound's performance by fine-tuning it on novel tasks and datasets, and then comparing it with existing traditional models to evaluate its performance.

Chen et al. in a recent study have compared the performance of RETFound to that of Vision Transformers(ViT) and VGG16 models, focusing on a task different than ours. Their task was to predict the CDR and the average RNFL thickness from fundus photos. They found that RETFound outperforms both models without being explicitly trained on the task, highlighting the foundation model's ability to generalize across retinal images tasks. [2]

In a different study, Kuo et al. did similar work to ours, trying to test the performance of RETFound further on a different task. The idea was to classify normal and abnormal cases from OCT scans. The authors didn't just find that RETFound outperforms ResNet-50 and

ViT-L on a small imbalanced dataset, but even match their results with only 27% of the data, which proves that the foundation model can have a high performance even with limited data. [3]
Furthermore, Yow et al. have also tackled this problem, perhaps in the most similar fashion to our study. They did a disease classification and found that traditional computer vision algorithms proved similar to RETFound on large datasets, while the latter outperforms the former on small datasets, which similar to the study mentioned before, shows that RETFound can maintain a satisfying performance even if the dataset is relatively small. [4]
These studies provide important insights about the strengths of RETFound, particularly in its ability to outperform traditional models in various scenarios. However, it doesn't prevent us from testing these findings on a new retinal image classification task and new dataset and compare it to three traditional models.

3 METHODOLOGY

In this section, we will describe the methods we have used in the study, which is summarized in Figure 1.

3.1 Data

We first start by choosing an appropriate dataset for our task. We decided to pick a publicly available dataset on Kaggle that consists of 4217 retinal images organized into four classes in the following way: 1038 images are labeled as "Cataract", 1098 images as "Diabetic Retinopathy", 1007 images as "Glaucoma", and 1074 images as "Normal". We found this dataset to be ideal for our project since it is almost perfectly balanced and represents not only a binary classification, but a multiclass one.
To prepare the data, we form a script to organize the data in the conventional way of three subsets: training, validation, and testing. The dataset was split into 80% for training, 10% for validation, and 10% for testing. This is conventional as it allows the models to learn from the majority of the data using the training set. The validation set is used to test the model's performances after each epoch, and then the epoch on which the model has performed the best will be saved so we take this model for evaluation as it will most probably perform best on the testing dataset too. As for the testing dataset, it is unseen by the model until the evaluation step in order to ensure that the model is as unbiased as possible.
Before feeding the data to the model, it has been subject to some preprocessing. The images were kept on their original size of 256x256, which is not really conventional as we discovered after training the models, that resizing the images to 224x224 would've been more conventional and would allow for more models to be used in the study. The images were then transformed into tensors, since deep learning models only understand numerical form of data. Lastly, the images were normalized and standardized using the ImageNet mean and standard deviation values, which are conventional values to use. This last step of preprocessing helps reduce bias.

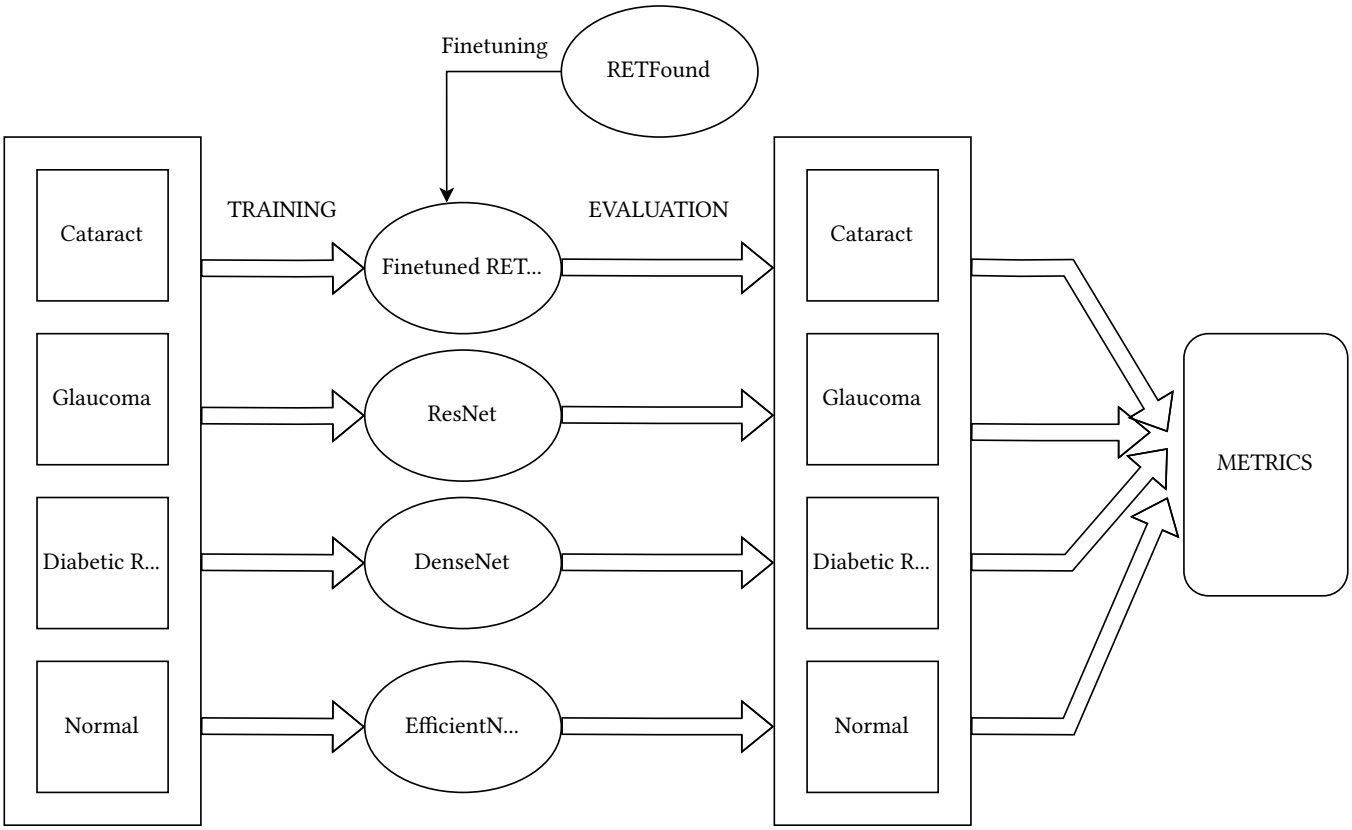

Figure 1: A figure summarizing our methodology starting from training the models on the training dataset, then finetuning RETFound to evaluating the models on the testing dataset.

3.2 Models Training

3.2.1 RETFound. In order to train RETFound, we have followed the instructions provided on the authors' github repository [5], which includes detailed instructions for fine-tuning the model on a downstream task. We have chosen the RETFound_mae model specific for CFP images, since our dataset is formed of retinal CFP images. We have then finetuned it as specified in the instructions on our dataset, ensuring that pretrained weights were modified to handle the representation of our dataset. Some specific hyper-parameters for the fine-tuning process of RETFound are a layer decay of 0.65, a weight decay of 0.05, and a drop path of 0.2. The rest of the hyper-parameters will be mentioned in Section 3.2.3.

3.2.2 Traditional models. For the rest of the models, we will be using ResNet50, DenseNet121, and EfficientNet_b0 in order to compare their performances to that of the fine-tuned RETFound model so we can evaluate it. These three models were chosen for their proven effectiveness in the domain of Computer Vision. Therefore, these models will serve as our baselines. ResNet50 and DenseNet121 are imported from the famous PyTorch library, specifically torchvision, while EfficientNet will be imported from timm.
For all three models, we use the Cross Entropy Loss as our loss function, which is commonly used for multi-classification tasks. Also, we use AdamW as our optimizer, a variant of the famous Adam algorithm that also implements weight decay. Both our choices of the loss function and the optimizer are done to match the RETFound original loss function and optimizer, ensuring consistence and fair comparison. We use PyTorch to train the models.
We will iterate over the three models using the same loop. Each model will go through the training loop and will be evaluated as we save the results before passing to the next model. During the training loop, after each training epoch, the model in the iteration will be evaluated on the validation dataset, and the best model so far in terms of the accuracy metric will be saved. Thus, when evaluating the model, the best model parameters on the validation dataset among all the epochs will be considered, to ensure the evaluation step is done using the best model.

3.2.3 Hyper-parameters. To ensure the comparison is as fair as possible, we try to make sure that the models have the same sets of hyper-parameters. For both RETFound and the traditional models we will have the batch size set at 16, the starting learning rate at $1e^{-3}$ and 10 training epochs, so we allow the same time for all models to converge to the optimum.

3.3 Models Evaluation

To compare the performance of the models, we evaluate the best checkpoint of each model, which is decided using the validation dataset on the testing dataset that is unseen by the models before this stage.
To conduct the evaluation we use three widely used metrics in the field:

(1) Accuracy: Calculates the proportion of correct predictions out of all predictions.
(2) F1 Score: Takes into account both Precision and Recall.
(3) ROC-AUC: Gives an overall measure of how well is the model differentiating between classes.

By using different metrics, we aim to obtain a bigger overview of the models' respective performances.

4 RESULTS

4.1 Performance Results

As illustrated in Figure 2, RETFound achieved a solid performance across all three metrics. Specifically, it achieved an accuracy of 0.89, an F1 score of 0.89 and an ROC-AUC of 0.98. These results are similar to those of ResNet50, achieving the same results on accuracy and F1 score, while slightly being outperformed by RETFound on the ROC-AUC metric as ResNet50 has scored 0.97.
However, DenseNet121 and EfficientNet_b0 both outperformed RETFound on all metrics. Both models achieved the same results on the three metrics with 0.92 on both accuracy and F1 score and 0.99 on ROC-AUC. These results show that DenseNet121 and EfficientNet_b0 have outperformed RETFound by a small margin.
While the results achieved by RETFound are high and comparable to those of traditional models like ResNet50 which it slightly outperformed, it falls short of our expectations which were that RETFound would outperform traditional computer vision models. Given the high expectations surrounding foundation models, in their ability to generalize across a variety of tasks, we anticipated a better performances over traditional computer vision models. The expectations we had come from RETFound being specifically tailored for tasks related to retinal images.
Briefly, despite it scoring satisfying results similar to those of the three traditional models we have chosen, RETFound couldn't meet our expectations of outperforming these models.

4.2 Training Time

An important consideration in our comparison of the foundation model and the traditional models is the training and evaluation time each of them needed. Traditional models such as ResNet50, DenseNet121, and EfficientNet_b0 showed significantly faster training and evaluation times compared to RETFound. The training and evaluation of all three traditional models combined took approximately two hours. RETFound, on the other hand, required around nine hours to complete both fine-tuning and evaluation on our dataset.
This huge difference in the time needed for completing training and evaluation between RETFound and our three traditional models is not reflected in our performance results. As we expect that the longer time needed is worth it and could be compromised with an increase in the performance of traditional models, that wasn't the case in our study as we have discussed in Section 4.1.
This longer training time for RETFound could impact its use in several applications.

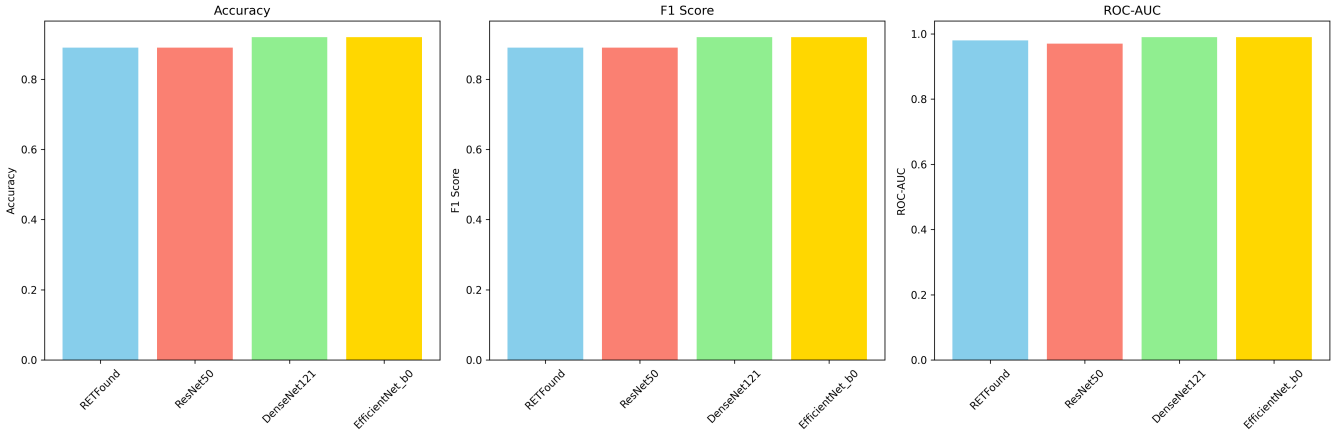

Figure 2: A figure summarizing the results we have gotten on our dataset. Three metrics were used from the left to the right: Accuracy, F1, and ROC-AUC on all 4 models: RETFound in blue, ResNet in red, DenseNet in green and EfficientNet in yellow

5 CONCLUSION

In this work, we evaluated the performance of RETFound fine-tuned on a multi-class task of eye diseases using retinal images for which it was specifically customized and is expected to have great results. In order to do so, we have compared it to three traditional models: ResNet50, DenseNet121, and EfficientNet_b0 on the same dataset and while trying to have the same training environment. Our results show that RETFound did not outperform the traditional models, which goes against our expectations. More in detail, it has barely outperformed ResNet50 but was slightly outperformed by DenseNet121 and EfficientNet_b0 even though RETFound took more than four times more time to train and evaluate than all three models combined.

However, it is undeniable that RETFound have shown better performances than traditional models in other studies as mentioned in section 2. The difference between our results and related work could be the consequence of several factors. First, due to our time constraint on that study, we were unable to test different sets of hyper-parameters as we have fixed one set and tested it for all the models. This is inconvenient and could result in not getting to the full potential of the model as a more thorough hyper-parameter search could have led to better RETFound results.

Another possible improvement would be to dig deeper into the metrics significance and use more metrics that fit better to our problem so we understand exactly where each model's strengths and weaknesses are and tweak each model accordingly.

While RETFound have shown good results, traditional methods did slightly outperform it despite being trained and evaluated faster which make them, based on our results, more practical.

REFERENCES

[1] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258 [cs.LG] https://arxiv.org/abs/2108.07258

[2] Maggie S. Chen, Rohith Ravindranath, Robert Chang, Yukun Zhou, Pearse A. Keane, and Sophia Y. Wang. 2025. Independent Evaluation of RETFound Foundation Model's Performance on Optic Nerve Analysis Using Fundus Photography. *Ophthalmology Science* 5, 3 (2025), 100720. https://doi.org/10.1016/j.xops.2025.100720

[3] David Kuo, Qitong Gao, Dev Patel, Miroslav Pajic, and Majda Hadziahmetovic. 2025. How Foundational Is the Retina Foundation Model? Estimating RETFound's Label Efficiency on Binary Classification of Normal versus Abnormal OCT Images. *Ophthalmology Science* 5, 3 (2025), 100707. https://doi.org/10.1016/j.xops.2025.100707

[4] Samantha Min Er Yew, Xiaofeng Lei, Jocelyn Hui Lin Goh, Yibing Chen, Sahana Srinivasan, Miao li Chee, Krithi Pushpanathan, Ke Zou, Qingshan Hou, Zhi Da Soh, Cancan Xue, Marco Chak Yan Yu, Charumathi Sabanayagam, E Shyong Tai, Xueling Sim, Yaxing Wang, Jost B. Jonas, Vinay Nangia, Gabriel Dawei Yang, Emma Anran Ran, Carol Yim-Lui Cheung, Yangqin Feng, Jun Zhou, Rick Siow Mong Goh, Yukun Zhou, Pearse A. Keane, Yong Liu, Ching-Yu Cheng, and Yih-Chung Tham. 2025. Are Traditional Deep Learning Model Approaches as Effective as a Retinal-Specific Foundation Model for Ocular and Systemic Disease Detection? arXiv:2501.12016 [cs.CV] https://arxiv.org/abs/2501.12016

[5] Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. 2023. A foundation model for generalizable disease detection from retinal images. *Nature* 622, 7981 (2023), 156–163.