

Retinal Classification in the Transformer Age

Derek Burrola
djb4994@utexas.edu
University of Texas at Austin
Austin, Texas, USA

Retinal image of: Glaucoma | Prediction: Glaucoma

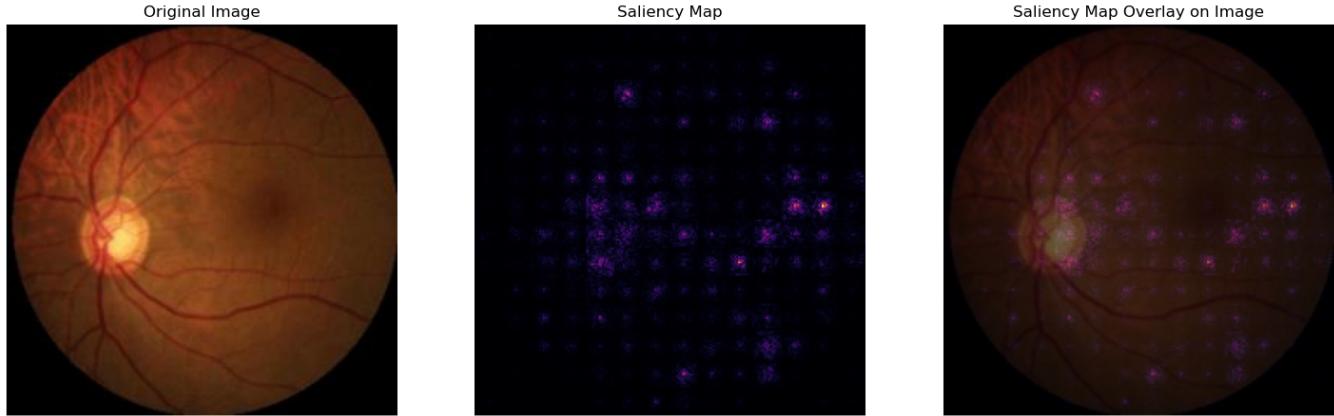


Figure 1: Retinal Image with Salience Map

Abstract

Early detection of eye diseases is critical but often limited due to the shortage of ophthalmologists and the high cost of screenings. This study explores the application of Vision Transformer (ViTs) as a preliminary classification of human retinal images. This research builds on previous studies that used convolutional neural networks (CNNs). Through transfer learning from pre-trained ViT models trained on ImageNet-21k, the model is fine-tuned on public datasets to show improvements in precision, recall, and overall accuracy through the use of the ViT architecture reaching accuracies of 94.98%. The study briefly explores the gained benefits of using ViT over CNNs including attention mechanism and how they can be applied to images to further facilitate the work of ophthalmologists.

Keywords

ViT, retinal, image classification, fundus, vision transformer, medical, AI, medical image classification, eye classification, disease, retinal disease, retinal disease image classification

ACM Reference Format:

Derek Burrola. 2025. Retinal Classification in the Transformer Age. In . ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 Introduction

Eye-related diseases globally account for millions of affected patients and their treatment exceeds \$400 billion in the United States alone [4]. Many of these diseases are either preventable if detected early and can be detected through relatively inexpensive methods such as eye screenings [7]. While these diseases are easy to detect, and most are simple to treat and even prevent if detected early, the amount of well-trained ophthalmologists is experiencing shortages and the cost for these screenings has become too expensive for most people to consider [3]. In comparison, a computer trained eye screening model can be created at a relatively inexpensive price. This research paper builds on the learned knowledge from Burrola [2], where pre-trained CNNs were used to classify fundus images.

Vision transformers (ViT) have garnered an increase in popularity since the success of Dosovitsky et al. [6]. Traditional methods of image classification have relied mostly on convolutional neural networks (CNNs) as can be seen in most research papers regarding image classification. CNNs can perform optimally even on limited datasets, even with the need to use pre-trained models. These models however lack a stable way of decrypting what the model is thinking and how it is classifying the images. Although that is not the focus of this research, it is a vast improvement that will be leveraged and used in the future of medical image research with artificial intelligence (AI).

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, Washington, DC, USA

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

A massive improvement of ViT over CNNs is they include an attention mechanism just like they do in Natural Language Processing (NLP). This attention mechanism allows the model to capture relationships in the images in short and long-range. Since the image is treated as tokens, self-attention can capture relationships between tokens which can be used for visualizing where the model is giving focus. This simple mechanism can allow a trained professional to look at an image and quickly focus their attention on aspects of the image that can lead to a diagnosis and even allow professional to notice something that they may otherwise miss.

Our goal is to create a model that can be used across medical facilities to quickly give a pre-diagnosis for patients in a matter of seconds at an inexpensive price. Early detection of diseases should be possible to do even at home, using the camera on a phone and a well-trained model. This goal is in progress and this research will not achieve that goal, though it will lead one step closer to making early detection from home a reality. For this paper, the goal will be to train a model on images of fundus known to have Glaucoma, Diabetic Retinopathy, or Cataracts and improve on previous accuracies done by CNNs [1] [3].

2 Related Work

The use of Vision transformers has been used some info about how its used in a bunch of other thigns here ViTs have been shown to work surprisingly well in comparison to CNNs [6] and have not been fully explored in the medical field [8]. ViTs require significantly more data than CNNs do to yield good results as compared to CNNs. CNNs can train on smaller data sets, but it is common for them to use transfer learning as a starting point. By that same methodology ViTs can take advantage of transfer learning and show that their improve their performance by doing so [8]

Work done by Sai Divya Battalapalli showed a CNN model trained on 7 epochs could yield results of approximately 80% accuracy [1]. This is the comparison benchmark this paper is attempting to improve upon by utilizing the same dataset, epoch count, and evaluation methods, where applicable.

3 Methodology

3.1 Model

Due to ViTs requiring large amounts of data to be trained on, we will be using a process called transfer learning. Transfer learning is the process of taking the weights of an existing/pre-trained model and then training/fine-tuning from that starting point on a new set of data for faster learning and improved performance. This technique allows smaller datasets to be useful where normally a ViT would only become useful and efficient with datasets like Google's JFT-300M, which contains over 300 million images [8]. Medical pre-trained models are extremely rare to find available publicly due to all the regulations around patient information, therefore we opted to use a generic image model. We chose the "vit_base_patch16_224" PyTorch Image Model. This model was trained on the ImageNet-21k dataset using approximately 14 million images of size 224x224 and used 16x16 size patches during its training [9].

3.2 Dataset

We will be using the same retinal image dataset Battalapalli used when training their CNN model [3]. The dataset is called "Eye diseases classification". It contains 2-dimensional colored retinal images taken from various datasets including IDRiD, Ocular recognition, HRF [5]. The is well-balanced with approximately 1,000 images of the following classes: "Normal", "Cataracts", "Glaucoma", and "Diabetic Retinopathy". The dataset on its own, though relatively small, is expected to be large enough for the ViT model to get performance equal to or greater than Battalapalli's 80% accuracy. If need be, data augmentation is considered to increase the amount of training data.

We prepare the data for training by splitting each class into using 70% of the set for training, 15% for validation (this will be used during training), and 15% for testing. The images from all the classes will be in a single training group and the model will learn those classes at the same time. The images will then be transformed into a size 224x224 pixels to make the images compatible the pre-trained model, the images will be converted to tensors and finally they will be normalized by the mean and standard deviation of the ImageNet dataset to increase the model's likelihood of good convergence. No data augmentation on the dataset, and the images color scheme was not altered.

3.3 Training

For training, we used cross entropy loss, AdamW as the optimizer, CosineAnnealingLR scheduler to fine-tune the learning rate to ensure we could reach the best performance during training. To have our results be comparable to Battalapalli's training, we opted to train for 7 epochs. We optimized the rest of our hyperparameters to train on a learning rate of 3e-5 using batches of 64. The best results came from freezing half of the layers of the model (block0-block5). Once the hyperparameters were optimized, the model was trained 30 different times with a new randomized training, validation, and testing set for each run to achieve the results.

3.4 Validation

For validation, we used the remaining 30% of the dataset and split it into 2 groups of 15%. The first group was run during the training to test the model's performance. The information gathered from the validation allowed us to make decisions about which models were performing best at with what loss. This helped narrow the learning rate and batch sizes early on.

3.5 Testing

For testing we used the last 15% of the dataset. This set was used to test the model using batches of 64, and metrics were gathered including accuracy, precision, recall, f1-scores across each class. To ensure statistically valid results and as close to true accuracy as possible, the model ran 30 times each with a different training, validation, and test set. Metrics were collected from each run. The pre-trained model was also tested without training to get a baseline of how well the model alone would perform. The accuracy of the pre-trained model ranged from 15% to 35% with an average accuracy of 24.55% across 30 runs (all different test sets).

After training our ViT model's accuracies ranged from 92% to 94.98% accuracy with an average of 93.74% across 30 runs.

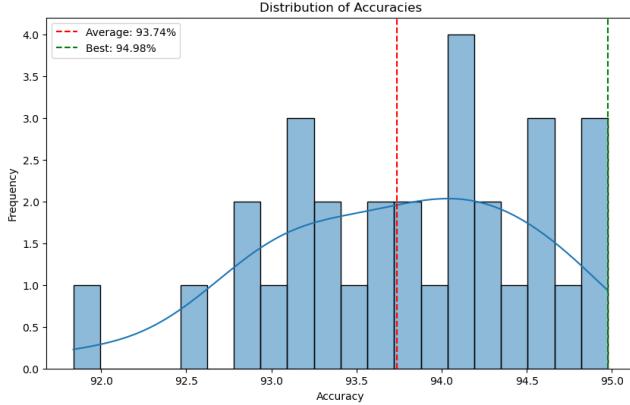


Figure 2: Distribution of accuracies on trained ViT model.

4 Results

Our ViT model did an excellent job in classifying the retinal images from the dataset. The best performing model reached almost 95% accuracy on the test set. This alone shows a drastic improvement over the CNNs accuracy we were comparing to. A breakdown of the metrics for our best performing model can be seen in figure X(metrics). The model not only improved overall in the accuracy of classification, but it also drastically improved its accuracy in regard to detecting Glaucoma. The CNN model achieved 46% precision while our ViT model achieved 93% precision on the same set trained on the same number of epochs.

Our macro average precision, recall, and f1-score were almost identical to the weighted averages. This shows that the model was able to generalize with high accuracy. The dataset was well-balanced which would indicate that the macro and weighted average should be very similar, this however was not the case when trained on the CNN architecture where these values differed by 1-5

	precision	recall	f1-score	support
Normal	0.8882	0.9321	0.9096	162
Cataract	0.9805	0.9618	0.9711	157
Diabetic Retinopathy	1.0000	0.9940	0.9970	166
Glaucoma	0.9324	0.9079	0.9200	152
accuracy	0.9498	0.9498	0.9498	0
macro avg	0.9503	0.9489	0.9494	637
weighted avg	0.9507	0.9498	0.9500	637

Figure 3: ViT Metrics for best model

A confusion matrix can be seen in figure X2 showing similar results for the accuracy of the model's predictive ability. With very few false positive and few false negatives, the results show high probability of success for this classification group with a relatively small dataset. Larger datasets and more classes (different retinal disease) are likely to show similar results.

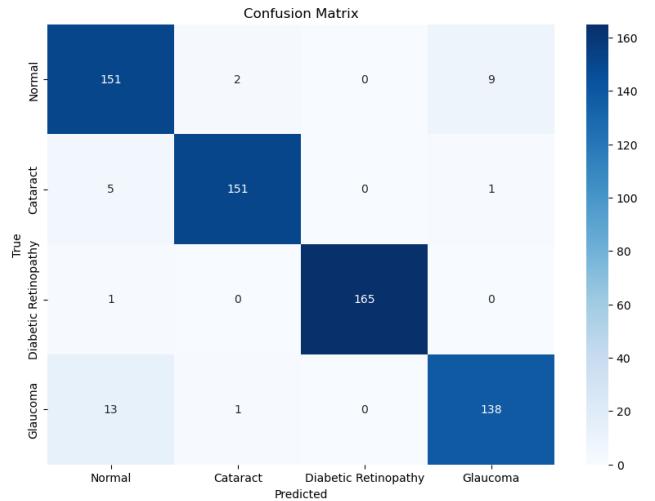
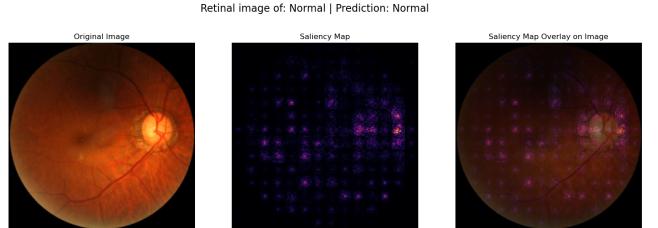


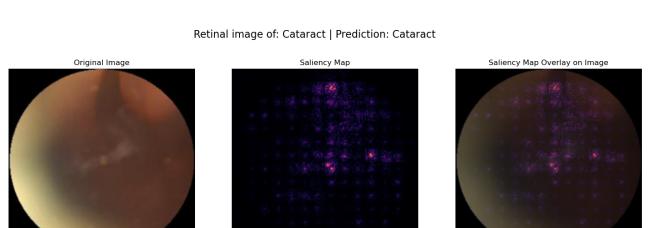
Figure 4: ViT Metrics for best model



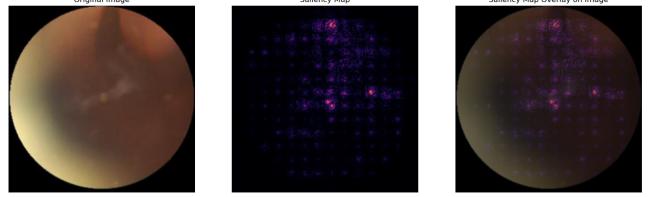
Figure 5: Color map legend



Retinal image of: Normal | Prediction: Normal

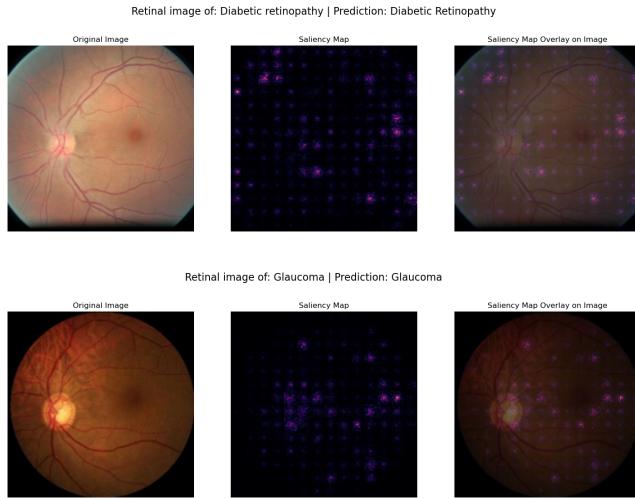
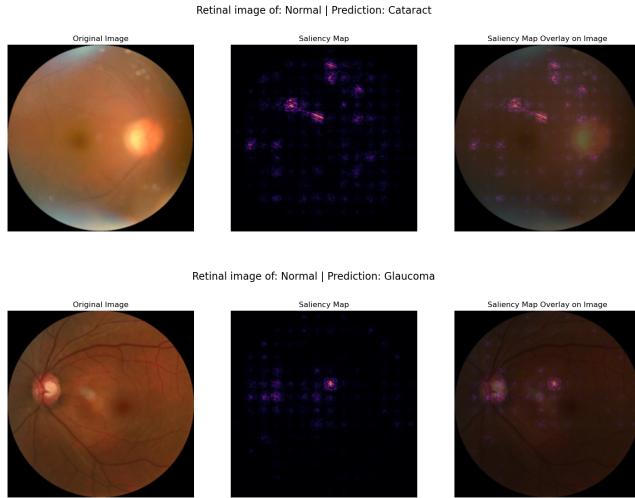


Retinal image of: Cataract | Prediction: Cataract



The results expand beyond the metrics of accuracy. As mentioned earlier, vision transformers have an attention mechanism that we can extract to "understand" where the model was giving focus on an image for its prediction. The following figures show a few examples of the image given to the model, the saliency map, which is represented a visual of where the model gives attention to the most on the image, and the saliency map overlay-ed on the original image. This is aspect alone allows the model to be used by ophthalmologists to review a patient's retinal and have an idea of where to focus or to catch details that would otherwise be hard to spot.

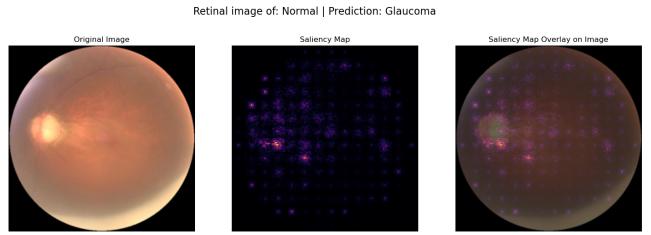
Included are also some examples were the model predicted incorrectly, these can be used to further train the model. The images

**Figure 6: Retinal Images with Saliency Maps**

as well could in theory be studied more thoroughly as they could be showing early signs of these diseases that the model was able to detect while a human observer may have missed, due to negligible differences or cases where the disease is too early for human observers to detect.

5 Conclusion

The results showed a high level of accuracy and massive improvements over previous retinal classification. The use of visual transformers in image classification in the medical field has a need for more exploration, and improvements on this architecture will only improve the level of detection. Even this experiment can be improved. The data was limited and not data augmentation was used, this could result in higher precision. The disease list was also limited to three diseases and a control group (normal). The model also only tried one pre-trained model and some preliminary training showed that the model was improving past 7 epochs. A deeper dive into where the model's focus through its attention mechanism can

**Figure 7: Retinal Images (incorrect predictions) with Saliency Maps**

be explored for further analysis. One last important improvement to this experiment would be in the use of multimodal models. The classification can further be expanded with an LLM to clarify areas of interest from the saliency maps and why those areas matter. Multimodal training should be considered to improve this model in a way that can be used publicly.

The results from this experiment show that a pre-liminary eye exam conducted by a ViT model can indeed yield highly accurate results. This can reduce cost for patients doing simple exams; while also increasing the amount of patients an ophthalmologists can meet with that require direct help. Less time can be spent doing preliminary work, and more time can be directed at treating the patients who need help. Technology like this also lowers the bar of entry and fear that most people have about having to spend high amounts of money to check their health. This advancement can keep being improved and potentially allow for this exam to be done at home with a phone camera. A few seconds to take a picture of an eye at home, and a patient may have just prevented future blindness due to early detection without any inconvenience.

In the future we hope to expand on this research, with different datasets or with funding from medical institutions. Labeled datasets are massively lacking in the medical field, especially when it comes to public datasets. We hope that we will be able to partner with medical and research facilities to expand the datasets we have, add more diseases and work alongside ophthalmologists to diagnose and find new areas of interest in the retina.

6 Acknowledgments

I'd like to thank the faculty and staff in the Healthcare in Artificial Intelligence class of Spring 2025. A big thank you to Dr. Ying Ding for her lectures knowledge.

References

- [1] bsdr18. 2021. Image Classification on Eye Disease Dataset. GitHub repository. Available: <https://github.com/bsdr18/Image-Classification-on-Eye-Disease-Dataset> (accessed 2025-04-28).
- [2] Derek Burrola. 2024. Ocular Disease Classification Using Machine Learning: A Comparative Study of CNN-Based Models. Manuscript. Accessed: 2025-04-28.
- [3] S. Chelaramani, M. Gupta, V. Agarwal, P. Gupta, and R. Habash. 2021. Multi-task Knowledge Distillation for Eye Disease Prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 3983–3993.
- [4] Margaret M. Deangelis, Ana C. Silveira, Elise A. Carr, and Ivana K. Kim. 2011. Genetics of Age-Related Macular Degeneration: Current Concepts, Future Directions. *Seminars in Ophthalmology* 26, 3 (May 2011), 77–93. doi:10.3109/08820538.2011.577129
- [5] Guna Venkat Doddi. 2023. Eye Diseases Classification Dataset. Kaggle dataset. Available: <https://www.kaggle.com/datasets/gunavenkatdoddi/eye-diseases-classification> (accessed 2025-04-28).

- [6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929. Available: <https://arxiv.org/abs/2010.11929>.
- [7] Nita G. Forouhi, Anoop Misra, Viswanathan Mohan, Roy Taylor, and William Yancy. 2018. Dietary and Nutritional Approaches for Prevention and Management of Type 2 Diabetes. *BMJ* (2018), k2234. doi:10.1136/bmj.k2234
- [8] Christos Matsoukas, Johan Fredin Haslum, Magnus Söderberg, and Kevin Smith. 2021. Is it Time to Replace CNNs with Transformers for Medical Images? arXiv preprint arXiv:2108.09038. Available: <https://arxiv.org/abs/2108.09038>.
- [9] Ross Wightman. 2019. PyTorch Image Models. GitHub repository. Available: <https://github.com/huggingface/pytorch-image-models>, DOI: <https://doi.org/10.5281/zenodo.4414861>.

Received 28 April 2025