ISSN 1870-4069

Detección temprana de degeneración macular asociada con la edad mediante arquitecturas basadas en Transformadores de Visión: Un estudio comparativo

Augusto Javier Reyes-Delgado[1], Jorge Ernesto González-Díaz[1], José Luis Sánchez-Cervantes[1], Yara Anahí Jiménez-Nieto[2], Adolfo Rodríguez-Parada[2], José Luis Rodríguez-Loaiza[3]

[1] Tecnológico Nacional de México, México

[2] Universidad Veracruzana, Facultad de Negocios y Tecnologías, Campus Ixtaczoquitlán, México

[3] Instituto de Oftalmología Conde de Valenciana, Departamento de retina, México

{M17010207,D04010291,jose.sc}@orizaba.tecnm.mx
{yjimenez.adrodriguez}@uv.mx,
jose.rodriguez@institutodeoftalmologia.org

Resumen. La Degeneración Macular Asociada con la Edad (DMAE) es una de las principales causas de pérdida de visión en personas mayores a nivel mundial y catalogada entre las primeras seis afecciones visuales en México. La dificultad de diagnosticar la DMAE en etapas iniciales debido a las sutiles características patológicas en las imágenes retinianas motiva el uso de métodos avanzados de Deep Learning, los cuales ofrecen un potencial significativo para mejorar la precisión del diagnóstico. Recientemente, las arquitecturas basadas en transformadores de visión, como Vision Transformer (ViT), Swin Transformer BERT Pre-training of Image Transformers (BEiT), han emergido, ofreciendo una nueva perspectiva en el análisis de imágenes al aprender relaciones espaciales complejas. Este estudio presenta un análisis comparativo de estas arquitecturas de transformadores de visión aplicadas a la detección de la DMAE, enfocándose en la capacidad de cada modelo para identificar y clasificar las etapas tempranas de la enfermedad. A pesar de los desafíos asociados con el tamaño reducido de los conjuntos de datos de imágenes médicas. Nuestros resultados sugieren que las arquitecturas basadas en ViT y sus derivados logran un rendimiento significativo en la detección de la DMAE, siendo el BEiT particularmente destacado por su consistencia en el desempeño superior. No obstante, es importante resaltar que el ViT mantiene una eficacia notable en la clasificación multiclase de la DMAE, con la ventaja adicional de requerir un menor consumo de recursos computacionales.

Palabras clave: Clasificación multiclase, degeneración macular asociada con la edad (DMAE), detección temprana, transformadores de visión.

Early Detection of Age-related Macular Degeneration Using Vision Transformer-based Architectures: A Comparative Study

Abstract. Age-related macular degeneration (AMD) is one of the leading causes of vision loss in older adults worldwide and is among the top six visual impairments in Mexico. The difficulty in diagnosing AMD in its early stages, due to subtle pathological features in retinal images, motivates the use of advanced deep learning methods that offer significant potential to improve diagnostic accuracy. Recently, vision transformer architectures such as Vision Transformer (ViT) and Swin Transformer BERT Pre-training of Image Transformers (BEiT) have emerged, providing a novel perspective in image analysis by learning complex spatial relationships. This study presents a comparative analysis of these Vision Transformer architectures applied to the detection of AMD, focusing on each model's capability to identify and classify the early stages of the disease. Despite the challenges associated with the small size of medical image datasets, our results suggest that ViT-based architectures and their derivatives achieve significant performance in AMD detection, with BEiT in particular notable for its consistently superior performance. Nevertheless, it is important to highlight that ViT retains remarkable effectiveness in multi-class classification of AMD, with the added advantage of requiring fewer computational resources.

Keywords: Multiclass classification, age-related macular degeneration (AMD), early detection, vision transformers.

1. Introducción

Según la Organización Mundial de la Salud [1], la DMAE afecta significativamente la calidad de vida de las personas mayores, impactando su independencia y capacidad para realizar actividades diarias. En México la DMAE ha sido catalogada por la Secretaría de Salud como el tercero de los seis principales problemas oculares que afectan a la población [2]. La detección temprana de la DMAE es fundamental para prevenir la progresión de la enfermedad y preservar la visión. La literatura sugiere que la integración de métodos de Deep Learning, especialmente las redes neuronales convolucionales (CNN), ha mejorado el rendimiento en la detección y clasificación de imágenes médicas, superando en algunos casos a las evaluaciones hechas por especialistas [3].

El avance en el procesamiento de imágenes y la incorporación de técnicas de Deep Learning han ofrecido nuevas perspectivas para su uso en ciencias médicas. El procesamiento de imágenes utilizando CNN y sus métodos automáticos de análisis han demostrado ser herramientas de alta eficiencia, proporcionando sistemas inteligentes y amigables para el escaneo y diagnóstico de enfermedades, incluida la DMAE, fuera de un entorno clínico [4]. Además, se han propuesto diferentes enfoques para detectar las características patológicas de la DMAE utilizando imágenes de alta resolución, analizando patrones de color y textura [5]. La clasificación automática del nivel de

progresión de la enfermedad enfrenta desafíos, especialmente con características sutiles o similares a condiciones no patológicas.

Esto se complica por las altas resoluciones de imagen y la intensidad de recursos necesarios, lo que da la posibilidad de afectar la precisión del diagnóstico y ralentizar el proceso [6]. En este contexto, con base en el análisis comparativo que abordamos en este articulo consideramos que las arquitecturas de transformadores de visión como ViT [7], Swin Transformer [8] y BEiT [9] sugieren una evolución prometedora en la detección de la DMAE, ofreciendo la capacidad de entender las complejas relaciones espaciales en las imágenes retinianas. Este estudio presenta un análisis comparativo de estas arquitecturas de transformadores de visión aplicadas a la detección de la DMAE, enfocándose en la capacidad de cada modelo para identificar y clasificar las etapas de la enfermedad en No DMAE, Leve, Moderada y Avanzada. Estos modelos pueden ser clave para mejorar la detección temprana y la precisión diagnóstica de la DMAE, lo que es crucial para el tratamiento efectivo y la preservación de la visión en las poblaciones envejecidas incluyendo la de nuestro país.

El trabajo se organiza con base en las siguientes secciones: La Sección 2 muestra los trabajos relacionados. En la Sección 3 se describen los detalles sobre las arquitecturas comparadas, ViT, Swin Transformer y BEiT. En la Sección 4 se detallan los experimentos realizados y la evaluación comparativa del rendimiento de las arquitecturas incluidas en el estudio. Finalmente, la Sección 5 presenta conclusiones y trabajo futuro.

2. Trabajos relacionados

Se han identificado en la literatura algunos trabajos relacionados, que, si bien no realizan una comparativa entre diferentes arquitecturas de transformadores de visión para DMAE, destacan la eficiencia de estos sobre las CNN. Entre los de mayor relevancia se encuentra el presentado en [10], que abordó la aplicación de ViT para detectar glaucoma mediante el uso de imágenes de fondo del ojo. Se evaluaron varios modelos: ViT, Swin Transformer, Twins-PCPVT y Atención de clase en transformadores de imágenes (CaiT) con algoritmos de aprendizaje con pocas imágenes y se analizó el impacto de las técnicas de aumento de datos.

Los resultados del estudio mostraron que ViT, combinado con ProtoNets, superó a las contrapartes basadas en CNN y logró un rendimiento competitivo en conjuntos de datos de referencia. Por otro lado, en [11] se evaluó la eficacia de las CNN y Sistemas Basados en ViT para detectar glaucoma en imágenes de fondo de ojo. Los autores probaron diversas arquitecturas de CNN como VGG19, ResNet50, InceptionV3 y Xception, junto con variantes de ViT como Swin Transformer y Twins-PCPVT, así como sistemas híbridos como CaiT, Transformadores de imágenes con eficiencia de datos (DeiT), Transformador de imagen mejorada por convolución (CeiT) y Transformador de visión convolucional (ConViT), y la arquitectura ResMLP. Los resultados mostraron un rendimiento similar entre CNN y ViT en el conjunto de pruebas, pero las CNN demostraron mejor generalización en conjuntos externos. Asimismo, en [12] se exploró el potencial de las arquitecturas ViT en aplicaciones de imagen médica. Se compararon las capacidades de ViT con las de las CNN en tareas como segmentación, reconocimiento y clasificación de imágenes médicas. Se

destacaron arquitecturas como Conformer, U-Net Transformer y Multi-transSP, mostrando su eficacia en la mejora de la precisión y la eficiencia en diversas aplicaciones médicas.

Los resultados mostraron que ViT superó a las CNN en la segmentación de imágenes médicas, gracias a su capacidad para modelar dependencias a largo plazo y su escalabilidad.

Además, un método para clasificar enfermedades retinianas mediante imágenes de tomografía de coherencia óptica (OCT) fue introducido en [13], utilizando una red Swin-Poly Transformer. Los hallazgos indicaron que el método propuesto facilitó una clasificación retiniana precisa y eficiente, subrayando el valor de la inteligencia artificial en diagnósticos oftalmológicos y el potencial de las redes ViT en este ámbito.

Del mismo modo, en [14] los autores se centraron en la comparación de CNNs y ViTs para la clasificación de radiografías de tórax (CXR) en casos de COVID-19, neumonía viral y casos sanos. Los autores utilizaron el conjunto de datos COVID-QU-Ex, dividiendo aleatoriamente el 80% para entrenamiento y el 20% para pruebas. Evaluaron la efectividad en casos balanceados y desbalanceados, implementaron modelos ViT como Twins, Swin y Segformer.

Los resultados mostraron que los modelos basados en CNN y ViT tenían un rendimiento comparable, con una precisión máxima del 99.82% para EfficientNetB7 (CNN) y un rendimiento destacado para SegFormer (ViT).

De manera similar en [15] , se evaluó el rendimiento de las arquitecturas de ViT, específicamente ViT-B y Swin-B, en clasificación de imágenes médicas, contrastando su efectividad con los modelos basados en CNNs para diagnosticar enfermedades como enfermedades torácicas, embolia pulmonar y tuberculosis usando radiografías y tomografías computarizadas. Aquí plantearon que la inicialización adecuada es esencial para los Transformadores de Visión en el ámbito médico, y que los enfoques de autoaprendizaje que utilizan información mutua generan representaciones más precisas para la clasificación médica.

En el mismo sentido, los autores de [16] exploraron el uso de ViT, Swin Transformer y ConvNext, aplicando técnicas de transfer learning para la detección de Glaucoma a partir de imágenes del fondo de ojo. Este esfuerzo buscó crear un método automatizado que permitió identificar el Glaucoma en fases tempranas, con el fin de prevenir la ceguera.

Finalmente, en [17] Wassel et al., reportaron un estudio centrado en la clasificación de condiciones oculares glaucomatosas utilizando modelos de ViT en imágenes de fondo de ojo completas y recortadas en el disco óptico. Evaluaron las arquitecturas ViT, Swin, CaiT, crossViT, XciT, ResMlp y DeiT, tanto de forma individual como en ensambles. Además del glaucoma, abordaron otras enfermedades oftalmológicas como diabetes, cataratas, hipertensión, miopía patológica y otras anomalías.

Los resultados mostraron que Swin y CaiT obtuvieron altos niveles de precisión, sensibilidad y especificidad en la validación y prueba de los conjuntos de datos combinados, destacando su eficacia en la detección de glaucoma en imágenes oftalmológicas, lo que sugiere su potencial utilidad en la práctica clínica.

La tabla 1 muestra los datos resumidos de los trabajos relacionados identificados en la literatura para este trabajo de investigación.

Tabla 1. Trabajos relacionados de comparativas de arquitecturas ViT en medicina.

Año	Autor	Arquitecturas ViT	Enfermedades	Tipo Imágenes
2023	Nurgazin M. et al [10]	Variantes del ViT clásico: ViT_tiny ViT_small ViT_base	Melanoma, Carcinoma basocelular, Carcinoma espinocelular, Nevus, Queratosis actínica, Dermatofibroma, Quiste epidermoide, Psoriasis, Dermatitis atópica, Rosácea, Cáncer de mama	Lesiones cutáneas. Biopsias de tejido mamario. Citologías cervicales
2023	Alayon S. et. al [11]	ViT, Swin Transformer, Twins-PCPVT, CaiT	Glaucoma	Fondo de Ojo
2023	Li J. et al [12]	Conformer, U-Net Transformer, Módulo Residual Transformer Multi-transSP, TransPath, i-ViT BabyNet	Predicción de peso fetal. Detección de retinopatía diabética. Segmentación de cartílago de rodilla	Ultrasonido. Resonancia magnética, Tomografía computarizada, Rayos X, Histopatología
2023	He J. et al [13]	ViT Swin Transformer	Retinopatía diabética, Edema macular diabético, Glaucoma, Anormalidades oculares	Tomografía de coherencia óptica (OCT)
2023	Nafisah S. et al [14]	Twins, Swin, Segformer	COVID-19 Neumonía	Radiografías de tórax (CXR)
2022	Ma D. et al [15]	ViT-B, Swin-B.	Enfermedades torácicas, Embolia pulmonar, Tuberculosis.	Radiografías de tórax, Tomografías computarizadas.
022	Mallick S. et al [16]	ViT, Swin Transformer, ConvNext	Glaucoma	Fondo de Ojo
2022	Wassel M. et al [17]	Cait, crossViT, XciT, ResMlp, DeiT, ViT	Glaucoma, Diabetes Cataratas, Hipertensión, Miopía Patológica	Fondo de Ojo

La Tabla 1 muestra investigaciones recientes centradas en diversas aplicaciones de arquitecturas de ViT en el campo médico, abarcando desde enfermedades de la piel hasta condiciones oculares como glaucoma y retinopatía diabética.

Sin embargo, no se observan trabajos específicos que exploren el rendimiento de los modelos ViT, Swin Transformer y BEiT para la detección o clasificación de la DMAE. Esto permite inferir que la aplicación de BEiT en el diagnóstico de DMAE es un área poco explorada, lo cual destaca la oportunidad y necesidad de investigar esta dirección, dada la importancia de la DMAE como causa significativa de pérdida de visión en la población mayor.

Con base en lo anterior, realizar una comparativa entre ViT, Swin Transformer y BEiT ofrece una perspectiva poco abordada en la literatura para el análisis de imágenes,

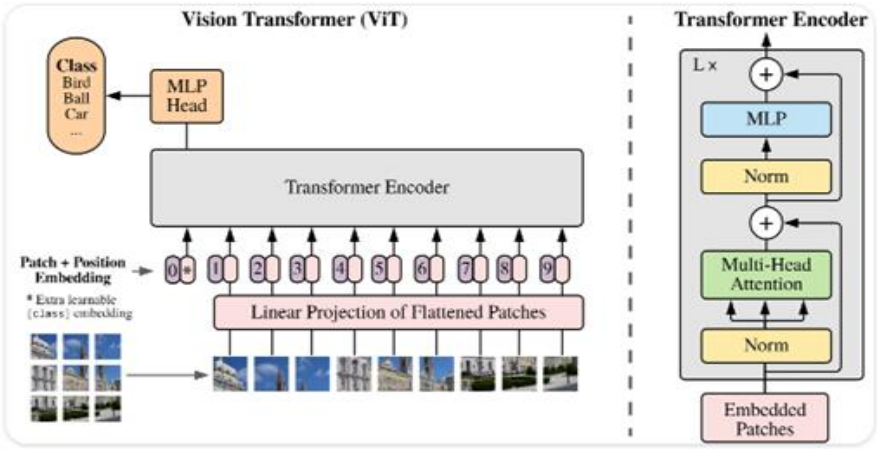

Fig. 1. Arquitectura ViT original (tomada de [7]).

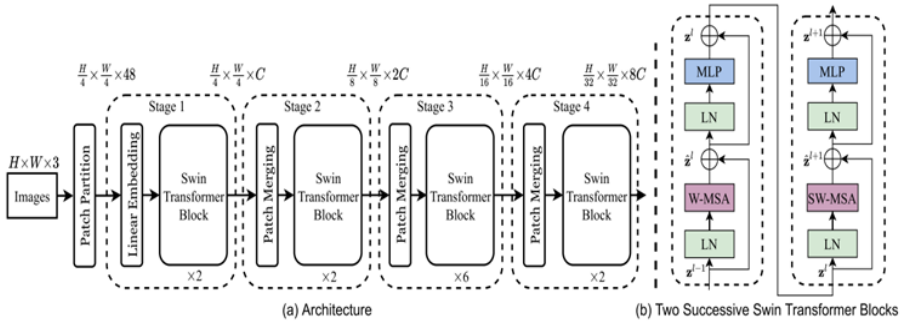

Fig. 2. Arquitectura de Swin Transformer (tomada de [8]).

aprovechando la atención global y las características jerárquicas, lo que podría mejorar significativamente la detección y clasificación de la DMAE.

3. Métodos

En este estudio se realiza una comparativa detallada entre las arquitecturas de ViT, Swin Transformer y BEiT para la detección y clasificación multiclase de la DMAE en imágenes de fondo del ojo. Se manejan cuatro categorías de clasificación: No DMAE, DMAE moderada, intermedia y avanzada.

Estas tecnologías de Deep Learning se seleccionaron por su potencial para procesar de manera eficaz las características visuales intrincadas, fundamentales para discernir los distintos estadios de la DMAE, lo que resulta clave para lograr un diagnóstico temprano y preciso.

La elección de estas arquitecturas se justifica por su avanzada capacidad para capturar patrones globales y locales en las imágenes, lo cual es esencial para una detección fiable y una clasificación precisa de la progresión de la DMAE.

3.1. Vision Transformer (ViT)

ViT [7] es una innovadora arquitectura que aplica el mecanismo de transformadores, usual en el procesamiento del lenguaje, al campo de la visión por computadora. ViT parte las imágenes en parches y los procesa como si fueran tokens en una secuencia.

Utiliza la atención para ponderar la importancia de diferentes partes de la imagen, permitiendo al modelo captar patrones complejos y relaciones a larga distancia (Fig. 1). Su enfoque en las relaciones globales lo hace especialmente adecuado para identificar patrones en imágenes médicas, como las relacionadas con la DMAE, donde las manifestaciones de la enfermedad pueden estar sutiles y distribuidas por toda la imagen.

3.2. Swin Transformer

Swin Transformer fue introducido por Ze Liu et al. en su trabajo "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows" [8] en 2021. El Swin Transformer surge como una respuesta a algunas limitaciones de los modelos de transformadores puros como ViT, especialmente en términos de eficiencia computacional y la capacidad para manejar tamaños de imagen variables. Aunque ViT demostró que los transformadores podían ser poderosos para tareas de visión, su enfoque de tratar la imagen como una secuencia de parches fijos planteaba desafíos en términos de escalabilidad y adaptabilidad a diferentes resoluciones y tamaños de imagen.

El Swin Transformer introduce varios conceptos innovadores para abordar estas limitaciones:

- **Ventanas Desplazadas (Shifted Windows):** Una de las innovaciones clave del Swin Transformer es su uso de ventanas desplazadas. Divide la imagen en ventanas no superpuestas para la atención local, lo cual reduce la complejidad computacional.

- **Jerarquía**: Al igual que en las CNN, el Swin Transformer procesa las imágenes en varias resoluciones. Comienza con una alta resolución y va reduciéndola progresivamente, permitiendo al modelo capturar características a diferentes escalas y mejorar la eficiencia al reducir la resolución en las capas más profundas.

- **Flexibilidad y Generalidad**: A diferencia de ViT, que utiliza parches de tamaño fijo, el Swin Transformer puede manejar de manera más efectiva diferentes tamaños de imagen y resoluciones, lo que lo hace más flexible y adaptable para diversas aplicaciones en visión por computadora.

La selección de Swin Transformer se justifica por su diseño que aborda eficientemente la jerarquía y la localidad en imágenes. A diferencia del ViT, que considera toda la imagen de manera global, Swin Transformer procesa las imágenes en ventanas locales, lo que permite una representación más detallada de las características locales (Fig.2).

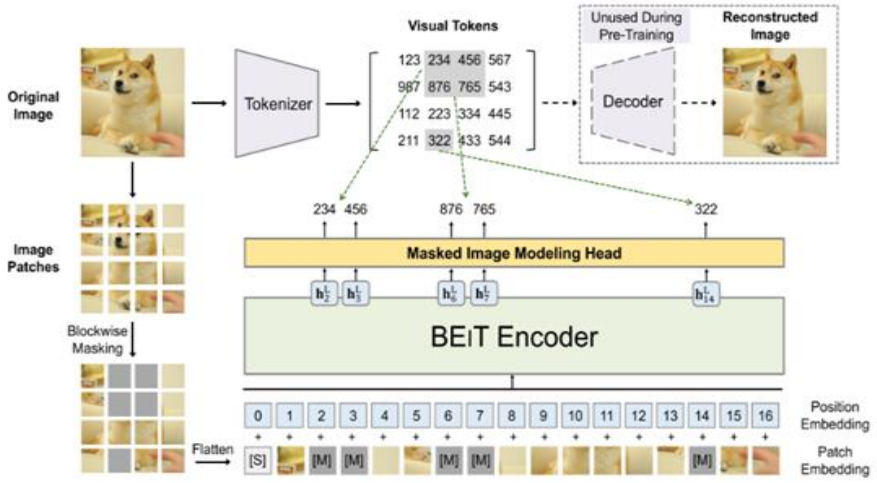

Fig. 3. Arquitectura de BEiT (tomada de [9]).

3.3. BEiT (BERT Pre-training of Image Transformers)

Fue presentado en un trabajo por Bao et al. [9] en 2021, titulado "BEiT: BERT Pre-training of Image Transformers".

BEiT se inspira en el éxito de BERT (Bidirectional Encoder Representations from Transformers) en el campo de Procesamiento de lenguaje natural (NLP). BERT revolucionó NLP mediante el preentrenamiento de transformadores en grandes corpus de texto usando tareas de predicción de palabras ocultas, donde el modelo aprende a predecir partes del texto que han sido intencionalmente ocultadas. BEiT traslada este enfoque de preentrenamiento al dominio de las imágenes (Fig.3). En lugar de predecir palabras ocultas, BEiT se entrena para predecir partes ocultas de una imagen. Este proceso implica dos etapas principales:

– **Tokenización de Imágenes**: BEiT convierte una imagen en un conjunto de tokens visuales utilizando un modelo de tokenización de imágenes (como un VQ-VAE, un autoencoder variacional cuantizado). Esto resulta en una representación de la imagen en forma de tokens, similar a cómo se tokeniza un texto en NLP.

– **Preentrenamiento de Modelo**: El modelo se preentrena con la tarea de predecir los tokens visuales de partes de la imagen que han sido ocultadas, similar a la predicción de palabras faltantes en BERT. Esto enseña al modelo a entender y predecir la estructura y el contenido visual basándose en el contexto proporcionado por las partes visibles de la imagen.

El uso de BEiT en este estudio está justificado por su enfoque en el aprendizaje de representaciones visuales mediante la predicción de píxeles ocultos, lo cual representa una innovación para el análisis de imágenes de fondo de ojo en DMAE. BEiT es capaz de captar sutilezas en las texturas y patrones de las imágenes, aspectos cruciales para identificar las etapas de la DMAE.

4. Resultados experimentales

Para la comparativa, se entrenaron tres modelos seleccionados, mediante un proceso de transferencia de entrenamiento (fine tuning), utilizando el mismo conjunto de datos. Se llevaron a cabo múltiples iteraciones para ajustar los hiperparámetros de cada modelo. A continuación, se detallan los aspectos más específicos de la implementación.

4.1. Detalles de implementación

La implementación de los modelos ViT y Swin Transformer se realizó en Google Colab®, utilizando aceleración por GPU. En contraste, el modelo BEiT se entrenó en una computadora de escritorio con un GPU Nvidia RTX 3070 de 8GB, debido a limitaciones de recursos en Colab®. Para el entrenamiento, se empleó la biblioteca PyTorch Transformers en todos los modelos.

Para determinar los hiperparámetros óptimos, se realizaron múltiples experimentos, concluyendo que 42 épocas y lotes de 32 imágenes resultan ser los más adecuados. Se observó que incrementar el número de épocas más allá de 42 no generaba mejoras significativas en el rendimiento, identificándose una meseta en el desempeño cerca de las 20 épocas. Adicionalmente, la tasa de aprendizaje se estableció en $5e^{-05}$ después de exhaustivas evaluaciones.

4.2. Conjunto de datos

Para construir el conjunto de datos, se utilizaron inicialmente 305 imágenes, repartidas en 185 para entrenamiento y 60 para validación y pruebas, siguiendo una distribución de 60%/20%/20%. Para potenciar la generalización del modelo y reducir el sobreajuste, se aplicaron técnicas de aumento de datos al lote de entrenamiento, incluyendo cambios en tamaño, rotaciones y ajustes de brillo y contraste. Esto incrementó el conjunto a 1,094 imágenes, con 974 dedicadas al entrenamiento. Las imágenes se obtuvieron del conjunto de datos iChallenge-AMD [18] y de un conjunto en Kaggle publicado por Mujib [19], que incluye imágenes de DMAE extraídas de varios conjuntos de imágenes de fondo de ojo con patologías retinianas.

La clasificación de las imágenes se basó en la literatura existente [20,21], y posteriormente fue validada por expertos en el area medica [22].

4.3. Resultados

Al concluir el entrenamiento de cada modelo, se registró la exactitud, observando un incremento hasta la 40ª época. Este desempeño fue monitoreado y documentado utilizando la plataforma Weights & Biases® (W&B), como se ilustra en la Fig.4.

Todos los modelos alcanzaron una exactitud superior al 0.7500. Para fortalecer la evaluación de los resultados, se calcularon métricas adicionales de rendimiento, incluyendo precisión, sensibilidad y F1 Score, permitiendo un análisis más profundo de las capacidades de cada modelo. Estas métricas se obtuvieron utilizando el conjunto de validación y el conjunto de pruebas previamente separado del conjunto de datos inicial. Los resultados de estas métricas se presentan en la tabla 2.

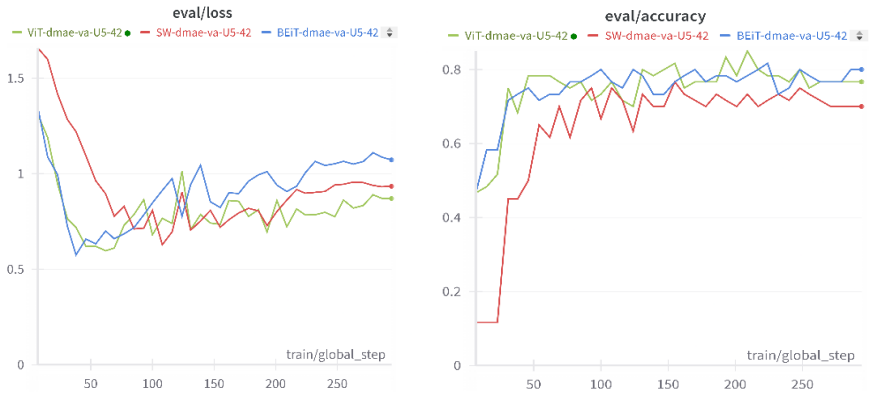

Fig. 4. Resultados obtenidos por los modelos durante el proceso de entrenamiento.

Los resultados presentados en la tabla 2 muestran diferencias notables en el rendimiento de los modelos. Al promediar los resultados de precisión de los modelos para los conjuntos de validación y prueba sobre los que fueron evaluados, se obtiene que ViT alcanza una exactitud del 0.7833, Swin Transformer obtiene el 0.7583, y BEiT llega hasta el 0.8166.

Es importante destacar que, para todos los modelos, el rendimiento en la clasificación de imágenes de fondo de ojo de las clases "leve" y "moderada" es consistente y muestra mejores resultados en comparación con las clases "No dmae" y "Avanzada", las cuales presentan una mayor varianza. Las Figs. 5 y 6 muestran las gráficas comparativas de las principales métricas aplicadas a los modelos.

4.4. Discusión

En nuestro estudio comparativo, se evaluaron tres modelos avanzados: ViT, Swin Transformer y BEiT. Los resultados indican variaciones significativas en el rendimiento de cada modelo, subrayando la importancia de la selección de la arquitectura en aplicaciones clínicas.

El modelo ViT exhibió una elevada precisión en el conjunto de validación, aunque se observó una disminución en su rendimiento al evaluarlo en el conjunto de pruebas. Destacó particularmente en la clasificación de casos avanzados de degeneración macular dentro del conjunto de validación, lo cual sugiere que esta arquitectura posee una aptitud específica para la identificación de manifestaciones severas de la enfermedad. No obstante, se registró una sensibilidad reducida en la detección de casos en etapas tempranas, lo que podría reflejar una predisposición hacia la sobre clasificación en las fases más graves.

En contraste, el Swin Transformer presentó una precisión general ligeramente inferior en comparación con el ViT, especialmente notable en el conjunto de validación. En el conjunto de pruebas, esta arquitectura enfrentó retos al clasificar de manera acertada los casos avanzados, evidenciado por su baja sensibilidad y puntuación F1 en dicha categoría. Sin embargo, mostró un desempeño competitivo en la identificación de etapas iniciales de la enfermedad, lo que indica su potencial utilidad en la detección precoz de la misma.

Tabla 2. Resultados en las métricas de evaluación de los diferentes modelos.

Modelo	Exactitud	Conjunto	Clase	Precisión	Sensibilidad	F1-Score
ViT	0.8500	Validación	No dmae	0.8000	0.6666	0.7272
			Leve	0.8333	0.9259	0.8771
			Moderada	0.8500	0.8500	0.8500
			Avanzada	1.0000	0.7142	0.8333
	0.7166	Pruebas	No dmae	1.0000	0.5000	0.6666
			Leve	0.7187	0.8518	0.7796
			Moderada	0.6666	0.7000	0.6829
			Avanzada	0.7500	0.4285	0.5454
Swin Transformer	0.7500	Validación	No dmae	0.5555	0.8333	0.6666
			Leve	0.7777	0.7777	0.7777
			Moderada	0.8235	0.7000	0.7567
			Avanzada	0.7142	0.7142	0.7142
	0.7666	Pruebas	No dmae	0.8333	0.8333	0.8333
			Leve	0.9583	0.8518	0.9019
			Moderada	0.6666	0.7000	0.6829
			Avanzada	0.4444	0.5714	0.5000
BEiT	0.8166	Validación	No dmae	0.8000	0.6666	0.7272
			Leve	0.7931	0.8518	0.8214
			Moderada	0.8333	0.7500	0.7894
			Avanzada	0.8750	1.0000	0.9333
	0.8166	Pruebas	No dmae	1.000	0.6666	0.8000
			Leve	0.8518	0.8518	0.8518
			Moderada	0.7391	0.8500	0.7906
			Avanzada	0.8333	0.7142	0.7692

Por su parte, el BEiT demostró ser el modelo más eficaz en el conjunto de pruebas, superando a los modelos ViT y Swin Transformer en términos de precisión general. A pesar de un rendimiento inicialmente inferior al ViT en el conjunto de validación, el BEiT evidenció una consistencia notable entre ambos conjuntos y una mejora significativa en la detección de todas las etapas de la enfermedad en comparación con el ViT durante las pruebas.

Esto revela una capacidad de generalización y robustez superior, posicionando al BEiT como una alternativa prometedora para la detección práctica de la degeneración macular en sus diversas etapas.

La variabilidad en el rendimiento entre estos modelos destaca la complejidad de aplicar Deep Learning para diagnósticos médicos. Mientras que ViT y Swin Transformer ofrecen ventajas en la detección de etapas específicas de la enfermedad, BEiT muestra un balance entre sensibilidad y precisión en un rango más amplio de condiciones.

Lo anterior subraya la necesidad de considerar múltiples factores, como la exactitud, sensibilidad y especificidad, al seleccionar un modelo de Deep Learning para la detección de enfermedades oftalmológicas.

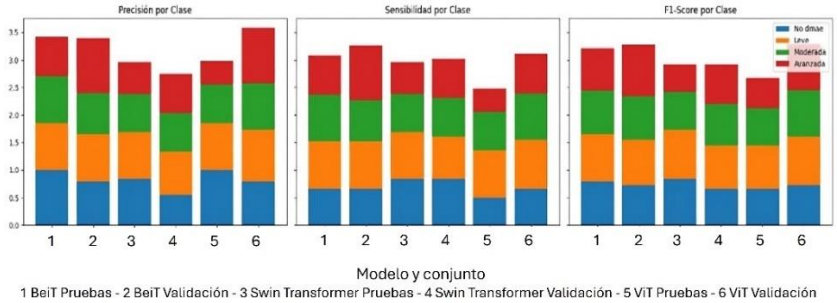

Fig. 5. Comparativa de precisión, sensibilidad y F1-score por clase de diferentes modelos.

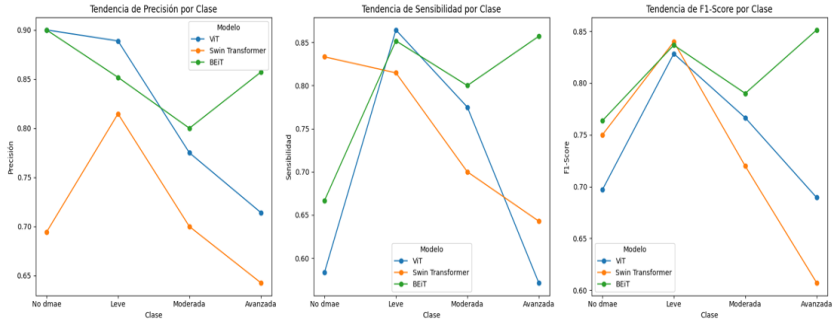

Fig. 6. Comparativa de tendencias por clase y modelo de las métricas de evaluación de los modelos.

5. Conclusiones

Aunque ViT sigue presentando un alto desempeño, BEiT se presenta como una alternativa más consistente en el presente caso de estudio.

Los resultados promedio de precisión para los conjuntos de validación y prueba muestran que el BEiT lidera con una exactitud del 81.66%, seguido por ViT con el 78.33%, y finalmente Swin Transformer con el 75.48%. Estas cifras reflejan no solo la capacidad de generalización de cada modelo sino también su fiabilidad en el reconocimiento de patrones asociados con condiciones oculares específicas.

Resulta particularmente interesante que los modelos muestren una consistencia en el rendimiento en la clasificación de condiciones clasificadas como “leve” y “moderada”. Este fenómeno indica que las características visuales presentes en estas etapas de la enfermedad son más distintivas y, por lo tanto, más fácilmente reconocibles por los modelos de aprendizaje profundo. Por otro lado, las categorías “No dmae” y “Avanzada” exhiben una mayor variabilidad en los resultados, lo que sugiere que las manifestaciones visuales de estas etapas pueden ser más sutiles o menos diferenciadas, dificultando así la clasificación precisa.

Los hallazgos subrayan la importancia de la selección de modelo en aplicaciones de diagnóstico médico basadas en inteligencia artificial. Aunque el BEiT supera en rendimiento general, por su balance entre precisión y capacidad de generalización, ViT

continúa mostrando un desempeño superior en determinados casos. El Swin Transformer, a pesar de tener un rendimiento ligeramente inferior, aún podría ser valioso en un contexto clínico cuando se combina con otras modalidades o como parte de un sistema de ensamble.

Nuestros resultados sugieren que, aunque no existe una solución única para la detección de todas las etapas de la degeneración macular asociada con la edad, la selección cuidadosa de la arquitectura de Deep Learning puede mejorar significativamente los resultados de diagnóstico. Futuras investigaciones deberán explorar la integración de estas arquitecturas con otras modalidades de datos y técnicas de aprendizaje, para desarrollar sistemas de diagnóstico más precisos y confiables.

Referencias

1. Organización Mundial de la Salud: Informe mundial sobre la visión, vol. 214, no. 14 (2020)
2. Día Mundial de la Visión 2020: Secretaría de Salud. Available: https://www.gob.mx/salud/es/articulos/dia-mundial-de-la-vision-2020?idiom=es (2024)
3. He, T., Zhou, Q., Zou, Y.: Automatic Detection of Age-related Macular Degeneration based on deep Learning and Local outlier Factor Algorithm. Diagnostics, vol. 12, no. 2, pp. 532 (2022). DOI: 10.3390/DIAGNOSTICS12020532.
4. Abd-El-Khalek, A.A., Balaha, H.M., Alghamdi, N.S., Ghazal, M., Khalil, A.T., Abo-Elsoud, M.E.A., El-Baz, A.: A Concentrated Machine Learning-based Classification System for Age-related Macular Degeneration (AMD) Diagnosis using Fundus Images. Scientific Reports, vol. 14, no. 1, pp. 2434 (2024). DOI: 10.1038/s41598-024-52131-2.
5. Leingang, O., Riedl, S., Mai, J., Reiter, G.S., Faustmann, G., Fuchs, P., Bogunović, H.: Automated deep Learning-based AMD Detection and Staging in Real-World OCT Datasets (PINNACLE study report 5). Scientific Reports, vol. 13, no. 1, pp. 1–13 (2023). DOI: 10.1038/s41598-023-46626-7.
6. Retina-Specialist: Deep learning for AMD screening and detection. Available: https://www.retina-specialist.com/article/deep-learning-for-amd-screening-and-detection (2024)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929 (2020)
8. Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR´21, 9th International Conference on Learning Representations (2020). Available: https://arxiv.org/abs/2010.11929v2.
9. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows (2020). DOI: 10.48550/arXiv.2103.14030.
10. Bao, H., Dong, L., Piao, S., Wei, F.: BEiT: BERT Pre-Training of Image Transformers. ICLR´22 - 10th International Conference on Learning Representations (2021). https://arxiv.org/abs/2106.08254v2.
11. Nurgazin, M., Tu, N.A.: A Comparative Study of Vision Transformer Encoders and Few-shot Learning for Medical Image Classification. In Proceedings IEEE/CVF International Conference on Computer Vision Workshops, ICCVW´23, pp. 2505–2513 (2023). DOI: 10.1109/ICCVW60793.2023.00265.
12. Alayón, S., Hernández, J., Fumero, J.F., Sigut, F.J., Díaz-Alemán, T.: Comparison of the Performance of Convolutional Neural Networks and Vision Transformer-Based Systems for Automated Glaucoma Detection with Eye Fundus Images. Applied Sciences, vol. 13, no. 23, pp. 12722 (2023). DOI: 10.3390/app132312722.

13. Li, J., Chen, J., Tang, Y., Wang, C., Landman, B.A., Zhou, S.K.: Transforming Medical Imaging with Transformers? A Comparative Review of Key Properties, current Progresses, and Future Perspectives. Medical Image Analysis, vol. 85 (2023). DOI: 10.1016/j.media.2023.102762.
14. He, J., Wang, J., Han, Z., Ma, J., Wang, C., Qi, M.: An Interpretable Transformer Network for the Retinal Disease Classification using Optical Coherence Tomography. Sci Rep, vol. 13, no. 1 (2023). DOI: 10.1038/s41598-023-30853-z.
15. Nafisah, S.I., Muhammad, G., Hossain, M.S., AlQahtani, S.A.: A Comparative Evaluation between Convolutional Neural Networks and Vision Transformers for COVID-19 Detection. Mathematics, vol. 11, no. 6 (2023). DOI: 10.3390/math11061489.
16. Ma, D. A., Hosseinzadeh-Taher, M.R., Pang, J., Islam, N.U., Haguigui, F., Gotway, M.B., Liang, J.: Benchmarking and Boosting Transformers for Medical Image Classification, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Science and Business Media Deutschland GmbH, pp. 12–22 (2022). DOI: 10.1007/978-3-031-16852-9_2.
17. Mallick, S., Paul, J., Sengupta, N., Sil, J.: Study of Different Transformer based Networks For Glaucoma Detection. In IEEE Region 10 Annual International Conference, Proceedings/TENCON, Institute of Electrical and Electronics Engineers Inc. (2022). DOI: 10.1109/TENCON55691.2022.9977730.
18. Wassel, M., Hamdi, A.M., Adly, N., Torki, M.: Vision Transformers Based Classification for Glaucomatous Eye Condition. In Proceedings International Conference on Pattern Recognition, Institute of Electrical and Electronics Engineers Inc., pp. 5082–5088 (2022). DOI: 10.1109/ICPR56361.2022.9956086.
19. Baidu: Research Open-Access Dataset - download. Available: https://ai.baidu.com/broad/download (2024)
20. Mujib, R.: ARMD curated dataset 2023, Kaggle. [Online]. Available: https://www.kaggle.com/datasets/rakhshandamujib/armd-curated-dataset-2023 (2023)
21. National Eye Institute: U.S. Department of Health and Human Services, Age-related macular degeneration (AMD). https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/age-related-macular-degeneration (2024)
22. Al-Zamil, W., Yassin, S.: Recent Developments in Age-related Macular Degeneration: A review. Clin Interv Aging, vol. 12, pp. 1313–1330 (2017). DOI: 10.2147/CIA.S143508.
23. CONDE: Investigación – Unidad de Investigación. Available: https://www.condeinvestigacion.org/ (2023)