# RetinFormer: Augmented ViT-based Retinopathy Classification

Hyeonwook Kim and Puneeth Batchu, M.S. in Artificial Intelligence

**FACULTY MENTOR: Ming Ma, Ph.D.**

Katz School of Science and Health

## Introduction

- Early detection of diabetic or hypertensive retinopathy helps prevent blindness.
- Artificial intelligence (AI) harnessing Vision transformers (ViTs) technology may offer a more accurate tool for early diagnosis, which could lead to early interventions that prevent vision loss—and mitigate the broader impact of these diseases.

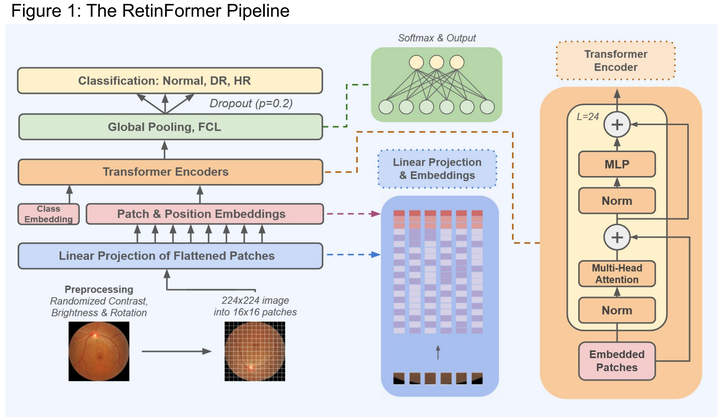
**Differences in AI Architecture**

- Existing methods use Convolutional Neural Networks (CNN) for Retinopathy classification. Models such as ResNet-50 were used for feature extraction, achieving 96% and 75% accuracy on MESSIDOR-2 and EYEPACS datasets respectively (Yaqoob et al., 2021).
- Vision Transformers (ViTs), which leverage the transformer architecture from Natural Language Processing, are increasingly being used as an alternative to CNNs for image processing, demonstrating competitive or even superior performance by capturing global relationships in images (Dosovitskiy et al., 2020).
- This paper proposes a new augmented Vision Transformer-based diabetic and hypertensive retinopathy classification method.

## Method

Our ViT model was trained on three public datasets; ODIR-5K, Messidor-2 and CGI-HRDC 2023, totaling up to over 7,000 unique fundus (non-invasive retinal) scans. Each input image was preprocessed to have randomized contrast, rotation, and brightness to prevent the model from overfitting to the data.

Figure 1: The RetinFormer Pipeline




The model performance was compared to other ViT-based classifiers, as well as an existing EfficientNet-based model. Each model was trained with Adam Optimizer, with a learning rate of 0.000004 and batch size of 8. Accuracy, F-1 Scores, and AUC was measured to benchmark each model.

## Results

| Model | Accuracy | Precision | F1-Score | AUC |
|---|---|---|---|---|
| HDR-EfficientNet | 0.8963 | 0.8983 | 0.8973 | 0.9548 |
| ViTAEv2 | 0.9018 | 0.9013 | 0.9015 | 0.9646 |
| ViT_b_16 | 0.9127 | 0.9168 | 0.9147 | 0.9847 |
| ViT_b_32 | 0.9147 | 0.9165 | 0.9156 | 0.9814 |
| ViT_l_32 | 0.9167 | 0.9165 | 0.9166 | 0.9784 |
| **RetinFormer** | **0.9271** | **0.9270** | **0.9270** | **0.9866** |

Table 1: Classification Performance Comparison

Accuracy, Precision, F1-Score, and AUC (area under the curve) comparisons between models. RetinFormer outperforms existing models in all metrics.

| Pretrained Weights | Pre-Process | Accuracy | Precision | F1-Score | AUC |
|---|---|---|---|---|---|
| ✗ | ✗ | 0.8884 | 0.8892 | 0.8887 | 0.9657 |
| ✗ | ✓ | 0.9008 | 0.9027 | 0.9017 | 0.9687 |
| ✓ | ✗ | 0.9261 | 0.9261 | 0.9261 | 0.9812 |
| ✓ | ✓ | 0.9271 | 0.9270 | 0.9271 | 0.9866 |

Table 2: Results of Ablation Study

Results of RetinFormer ablation study: pretrained weights and image pre-processing increased all measured benchmarks, with pretrained weights offering a slight advantage. Combining both processes has a minimal effect, indicating that the advantages they bring individually potentially overlap.

| Model | Training Time (Seconds) |
|---|---|
| HDREfficientNet | 11763.46 |
| ViTAEv2 | 27884.74 |
| ViT_b_16 | 17408.64 |
| ViT_b_32 | 10940.84 |
| ViT_l_32 | 19690.58 |
| RetinFormer | 38459.32 |

Table 3: Training Time Comparison

Training time comparisons between models: RetinFormer, a model based primarily on a ViT_l_16 model, requires more time to train compared to smaller models.

## Conclusions

- The study demonstrates the potential of Vision Transformers in automating the diagnosis of diabetic and hypertensive retinopathy.
- The results also highlight the importance of pre-trained weights and data preprocessing for better model performance and generalizability.

**Limitations**

- The data size was relatively small (7,000 images) and a larger sample size could enhance the model's robustness.
- Vision Transformer models require more time to train and test, depending on model size, when compared to their CNN-based counterparts.

**Recommendations**

- Future work could explore integrating multimodal data, such as patient medical history or other data points, to improve diagnostic accuracy.

## Acknowledgements

We would like to express our gratitude to Professor Ming Ma, Ph.D., for his guidance and mentorship throughout the entire project.

## References

Yaqoob, M. K., Ali, S. F., Bilal, M., Hanif, M. S., & Al-Saggaf, U. M. (2021). *ResNet-based deep features and random forest classifier for diabetic retinopathy detection*. Sensors, 21(11), 3883. https://doi.org/10.3390/s21113883

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale*. https://arxiv.org/abs/2010.11929

Peking University. (2019). *Peking University International Competition on Ocular Disease Intelligent Recognition* (ODIR-2019).

Computer Graphics International. (2023). *CGI-HRDC 2023 - Hypertensive Retinopathy Diagnosis Challenge*. https://codalab.lisn.upsaclay.fr/competitions/11877