www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 15(4), Nov 2024, 72-84. ISSN: 2007-1558. https://doi.org/10.61467/2007.1558.2024.v15i4.522.

Early detection of age-related macular degeneration using Vision Transformer-based Architectures – A comparative study with offline metrics and data augmenting

Augusto Javier Reyes Delgado[1], Jorge Ernesto González Díaz[1], José Luis Sánchez Cervantes[1], Giner Alor Hernández[1], Lisbeth Rodríguez Mazahua[1], Adolfo Rodríguez Parada[2], Yara Anahí Jiménez Nieto[2]*

[1] Tecnológico Nacional de México/I. T. Orizaba, 94320, México.
[2] Facultad de Negocios y Tecnologías, Universidad Veracruzana. Ixtaczoquitlán, 94452, México.

E-mails: m17010207@orizaba.tecnm.mx, jose.sc@orizaba.tecnm.mx, giner.ah@orizaba.tecnm.mx, lisbeth.rm2@orizaba.tecnm.mx, adrodriguez@uv.mx, yjimenez@uv.mx

*Corresponding author: d04010291@orizaba.tecnm.mx

Abstract. Age-related macular degeneration (AMD) is one of the leading causes of vision loss in elderly adults around the world and is among the main visual impairments in Mexico. The difficulty of diagnosing AMD in its early stages motivates the use of advanced deep-learning methods that offer significant potential to improve diagnostic accuracy in retinal image analysis. In recent years, Transformer architectures for computer vision, such as Vision Transformer (ViT), Swin Transformer and BERT Pre-training of Image Transformers (BEiT) have provided a novel perspective for image analysis. This study presents a comparative analysis of these architectures, applied to AMD detection, focusing on each model's capability to classify the early stages of the disease. Although the small size of medical image datasets represented a challenge, our results suggest that ViT-based architectures and their derivatives achieve significant performance in AMD detection. BEiT is particularly notable for its consistently superior performance.

Keywords: Multiclass classification, Age-Related Macular Degeneration (AMD), Early Detection, Vision Transformers.

Article Info
Received Jul 24, 2024.
Accepted Oct 4, 2024.

1 Introduction

As is pointed out in (WHO, 2020), AMD significantly affects the life quality of life of elderly persons who suffer from such conditions, affecting their independence and capability to perform their daily activities. In Mexico, AMD has been cataloged as the third of the six leading causes of ocular issues that affect the population, as reported by the Mexican Health Secretary (World Sight Day 2020). The early detection of AMD is fundamental to prevent the advance of the disease and preserve vision. The literature suggests that integrating Deep Learning methods (mainly Convolutional Neural Networks or CNN) has improved performance in detecting and classifying medical images and, in some cases, has surpassed evaluations made by specialists (T. He et al., 2022).

With the advance of image processing and the implementation of Deep Learning techniques, new perspectives for its use in medical sciences have appeared. Image processing using CNN and its automatic analysis methods have proven to be highly efficient tools, providing intelligent and user-friendly systems for the scanning and diagnosing diseases, including AMD, outside of a clinical environment (Abd El-Khalek et al., 2024). Furthermore, different approaches have been proposed to detect the

© Editorial Académica Dragón Azteca S. de R.L. de C.V. (EDITADA.ORG), All rights reserved.

pathological characteristics of AMD captured in high-resolution images by analyzing texture patterns and colors (Leingang et al., 2023). The automatic classification of the disease level progression faces challenges, especially with those features that are subtle or like non-pathological conditions. This is complicated by the high image resolutions and the high resources required, possibly affecting the accuracy of the diagnosis and slowing down the process (Deep Learning for AMD Screening and Detection, n.d.). In this context, based on the comparative analysis presented in this article, we consider that vision transformer architectures such as ViT (Dosovitskiy et al., 2020), Swin Transformer (Liu et al., 2021), and BEiT (Bao et al., 2021) suggest a promising evolution in the detection of AMD, giving the capability to understand the complex spatial relationships in retinal images. This study presents a comparative analysis of these vision transformer architectures applied to the detection of AMD, focusing on the capability of each model to identify and classify the stages of the disease into No AMD, Mild, Moderate, and Advanced. These models may be essential to improving early detection and diagnostic accuracy of AMD, which is crucial for effective treatment and vision preservation in elderly populations, Mexico included.

This work is organized in the following sections: Section 2 shows the related works. Section 3 compares architectures, ViT, Swin Transformer, and BEiT, and provides details for their implementation in the AMD case of study. Section 4 details the experiments carried out, their results, and the comparative evaluation of the performance of the architectures included in the study. Finally, Section 5 presents the conclusions and future work.

2 Related works

Several related works have been identified in the literature. Nevertheless, they do not compare transformers with different architectures for computer vision applied to DMAE, highlighting their efficiency over CNNs.

One of the most relevant works is presented in (Tu, 2023); such work addressed the application of ViT to detect glaucoma through fundus images. The authors evaluated several architectures: ViT, Swin Transformer, Twins-PCPVT, and Class Attention on Image Transformers (CaiT). They used learning algorithms with few images, and the impact of data augmentation techniques was also analyzed. The study results showed that ViT, combined with ProtoNets, outperformed CNN-based counterparts and achieved competitive performance on benchmark data sets. In (Alayón et al., 2023), the authors evaluated the effectiveness of CNNs and Vit-based systems in detecting glaucoma in fundus images. The authors tested several CNN architectures such as VGG19, ResNet50, InceptionV3 and Convolution Enhanced Image Shaper (CeiT), Convolutional Vision Transformer (ConViT), and the ResMLP architecture. The results show that CNN and ViT performed similarly on the test set, although CNNs demonstrated better generalization on external datasets. Likewise, in (Li et al., 2023), the efficiency of ViT architectures was explored in medical imaging applications. The capabilities of ViT were compared with those of CNNs in tasks such as segmentation, recognition, and classification of medical images. Architectures such as Conformer, U-Net Transformer, and Multi-transSP were highlighted, as they showed superior effectiveness in improving precision and efficiency in various medical applications. The results showed that ViT outperformed CNNs in medical image segmentation due to its ability to model long-term dependencies and scalability. Furthermore, J. He et al. (2023) introduced a method to classify retinal diseases using optical coherence tomography (OCT) images. In this work, a Swin-Poly Transformer network was proved. The findings indicated that the proposed method facilitated accurate and efficient retinal classification and highlighted the value of artificial intelligence in ophthalmological diagnoses and the potential of ViT networks in the medical area. Similarly, in (Nafisah et al., 2023), the researchers compared CNNs and ViT architectures to classify chest radiographs (CXR) in COVID-19 cases, viral pneumonia, and healthy cases. This study used the COVID-QU-Ex dataset, randomly splitting 80% for training and 20% for testing. They evaluated the effectiveness in balanced and unbalanced cases, implementing ViT models such as Twins, Swin, and Segformer. The results showed that the CNN and ViT-based models had similar performance, with a maximum accuracy of 99.82% for EfficientNetB7 (CNN) and outstanding performance for SegFormer (ViT). In work (Ma et al., 2022), the authors evaluated the performance of ViT architectures, specifically ViT-B and Swin-B, in medical image classifications, contrasting their effectiveness with models based on CNNs to diagnose diseases such as thoracic diseases, pulmonary embolisms, and tuberculosis, using x-rays and CT scans. This work proposed that adequate initialization is essential for Vision Transformers in the medical field and that self-learning approaches that use mutual information generate more accurate representations for medical classification. In the same sense, the authors of Mallick et al. (2022) explored the use of ViT, Swin Transformer, and ConvNext by applying transfer learning techniques to detect Glaucoma from fundus images. This effort sought to create an automated method that allowed the identification of Glaucoma in its early stages to prevent blindness. Finally, (Wassel et al., 2022), Wassel et al. reported a study that classified glaucomatous ocular conditions using ViT-based models, using whole and cropped optic disc fundus images. ViT, Swin, CaiT, CrossFit, XciT, ResMlp, and DeiT were evaluated in both cases, individually and in assemblies. In addition to glaucoma, they addressed other ophthalmological diseases such as diabetes, cataracts, hypertension, pathological myopia, and other anomalies. Their results showed that Swin and CaiT obtained the highest precision, sensitivity, and specificity levels in validating and testing the combined data sets, underscoring their effectiveness for glaucoma detection in ophthalmological images and suggesting their potential usefulness in clinical practice. The following table shows the summary data of the related works identified in the literature for this research work.

Table 1. Related works on comparisons of ViT architectures in medicine.

Author	Architecture	Disease	Image Type
Nurgazin M. et al. (2023)	Variantes del ViT clásico: ViT_tiny ViT_small ViT_base	Melanoma, Basal cell carcinoma, Squamous cell carcinoma, Nevus, Actinic keratosis, Dermatofibroma, Epidermoid cyst, Psoriasis, Atopic dermatitis, Rosacea, Breast cancer.	Skin lesions. Breast tissue biopsies. Cervical cytology.
Alayon S. et. al. (2023)	ViT, Swin Transformer, Twins-PCPVT, CaiT.	Glaucoma	Fondus.
Li J. et al. (2023)	Conformer, U-Net Transformer, Módulo Residual Transformer Multi-transSP, TransPath, i-ViT BabyNet	Fetal weight prediction. Detection of diabetic retinopathy. Knee cartilage segmentation.	Ultrasound. Magnetic resonance imaging, CT scan, X-ray, Histopathology.
He J. et al (2023)	ViT Swin Transformer	Diabetic retinopathy, Diabetic macular edema, Glaucoma, Ocular abnormalities.	Optical coherence tomography (OCT).
Nafisah S. et al. (2023)	Twins, Swin, Segformer	COVID-19 Pneumonia	Chest x-rays (CXR)
Ma D. et al. (2022)	ViT-B, Swin-B.	Chest diseases, Pulmonary embolism, Tuberculosis.	Chest x-rays, CT scans.
Mallick S. et al. (2022)	ViT, Swin Transformer, ConvNext	Glaucoma	Fondus.
Wassel M. et al. (2022)	Cait, crossViT, XciT, ResMlp, DeiT, ViT	Glaucoma, Diabetes, Cataracts, Hypertension, Pathological Myopia	Fondus.

3 Methods and Implementation

In this section, the compared transformer architectures are depicted, and the details for their implementation in the study, namely Dry-AMD, are described. This includes the acquisition of the image datasets, the preprocessing applied to them, and the parameters for the training process. The pipeline of the process is summarized in the next figure.

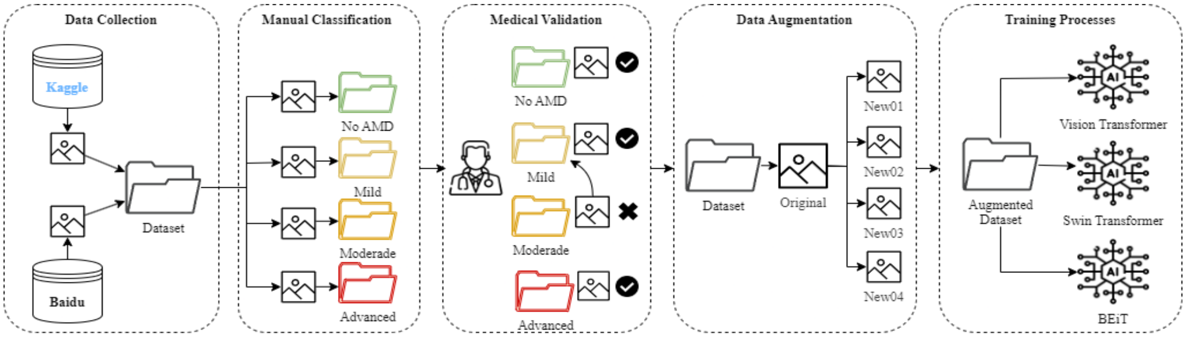


Fig. 1 Process pipeline.

3.1 Transformer Architectures

In this study, a detailed comparison between the ViT, Swin Transformer, and BEiT architectures for the detection and multiclass classification of AMD in fundus images is made. The classification categories used are No AMD, mild, moderate, and advanced AMD. The selected Deep Learning technologies were chosen for their potential to effectively process intricate visual features critical to discerning the different stages of AMD, which is crucial to achieving an early and accurate diagnosis. Besides, these architectures present an advanced ability to capture global and local image patterns, essential for reliable detection and accurate classification of AMD progression.

3.1.1 Vision Transformer (ViT)

ViT (Dosovitskiy et al., 2020) is an innovative architecture that applies the transformer mechanism, which is common in language processing, to computer vision. ViT breaks images into patches and processes them as tokens in a sequence. It uses attention to weigh the importance of different parts of the image, allowing the model to capture complex patterns and long-distance relationships. Its focus on global relationships makes it especially suitable for identifying patterns in medical images, such as those related to AMD, where the manifestations of the disease may be subtle and distributed throughout the image.

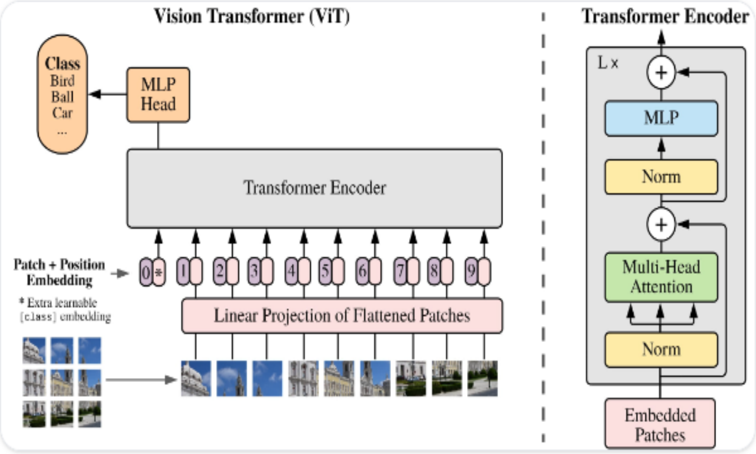


Fig. 2 Original ViT architecture (extracted from (Dosovitskiy et al., 2020))

3.1.2 Swin Transformer

Swin Transformer was introduced by Ze Liu et al. in their work "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows" (Liu et al., 2021). The Swin Transformer arises in response to some limitations of pure transformer models such as ViT, especially regarding computational efficiency and the capacity to handle varying image sizes. Although ViT demonstrated that transformers could be decisive for vision tasks, its approach of treating the image as a sequence of fixed patches posed challenges in terms of scalability and adaptability to different resolutions and image sizes.

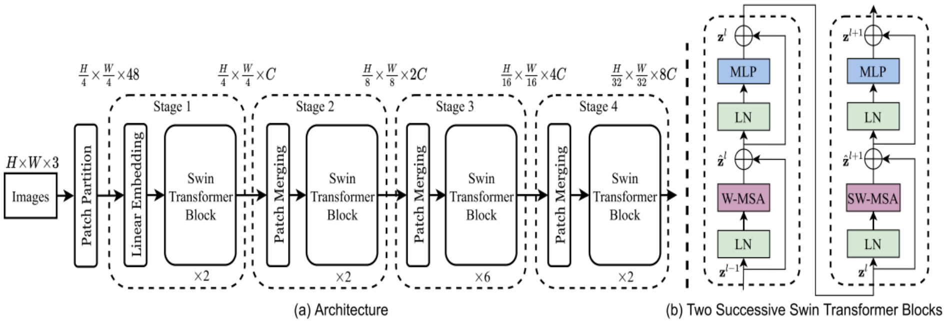


Fig. 3. Swin Transformer architecture (extracted from (Liu et al., 2021))

The Swin Transformer introduces several innovative concepts to address these limitations:

- **Shifted Windows:** One of the key innovations of the Swin Transformer is its use of shifted windows. It divides the image into non-overlapping windows for local attention, which reduces computational complexity.
- **Hierarchy:** As in CNNs, the Swin Transformer processes images at various resolutions. It starts with high resolution and progressively reduces it, allowing the model to capture features at different scales and improve efficiency by reducing resolution in deeper layers.
- **Flexibility and Generality**: Unlike ViT, which uses fixed-size patches, the Swin Transformer can more effectively handle different image sizes and resolutions, making it more flexible and adaptable for various vision applications. by computer.

The Swin Transformer's selection is justified by its design, which efficiently addresses hierarchy and locality in images. Unlike ViT, which considers the entire image globally, the Swin Transformer processes images in local windows, allowing a more detailed representation of local features.

3.1.3 BEiT (BERT Pre-training of Image Transformers)

It was presented in a work by (Bao et al., 2021) titled "BEiT: BERT Pre-training of Image Transformers." BEiT is inspired by the success of BERT (Bidirectional Encoder Representations from Transformers) in Natural Language Processing (NLP). BERT revolutionized NLP by pre-training transformers on large text corpora using hidden word prediction tasks, where the model learns to predict parts of the text that have been intentionally hidden. BEiT brings this pre-training approach to the image domain. Instead of predicting hidden words, BEiT is trained to predict hidden parts of an image. This process involves two main stages:

- **Image Tokenization:** BEiT converts an image into a set of visual tokens using an image tokenization model (such as a VQ-VAE, a quantized variational autoencoder). This results in a representation of the image in tokens, similar to how text is tokenized in NLP.
- **Model Pre-Training:** The model is pre-trained to predict visual tokens from hidden parts of the image, similar to the prediction of missing words in BERT. This teaches the model to understand and predict visual structure and content based on the context provided by the visible parts of the image.

The use of BEiT in this study is justified by its focus on learning visual representations by predicting hidden pixels. This innovation for analyzing fundus images in AMD allows BEiT to capture subtleties in the textures and patterns of images, which are crucial for identifying the stages of AMD.

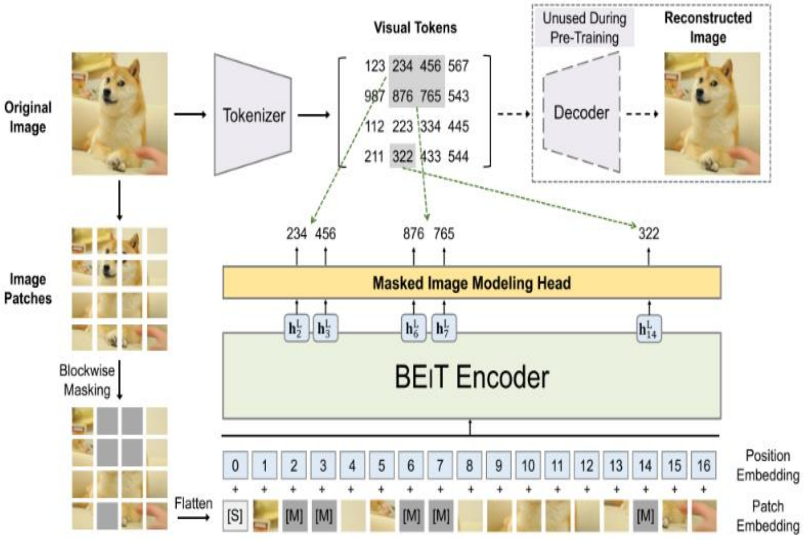


Fig. 4. BEiT architecture (Extracted from (Bao et al., 2021))

3.2 Implementation Details

For comparison, models of the three architectures were trained through a fine-tuning process due to the limited size of the dataset and the computational resources available. Multiple iterations were carried out to tune the hyperparameters of each model. The same set was used for all the training processes. Most of the implementation aspects are detailed below.

3.2.1 Hardware Specifications

The ViT and Swin Transformer models were implemented in Google Colab®, using GPU acceleration. The GPU available in Colab® was a NVidia Tesla T4 with 15GB of VRAM. In contrast, due to resource limitations in Colab®, the BEiT model was trained on a desktop computer with an Nvidia RTX 3070 8GB GPU and an AMD Ryzen 5 3500X CPU. The PyTorch Transformers library was used for the training process of all the models.

3.2.2 Datasets

To build the dataset, 305 images were initially used, divided into 185 for training and 60 for validation and testing, following a 60% / 20% / 20% distribution. The images were obtained from the iChallenge-AMD dataset (Broad (Baidu Research Open-Access Dataset), 2020) and a set on Kaggle published by Mujib (Rakhshanda Mujib, 2023), which includes AMD images extracted from several fundus image sets with retinal pathologies. The classification of the images was based on existing literature (U.S. Department of Health and Human Services, n.d.) (Al-Zamil & Yassin, 2017). The labels for the different classes are: 1. Mild, 2. Moderate, 3. Advanced & 4. No AMD. After the manual classification of the images, the classification was corrected and validated by experts in the medical area (CONDE Investigación – Unidad de Investigación, n.d.).

Data Augmenting

Data augmentation techniques are applied to the training batch to enhance model generalization and reduce overfitting due to the limited dataset size. Such transformations include resizing, rotations, and brightness and contrast adjustments. This increased the set to 1,094 images, with 974 dedicated to training.

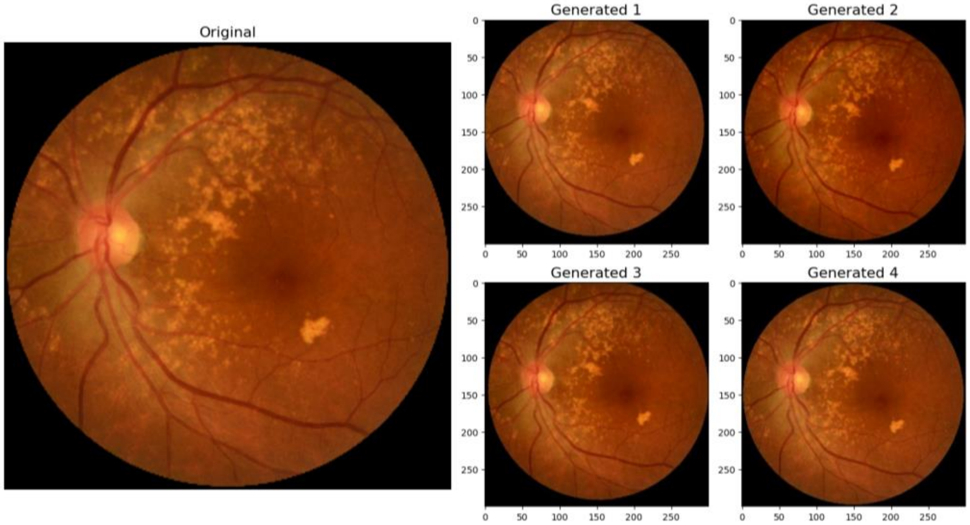


Fig 5. Example of the transformations applied.

Several data augmenting processes were made, which gave different results in the training process. Finally, after several proofs, the transformations applied to the original set were not drastic enough to maintain the subtle patterns in the images; it turned out that very drastic changes in the original images produced noise in the training process.

3.2.3 Training Parameters

Multiple experiments were carried out to determine the optimal hyperparameters, concluding that 42 epochs and batches of 32 images were the most appropriate. Increasing the number of epochs beyond 42 did not generate significant improvements in performance, identifying a plateau in performance around 20 epochs. Additionally, the learning rate was set to 5e-05 after extensive evaluations.

Table 2. Hyperparameters in the models' training process.

Hyperparameter	Value
Learning rate	5e-05
Train batch size	32
Evaluation batch size	32
Seed	42
Optimizer	Adaptive Moment Estimation (Adam)
Warmup ratio	0.1
Epochs	42

4 Experiment Results

The following section shows the results obtained from the data and parameters previously described, in addition to the graphics and metrics used to evaluate the trained models.

4.1 Results

The accuracy metric of any model trained was recorded. An increase in this metric was observed until the 40th epoch. This performance was monitored and documented using the Weights & Biases® (W&B) platform, as illustrated in the image below.

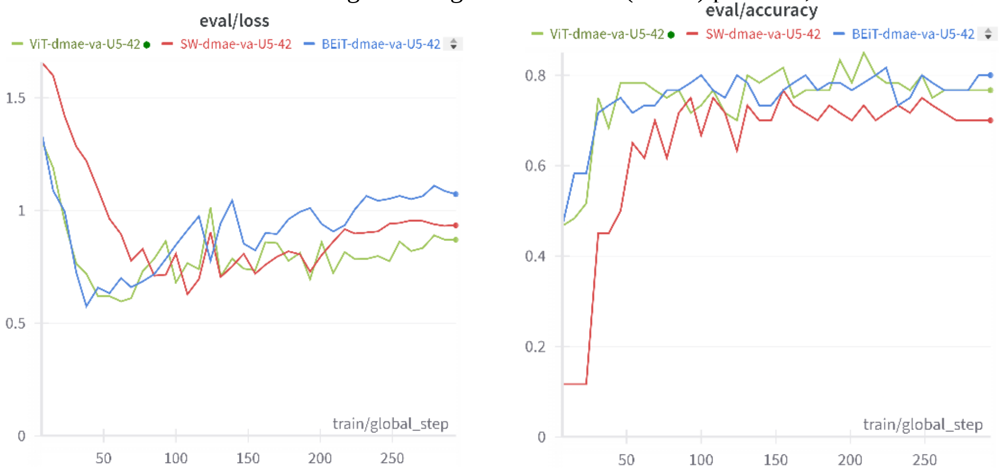


Fig. 6. Results obtained by the models during the training process.

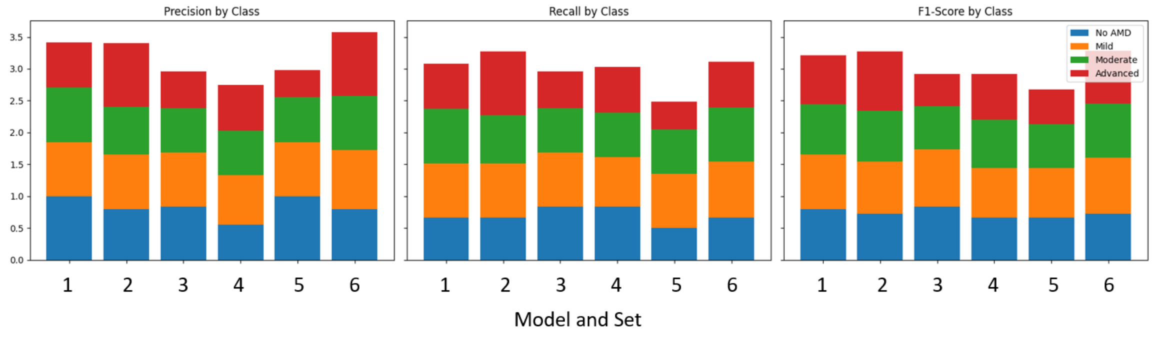
The presented models achieved an accuracy greater than 0.7500. To strengthen the evaluation of the results, additional performance metrics were calculated, including precision, sensitivity, and F1 Score, allowing a deeper analysis of each model's capabilities. These metrics were obtained using the validation and test sets previously separated from the initial data set. The results of these metrics are presented in the following table.

Table 3. Results in the evaluation metrics of the different models.

Model	Accuracy	Set	Class	Precision	Recall	F1-Score
ViT	**0.8500**	Validation	No AMD	0.8000	0.6666	0.7272
			Mild	0.8333	**0.9259**	**0.8771**
			Moderate	0.8500	0.8500	0.8500
			Advanced	1.0000	0.7142	0.8333
	0.7166	Test	No AMD	1.0000	0.5000	0.6666
			Mild	0.7187	**0.8518**	**0.7796**
			Moderate	0.6666	0.7000	0.6829
			Advanced	0.7500	0.4285	0.5454
Swin Transformer	0.7500	Validation	No AMD	0.5555	**0.8333**	0.6666
			Mild	0.7777	0.7777	**0.7777**
			Moderate	0.8235	0.7000	0.7567
			Advanced	0.7142	0.7142	0.7142
	0.7666	Test	No AMD	0.8333	0.8333	0.8333
			Mild	0.9583	**0.8518**	**0.9019**
			Moderate	0.6666	0.7000	0.6829
			Advanced	0.4444	0.5714	0.5000
BEiT	**0.8166**	Validation	No AMD	0.8000	0.6666	0.7272
			Mild	0.7931	0.8518	0.8214
			Moderate	0.8333	0.7500	0.7894
			Advanced	0.8750	**1.0000**	**0.9333**
	0.8166	Test	No AMD	1.000	0.6666	0.8000
			Mild	0.8518	**0.8518**	**0.8518**
			Moderate	0.7391	0.8500	0.7906
			Advanced	0.8333	0.7142	0.7692

The results presented in the table above show notable differences in the models' performance. By averaging the precision results of the models for the validation and test sets on which they were evaluated, ViT reaches an accuracy of 0.7833, Swin Transformer obtains 0.7583, and BEiT reaches 0.8166.

Importantly, for all models, the performance in the classification of fundus images of the "mild" and "moderate" classes is consistent and shows better results compared to the "No dmae" and "moderate" classes. Advanced", which presents a more significant variance. The Figs. 7 to 11 show the comparative graphs of the primary metrics applied to the models.



1 BEiT Test – 2 BEiT Validation – 3 Swin Transformer Test – 4 Swin Transformer Validation – 5 ViT Test – 6 ViT Validation

Fig. 7. Precision, sensitivity, and F1-score comparison by class of different models.

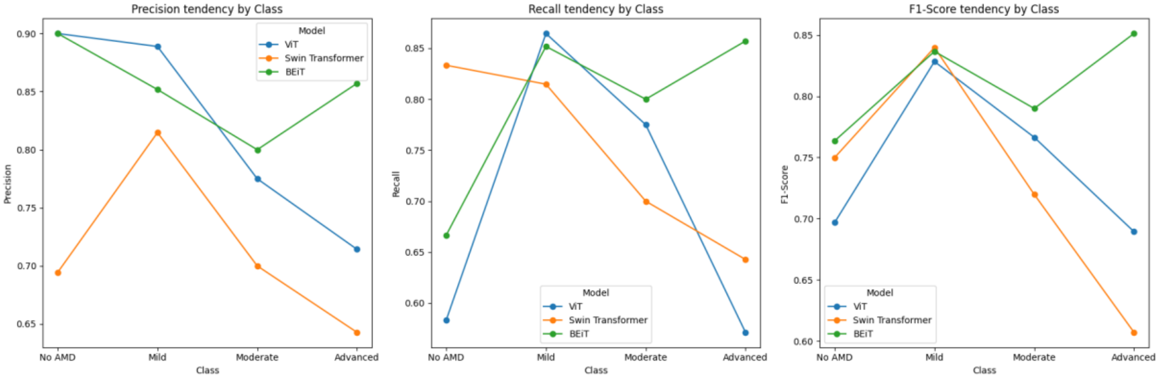


Fig. 8. Comparison of trends by class and model of the model evaluation metrics.

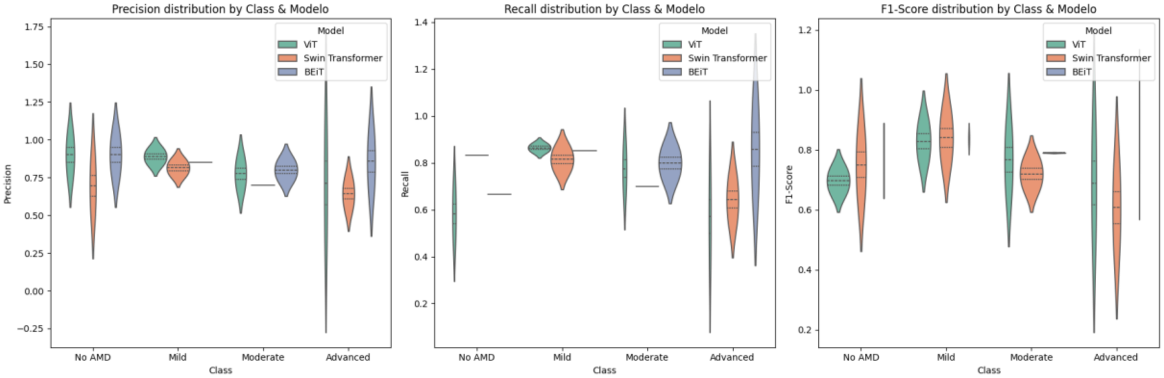


Fig. 9. General distribution by model of the metrics used in the comparative study.

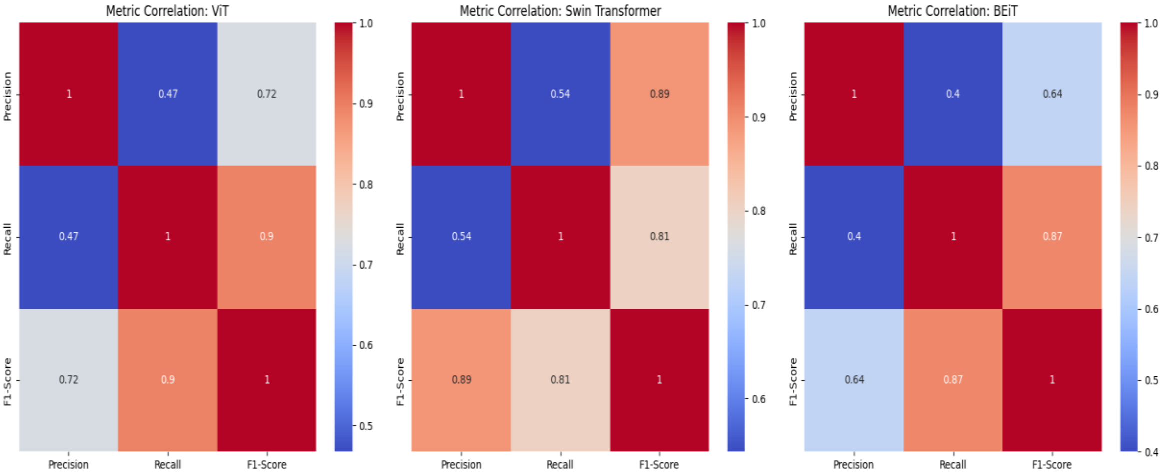


Fig. 10. Correlation matrices of metrics of the evaluated models.

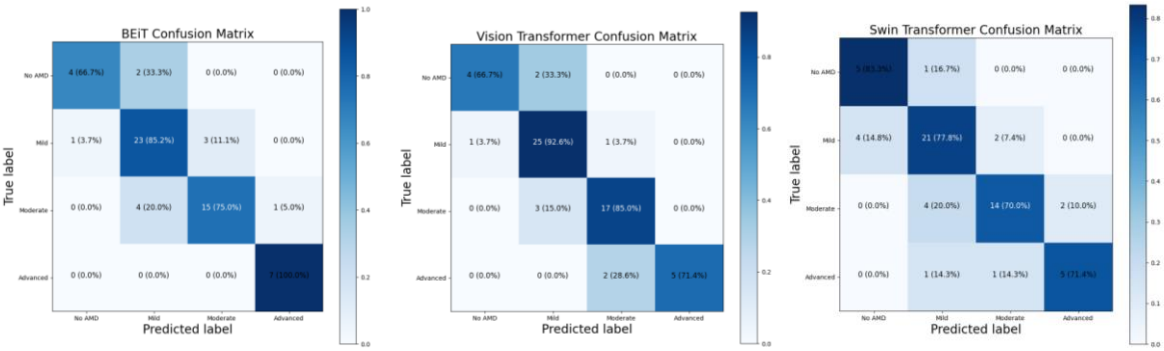


Fig. 11. Confusion matrices of the evaluated models (validation set).

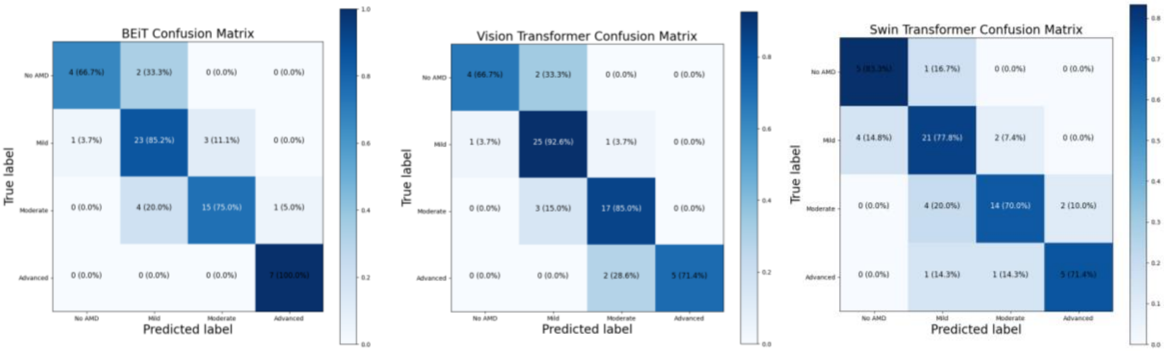


Fig. 12. Confusion matrices of the evaluated models (test set).

4.3 Computational Efficiency

The purpose of creating machine learning mechanisms is to be helpful as a support tool for ophthalmological diagnosis, so it is essential to consider the computation efficiency of the different models generated. The knowledge of computational efficiency provides perspectives about how scalable, flexible, and adaptable the models are for their implementation in medical applications. Next, the metrics were extracted from the Weights & Biases® (W&B) platform (memory consumption and training time) or obtained using the Google Colab® platform.

Table 4. Computational Efficiency of the different models.

Model	Training time (minutes)	Inference Time (ms/image)	Memory Consumption (GB)	Parameters(M)	FLOPs (GFLOPs)
Vision Transformer	24.25	**7.95**	6.59	86.39	**17.58**
Swin Transformer	**9.05**	17.64	**5.35**	27.52	4.51
BEiT	28.10	11.74	7.08	85.76	17.58

Due to the models being trained in different hardware environments, it's necessary to make a distinction for the training time metric, as the GPUs used during the training process were different. In addition, the inference time may vary depending on the hardware used to run the model. In this case, the GPU used was the Nvidia Tesla T4 available in Colab®.

4.3 Discussion

Our comparative study evaluated three advanced models: ViT, Swin Transformer, and BEiT. The results indicate significant variations in each model's performance, underscoring the importance of architecture selection in clinical applications.

The ViT model exhibited high accuracy on the validation set, although a decrease in its performance was observed when evaluated on the test set. It particularly stood out in classifying advanced cases of macular degeneration within the validation set, which suggests that this architecture has a specific aptitude for identifying severe manifestations of the disease. However, a reduced sensitivity was recorded in detecting cases in the early stages, which could reflect a predisposition towards overclassification in the most serious phases.

In contrast, the Swin Transformer presented a slightly lower overall accuracy than the ViT, which was especially notable on the validation set. This architecture faced challenges in accurately classifying advanced cases in the test set, evidenced by its low sensitivity and F1 score in said category. However, it showed competitive performance in identifying the initial stages of the disease, which indicates its potential usefulness in early detection.

BEiT proved to be the most effective model in the test set, outperforming the ViT and Swin Transformer models regarding overall accuracy. Despite initially inferior performance to the ViT in the validation set, the BEiT showed notable consistency between both sets and a significant improvement in detecting all stages of the disease compared to the ViT during testing. This reveals a superior generalization capacity and robustness, positioning BEiT as a promising alternative for the practical detection of macular degeneration in its various stages.

The variability in performance between these models highlights the complexity of applying Deep Learning to medical diagnoses. While ViT and Swin Transformer offer advantages in detecting specific stages of the disease, BEiT shows a balance between sensitivity and accuracy over a broader range of conditions. The above highlights the need to consider multiple factors, such as accuracy, sensitivity, and specificity, when selecting a Deep Learning model to detect ophthalmological diseases.

5 Conclusions

Although ViT continues to present high performance, BEiT is a more consistent alternative in the present case study.

The average accuracy results for the validation and test sets show that BEiT leads with an accuracy of 81.66%, followed by ViT with 78.33%, and Swin Transformer with 75.48%. These figures reflect the generalization capacity of each model and its reliability in recognizing patterns associated with specific ocular conditions.

Interestingly, the models show consistency in performance in classifying conditions classified as “mild” and “moderate.” This phenomenon indicates that the visual features present in these stages of the disease are more distinctive and, therefore, more easily recognized by deep learning models. On the other hand, the “No dmae” and “Advanced” categories exhibit more significant variability in the results, which suggests that the visual manifestations of these stages may be more subtle or less differentiated, thus making accurate classification difficult.

The findings underline the importance of model selection in artificial intelligence-based medical diagnostic applications. Although BEiT outperforms in general performance due to its balance between precision and generalization capacity, ViT still shows superior performance in some instances. Despite slightly lower performance, the Swin Transformer could still be valuable in a clinical context when combined with other modalities or as part of an assembly system.

Our results suggest that, although there is no single solution for detecting all stages of age-related macular degeneration, careful selection of Deep Learning architecture can significantly improve diagnostic results. Future research should explore the integration of these architectures with other data modalities and learning techniques to develop more accurate and reliable diagnostic systems.

Acknowledgments

All of the researchers thank Mexico’s Secretariat of Public Education (SEP) through the PRODEP program, Tecnológico Nacional de México (TecNM), Instituto de oftalmología Fundación Conde de Valenciana, and Mexico’s NationalCouncil of Humanities, Sciences and Technologies (CONAHCYT) for supporting this work.

References

Abd El-Khalek, A. A., Balaha, H. M., Alghamdi, N. S., Ghazal, M., Khalil, A. T., Abo-Elsoud, M. E. A., & El-Baz, A. (2024). A concentrated machine learning-based classification system for age-related macular degeneration (AMD) diagnosis using fundus images. *Scientific Reports 2024 14:1*, *14*(1), 1–20. https://doi.org/10.1038/s41598-024-52131-2

Alayón, S., Hernández, J., Fumero, F. J., Sigut, J. F., & Díaz-Alemán, T. (2023). Comparison of the Performance of Convolutional Neural Networks and Vision Transformer-Based Systems for Automated Glaucoma Detection with Eye Fundus Images. *Applied Sciences*, *13*(23), 12722. https://doi.org/10.3390/app132312722

Al-Zamil, W., & Yassin, S. (2017). Recent developments in age-related macular degeneration: a review. *Clinical Interventions in Aging*, *Volume 12*, 1313–1330. https://doi.org/10.2147/CIA.S143508

Bao, H., Dong, L., Piao, S., & Wei, F. (2021). BEiT: BERT Pre-Training of Image Transformers. *ICLR 2022 - 10th International Conference on Learning Representations*. https://arxiv.org/abs/2106.08254v2

Broad (Baidu Research Open-Access Dataset). (2020). *Refuge - Grand Challenge*. Grand Challenge. https://refuge.grand-challenge.org/iChallenge-AMD/

CONDE Investigación – Unidad de Investigación. (n.d.). Retrieved April 1, 2024, from https://www.condeinvestigacion.org/

Deep learning for AMD screening and detection. (n.d.). Retrieved March 16, 2024, from https://www.retina-specialist.com/article/deep-learning-for-amd-screening-and-detection

Día Mundial de la Visión 2020 | Secretaría de Salud | Gobierno | gob.mx. (n.d.). Retrieved March 16, 2024, from https://www.gob.mx/salud/es/articulos/dia-mundial-de-la-vision-2020?idiom=es

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. *ICLR 2021 - 9th International Conference on Learning Representations*. https://arxiv.org/abs/2010.11929v2

He, J., Wang, J., Han, Z., Ma, J., Wang, C., & Qi, M. (2023). An interpretable transformer network for the retinal disease classification using optical coherence tomography. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-30853-z

He, T., Zhou, Q., & Zou, Y. (2022). Automatic Detection of Age-Related Macular Degeneration Based on Deep Learning and Local Outlier Factor Algorithm. *Diagnostics 2022, Vol. 12, Page 532*, *12*(2), 532. https://doi.org/10.3390/DIAGNOSTICS12020532

Leingang, O., Riedl, S., Mai, J., Reiter, G. S., Faustmann, G., Fuchs, P., Scholl, H. P. N., Sivaprasad, S., Rueckert, D., Lotery, A., Schmidt-Erfurth, U., & Bogunović, H. (2023). Automated deep learning-based AMD detection and staging in real-world OCT datasets (PINNACLE study report 5). *Scientific Reports 2023 13:1*, *13*(1), 1–13. https://doi.org/10.1038/s41598-023-46626-7

Li, J., Chen, J., Tang, Y., Wang, C., Landman, B. A., & Zhou, S. K. (2023). Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives. In *Medical Image Analysis* (Vol. 85). https://doi.org/10.1016/j.media.2023.102762

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. http://arxiv.org/abs/2103.14030

Ma, D. A., Hosseinzadeh Taher, M. R., Pang, J., Islam, N. U., Haghighi, F., Gotway, M. B., & Liang, J. (2022). Benchmarking and Boosting Transformers for Medical Image Classification. *Lecture Notes in Computer Science*

(Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), *13542 LNCS*, 12–22. https://doi.org/10.1007/978-3-031-16852-9_2

Mallick, S., Paul, J., Sengupta, N., & Sil, J. (2022). Study of Different Transformer based Networks For Glaucoma Detection. *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, *2022-Novem*. https://doi.org/10.1109/TENCON55691.2022.9977730

Nafisah, S. I., Muhammad, G., Hossain, M. S., & AlQahtani, S. A. (2023). A Comparative Evaluation between Convolutional Neural Networks and Vision Transformers for COVID-19 Detection. *Mathematics*, *11*(6). https://doi.org/10.3390/math11061489

Nurgazin, M., & Tu, N. A. (2023). A Comparative Study of Vision Transformer Encoders and Few-shot Learning for Medical Image Classification. *Proceedings - 2023 IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2023*, 2505–2513. https://doi.org/10.1109/ICCVW60793.2023.00265

Organización Mundial de la Salud. (2020). Informe mundial sobre la visión. In *World health Organisation* (Vol. 214, Issue 14).

Rakhshanda Mujib. (2023, May 11). *ARMD curated dataset 2023*. Kaggle. https://www.kaggle.com/datasets/rakhshandamujib/armd-curated-dataset-2023

U.S. Department of Health and Human Services. (n.d.). *Age-related macular degeneration (AMD)*. National Eye Institute. Retrieved September 26, 2023, from https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/age-related-macular-degeneration

Wassel, M., Hamdi, A. M., Adly, N., & Torki, M. (2022). Vision Transformers Based Classification for Glaucomatous Eye Condition. *Proceedings - International Conference on Pattern Recognition*, *2022-Augus*, 5082–5088. https://doi.org/10.1109/ICPR56361.2022.9956086