Comparative Analysis of Vision Transformer and ResNet50 for Glaucoma Detection: Balancing Performance and Efficiency

Eric Hwang
09eric.hwang@dulwich-seoul.kr
Dulwich College Seoul
Seoul, Republic of Korea

Abstract

This study develops and evaluates the performance and computational efficiency of ResNet50 (CNN-based architecture) and Vision Transformer (Transformer-based architecture) for detecting glaucoma from fundus photographs. Glaucoma is one of the leading causes of irreversible blindness that affects millions of people around the world. Using deep learning methods, both models are trained to learn indicators of glaucomatous fundus photographs, such as thinning of the retinal nerve fiber layer (RNFL) and nasalization of blood vessels, to classify healthy and glaucomatous eyes. Gradient-weighted Class Activation Mapping (Grad-CAM) was used to interpret the model predictions by visualizing the regions of fundus photographs that contributed most significantly to the classification. With a publicly available dataset, we fine-tuned both models by leveraging transfer learning with a small learning rate (0.0001) on pre-trained layers. Both models were assessed with metrics such as accuracy, F1 score, inference time, throughput, and maximum GPU memory usage under controlled conditions. ResNet50 outperformed ViT, achieving higher accuracy (90.72% vs 87.64%) and an F1 score (0.9104 vs 0.8614) while being significantly more computationally efficient, with 68.48% faster inference and 50.52% lower GPU usage. These findings highlight the suitability of ResNet50 over ViT for use in resource-constrained medical settings to assist ophthalmologists diagnosing patients, as it effectively balances performance and efficiency.

Keywords

Artificial Intelligence, Computer Vision, Deep Learning, Neural Networks, Transfer Leanring, Medical Imaging, Glaucoma Detection

ACM Reference Format:
Eric Hwang. 2025. Comparative Analysis of Vision Transformer and ResNet50 for Glaucoma Detection: Balancing Performance and Efficiency . In *Proceedings of International Journal of Secondary Computing and Applications Research (IJSCAR VOL. 2, ISSUE 1).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.5281/zenodo.15149831

1 Introduction

Glaucoma is a chronic eye disease that is one of the leading causes of blindness and affects more than 80 million people around the world, with early diagnosis and treatment being the most important steps to prevent vision loss [7]. However, traditional approaches to manually diagnosing glaucoma can be time consuming and subject to high variability as they are based on subjective analysis [6]. As a result, these limitations can lead to delayed diagnosis, particularly in high-footfall and resource-constrained medical settings where ophthalmologists are in short supply.

With recent advancements in Artificial Intelligence (AI), Deep Learning (DL) models have emerged as transformative tools for medical imaging, offering accurate and time-efficient disease diagnosis. Architectures based on Convolutional Neural Networks (CNNs) have set the benchmark for medical image classification, while Vision Transformers (ViTs) [2] showcase the potential to outperform CNN-based architectures. Despite advancements in AI for healthcare, few studies have systemically compared CNNs and ViTs on both computational efficiency and performance for glaucoma detection. Along with performance itself, computational efficiency is also a crucial metric, often overlooked in previous research due to its focus on surpassing human expertise, particularly when evaluating a model's suitability to be used in resource-constrained medical environments, which is where the implementation of an automated diagnosis system should be of a higher priority.

Therefore, to address this gap, our study evaluates ResNet50 [4] and ViT using a collection of publicly available fundus photograph datasets. By leveraging transfer learning and fine-tuning both architectures, we rigorously analyze performance metrics (accuracy, F1 score) and computational efficiency metrics (inference time, throughput, maximum GPU memory usage) to determine their suitability for being implemented as an assistive tool for glaucoma diagnosis. Through this analysis, we establish a benchmark for deploying deep-learning tools for retinal imaging in healthcare, specifically tailored for glaucoma detection in diverse clinical settings with implications for broader medical applications.

2 Related Work

In the medical field, CNNs traditionally have performed very well in medical image classification and segmentation. In 2020, a comprehensive study by García et al. assessed two CNN architectures using a circumpapillary Optical Coherence Tomography (OCT) test dataset (glaucomatous: 93, healthy: 156 B-scans), concluding that fine-tuned CNNs, especially VGG architectures achieved a high area under the ROC curve [3]. Similarly, a study by Kulkarni and Ahmed evaluated a more exhaustive list of CNNs and histogram equalization techniques using a small (glaucomatous: 396, healthy: 309) dataset, with VGG-19 achieving the highest accuracy (97.9%) and ResNet architectures achieving the highest specificity score (98.4%) [8]. These studies highlight the effectiveness of CNN architectures in glaucoma detection.

This paper is published under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International (CC-BY-NC-ND 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
IJSCAR VOL. 2, ISSUE 1
© 2025 The Author(s). Published by IJSCAR.

Recently, there has been active research on utilizing ViTs for medical imaging. A study by Nurgazin et al. demonstrated that ViT with ProtoNet, a few-shot learning algorithm, can achieve competitive performance against state-of-the-art CNN-based approaches in medical benchmark datasets [10]. More specifically for glaucoma detection, Hwang et al. demonstrated that ViTs perform comparably to CNNs in glaucoma detection using fundus photographs (total: 3,142), achieving a comparable area under the ROC curve and accuracy. They concluded that ViTs are preferable when prioritizing sensitivity over minimizing false positives [5].

Overall, the existing literature underscores the efficacy of both CNNs and ViTs in medical imaging, with emerging evidence suggesting that ViTs may offer competitive, if not superior, performance in medical imaging. While a comparative analysis of diagnostic performance exists such as [5], a comprehensive evaluation that incorporates computational metrics for glaucoma detection and considers the constraints of resource-limited clinical settings is lacking. Addressing this gap is crucial for informing the selection of appropriate models for deployment in varying healthcare environments.

Furthermore, although these studies demonstrate the effectiveness of CNNs and ViTs, their small dataset sizes limit their generalizability to larger, more diverse clinical populations. By leveraging a significantly larger collection of fundus photograph datasets (total: 7815 without sampling), this study addresses this limitation, providing a more rigorous evaluation of CNN and Vision Transformer (ViT) architectures. This larger dataset enables more robust performance comparisons and enhances the applicability of findings to real-world clinical settings.

Finally, while [5] used a collection of datasets separately for comparison and without balancing the dataset, which yielded valuable insights on architectural differences, we balanced the dataset using sampling techniques with data augmentations to ensure equitable representation of glaucomatous and healthy cases, reducing potential biases and improving the reliability of the model evaluation. This methodological enhancement strengthens the robustness of the results, making them more applicable to diverse clinical settings.

3 Methods

3.1 Data Collection and Preprocessing

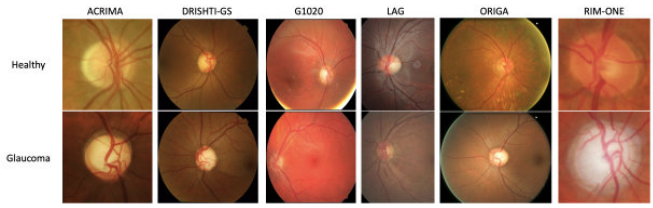


Figure 1: Sample Images of Fundus Photographs from Each Dataset

The dataset used for this study is a collection of 6 public fundus photograph datasets, downloaded from Kaggle [11] which consists of the following: ACRIMA, DRISHTI-GS, G1020, LAG, ORIGA, and RIM-ONE [1]. Using a collection of datasets increases the total dataset size and also increases data diversity which ensures that the models learn to generalize to variations such as differences in fundus photography protocols, imaging equipment, and patient demographics.

In total, There are 7815 images (healthy: 5002, glaucoma: 2813) of fundus photographs. Figure 1 demonstrates sample fundus photographs from each dataset. Since there is a dataset imbalance, with a higher number of healthy eye scans, random oversampling was applied to balance the dataset, preventing the models from becoming biased toward the majority class. This step is critical in the medical context because false negatives (failing to identify glaucoma correctly) can have severe consequences for patients as the early detection of glaucoma is essential. With random oversampling, the dataset becomes (healthy: 5002, glaucoma: 5002).

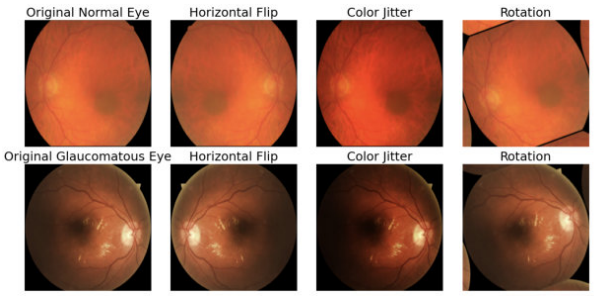


Figure 2: Sample Images of Data Augmentation Methods

As shown in Figure 2, the multiple data augmentation methods (horizontal flip, random rotation, and color jitter) were used to simulate real-world variations in retinal imaging conditions while also increasing the total dataset size which prevents overfitting. Horizontal flips simulate eye symmetry, rotations mimic misalignment during image capture, and color jitter compensates for variations in imaging equipment and lighting conditions. As a result, it makes the models more robust, which is especially important in resource-constrained environments where imaging quality may vary significantly.

Moreover, All images were resized to 224x224 pixels to standardize input dimensions which allows compatibility with ResNet50 and ViT architectures. Furthermore, the fundus photograph can still preserve the indicators of glaucoma, such as nasalization of the blood vessels and thinning of the retinal nerve fiber layer (RNFL) [9], while optimizing computational efficiency, a key consideration for deployment in resource-constrained medical environments. Finally, the dataset was randomly split into training data and validation data with a 75:25 ratio which allows sufficient data for training while having enough data for an unbiased validation to evaluate the metrics.

3.2 Model Selection and Modification

Since ResNet achieved the highest specificity in [3] and is widely researched in medical classification tasks such as ResNet-22 for breast cancer classification [14], ResNet50, an architecture more robust



Figure 3: ResNet50 Model Schematic (Illustration inspired from [12])

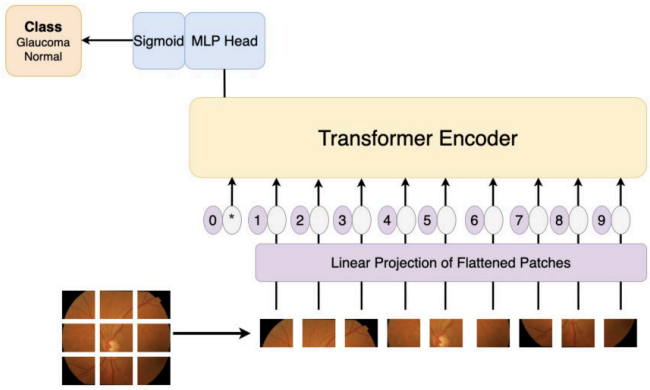


Figure 4: ViT Model Schematic (Illustration inspired from [2])

than the original ResNet34, was selected as a representative CNN model against ViT. This comparison highlights the strengths and limitations of convolutional and transformer-based architectures in glaucoma detection.

Both models were initialized with ImageNet pre-trained weights and then were trained with the glaucoma dataset to facilitate transfer learning. Their entire network was fine-tuned with a small learning rate to allow them to adapt to the specific task of detecting glaucoma from general image classification.

The activation function of the output layer was changed to a sigmoid function as the task is binary classification. Furthermore, it makes the models output a prediction between 0 and 1 for detecting glaucoma, which allows the model to be used as a screening tool or to assist ophthalmologists in diagnosing patients.

L2 Regularization (weight decay) was used as this value minimized overfitting by penalizing large weights which encourages simpler model representations. Regularization is particularly crucial in medical imaging tasks, where preserving the model's ability to generalize across diverse patient data is vital.

Furthermore, The number of epochs was set to 10 for both models as during preliminary experiments, ViT's training and validation loss began diverging, indicating overfitting, after the 10th epoch due to the relatively small dataset and transformer-based architecture's high capacity for pattern memorization. Hence, by limiting the number of epochs, we balanced the models' learning capacity and generalization.

3.3 Metrics

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Accuracy and F1 score were chosen as primary metrics for measuring model performance as accuracy provides an overall measure of each model's ability to accurately detect glaucoma from fundus photographs. Furthermore, the F1 score balances sensitivity and precision which is crucial in medical classification tasks.

- Inference: time taken for the model to make a prediction on a given input, measured in seconds.
- Throughput: number of predictions made by the model per unit of time, measured in images/second.
- Max GPU Memory Usage: the peak amount of GPU memory used during training, measured in MB.

On the other hand, Computational efficiency was assessed using inference time per image, throughput (images processed per second), and maximum GPU memory usage during inference. These metrics were chosen to evaluate the feasibility of deploying these models in medical environments with restricted computational resources or a high density of patients.

3.4 Controls

Both models were trained in equal hardware and software conditions. Both were run on Google Colaboratory with T4 GPU (High-RAM 51GB) as a hardware accelerator. Equal versions of Python (3.10.12) and Pytorch (2.3.1+cu121) were used to prevent discrepancies. Moreover, the same data set was used with the same ratio of train-validation split.

In terms of model design, the pre-trained weights for both models were generated using the ImageNet dataset which gives them a fair baseline to begin transfer learning. Both models were trained with Binary Cross Entropy Loss (BCELoss) as the loss function and Adam optimizer as the optimization algorithm to ensure they are both minimizing the same loss function and optimizing towards the same goal.

Overall, the following hyperparameters were kept constant for equal training conditions.

- Epoch: 6
- Learning Rate: 0.0001
- L2 Regularization - Weight Decay: 0.0001
- Batch Size: 32

4 Results

Epoch	Train Loss	Validation Loss	Accuracy	F1	Inference Time	Throughput	GPU usage
1	0.3341	0.3852	82.37%	0.8301	8.4035 s	297.61 img/s	782.34 MB
2	0.2243	0.2287	88.56%	0.8844	8.4699 s	295.28 img/s	782.04 MB
3	0.1764	0.2181	89.40%	0.8912	8.4621 s	295.55 img/s	782.04 MB
4	0.1602	0.2385	89.00%	0.8907	8.4657 s	295.43 img/s	782.04 MB
5	0.1392	0.2020	89.96%	0.8984	8.4778 s	295.01 img/s	782.04 MB
6	0.1116	0.2034	90.72%	0.9104	8.4417 s	296.27 img/s	782.04 MB

Table 1: Results of ResNet50 Training and Validation

Epoch	Train Loss	Validation Loss	Accuracy	F1	Inference Time	Throughput	GPU usage
1	0.4704	0.3387	84.69%	0.8340	26.85 s	93.14 img/s	1580.54 MB
2	0.2972	0.2621	87.60%	0.8693	26.80 s	93.32 img/s	1580.54 MB
3	0.2403	0.2207	88.64%	0.8822	26.83 s	93.21 img/s	1580.54 MB
4	0.2145	0.2359	88.68%	0.8801	26.86 s	93.13 img/s	1580.54 MB
5	0.1993	0.2588	88.04%	0.8781	26.86 s	93.12 img/s	1580.54 MB
6	0.1925	0.2612	87.64%	0.8614	26.75 s	93.49 img/s	1580.54 MB

Table 2: Results of ViT Training and Validation

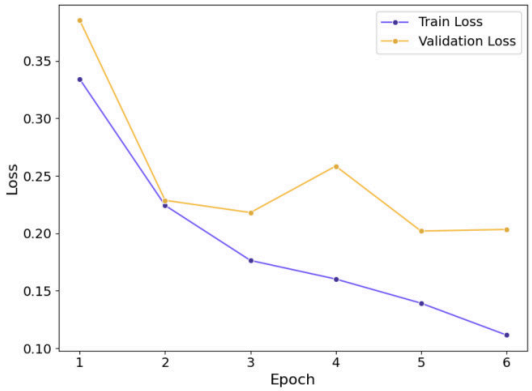


Figure 5: Train and Validation Loss over Epoch, ResNet50

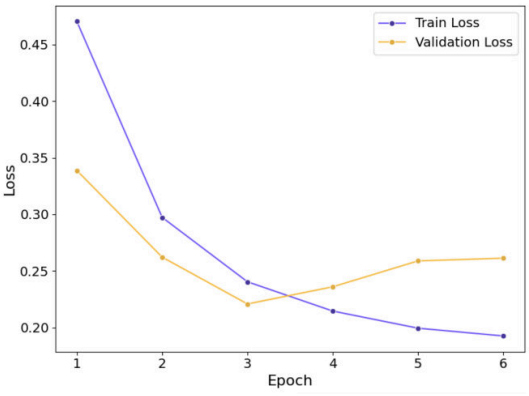


Figure 6: Train and Validation Loss over Epoch, ViT

5 Discussion

5.1 Training and Performance Analysis

For ResNet50, the training loss steadily decreases from 0.3341 to 0.1116 over 6 epochs, and the validation loss also decreases from 0.3852 to 0.2034 with only a slight increase from Epoch 3 to 4. Therefore, it indicates that ResNet50 generalized well without significant overfitting issues. For ViT, the training loss drops sharply from 0.4704 to 0.2972 over the first 2 epochs but decreases much more slowly after reaching the final training loss of 0.1925. Its validation loss initially decreases from 0.3387 to 0.2207 for the first 3 epochs but increases slightly afterward to reach 0.2612 at the final epoch. This indicates that ViT had slight overfitting at later epochs as validation loss starts to increase while train loss decreases but overall, it generalized quite well.

Furthermore, ResNet50 achieved a 90.72% accuracy for the final epoch with an F1 score of 0.8614. This demonstrates that the model had a very good performance as it predicted most of the dataset correctly without overly predicting one side, indicated by a high F1 score. Although ViT also had a strong performance with 87.64% accuracy and 0.8614 F1, ResNet50 exceeded ViT in performance metrics with 3.08% higher accuracy and 0.049 higher F1. It can also be observed that the accuracy of ViT stabilizes from Epoch 2 onward which implies that the model is converging where it is no longer learning more information from the given dataset. Lastly, the average of each computational metric over 6 epochs was calculated as shown in Table 3.

5.2 Computational Efficiency and Practical Implications

Model	Average Inference Time	Average Throughput	Max GPU Usage
ResNet50	8.45 s	295.86 img/s	782.09 MB
ViT	26.83 s	93.24 img/s	1580.54 MB

Table 3: Comparison of Computational Metrics

In general, ResNet50 was 68. 48% faster in inference making and predicted 217. 31% more images/second while using 50. 52% less GPU, outperforming ViT also in computational metrics. The difference between the two models was significant, where ViT was using approximately double the GPU memory as ResNet50 while making predictions much slower. Furthermore, ResNet50 even exceeded ViT in performance metrics.

The big gap in computational efficiency between ResNet50 and ViT highlights that ResNet50 is much better suited to be deployed in resource-constrained medical environments such as rural clinics and mobile diagnostic units or high-footfall environments. While ViT showcased promising results for diagnostic performance, its computational inefficiency directs to the need for further optimizing Transformer-based models for real-world applications.

The computational inefficiency of ViT is likely caused by the self-attention mechanisms which enables global feature extraction, but leads to quadratic complexity. This explains the significantly higher GPU memory usage and slower inference time of ViT as ResNet50 has more linear computation process. Furthermore, the ViT used in the investigation, ViT_B_16, has notably higher number of parameters with 86,567,656 [16] compared to 25,557,032 parameters [15] of ResNet50. This likely explains the slight overfitting for the last 3 epochs as it could've begun memorizing the dataset.

5.3 Grad-CAM Visualization

Figure 7 demonstrates the Gradient-weighted Class Activation Mapping (Grad-CAM) [13] visualization for ResNet50, mapping different regions on the fundus photograph that the model identified as most influential in detecting glaucoma. While the optical nerve head is visible on the activation map, the model appears to focus more on the surrounding peripapillary regions which suggests that it relies on subtle retinal features beyond the optical nerve head. This Grad-CAM visualization provides insights to the model's decision-making process. The focus on peripapillary regions in the Grad-CAM visualization could explain why ResNet50 outperformed ViT

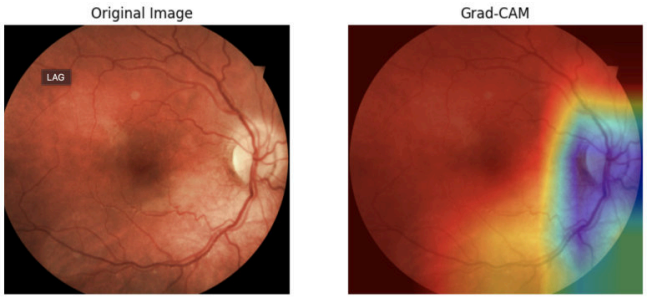


Figure 7: Grad-CAM Visualization of ResNet50

as it localized these small but critical indicators of glaucoma compared to ViT which architecturally focuses more on global patterns. Moreover, utilizing Grad-CAM visualizations for medical diagnosis have multiple benefits as it provide ophthalmologists visual justifications for the model's predictions, making it more effective for assisting medical diagnosis by human doctors.

Therefore, Figure 7 indicates that the CNN architecture of ResNet50 is better at extracting small features. Especially with ResNet allowing the layers to be much deeper than other CNNs by eliminating the vanishing gradient problem, it is clear that ResNet50 was better at identifying small indicators of glaucoma than ViT. Future work could involve validating Grad-CAM visualizations with expert annotations to confirm alignment with clinical markers of glaucoma.

5.4 Limitations and Future Directions

Despite utilizing a significantly larger dataset (total: 10004 after oversampling) with various augmentation methods to increase the size further, ViT's performance plateaued after Epoch 3, suggesting that the model may require even larger datasets or different augmentation strategies to fully converge and reach its potential. Future studies could explore training ViT on substantially bigger private datasets from multiple institutions or using synthetic data augmentations.

In addition, classifying glaucoma as binary oversimplifies the real-world problem of diagnosing glaucoma as it can be at different stages or conditions, requiring more elaboration than just binary prediction. Hence, developing multi-classification datasets for glaucoma detection could be crucial for identifying nuanced progression of the disease.

Future works should also investigate modifying the ViT architecture such as hybrid CNN-Transformer models that can capture both local and global images.

6 Conclusions

This study develops and evaluates the model performance and computational efficiency of ResNet50 and Vision Transformer for detecting glaucoma from fundus photographs. With a publicly available dataset, we fine-tuned both models by leveraging transfer learning with a small learning rate (0.0001) on pre-trained layers.

For diagnostic performances, ResNet50 achieved a 90.72% accuracy for the final epoch with F1 score of 0.8614. Although ViT also had strong performance with 87.64% accuracy and 0.8614 F1, ResNet50 exceeded ViT in performance metrics with 3.08% higher accuracy and 0.049 higher F1. Furthermore, in terms of computational efficiency, ResNet50 was 68.48% faster at making inferences and predicted 217.31% more images/second while using 50.52% less GPU, outperforming ViT also in the computational metrics. These findings highlight the suitability of ResNet50 over ViT for use in resource-constrained medical settings, where both reliability and efficiency are crucial, to assist ophthalmologists in diagnosing patients.

7 Acknowledgement

I would like to express my deepest thanks to Ms. Catherine Robinson for her guidance and support throughout the research. Her insights and feedback were greatly appreciated and helped me stay on track during key stages of my work.

References

[1] Francisco José Fumero Batista, Tinguaro Diaz-Aleman, Jose Sigut, Silvia Alayon, Rafael Arnay, and Denisse Angel-Pereira. 2020. RIM-ONE DL: A Unified Retinal Image Database for Assessing Glaucoma Using Deep Learning. *Image Analysis and Stereology* 39, 3 (2020), 161–167. https://doi.org/10.5566/ias.2346

[2] 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* (June 2021). https://doi.org/10.48550/arXiv.2010.11929 ICLR camera-ready version.

[3] Gabriel García, Rocío del Amor, Adrián Colomer, and Valery Naranjo. 2020. Glaucoma Detection From Raw Circumpapillary OCT Images Using Fully Convolutional Neural Networks. In *Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP)* (Abu Dhabi, United Arab Emirates). IEEE, 2526–2530. https://doi.org/10.1109/ICIP40778.2020.9190916

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385* (December 2015). https://doi.org/10.48550/arXiv.1512.03385 Tech report.

[5] Elizabeth E. Hwang, Dake Chen, Ying Han, Lin Jia, and Jing Shan. 2023. Multi-Dataset Comparison of Vision Transformers and Convolutional Neural Networks for Detecting Glaucomatous Optic Neuropathy from Fundus Photographs. *Bioengineering (Basel)* 10, 11 (October 2023), 1266. https://doi.org/10.3390/bioengineering10111266

[6] Shajila Beegam M. K. and Mala Kalra. 2024. Leveraging CNN and Fundus Imaging for Enhanced Glaucoma Detection. *SN Computer Science* 5, 1 (2024), Article 1137. https://doi.org/10.1007/s42979-024-03527-4

[7] Mona Kaleem. n.d.. Glaucoma. [Online]. Available: https://www.hopkinsmedicine.org/health/conditions-and-diseases/glaucoma. Johns Hopkins Medicine. [Accessed: 2024-12-00].

[8] Ashish Kulkarni and H. Shafeeq Ahmed. 2024. A Comparative Study on Deep Convolutional Neural Networks and Histogram Equalization Techniques for Glaucoma Detection From Fundus Images. *medRxiv* (October 2024). https://doi.org/10.1101/2024.10.25.24316109

[9] Jordan Mandell, Benjamin Lin, and Ann Quan. 2024. *Evaluating the cup and disc in glaucoma.* https://eyeguru.org/blog/evaluating-cup-and-disc/ Accessed: 2024-12-01.

[10] Maxat Nurgazin and Nguyen Anh Tu. 2023. A Comparative Study of Vision Transformer Encoders and Few-shot Learning for Medical Image Classification. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (Paris, France). IEEE, 2505–2513. https://doi.org/10.1109/ICCVW60793.2023.00265

[11] Ayush Roy. 2023. Glaucoma Classification Datasets: DRISHTI-GS, RIM-ONE, ACRIMA, ORIGA, and G1020. https://www.kaggle.com/datasets/ayush02102001/glaucoma-classification-datasets?resource=download Accessed: 2024-01-01.

[12] Srinivas Rahul Sapireddy. 2023. *ResNet-50: Introduction.* https://srsapireddy.medium.com/resnet-50-introduction-b5435fdba66f Accessed: 2024-12-00.

[13] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2019. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* 128, 2 (2019), 336–359. https://doi.org/10.1007/s11263-019-01228-7

[14] Yiqiu Shen, Nan Wu, Jason Phang, and Jungkyu Park. 2020. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *arXiv preprint arXiv:2002.07613* (February 2020). https://doi.org/10.48550/arXiv.2002.07613

[15] Torchvision Contributors. 2023. *torchvision.models.resnet50.* https://pytorch.org/vision/main/models/generated/torchvision.models.resnet50.html
[16] Torchvision Contributors. 2023. *torchvision.models.vit_b_16.* https://pytorch.org/vision/main/models/generated/torchvision.models.vit_b_16.html

Received 01 January 2025; accepted 27 March 2025