A ES
ISSN: 2096-3246
Volume 58, Issue 01, 2026

A Comparative Evaluation of Lightweight Vision Transformer Models for Retinal Disease Classification

Keerthi Guttikonda[1], Baburao Markapudi[2],Bala Brahmeswara Kadaru[3], Sravana Kumar Komma[4], Dr.Shobana Gorintila[5]

[1,2,3]Department of Computer Science and Engineering,
Seshadri Rao Gudlavalleru Engineering College,
Gudlavalleru, India.

[4]AI-Department, VIT, Bhimavaram, West Godavari(D.T), Andhra Pradesh, India.

[5] Professor, CSE-Department, NRI Institute of Technology, Pothavarappadu(V), Via Nunna, Agiripalli(M), Vijayawada Rural, Krishna(D.T), Andhra Pradesh, India.

keerthi.guttikonda@gmail.com[1], baburaompd@gmail.com[2], balukadaru2@gmail.com[3], komma1506@gmail.com[4], drgshobana@gmail.com[5]

Abstract. The retinal disease screening systems require a high level of diagnostic accuracy along with computational efficiency to be implemented in resource-constrained clinical settings. Although the large vision transformer models have shown promising results in retinal image analysis tasks, their computational complexity restricts their applicability. This paper presents a comparative study of two light-weight Vision Transformer (ViT) models, namely Swin-Tiny and ViT-Small, for multi-class retinal disease classification problems using color fundus images. A judiciously selected dataset of 4,217 fundus images from four classes: Normal, Diabetic Retinopathy (DR), Age-related Macular Degeneration (AMD), and Branch Retinal Vein Occlusion (BRVO) is used for the experimental study. Both models are trained in the same manner for preprocessing, training, and testing to make a fair and impartial comparison. The performance of the models is assessed using standard classification performance metrics like accuracy, precision, recall, and F1-score, and model complexity metrics like the number of parameters and computational complexity. The experimental study reveals that both models are able to achieve a competitive performance with varying trade-offs between classification performance and computational complexity.

Keywords: Retinal disease classification · Vision Transformer · Swin Transformer · Fundus imaging· Computer-aided diagnosis·

1 Introduction

The analysis of retinal fundus images has been identified as a promising approach to mass screening, especially in areas where access to ophthalmologists is limited. Early approaches in the field of deep learning demonstrated that automated systems can be equally accurate to human experts in detecting diabetic retinopathy from retinal fundus images [1]. Further validation across multiethnic populations confirmed the robustness and generalizability of deep learning frameworks for retinal disease detection [2].

Deep learning models have also been successfully applied to the grading and severity assessment of AMD from color fundus photographs [3]. Convolutional Neural Network (CNN) -based image mining and automated detection systems improved sensitivity and specificity for diabetic retinopathy screening tasks [4–5]. At the same time, studies highlighted the importance of consistent reference standards and reliable annotations for accurate evaluation of machine learning models in clinical settings [6].

Apart from disease diagnosis, the diagnostic potential of retinal fundus images has been explored to predict systemic risk factors using deep learning techniques [7]. End-to-end retinal diagnosis and referral systems further explored the feasibility of AI-based retinal care [8]. Although significant progress has been made, traditional CNN models are mostly based on local receptive fields, which may not be adequate to capture long-range dependencies in high-resolution retinal images.

Vision Transformers (ViTs) proposed a global self-attention mechanism that allows capturing relationships across the image [9]. Hierarchical transformer models with shifted window attention improved the efficiency of the model without compromising its representational power, making it applicable for medical image analysis tasks [10]. Inspired by the recent progress, this work investigates the use of lightweight vision transformer models for multi-class retinal disease classification, aiming to strike a balance between diagnostic performance and efficiency.

The rest of this paper is organized as follows. Section 2 discusses the related work on retinal disease diagnosis and lightweight vision transformer models. Section 3 describes the dataset and preprocessing steps. Section 4 introduces the Swin-Tiny and ViT-Small models. Section 5 shows the experimental setup, followed by the results and discussion in Section 6. Section 7 concludes the paper with future work, and Section 8 concludes the paper.

2. Literature Review

More recently, transformer architectures have had a big impact on medical image analysis. Cao et al. proposed Swin-Unet, a pure transformer-based encoder-decoder architecture for medical image segmentation, demonstrating that hierarchical self-attention mechanisms are efficient for capturing multi-scale contextual information in biomedical images [11]. More recently, Chen et al. proposed TransUNet, a transformer encoder-based U-Net architecture, demonstrating that transformers can be strong feature extractors for medical image analysis tasks by modeling long-range dependencies [12]. These works proved the feasibility of transformer-based architectures for structured medical image understanding.

Before the rise of transformers, convolutional neural networks were the foundation of medical image classification models. He et al. introduced deep residual learning using ResNet, solving the degradation problem in deep networks and allowing for greatly improved feature representation for visual recognition tasks [13]. Scaling up from these ideas, Tan and Le proposed EfficientNet, a compound scaling approach that balances depth, width, and resolution for better performance-efficiency trade-offs [14]. These CNN-based advances provided robust baselines for retinal disease classification and highlighted the need for computational efficiency in practical applications.

With the increasing popularity of vision transformers, the focus has been on improving the training efficiency and generalization of vision transformers. Touvron et al. proposed data-efficient image transformers (DeiT), proving that transformer models can match the performance of CNNs without requiring large amounts of training data using knowledge distillation techniques [15]. This is particularly important in the medical imaging community, where labeled data are scarce.

In retinal disease analysis in particular, Li et al. utilized vision transformer models for fundus image classification and found better performance in identifying multiple retinal diseases by utilizing global contextual modeling [16]. Yu et al. designed a new model called MIL-VT, which combined multiple instance learning and vision transformers to improve robustness in fundus image classification, particularly when lesion localization is ambiguous [17]. Further advancing hierarchical transformer models, Xu et al. showed successful differential diagnosis of age-related macular degeneration using multi-scale attention mechanisms specifically designed for retinal images [18].

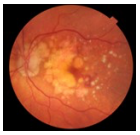
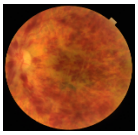
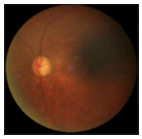
Hybrid models have also been developed to leverage the complementary benefits of convolutional operations and transformer models for global representation. Dutta et al. designed a new Conv-ViT model that combined convolutional layers and vision transformer modules to improve feature extraction for retinal disease diagnosis [19]. Moreover, Jiang et al. designed a vision transformer model-based computer-aided diagnosis system for retinopathy, which demonstrated the real-world applicability of transformer models in retinal screening systems [20].

Taken together, these studies clearly reflect the shift from traditional CNN-based models to transformer and hybrid models in retinal image analysis. They also underscore the need to achieve a balance between representation capability and computational complexity, which further encourages comparative analysis of light-weight transformer models for multi-class retinal disease classification.

3. Dataset and Preprocessing

This work uses a carefully compiled dataset of color retinal fundus images sourced from various publicly available sources. The dataset consists of a total of 4,217 fundus images belonging to four significant categories: Normal, DR, AMD, and BRVO. The chosen dataset contains both normal and diseased retinal images, covering various manifestations of the diseases that are usually encountered in ophthalmological screening. The class-organized structure is maintained to facilitate multi-class classification. Table 1 illustrates the compassion of retinal conditions in the dataset.

Table 1. Comparison of retinal conditions including normal retina, AMD, BRVO, and DR

Feature	Eye Condition			
	Normal	**AMD**	**BRVO**	**DR**
Cause	Healthy retinal circulation	Degeneration of macula due to aging, oxidative stress, genetic factors	Blockage of branch retinal vein	Chronic hyperglycemia damaging retinal blood vessels
Affected Areas	Entire retina appears healthy	Macula (central vision area)	Specific branch vein territory (localized sector of retina)	Diffuse retina, especially microvasculature
Sample Images				

To maintain uniformity in experiments, all the fundus images are resized to a fixed spatial dimension of 224 × 224 pixels, as required by the transformer architectures considered in this work. Since the fundus images are usually captured in varying illumination conditions and imaging modalities, a uniform preprocessing step is adopted to reduce variability among images while retaining clinically significant details like vascular patterns, lesions, and macular areas.

Normalization of the images is done based on mean and standard deviation values obtained from the ImageNet dataset to normalize the input distribution to match the pretrained weights for efficient transfer learning. The data preprocessing technique does not involve drastic geometric and intensity transformation to avoid any degradation of the minute pathological features essential for precise retinal disease segmentation.

The dataset is split into training and testing sets with class-wise stratification to avoid any leakage of information and ensure an unbiased performance assessment. The same dataset split and preprocessing steps are uniformly followed for all models being compared to ensure a fair

comparison of their relative performance. This uniform data preparation framework will help in making a fair comparison between Swin-Tiny and ViT-Small models in the experimental analysis.

4. Models Under Comparison

This section describes the two lightweight vision transformer architectures evaluated in this study: Swin-Tiny and ViT-Small. Both models are designed to reduce computational complexity while retaining the representational advantages of transformer-based learning. Their architectural differences, parameter efficiency, and suitability for retinal disease classification are discussed below.

4.1 Swin-Tiny Architecture

Swin-Tiny is a light version of the Swin Transformer series that follows a hierarchical structure with window-based self-attention. Swin-Tiny differs from the general vision transformer in that it only computes self-attention within non-overlapping local windows, which is much more computationally efficient. However, to overcome the locality constraint imposed by the window-based self-attention mechanism, shifted window strategies are used between consecutive transformer layers to allow information flow between windows efficiently. The Swin Transformer architecture is represented in figure 1.

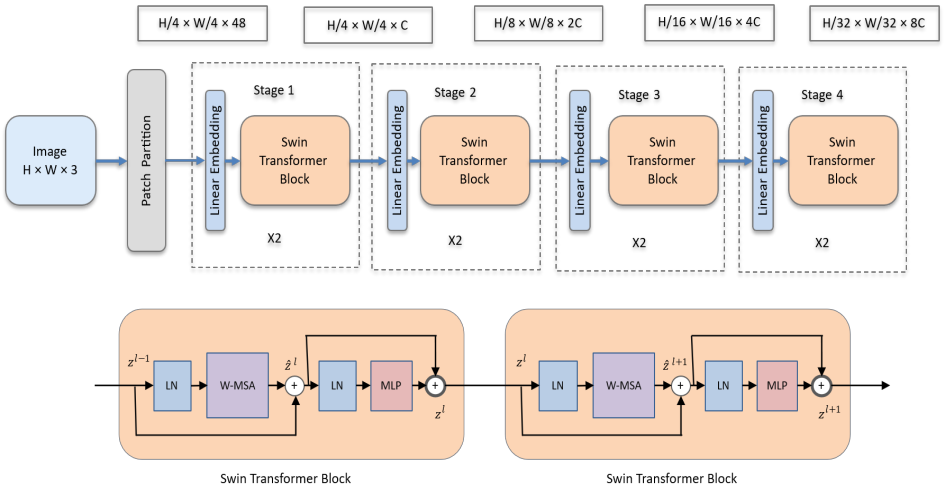


Fig 1: Architecture of the Swin-Tiny model illustrating hierarchical patch embedding, window-based multi-head self-attention, shifted window mechanism, and patch merging layers for multi-scale feature extraction.

For a given input fundus image of 224 × 224 resolution, Swin-Tiny first divides the input image into smaller patches and then embeds them into a low-dimensional feature space. The model then

processes these embeddings through a series of hierarchical levels, each of which comprises Swin Transformer layers and patch merging layers. As the network progresses, the spatial resolution is gradually reduced, and the feature dimension is gradually increased to capture multi-scale retinal features such as fine vascular details and larger pathological areas.

Swin-Tiny has roughly 28 million trainable parameters, which is much smaller compared to other variants of the Swin Transformer series while retaining excellent feature extraction performance. The hierarchical structure and efficient attention mechanism of Swin-Tiny are particularly beneficial for the analysis of retinal fundus images, where both localized lesions and global structural information are essential for distinguishing between different diseases.

4.2 ViT-Small Architecture

ViT-Small is a light version of the original Vision Transformer model, which applies global self-attention to all the patches in the image. In the ViT-Small model, the input fundus image is split into patches of a fixed size, which are linearly embedded and then embedded again using positional embeddings. The embeddings are then passed through a series of transformer encoder layers that consist of multi-head self-attention and feed-forward networks. The Vision Transformer model is shown in figure 2.

Unlike the Swin-Tiny model, the ViT-Small model does not employ a hierarchical feature map or attention windows. Rather, the ViT-Small model applies global self-attention at every layer of the transformer, which enables the model to attend to the entire image for long-range dependencies. This is especially useful for the ViT-Small model in effectively capturing global retinal information, which can be extremely beneficial in detecting global pathological patterns. However, the application of global self-attention also increases the computational cost compared to attention windows.

The ViT-Small model has approximately 22 million parameters, which is much more parameter-efficient compared to the Swin-Tiny model. The ViT-Small model has a relatively simpler architecture and fewer parameters, which makes it extremely attractive from an efficiency standpoint. With the application of transfer learning, the ViT-Small model can be extremely beneficial in medical imaging tasks, which was originally intended for natural image classification tasks.

ISSN: 2096-3246
Volume 58, Issue 01, 2026

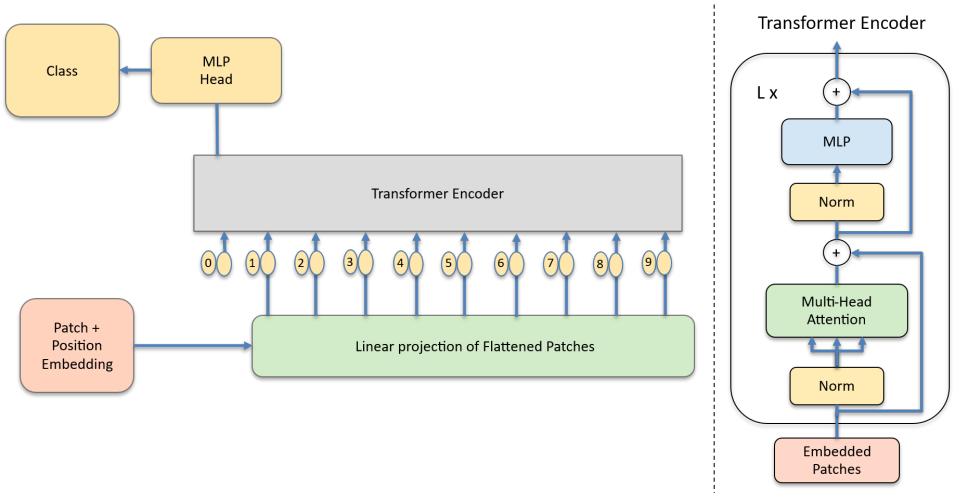


Fig 2: Architecture of the Vision Transformer (ViT-Small) model showing patch embedding, positional encoding, stacked transformer encoder blocks with global multi-head self-attention, and classification head.

4.3 Architectural Comparison and Suitability

While both Swin-Tiny and ViT-Small are lightweight transformer architectures, they are vastly different in terms of their attention mechanisms and feature representation approaches. Swin-Tiny focuses on hierarchical learning using localized and shifted window attention, which allows for efficient multi-scale feature extraction. On the other hand, ViT-Small focuses on global context representation using full self-attention, which incurs higher per-layer computational complexity.

Based on the architectural differences between the two models, there is a need to conduct a comparative assessment of their performance capabilities on retinal disease classification tasks. Through the assessment of both performance and efficiency-related metrics, it is hoped that this research will be able to identify which of the two lightweight vision transformer models has a superior balance between performance and computational complexity for practical retinal screening applications.

5. Experimental Setup

The experimental evaluation is designed to ensure a fair and consistent comparison between the Swin-Tiny and ViT-Small models for multi-class retinal disease classification. Both models are trained and evaluated under identical conditions, with the same dataset splits, preprocessing pipeline, training strategy, and evaluation metrics.

All experiments are implemented using the PyTorch deep learning framework. Transfer learning is employed by initializing both transformer models with weights pretrained on the ImageNet dataset. This approach enables faster convergence and improves feature generalization,

particularly given the limited size of medical imaging datasets. The pretrained backbones are fine-tuned end-to-end using the retinal fundus images.

The input images are resized and normalized along each color channel, with a size of 224 × 224 pixels when converted to model, according to the preprocessing process used in Section 3. To prevent leakage, we split the dataset into training and testing subsets in a class-wise manner. We do no other domain-specific preprocessing in either training or evaluation.

The model is trained in a supervised learning setting with the loss function defined on the categorical cross entropy suitable for single-label multi-class classification. The parameters are updated using an adaptive gradient based optimizer Both models have trained for a certain number of epochs with the same batch sizes to maintain consistency between experiments. The models were subjected to the same learning rate schedule and regularization techniques to maintain a solid training process, as well as prevent them from overfitting.

To accelerate computation, we perform experiments in a GPU-enabled environment. Evaluation is performed solely on unseen test data without any additional fine-tuning of the model. We evaluate performance according to standard classification metrics, like accuracy, macro-averaged precision, macro-averaged recall and macro-averaged F1-score, for a balanced evaluation over all disease classes. To facilitate a holistic comparison between Swin-Tiny versus ViT-Small, model complexity metrics (i.e., parameters and inference) are also included along with classification performance.

6. Results and Comparative Analysis

In this section we conduct a comparative evaluation of the Swin-Tiny and ViT-Small models on multi-class retinal disease classification. We evaluate both models with the same training and testing setups for a fair comparison. We use standard classification metrics, namely accuracy as well as macro-averaged precision, macro-averaged recall, and macro F1-score to evaluate performance. Moreover, trade-offs are emphasised with the analysis of model complexity and efficiency characteristics that are relevant to real-world deployment. Figure 3 and 4 compares training accuracies and losses between the two models respectively.

ISSN: 2096-3246
Volume 58, Issue 01, 2026

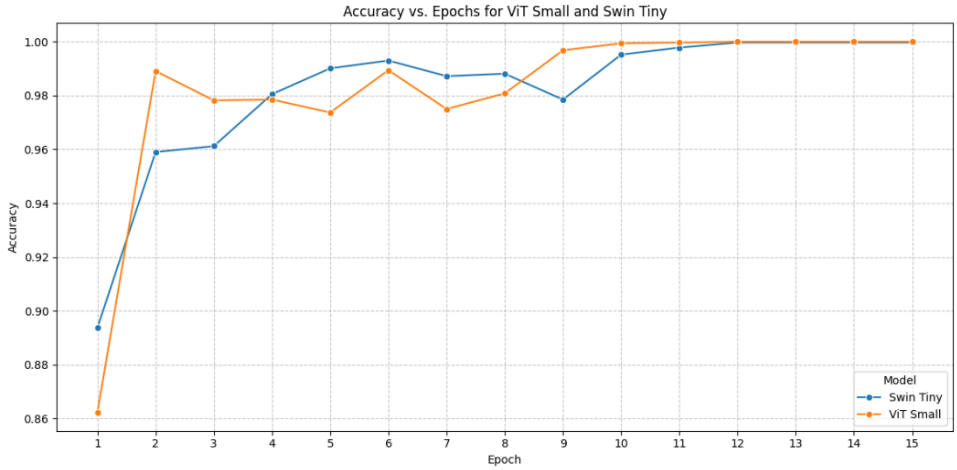


Fig:3 Training accuracy curves of Swin-Tiny and ViT-Small across epochs, demonstrating convergence behavior and comparative learning stability.

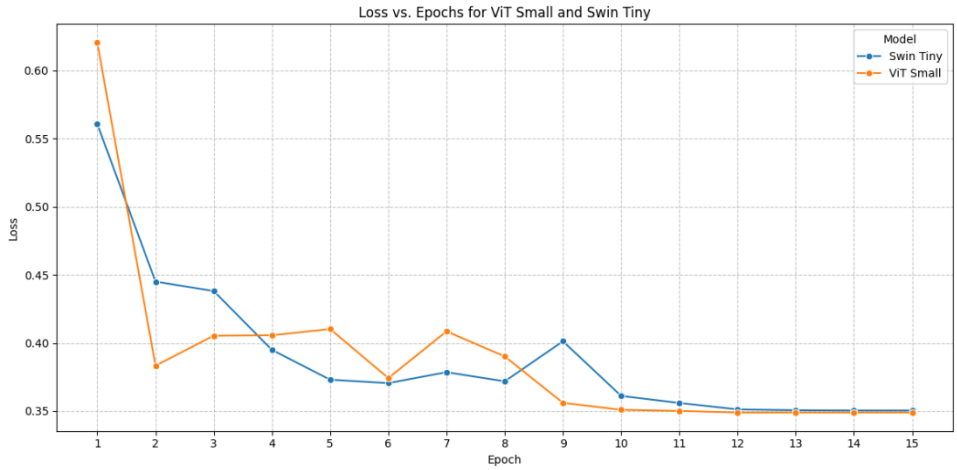


Fig:4. Training and validation loss curves of Swin-Tiny and ViT-Small across epochs, illustrating optimization dynamics and overfitting characteristics.

6.1 Quantitative Performance Comparison

The classification results are shown in Table 1, using Swin-Tiny and ViT-Small on the test dataset. To properly weigh each disease class the same, thereby eliminating biasing towards classes with a larger sample size, macro-averaged metrics are reported. Figure 5 Visualizes the performance metrics of both models.

Table 2. Performance comparison of lightweight vision transformer models

Model	Accuracy	Precision	Recall	F1-Score
Swin-Tiny	0.988	0.985	0.989	0.987
ViT-Small	0.978	0.969	0.979	0.973

Both models perform excellently on all metrics indicating that lightweight vision transformers are effective when it comes to the classification of retinal diseases. Overall, Swin-Tiny achieves slightly higher accuracy and F1-score than ViT-Small, showing its superiority in learning discriminative multi-scale features from fundus images. On the other hand, ViT-Small obtains competitive performance with fewer parameters than EGL, indicating good parameter efficiency of our model.

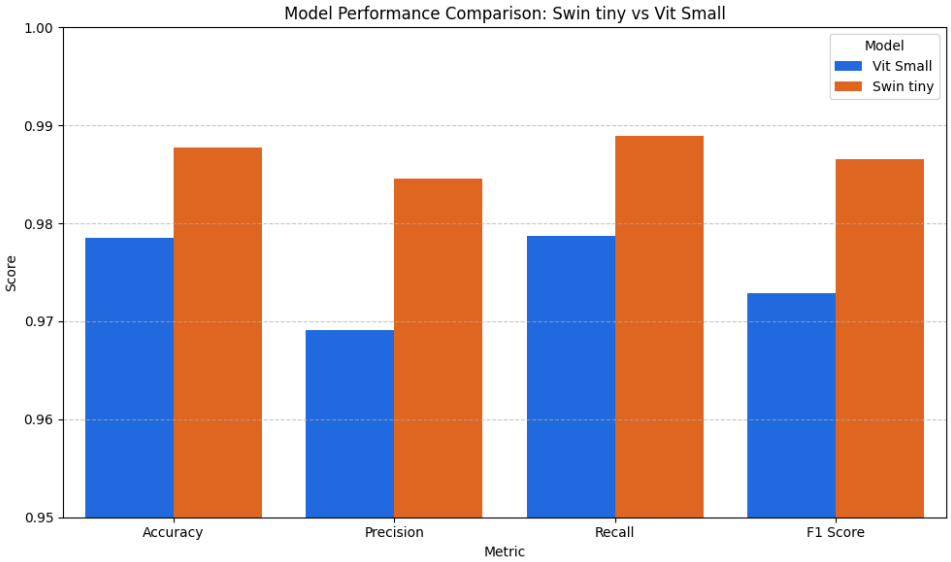


Fig:5.Comparative performance of Swin-Tiny and ViT-Small models on the retinal disease test dataset in terms of accuracy, precision, recall, and F1-score.

6.2 Efficiency and Computational Analysis

Besides accuracy, computational efficiency is a major factor for model selection in clinical screening systems. Swin-Tiny utilizes window-based selfattention and hierarchical feature representations, allowing to alleviate the computational cost of global attention methods. Even though Swin-Tiny has a larger number of parameters than ViT-Small, the localized attention allows us to have faster inference on high-resolution retinal images.

ViT-Small has better parameters efficiency but uses global self-attention among all image patches for each transformer layer. This design increases the per-layer computational cost, and can lead to a slower inference latency when processing large batches or operating on constrained hardware. Its simpler architecture with smaller memory footprints could make it more attractive in situations where the number of parameters and a large model size are the primary concerns.

Optional efficiency indicators such as average inference time per image and GPU memory usage further highlight these trade-offs. In practical settings, Swin-Tiny demonstrates more stable inference performance, whereas ViT-Small offers a compact alternative with marginally reduced accuracy.

6.3 Class-wise Performance Analysis

Analysis of confusion matrices reveals that both models perform reliably across all four Confusion matrix analysis shows that both models are reliable for all four categories of retinal diseases. The majority of errors occur between disease categories that are related to each other, which is due to the similarity in pathological patterns of certain diseases in fundus images. Swin-Tiny has better discrimination capabilities for subtle characteristics of diseases, which can be attributed to the hierarchical attention mechanism of the model that is capable of capturing both local and global context of retinal images. ViT-Small has occasional confusion in situations where local feature differentiation is required. Figure 6 and 7 shows the confusion matrices of Swin tiny and Vit small models respectively.

Overall, the comparative results indicate that both Swin-Tiny and ViT-Small are viable candidates for lightweight retinal disease classification. The observed performance and efficiency differences motivate further discussion on model selection based on deployment requirements, which is explored in the following section.

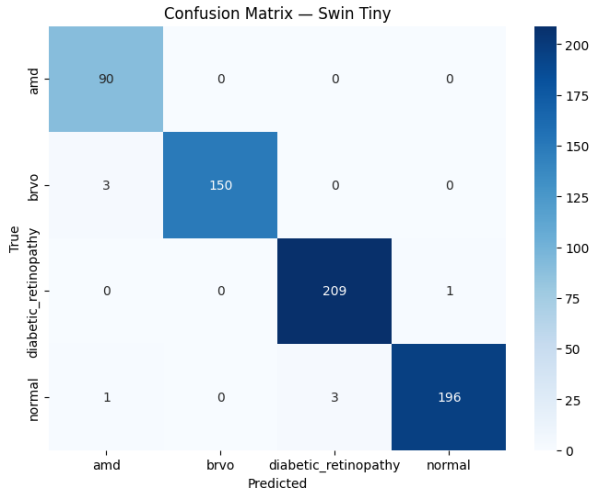


Fig.6. Confusion matrix of the Swin-Tiny model on the test dataset showing class-wise prediction performance across Normal, DR, AMD, and BRVO categories.

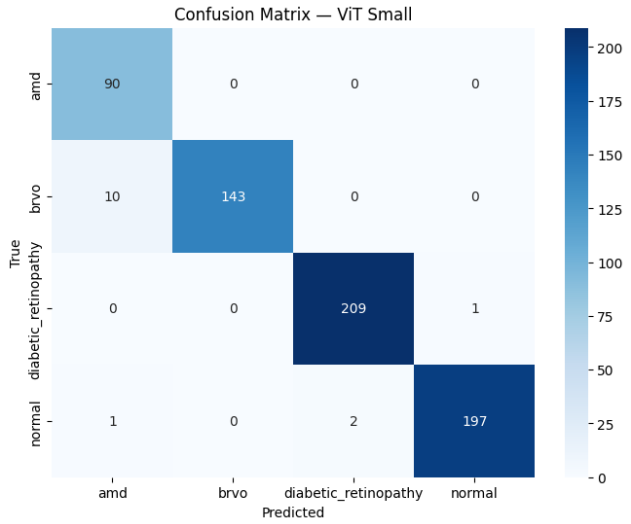


Fig.7. Confusion matrix of the ViT-Small model on the test dataset highlighting inter-class misclassification patterns among retinal disease categories.

7. Discussion

The comparative analysis of Swin-Tiny and ViT-Small brings forth significant considerations regarding the trade-offs between classification accuracy and computational complexity, which are critical to practical retinal screening tasks. While both Swin-Tiny and ViT-Small show excellent ability in multi-class retinal disease classification, their differences in architecture cause them to have different advantages and disadvantages.

Swin-Tiny has a slightly better performance in terms of accuracy, recall, and F1-score. This is because Swin-Tiny has a hierarchical architecture and a window-based self-attention mechanism that helps it learn multi-scale features efficiently. Retinal fundus images tend to have both local and global pathologies, such as microaneurysms and hemorrhages, and larger abnormalities of the macula and retinal vasculature. The shifted window attention mechanism in Swin-Tiny helps the model focus on both local and global information in the images, resulting in better separation of visually similar disease classes.

On the other hand, ViT-Small relies on the global self-attention mechanism among all the patches in the image at each layer of the transformer architecture. While this design makes it relatively easier to directly capture long-range dependencies in the image, it may not be as effective in capturing localized details of the image that are of high importance in distinguishing certain types of retinal diseases. Additionally, the absence of hierarchical feature aggregation may make it difficult for the model to capture the multi-scale geometry of the retina, which is commonly observed in fundus images.

In terms of efficiency, ViT-Small has advantages in terms of the number of parameters and model complexity. Since it has fewer parameters to be trained, it consumes less memory and is potentially easier to deploy on memory-limited hardware. Nevertheless, its full self-attention mechanism incurs higher computational overhead per inference, which can be a concern for real-time diagnosis applications. Swin-Tiny, which has a higher parameter overhead, enjoys the advantage of localized attention computation, leading to relatively more balanced inference efficiency when handling high-resolution retinal images.

The results indicate that the selection between Swin-Tiny and ViT-Small should be based on the deployment conditions. Swin-Tiny is more appropriate in conditions where the classification performance and robustness to various retinal pathologies are of prime importance. ViT-Small can be considered in scenarios where the importance of compactness and lower storage requirements is of high significance.

In conclusion, this comparative analysis study clearly indicates that vision transformers with low complexity can perform equally well in retinal disease classification tasks and also resolve the issues of efficiency. It is important to comprehend the architectural differences among various low-complexity transformer models to develop efficient retinal screening solutions.

8. Conclusion

This paper has offered a comparative assessment of two lightweight vision transformer models, namely Swin-Tiny and ViT-Small, for multi-class classification of retinal diseases from color fundus images. The main aim of this paper has been to critically assess the trade-offs between the classification accuracy and computational complexity of the proposed models in order to select suitable models for implementation in efficiency-constrained retinal disease screening applications.

The experimental outcomes have shown that Swin-Tiny and ViT-Small are effective in achieving high classification accuracy for four retinal disease classes, namely Normal, Diabetic Retinopathy, Age-related Macular Degeneration, and Branch Retinal Vein Occlusion. Swin-Tiny has been observed to marginally outperform ViT-Small in terms of classification accuracy and robustness, which can be attributed to its hierarchical architecture and window-based self-attention mechanism capable of capturing multi-scale retinal image features. ViT-Small, on the other hand, has been found to be more compact with lower parameter complexity, which is advantageous for applications where model size and storage complexity are important implementation constraints.

The comparative study emphasizes that there is no universally best lightweight transformer model for all application scenarios. Rather, the choice of the transformer model depends on the requirements of the respective application, such as accuracy, computational power, and real-time

processing capabilities. Swin-Tiny is more suitable for applications where the diagnostic accuracy is of prime importance, whereas ViT-Small is a decent alternative for resource-constrained systems.

In conclusion, the results of this research study offer valuable information regarding the application of lightweight vision transformer models for the automatic retinal disease diagnosis system. Future research studies may focus on the extension of this comparative study to other variants of the transformer models, testing the performance of the models on larger datasets, and evaluating the performance of the models in real-world applications.

References

1. V. Gulshan et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.
2. D. S. W. Ting et al., "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations," *JAMA*, vol. 318, no. 22, pp. 2211–2223, 2017.
3. A. Grassmann, J. Mengelkamp, C. Brandl, S. Harsch, M. E. Zimmermann, and A. Linkohr, "A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography," *Ophthalmology*, vol. 125, no. 9, pp. 1410–1420, 2018.
4. J. Quellec, K. Charrière, Y. Boudi, B. Cochener, and G. Lamard, "Deep image mining for diabetic retinopathy screening," *Med. Image Anal.*, vol. 39, pp. 178–193, 2017.
5. W. Li et al., "A deep learning system for automated detection of diabetic retinopathy using color fundus photographs," *Comput. Biol. Med.*, vol. 99, pp. 92–101, 2018.
6. J. Krause et al., "Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy," *Ophthalmology*, vol. 125, no. 8, pp. 1264–1272, 2018.
7. A. Poplin et al., "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning," *Nat. Biomed. Eng.*, vol. 2, no. 3, pp. 158–164, 2018.
8. J. De Fauw et al., "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nat. Med.*, vol. 24, no. 9, pp. 1342–1350, 2018.
9. A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.
10. Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. ICCV*, pp. 10012–10022, 2021.
11. H. Cao et al., "Swin-Unet: Unet-like pure transformer for medical image segmentation," in *Proc. ECCV Workshops*, pp. 205–218, 2022.

12. J. Chen et al., “TransUNet: Transformers make strong encoders for medical image segmentation,” arXiv:2102.04306, 2021.
13. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. CVPR*, pp. 770–778, 2016.
14. [14] M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proc. ICML*, pp. 6105–6114, 2019.
15. H. Touvron et al., “Training data-efficient image transformers and distillation through attention,” in *Proc. ICML*, pp. 10347–10357, 2021.
16. X. Li, D. Xie, J. Li, and L. Shen, “Vision transformer-based classification of retinal diseases from fundus images,” *IEEE Access*, vol. 10, pp. 45678–45689, 2022.
17. S. Yu et al., “MIL-VT: Multiple instance learning enhanced vision transformer for fundus image classification,” in *Proc. MICCAI*, pp. 45–54, 2021.
18. K. Xu et al., “Automatic detection and differential diagnosis of age-related macular degeneration using hierarchical vision transformers,” *Comput. Biol. Med.*, vol. 167, p. 107573, 2023.
19. P. Dutta, K. A. Sathi, M. A. Hossain, and M. A. A. Dewan, “Conv-ViT: A convolution and vision transformer-based hybrid feature extraction method for retinal disease detection,” *J. Imaging*, vol. 9, no. 7, p. 140, 2023.
20. Y. Jiang et al., “Computer-aided diagnosis of retinopathy based on vision transformer,” *J. Innov. Opt. Health Sci.*, vol. 15, no. 2, p. 2250006, 2022.