UNIVERSIDAD NACIONAL DEL ALTIPLANO

FACULTAD DE INGENIERÍA ESTADÍSTICA E INFORMÁTICA

ESCUELA PROFESIONAL DE INGENIERÍA ESTADÍSTICA E INFORMÁTICA

ANÁLISIS COMPARATIVO DE REDES NEURONALES CONVOLUCIONALES Y VISION TRANSFORMERS PARA EL DIAGNÓSTICO AUTOMATIZADO EN IMÁGENES RADIOGRÁFICAS

TESIS

PRESENTADA POR:

YEFER ANDERSSON MAMANI CHAMBI

PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO ESTADÍSTICO E INFORMÁTICO

PUNO - PERU

2024

Yefer Andersson Mamani Chambi

ANÁLISIS COMPARATIVO DE REDES NEURONALES CONVOLUCIONALES Y VISION TRANSFORMERS PARA EL DIA...

Universidad Nacional del Altiplano

Detalles del documento

Identificador de la entrega
trn:oid:::8254:416611941

Fecha de entrega
16 dic 2024, 12:46 p.m. GMT-5

Fecha de descarga
16 dic 2024, 12:51 p.m. GMT-5

Nombre de archivo
ANÁLISIS COMPARATIVO DE REDES NEURONALES CONVOLUCIONALES Y VISION TRANSFORMER....docx

Tamaño de archivo
6.5 MB

199 Páginas

34,008 Palabras

202,434 Caracteres

turnitin Página 1 of 211 - Portada Identificador de la entrega trn:oid:::8254:416611941

11% Similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para ca...

Filtrado desde el informe

- Bibliografía
- Coincidencias menores (menos de 12 palabras)

Fuentes principales

6% Fuentes de Internet
2% Publicaciones
10% Trabajos entregados (trabajos del estudiante)

Marcas de integridad

N.° de alerta de integridad para revisión

Caracteres reemplazados
149 caracteres sospechosos en N.° de páginas
Las letras son intercambiadas por caracteres similares de otro alfabeto.

Los algoritmos de nuestro sistema analizan un documento en profundidad para buscar inconsistencias que permitirían distinguirlo de una entrega normal. Si advertimos algo extraño, lo marcamos como una alerta para que pueda revisarlo.

Una marca de alerta no es necesariamente un indicador de problemas. Sin embargo, recomendamos que preste atención y la revise.

Universidad Nacional del Altiplano
Firmado digitalmente por PARI CONDORI Elqui Yeye FAU 20145496170 soft
Motivo: Soy el autor del documento
Fecha: 16.12.2024 13:41:13 -05:00

Universidad Nacional del Altiplano de Puno
Firmado digitalmente por JUAREZ VARGAS Juan Carlos FAU 20145496170 soft
Motivo: Soy el autor del documento
Fecha: 16.12.2024 14:26:08 -05:00

turnitin Página 2 of 211 - Descripción general de integridad Identificador de la entrega trn:oid:::8254:416611941

DEDICATORIA

Esta tesis está dedicada con todo mi amor y gratitud a mi familia, quienes han sido el pilar fundamental de mi vida.

A mis padres, Edgar M. C. y Olga C. M., cuya dedicación, esfuerzo y amor incondicional me han brindado la oportunidad de convertirme en quien soy hoy. Su sacrificio constante para darme una educación y un futuro mejor no pasa desapercibido.

A mi hermana, Sherly M. C., quien con su apoyo y cariño, ha sido una fuente constante de motivación en mi vida. Gracias por estar siempre a mi lado y ser parte fundamental de este camino.

Yefer Andersson Mamani Chambi

AGRADECIMIENTOS

En primer lugar, agradezco a Dios por darme la fuerza y la sabiduría para superar cada obstáculo en este camino.

A mis padres, Edgar M. C. y Olga C. M., gracias por creer en mí incluso en los momentos en los que yo dudé de mis propias capacidades. Su fe en mi potencial fue el motor que me impulsó a seguir adelante, aún cuando el camino se tornó complicado. Su apoyo constante y su confianza en que yo sería capaz de lograrlo me han dado la seguridad necesaria para superar cada obstáculo.

A mi hermana, Sherly M. C., gracias por tu constante aliento y por confiar siempre en que sería capaz de alcanzar esta meta. Tus palabras de ánimo y tu certeza en mis capacidades me llenaron de esperanza en los momentos más desafiantes.

A mis maestros y asesores, quienes compartieron su conocimiento y me guiaron durante este proceso. Sus enseñanzas han sido fundamentales para mi crecimiento académico y personal.

A mis amigos y compañeros, quienes de una forma u otra estuvieron presentes a lo largo de esta travesía, formando un equipo sólido que me acompañó en cada desafío.

Y finalmente a ti, Jimena L. P., gracias por tu apoyo incondicional y por brindarme ánimo en los momentos más difíciles. Por estar siempre presente cuando más te necesitaba y por recordarme que, sin importar los obstáculos que enfrentemos, siempre podemos cumplir nuestra promesa de 'vamos a hacer que funcione'.

A todos ustedes, mi más sincero agradecimiento por formar parte de este capítulo tan importante en mi vida.

Yefer Andersson Mamani Chambi

ÍNDICE GENERAL

REVISIÓN DE LITERATURA

ÁREA: Inteligencia Artificial.

TEMA: Análisis comparativo de arquitecturas de Deep Learning en diagnóstico médico.

FECHA DE SUSTENTACIÓN: 19 de diciembre del 2024

ÍNDICE DE TABLAS

Pág.

ÍNDICE DE FIGURAS

Pág.

ÍNDICE DE ANEXOS

Pág.

ACRÓNIMOS

AUC:	Area Under the Curve (Área Bajo la Curva)
CNN:	Convolutional Neural Networks (Redes Neuronales Convolucionales)
DL:	Deep Learning (Aprendizaje Profundo)
IA:	Inteligencia Artificial
ML:	Machine Learning (Aprendizaje Automático)
ResNet:	Residual Neural Network (Red Neural Residual)
ROC:	Receiver Operating Characteristic (Característica Operativa del Receptor)
TB:	Tuberculosis
UNA:	Universidad Nacional del Altiplano
ViT:	Vision Transformer (Transformador de Visión)
XAI:	eXplainable Artificial Intelligence (Inteligencia Artificial Explicable)

RESUMEN

El diagnóstico médico automatizado mediante técnicas de aprendizaje profundo representa un campo en constante evolución, donde la selección de arquitecturas óptimas es importante para garantizar diagnósticos precisos y confiables. Esta investigación de tipo comparativo evaluó el desempeño predictivo de las Redes Neuronales Convolucionales (CNN) y los Vision Transformers (ViT) en el diagnóstico automatizado de imágenes radiográficas. Se analizaron 15,834 imágenes distribuidas entre casos de artrosis de rodilla, neumonía y tuberculosis, implementando cuatro variantes de cada arquitectura mediante validación cruzada de 5 folds y métricas exhaustivas de rendimiento. Los resultados demostraron una superioridad significativa de los Vision Transformers, con el modelo ViT-S/16 alcanzando un accuracy medio de 0.9132 (± 0.0144) en patología única y 0.9313 (± 0.0281) en múltiples patologías, superando al mejor modelo CNN (VGG16). El análisis inferencial mediante pruebas t de Student confirmó la significancia estadística de estas diferencias ($p < 0.05$). Se concluye que los Vision Transformers ofrecen un rendimiento superior y más estable para el diagnóstico automatizado mediante imágenes radiográficas, estableciendo una base sólida para su implementación en entornos clínicos.

Palabras Clave: Vision Transformers, Redes Neuronales Convolucionales, Diagnóstico automatizado, Imágenes radiográficas, Aprendizaje profundo.

ABSTRACT

Automated medical diagnosis through deep learning techniques represents a constantly evolving field, where the selection of optimal architectures is crucial to ensure accurate and reliable diagnoses. This comparative research evaluated the predictive performance of Convolutional Neural Networks (CNN) and Vision Transformers (ViT) in automated diagnosis using radiographic images. A total of 15,834 images distributed among cases of knee osteoarthritis, pneumonia, and tuberculosis were analyzed, implementing four variants of each architecture through 5-fold cross-validation and comprehensive performance metrics. The results demonstrated significant superiority of Vision Transformers, with the ViT-S/16 model achieving a mean accuracy of 0.9132 (± 0.0144) in single pathology and 0.9313 (± 0.0281) in multiple pathologies, outperforming the best CNN model (VGG16). Inferential analysis using Student's t-tests confirmed the statistical significance of these differences ($p < 0.05$). It is concluded that Vision Transformers offer superior and more stable performance for automated diagnosis using radiographic images, establishing a solid foundation for their implementation in clinical settings.

Keywords: Vision Transformers, Convolutional Neural Networks, Automated diagnosis, Radiographic images, Deep learning.

CAPÍTULO I

INTRODUCCIÓN

En los últimos años, la aplicación de la inteligencia artificial (IA) en el ámbito de la salud ha experimentado un crecimiento exponencial, impulsado por los avances tecnológicos y la creciente disponibilidad de datos médicos digitales (Rajpurkar et al., 2018). La radiología, en particular, se ha posicionado como una de las especialidades médicas en las que la IA tiene el potencial de generar un impacto significativo, debido a su dependencia en el análisis de imágenes diagnósticas (Qin et al., 2018). En este contexto, las técnicas de aprendizaje profundo, como las redes neuronales convolucionales (CNN) y los Vision Transformers (ViT), han demostrado un rendimiento excepcional en tareas de clasificación y detección de patologías en imágenes médicas (Yang et al., 2020).

A pesar de los avances prometedores en la aplicación de la IA en radiología, aún existen desafíos importantes que deben abordarse para garantizar su implementación efectiva en la práctica clínica. Uno de los principales retos radica en la comprensión comparativa del rendimiento de diferentes arquitecturas de aprendizaje profundo en contextos específicos del diagnóstico radiológico (Pauly y Ashok, 2018). Aunque las CNN han sido ampliamente estudiadas y aplicadas en este campo, el surgimiento de los ViT ha planteado nuevas oportunidades y preguntas sobre su capacidad para capturar dependencias globales y manejar casos complejos en el análisis de imágenes médicas (Yoo et al., 2021).

En este sentido, la presente investigación se propone abordar la brecha de conocimiento existente en la comprensión comparativa del rendimiento entre CNN y ViT en el diagnóstico radiológico, centrándose específicamente en la detección de patologías

únicas frente a múltiples por región anatómica. Para ello, se llevará a cabo un análisis riguroso en escenarios clínicos reales, utilizando bases de datos de imágenes radiográficas etiquetadas y validadas por expertos (Norman et al., 2018). Los resultados de este estudio proporcionarán evidencia empírica sobre las fortalezas y limitaciones de cada arquitectura en este contexto específico, lo que permitirá orientar la toma de decisiones informadas sobre la implementación de estas tecnologías en sistemas de diagnóstico automatizado.

La relevancia de esta investigación se fundamenta en su potencial para contribuir al avance del diagnóstico médico automatizado, así como en su impacto en la mejora de la calidad de atención al paciente y la optimización de los recursos en los sistemas de salud (Jiang y Zhang, 2019). Además, los hallazgos de este estudio sentarán las bases para el desarrollo de arquitecturas híbridas que combinen las fortalezas de CNN y ViT, lo que podría conducir a un mayor rendimiento en el análisis de imágenes médicas y, en última instancia, a mejores resultados clínicos.

1.1. PLANTEAMIENTO DEL PROBLEMA

El diagnóstico médico basado en imágenes radiográficas desempeña un papel fundamental en la atención sanitaria moderna, siendo esencial para la detección temprana y el seguimiento de diversas patologías. No obstante, el aumento significativo en la cantidad de estudios radiológicos, sumado a la escasez de especialistas en radiología, ha generado cuellos de botella en los sistemas de salud a nivel global (Rajpurkar et al., 2018). Esta problemática se ha intensificado en las últimas décadas, donde el tiempo promedio de interpretación de imágenes puede extenderse hasta 72 horas en regiones con recursos limitados, impactando negativamente en la calidad de atención al paciente (Qin et al., 2018). En consecuencia, Qin et al. (2018) señalan que la demora en el diagnóstico puede

conducir a un deterioro en la salud del paciente y a un incremento en los costos asociados al tratamiento de enfermedades avanzadas.

En este contexto, la integración de la inteligencia artificial (IA) en el diagnóstico médico se presenta como una solución prometedora para mitigar estos desafíos. Yang et al. (2020) indican que las Redes Neuronales Convolucionales (CNN) han mostrado su efectividad en el análisis de imágenes médicas, alcanzando niveles de precisión que rivalizan con los de los expertos humanos en ciertas tareas diagnósticas. Sin embargo, Yoo et al. (2021) señalan que estas arquitecturas presentan limitaciones, especialmente en la captura de relaciones de largo alcance y en el procesamiento simultáneo de múltiples patologías, aspectos relevantes en el contexto del diagnóstico radiológico. Por lo tanto, es importante explorar alternativas que puedan superar estas limitaciones.

En este sentido, el reciente surgimiento de los Vision Transformers (ViT) introduce un nuevo paradigma en el procesamiento de imágenes médicas, ofreciendo ventajas en la captura de dependencias globales y en el manejo de casos complejos (Matsuda et al., 2020). A pesar de su potencial, Pauly y Ashok (2018) destacan que existe una brecha significativa en la comprensión comparativa del rendimiento entre CNN y ViT en contextos específicos del diagnóstico radiológico, particularmente en la detección de patologías únicas frente a múltiples por región anatómica. Norman et al. (2018) indican que esta falta de un análisis comparativo riguroso dificulta la toma de decisiones informadas sobre la implementación de estas tecnologías en sistemas de diagnóstico automatizado. Por consiguiente, esta investigación se propone abordar esta brecha de conocimiento, proporcionando evidencia empírica sobre el rendimiento relativo de CNN y ViT en casos específicos de diagnóstico radiológico.

1.2. FORMULACIÓN DEL PROBLEMA

1.2.1. Problema general

- ¿Cuál es la arquitectura con mejor desempeño predictivo entre las Redes Neuronales Convolucionales (CNN) y los Vision Transformers (ViT) en el diagnóstico automatizado de imágenes radiográficas para la detección de patologías únicas y múltiples por región anatómica?

1.2.2. Problemas específicos

- ¿Cuál es la arquitectura con mejor desempeño predictivo entre las Redes Neuronales Convolucionales (CNN) y los Vision Transformers (ViT) en el diagnóstico de artrosis de rodilla, como caso de estudio para la detección de una patología única en una región anatómica específica?
- ¿Cuál es la arquitectura con mejor desempeño predictivo entre las Redes Neuronales Convolucionales (CNN) y los Vision Transformers (ViT) en el diagnóstico de neumonía y tuberculosis, como caso de estudio para la detección de múltiples patologías en una misma región anatómica?

1.3. JUSTIFICACIÓN DE LA INVESTIGACIÓN

La presente investigación se justifica desde múltiples dimensiones que resaltan su relevancia para el avance del diagnóstico médico automatizado. Desde una perspectiva científica, este estudio contribuye significativamente al entendimiento de las capacidades y limitaciones de las arquitecturas de deep learning en el procesamiento de imágenes médicas. Kök et al. (2019) señalan que la comprensión detallada de estos modelos es esencial para optimizar su implementación en entornos clínicos reales. Además, la mejora

en la precisión diagnóstica puede tener un impacto directo en la calidad de atención al paciente y en los resultados de salud (Heinrich et al., 2018).

Desde el ámbito tecnológico, la comparación sistemática entre CNN y ViT proporciona información valiosa para el desarrollo de sistemas de diagnóstico más eficientes y precisos. Los resultados de esta investigación pueden orientar el diseño de arquitecturas híbridas que aprovechen las fortalezas complementarias de ambos enfoques, mejorando potencialmente la precisión diagnóstica en casos complejos (Shapiro et al., 2020). Asimismo, Patel et al. (2019) señalan que la integración de estas tecnologías en la práctica clínica podría facilitar la detección temprana de patologías, lo que es fundamental para mejorar los pronósticos de los pacientes.

Desde una perspectiva social y económica, la automatización efectiva del diagnóstico radiológico tiene el potencial de reducir significativamente los costos en salud y mejorar el acceso a servicios diagnósticos especializados. Según Bressem et al. (2020), la implementación de sistemas de IA en radiología puede reducir los tiempos de espera en hasta un 40% y los costos operativos en un 30%. Este impacto no solo beneficia a los sistemas de salud, sino que también mejora la experiencia del paciente al facilitar un acceso más rápido a diagnósticos precisos y tratamientos adecuados (Jiang y Zhang, 2019).

Por último, la elección de imágenes radiográficas como campo de aplicación se justifica por su ubicuidad en el diagnóstico médico y su papel fundamental en la detección temprana de patologías. Kanuri et al. (2022) indican que las radiografías continúan siendo la modalidad de imagen más accesible y utilizada globalmente, representando más del 60% de todos los estudios de imagen médica realizados anualmente. En este sentido, este contexto resalta la importancia de investigar y comparar las tecnologías emergentes en el

análisis de imágenes radiográficas, asegurando que se utilicen las herramientas más efectivas para mejorar los resultados en salud.

1.4. HIPOTESIS DE INVESTIGACIÓN

1.4.1. Hipótesis General

- Los Vision Transformers (ViT) presentan un mejor desempeño predictivo que las Redes Neuronales Convolucionales (CNN) en el diagnóstico automatizado de imágenes radiográficas para la detección de patologías únicas y múltiples por región anatómica.

1.4.2. Hipótesis Específicas

- Los Vision Transformers (ViT) superan a las Redes Neuronales Convolucionales (CNN) en el desempeño predictivo para el diagnóstico de artrosis en imágenes radiográficas de rodilla, como caso de estudio de detección de una patología única en una región anatómica específica.
- Los Vision Transformers (ViT) presentan un mejor desempeño predictivo que las Redes Neuronales Convolucionales (CNN) en el diagnóstico de neumonía y tuberculosis a partir de imágenes radiográficas, como caso de estudio de detección de múltiples patologías en una misma región anatómica.

1.5. OBJETIVOS DE LA INVESTIGACIÓN

1.5.1. Objetivo general

- Evaluar y comparar el desempeño predictivo de las Redes Neuronales Convolucionales (CNN) y los Vision Transformers (ViT) en el diagnóstico

automatizado de imágenes radiográficas para la detección de patologías únicas y múltiples por región anatómica.

1.5.2. Objetivos Específicos

- Evaluar y comparar el desempeño predictivo de las Redes Neuronales Convolucionales (CNN) y los Vision Transformers (ViT) en el diagnóstico de artrosis de rodilla como caso de estudio de detección de patología única por región anatómica.
- Evaluar y comparar el desempeño predictivo de las Redes Neuronales Convolucionales (CNN) y los Vision Transformers (ViT) en el diagnóstico de neumonía y tuberculosis como caso de estudio de detección de patologías múltiples por región anatómica.

CAPÍTULO II

REVISIÓN DE LITERATURA

2.1. ANTECEDENTES DE LA INVESTIGACIÓN

2.1.1. Antecedentes internacionales

Sarmadi et al. (2024) evaluaron el desempeño de los Vision Transformers (ViTs) y las Redes Neuronales Convolucionales (CNNs) en la detección de osteoporosis a partir de imágenes de rayos X, con el objetivo de identificar el modelo más eficaz para clasificar a los sujetos como normales, osteopénicos u osteoporóticos. Utilizando un enfoque basado en "transfer learning", entrenaron cuatro variantes de ViTs y CNNs preentrenados con ImageNet, aplicándolos a un dataset de 240 imágenes, desbalanceado en un 64,1 % para sujetos osteopénicos, 20,5 % para osteoporóticos y 15,4 % para normales. Los ViTs, particularmente la variante ViT_s16, alcanzaron una precisión promedio de 63,83 %, superando a la mejor CNN (VGG16) que obtuvo un 61,7 %. Además, los ViTs mostraron mejor desempeño en la clase mayoritaria (osteopénicos), mientras que ambas arquitecturas presentaron limitaciones significativas en la identificación de sujetos osteoporóticos debido al desbalance del dataset. Los autores concluyen que los ViTs, gracias a su capacidad para capturar relaciones de largo alcance en las imágenes, son una herramienta prometedora para el diagnóstico automatizado de osteoporosis, aunque subrayaron la necesidad de datos más amplios y balanceados para mejorar la generalización de los modelos.

Takahashi et al. (2024) realizaron un análisis comparativo exhaustivo entre las Redes Neuronales Convolucionales (CNN) y los Vision Transformers

(ViT) en el análisis de imágenes médicas. A través del análisis de 36 estudios cuidadosamente seleccionados, los investigadores encontraron que los modelos basados en atención (ViT) generalmente superaron a los modelos basados en convolución (CNN), con un 53% de los estudios favoreciendo los transformers, 11% las CNN, y 36% mostrando resultados comparables. Un hallazgo significativo fue que los resultados mostraron que los ViTs destacaron en tareas que requieren la identificación de patrones globales y relaciones de largo alcance, logrando en algunos casos precisiones superiores al 95 %, como en la clasificación de subtipos de enfisema, donde los ViTs lograron un 95,95 % de precisión frente a un 66,07 % de las CNNs en conjuntos de datos públicos. No obstante, las CNNs, como DenseNet121, fueron ligeramente más eficaces en ciertas aplicaciones clínicas, especialmente con conjuntos de datos más pequeños o altamente especializados.

Nafisah et al. (2023) desarrollaron una investigación centrada en la detección de COVID-19 mediante imágenes radiográficas de tórax, implementando un sistema basado en inteligencia artificial explicable (XAI) que integró Redes Neuronales Convolucionales (CNN) y Vision Transformers (ViT). El estudio, que utilizó el conjunto de datos COVID-QU-Ex con 21,165 imágenes, implementó técnicas de preprocesamiento que incluyeron segmentación mediante UNet y aumentación por rotación. Los resultados demostraron un rendimiento excepcional, con el modelo EfficientNetB7 alcanzando una precisión del 99.82%, seguido por el modelo SegFormer. La investigación destacó por incorporar herramientas de visualización para asistir a los profesionales médicos y demostró que ambas arquitecturas pueden alcanzar rendimientos comparables en

clasificación de imágenes médicas, estableciendo un precedente significativo en el diagnóstico médico asistido por computadora.

Abou Ali et al. (2023) realizaron un estudio comparativo sobre la clasificación de glóbulos blancos utilizando redes neuronales convolucionales (CNN) y Vision Transformers (ViT) mediante imágenes microscópicas. La investigación empleó dos conjuntos de datos públicos - PBC (Peripheral Blood Cells) y BCCD (Blood Cell Count and Detection) - para evaluar el rendimiento de varios modelos pre-entrenados de ImageNet ILSVRC y Google ViT. Los resultados demostraron que Google ViT superó consistentemente a los modelos CNN, alcanzando una precisión del 100 % en el conjunto de datos PBC y manteniendo un rendimiento superior incluso con datos ruidosos y de baja calidad del conjunto BCCD, donde logró una precisión del 88.36 % en comparación con un máximo del 60.25 % obtenido por las CNNs. El estudio destaca particularmente la robustez de ViT para manejar conjuntos de datos tanto pequeños como grandes, y su resistencia al ruido en los datos, estableciendo un precedente significativo para futuras aplicaciones en el análisis automatizado de imágenes médicas.

Hwang et al. (2023) desarrollaron un estudio comparativo entre modelos de Vision Transformer (ViT) y Redes Neuronales Convolucionales (CNN) para la detección de neuropatía óptica glaucomatosa a partir de fotografías de fondo de ojo. La investigación evaluó el rendimiento de ambas arquitecturas utilizando seis bases de datos públicas independientes, que incluían imágenes recopiladas de cuatro países y variaban en tamaño desde 101 hasta 800 imágenes. Mediante modelos pre-entrenados con ImageNet y un riguroso preprocesamiento de imágenes que incluía la segmentación del nervio óptico, los resultados

demostraron que los modelos ViT igualaron o superaron a los CNN en cinco de las seis bases de datos en términos de área bajo la curva ROC y precisión, destacándose especialmente en conjuntos de datos con mayor representación de controles sanos. Esta investigación aporta evidencia significativa para la selección de arquitecturas en la detección automatizada del glaucoma, sugiriendo que los modelos ViT son más apropiados para entornos con datos heterogéneos, mientras que los CNN podrían ser preferibles cuando se requiere alta especificidad.

H. E. Kim et al. (2023) realizaron un estudio comparativo entre Vision Transformer (ViT)y Redes Neuronales Convolucionales (CNN) para la clasificación automatizada de imágenes bacterianas teñidas con Gram. Los investigadores evaluaron seis modelos ViT y dos CNN utilizando dos conjuntos de datos: uno local de la Facultad de Medicina de Mannheim (n=8500) y el conjunto público DIBaS (n=660). Mediante un enfoque metodológico que exploró diferentes configuraciones de modelos, incluyendo tamaños variables, épocas de entrenamiento (1 vs. 100) y esquemas de cuantización, los resultados mostraron que los modelos ViT, particularmente aquellos con arquitectura BEiT, DeiT y ViT, superaron consistentemente a las CNN con menos épocas de entrenamiento. El modelo DeiT demostró ser el más prometedor, alcanzando una precisión de hasta 95.1% en el conjunto MHU y procesando hasta seis imágenes por segundo en configuración int8. Esta investigación proporciona pautas fundamentales para la selección y optimización de modelos VT en análisis bacteriológicos time-critical.

Ghaffari Laleh et al. (2022) desarrollaron una investigación para evaluar y mitigar la vulnerabilidad de los sistemas de inteligencia artificial (IA) frente a ataques adversarios en patología computacional antes de su implementación

clínica generalizada. Utilizando dos tareas de clasificación clínicamente relevantes (subtipos de carcinoma de células renales y cáncer gástrico), los investigadores compararon el rendimiento y la robustez de las redes neuronales convolucionales (CNN) y los transformadores de visión (ViT) ante ataques adversarios de tipo "white-box" y "black-box". Los resultados mostraron que ambos modelos alcanzaron una precisión comparable en la clasificación base, con un AUROC de 96.0% para CNN y 95.8% para ViT en carcinoma renal, y 78.2% para CNN y 76.8% para ViT en cáncer gástrico. Sin embargo, los ViT demostraron ser significativamente más robustos frente a ataques adversarios, manteniendo su rendimiento incluso bajo perturbaciones sustanciales. Esta investigación proporciona evidencia empírica crucial para priorizar el uso de arquitecturas ViT en el desarrollo de sistemas de IA para patología computacional, ofreciendo una protección inherente contra la manipulación de datos de entrada.

Dierickx et al. (2023) realizaron un estudio comparativo entre redes neuronales convolucionales (CNN), transformadores visuales (ViT) y transformadores convolucionales compactos (CCT) para la interpretación de la respuesta en frecuencia del canal en el contexto de G.Fast, una tecnología de línea digital de suscriptor. La investigación se centró en evaluar el rendimiento de estos tres modelos en la clasificación de 17 tipos diferentes de defectos en la línea mediante el análisis de la respuesta del canal (Hlog). Utilizando un conjunto de datos de 34 millones de realizaciones generadas por simulación, los investigadores compararon el rendimiento, el tiempo de entrenamiento y los recursos computacionales requeridos por cada arquitectura. Los resultados demostraron que el modelo CCT logró la mayor precisión, alcanzando un 86.04% en datos no aumentados y 76.32% en datos aumentados, superando tanto a CNN (83.35% y

65.97%) como a ViT (83.58% y 75.64%). Este estudio proporciona evidencia crucial para la selección de arquitecturas en aplicaciones de telecomunicaciones, destacando las ventajas del CCT en términos de precisión y eficiencia computacional.

Arshed et al. (2023) desarrollaron una investigación centrada en la clasificación multiclase de cáncer de piel mediante redes transformadoras de visión (ViT) y modelos preentrenados basados en redes neuronales convolucionales (CNN). El estudio abordó el desafío del desbalance de clases en los datos implementando técnicas de aumento de datos para incrementar artificialmente las muestras del conjunto HAM10000. Los investigadores evaluaron el rendimiento del modelo ViT propuesto comparándolo con 11 modelos CNN preentrenados, incluyendo variantes de ResNet, DenseNet y VGG. Los resultados demostraron que el modelo ViT superó significativamente a los modelos CNN tradicionales, alcanzando una precisión del 92.61%, una exactitud del 92.14% y una puntuación F1 del 92.17%. La principal contribución de este trabajo radica en demostrar la efectividad superior de la arquitectura ViT sobre los métodos CNN convencionales para la clasificación multiclase de cáncer de piel, estableciendo un nuevo estándar en el diagnóstico automatizado de esta enfermedad.

Garcia-Martin y Sanchez-Reillo (2023) realizaron un estudio innovador que evaluaba la aplicación de redes Vision Transformer (ViT) para el reconocimiento biométrico de venas en múltiples modalidades. Los investigadores apuntaron a mejorar la autenticación basada en venas implementando el aprendizaje por transferencia con ViTs preentrenados en conjuntos de datos ImageNet y ajustados en 14 bases de datos de venas diferentes

que cubren variantes de dedos, palma, dorso de la mano y muñeca. Utilizando una metodología integral, primero preentrenaron cuatro arquitecturas ViT diferentes (ViT-S/16, ViT-B/32, ViT-L/16 y ResNet50+ViT-L/32) en ImageNet-21k e ImageNet-1k, luego las ajustaron en las bases de datos vasculares. Sus resultados demostraron que la arquitectura ViT-L/16 logró un rendimiento superior en la mayoría de los conjuntos de datos, obteniendo tasas de identificación positiva verdadera (TPIR) que van desde el 96,00% hasta más del 99,67%. Además, introdujeron una nueva base de datos de venas de muñeca sin contacto (UC3M-CV3) con 4.800 imágenes de 100 sujetos. Esta investigación representa un avance significativo en el reconocimiento biométrico de venas al establecer a Vision Transformers como una alternativa altamente efectiva a las redes neuronales convolucionales tradicionales para esta aplicación.

2.1.2. Antecedentes nacionales

Acenjo et al. (2023) tuvieron como objetivo comparar cuatro modelos preentrenados por aprendizaje por transferencia (RESNET-50, VGG-16, Vision Transformer y Swin Transformer) para evaluar su precisión en el reconocimiento facial con oclusión de mascarillas. La metodología incluyó la creación de un conjunto de datos propio de 30 sujetos, preprocesamiento de las imágenes y entrenamiento de los modelos en diferentes escenarios. Los resultados mostraron que las arquitecturas transformers obtuvieron una mayor precisión (87-96% sin mascarilla, 61-87% con mascarilla) en comparación con las redes neuronales convolucionales (24-25% sin mascarilla, 32-53% con mascarilla). La principal contribución fue la experimentación con arquitecturas CNN y Vision Transformers, así como la creación de un conjunto de datos público,

robusteciendo el estado del arte en reconocimiento facial con oclusión por mascarillas.

Cabrejos Yalán (2022) tuvo como objetivo automatizar el proceso de diagnóstico de neumonía mediante la implementación de un sistema basado en inteligencia artificial. La metodología incluyó el desarrollo de tres motores cognitivos basados en algoritmos del estado del arte de redes neuronales convolucionales. Además, se desplegó el sistema a través de una aplicación web que permite visualizar los porcentajes de probabilidad de padecer neumonía por cada imagen de múltiples pacientes. Los resultados mostraron que se logró disminuir el número de diagnósticos incorrectos en un 80% utilizando el mejor algoritmo basado en la arquitectura de redes neuronales AlexNet, y se redujo los tiempos de espera de los pacientes en un 32%. La principal contribución fue el desarrollo de un sistema de diagnóstico automatizado de neumonía que mejora la precisión y velocidad del proceso, lo cual es crítico dado que la neumonía puede causar la muerte en niños en tan solo 2 días después del inicio de la enfermedad.

Yan An Montoya y Sofía Alejandra Cornejo (2022) tuvieron como objetivo recopilar información sobre herramientas de diagnóstico que utilizan Deep Learning (DL) en imágenes médicas para detectar COVID-19. La metodología empleada fue un estudio observacional descriptivo, basado en una Revisión Sistemática de Literatura (RSL) siguiendo la metodología de Barbara Kinchenhan. Los resultados mostraron que las Redes Neuronales Convolucionales (CNN), en sus diferentes algoritmos, alcanzaron un alto grado de precisión de más del 90% en el análisis de imágenes radiográficas para el diagnóstico de COVID-19. La principal contribución del estudio fue evidenciar la utilidad de las CNN para emitir diagnósticos oportunos de COVID-19, aunque se destaca que en la

mayoría de los trabajos revisados se aplicaron protocolos de evaluación que sobreestimaron los resultados debido a limitaciones en la cantidad y calidad de las imágenes utilizadas para entrenar los algoritmos.

2.1.3. Antecedentes locales

Huanco Ramos (2023) tuvo como objetivo determinar el modelo de proceso diagnóstico del COVID-19 mediante la aplicación de técnicas de deep learning a partir de imágenes de rayos X de tórax de los pulmones de los pacientes. La metodología incluyó la obtención de 21,165 imágenes de radiografías de tórax de pacientes, de las cuales se seleccionaron aquellas con COVID-19. Posteriormente, se realizó el preprocesamiento y procesamiento de las imágenes utilizando modelos de redes neuronales convolucionales como VGG19, DenseNet169, ResNet101 y EfficientNetB0, codificados en lenguaje Python. Los resultados mostraron que el modelo EfficientNetB0 obtuvo el desempeño más efectivo, con una exactitud del 99.130% y una precisión del 99% en la implementación del algoritmo. La principal contribución del estudio fue proporcionar una herramienta que facilite a los expertos un diagnóstico eficiente para identificar pacientes infectados con COVID-19, lo cual es fundamental en la lucha contra esta enfermedad que ocasiona neumonía y diversos síntomas.

2.2. FUNDAMENTACIÓN TEÓRICA

2.2.1. Inteligencia Artificial (IA)

La inteligencia artificial representa la capacidad de los sistemas computacionales para emular procesos cognitivos humanos, realizando tareas que tradicionalmente requerían inteligencia humana (Lecun et al., 2015). Este campo abarca un amplio espectro de capacidades, desde el reconocimiento de patrones

hasta la toma de decisiones complejas, fundamentándose en algoritmos y modelos matemáticos que permiten a las máquinas "aprender" de los datos disponibles.

En el contexto actual, la IA ha evolucionado desde sistemas basados en reglas predefinidas hacia sistemas adaptativos capaces de aprender y mejorar a partir de la experiencia. Esta evolución ha sido posible gracias al desarrollo de algoritmos más sofisticados, la disponibilidad de grandes cantidades de datos y el incremento en la capacidad de procesamiento computacional (Miralles Linares et al., 2024).

2.2.2. Evolución del aprendizaje automático hacia el aprendizaje profundo

El aprendizaje automático, como subconjunto de la IA, ha experimentado una notable evolución hasta llegar al aprendizaje profundo. Esta progresión representa un cambio paradigmático en la forma en que las máquinas procesan y aprenden de los datos. Mientras que los métodos tradicionales de aprendizaje automático requerían una extensa ingeniería de características manual, el aprendizaje profundo permite el descubrimiento automático de representaciones necesarias para la detección de patrones o la clasificación (Leoni et al., 2024).

La transición hacia el aprendizaje profundo ha sido impulsada por tres factores fundamentales: el incremento exponencial en la disponibilidad de datos digitales, el desarrollo de arquitecturas neuronales más eficientes, y la aparición de hardware especializado como las unidades de procesamiento gráfico (GPU) y las unidades de procesamiento tensorial (TPU) (Krizhevsky et al., 2017).

2.2.3. Aprendizaje profundo en el análisis de imágenes

El análisis de imágenes constituye uno de los dominios en los que el aprendizaje profundo ha evidenciado su máximo potencial. La aptitud de dichas redes para adquirir jerarquías de características cada vez más abstractas las hace particularmente eficaces en el manejo de información visual de alta complejidad. Esta característica adquiere particular relevancia en el ámbito médico, en el que la interpretación exacta de imágenes diagnósticas es de vital importancia (Slimi et al., 2024).

Los modelos de aprendizaje profundo han demostrado una capacidad excepcional para identificar patrones sutiles en imágenes médicas que podrían pasar desapercibidos incluso para especialistas experimentados. Esta capacidad se fundamenta en su habilidad para procesar y analizar grandes cantidades de datos visuales de manera sistemática y objetiva (Dang et al., 2024).

2.2.4. Aprendizaje profundo en la medicina

La incorporación del aprendizaje profundo en la práctica médica ha revolucionado diversos aspectos del diagnóstico y tratamiento. En el campo de la radiología, estos sistemas han demostrado una precisión comparable o superior a la de los radiólogos expertos en tareas específicas como la detección de nódulos pulmonares o la identificación de lesiones mamográficas (Tian et al., 2024).

Las aplicaciones actuales incluyen:

- Diagnóstico automatizado de patologías en imágenes radiográficas
- Segmentación de órganos y tumores en imágenes de resonancia magnética
- Clasificación de lesiones dermatológicas

- Análisis de imágenes histopatológicas

El impacto de estas tecnologías se extiende más allá del diagnóstico, contribuyendo también a la planificación de tratamientos, el pronóstico de enfermedades y la investigación médica (Targonski et al., 2020).

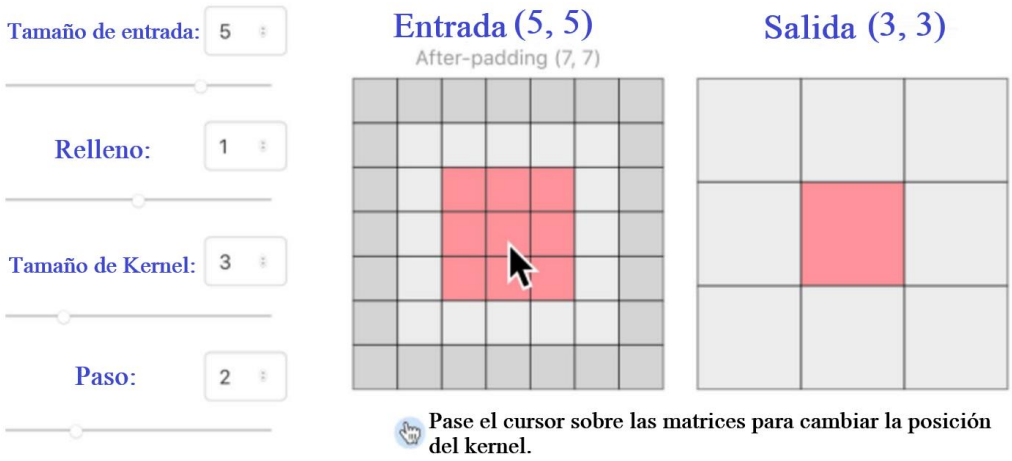
2.2.5. Redes Neuronales Convolucionales (CNN)

Las Redes Neuronales Convolucionales son un tipo especializado de red neuronal artificial diseñada específicamente para procesar datos con estructura de cuadrícula, como las imágenes. A diferencia de las redes neuronales tradicionales, las CNN incorporan operaciones de convolución que permiten capturar eficientemente patrones espaciales y jerárquicos en los datos visuales (Leite et al., 2024).

El principio fundamental de las CNN se basa en la aplicación de filtros convolucionales que se deslizan sobre la imagen de entrada, detectando características específicas en diferentes niveles de abstracción. Este proceso se inspira en el funcionamiento del córtex visual biológico, donde las neuronas responden a estímulos dentro de regiones específicas del campo visual, conocidas como campo receptivo (Ilesanmi et al., 2023).

Figura 1

Funcionamiento de un bloque de CNN.



Fuente: (Sarmadi et al., 2024)

2.2.6. Arquitectura general de una CNN

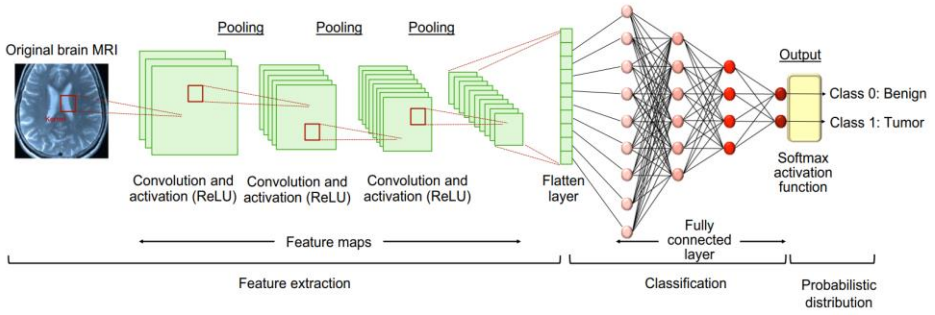
La arquitectura de una CNN se constituye mediante una disposición secuencial de capas especializadas, cada una con una función específica en el proceso de extracción y análisis de características visuales.

2.2.6.1. Capas convolucionales (convolutional layers)

Las capas convolucionales constituyen el núcleo fundamental de una CNN. En estas capas, se aplican filtros (también llamados kernels) que se deslizan sobre la imagen de entrada, realizando operaciones de convolución para detectar características específicas. Cada filtro aprende a reconocer patrones particulares, desde bordes y texturas simples en las primeras capas hasta estructuras más complejas en las capas superiores (Zhao et al., 2023).

Figura 2

Un ejemplo de análisis de imágenes médicas utilizando CNN (resonancia magnética cerebral).



Fuente: (Takahashi et al., 2024)

La operación de convolución se define matemáticamente como:

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i+m, j+n) \cdot K(m.n)$$

Donde:

- $S(i,j)$ es el valor en la posición (i,j) de la salida (o mapa de activación) después de aplicar la convolución.
- $I(i+m, j+n)$ es el valor del píxel de la imagen de entrada I en la posición $(i+m, j+n)$.
- $K(m.n)$ es el valor del filtro en la posición $(m.n)$.
- $\sum_m \sum_n$ ... indica que estamos realizando una suma sobre las posiciones del filtro (a lo largo de sus filas y columnas).
- El filtro K se desliza por toda la imagen de entrada I, y en cada posición se calcula un valor de salida $S(i,j)$ realizando la suma ponderada de los valores de I y K.

2.2.6.2. Capas completamente conectadas (fully connected layers)

Las capas completamente conectadas, típicamente ubicadas al final de la arquitectura, integran la información espacial extraída por las capas anteriores para realizar la clasificación final. Cada neurona en estas capas está conectada con todas las neuronas de la capa anterior, permitiendo un aprendizaje global de las características extraídas (Abou Ali et al., 2023).

2.2.7. Funcionamiento de una CNN

El proceso de entrenamiento de una CNN involucra dos fases principales: la propagación hacia adelante y la retropropagación. Durante la propagación hacia adelante, la imagen de entrada se procesa secuencialmente a través de todas las capas de la red, generando una predicción final. La retropropagación, por su parte, ajusta los pesos de la red minimizando una función de pérdida mediante el descenso del gradiente (Abou Ali et al., 2023).

$$Perdida\ de\ Entropia\ Cruzada = -\sum_{i=1}^{K} y_i \cdot \log(p_i)$$

Donde:

- K es el número de clases posibles.
- y_i es el valor verdadero (etiqueta) para la clase i, que generalmente es 1 si la clase i es la correcta y 0 en caso contrario (es decir, la codificación one-hot).
- p_i es la probabilidad predicha para la clase i, que es la salida del modelo, típicamente obtenida aplicando una función softmax a las salidas de la red.

- log es el logaritmo natural.

2.2.8. Ventajas de las CNN frente a otros enfoques para procesamiento de imágenes

Las CNN presentan ventajas significativas sobre otros métodos de procesamiento de imágenes. Su capacidad para aprender automáticamente jerarquías de características elimina la necesidad de diseñar extractores de características manualmente. Además, la invarianza espacial inherente a las operaciones de convolución y pooling permite reconocer objetos independientemente de su posición en la imagen (Ilesanmi et al., 2023).

En el contexto médico, las CNN han demostrado una capacidad excepcional para detectar patrones sutiles en imágenes radiográficas, superando en muchos casos el rendimiento de los métodos tradicionales de procesamiento de imágenes y aproximándose a la precisión diagnóstica de especialistas humanos.

2.2.9. Vision Transformers (ViT)

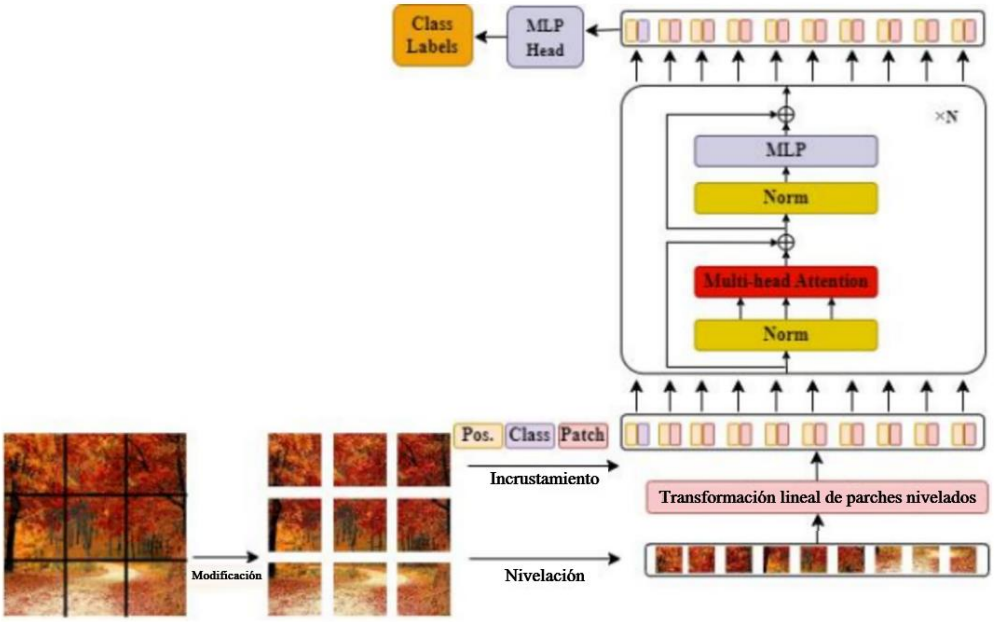
Los Vision Transformers representan un cambio paradigmático en el procesamiento de imágenes médicas, alejándose de la dependencia exclusiva de las operaciones de convolución. Esta arquitectura, introducida por Dierickx et al. (2023), demuestra que es posible aplicar el mecanismo de atención directamente a las imágenes, tratándolas como secuencias de parches. Este enfoque ha demostrado resultados excepcionales en diversas tareas de visión por computador, incluyendo la clasificación y segmentación de imágenes médicas.

La principal innovación de los ViT radica en su capacidad para capturar relaciones globales en la imagen sin las limitaciones del campo receptivo inherentes a las CNN. Esta característica resulta particularmente valiosa en el

análisis de imágenes médicas, donde las relaciones espaciales a larga distancia pueden ser cruciales para el diagnóstico (Saha et al., 2024).

Figura 3

Descripción general de una arquitectura ViT



Fuente: (Sarmadi et al., 2024)

2.2.10. Arquitectura básica de vision transformers

2.2.10.1. División de imágenes en parches

El primer paso en el procesamiento de imágenes mediante ViT consiste en la división de la imagen de entrada en parches no superpuestos de tamaño fijo. Este proceso se puede expresar matemáticamente como:

$$x \in \mathbb{R}^{H \times W \times C} \rightarrow x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$$

- H y W son la altura y anchura de la imagen.
- C es el número de canales.
- P es el tamaño de los patches.

- *N* es el número de patches.

2.2.10.2. Codificación posicional

La información posicional resulta crucial en el procesamiento de imágenes, ya que las relaciones espaciales entre diferentes regiones de la imagen contienen información diagnóstica vital. Los ViT incorporan esta información mediante embeddings posicionales que se suman a los embeddings de los parches:

$$z_0 = \left[x_{class}; x_{p^1}E;\ x_{p^2}E; \ldots;\ x_{p^N}E\right] + E_{pos}$$

- z_0 es el tensor de entrada para el primer bloque de la red Transformer, que incluye los embeddings de todos los parches y el embedding de clase.
- x_{class} es el embedding correspondiente a un token de clase (generalmente utilizado para representar la predicción final del modelo). Este token se agrega al principio de la secuencia de parches.
- x_{p^i} E son los embeddings de los parches p^i, donde i es el índice de cada parche en la imagen dividida. Estos embeddings corresponden a las representaciones vectoriales de los parches.
- E_{pos} es la matriz de embeddings posicionales aprendibles. Esta matriz tiene la misma dimensión que los embeddings de los parches y se suma a cada uno de ellos para codificar su información posicional relativa. Cada posición en E_{pos} corresponde a una posición particular de un parche en la imagen.

- [·] denota la concatenación de los diferentes embeddings, que resulta en una secuencia completa de embeddings que serán procesados por el modelo Transformer.

2.2.10.3. Mecanismo de atención (self-attention)

El mecanismo de atención constituye el núcleo de los Vision Transformers, permitiendo que cada parche de la imagen interactúe con todos los demás parches de manera directa. La atención se calcula mediante tres matrices: Query (Q), Key (K) y Value (V):

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- Q es la matriz de consultas (Queries). En el contexto de ViT, representa las representaciones de los parches de la imagen que estamos "consultando".
- K es la matriz de claves (Keys). Representa las representaciones de los parches con los cuales estamos comparando las consultas.
- V es la matriz de valores (Values). Contiene las representaciones que se utilizarán para la salida de la atención.
- d_k es la dimensión de las claves (y también de las consultas, ya que Q y K tienen la misma dimensión). El término $\sqrt{d_k}$ es utilizado para normalizar el producto interno entre las consultas y las claves, ayudando a estabilizar los gradientes durante el entrenamiento.
- softmax: La función softmax es aplicada a la matriz resultante de $\frac{QK^T}{\sqrt{d_k}}$ para obtener una distribución de probabilidad que indica la importancia relativa de cada parche respecto a los demás. Este paso

asegura que las sumas de los pesos sean 1, proporcionando una medida de la relevancia entre los parches.

2.2.11. Ventajas y limitaciones de ViT frente a CNN

Los Vision Transformers presentan ventajas significativas sobre las CNN en varios aspectos. Su capacidad para capturar relaciones globales de manera directa resulta especialmente beneficiosa en el análisis de imágenes médicas, donde las características diagnósticas pueden estar distribuidas en regiones distantes de la imagen. Esta propiedad ha demostrado ser particularmente útil en la detección de patologías que requieren la comprensión de relaciones espaciales complejas (Abou Ali et al., 2023).

Sin embargo, los ViT también presentan ciertas limitaciones. Requieren conjuntos de datos de entrenamiento más grandes que las CNN para alcanzar un rendimiento óptimo, y su coste computacional puede ser significativamente mayor. Además, la ausencia de las propiedades de invarianza a la traslación inherentes a las CNN puede afectar su rendimiento en ciertas tareas de visión por computador (Saha et al., 2024).

En el contexto específico de las imágenes médicas, los ViT han demostrado resultados prometedores, especialmente en tareas que requieren la comprensión de relaciones anatómicas complejas. No obstante, su implementación efectiva a menudo requiere adaptaciones específicas para abordar las particularidades de las imágenes médicas, como la alta resolución y la necesidad de preservar detalles finos (Dierickx et al., 2023).

2.2.12. Modelos preentrenados

Los modelos preentrenados representan redes neuronales que han sido entrenadas previamente en grandes conjuntos de datos, generalmente en tareas de clasificación de imágenes naturales. Este entrenamiento inicial permite a los modelos desarrollar representaciones jerárquicas robustas de características visuales, desde patrones básicos hasta estructuras más complejas. Según Rao et al. (2024) estos modelos constituyen una base sólida para el aprendizaje de tareas específicas en dominios especializados como la imagen médica.

El proceso de preentrenamiento típicamente involucra el uso de conjuntos de datos masivos como ImageNet, que contiene millones de imágenes etiquetadas. Durante este proceso, los modelos aprenden representaciones genéricas que pueden resultar útiles para una amplia gama de tareas de visión por computador. Este aprendizaje inicial resulta especialmente valioso cuando se trabaja con conjuntos de datos limitados en dominios específicos (Tseng y Jiang, 2024).

2.2.13. Aprendizaje por transferencia (transfer learning)

El aprendizaje por transferencia constituye un paradigma que permite aprovechar el conocimiento adquirido en una tarea fuente para mejorar el rendimiento en una tarea objetivo relacionada (L. Yang et al., 2024). En el contexto de las imágenes médicas, esto significa utilizar modelos preentrenados en grandes conjuntos de datos de imágenes naturales y adaptarlos para tareas específicas de diagnóstico médico.

Matemáticamente, el transfer learning se puede formalizar como:

2.2.13.1. Dominio fuente (Ds)

$$Ds = \{x_s, P(X_s)\}$$

- x_s es el espacio de entrada del dominio fuente (por ejemplo, el espacio de las imágenes naturales).
- $P(X_s)$ es la distribución de probabilidad de los datos en el dominio fuente.

2.2.13.2. Tarea Fuente (Ts)

$$Ts = \{Y_s, f_s(\cdot)\}$$

- Y_s son las etiquetas o resultados asociados a los datos en el dominio fuente (por ejemplo, las categorías de imágenes naturales).
- $f_s(\cdot)$ son las etiquetas o resultados asociados a los datos en el dominio fuente (por ejemplo, las categorías de imágenes naturales).

2.2.13.3. Tarea fuente (Dt)

$$Dt = \{x_t, P(X_t)\}$$

- x_t son las etiquetas o resultados asociados a los datos en el dominio fuente (por ejemplo, las categorías de imágenes naturales).
- $P(X_t)$ son las etiquetas o resultados asociados a los datos en el dominio fuente (por ejemplo, las categorías de imágenes naturales).

2.2.13.4. Tarea objetivo (Tt)

$$Tt = \{Y_t, f_t(\cdot)\}$$

- Y_t son las etiquetas o resultados asociados a los datos en el dominio objetivo (por ejemplo, categorías relacionadas con diagnósticos médicos).
- $f_t(\cdot)$ es la función predictiva para la tarea objetivo, que se desea mejorar usando el conocimiento de la tarea fuente.

Este proceso resulta especialmente efectivo cuando existe una relación entre las características de bajo nivel aprendidas en el dominio fuente y las necesarias en el dominio objetivo (C. Kim et al., 2024) .

2.2.14. Estrategias en transfer learning

2.2.14.1. Uso de modelos preentrenados como extractores de características

La estrategia de extracción de características implica utilizar las capas convolucionales o los bloques de atención de un modelo preentrenado como un extractor fijo de características, manteniendo sus pesos congelados. Esta aproximación resulta particularmente útil cuando el conjunto de datos objetivo es pequeño o cuando las características de bajo nivel son similares entre los dominios fuente y objetivo.

Las características extraídas pueden representarse como:

$$f(x) = \emptyset\, L(x;\, \theta_{fixed})$$

- $f(x)$ es el vector de características extraído de la imagen x, el cual es utilizado por las capas posteriores del modelo para tareas como clasificación, segmentación, etc.

- $\emptyset\, L$ es la función de transformación que toma la entrada x (la imagen de entrada) y la pasa a través de las primeras L capas del modelo preentrenado. Este proceso extrae las características relevantes de la imagen hasta la capa L.
- θ_{fixed} son los parámetros (pesos) del modelo preentrenado, que se mantienen fijos durante el entrenamiento. Esto significa que no se actualizan durante el proceso de entrenamiento para la tarea objetivo.

2.2.14.2. Ajuste fino (fine-tuning)

El ajuste fino representa una estrategia más flexible que permite la actualización de los pesos del modelo preentrenado durante el entrenamiento en la tarea objetivo. Este proceso puede involucrar la actualización de todas las capas del modelo o solo de un subconjunto específico, dependiendo de factores como el tamaño del conjunto de datos objetivo y su similitud con el dominio fuente.

La función objetivo para el fine-tuning se puede expresar como:

$$L(\theta) = L_{task}(f(x;\theta), y) + \lambda R(\theta)$$

- L_{task} es la función perdida específica de la tarea.
- $R(\theta)$ es el término de regularización.
- λ es el parámetro que controla el equilibrio entre la pérdida de la tarea y el término de regularización.

2.2.15. Importancia de los modelos preentrenados

En el contexto de las imágenes médicas, donde los conjuntos de datos etiquetados suelen ser limitados debido a restricciones prácticas y éticas, los modelos preentrenados cobran especial relevancia. Estos modelos permiten aprovechar el conocimiento adquirido en grandes conjuntos de datos generales para mejorar el rendimiento en tareas específicas de diagnóstico médico con conjuntos de datos más pequeños.

La efectividad de esta aproximación se fundamenta en la transferibilidad de las características aprendidas. Los estudios han demostrado que las características de bajo nivel aprendidas en imágenes naturales (como bordes, texturas y patrones básicos) resultan también útiles para el análisis de imágenes médicas. Esta transferibilidad permite obtener resultados robustos incluso con conjuntos de datos relativamente pequeños (Ahsan y Siddique, 2022).

2.2.16. Modelos CNN

2.2.16.1. VGG16 y VGG19

La familia de modelos VGG, desarrollada por el Visual Geometry Group de la Universidad de Oxford, representa una de las arquitecturas más influyentes en el campo de la visión por computador. Estos modelos se caracterizan por su diseño simple pero efectivo, empleando una serie de capas convolucionales 3x3 seguidas de capas de max-pooling (Madhur Jain et al., 2023).

La estructura de estos modelos se fundamenta en bloques convolucionales repetitivos que incrementan progresivamente el número

de filtros mientras reducen las dimensiones espaciales. VGG16 consta de 16 capas con pesos entrenables, mientras que VGG19 extiende esta arquitectura a 19 capas. La profundidad adicional de VGG19 permite la extracción de características más complejas, aunque también incrementa el costo computacional (Madhur Jain et al., 2023).

En el contexto de las imágenes médicas, estos modelos han demostrado una notable capacidad para detectar patrones relevantes en radiografías. Su arquitectura uniforme facilita la interpretabilidad de las características aprendidas, un aspecto crucial en aplicaciones médicas. Estudios recientes han demostrado que las características aprendidas por las capas intermedias de VGG son particularmente efectivas para la detección de anomalías en imágenes radiográficas (Ferdousi et al., 2024).

2.2.16.2. ResNet50 y ResNet101

Las arquitecturas ResNet introdujeron el concepto revolucionario de conexiones residuales, permitiendo el entrenamiento efectivo de redes significativamente más profundas. Estos modelos abordan el problema de la degradación del gradiente mediante conexiones de salto (skip connections) que permiten que la información fluya directamente a través de bloques residuales (Zhang et al., 2023).

La estructura básica de un bloque residual se puede expresar matemáticamente como:

$$y = F(x, \{W_i\}) + x$$

Donde $F(x, \{W_i\})$ es la transformación residual que se aprende en el bloque y x es la entrada al bloque.

ResNet50 y ResNet101, con 50 y 101 capas respectivamente, han demostrado una excepcional capacidad para capturar jerarquías complejas de características en imágenes médicas. Su arquitectura profunda, combinada con las conexiones residuales, permite la identificación de patrones sutiles en diferentes escalas, una característica particularmente valiosa en el análisis de imágenes radiográficas (Madhur Jain et al., 2023).

2.2.17. Modelos ViT

Los modelos de Vision Transformer (ViT) desarrollados por Sayak Paul y disponibles en Kaggle representan implementaciones específicas de la arquitectura ViT, cada una con características particulares que las adaptan a diferentes tareas de visión por computadora. A continuación, se presenta una explicación detallada de cada modelo:

2.2.17.1. vit_s16_fe

El modelo vit_s16_fe es una implementación del Vision Transformer que utiliza una configuración de "patch size" de 16x16 píxeles. Esto significa que la imagen de entrada se divide en parches de 16x16 píxeles antes de ser procesada por el modelo. Esta elección de tamaño de parche permite al modelo capturar características locales con un nivel de detalle moderado, equilibrando la complejidad computacional y la capacidad de representación. El sufijo "fe" indica que este modelo está diseñado para la extracción de características (feature extraction), lo que lo hace adecuado para tareas donde las representaciones aprendidas se

utilizan en etapas posteriores de procesamiento o clasificación (Sarmadi et al., 2024).

2.2.17.2. vit_r26_s32_medaug_fe

Este modelo combina una arquitectura híbrida que integra una red ResNet-26 como extractor de características inicial, seguida de un Vision Transformer que procesa parches de 32x32 píxeles. Esta combinación permite al modelo beneficiarse de las capacidades de extracción de características locales de la ResNet y las capacidades de modelado de dependencias globales del Transformer. El término "medaug" sugiere que se han aplicado técnicas de aumento de datos de intensidad media durante el entrenamiento, lo que ayuda a mejorar la robustez del modelo frente a variaciones en los datos de entrada. Al igual que el modelo anterior, el sufijo "fe" indica su uso para la extracción de características (Sarmadi et al., 2024).

2.2.17.3. vit_b32_fe

El modelo vit_b32_fe corresponde a la variante base del Vision Transformer que utiliza un tamaño de parche de 32x32 píxeles. Esta configuración implica que la imagen se divide en parches más grandes, lo que reduce el número total de parches y, por ende, la carga computacional. Sin embargo, esta elección también puede limitar la capacidad del modelo para capturar detalles finos en la imagen. Este modelo está diseñado para la extracción de características, proporcionando representaciones que pueden ser utilizadas en diversas aplicaciones de visión por computadora (Sarmadi et al., 2024).

2.2.17.4.vit_r50_l32_fe

El modelo vit_r50_l32_fe es otra arquitectura híbrida que combina una ResNet-50 como extractor de características inicial con un Vision Transformer que procesa parches de 32x32 píxeles. La ResNet-50, siendo más profunda que la ResNet-26, permite una extracción de características más compleja y detallada. La combinación con el Transformer permite al modelo capturar tanto características locales detalladas como dependencias globales en la imagen. El sufijo "fe" indica que este modelo está orientado a la extracción de características, siendo útil en escenarios donde se requieren representaciones ricas de las imágenes para tareas posteriores (Sarmadi et al., 2024).

2.2.18. Características de los datos médicos (imágenes de rayos x)

Las imágenes radiográficas presentan características únicas que las distinguen de las imágenes naturales convencionales. La naturaleza monocromática de estas imágenes, combinada con su alta resolución espacial y amplio rango dinámico, plantea desafíos específicos para su procesamiento digital. Las radiografías típicamente se caracterizan por una resolución de 8 a 16 bits por píxel, lo que resulta en un rango dinámico significativamente mayor que las imágenes convencionales de 8 bits (Takahashi et al., 2024).

La calidad de las imágenes radiográficas puede verse afectada por diversos factores técnicos y físicos, incluyendo:

- Ruido cuántico inherente al proceso de adquisición
- Variaciones en la exposición y el contraste
- Artefactos de dispersión y absorción

- Superposición de estructuras anatómicas

Estos factores deben considerarse durante el preprocesamiento para asegurar la calidad óptima de los datos de entrada a los modelos de aprendizaje profundo.

2.2.19. Preprocesamiento de datos para redes neuronales

2.2.19.1. Normalización de imágenes

La normalización de imágenes es un paso esencial en la preparación de datos para el análisis mediante redes neuronales profundas. Este proceso busca estandarizar las características estadísticas de las imágenes para facilitar el aprendizaje del modelo. Una de las técnicas más simples y ampliamente utilizadas consiste en dividir los valores de los píxeles entre 255, lo que transforma los valores originales, típicamente en el rango [0, 255], a un rango normalizado de [0, 1]. Matemáticamente, esto se puede expresar como:

$$x_{norm} = \frac{x}{255}$$

Donde:

- x es el valor original del píxel.

2.2.19.2. Conversión a un formato adecuado para modelos

La preparación de imágenes radiográficas para su procesamiento mediante modelos de aprendizaje profundo requiere una serie de transformaciones específicas. Estas incluyen el redimensionamiento de las imágenes a las dimensiones requeridas por cada arquitectura, la conversión

de formato y la organización de los datos en estructuras eficientes para el entrenamiento (Takahashi et al., 2024).

El proceso de conversión debe preservar la información diagnóstica relevante mientras se adapta a las restricciones técnicas de los modelos. Esto puede implicar:

- Conversión de profundidad de bits (por ejemplo, de 16 a 8 bits)
- Redimensionamiento con interpolación específica para imágenes médicas
- Organización en lotes optimizada para el entrenamiento

2.2.20. Aumentación de datos (data augmentation)

2.2.20.1. Definición y beneficios

La aumentación de datos representa una estrategia fundamental para abordar la limitada disponibilidad de datos etiquetados en el contexto médico. Esta técnica implica la generación de nuevas muestras de entrenamiento mediante transformaciones que preservan la información diagnóstica relevante mientras introducen variaciones realistas en los datos (Gupta y Gupta, 2019).

Los beneficios de la aumentación de datos incluyen:

- Incremento en el tamaño efectivo del conjunto de entrenamiento
- Mejora en la robustez del modelo frente a variaciones en la adquisición
- Reducción del sobreajuste
- Mayor generalización a casos clínicos diversos

2.2.20.2. Técnicas Comunes de Aumentación

Las técnicas de aumentación para imágenes radiográficas deben diseñarse considerando las características específicas de estas imágenes y la preservación de la información diagnóstica. Las transformaciones más comunes incluyen rotaciones limitadas, traslaciones y ajustes de contraste, implementadas de manera que mantengan la validez clínica de las imágenes (Wang et al., 2023).

Las transformaciones típicamente empleadas incluyen:

Transformaciones Geométricas:

$$T(x,y) = \begin{bmatrix} \cos(\theta) & -sen(\theta) \\ sen(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix}$$

Donde:

- θ es el ángulo de rotación.
- $[t_x, t_y]$ es el vector de traslación.

Transformaciones de Intensidad:

$$I'(x,y) = \alpha \cdot I(x,y) + \beta$$

Donde:

- α controla el contraste de la imagen.
- β controla el brillo de la imagen.

Estas transformaciones deben aplicarse dentro de rangos clínicamente relevantes, determinados mediante consulta con expertos radiológicos y validación empírica.

2.2.21. Métricas y Evaluación de Modelos en Clasificación

La evaluación rigurosa del rendimiento de los modelos de aprendizaje profundo en el contexto del diagnóstico médico requiere la aplicación de métricas específicas y protocolos de validación robustos. Este capítulo examina las diferentes métricas de evaluación y su relevancia en el contexto del diagnóstico automatizado mediante imágenes radiográficas (Chicco y Jurman, 2020).

2.2.22. Accuracy

La precisión general representa la proporción de predicciones correctas sobre el total de casos evaluados. En el contexto médico, si bien esta métrica proporciona una visión general del rendimiento del modelo, debe interpretarse con cautela, especialmente en situaciones con clases desequilibradas (Leite et al., 2024). El accuracy se expresa matemáticamente como:

$$Accuracy: \frac{VP + VN}{VP + VN + FP + FN}$$

Donde:

- VP: Verdaderos Positivos
- VN: Verdaderos Negativos
- FP: Falsos Positivos
- FN: Falsos Negativos

2.2.23. Precisión

Como señalan Grandini et al. (2020), la precisión, a diferencia de la accuracy, mide la proporción de predicciones positivas correctas sobre el total de predicciones positivas realizadas. Esta métrica es particularmente relevante en el

diagnóstico médico cuando el costo de los falsos positivos es alto. La precisión se expresa matemáticamente como:

$$Precisión = \frac{VP}{(VP + FP)}$$

2.2.24. Sensibilidad (Recall) y Especificidad

La sensibilidad y especificidad son métricas particularmente relevantes en el contexto médico, ya que proporcionan una visión más detallada del rendimiento del modelo en diferentes aspectos del diagnóstico (Akwo et al., 2024).

La sensibilidad (también conocida como recall o tasa de verdaderos positivos) mide la capacidad del modelo para identificar correctamente los casos positivos:

$$Recall = \frac{VP}{(VP + FN)}$$

La especificidad, por otro lado, mide la capacidad del modelo para identificar correctamente los casos negativos:

$$Especificidad = \frac{VN}{(VN + FP)}$$

En el contexto del diagnóstico médico, estas métricas tienen implicaciones clínicas directas. Una alta sensibilidad es crucial cuando el costo de no detectar una condición patológica es elevado, mientras que una alta especificidad es importante para evitar falsos positivos que podrían llevar a intervenciones innecesarias (Grandizio et al., 2024).

2.2.25. F1-Score

El F1-score representa una media armónica entre la precisión (precision) y la sensibilidad (recall), proporcionando una métrica única que balance ambos aspectos del rendimiento del modelo:

$$F1 - Score = 2 \times \frac{(Precisión \times Recall)}{(Precisión + Recall)}$$

Donde:

- Precisión = VP / (VP + FP)
- Recall = VP / (VP + FN)

Esta métrica resulta particularmente útil cuando se busca un equilibrio entre la identificación correcta de casos positivos y la minimización de falsos positivos, una consideración crucial en el diagnóstico médico (Chicco y Jurman, 2020).

2.2.26. Matriz de Confusión

La matriz de confusión proporciona una visión completa del rendimiento del modelo, permitiendo la identificación de patrones específicos en los errores de clasificación. En el contexto del diagnóstico médico, la interpretación detallada de la matriz de confusión puede revelar sesgos o debilidades específicas del modelo que requieren atención.

Para una clasificación binaria, la matriz de confusión se estructura como:

Tabla 1

Estructura de una matriz de confusión

	Predicción Positiva	**Predicción Negativa**
Clase Positiva	VP	FN
Clase Negativa	FP	VN

Fuente: (Grandini et al., 2020)

La interpretación de esta matriz debe considerar el contexto clínico específico y las implicaciones de diferentes tipos de errores en la práctica médica (Grandini et al., 2020).

2.2.27. Importancia de la Validación y Prueba en el Rendimiento del Modelo

La evaluación robusta de modelos de aprendizaje profundo en aplicaciones médicas requiere una estrategia de validación cuidadosamente diseñada. La división tradicional de datos en conjuntos de entrenamiento, validación y prueba debe realizarse considerando la distribución de casos clínicos y la variabilidad inherente en las imágenes médicas.

El proceso de validación cruzada estratificada, expresado como:

$$CV_{score} = \frac{1}{K}\sum_{k=1}^{K} score_k$$

Donde K es el número de pliegues (subconjuntos) en la validación cruzada, y $score_k$ es el rendimiento del modelo en el k-ésimo pliegue.

La evaluación final en un conjunto de prueba independiente, no utilizado durante el desarrollo del modelo, resulta esencial para estimar el rendimiento real en la práctica clínica. Este conjunto debe ser representativo de la población objetivo y las condiciones de uso previstas para el modelo (Chicco y Jurman, 2020).

2.2.28. Curva ROC (Receiver Operating Characteristic)

La curva ROC representa una herramienta fundamental para la evaluación del rendimiento de modelos de clasificación en medicina. Como explican Grandini et al. (2020) en sus trabajos seminales, esta curva visualiza la relación entre la tasa de verdaderos positivos (sensibilidad) y la tasa de falsos positivos (1 - especificidad) a diferentes umbrales de clasificación.

El área bajo la curva ROC (AUC-ROC) proporciona una medida única del rendimiento del modelo, donde:

- AUC = 1.0 representa una clasificación perfecta
- AUC = 0.5 indica un rendimiento equivalente al azar
- AUC > 0.9 generalmente indica un excelente rendimiento diagnóstico

2.3. MARCO CONCEPTUAL

2.3.1. Inteligencia artificial

La Inteligencia Artificial (IA) se define como la disciplina tecnológica que busca desarrollar sistemas capaces de realizar tareas que típicamente requieren de la inteligencia humana, tales como el razonamiento, la percepción, la toma de decisiones y el aprendizaje. La IA comprende subcampos como el aprendizaje automático (Machine Learning, ML) y el aprendizaje profundo (Deep Learning,

DL), que son fundamentales para aplicaciones avanzadas en diversos dominios, incluidos el procesamiento de imágenes y la medicina.

2.3.2. Aprendizaje automático (Machine Learning)

El aprendizaje automático es un subconjunto de la IA que se enfoca en el diseño y desarrollo de algoritmos capaces de aprender y mejorar su desempeño en tareas específicas a través de la experiencia y los datos. Estos algoritmos construyen modelos matemáticos a partir de datos de entrada, con el objetivo de realizar predicciones o tomar decisiones sin ser explícitamente programados.

2.3.3. Aprendizaje profundo (Deep Learning)

El aprendizaje profundo es una rama del aprendizaje automático que utiliza redes neuronales artificiales de múltiples capas para modelar relaciones complejas y aprender representaciones jerárquicas de datos. Este enfoque ha mostrado un desempeño sobresaliente en tareas como la clasificación de imágenes, el procesamiento del lenguaje natural y el diagnóstico médico.

2.3.4. Redes Neuronales Convolucionales (CNN)

Las Redes Neuronales Convolucionales (Convolutional Neural Networks, CNN) son arquitecturas de aprendizaje profundo diseñadas específicamente para el procesamiento y análisis de datos visuales. A través de capas convolucionales, estas redes extraen características relevantes de las imágenes, preservando la estructura espacial y reduciendo la dimensionalidad de los datos.

2.3.5. Vision Transformers (ViT)

Los Vision Transformers (ViT) representan un paradigma emergente en el análisis de imágenes, basado en el uso del mecanismo de atención para modelar relaciones entre diferentes regiones de una imagen. Este enfoque transforma las imágenes en secuencias de parches y utiliza capas de autoatención para procesar la información, logrando resultados comparables o superiores a las CNN en diversas tareas de visión computacional.

2.3.6. Aprendizaje por transferencia (Transfer Learning)

El aprendizaje por transferencia se refiere al proceso mediante el cual un modelo preentrenado en una tarea o dominio específico es adaptado para resolver problemas en un nuevo dominio o tarea. Este enfoque permite aprovechar representaciones previamente aprendidas, reduciendo los requisitos computacionales y de datos en la nueva tarea.

2.3.7. Preprocesamiento de datos

El preprocesamiento de datos es una etapa crítica en la implementación de modelos de aprendizaje profundo, especialmente en imágenes médicas. Incluye técnicas como la normalización, el escalado y la conversión de los datos a formatos compatibles con las redes neuronales. Estas prácticas aseguran que los datos estén en condiciones óptimas para su análisis y modelado.

2.3.8. Aumentación de datos

La aumentación de datos es una técnica que genera nuevas instancias a partir de datos existentes mediante transformaciones como rotaciones, escalado, recortes o cambios en la iluminación. Esto ayuda a mitigar problemas relacionados

con el sobreajuste y mejora la capacidad del modelo para generalizar en datos no vistos.

2.3.9. Métricas de evaluación de modelos

Las métricas de evaluación son fundamentales para medir el desempeño de los modelos de clasificación. Entre las métricas comunes se encuentran el Accuracy, la Precisión, el Recall, el F1-Score y el análisis de la matriz de confusión. Estas métricas proporcionan una evaluación integral, identificando fortalezas y áreas de mejora en el modelo.

2.3.10. Promedio del ROC AUC en clasificación multiclase

En problemas de clasificación multiclase, como los presentes en esta investigación, el ROC AUC promedio es una métrica integral que evalúa el desempeño global del modelo al distinguir entre múltiples categorías. Este promedio se calcula utilizando enfoques como el Macro-Averaging, que otorga igual peso a todas las clases al promediar el AUC de cada una, o el Weighted-Averaging, que ajusta el promedio según el tamaño relativo de las clases para manejar desbalances. Al resumir el rendimiento del modelo en un único valor, el ROC AUC promedio permite una evaluación comparativa justa entre diferentes enfoques y arquitecturas, proporcionando una visión clara de la capacidad discriminativa del modelo en escenarios complejos, como el diagnóstico médico multiclase, y guiando mejoras en su diseño y entrenamiento.

CAPÍTULO III

MATERIALES Y MÉTODOS

3.1. DISEÑO Y TIPO DE INVESTIGACIÓN

3.1.1. Tipo de investigación

La presente investigación se enmarca dentro del tipo comparativo, el cual tiene como propósito contrastar dos o más grupos o situaciones respecto a una variable de interés (Arias, 2012). En este caso particular, se busca evaluar y comparar el desempeño de dos tipos de modelos de aprendizaje profundo: las Redes Neuronales Convolucionales (CNN) y los Vision Transformers (ViT), en la tarea específica de diagnóstico automatizado de imágenes radiográficas. Este enfoque comparativo permite identificar las fortalezas y debilidades de cada modelo, así como determinar cuál de ellos ofrece un mejor rendimiento en la tarea mencionada. De esta manera, se contribuye al conocimiento en el área de la inteligencia artificial aplicada al campo de la medicina.

3.1.2. Diseño de investigación

El diseño de investigación empleado es no experimental y cuantitativo. En un estudio no experimental, no se realiza manipulación deliberada de las variables, sino que se observan los fenómenos tal y como se presentan en su contexto natural para posteriormente analizarlos (Hernández-Sampieri et al., 2014).

Además, el enfoque cuantitativo se caracteriza por la recolección de datos numéricos y el análisis estadístico para establecer patrones de comportamiento y probar hipótesis (Hernández-Sampieri y Mendoza-Torres, 2018). En esta investigación, se recopilan métricas de desempeño como exactitud, precisión,

recall, F1-score y área bajo la curva ROC, las cuales son analizadas estadísticamente para comparar el rendimiento de las CNN y los ViT en el diagnóstico automatizado de imágenes radiográficas.

3.2. POBLACIÓN Y MUESTRA

3.2.1. Población

La población objetivo de esta investigación está conformada por imágenes radiográficas de rodilla y tórax. Estas imágenes provienen de bases de datos públicas especializadas en el diagnóstico de patologías específicas:

- Artrosis de rodilla: La base de datos “Knee Osteoarthritis Dataset with Severity Grading” (Chen, 2018) contiene imágenes de rodilla con diferentes grados de severidad de osteoartritis. Estas imágenes fueron recopiladas de la Osteoarthritis Initiative (OAI) y etiquetadas por expertos según la escala de Kellgren-Lawrence (KL).
- Neumonía: Las imágenes de neumonía se obtuvieron de la base de datos “Chest X-Ray Images (Pneumonia)” (Kermany et al., 2018), que forma parte del conjunto de datos “Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification”. Estas imágenes fueron recolectadas de pacientes pediátricos del Guangzhou Women and Children’s Medical Center.
- Tuberculosis: La base de datos “Tuberculosis (TB) Chest X-ray Database” (Rahman et al., 2020) contiene imágenes de tórax de pacientes con tuberculosis y casos normales. Las imágenes provienen de diversas fuentes, incluyendo la National Library of Medicine (NLM), el Belarus Tuberculosis Portal y el NIAID TB Portal.

Estas bases de datos incluyen imágenes de pacientes de diferentes edades y géneros, capturadas bajo distintos protocolos y equipos radiográficos, lo que asegura una amplia representatividad de los casos reales encontrados en la práctica clínica.

3.2.2. Muestra

La muestra seleccionada para este estudio fue obtenida mediante un muestreo no probabilístico por conveniencia, seleccionando imágenes de bases de datos accesibles y relevantes para las patologías estudiadas. Este enfoque, común en estudios de salud y tecnología, permite optimizar los recursos disponibles y garantizar la calidad de los datos (Hernández-Sampieri y Mendoza-Torres, 2018). La muestra consta de un total de 15,834 imágenes radiográficas, distribuidas de la siguiente manera:

- Artrosis de rodilla: 5,778 imágenes (Chen, 2018)
- Neumonía: 5,863 imágenes (Kermany et al., 2018)
- Tuberculosis: 4,193 imágenes, compuestas por 700 imágenes públicas, 2,800 imágenes del NIAID TB Portal y 3,500 imágenes normales (Rahman et al., 2020)

El uso de un muestreo no probabilístico por conveniencia es común en investigaciones basadas en bases de datos preexistentes, debido a la accesibilidad y la calidad de las imágenes disponibles (Etikan, Musa y Alkassim, 2016). Este tamaño de muestra es significativo y permite realizar un análisis estadístico robusto para obtener resultados confiables (Hernández-Sampieri y Mendoza-Torres, 2018).

Figura 4

Imágenes radiográficas del conjunto de datos de artrosis de rodilla

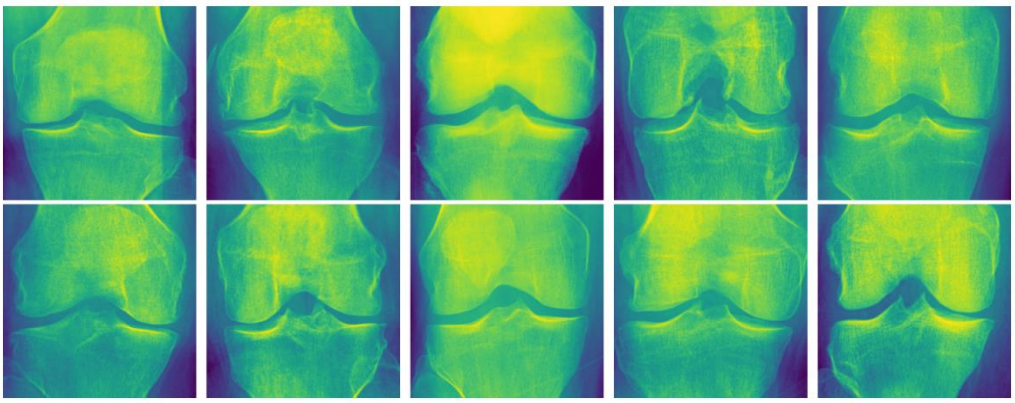


Figura 5

Imágenes radiográficas del conjunto de datos de neumonía

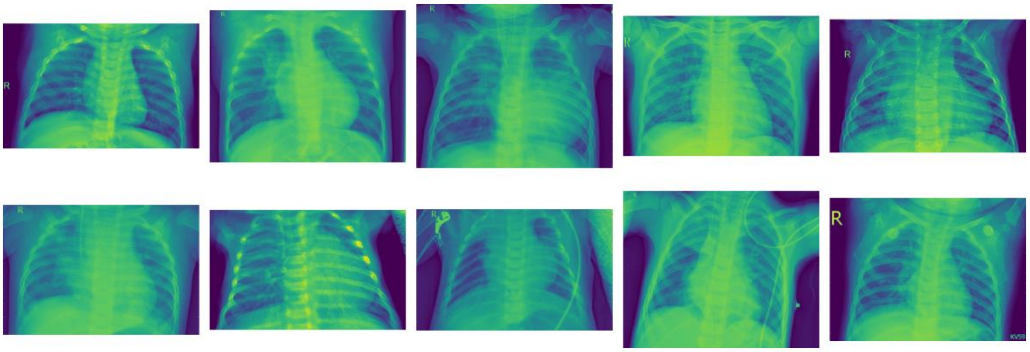
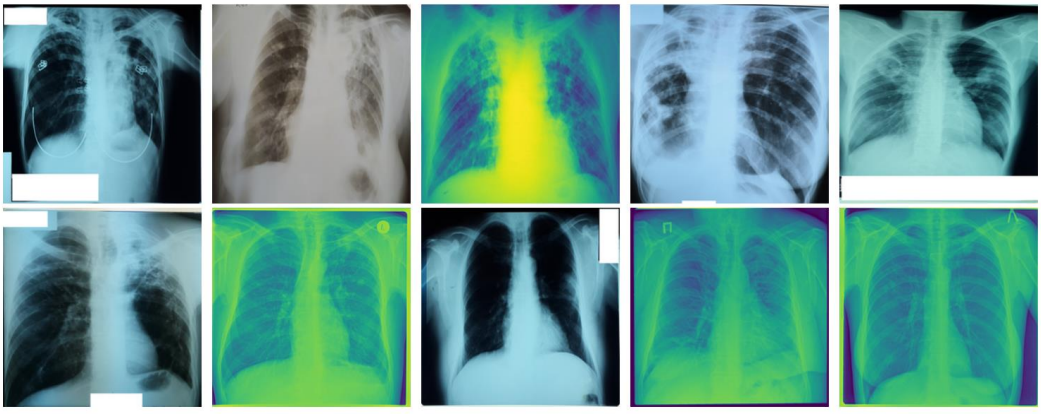


Figura 6

Imágenes radiográficas del conjunto de datos de tuberculosis



3.3. TÉCNICAS Y MÉTODOS

3.3.1. Técnicas de recolección de datos

La recolección de datos para este estudio se basa en la obtención de imágenes radiográficas provenientes de bases de datos públicas especializadas en el diagnóstico de patologías específicas, como se detalló en la sección 3.2.1. Estas bases de datos constituyen fuentes secundarias de información, ya que los datos fueron recopilados previamente por otros investigadores con fines similares a los del presente estudio (Hernández-Sampieri y Mendoza-Torres, 2018).

La técnica de recolección empleada es la revisión documental, que consiste en la consulta y extracción de información relevante de documentos, registros y materiales diversos (Arias, 2012). En este caso, se accedió a las bases de datos mencionadas y se seleccionaron las imágenes radiográficas que cumplían con los criterios de inclusión y exclusión predefinidos, obteniendo así la muestra final para el análisis.

3.3.2. Métodos de análisis

El análisis de los datos se fundamenta en la aplicación de técnicas de aprendizaje profundo para la clasificación automatizada de las imágenes radiográficas, específicamente se emplean dos tipos de modelos:

- **Redes Neuronales Convolucionales (CNN):** Se entrenan y evalúan cuatro arquitecturas de CNN ampliamente utilizadas en tareas de visión por computadora: VGG16, VGG19, ResNet50 y ResNet101.
- **Vision Transformers (ViT):** Se emplean cuatro variantes de la arquitectura ViT: ViT-S/16, ViT-R26-S32, ViT-B/32 y ViT-R50-L32.

El proceso de análisis incluye las siguientes etapas:

a) **Preprocesamiento de las imágenes:** Las imágenes se redimensionan a un tamaño uniforme de 224x224 píxeles, se normalizan y, en el caso de los ViT, se realiza una expansión de dimensiones para adecuarlas a la entrada requerida por los modelos.

b) **Entrenamiento de los modelos:** Los modelos CNN y ViT se entrenan utilizando la técnica de transfer learning, aprovechando los pesos pre-entrenados en grandes conjuntos de datos como ImageNet. Se emplean generadores de datos con aumento para el entrenamiento, y se dividen los datos en conjuntos de entrenamiento, validación y prueba (20%).

c) **Evaluación de los modelos:** Se utilizan métricas estándar de desempeño para evaluar y comparar los modelos, incluyendo exactitud (accuracy), precisión (precision), recall, F1-score y área bajo la curva ROC (ROC AUC). Además, se realiza una validación cruzada de 5 folds para obtener la media y desviación estándar de la exactitud de cada modelo.

d) **Comparación de modelos:** Se contrastan los resultados obtenidos por las CNN y los ViT en términos de las métricas de desempeño mencionadas. Se realizan comparaciones pareadas entre modelos de complejidad similar: VGG16 vs. ViT-S/16, VGG19 vs. ViT-S/32, ResNet50 vs. ViT-B/32 y ResNet101 vs. ViT-R50-L32. Se aplica la prueba t de Student para determinar si existen diferencias significativas en el desempeño, verificando previamente los supuestos de normalidad (Shapiro-Wilk) y homocedasticidad (Levene).

3.3.3. Procedimientos específicos

El desarrollo de la investigación sigue una serie de procedimientos específicos para garantizar la reproducibilidad y la validez de los resultados:

- **Preprocesamiento de imágenes:** Las imágenes se cargan y preprocesan utilizando las bibliotecas OpenCV y NumPy de Python. Se aplican las transformaciones necesarias, como redimensionamiento, normalización y expansión de dimensiones, según los requerimientos de cada modelo.
- **Implementación de modelos:** Los modelos CNN y ViT se implementan utilizando las bibliotecas TensorFlow y Keras de Python. Se aprovechan las implementaciones pre-entrenadas disponibles en la biblioteca TensorFlow Hub para los modelos ViT.
- **Entrenamiento y validación:** Se utiliza la biblioteca TensorFlow para definir los generadores de datos con aumento, los callbacks (ModelCheckpoint y EarlyStopping) y las métricas de desempeño. Los modelos se entrenan y validan utilizando los conjuntos de datos correspondientes, ajustando los hiperparámetros según sea necesario.
- **Evaluación y comparación:** Se evalúa el desempeño de los modelos en el conjunto de prueba utilizando las métricas seleccionadas. Se calculan la media y la desviación estándar de la exactitud a través de la validación cruzada de 5 folds. Se realizan las comparaciones pareadas entre modelos y se aplica la prueba t de Student para determinar la significancia estadística de las diferencias observadas.
- **Visualización de resultados:** Se generan gráficos de las métricas de evaluación para cada modelo utilizando la biblioteca Matplotlib de Python.

Se guardan las métricas y resultados en archivos de texto para su posterior análisis y reporte.

CAPITULO IV

RESULTADOS Y DISCUSIÓN

4.1. RESULTADOS DEL DESEMPEÑO PREDICTIVO DE LAS CNN Y LOS VIT EN LA DETECCIÓN DE UNA PATOLOGÍA ESPECÍFICA POR REGIÓN ANATÓMICA

4.1.1. Desempeño predictivo de las redes neuronales convolucionales (CNN)

4.1.1.1. Métricas de evaluación del modelo basado en VGG16

El modelo basado en la arquitectura VGG16 mostró un desempeño predictivo notable en la clasificación de las diferentes categorías relacionadas con la artrosis de rodilla. Este desempeño se evaluó utilizando métricas clave como precisión, recall, F1-score y accuracy general, permitiendo un análisis detallado de la capacidad del modelo para distinguir entre las distintas condiciones de artrosis en imágenes radiográficas. Además, para entender de forma más específica el comportamiento del modelo, se incluyó una matriz de confusión que detalla los aciertos y errores en cada categoría.

Los resultados generales indican que el modelo logró un accuracy global del 86.8 %, lo que significa que aproximadamente el 87 % de las predicciones realizadas por el modelo fueron correctas. La precisión ponderada fue de 0.871, lo que refleja la capacidad del modelo para realizar predicciones confiables, mientras que el recall ponderado de 0.868 indica que el modelo identificó correctamente la mayoría de las instancias

reales. Finalmente, el F1-score ponderado alcanzó un valor de 0.869, lo que muestra un buen equilibrio general entre precisión y recall.

En la Tabla 2, que se incluye al final de este párrafo, se muestran las métricas desglosadas por categoría. Para la categoría "SIN ARTROSIS", el modelo mostró un excelente desempeño, con una precisión de 0.904 y un recall de 0.916. Esto significa que el modelo no solo predijo correctamente la mayoría de las instancias etiquetadas como "SIN ARTROSIS", sino que también evitó en gran medida clasificaciones incorrectas de otras categorías en esta clase. Este resultado se tradujo en un F1-score de 0.910, lo que representa un balance óptimo entre precisión y sensibilidad para esta clase.

Tabla 2

Métricas de evaluación del modelo basado en VGG16 en la detección de una patología por región anatómica

	Precision	**Recall**	**F1-score**	**Support**
SIN ARTROSIS	0.904	0.916	0.910	585
DUDOSA	0.930	0.799	0.859	268
LEVE	0.864	0.906	0.884	393
MODERADA	0.798	0.798	0.812	196
SEVERA	0.469	0.511	0.489	45
Accuracy			0.868	1487
Macro avg	0.793	0.792	0.791	1487
Weighted avg	0.871	0.868	0.869	1487

Fuente: Elaboración propia

La categoría "DUDOSA" presentó un comportamiento interesante. Con una precisión de 0.930, el modelo fue altamente confiable al predecir esta clase, pero el recall más bajo (0.799) indicó que hubo una proporción considerable de instancias reales de esta categoría que no fueron detectadas. Esto sugiere que el modelo tiende a ser más conservador en la identificación de casos "DUDOSA", probablemente clasificándolos en otras categorías cercanas, como "LEVE". A pesar de esto, el F1-score de 0.859 refleja un desempeño sólido, aunque con cierto margen de mejora en la sensibilidad para esta categoría.

En el caso de la categoría "LEVE", el modelo mostró una notable capacidad para identificar correctamente esta condición. El recall de 0.906 fue el más alto entre todas las clases, lo que indica una gran sensibilidad para detectar casos "LEVE". La precisión de 0.864 complementa este resultado, mostrando que el modelo fue bastante confiable al predecir esta clase. El F1-score de 0.884 confirma un desempeño equilibrado y robusto en esta categoría, destacando su relevancia como una de las clases mejor identificadas por el modelo.

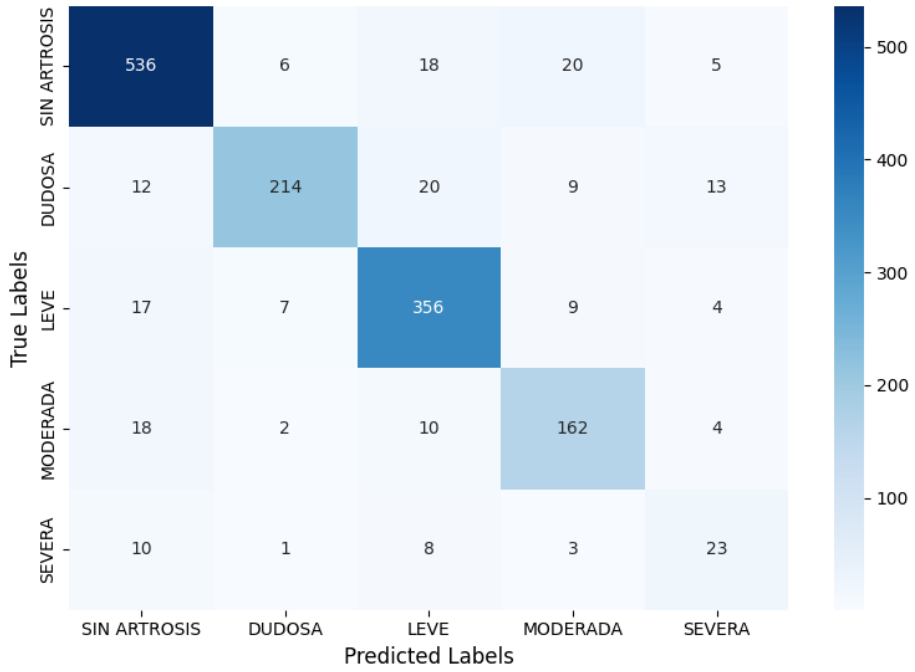
Para las categorías "MODERADA" y "SEVERA", el modelo enfrentó mayores dificultades, lo cual se refleja en métricas más bajas. En el caso de la categoría "MODERADA", el recall fue de 0.827, lo que significa que el modelo identificó correctamente una buena proporción de los casos reales. Sin embargo, la precisión de 0.798 y el F1-score de 0.812 sugieren que todavía existen desafíos para diferenciar esta categoría de otras, especialmente "LEVE" y "SEVERA". Por otro lado, la categoría "SEVERA" presentó el desempeño más limitado, con una precisión de

0.469 y un recall de 0.511. Esto indica que el modelo tuvo dificultades significativas tanto para identificar esta condición como para evitar clasificaciones erróneas hacia otras categorías. El F1-score de 0.489 refuerza esta conclusión y evidencia que el bajo soporte de datos para esta clase (45 imágenes) pudo haber afectado negativamente el desempeño.

La matriz de confusión, presentada en la Figura 7, complementa las métricas y ofrece una visión más detallada de los aciertos y errores del modelo. En esta figura, se observa que para la categoría "SIN ARTROSIS", la mayoría de las imágenes se clasificaron correctamente, con un total de 536 aciertos sobre 585 instancias reales. Las confusiones principales ocurrieron hacia las categorías "LEVE" y "MODERADA", lo que sugiere que el modelo ocasionalmente confunde condiciones cercanas en la escala de severidad. En la categoría "DUDOSA", 214 de las 268 instancias fueron correctamente identificadas, pero el modelo mostró una tendencia a clasificarlas como "LEVE". Este comportamiento podría reflejar similitudes radiográficas entre estas categorías, lo que dificulta la distinción.

Figura 7

Matriz de confusión para el modelo basado en VGG16 en la detección de una patología por región anatómica



La categoría "LEVE" fue una de las mejor identificadas, con 356 aciertos sobre 393 casos reales, y los errores se distribuyeron de manera uniforme entre las demás categorías. Esto refuerza la idea de que el modelo tiene una alta sensibilidad para esta condición. En contraste, las categorías "MODERADA" y "SEVERA" evidenciaron mayores dificultades, con un mayor número de confusiones entre ellas. Por ejemplo, en la clase "SEVERA", solo 23 de las 45 instancias reales se clasificaron correctamente, mientras que los errores se distribuyeron principalmente hacia "MODERADA" y "LEVE". Este patrón resalta la complejidad de diferenciar entre niveles avanzados de artrosis por este modelo.

En términos generales, los resultados sugieren que el modelo basado en VGG16 es altamente efectivo para detectar condiciones comunes y menos severas, como "SIN ARTROSIS" y "LEVE". Sin embargo, presenta limitaciones importantes en la identificación de categorías menos representadas, como "SEVERA". Además, el desempeño global del modelo, con métricas ponderadas superiores al 86%, demuestra su potencial como herramienta de diagnóstico automatizado.

4.1.1.2. Métricas de evaluación del modelo basado en VGG19

El desempeño del modelo basado en la arquitectura VGG19 se evaluó utilizando las métricas de precisión, recall, F1-score y accuracy general, así como la matriz de confusión, para analizar tanto los aciertos como los errores en la clasificación de las distintas categorías de artrosis en imágenes radiográficas. Los resultados mostraron un desempeño moderado, con limitaciones claras en algunas categorías, especialmente aquellas con menor soporte de datos.

El modelo alcanzó un accuracy global del 71.4 %, lo que significa que aproximadamente el 71 % de las predicciones realizadas fueron correctas. Las métricas ponderadas ofrecen un panorama más amplio del desempeño, con una precisión promedio de 0.735, un recall promedio de 0.714 y un F1-score promedio de 0.721. Estas cifras indican que, aunque el modelo logra un nivel razonable de precisión global, hay categorías en las que presenta un desempeño considerablemente inferior.

En la Tabla 3, que se incluye a continuación, se observan las métricas desglosadas por categoría. En la clase "SIN ARTROSIS", el

modelo mostró la mayor precisión entre todas las categorías (0.854), reflejando su capacidad para identificar correctamente una proporción significativa de los casos pertenecientes a esta clase. Sin embargo, el recall fue de 0.720, lo que indica que el modelo no detectó el 28 % de las instancias reales de esta categoría. Este desequilibrio entre precisión y recall produjo un F1-score de 0.781, que si bien es adecuado, denota que el modelo no es completamente consistente en la identificación de esta clase.

Tabla 3

Métricas de evaluación del modelo basado en VGG19 en la detección de una patología por región anatómica

	Precision	Recall	F1-score	Support
SIN ARTROSIS	0.854	0.720	0.781	585
DUDOSA	0.679	0.672	0.675	268
LEVE	0.720	0.779	0.748	393
MODERADA	0.608	0.719	0.659	196
SEVERA	0.194	0.311	0.239	45
Accuracy			0.714	1487
Macro avg	0.611	0.640	0.621	1487
Weighted avg	0.735	0.714	0.721	1487

Fuente: Elaboración propia

La clase "DUDOSA" mostró métricas uniformes, con una precisión de 0.679 y un recall de 0.672, lo que sugiere que el modelo tiene una capacidad moderada para identificar y clasificar correctamente esta categoría. Sin embargo, estas métricas también indican que una proporción significativa de casos reales no fue identificada correctamente. El F1-score

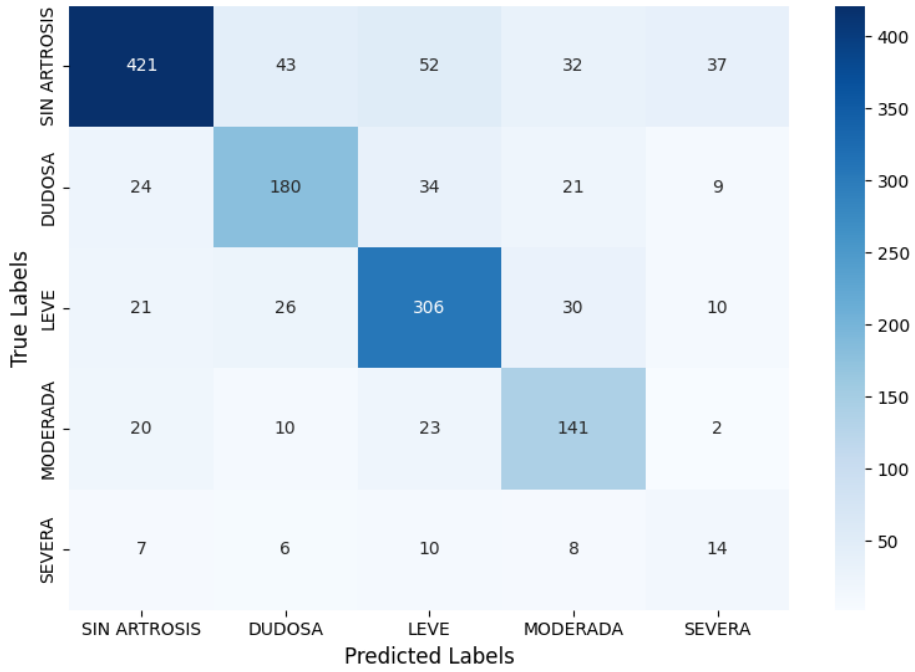
de 0.675, aunque razonable, refleja las limitaciones del modelo al trabajar con esta clase, especialmente considerando que las confusiones con otras categorías más comunes, como "LEVE", son notorias.

La categoría "LEVE" mostró un desempeño relativamente mejor en comparación con otras clases. Con un recall de 0.779, el modelo identificó correctamente el 77.9 % de las instancias reales, mientras que la precisión fue de 0.720, indicando que una proporción significativa de las predicciones realizadas para esta clase fue acertada. Esto se tradujo en un F1-score de 0.748, que señala un desempeño sólido y equilibrado, aunque todavía con margen para la mejora.

Las clases "MODERADA" y "SEVERA" presentaron un desempeño más limitado. En la clase "MODERADA", el recall de 0.719 indica que el modelo logró identificar una buena proporción de los casos reales, pero con una precisión de 0.608, lo que sugiere que un porcentaje importante de las predicciones realizadas para esta clase fue incorrecto. El F1-score de 0.659 refuerza esta conclusión, evidenciando que la capacidad del modelo para clasificar esta clase de forma consistente es moderada. La categoría "SEVERA" presentó el desempeño más bajo entre todas, con una precisión de 0.194, un recall de 0.311 y un F1-score de apenas 0.239. Estos resultados indican que el modelo tuvo grandes dificultades tanto para identificar esta clase como para evitar clasificaciones erróneas hacia otras categorías.

Figura 8

Matriz de confusión para el modelo basado en VGG19 en la detección de una patología por región anatómica



La matriz de confusión, presentada en la Figura 8, proporciona información más detallada sobre las clasificaciones realizadas por el modelo. En la clase "SIN ARTROSIS", se observa que de las 585 instancias reales, 421 fueron clasificadas correctamente. Sin embargo, las confusiones hacia las categorías "DUDOSA" y "LEVE" fueron significativas, con 43 y 52 casos respectivamente. Esto pone de manifiesto que, aunque el modelo tiene un desempeño aceptable en esta clase, todavía presenta problemas para diferenciarla de otras condiciones relacionadas.

La clase "DUDOSA" mostró una tasa de aciertos moderada, con 180 instancias clasificadas correctamente. Las confusiones principales ocurrieron hacia la categoría "LEVE", con 34 casos erróneos, lo que sugiere que estas dos categorías comparten características similares que

dificultan la diferenciación. En la clase "LEVE", el modelo clasificó correctamente 306 de las 393 instancias reales, pero se observan errores hacia las categorías "DUDOSA" y "MODERADA", lo que nuevamente refuerza la idea de que estas categorías presentan patrones radiográficos superpuestos.

En cuanto a la categoría "MODERADA", el modelo clasificó correctamente 141 de las 196 instancias reales. Sin embargo, hubo confusiones significativas hacia "LEVE" y "SIN ARTROSIS", indicando que el modelo tiene dificultades para identificar características distintivas de esta clase. La categoría "SEVERA" fue la más problemática, con solo 14 aciertos de las 45 instancias reales. Las confusiones estuvieron mayoritariamente dirigidas hacia "MODERADA" y "LEVE", lo que sugiere que el modelo no pudo aprender de manera efectiva los patrones característicos de esta clase.

De estos resultados se puede generalizar de la siguiente manera, el desempeño del modelo VGG19 muestra una capacidad moderada para clasificar correctamente las categorías más representadas, como "SIN ARTROSIS" y "LEVE", mientras que su rendimiento es considerablemente más bajo en clases menos representadas, como "SEVERA". Estas limitaciones son evidentes tanto en las métricas de evaluación como en la matriz de confusión, lo que refleja la dificultad del modelo para manejar categorías complejas y con menor soporte de datos. Los resultados obtenidos permiten observar patrones en el comportamiento del modelo, destacando fortalezas en la clasificación de

condiciones comunes y limitaciones en el manejo de categorías con características más desafiantes

4.1.1.3. Métricas de evaluación del modelo basado en ResNet50

El modelo basado en la arquitectura ResNet50 fue evaluado empleando métricas de precisión, recall, F1-score y accuracy general, junto con una matriz de confusión, para analizar en detalle el comportamiento del modelo en la clasificación de las distintas categorías de artrosis de rodilla. Los resultados indican un desempeño general adecuado, con métricas más consistentes que otros modelos previamente evaluados, aunque con las mismas limitaciones observadas en las categorías menos representadas.

El modelo alcanzó un accuracy global de 78.5 %, lo que indica que el 78.5 % de las predicciones realizadas por el modelo fueron correctas. Las métricas ponderadas muestran una precisión promedio de 0.802, un recall promedio de 0.785 y un F1-score promedio de 0.792. Estas cifras reflejan un buen desempeño general, especialmente en las categorías mayoritarias, aunque persisten dificultades en las categorías con menor soporte.

En la Tabla 4, que se presenta a continuación, se muestran las métricas desglosadas por categoría. Para la clase "SIN ARTROSIS", el modelo logró una precisión de 0.881 y un recall de 0.798, lo que indica que identificó correctamente el 79.8 % de las instancias reales de esta clase. El F1-score de 0.838 refuerza esta observación, mostrando un

desempeño sólido en la detección de esta categoría, aunque con margen de mejora en la sensibilidad.

Tabla 4

Métricas de evaluación del modelo basado en ResNet50 en la detección de una patología por región anatómica

	Precision	Recall	F1-score	Support
SIN ARTROSIS	0.881	0.798	0.838	585
DUDOSA	0.733	0.787	0.759	268
LEVE	0.834	0.804	0.819	393
MODERADA	0.724	0.776	0.749	196
SEVERA	0.263	0.467	0.336	45
Accuracy			0.785	1487
Macro avg	0.687	0.726	0.700	1487
Weighted avg	0.802	0.785	0.792	1487

Fuente: Elaboración propia

En la categoría "DUDOSA", el modelo logró métricas equilibradas, con un recall de 0.787, lo que indica que identificó correctamente el 78.7 % de las instancias reales. La precisión de 0.733 muestra un nivel razonable de confianza en las predicciones realizadas para esta clase. El F1-score de 0.759 evidencia un buen desempeño general, aunque con cierta proporción de errores, probablemente hacia categorías cercanas como "LEVE".

La clase "LEVE" fue una de las mejor clasificadas por el modelo. Con un recall de 0.804, el modelo fue capaz de identificar correctamente más del 80 % de los casos reales, mientras que la precisión alcanzó un

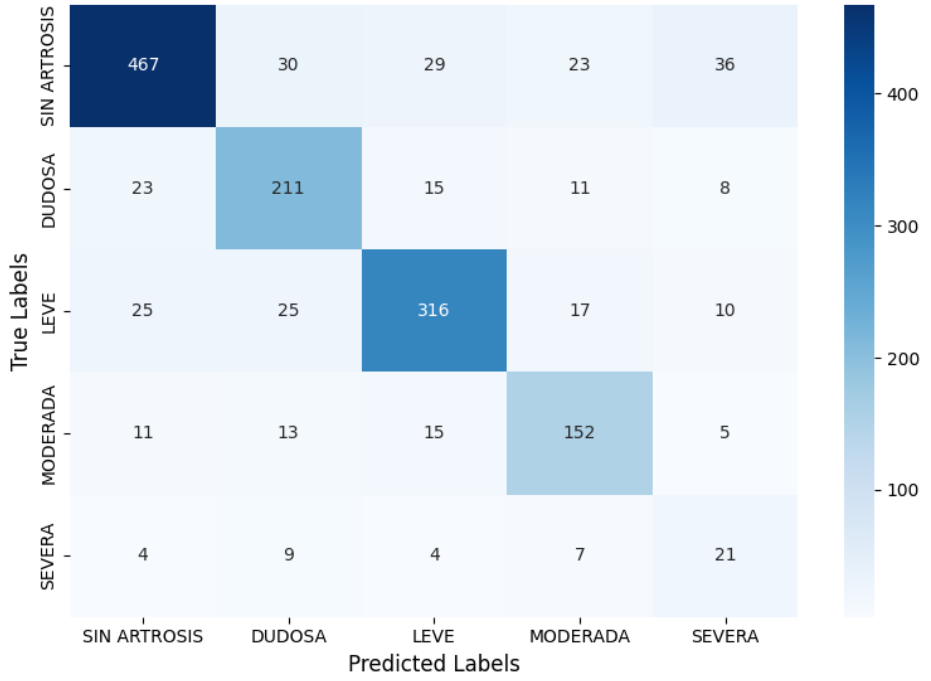
valor de 0.834, lo que indica que la mayoría de las predicciones para esta clase fueron correctas. El F1-score de 0.819 refuerza el buen desempeño del modelo en esta categoría, destacando como una de las más consistentes.

En cuanto a la categoría "MODERADA", el modelo logró un recall de 0.776, lo que sugiere una alta sensibilidad para detectar esta clase. Sin embargo, la precisión de 0.724 indica que una proporción significativa de las predicciones realizadas fueron incorrectas. El F1-score de 0.749 refleja un desempeño adecuado, aunque limitado por confusiones con categorías como "LEVE" y "SIN ARTROSIS".

La clase "SEVERA" fue nuevamente la más desafiante para el modelo. Con una precisión de 0.263 y un recall de 0.467, los resultados indican que el modelo tuvo dificultades tanto para identificar correctamente esta clase como para evitar clasificaciones erróneas hacia otras categorías. El F1-score de 0.336 evidencia estas limitaciones, mostrando que la baja representación de esta clase en los datos afecta significativamente el desempeño.

Figura 9

Matriz de confusión para el modelo basado en ResNet50 en la detección de una patología por región anatómica



La matriz de confusión, presentada en la Figura 9, ofrece un análisis más detallado de los aciertos y errores del modelo. En la clase "SIN ARTROSIS", el modelo clasificó correctamente 467 de las 585 instancias reales. Sin embargo, se observaron confusiones hacia las categorías "DUDOSA" y "LEVE", con 30 y 29 casos respectivamente. Esto indica que, aunque el modelo tiene un buen desempeño en esta categoría, persisten desafíos para diferenciarla de otras condiciones relacionadas.

En la clase "DUDOSA", el modelo identificó correctamente 211 de las 268 instancias reales, pero los errores se distribuyeron principalmente hacia "SIN ARTROSIS" y "LEVE", lo que sugiere similitudes radiográficas entre estas categorías. La categoría "LEVE" tuvo un

desempeño destacado, con 316 de las 393 instancias clasificadas correctamente. Los errores estuvieron relacionados principalmente con "DUDOSA" y "MODERADA", lo que refuerza la idea de que estas clases comparten características que dificultan la distinción.

Para la clase "MODERADA", el modelo clasificó correctamente 152 de las 196 instancias reales, pero presentó errores hacia "LEVE" y "SIN ARTROSIS". Finalmente, la categoría "SEVERA" mostró un desempeño limitado, con solo 21 de las 45 instancias reales clasificadas correctamente. La mayoría de los errores se distribuyeron hacia "MODERADA" y "LEVE", lo que refleja una dificultad persistente para identificar esta clase.

En general, los resultados de ResNet50 destacan un buen desempeño en las categorías más representadas, como "SIN ARTROSIS" y "LEVE", mientras que las categorías con menor soporte, como "SEVERA", continúan siendo un desafío. Estas métricas y patrones reflejan una capacidad sólida del modelo para manejar datos complejos, aunque con limitaciones evidentes en la diferenciación de categorías menos comunes o con características similares.

4.1.1.4. Métricas de evaluación del modelo basado en ResNet101

El modelo basado en la arquitectura ResNet101 fue evaluado utilizando las métricas de precisión, recall, F1-score y accuracy general, además de una matriz de confusión para analizar detalladamente los aciertos y errores en la clasificación de las categorías relacionadas con la artrosis de rodilla. Los resultados obtenidos muestran un desempeño

moderado, con fortalezas en algunas categorías más representadas y limitaciones significativas en las menos comunes.

El modelo alcanzó un accuracy global del 73.2 %, lo que significa que el 73.2 % de las predicciones realizadas fueron correctas. Las métricas ponderadas muestran una precisión promedio de 0.773, un recall promedio de 0.732 y un F1-score promedio de 0.743, reflejando un desempeño aceptable en términos generales, aunque con variaciones notables entre las categorías.

En la Tabla 5, que se incluye a continuación, se presentan las métricas detalladas por clase. La categoría "SIN ARTROSIS" logró una precisión de 0.914, la más alta entre todas las categorías, lo que indica que el modelo fue muy confiable al clasificar esta clase. Sin embargo, el recall fue de 0.687, lo que implica que el modelo no identificó correctamente el 31.3 % de los casos reales pertenecientes a esta categoría. Este desequilibrio resultó en un F1-score de 0.784, lo que, aunque aceptable, pone en evidencia cierta inconsistencia en la capacidad del modelo para detectar todas las instancias reales de esta clase.

Tabla 5

Métricas de evaluación del modelo basado en ResNet101 en la detección de una patología por región anatómica

	Precision	Recall	F1-score	Support
SIN ARTROSIS	0.914	0.687	0.784	585
DUDOSA	0.674	0.810	0.736	268
LEVE	0.778	0.758	0.768	393
MODERADA	0.610	0.791	0.689	196
SEVERA	0.193	0.378	0.256	45
Accuracy			0.732	1487
Macro avg	0.634	0.685	0.647	1487
Weighted avg	0.773	0.732	0.743	1487

Fuente: Elaboración propia

La categoría "DUDOSA" mostró un comportamiento destacable en términos de recall, con un valor de 0.810, lo que significa que el modelo fue capaz de identificar correctamente el 81 % de las instancias reales pertenecientes a esta clase. Sin embargo, la precisión de 0.674 indica que una proporción considerable de predicciones realizadas para esta categoría fueron incorrectas, probablemente clasificadas erróneamente hacia otras clases cercanas. El F1-score de 0.736 refleja un desempeño general bueno pero no exento de errores.

En cuanto a la categoría "LEVE", el modelo mostró métricas equilibradas. El recall de 0.758 indica que el modelo identificó correctamente la mayoría de los casos reales de esta categoría, mientras que la precisión de 0.778 refleja un alto nivel de confianza en las predicciones realizadas. Este balance entre precisión y recall resultó en un

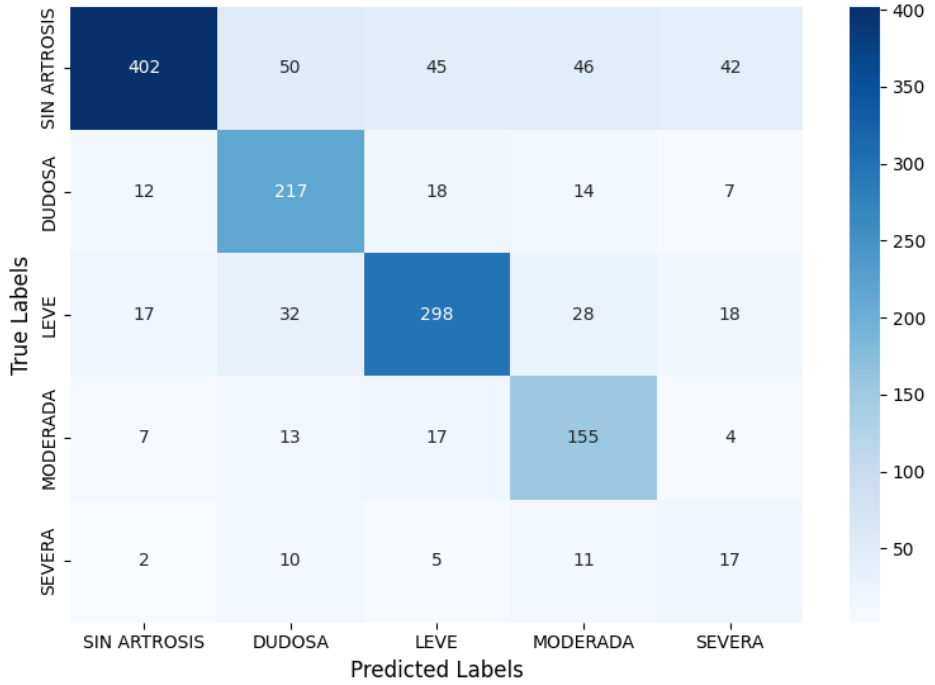
F1-score de 0.768, posicionando a esta categoría como una de las mejor clasificadas por ResNet101.

Para la clase "MODERADA", el modelo presentó un desempeño interesante. El recall de 0.791 indica que el modelo logró identificar correctamente casi el 80 % de las instancias reales de esta clase, lo que refleja una alta sensibilidad para detectar casos de artrosis moderada. Sin embargo, la precisión fue más baja, con un valor de 0.610, lo que sugiere que una proporción significativa de las predicciones realizadas para esta clase fueron incorrectas. Esto se tradujo en un F1-score de 0.689, que aunque moderado, evidencia las dificultades del modelo para separar claramente esta clase de otras categorías relacionadas.

La categoría "SEVERA" fue nuevamente la que presentó mayores desafíos para el modelo. Con una precisión de 0.193 y un recall de 0.378, los resultados indican que el modelo tuvo grandes dificultades tanto para identificar correctamente los casos de esta clase como para evitar clasificaciones erróneas hacia otras categorías. El F1-score de 0.256 refleja estas limitaciones, mostrando que esta clase sigue siendo un reto importante para la arquitectura ResNet101, probablemente debido a la baja representación de esta categoría en los datos.

Figura 10

Matriz de confusión para el modelo basado en ResNet10 en la detección de una patología por región anatómica



La matriz de confusión, presentada en la Figura 10, permite un análisis más detallado de los aciertos y errores. En la clase "SIN ARTROSIS", se observa que de las 585 instancias reales, 402 fueron clasificadas correctamente. Sin embargo, hubo confusiones frecuentes hacia las categorías "DUDOSA" y "LEVE", con 50 y 45 casos respectivamente, lo que pone de manifiesto que el modelo tuvo dificultades para diferenciar esta clase de otras condiciones similares. En la clase "DUDOSA", el modelo clasificó correctamente 217 de las 268 instancias reales, pero se observaron errores hacia "LEVE" y "SIN ARTROSIS", lo que sugiere que estas clases presentan características radiográficas superpuestas.

La categoría "LEVE" mostró un buen desempeño, con 298 de las 393 instancias clasificadas correctamente. Sin embargo, hubo errores significativos hacia las clases "DUDOSA" y "MODERADA", lo que refuerza la idea de que estas categorías tienen patrones similares que dificultan su distinción. Para la clase "MODERADA", el modelo clasificó correctamente 155 de las 196 instancias reales, pero presentó errores hacia "LEVE" y "SIN ARTROSIS". Finalmente, la categoría "SEVERA" mostró un desempeño muy limitado, con solo 17 aciertos de las 45 instancias reales. La mayoría de los errores se distribuyeron hacia "MODERADA" y "LEVE", lo que refleja una dificultad importante para identificar esta clase con precisión.

4.1.2. Desempeño predictivo de los Vision Transformers (ViT)

4.1.2.1. Métricas de evaluación del modelo basado en ViT-S/16

El modelo basado en Vision Transformers, específicamente en la arquitectura ViT-S/16, mostró un desempeño notablemente elevado en la clasificación de las distintas categorías relacionadas con la artrosis de rodilla. Este modelo fue evaluado utilizando métricas de precisión, recall, F1-score y accuracy general, complementado con una matriz de confusión que permitió analizar de manera detallada los aciertos y errores del modelo. Los resultados obtenidos reflejan un desempeño robusto y consistente, especialmente en las categorías mayoritarias.

El modelo alcanzó un accuracy global del 90.1 %, lo que significa que más del 90 % de las predicciones realizadas fueron correctas. Las métricas ponderadas muestran una precisión promedio de 0.904, un recall

promedio de 0.901 y un F1-score promedio de 0.902. Estos resultados indican un desempeño superior en términos generales, con una capacidad sobresaliente para identificar correctamente las instancias reales en la mayoría de las categorías.

En la Tabla 6, que se incluye a continuación, se presentan las métricas desglosadas por clase. Para la categoría "SIN ARTROSIS", el modelo logró una precisión de 0.901 y un recall de 0.966, lo que significa que el 96.6 % de las instancias reales de esta clase fueron correctamente identificadas. El F1-score de 0.932 refleja un excelente equilibrio entre precisión y sensibilidad, destacando el desempeño excepcional del modelo en esta categoría.

Tabla 6

Métricas de evaluación del modelo basado en ViT-S/16 en la detección de una patología por región anatómica

	Precision	**Recall**	**F1-score**	**Support**
SIN ARTROSIS	0.901	0.966	0.932	585
DUDOSA	0.951	0.869	0.908	268
LEVE	0.948	0.873	0.909	393
MODERADA	0.842	0.867	0.854	196
SEVERA	0.569	0.644	0.604	45
Accuracy			0.901	1487
Macro avg	0.842	0.844	0.842	1487
Weighted avg	0.904	0.901	0.902	1487

Fuente: Elaboración propia

La categoría "DUDOSA" presentó resultados igualmente sólidos, con una precisión de 0.951 y un recall de 0.869, lo que indica que el modelo identificó correctamente el 86.9 % de las instancias reales. El F1-score de 0.908 reafirma el excelente desempeño de esta categoría, reflejando una capacidad sobresaliente del modelo para clasificar correctamente esta clase.

En la categoría "LEVE", el modelo continuó mostrando métricas consistentes y elevadas. El recall de 0.873 significa que el modelo fue capaz de identificar correctamente casi el 87.3 % de los casos reales, mientras que la precisión de 0.948 indica un alto nivel de confianza en las predicciones realizadas. El F1-score de 0.909 refuerza la percepción de un desempeño robusto y equilibrado en esta categoría.

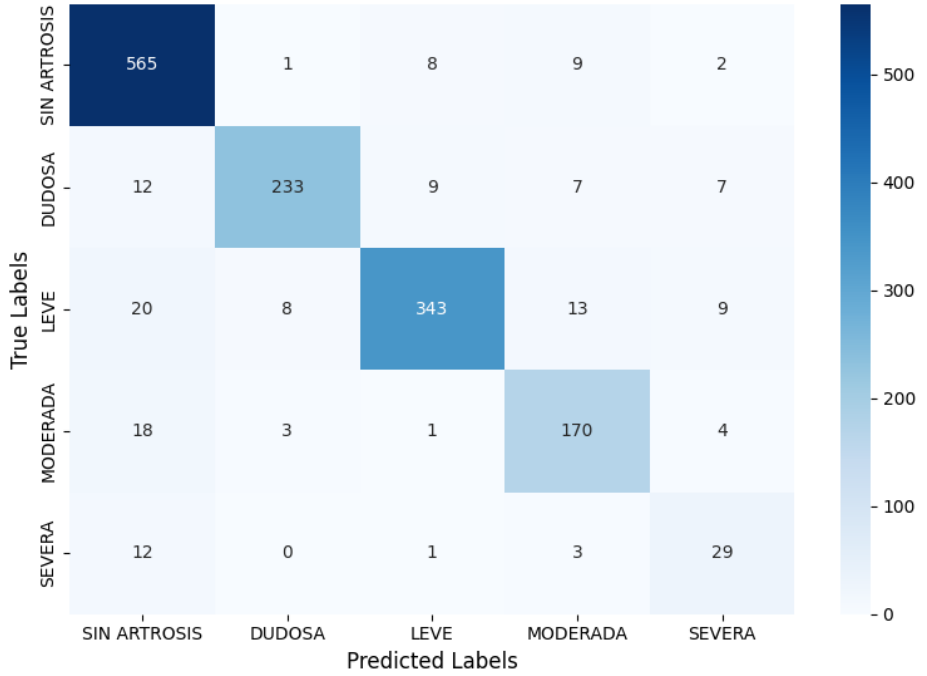
Para la categoría "MODERADA", el modelo logró un recall de 0.867, lo que refleja una alta sensibilidad en la detección de esta clase. La precisión de 0.842, aunque ligeramente inferior a las categorías anteriores, sigue siendo notable, lo que indica que la mayoría de las predicciones realizadas para esta clase fueron correctas. El F1-score de 0.854 muestra un desempeño sólido, aunque con un ligero margen de mejora en la especificidad.

La categoría "SEVERA", como en otros modelos, presentó las mayores dificultades. Con una precisión de 0.569 y un recall de 0.644, el modelo logró identificar correctamente el 64.4 % de las instancias reales, aunque con un nivel considerable de errores. El F1-score de 0.604 refleja

estas limitaciones, aunque sigue siendo superior a los resultados obtenidos para esta categoría con modelos previos.

Figura 11

Matriz de confusión para el modelo basado en ViT-S/16 en la detección de una patología por región anatómica



La matriz de confusión, presentada en la Figura 11, ofrece un análisis detallado de los aciertos y errores del modelo. En la categoría "SIN ARTROSIS", el modelo clasificó correctamente 565 de las 585 instancias reales, con errores mínimos hacia las categorías "DUDOSA" y "LEVE". Este resultado destaca la precisión del modelo para distinguir esta clase de las demás. En la categoría "DUDOSA", se identificaron correctamente 233 de las 268 instancias reales, mientras que los errores se distribuyeron principalmente hacia las categorías "SIN ARTROSIS" y "LEVE", lo que sugiere que estas clases pueden compartir ciertas características radiográficas.

En la categoría "LEVE", el modelo mostró un desempeño notable, con 343 de las 393 instancias clasificadas correctamente. Los errores se distribuyeron principalmente hacia "DUDOSA" y "MODERADA", reflejando cierta dificultad para diferenciar estas clases en algunos casos. Para la categoría "MODERADA", el modelo clasificó correctamente 170 de las 196 instancias reales, con un bajo nivel de confusiones hacia las categorías "LEVE" y "SIN ARTROSIS". Finalmente, la categoría "SEVERA" mostró un desempeño limitado, con 29 de las 45 instancias clasificadas correctamente. Los errores se distribuyeron principalmente hacia "MODERADA" y "LEVE", lo que indica que las características radiográficas de esta clase pueden solaparse con otras categorías.

En general, el modelo ViT-S/16 mostró un desempeño sobresaliente en la mayoría de las categorías, con métricas consistentemente altas que reflejan su capacidad para clasificar correctamente las condiciones mayoritarias y, en menor medida, las menos representadas. La capacidad del modelo para manejar datos complejos y diferenciar entre categorías similares se evidencia en sus resultados, especialmente en comparación con arquitecturas más tradicionales. Sin embargo, las dificultades persistentes en la clasificación de la categoría "SEVERA" destacan la necesidad de un enfoque más específico para manejar casos menos representados.

4.1.2.2. Métricas de evaluación del modelo basado en ViT-R26-S32

El modelo basado en la arquitectura ViT-R26-S32 fue evaluado utilizando las métricas de precisión, recall, F1-score y accuracy general,

complementadas con una matriz de confusión para proporcionar una visión más detallada del desempeño del modelo en la clasificación de las diferentes categorías relacionadas con la artrosis de rodilla. Los resultados reflejan un desempeño sólido y equilibrado en la mayoría de las clases, con limitaciones en la categoría "SEVERA".

El modelo logró un accuracy global de 88.2 %, lo que indica que más del 88 % de las predicciones realizadas fueron correctas. Las métricas ponderadas muestran una precisión promedio de 0.888, un recall promedio de 0.882 y un F1-score promedio de 0.884. Estos valores indican un modelo con alta capacidad para clasificar correctamente la mayoría de las instancias, mostrando especial fortaleza en las categorías más representadas.

En la Tabla 7, que se presenta a continuación, se muestran las métricas desglosadas por clase. En la categoría "SIN ARTROSIS", el modelo alcanzó una precisión de 0.966 y un recall de 0.875, lo que indica que el 87.5 % de las instancias reales fueron identificadas correctamente. El F1-score de 0.918 evidencia un desempeño robusto y consistente en esta clase, destacando la capacidad del modelo para diferenciar esta categoría de las demás.

Tabla 7

Métricas de evaluación del modelo basado en ViT-R26-S32 en la detección de una patología por región anatómica

	Precision	Recall	F1-score	Support
SIN ARTROSIS	0.966	0.875	0.918	585
DUDOSA	0.820	0.903	0.860	268
LEVE	0.888	0.929	0.908	393
MODERADA	0.826	0.847	0.836	196
SEVERA	0.540	0.600	0.568	45
Accuracy			0.882	1487
Macro avg	0.808	0.831	0.818	1487
Weighted avg	0.888	0.882	0.884	1487

Fuente: Elaboración propia

En la categoría "DUDOSA", el modelo mostró métricas consistentes y elevadas. Con una precisión de 0.820 y un recall de 0.903, el modelo identificó correctamente el 90.3 % de las instancias reales, aunque con cierto nivel de errores hacia otras categorías. El F1-score de 0.860 refleja un desempeño muy sólido y equilibrado para esta clase.

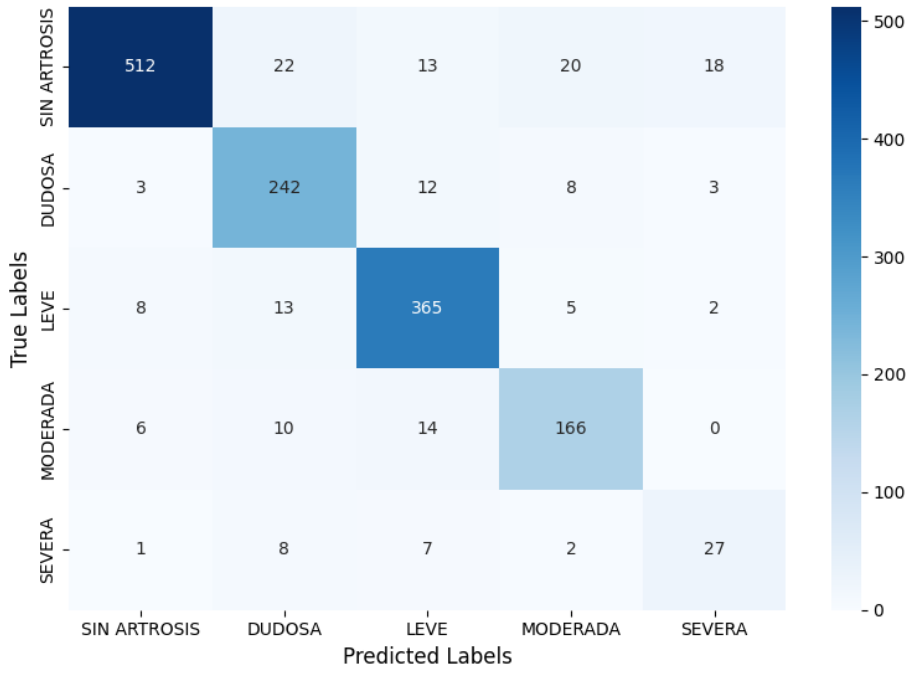
La categoría "LEVE" se destacó como una de las mejor clasificadas por el modelo. Con un recall de 0.929, el modelo logró identificar correctamente el 92.9 % de las instancias reales, mientras que la precisión de 0.888 refleja que la mayoría de las predicciones realizadas para esta clase fueron acertadas. El F1-score de 0.908 refuerza la percepción de un desempeño excepcional en esta categoría.

En la categoría "MODERADA", el modelo obtuvo métricas ligeramente inferiores a las de "LEVE" y "DUDOSA", pero aún consistentes. Con un recall de 0.847, el modelo fue capaz de identificar correctamente el 84.7 % de los casos reales, mientras que la precisión de 0.826 sugiere una proporción significativa de aciertos en las predicciones. El F1-score de 0.836 evidencia un buen equilibrio en esta categoría.

La categoría "SEVERA", como en otros modelos, presentó el mayor desafío. Con una precisión de 0.540 y un recall de 0.600, el modelo logró identificar correctamente el 60 % de las instancias reales de esta categoría, aunque con una cantidad considerable de errores. El F1-score de 0.568 muestra las limitaciones persistentes en la clasificación de esta clase, posiblemente debido a su baja representación en los datos.

Figura 12

Matriz de confusión para el modelo basado en ViT-R26-S32 en la detección de una patología por región anatómica



La matriz de confusión, presentada en la Figura 12, proporciona un análisis detallado del desempeño del modelo en cada clase. En la categoría "SIN ARTROSIS", el modelo clasificó correctamente 512 de las 585 instancias reales, con errores menores hacia las categorías "DUDOSA" y "LEVE", lo que indica un alto nivel de precisión para esta clase. En la categoría "DUDOSA", se identificaron correctamente 242 de las 268 instancias reales, con errores mínimos hacia las categorías "SIN ARTROSIS" y "LEVE", lo que refleja una capacidad sólida para diferenciar esta clase.

En la categoría "LEVE", el modelo mostró un desempeño sobresaliente, clasificando correctamente 365 de las 393 instancias reales. Los errores se distribuyeron principalmente hacia "DUDOSA" y "MODERADA", lo que sugiere que estas categorías pueden compartir características que dificultan su diferenciación. En la categoría "MODERADA", el modelo identificó correctamente 166 de las 196 instancias reales, con errores menores hacia "LEVE" y "SIN ARTROSIS". Finalmente, la categoría "SEVERA" mostró un desempeño limitado, con 27 de las 45 instancias clasificadas correctamente. Los errores se distribuyeron hacia "MODERADA" y "LEVE", lo que indica una dificultad persistente para identificar esta clase con precisión.

En general, el modelo ViT-R26-S32 mostró un desempeño consistente y sólido en las categorías mayoritarias, como "SIN ARTROSIS", "LEVE" y "DUDOSA". Aunque hubo limitaciones en la categoría "SEVERA", el modelo demostró una alta capacidad para manejar datos complejos y diferenciar entre la mayoría de las condiciones

radiográficas evaluadas. Estos resultados refuerzan la eficacia del modelo en la tarea de diagnóstico automatizado de artrosis.

4.1.2.3. Métricas de evaluación del modelo basado en ViT-B/32

El modelo basado en la arquitectura ViT-B/32 fue evaluado empleando las métricas de precisión, recall, F1-score y accuracy general, complementado con una matriz de confusión para un análisis detallado del desempeño en la clasificación de las distintas categorías de artrosis de rodilla. Los resultados obtenidos indican un desempeño global bueno, con valores consistentes en la mayoría de las categorías, aunque se observaron dificultades para las menos representadas, como "SEVERA".

El modelo alcanzó un accuracy global de 85.1 %, lo que refleja que más del 85 % de las predicciones realizadas fueron correctas. Las métricas ponderadas muestran una precisión promedio de 0.860, un recall promedio de 0.851 y un F1-score promedio de 0.854. Estas cifras destacan un modelo con capacidad confiable para clasificar correctamente la mayoría de las instancias evaluadas, con ligeras variaciones en el rendimiento entre las distintas categorías.

En la Tabla 8, presentada a continuación, se muestran las métricas desglosadas por categoría. La clase "SIN ARTROSIS" obtuvo una precisión de 0.889 y un recall de 0.894, lo que indica que el modelo identificó correctamente el 89.4 % de las instancias reales pertenecientes a esta categoría. El F1-score de 0.892 refuerza la consistencia del modelo en esta clase, mostrando un buen equilibrio entre precisión y sensibilidad.

Tabla 8

Métricas de evaluación del modelo basado en ViT-B/32 en la detección de una patología por región anatómica

	Precision	Recall	F1-score	Support
SIN ARTROSIS	0.889	0.894	0.892	585
DUDOSA	0.874	0.802	0.837	268
LEVE	0.904	0.835	0.868	393
MODERADA	0.764	0.878	0.817	196
SEVERA	0.431	0.622	0.509	45
Accuracy			0.851	1487
Macro avg	0.772	0.806	0.784	1487
Weighted avg	0.860	0.851	0.854	1487

Fuente: Elaboración propia

En la categoría "DUDOSA", el modelo alcanzó una precisión de 0.874 y un recall de 0.802, lo que implica que identificó correctamente el 80.2 % de las instancias reales pertenecientes a esta clase. Sin embargo, hubo un nivel significativo de confusiones hacia categorías cercanas, como "SIN ARTROSIS". El F1-score de 0.837 refleja un buen desempeño general, aunque con margen de mejora en la sensibilidad.

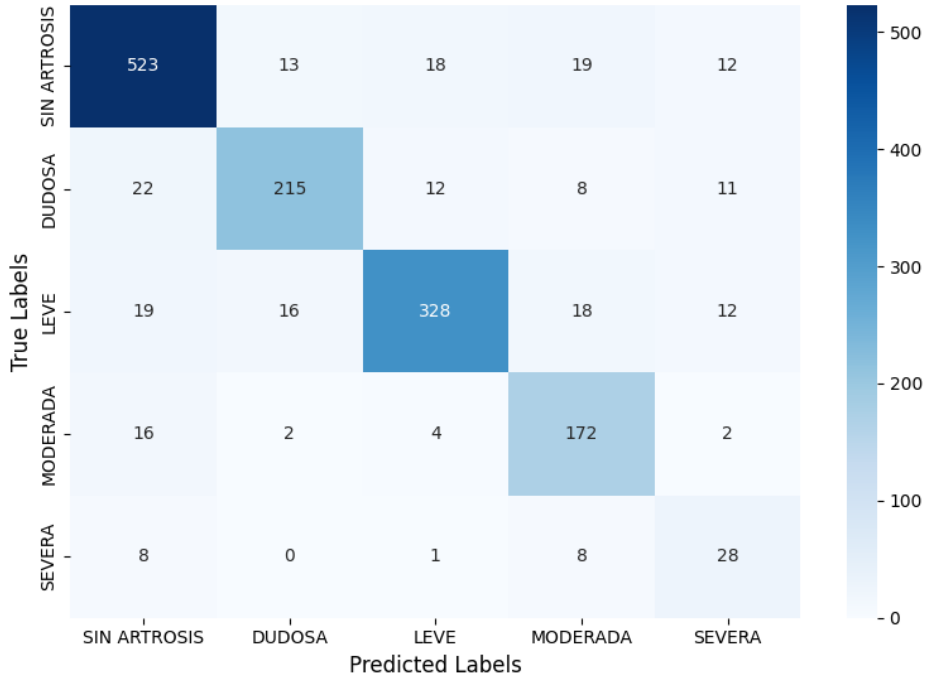
La clase "LEVE" mostró un desempeño consistente y elevado, destacándose como una de las mejor clasificadas. Con un recall de 0.835, el modelo identificó correctamente el 83.5 % de las instancias reales, mientras que la precisión de 0.904 indica un alto nivel de confianza en las predicciones realizadas para esta clase. El F1-score de 0.868 refuerza la percepción de un desempeño sólido y confiable en esta categoría.

En cuanto a la categoría "MODERADA", el modelo alcanzó métricas destacables, con un recall de 0.878, lo que indica que el 87.8 % de las instancias reales fueron identificadas correctamente. La precisión de 0.764 muestra que algunas predicciones fueron incorrectas, principalmente hacia categorías cercanas. El F1-score de 0.817 evidencia un desempeño adecuado, con una alta sensibilidad para identificar esta clase.

La categoría "SEVERA", como en otros modelos, presentó mayores dificultades. Con una precisión de 0.431 y un recall de 0.622, el modelo logró identificar correctamente el 62.2 % de las instancias reales, aunque con una cantidad significativa de errores. El F1-score de 0.509 refleja las limitaciones persistentes en la clasificación de esta clase, posiblemente influenciadas por su baja representación en los datos.

Figura 13

Matriz de confusión para el modelo basado en ViT-B/32 en la detección de una patología por región anatómica



La matriz de confusión, presentada en la Figura 13, proporciona un análisis detallado del desempeño del modelo. En la categoría "SIN ARTROSIS", el modelo clasificó correctamente 523 de las 585 instancias reales, con errores menores hacia las categorías "DUDOSA" y "LEVE". Esto evidencia un buen nivel de precisión para esta clase. En la categoría "DUDOSA", se identificaron correctamente 215 de las 268 instancias reales, aunque se observaron errores hacia "SIN ARTROSIS" y "LEVE", lo que sugiere características compartidas entre estas clases.

En la categoría "LEVE", el modelo clasificó correctamente 328 de las 393 instancias reales, con errores distribuidos principalmente hacia "DUDOSA" y "MODERADA". Este comportamiento refuerza la idea de que estas categorías pueden presentar patrones radiográficos similares que

dificultan la diferenciación. Para la clase "MODERADA", el modelo clasificó correctamente 172 de las 196 instancias reales, con un bajo nivel de confusiones hacia "LEVE". Finalmente, la categoría "SEVERA" mostró un desempeño limitado, con solo 28 de las 45 instancias clasificadas correctamente. Los errores se distribuyeron hacia "MODERADA" y "LEVE", reflejando una dificultad para identificar esta clase con precisión.

En aspectos genéricos, el modelo ViT-B/32 mostró un desempeño confiable en la clasificación de las categorías más representadas, como "SIN ARTROSIS", "DUDOSA" y "LEVE". Aunque la categoría "SEVERA" continúa presentando desafíos significativos, el modelo demostró una capacidad sólida para manejar datos complejos y diferenciar entre las categorías evaluadas, reflejando un rendimiento consistente y adecuado para esta tarea de diagnóstico automatizado.

4.1.2.4. Métricas de evaluación del modelo basado en ViT-R50-L32

El modelo basado en la arquitectura ViT-R50-L32 fue evaluado utilizando las métricas de precisión, recall, F1-score y accuracy general, complementadas con una matriz de confusión para analizar detalladamente su desempeño en la clasificación de las diferentes categorías de artrosis de rodilla. Los resultados reflejan un desempeño global moderado, con fortalezas en categorías más representadas y limitaciones significativas en las menos representadas, como "SEVERA".

El modelo alcanzó un accuracy global de 78.0 %, lo que significa que el 78 % de las predicciones realizadas fueron correctas. Las métricas

ponderadas muestran una precisión promedio de 0.786, un recall promedio de 0.780 y un F1-score promedio de 0.777, lo que evidencia un desempeño aceptable, aunque con variaciones notables entre las distintas clases.

En la Tabla 9, presentada a continuación, se observan las métricas desglosadas por clase. En la categoría "SIN ARTROSIS", el modelo obtuvo una precisión de 0.780 y un recall de 0.926, lo que indica que identificó correctamente el 92.6 % de las instancias reales pertenecientes a esta clase. El F1-score de 0.847 refleja un buen desempeño general, aunque con margen de mejora en la especificidad de las predicciones.

Tabla 9

Métricas de evaluación del modelo basado en ViT-R50-L32 en la detección de una patología por región anatómica

	Precision	**Recall**	**F1-score**	**Support**
SIN ARTROSIS	0.780	0.926	0.847	585
DUDOSA	0.860	0.664	0.749	268
LEVE	0.853	0.723	0.782	393
MODERADA	0.671	0.719	0.695	196
SEVERA	0.357	0.333	0.345	45
Accuracy	0.780	1487		
Macro avg	0.704	0.673	0.684	1487
Weighted avg	0.786	0.780	0.777	1487

Fuente: Elaboración propia

En la categoría "DUDOSA", el modelo mostró un desempeño moderado, con una precisión de 0.860 y un recall de 0.664. Esto indica que, aunque el modelo fue confiable en las predicciones realizadas para

esta clase, no logró identificar correctamente un porcentaje considerable de las instancias reales. El F1-score de 0.749 refleja estas limitaciones, con un balance adecuado pero no óptimo entre precisión y sensibilidad.

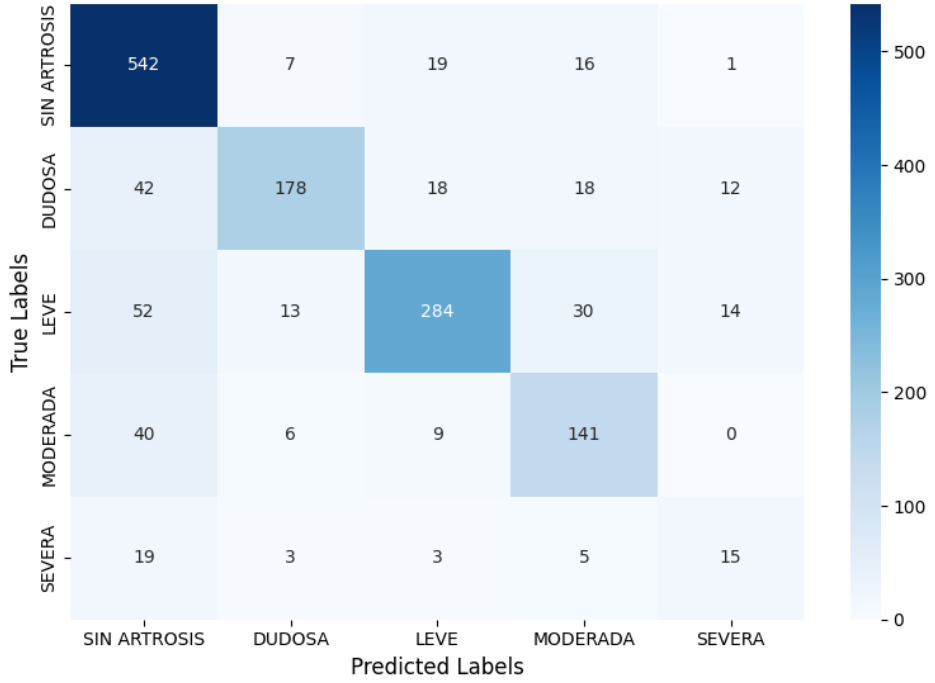
La categoría "LEVE" presentó métricas consistentes, con un recall de 0.723, lo que significa que el modelo identificó correctamente el 72.3 % de las instancias reales. La precisión de 0.853 indica que la mayoría de las predicciones realizadas para esta clase fueron correctas. El F1-score de 0.782 evidencia un desempeño equilibrado, aunque con margen para mejorar en la sensibilidad de la detección.

En cuanto a la categoría "MODERADA", el modelo obtuvo un recall de 0.719, lo que refleja que el 71.9 % de las instancias reales fueron correctamente identificadas. La precisión fue de 0.671, lo que indica que un porcentaje significativo de las predicciones realizadas fueron incorrectas, principalmente hacia categorías cercanas como "LEVE". El F1-score de 0.695 evidencia un desempeño moderado, limitado por confusiones entre clases similares.

La categoría "SEVERA" fue nuevamente la que presentó mayores desafíos. Con una precisión de 0.357 y un recall de 0.333, el modelo identificó correctamente solo el 33.3 % de las instancias reales de esta categoría. El F1-score de 0.345 refleja estas limitaciones, destacando la dificultad persistente para clasificar correctamente esta clase, probablemente debido a su baja representación en el conjunto de datos.

Figura 14

Matriz de confusión para el modelo basado en ViT-R50-L32 ResNet50 en la detección de una patología por región anatómica



La matriz de confusión, presentada en la Figura 14, ofrece un análisis detallado del comportamiento del modelo. En la categoría "SIN ARTROSIS", el modelo clasificó correctamente 542 de las 585 instancias reales, con errores distribuidos principalmente hacia "DUDOSA" y "LEVE". En la categoría "DUDOSA", el modelo identificó correctamente 178 de las 268 instancias reales, aunque se observaron errores hacia "SIN ARTROSIS" y "LEVE", reflejando características compartidas entre estas clases.

En la categoría "LEVE", el modelo clasificó correctamente 284 de las 393 instancias reales, aunque los errores estuvieron principalmente relacionados con las clases "DUDOSA" y "MODERADA". Esto sugiere que estas categorías pueden presentar patrones radiográficos similares que

dificultan la diferenciación. En la categoría "MODERADA", el modelo clasificó correctamente 141 de las 196 instancias reales, con errores distribuidos hacia "LEVE". Finalmente, en la categoría "SEVERA", el modelo identificó correctamente solo 15 de las 45 instancias reales, con errores significativos hacia "MODERADA" y "LEVE".

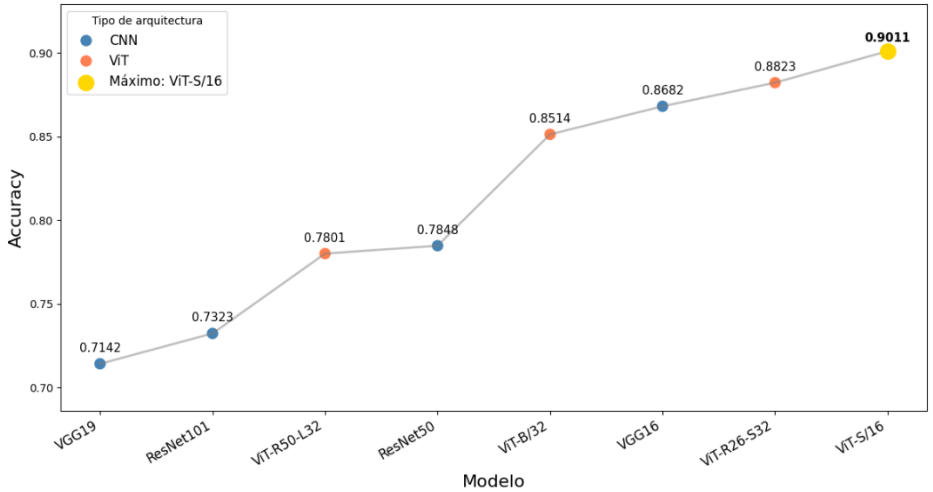
Se puede apreciar que, el modelo ViT-R50-L32 mostró un desempeño adecuado en las categorías más representadas, como "SIN ARTROSIS" y "LEVE", mientras que las categorías menos comunes, como "SEVERA", continúan siendo un desafío. Estas métricas y patrones resaltan la capacidad del modelo para clasificar datos complejos, aunque con limitaciones en la sensibilidad y especificidad de algunas clases.

4.1.3. Comparación de las métricas de evaluación de ambas arquitecturas

4.1.3.1. Comparativa del desempeño en términos de accuracy

Figura 15

Comparación de la métrica accuracy entre arquitecturas CNN y Vision Transformers para la detección de patología única



El análisis comparativo de la métrica accuracy entre las arquitecturas CNN y Vision Transformers revela patrones distintivos en el rendimiento de ambos enfoques para la detección de patología única en imágenes radiográficas. Las arquitecturas CNN evaluadas presentan valores de accuracy que oscilan entre 0.7142 y 0.8682, donde el modelo VGG16 alcanzó el máximo rendimiento con un valor de 0.8682, seguido por ResNet50 con 0.7848, ResNet101 con 0.7323, y VGG19 con 0.7142.

En el caso de las arquitecturas Vision Transformers, los resultados experimentales muestran un rango de accuracy entre 0.7801 y 0.9011. El modelo ViT-S/16 registró el valor más alto de accuracy (0.9011) entre todas las arquitecturas evaluadas, superando tanto a sus contrapartes ViT como a los modelos CNN. Los demás modelos ViT exhibieron valores de accuracy de 0.8823 (ViT-R26-S32), 0.8514 (ViT-B/32), y 0.7801 (ViT-RS0-L32).

La diferencia en accuracy entre los mejores modelos de cada arquitectura (ViT-S/16 y VGG16) es de 3.29 puntos porcentuales, lo cual representa una mejora significativa en la capacidad de clasificación. Este hallazgo es particularmente relevante considerando que la métrica accuracy refleja la proporción total de predicciones correctas realizadas por el modelo, tanto para casos positivos como negativos.

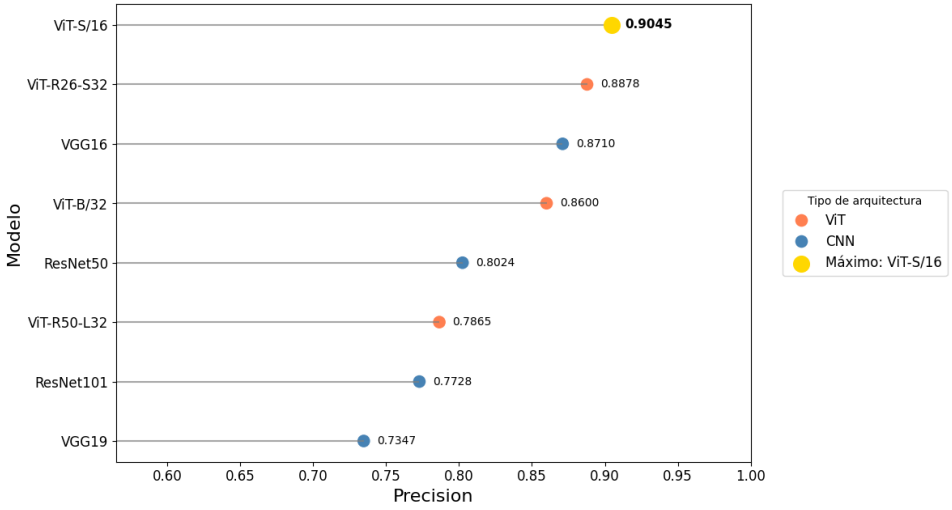
Los resultados obtenidos demuestran que ambas arquitecturas mantienen niveles de accuracy superiores a 0.70, estableciendo una base sólida para la detección automatizada de patología única en imágenes radiográficas. Sin embargo, la arquitectura Vision Transformer,

específicamente el modelo ViT-S/16, exhibe una superioridad cuantitativa en términos de accuracy, alcanzando un valor de 0.9011, lo que indica una mayor capacidad para la clasificación correcta de casos en el contexto específico de este estudio.

4.1.3.2. Comparativa de la precisión (precision)

Figura 16

Comparación de la métrica precision entre arquitecturas CNN y Vision Transformers para la detección de patología única



En continuidad con el análisis de métricas de rendimiento, los resultados de la métrica de precision proporcionan información adicional sobre la capacidad de los modelos para minimizar los falsos positivos. Las arquitecturas CNN exhiben valores de precision que fluctúan entre 0.7347 y 0.8710, donde el modelo VGG16 alcanza el valor más alto (0.8710), seguido por ResNet50 (0.8024), ResNet101 (0.7728), y VGG19 (0.7347).

Por otra parte, las arquitecturas Vision Transformers presentan un rango de precision entre 0.7865 y 0.9045. El modelo ViT-S/16 destaca

nuevamente al registrar el valor más elevado de precision (0.9045) entre todas las arquitecturas evaluadas. En orden descendente, los demás modelos ViT obtuvieron valores de 0.8878 (ViT-R26-S32), 0.8600 (ViT-B/32), y 0.7865 (ViT-RS0-L32).

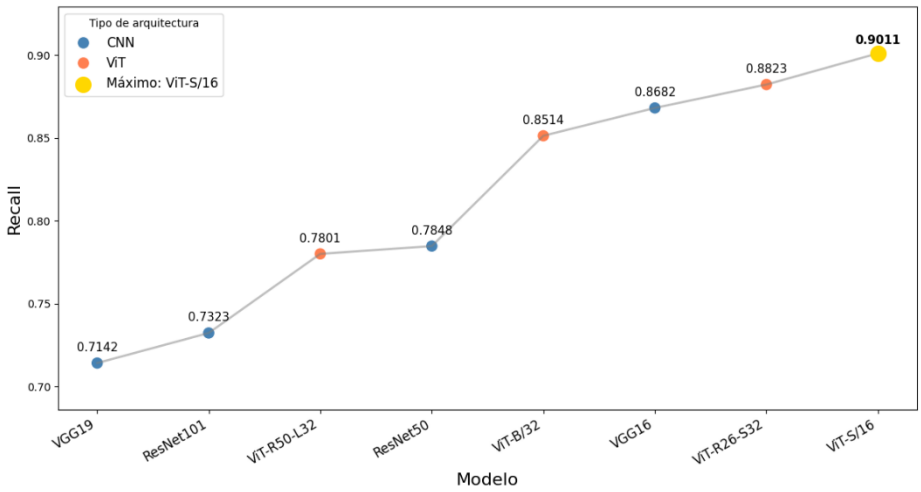
Es particularmente notable la diferencia de 3.35 puntos porcentuales en precision entre los modelos más destacados de cada arquitectura (ViT-S/16 y VGG16). Este resultado cobra especial relevancia considerando que la precision refleja la proporción de predicciones positivas correctas entre todas las predicciones positivas realizadas, lo cual es crucial en el contexto del diagnóstico médico donde la minimización de falsos positivos es esencial.

Asimismo, resulta significativo que ambas familias de arquitecturas mantienen niveles de precision superiores a 0.73, lo cual indica una capacidad robusta para evitar falsos positivos en la detección de patología única. No obstante, la arquitectura Vision Transformer, y específicamente el modelo ViT-S/16, demuestra una superioridad cuantitativa al alcanzar una precision de 0.9045, evidenciando una mayor capacidad para realizar predicciones positivas precisas en el contexto específico de la detección de patologías en imágenes radiográficas.

4.1.3.3. Comparativa del recall

Figura 17

Comparación de la métrica recall entre arquitecturas CNN y Vision Transformers para la detección de patología única



El análisis de la métrica recall revela patrones significativos en la capacidad de los modelos para identificar correctamente los casos positivos. Las arquitecturas CNN exhiben valores que oscilan entre 0.7142 y 0.8682, donde el modelo VGG16 alcanza el valor más alto (0.8682), seguido por ResNet50 (0.7848), ResNet101 (0.7323), y VGG19 (0.7142).

En cuanto a las arquitecturas Vision Transformers, los resultados muestran un rango de recall entre 0.7801 y 0.9011. Específicamente, el modelo ViT-S/16 registra el valor más alto (0.9011) entre todas las arquitecturas evaluadas, seguido por ViT-R26-S32 (0.8823), ViT-B/32 (0.8514), y ViT-R50-L32 (0.7801).

La diferencia en recall entre los mejores modelos de cada arquitectura (ViT-S/16 y VGG16) es de 3.29 puntos porcentuales, lo cual resulta particularmente relevante considerando que el recall refleja la

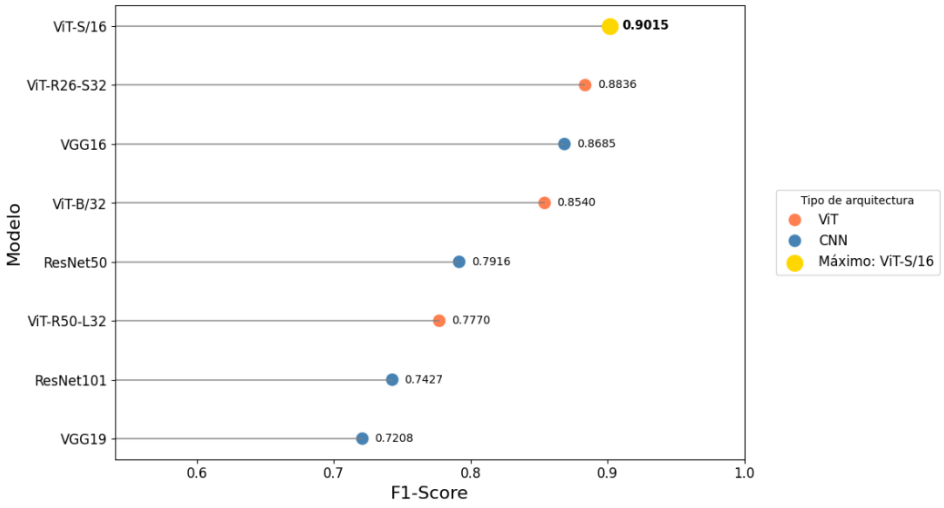
capacidad del modelo para identificar correctamente los casos positivos, aspecto crucial en el contexto del diagnóstico médico automatizado.

Resulta significativo que ambas familias de arquitecturas mantienen niveles de recall superiores a 0.71, lo cual indica una capacidad robusta para la detección de casos positivos. Sin embargo, la arquitectura Vision Transformer, específicamente el modelo ViT-S/16, demuestra una superioridad cuantitativa al alcanzar un recall de 0.9011, evidenciando una mayor sensibilidad en la detección de patologías en imágenes radiográficas.

4.1.3.4. Comparativa del F1-score

Figura 18

Comparación de la métrica F1-score entre arquitecturas CNN y Vision Transformers para la detección de patología única



Los resultados del F1-Score proporcionan una visión equilibrada del rendimiento de los modelos al combinar precision y recall. En el ámbito de las arquitecturas CNN, los valores fluctúan entre 0.7208 y

0.8685, donde el modelo VGG16 alcanza el máximo valor (0.8685), seguido por ResNet50 (0.7916), ResNet101 (0.7427), y VGG19 (0.7208).

Por otra parte, las arquitecturas Vision Transformers presentan valores de F1-Score que oscilan entre 0.7770 y 0.9015. El modelo ViT-S/16 destaca nuevamente al registrar el valor más alto (0.9015) entre todas las arquitecturas evaluadas, seguido por ViT-R26-S32 (0.8836), ViT-B/32 (0.8540), y ViT-R50-L32 (0.7770).

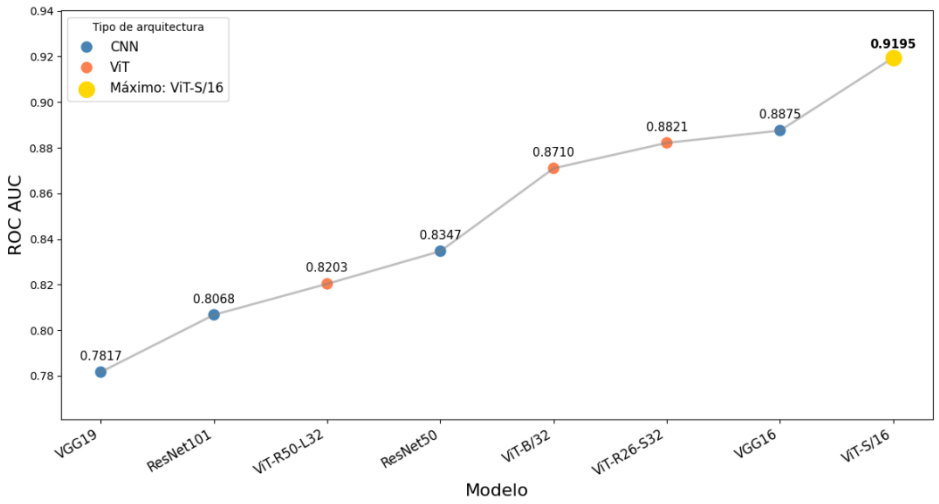
La diferencia de 3.30 puntos porcentuales en F1-Score entre los mejores exponentes de cada arquitectura resulta particularmente significativa, considerando que esta métrica penaliza los desequilibrios entre precision y recall, lo cual es fundamental en el contexto del diagnóstico médico donde se busca un balance óptimo entre ambas métricas.

Es importante destacar que ambas familias de arquitecturas mantienen valores de F1-Score superiores a 0.72, demostrando un rendimiento balanceado en la tarea de clasificación. No obstante, la arquitectura Vision Transformer, y específicamente el modelo ViT-S/16, exhibe una clara ventaja al mantener un F1-Score superior a 0.90, indicando un mejor equilibrio entre la precisión y la sensibilidad en sus predicciones.

4.1.3.5. Comparativa del ROC AUC

Figura 19

Comparación de la métrica ROC AUC entre arquitecturas CNN y Vision Transformers para la detección de patología única



El análisis del área bajo la curva ROC (ROC AUC) revela información crucial sobre la capacidad discriminativa general de los modelos evaluados. En el contexto de las arquitecturas CNN, los valores varían entre 0.7817 y 0.8875, donde el modelo VGG16 alcanza el valor más alto (0.8875), seguido por ResNet50 (0.8347), ResNet101 (0.8068), y VGG19 (0.7817).

En el caso de las arquitecturas Vision Transformers, los resultados muestran valores de ROC AUC entre 0.8203 y 0.9195. El modelo ViT-S/16 sobresale con el valor más alto (0.9195), seguido por ViT-R26-S32 (0.8821), ViT-B/32 (0.8710), y ViT-R50-L32 (0.8203).

Notablemente, ambas familias de arquitecturas mantienen valores de ROC AUC superiores a 0.78, demostrando una sólida capacidad discriminativa. Sin embargo, la arquitectura Vision Transformer, particularmente el modelo ViT-S/16, exhibe una superioridad cuantitativa

al alcanzar un ROC AUC de 0.9195, lo que sugiere una mayor robustez y fiabilidad en su capacidad de discriminación diagnóstica bajo diferentes umbrales de decisión.

4.1.4. Comparación por pares de los modelos con complejidad similar entre ambas arquitecturas mediante validación cruzada

4.1.4.1. Media y desviación estándar del accuracy en la validación cruzada

Tabla 10

Media y desviación estándar del accuracy en las arquitecturas CNN y Vision Transformers para la detección de patología única

Arquitectura	Modelo	Accuracy Medio	Desviación Estándar
CNN	VGG16	0.871658	0.020648
	VGG19	0.730749	0.019842
	ResNet50	0.774103	0.012874
	ResNet101	0.720880	0.016871
ViT	ViT-S/16	0.913159	0.014427
	ViT-R26-S32	0.864786	0.020422
	ViT-B/32	0.836052	0.017193
	ViT-R50-L32	0.759822	0.022083

Fuente: Elaboración propia

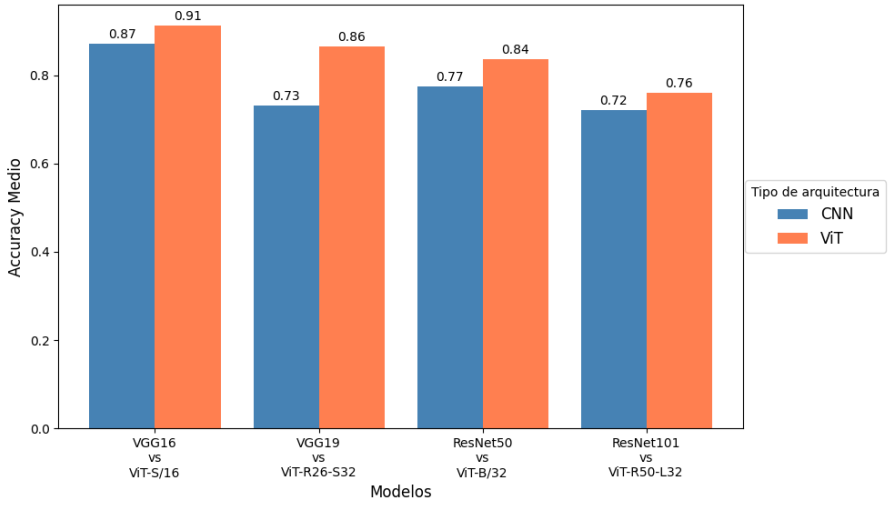
El análisis de los resultados obtenidos mediante validación cruzada revela patrones distintivos en la estabilidad y consistencia de las arquitecturas evaluadas. En el contexto de las CNN, se observa que el modelo VGG16 alcanza el accuracy medio más alto con un valor de 0.8717 y una desviación estándar de 0.0206, lo que indica una variabilidad moderada en sus predicciones. Por su parte, ResNet50 muestra un

comportamiento más estable con un accuracy medio de 0.7741 y la menor desviación estándar entre las CNN (0.0129), mientras que VGG19 y ResNet101 registran valores de 0.7307 ± 0.0198 y 0.7209 ± 0.0169 respectivamente, evidenciando una consistencia comparable en sus predicciones.

En lo que respecta a los Vision Transformers, los resultados denotan un rendimiento notablemente superior, donde el modelo ViT-S/16 destaca significativamente al alcanzar un accuracy medio de 0.9132 con una desviación estándar de 0.0144. Esta combinación de alto rendimiento y baja variabilidad sugiere una robustez sobresaliente en sus predicciones. Asimismo, los modelos ViT-R26-S32 (0.8648 ± 0.0204), ViT-B/32 (0.8361 ± 0.0172) y ViT-R50-L32 (0.7598 ± 0.0221) mantienen un patrón de rendimiento consistente, aunque con variaciones en sus niveles de estabilidad.

Figura 20

Comparación por pares del accuracy medio entre arquitecturas CNN y Vision Transformers para la detección de patología única



El análisis comparativo por pares entre modelos de complejidad similar revela diferencias sistemáticas en el rendimiento. La comparación VGG16 vs ViT-S/16 muestra una diferencia de 4 puntos porcentuales (0.87 vs 0.91) a favor del modelo Vision Transformer. De manera similar, el par VGG19 vs ViT-R26-S32 exhibe una diferencia aún más pronunciada de 13 puntos porcentuales (0.73 vs 0.86), evidenciando una superioridad significativa del modelo ViT.

En las comparaciones de arquitecturas basadas en ResNet, se observa que el par ResNet50 vs ViT-B/32 muestra una diferencia de 7 puntos porcentuales (0.77 vs 0.84) a favor del Vision Transformer. Esta tendencia se mantiene en el par ResNet101 vs ViT-R50-L32, donde se registra una diferencia de 4 puntos porcentuales (0.72 vs 0.76), aunque con una menor magnitud que en las comparaciones anteriores.

Un aspecto particularmente notable es que en todas las comparaciones por pares, los modelos Vision Transformer mantienen consistentemente un rendimiento superior a sus contrapartes CNN de similar complejidad. Esta superioridad sistemática, que oscila entre 4 y 13 puntos porcentuales, sugiere una ventaja inherente de la arquitectura Vision Transformer en la tarea de detección de patología única, independientemente del nivel de complejidad del modelo evaluado. Además, es importante destacar que la magnitud de esta diferencia se mantiene significativa incluso en los modelos de mayor complejidad, lo que refuerza la robustez de estos hallazgos.

4.1.4.2. Análisis inferencial para determinar diferencias significativas entre los modelos

Para validar la significancia estadística de las diferencias observadas en el rendimiento, se realizó un análisis inferencial sistemático. Previamente a la aplicación de las pruebas de hipótesis, se verificaron los supuestos fundamentales de normalidad y homocedasticidad.

Tabla 11

Supuestos de normalidad y homocedasticidad para la comparación entre modelos CNN y Vision Transformers en el caso de detección de patología única

CNN vs. ViT	Shapiro-Wilk CNN (p-value)	Shapiro-Wilk ViT (p-value)	Levene (p-value)
VGG16 vs. ViT-S/16	0.24739701	0.30043003	0.62486609
VGG19 vs. ViT-R26-S32	0.46125621	0.57637817	0.1706921
ResNet50 vs. ViT-B/32	0.64568216	0.80375868	0.40336032
ResNet101 vs. ViT-R50-L32	0.59800005	0.50294578	0.42554222

Fuente: Elaboración propia

El análisis de los supuestos estadísticos demuestra que los datos cumplen con los criterios necesarios para la aplicación de pruebas paramétricas. La prueba de Shapiro-Wilk para la normalidad en la comparación VGG16 vs. ViT-S/16 muestra valores p de 0.247 y 0.300 respectivamente, indicando que no hay evidencia para rechazar la normalidad en ambas distribuciones. Este patrón se mantiene consistente

en las demás comparaciones, con valores p que oscilan entre 0.461 y 0.804, sustancialmente superiores al nivel de significancia de 0.05. Paralelamente, la prueba de Levene para la homocedasticidad revela valores p entre 0.171 y 0.625, confirmando la homogeneidad de varianzas en todos los pares comparados. El cumplimiento robusto de ambos supuestos justifica la aplicación de la prueba t de Student para el análisis comparativo.

Tabla 12

Resultados de la prueba t de Student para la comparación entre modelos CNN y Vision Transformers en el caso de detección de patología única

CNN vs. ViT	T-Statistic	P-Value (one-tailed)
VGG16 vs. ViT-S/16	3.68413215	0.00309112
VGG19 vs. ViT-R26-S32	10.5259845	2.89E-06
ResNet50 vs. ViT-B/32	6.4492492	9.92E-05
ResNet101 vs. ViT-R50-L32	3.13338082	0.00697195

Fuente: Elaboración propia

Para evaluar formalmente la superioridad de los Vision Transformers, se establecieron las siguientes hipótesis estadísticas:

- H_0: $\mu ViT \leq \mu CNN$ (El accuracy medio de los Vision Transformers no es mayor que el de las CNN)
- H_1: $\mu ViT > \mu CNN$ (El accuracy medio de los Vision Transformers es mayor que el de las CNN)

Los resultados de la prueba t de Student unilateral revelan evidencia contundente para rechazar la hipótesis nula en todas las

comparaciones. La comparación VGG19 vs. ViT-R26-S32 muestra la evidencia más fuerte ($t = 10.526$, $p = 2.89\times10^{-6}$), indicando una diferencia altamente significativa en el rendimiento. Similar robustez estadística se observa en la comparación ResNet50 vs. ViT-B/32 ($t = 6.449$, $p = 9.92\times10^{-5}$), seguida por VGG16 vs. ViT-S/16 ($t = 3.684$, $p = 0.003$) y ResNet101 vs. ViT-R50-L32 ($t = 3.133$, $p = 0.007$).

La magnitud de los estadísticos t y sus correspondientes valores p, todos significativamente menores al nivel $\alpha = 0.05$, proporcionan evidencia estadística sólida para concluir que los Vision Transformers exhiben un rendimiento sistemáticamente superior a las CNN en todas las comparaciones realizadas. Estos resultados no solo validan las diferencias observadas en el análisis descriptivo previo, sino que también establecen con rigor estadístico la superioridad consistente de los Vision Transformers en la tarea de detección de patología única.

4.2. RESULTADOS DEL DESEMPEÑO PREDICTIVO DE LAS CNN Y LOS VIT EN LA DETECCIÓN DE MULTIPLES PATOLOGÍAS POR REGIÓN ANATÓMICA

4.2.1. Desempeño predictivo de las redes neuronales convolucionales (CNN)

4.2.1.1. Métricas de evaluación del modelo basado en VGG16

El desempeño del modelo basado en la arquitectura VGG16 fue evaluado para la detección de múltiples patologías en imágenes radiográficas, específicamente en las categorías "Normal", "Tuberculosis" y "Neumonía". Los resultados revelan una capacidad robusta para

identificar correctamente las condiciones mayoritarias, aunque se observaron desafíos en la clasificación de las patologías menos representadas, como la "Tuberculosis".

El modelo alcanzó un accuracy global del 88.0 %, indicando que el 88 % de las predicciones realizadas fueron correctas. Las métricas ponderadas muestran una precisión promedio de 0.890, un recall promedio de 0.880 y un F1-score promedio de 0.882, lo que demuestra una notable capacidad para manejar las diferentes categorías. Sin embargo, el análisis detallado por clase evidencia variaciones en el desempeño que merecen un estudio más profundo.

En la Tabla 13, se resumen las métricas específicas de cada categoría. Para la clase "Normal", el modelo obtuvo una precisión de 0.959, lo que indica que la mayoría de las predicciones realizadas para esta categoría fueron correctas. El recall fue de 0.853, lo que refleja que el modelo identificó correctamente el 85.3 % de las instancias reales de esta clase. El F1-score de 0.903 muestra un desempeño sobresaliente en esta categoría, con un buen equilibrio entre precisión y sensibilidad.

Tabla 13

Métricas de evaluación del modelo basado en VGG16 en la detección de múltiples patologías por región anatómica

	Precision	Recall	F1-score	Support
Normal	0.959	0.853	0.903	1017
Tuberculosis	0.651	0.800	0.718	140
Neumonía	0.846	0.926	0.884	855
Accuracy			0.880	2012
Macro avg	0.819	0.860	0.835	2012
Weighted avg	0.890	0.880	0.882	2012

Fuente: Elaboración propia

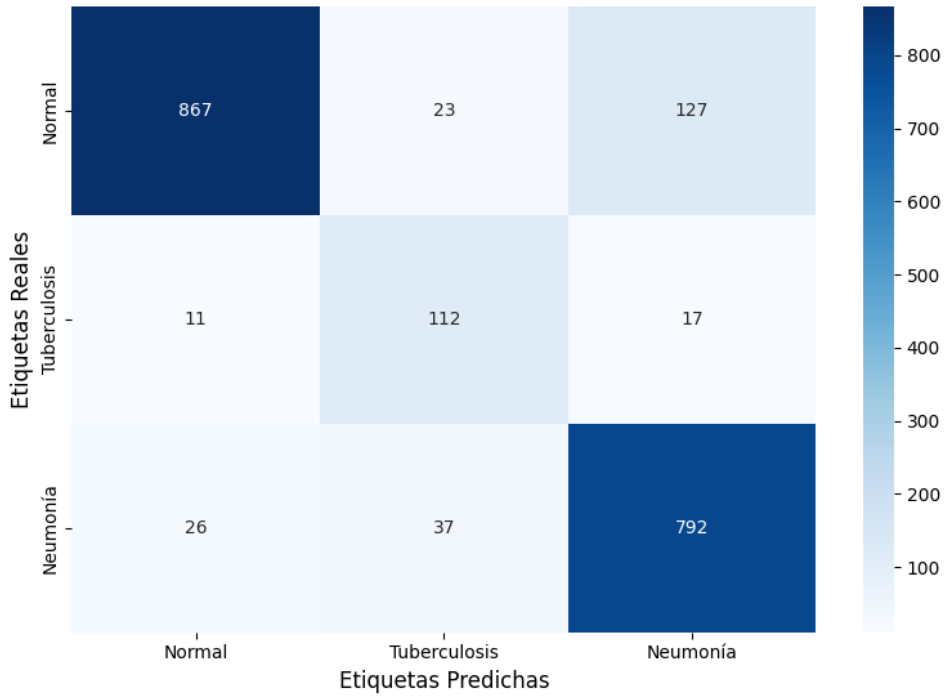
En cuanto a la categoría "Tuberculosis", el modelo mostró un desempeño moderado. La precisión alcanzada fue de 0.651, lo que indica que el 65.1 % de las predicciones realizadas para esta clase fueron correctas. Sin embargo, el recall de 0.800 señala que el modelo fue capaz de identificar correctamente el 80 % de los casos reales, aunque con un número significativo de confusiones hacia otras categorías, en particular "Normal". El F1-score de 0.718 refleja estas limitaciones, mostrando que aunque el modelo tiene una sensibilidad razonable, su capacidad para diferenciar esta clase de las demás no es completamente confiable.

Para la categoría "Neumonía", el modelo mostró un desempeño sólido y consistente. La precisión de 0.846 indica que la mayoría de las predicciones realizadas para esta clase fueron correctas, mientras que el recall de 0.926 destaca la alta sensibilidad del modelo al identificar esta condición. El F1-score de 0.884 evidencia un equilibrio satisfactorio entre

precisión y sensibilidad, destacando a esta clase como una de las mejor clasificadas por el modelo.

Figura 21

Matriz de confusión para el modelo basado en VGG16 en la detección de múltiples patologías por región anatómica



La Figura 21, que muestra la matriz de confusión, ofrece un análisis visual detallado de los aciertos y errores del modelo. En la categoría "Normal", el modelo clasificó correctamente 867 de las 1017 instancias reales. Sin embargo, se observan 127 confusiones hacia la clase "Neumonía".

En la categoría "Tuberculosis", el modelo identificó correctamente 112 de las 140 instancias reales, con 23 confusiones hacia "Normal" y 17 hacia "Neumonía". Para la categoría "Neumonía", se clasificaron correctamente 792 de las 855 instancias reales, con un bajo nivel de

confusiones hacia las categorías "Normal" (26 casos) y "Tuberculosis" (37 casos), lo que refuerza el sólido desempeño del modelo en esta clase.

Entonces se observó que, el modelo basado en VGG16 demostró un desempeño destacable en la clasificación de las categorías "Normal" y "Neumonía", con métricas que evidencian un alto nivel de precisión y sensibilidad. Sin embargo, la categoría "Tuberculosis" presentó mayores desafíos, con un número considerable de errores de clasificación hacia las otras dos clases. Estos resultados subrayan tanto las fortalezas del modelo en la detección de patologías mayoritarias como sus limitaciones en el manejo de categorías menos representadas. La matriz de confusión permite identificar patrones específicos de error, lo que ofrece valiosas oportunidades para futuras mejoras en el modelo.

4.2.1.2. Métricas de evaluación del modelo basado en VGG19

El modelo basado en la arquitectura VGG19 fue evaluado para la detección de múltiples patologías en imágenes radiográficas, específicamente en las categorías "Normal", "Tuberculosis" y "Neumonía". Los resultados obtenidos indican un desempeño moderado, con fortalezas en las categorías mayoritarias, como "Normal" y "Neumonía", y limitaciones significativas en la detección de "Tuberculosis".

El modelo alcanzó un accuracy global de 75.9 %, lo que significa que el 75.9 % de las predicciones realizadas por el modelo fueron correctas. Además, las métricas ponderadas reflejan una precisión promedio de 0.797, un recall promedio de 0.759 y un F1-score promedio

de 0.766, lo que sugiere que el modelo tiene una capacidad razonable para realizar predicciones, aunque no exenta de errores, especialmente en clases menos representadas.

En la Tabla 14, se presentan los resultados por categoría. Para la clase "Normal", el modelo alcanzó una precisión de 0.903, lo que significa que la mayoría de las predicciones realizadas para esta clase fueron correctas. Sin embargo, el recall de 0.690 muestra que el modelo solo pudo identificar el 69 % de las instancias reales de esta categoría. El F1-score de 0.783 refleja un desempeño sólido, aunque limitado por una sensibilidad menor a la esperada.

Tabla 14

Métricas de evaluación del modelo basado en VGG19 en la detección de múltiples patologías por región anatómica

	Precision	**Recall**	**F1-score**	**Support**
Normal	0.903	0.690	0.783	1017
Tuberculosis	0.368	0.636	0.466	140
Neumonía	0.741	0.861	0.797	855
Accuracy			0.759	2012
Macro avg	0.671	0.729	0.682	2012
Weighted avg	0.797	0.759	0.766	2012

Fuente: Elaboración propia

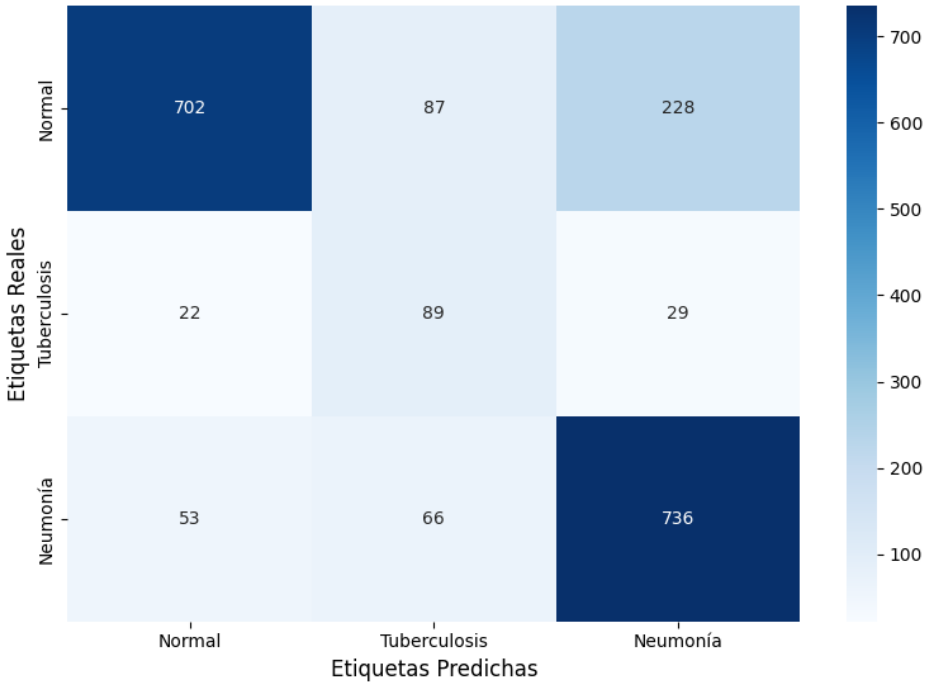
En la categoría "Tuberculosis", el modelo mostró una precisión de 0.368, lo que significa que solo una pequeña proporción de las predicciones realizadas para esta clase fueron correctas. El recall, aunque más alto (0.636), indica que el modelo logró identificar el 63.6 % de los

casos reales. Sin embargo, el F1-score de 0.466 refleja las limitaciones generales del modelo en esta categoría, con un nivel considerable de confusiones hacia otras clases, especialmente "Normal".

La categoría "Neumonía" destacó como una de las mejor clasificadas por el modelo. Con una precisión de 0.741 y un recall de 0.861, el modelo demostró una alta sensibilidad para identificar los casos reales, aunque con errores presentes. El F1-score de 0.797 evidencia un desempeño equilibrado y sólido en esta categoría, posicionándola como una de las más confiables del modelo.

Figura 22

Matriz de confusión para el modelo basado en VGG19 en la detección de múltiples patologías por región anatómica



La matriz de confusión, presentada en la Figura 22, proporciona un análisis más detallado de los aciertos y errores del modelo. En la categoría

"Normal", se observa que de las 1017 instancias reales, 702 fueron correctamente clasificadas. Sin embargo, el modelo confundió 228 instancias con la categoría "Neumonía" y 87 con "Tuberculosis".

En la categoría "Tuberculosis", el modelo clasificó correctamente 89 de las 140 instancias reales, mientras que confundió 87 casos con "Normal" y 29 con "Neumonía". Estos errores reflejan las dificultades del modelo para distinguir claramente esta categoría de las demás, lo que puede estar influido por la limitada cantidad de datos de soporte para esta clase. En la categoría "Neumonía", se clasificaron correctamente 736 de las 855 instancias reales, con 66 confusiones hacia "Tuberculosis" y 53 hacia "Normal".

En términos generales, el modelo basado en VGG19 mostró un desempeño adecuado en las categorías "Normal" y "Neumonía", con métricas que destacan su capacidad para identificar correctamente la mayoría de los casos en estas clases. Sin embargo, la categoría "Tuberculosis" presentó mayores desafíos, con errores significativos hacia las otras clases. La interpretación de la matriz de confusión proporciona una visión detallada de los puntos fuertes y débiles del modelo, ofreciendo una base sólida para comprender su desempeño.

4.2.1.3. Métricas de evaluación del modelo basado en ResNet50

La arquitectura ResNet50 se utilizó para analizar su capacidad de clasificación en tres condiciones principales: "Normal", "Tuberculosis" y "Neumonía". Los resultados revelan un desempeño aceptable en términos generales, con una alta precisión en las clases más representadas y ciertas

limitaciones al clasificar patologías menos frecuentes, como "Tuberculosis". Este análisis proporciona una visión integral del rendimiento del modelo, destacando tanto sus fortalezas como sus áreas de mejora.

El modelo alcanzó un accuracy global de 81.9 %, indicando que más del 80 % de las predicciones realizadas por el modelo fueron correctas. Las métricas promedio ponderadas reportaron una precisión de 0.841, un recall de 0.819 y un F1-score de 0.823, lo que sugiere un rendimiento estable en las tareas de clasificación, aunque no exento de errores significativos en algunas categorías.

En la Tabla 15, se presentan las métricas específicas por categoría. Para la clase "Normal", el modelo alcanzó una precisión de 0.931, destacando su capacidad para realizar predicciones confiables en esta categoría. Sin embargo, el recall de 0.766 muestra que el modelo identificó correctamente solo el 76.6 % de los casos reales de esta clase, lo que llevó a un F1-score de 0.840, un resultado sólido pero con margen para mejorar en la sensibilidad.

Tabla 15

Métricas de evaluación del modelo basado en ResNet50 en la detección de múltiples patologías por región anatómica

	Precision	Recall	F1-score	Support
Normal	0.931	0.766	0.840	1017
Tuberculosis	0.505	0.757	0.606	140
Neumonía	0.790	0.891	0.837	855
Accuracy			0.819	2012
Macro avg	0.742	0.805	0.761	2012
Weighted avg	0.841	0.819	0.823	2012

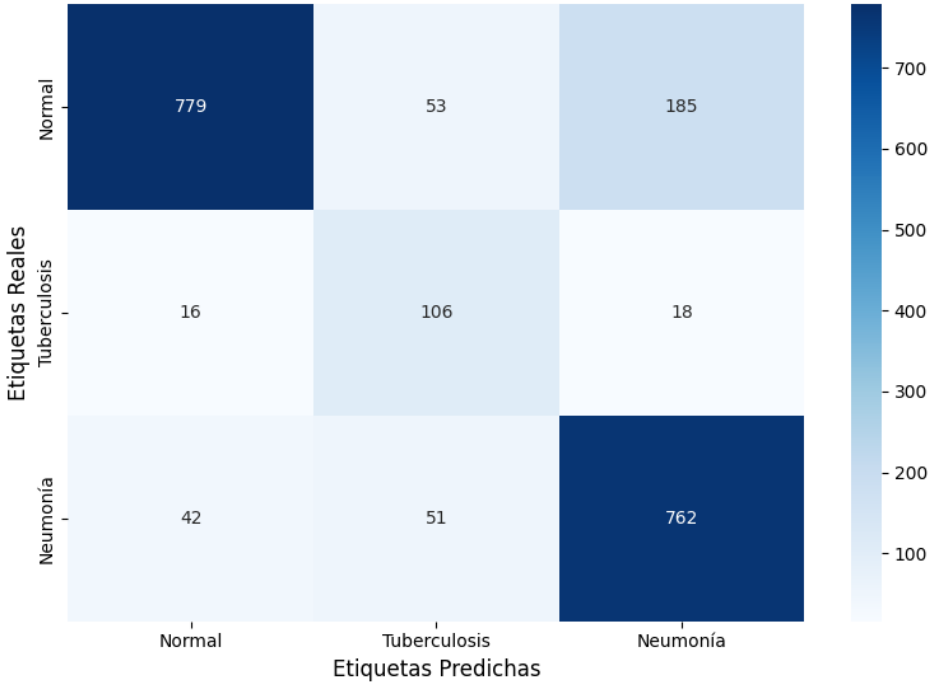
Fuente: Elaboración propia

En cuanto a la categoría "Tuberculosis", el modelo mostró un desempeño moderado, con una precisión de 0.505 y un recall de 0.757, lo que significa que logró identificar correctamente el 75.7 % de los casos reales, aunque la mitad de las predicciones realizadas para esta clase fueron erróneas. El F1-score de 0.606 resalta estas limitaciones, sugiriendo dificultades para distinguir esta patología de otras condiciones radiográficas.

Por otro lado, la categoría "Neumonía" obtuvo métricas notablemente positivas. Con una precisión de 0.790 y un recall de 0.891, el modelo identificó correctamente casi el 90 % de los casos reales de esta clase. El F1-score de 0.837 evidencia un equilibrio sólido entre precisión y sensibilidad, destacando a esta categoría como una de las mejor clasificadas.

Figura 23

Matriz de confusión para el modelo basado en ResNet50 en la detección de múltiples patologías por región anatómica



La matriz de confusión, presentada en la Figura 23, permite examinar de manera más detallada los aciertos y errores del modelo. Para la categoría "Normal", el modelo clasificó correctamente 779 de las 1017 instancias reales, aunque 185 casos fueron confundidos con "Neumonía" y 53 con "Tuberculosis".

En la categoría "Tuberculosis", el modelo identificó correctamente 106 de las 140 instancias reales, mientras que los errores se distribuyeron principalmente hacia "Neumonía" (18 casos) y "Normal" (16 casos). Esto resalta las dificultades para clasificar esta patología, lo que puede estar relacionado con la limitada representación de esta clase en los datos de entrenamiento. Por su parte, en la categoría "Neumonía", el modelo

clasificó correctamente 762 de las 855 instancias reales, aunque 42 casos fueron confundidos con "Normal" y 51 con "Tuberculosis".

Sintetizando estos resultados, el modelo ResNet50 demostró un buen desempeño en la clasificación de "Neumonía" y "Normal", mientras que la categoría "Tuberculosis" presentó mayores desafíos. Estos resultados destacan la solidez del modelo en tareas de clasificación de patologías mayoritarias, aunque también subrayan las limitaciones inherentes a la clasificación de condiciones con menor representación en los datos. La matriz de confusión proporciona una visión detallada de los errores comunes, permitiendo identificar áreas específicas para posibles mejoras en el rendimiento del modelo.

4.2.1.4. Métricas de evaluación del modelo basado en ResNet101

El modelo basado en la arquitectura ResNet101 fue evaluado para clasificar las categorías "Normal", "Tuberculosis" y "Neumonía", alcanzando un accuracy global del 78.8 %. Este valor refleja que, en promedio, casi el 79 % de las predicciones realizadas por el modelo coincidieron con las etiquetas reales. Las métricas ponderadas proporcionan mayor detalle: una precisión de 0.818, un recall de 0.788 y un F1-score de 0.795, lo cual indica que el modelo tiene un desempeño global aceptable, aunque presenta variaciones significativas según la clase evaluada.

En la Tabla 16, se presentan las métricas detalladas por categoría. En la clase "Normal", el modelo obtuvo una precisión de 0.906, lo que significa que la mayoría de las predicciones realizadas para esta categoría

fueron correctas. Sin embargo, el recall de 0.739 muestra que el modelo identificó correctamente el 73.9 % de las instancias reales de esta clase. Este desequilibrio entre precisión y sensibilidad resultó en un F1-score de 0.814, lo cual denota un desempeño adecuado pero con limitaciones en la capacidad del modelo para identificar todos los casos reales.

Tabla 16

Métricas de evaluación del modelo basado en ResNet101 en la detección de múltiples patologías por región anatómica

	Precision	**Recall**	**F1-score**	**Support**
Normal	0.906	0.739	0.814	1017
Tuberculosis	0.395	0.671	0.497	140
Neumonía	0.783	0.864	0.822	855
Accuracy			0.788	2012
Macro avg	0.695	0.758	0.711	2012
Weighted avg	0.818	0.788	0.795	2012

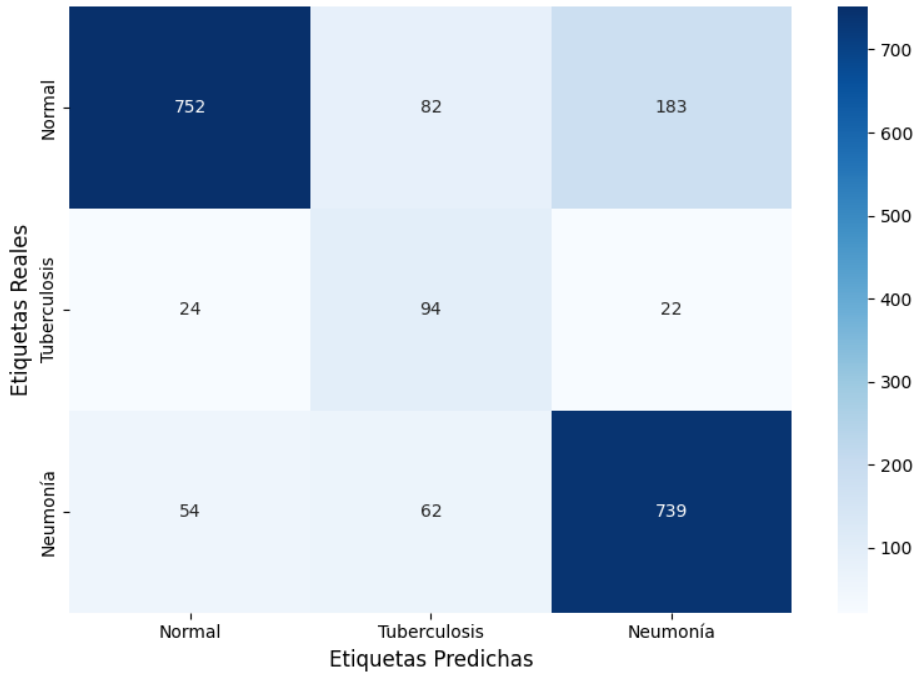
Fuente: Elaboración propia

La categoría "Tuberculosis" presentó un desempeño inferior comparado con las otras clases. El modelo logró una precisión de 0.395, lo que significa que menos del 40 % de las predicciones realizadas para esta clase fueron correctas. Sin embargo, el recall de 0.671 indica que el modelo fue capaz de identificar correctamente el 67.1 % de los casos reales de "Tuberculosis". Este contraste entre precisión y sensibilidad se refleja en un F1-score de 0.497, evidenciando una menor consistencia en la clasificación de esta categoría.

En cuanto a la categoría "Neumonía", el modelo mostró un desempeño sólido, con una precisión de 0.783 y un recall de 0.864, lo que indica que identificó correctamente el 86.4 % de las instancias reales. El F1-score de 0.822 pone de manifiesto un equilibrio satisfactorio entre precisión y sensibilidad, lo que posiciona a esta clase como una de las mejor clasificadas por el modelo.

Figura 24

Matriz de confusión para el modelo basado en ResNet101 en la detección de múltiples patologías por región anatómica



La matriz de confusión, presentada en la Figura 24, proporciona una representación visual de los aciertos y errores cometidos por el modelo. En la categoría "Normal", de las 1017 instancias reales, 752 fueron clasificadas correctamente, mientras que 183 casos se confundieron con "Neumonía" y 82 con "Tuberculosis". Estas confusiones indican que

algunas características radiográficas de estas condiciones pueden solaparse, dificultando la diferenciación precisa entre las clases.

En la clase "Tuberculosis", el modelo clasificó correctamente 94 de las 140 instancias reales. Sin embargo, se observaron 24 confusiones hacia la clase "Normal" y 22 hacia "Neumonía", lo que evidencia dificultades específicas para distinguir esta categoría. Por otro lado, en la clase "Neumonía", el modelo identificó correctamente 739 de las 855 instancias reales, mientras que 54 se clasificaron erróneamente como "Normal" y 62 como "Tuberculosis". Estos resultados reflejan que la mayor parte de los errores en esta categoría se distribuyen entre las otras dos, posiblemente debido a la similitud en ciertas características de las imágenes.

En palabras sencillas, el modelo ResNet101 mostró un desempeño adecuado en las categorías "Normal" y "Neumonía", con métricas que indican una buena capacidad para clasificar estas clases en términos generales. Por el contrario, la categoría "Tuberculosis" presentó un desempeño más limitado, con métricas que reflejan dificultades tanto en la precisión como en la sensibilidad. Estos resultados destacan la variabilidad en el desempeño del modelo dependiendo de la categoría, lo cual se evidencia claramente en la matriz de confusión, que detalla los patrones específicos de error en la clasificación.

4.2.2. Desempeño predictivo de los Vision Transformers (ViT)

4.2.2.1. Métricas de evaluación del modelo basado en ViT-S/16

El modelo ViT-S/16 mostró un desempeño destacado al clasificar las categorías "Normal", "Tuberculosis" y "Neumonía", alcanzando un accuracy global del 91.5 %. Esto significa que, en promedio, más del 91 % de las predicciones realizadas por el modelo fueron correctas. Adicionalmente, las métricas ponderadas indican una precisión promedio de 0.920, un recall promedio de 0.915 y un F1-score promedio de 0.916. Estos valores reflejan un desempeño global robusto, con un buen equilibrio entre precisión y sensibilidad en las tareas de clasificación.

En la Tabla 17, se presentan los valores específicos para cada categoría. La clase "Normal" obtuvo una precisión de 0.973, lo que evidencia un alto nivel de confianza en las predicciones realizadas para esta categoría. El recall de 0.897 indica que el modelo identificó correctamente el 89.7 % de las instancias reales, mientras que el F1-score de 0.933 pone de manifiesto un desempeño notablemente equilibrado en esta clase.

Tabla 17

Métricas de evaluación del modelo basado en ViT-S/16 en la detección de múltiples patologías por región anatómica

	Precision	Recall	F1-score	Support
Normal	0.973	0.897	0.933	1017
Tuberculosis	0.724	0.879	0.794	140
Neumonía	0.890	0.942	0.915	855
Accuracy			0.915	2012
Macro avg	0.862	0.906	0.881	2012
Weighted avg	0.920	0.915	0.916	2012

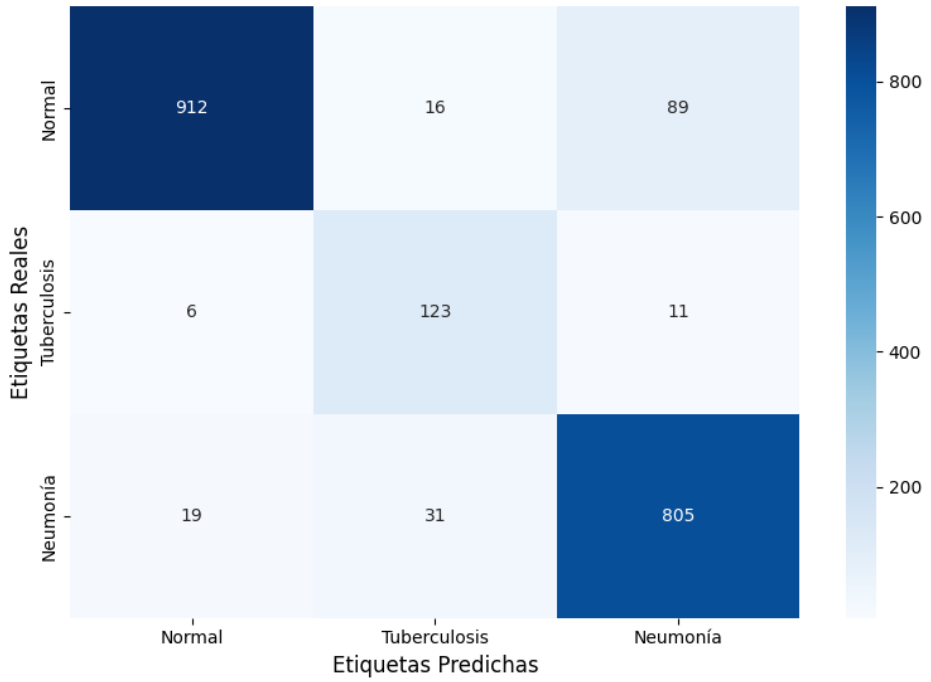
Fuente: Elaboración propia

En cuanto a la categoría "Tuberculosis", el modelo mostró un desempeño sólido, con una precisión de 0.724. Este valor indica que aproximadamente el 72.4 % de las predicciones realizadas para esta clase fueron correctas. El recall alcanzó un valor elevado de 0.879, evidenciando que el modelo fue capaz de identificar correctamente el 87.9 % de los casos reales. El F1-score de 0.794 refleja una buena capacidad general en esta categoría, aunque con cierta tendencia a confusiones menores con las otras clases.

Por su parte, la clase "Neumonía" también se destacó con métricas elevadas. La precisión de 0.890 indica un alto nivel de confianza en las predicciones, mientras que el recall de 0.942 muestra que el modelo identificó correctamente casi el 94.2 % de las instancias reales. El F1-score de 0.915 pone de relieve un desempeño equilibrado, posicionando a esta categoría como una de las mejor clasificadas por el modelo.

Figura 25

Matriz de confusión para el modelo basado en ViT-S/16en la detección de múltiples patologías por región anatómica



La matriz de confusión, presentada en la Figura 25, proporciona un análisis más detallado de los aciertos y errores del modelo. En la categoría "Normal", de las 1017 instancias reales, 912 fueron clasificadas correctamente, aunque 89 fueron confundidas con "Neumonía" y 16 con "Tuberculosis". Estas confusiones reflejan una alta precisión general, pero también destacan áreas específicas donde las características de las imágenes pueden haberse solapado con las de otras clases.

En la clase "Tuberculosis", el modelo clasificó correctamente 123 de las 140 instancias reales. Sin embargo, 11 casos fueron asignados erróneamente a "Neumonía" y 6 a "Normal", lo que indica un bajo nivel de confusión hacia otras clases. Finalmente, en la categoría "Neumonía", se observó que el modelo identificó correctamente 805 de las 855

instancias reales, con 31 confusiones hacia "Tuberculosis" y 19 hacia "Normal". Este patrón refuerza la alta capacidad del modelo para diferenciar esta condición de las otras dos.

En aspectos generales, el modelo ViT-S/16 demostró un desempeño sobresaliente en las tres categorías evaluadas, logrando un balance eficiente entre precisión y sensibilidad. Las métricas obtenidas reflejan un modelo confiable y robusto en términos de clasificación, mientras que la matriz de confusión ilustra los principales patrones de error, destacando áreas específicas donde el modelo podría haber encontrado mayores desafíos en la diferenciación de las condiciones radiográficas.

4.2.2.2. Métricas de evaluación del modelo basado en ViT-R26-S32

El modelo ViT-R26-S32 fue evaluado para la clasificación de las categorías "Normal", "Tuberculosis" y "Neumonía", alcanzando un accuracy global de 85.2 %. Este resultado refleja que el modelo logró clasificar correctamente más del 85 % de las instancias evaluadas. Además, los promedios ponderados de las métricas clave indican una precisión de 0.867, un recall de 0.852 y un F1-score de 0.855, lo que evidencia un desempeño adecuado y consistente en las diferentes clases evaluadas.

En la Tabla 18, se resumen las métricas específicas de cada categoría. La clase "Normal" obtuvo una precisión notable de 0.949, lo que indica que las predicciones realizadas para esta categoría fueron en su mayoría correctas. El recall, con un valor de 0.807, sugiere que el modelo

identificó correctamente el 80.7 % de las instancias reales de esta clase. El F1-score de 0.872 refleja un balance satisfactorio entre precisión y sensibilidad, consolidando un desempeño sólido en esta categoría.

Tabla 18

Métricas de evaluación del modelo basado en ViT-R26-S32 en la detección de múltiples patologías por región anatómica

	Precision	Recall	F1-score	Support
Normal	0.949	0.807	0.872	1017
Tuberculosis	0.617	0.807	0.700	140
Neumonía	0.810	0.913	0.859	855
Accuracy			0.852	2012
Macro avg	0.792	0.843	0.810	2012
Weighted avg	0.867	0.852	0.855	2012

Fuente: Elaboración propia

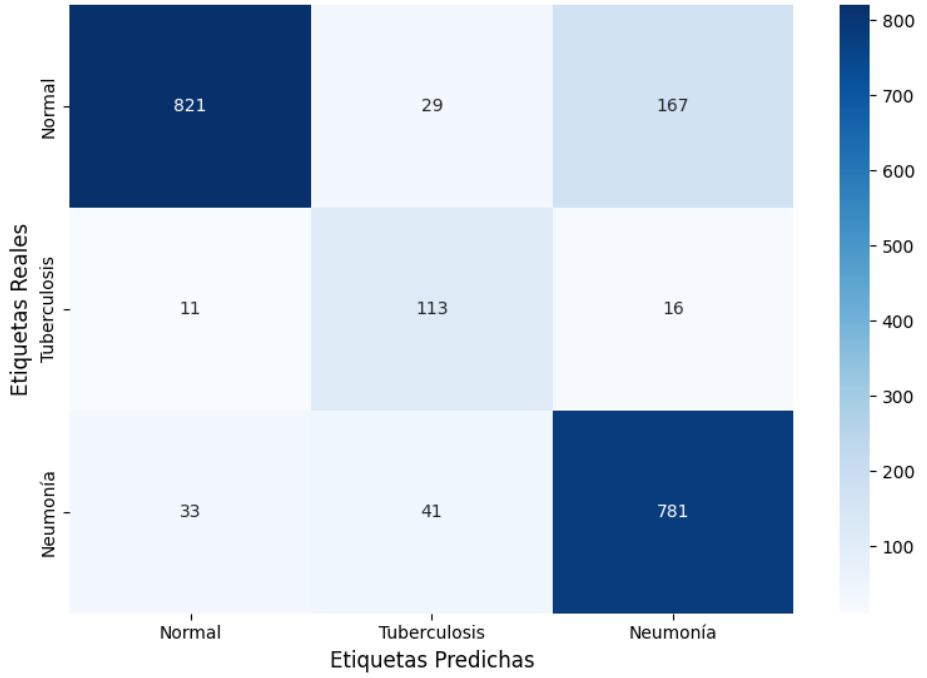
En la categoría "Tuberculosis", el modelo mostró un desempeño moderado. Con una precisión de 0.617, el modelo logró realizar predicciones correctas para esta clase en poco más del 61 % de los casos. Sin embargo, el recall de 0.807 destaca que el modelo identificó correctamente el 80.7 % de las instancias reales, lo que representa una sensibilidad considerablemente alta para una categoría con menor soporte en el conjunto de datos. El F1-score de 0.700 refleja estas características, posicionando a esta clase con un rendimiento aceptable dentro de las limitaciones inherentes a su bajo soporte.

Por su parte, la categoría "Neumonía" obtuvo métricas sólidas. La precisión de 0.810 indica que el modelo realizó predicciones confiables

para esta clase, mientras que el recall de 0.913 resalta la sensibilidad del modelo para identificar correctamente el 91.3 % de las instancias reales. El F1-score de 0.859 confirma un desempeño robusto y equilibrado en esta categoría, consolidando a "Neumonía" como una de las clases mejor clasificadas.

Figura 26

Matriz de confusión para el modelo basado en ViT-R26-S32 en la detección de múltiples patologías por región anatómica



La Figura 26, que presenta la matriz de confusión, permite visualizar los aciertos y errores cometidos por el modelo en cada categoría. Para la clase "Normal", el modelo clasificó correctamente 821 de las 1017 instancias reales, mientras que 167 casos fueron confundidos con "Neumonía" y 29 con "Tuberculosis". Estas confusiones pueden deberse a similitudes radiográficas entre las características de estas condiciones, que afectan la diferenciación precisa.

En la clase "Tuberculosis", el modelo identificó correctamente 113 de las 140 instancias reales. Sin embargo, se observaron 16 confusiones hacia "Neumonía" y 11 hacia "Normal". Este comportamiento sugiere que aunque el modelo tiene un nivel aceptable de sensibilidad en esta categoría, la precisión podría verse afectada por la presencia de patrones compartidos entre las clases. En cuanto a la categoría "Neumonía", se clasificaron correctamente 781 de las 855 instancias reales. No obstante, 41 casos fueron asignados a "Tuberculosis" y 33 a "Normal", lo que reafirma que el modelo tiene una sensibilidad alta, pero no completamente exenta de errores.

Esto evidencia que, el modelo ViT-R26-S32 demostró un desempeño global satisfactorio, con métricas consistentes que destacan especialmente en las clases "Normal" y "Neumonía". Aunque la categoría "Tuberculosis" presentó ciertas limitaciones en precisión, el elevado valor de recall evidencia una capacidad adecuada para identificar esta condición. La matriz de confusión detalla los errores más frecuentes, proporcionando una visión clara del comportamiento del modelo en la clasificación de cada categoría.

4.2.2.3. Métricas de evaluación del modelo basado en ViT-B/32

El modelo ViT-B/32 fue analizado en su capacidad para clasificar las categorías "Normal", "Tuberculosis" y "Neumonía", mostrando un desempeño global adecuado. El accuracy general alcanzado fue del 85.5 %, lo que implica que más del 85 % de las predicciones realizadas por el modelo fueron correctas. Las métricas ponderadas presentaron valores

consistentes: una precisión de 0.866, un recall de 0.855 y un F1-score de 0.857. Estos resultados reflejan una capacidad general favorable del modelo para distinguir entre las diferentes clases.

En la Tabla 19, se presentan las métricas específicas por clase. Para la categoría "Normal", el modelo obtuvo una precisión de 0.937, destacando una alta confianza en las predicciones realizadas. El recall fue de 0.818, lo que indica que el modelo identificó correctamente el 81.8 % de las instancias reales. El F1-score de 0.873 confirma un desempeño sólido en esta clase, aunque con cierta disminución en la sensibilidad.

Tabla 19

Métricas de evaluación del modelo basado en ViT-B/32 en la detección de múltiples patologías por región anatómica

	Precision	Recall	F1-score	Support
Normal	0.937	0.818	0.873	1017
Tuberculosis	0.632	0.771	0.695	140
Neumonía	0.820	0.913	0.864	855
Accuracy	0.855	2012		
Macro avg	0.796	0.834	0.811	2012
Weighted avg	0.866	0.855	0.857	2012

Fuente: Elaboración propia

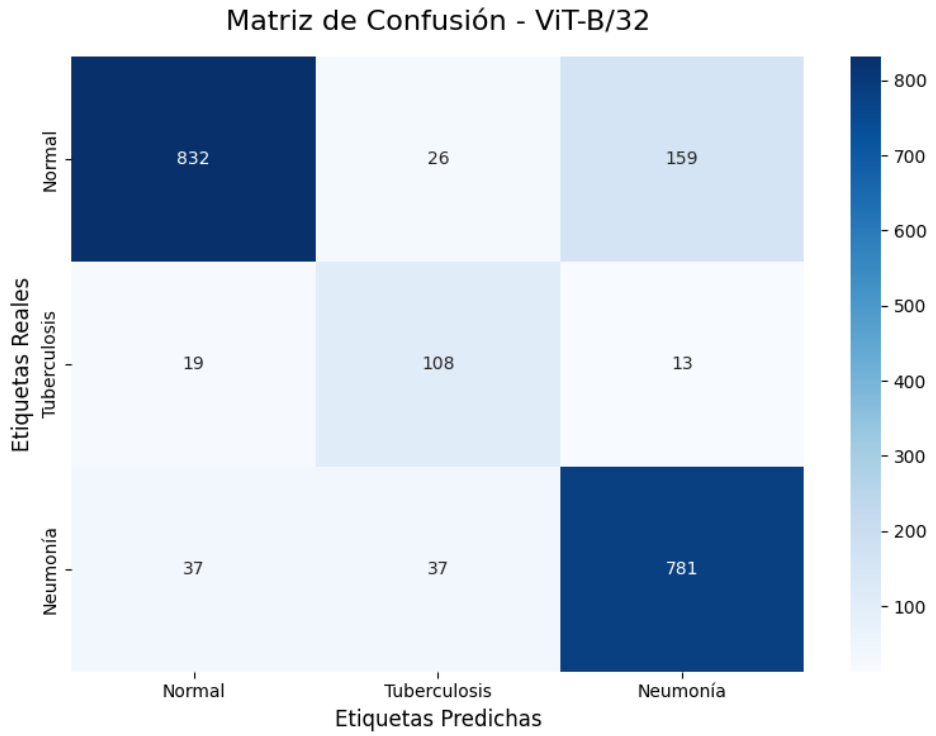
En la categoría "Tuberculosis", el modelo mostró un desempeño moderado. Con una precisión de 0.632, aproximadamente el 63.2 % de las predicciones realizadas para esta clase fueron correctas. Por otro lado, el recall alcanzó un valor de 0.771, indicando que el modelo identificó correctamente el 77.1 % de las instancias reales. El F1-score de 0.695

evidencia un desempeño equilibrado, aunque con espacio para mejorar en la capacidad de precisión al clasificar esta categoría.

En cuanto a la categoría "Neumonía", el modelo destacó con métricas elevadas. La precisión de 0.820 muestra un buen nivel de confianza en las predicciones realizadas, mientras que el recall de 0.913 refleja que el modelo identificó correctamente el 91.3 % de las instancias reales. El F1-score de 0.864 refuerza un desempeño robusto y balanceado en esta clase, posicionándola como una de las mejor manejadas.

Figura 27

Matriz de confusión para el modelo basado en ViT-B/32 en la detección de múltiples patologías por región anatómica



La Figura 27, que presenta la matriz de confusión, ofrece un análisis detallado de los aciertos y errores del modelo. Para la clase "Normal", de las 1017 instancias reales, 832 fueron clasificadas

correctamente. Sin embargo, 159 fueron confundidas con "Neumonía" y 26 con "Tuberculosis". Este patrón sugiere que en algunos casos, las características radiográficas compartidas entre estas clases dificultaron la diferenciación precisa.

En la categoría "Tuberculosis", el modelo clasificó correctamente 108 de las 140 instancias reales. No obstante, 19 casos fueron asignados erróneamente a "Normal" y 13 a "Neumonía". Estos errores reflejan las dificultades del modelo para discriminar esta categoría de las otras dos, posiblemente debido a la similitud en los patrones visuales de las imágenes. Por otro lado, en la categoría "Neumonía", se identificaron correctamente 781 de las 855 instancias reales, con 37 confusiones hacia "Tuberculosis" y 37 hacia "Normal". Esto pone en evidencia que, aunque el modelo fue altamente sensible en esta clase, las confusiones estuvieron distribuidas de manera uniforme hacia las otras categorías.

En términos generales, el modelo ViT-B/32 mostró un desempeño sólido en las categorías "Normal" y "Neumonía", mientras que "Tuberculosis" presentó un rendimiento más limitado, aunque aceptable. La matriz de confusión detalla claramente los patrones de error, proporcionando una visión integral del comportamiento del modelo en la clasificación de estas tres condiciones radiográficas. Estos resultados reflejan una capacidad global consistente del modelo para distinguir entre las categorías evaluadas.

4.2.2.4. Métricas de evaluación del modelo basado en ViT-R50-L32

El modelo ViT-R50-L32 fue evaluado para clasificar las categorías "Normal", "Tuberculosis" y "Neumonía", mostrando un desempeño global adecuado. El accuracy global alcanzado fue de 80.1 %, lo que indica que el modelo clasificó correctamente aproximadamente el 80 % de las instancias evaluadas. Adicionalmente, las métricas ponderadas destacaron una precisión promedio de 0.819, un recall promedio de 0.801 y un F1-score promedio de 0.803, lo que evidencia un rendimiento general sólido aunque con algunas áreas específicas de mejora.

En la Tabla 20, se detallan las métricas por categoría. La clase "Normal" obtuvo una precisión de 0.907, lo que refleja un alto nivel de acierto en las predicciones realizadas para esta categoría. El recall fue de 0.740, lo que significa que el modelo identificó correctamente el 74 % de las instancias reales de esta clase. El F1-score de 0.815 resalta un desempeño equilibrado entre precisión y sensibilidad, consolidando a esta categoría como una de las mejor clasificadas por el modelo.

Tabla 20

Métricas de evaluación del modelo basado en ViT-R50-L32 en la detección de múltiples patologías por región anatómica

	Precision	Recall	F1-score	Support
Normal	0.907	0.740	0.815	1017
Tuberculosis	0.547	0.700	0.614	140
Neumonía	0.759	0.890	0.819	855
Accuracy			0.801	2012
Macro avg	0.738	0.777	0.750	2012
Weighted avg	0.819	0.801	0.803	2012

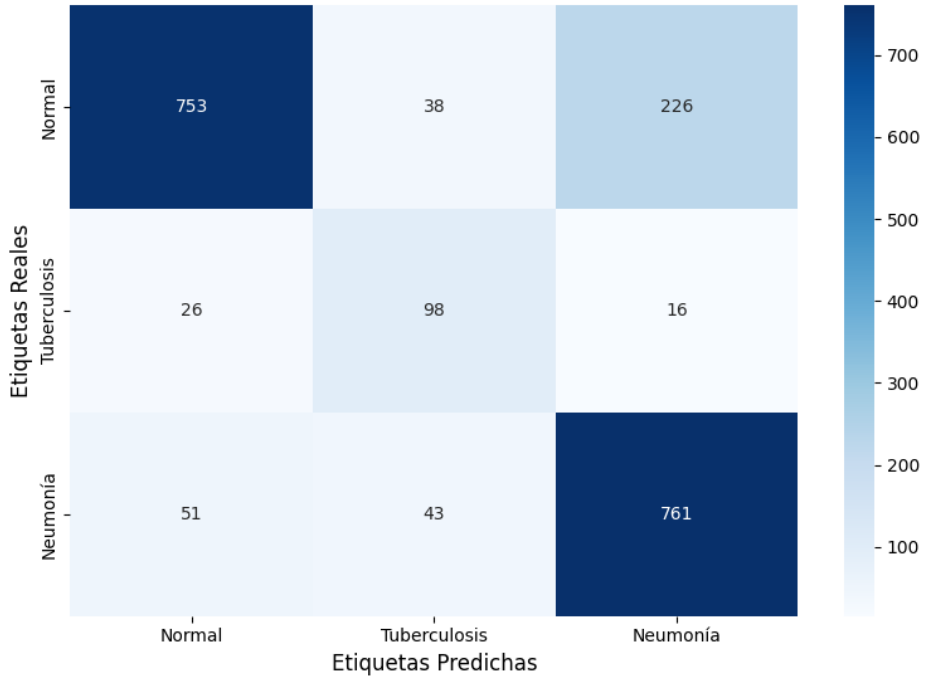
Fuente: Elaboración propia

La categoría "Tuberculosis" presentó un desempeño moderado, con una precisión de 0.547. Este valor indica que el 54.7 % de las predicciones realizadas para esta clase fueron correctas. Por otro lado, el recall de 0.700 refleja que el modelo identificó correctamente el 70 % de las instancias reales de esta categoría. El F1-score de 0.614 evidencia que, si bien el modelo tiene una sensibilidad aceptable, la precisión limitada impacta en la capacidad de clasificación general de esta clase.

En cuanto a la categoría "Neumonía", el modelo mostró un desempeño notable. La precisión de 0.759 indica que el modelo realizó predicciones confiables para esta clase, mientras que el recall de 0.890 demuestra que el 89 % de las instancias reales fueron correctamente identificadas. El F1-score de 0.819 subraya un balance efectivo entre precisión y sensibilidad, destacando esta clase como una de las mejor manejadas por el modelo.

Figura 28

Matriz de confusión para el modelo basado en ViT-R50-L32 en la detección de múltiples patologías por región anatómica



La Figura 28, que ilustra la matriz de confusión, proporciona un análisis más detallado de los aciertos y errores del modelo. Para la categoría "Normal", el modelo clasificó correctamente 753 de las 1017 instancias reales, aunque 226 fueron confundidas con "Neumonía" y 38 con "Tuberculosis". Estas confusiones indican que, en algunos casos, las características radiográficas pueden haberse solapado con las de las otras dos categorías, dificultando su correcta diferenciación.

En la categoría "Tuberculosis", el modelo identificó correctamente 98 de las 140 instancias reales. Sin embargo, se registraron 26 confusiones hacia la clase "Normal" y 16 hacia "Neumonía", lo que sugiere desafíos específicos en la separación de esta categoría de las otras. Por último, en la categoría "Neumonía", el modelo clasificó correctamente 761 de las 855

instancias reales, aunque 51 fueron clasificadas erróneamente como "Normal" y 43 como "Tuberculosis". Estos errores reflejan que, si bien el modelo tiene una alta sensibilidad para esta clase, persisten confusiones entre las categorías.

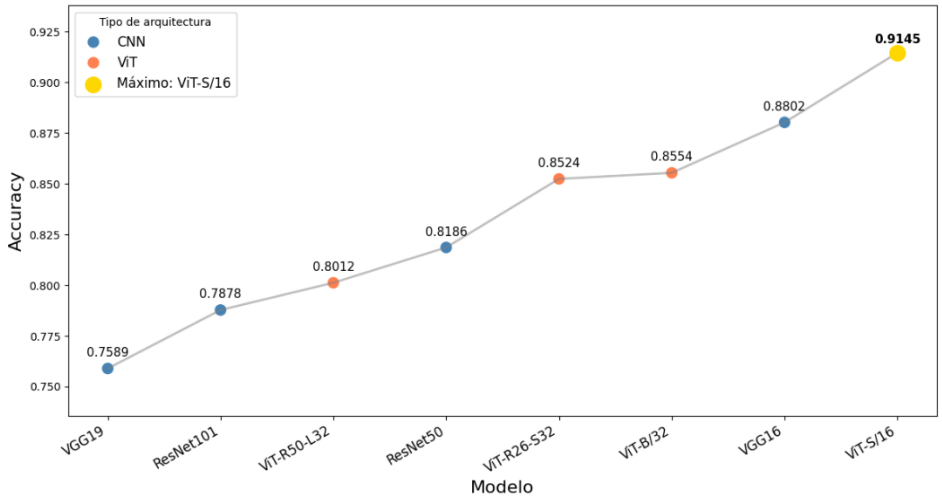
Como puede observarse, el modelo ViT-R50-L32 mostró un desempeño consistente en las categorías "Normal" y "Neumonía", mientras que "Tuberculosis" presentó una clasificación más limitada, principalmente en términos de precisión. La matriz de confusión evidencia los patrones específicos de error, proporcionando una visión detallada del comportamiento del modelo en cada una de las clases evaluadas. Estos resultados reflejan una capacidad global aceptable para clasificar las condiciones radiográficas, destacando tanto fortalezas como áreas de oportunidad en el rendimiento del modelo.

4.2.3. Comparación de las métricas de evaluación de ambas arquitecturas

4.2.3.1. Comparativa del desempeño en términos de accuracy

Figura 29

Comparación de la métrica accuracy entre arquitecturas CNN y Vision Transformers para la detección de patología múltiple



En el contexto de la detección de múltiples patologías, el análisis comparativo de la métrica accuracy revela patrones distintivos entre ambas arquitecturas. Las CNN presentan valores que oscilan entre 0.7589 y 0.8802, donde el modelo VGG16 alcanza el valor más alto (0.8802), seguido por ResNet50 (0.8186), ResNet101 (0.7878), y VGG19 (0.7589).

Por su parte, las arquitecturas Vision Transformers exhiben un rango de accuracy entre 0.8012 y 0.9145. Específicamente, el modelo ViT-S/16 registra el valor más alto (0.9145) entre todas las arquitecturas evaluadas, seguido por ViT-B/32 (0.8554), ViT-R26-S32 (0.8524), y ViT-R50-L32 (0.8012).

La diferencia en accuracy entre los mejores modelos de cada arquitectura (ViT-S/16 y VGG16) es de 3.43 puntos porcentuales, lo cual resulta particularmente significativo considerando la mayor complejidad

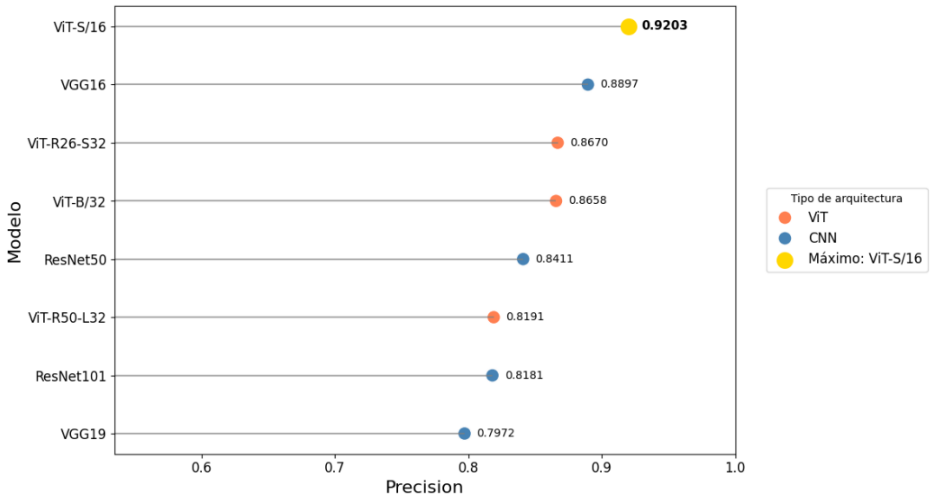
inherente a la clasificación de múltiples patologías en comparación con la detección de patología única.

Es notable que ambas familias de arquitecturas mantienen niveles de accuracy superiores a 0.75 en esta tarea más compleja. Sin embargo, la arquitectura Vision Transformer, específicamente el modelo ViT-S/16, demuestra una superioridad cuantitativa al alcanzar un accuracy de 0.9145, evidenciando una mayor capacidad para la clasificación correcta de múltiples patologías en imágenes radiográficas.

4.2.3.2. Comparativa de la precisión (precisión)

Figura 30

Comparación de la métrica precision entre arquitecturas CNN y Vision Transformers para la detección de patología múltiple



El análisis de la métrica precision en el contexto de múltiples patologías muestra resultados significativos en la capacidad de los modelos para minimizar los falsos positivos. Las arquitecturas CNN exhiben valores que fluctúan entre 0.7972 y 0.8897, donde el modelo

VGG16 alcanza el valor más alto (0.8897), seguido por ResNet50 (0.8411), ResNet101 (0.8181), y VGG19 (0.7972).

En el caso de las arquitecturas Vision Transformers, los resultados experimentales revelan un rango de precision entre 0.8191 y 0.9203. El modelo ViT-S/16 destaca nuevamente al registrar el valor más alto (0.9203) entre todas las arquitecturas evaluadas, seguido por ViT-R26-S32 (0.8670), ViT-B/32 (0.8658), y ViT-R50-L32 (0.8191).

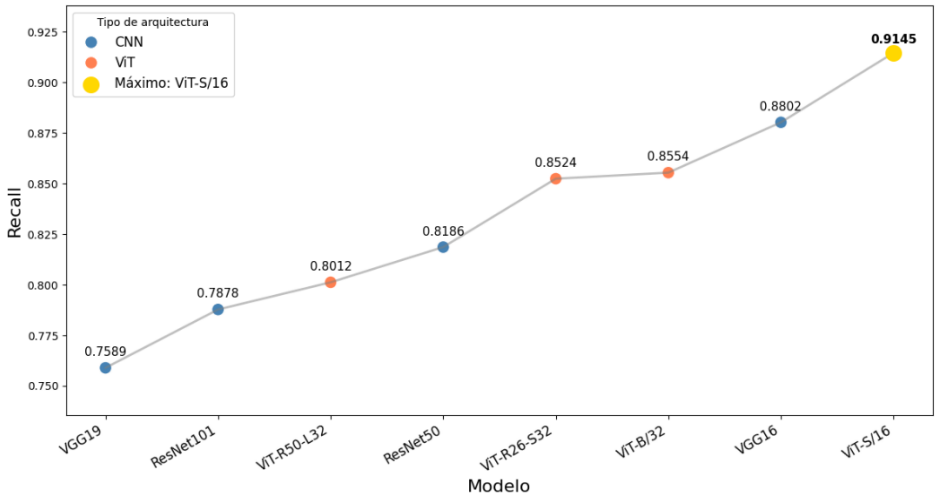
La diferencia de 3.06 puntos porcentuales en precision entre los mejores exponentes de cada arquitectura resulta particularmente relevante, considerando que esta métrica es crucial en escenarios clínicos donde la minimización de falsos positivos es esencial para evitar diagnósticos incorrectos de múltiples patologías.

Resulta significativo que ambas familias de arquitecturas mantienen niveles de precision superiores a 0.79, demostrando una robusta capacidad para evitar falsos positivos incluso en la tarea más compleja de detección múltiple. No obstante, la arquitectura Vision Transformer, especialmente el modelo ViT-S/16, exhibe una clara ventaja al mantener una precision superior a 0.92.

4.2.3.3. Comparativa del recall

Figura 31

Comparación de la métrica recall entre arquitecturas CNN y Vision Transformers para la detección de patología múltiple



Los resultados experimentales del recall en la detección de múltiples patologías proporcionan información esencial sobre la sensibilidad de los modelos. Las arquitecturas CNN presentan valores que oscilan entre 0.7589 y 0.8802, donde el modelo VGG16 logra el valor más alto (0.8802), seguido por ResNet50 (0.8186), ResNet101 (0.7878), y VGG19 (0.7589).

Paralelamente, las arquitecturas Vision Transformers muestran un rango de recall entre 0.8012 y 0.9145. El modelo ViT-S/16 sobresale al registrar el valor más alto (0.9145), seguido por ViT-B/32 (0.8554), ViT-R26-S32 (0.8524), y ViT-R50-L32 (0.8012).

En este contexto, la diferencia de 3.43 puntos porcentuales en recall entre los mejores modelos de cada arquitectura adquiere especial relevancia, dado que refleja una mayor capacidad para identificar

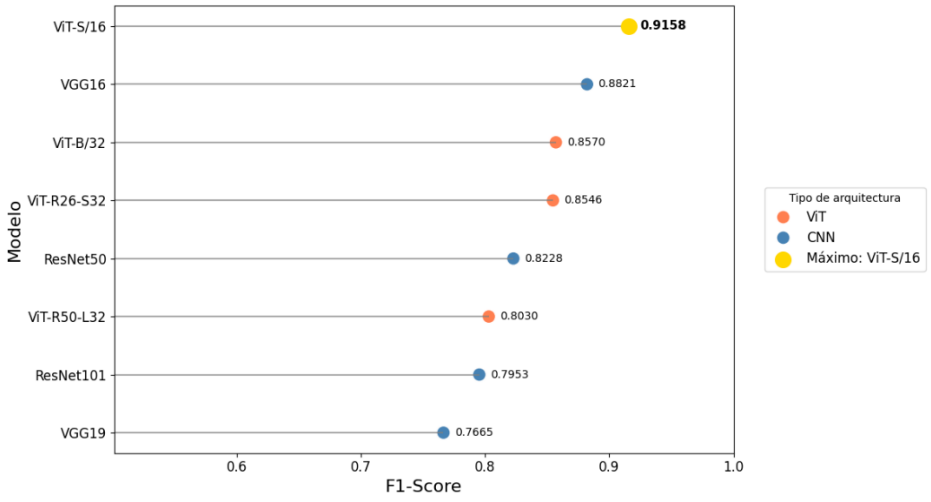
correctamente casos positivos en escenarios con múltiples patologías simultáneas.

Es destacable que ambas familias de arquitecturas mantienen valores de recall superiores a 0.75, demostrando una sólida capacidad de detección incluso en este escenario más complejo. Sin embargo, el modelo ViT-S/16 demuestra una superioridad significativa al alcanzar un recall de 0.9145, indicando una mayor sensibilidad en la detección de múltiples patologías.

4.2.3.4. Comparativa del F1-score

Figura 32

Comparación de la métrica F1-score entre arquitecturas CNN y Vision Transformers para la detección de patología múltiple



El análisis del F1-Score en el contexto de múltiples patologías revela patrones importantes sobre el equilibrio entre precision y recall. Las arquitecturas CNN muestran valores que fluctúan entre 0.7665 y 0.8821, donde el modelo VGG16 alcanza el valor más alto (0.8821), seguido por ResNet50 (0.8228), ResNet101 (0.7953), y VGG19 (0.7665).

En cuanto a las arquitecturas Vision Transformers, los resultados experimentales presentan un rango de F1-Score entre 0.8030 y 0.9158. Específicamente, el modelo ViT-S/16 registra el valor más alto (0.9158) entre todas las arquitecturas evaluadas, seguido por ViT-B/32 (0.8570), ViT-R26-S32 (0.8546), y ViT-R50-L32 (0.8030).

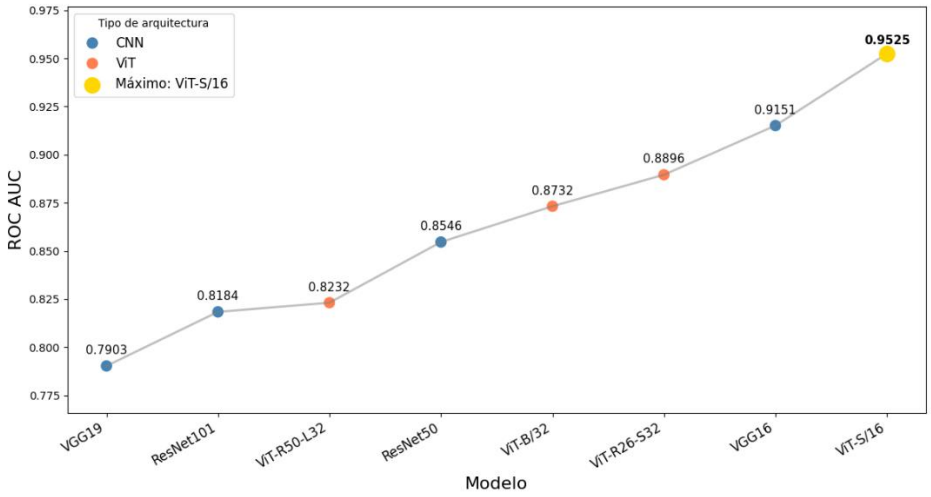
La diferencia de 3.37 puntos porcentuales en F1-Score entre los mejores exponentes de cada arquitectura resulta particularmente significativa, considerando que esta métrica refleja un balance óptimo entre la capacidad de minimizar tanto los falsos positivos como los falsos negativos en la detección de múltiples patologías.

Es notable que ambas familias de arquitecturas mantienen valores de F1-Score superiores a 0.76, demostrando un rendimiento balanceado incluso en esta tarea más compleja. Sin embargo, la arquitectura Vision Transformer, específicamente el modelo ViT-S/16, exhibe una clara superioridad al mantener un F1-Score superior a 0.91, indicando un mejor equilibrio general en sus predicciones de múltiples patologías.

4.2.3.5. Comparativa del ROC AUC

Figura 33

Comparación de la métrica ROC AUC entre arquitecturas CNN y Vision Transformers para la detección de patología múltiple



El análisis del área bajo la curva ROC (ROC AUC) proporciona una perspectiva integral sobre la capacidad discriminativa de los modelos en la detección de múltiples patologías. Las arquitecturas CNN exhiben valores que varían entre 0.7903 y 0.9151, donde el modelo VGG16 alcanza el valor más alto (0.9151), seguido por ResNet50 (0.8546), ResNet101 (0.8184), y VGG19 (0.7903).

Por su parte, las arquitecturas Vision Transformers presentan valores de ROC AUC entre 0.8232 y 0.9525. El modelo ViT-S/16 destaca significativamente al registrar el valor más alto (0.9525) entre todas las arquitecturas evaluadas, seguido por ViT-R26-S32 (0.8896), ViT-B/32 (0.8732), y ViT-R50-L32 (0.8232).

Resulta significativo que ambas familias de arquitecturas mantienen valores de ROC AUC superiores a 0.79, demostrando una sólida capacidad discriminativa. Sin embargo, la arquitectura Vision

Transformer, específicamente el modelo ViT-S/16, exhibe una notable superioridad al alcanzar un ROC AUC de 0.9525, evidenciando una mayor robustez y fiabilidad en su capacidad de discriminación en la detección de múltiples patologías bajo diferentes umbrales de decisión.

4.2.4. Comparación por pares de los modelos con complejidad similar entre ambas arquitecturas mediante validación cruzada

4.2.4.1. Media y desviación estándar del accuracy en la validación cruzada

Tabla 21

Media y desviación estándar del accuracy en las arquitecturas CNN y Vision Transformers para la detección de patología múltiple

Arquitectura	Modelo	Accuracy Medio	Desviación Estándar
CNN	VGG16	0.862197	0.029481
	VGG19	0.747124	0.015426
	ResNet50	0.792426	0.026923
	ResNet101	0.760880	0.032121
ViT	ViT-S/16	0.931264	0.028124
	ViT-R26-S32	0.832701	0.020316
	ViT-B/32	0.849192	0.012172
	ViT-R50-L32	0.796844	0.022154

Fuente: Elaboración propia

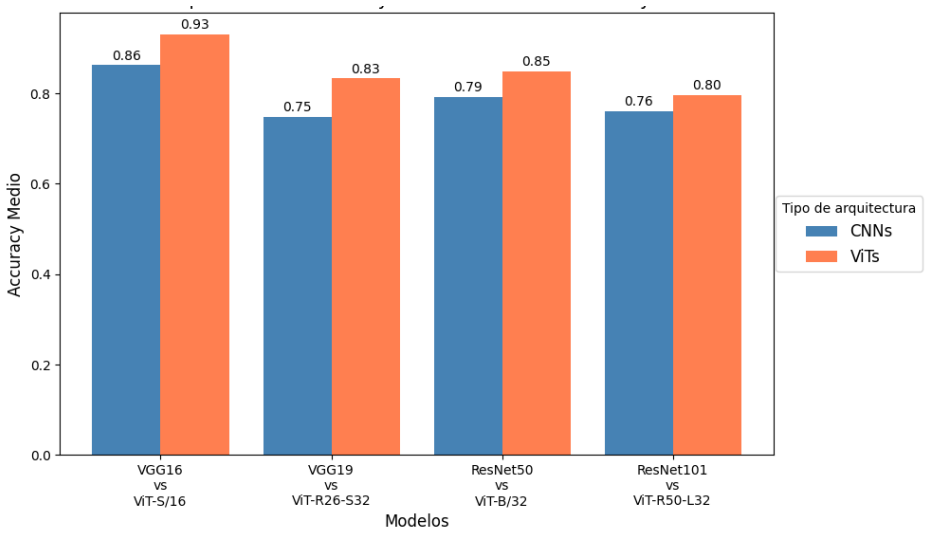
El análisis de los resultados obtenidos mediante validación cruzada en el contexto de múltiples patologías revela patrones distintivos en el rendimiento de ambas arquitecturas. En el ámbito de las CNN, el modelo VGG16 alcanza el accuracy medio más alto con un valor de 0.8622 y una desviación estándar de 0.0295, indicando una variabilidad considerable en

sus predicciones. Por su parte, ResNet50 muestra un rendimiento intermedio con un accuracy medio de 0.7924 y una desviación estándar de 0.0269, mientras que ResNet101 y VGG19 registran valores de 0.7609 ± 0.0321 y 0.7471 ± 0.0154 respectivamente, donde ResNet101 exhibe la mayor variabilidad entre los modelos CNN.

En lo que respecta a los Vision Transformers, los resultados evidencian un rendimiento notablemente superior, donde el modelo ViT-S/16 sobresale significativamente al alcanzar un accuracy medio de 0.9313 con una desviación estándar de 0.0281. Los modelos ViT-B/32 y ViT-R26-S32 mantienen un rendimiento robusto con valores de 0.8492 ± 0.0122 y 0.8327 ± 0.0203 respectivamente, donde es particularmente notable la baja variabilidad del ViT-B/32. El modelo ViT-R50-L32 registra un accuracy medio de 0.7968 con una desviación estándar de 0.0222, manteniendo un nivel de consistencia comparable a sus contrapartes.

Figura 34

Comparación por pares del accuracy medio entre arquitecturas CNN y Vision Transformers para la detección de patología múltiple



El análisis comparativo por pares entre modelos de complejidad similar revela diferencias sustanciales en el rendimiento. La comparación VGG16 vs ViT-S/16 muestra una diferencia significativa de 6.91 puntos porcentuales (0.8622 vs 0.9313) a favor del modelo Vision Transformer. Asimismo, la comparación VGG19 vs ViT-R26-S32 exhibe una diferencia aún más pronunciada de 8.56 puntos porcentuales (0.7471 vs 0.8327), mientras que ResNet50 vs ViT-B/32 presenta una diferencia de 5.68 puntos porcentuales (0.7924 vs 0.8492). La comparación ResNet101 vs ViT-R50-L32 mantiene esta tendencia con una diferencia de 3.60 puntos porcentuales (0.7609 vs 0.7968).

Resulta particularmente significativo que, en todas las comparaciones por pares, los modelos ViT mantienen una superioridad consistente sobre sus contrapartes CNN de similar complejidad. Esta ventaja sistemática, que oscila entre 3.60 % y 8.56 %, sugiere una

capacidad inherentemente superior de la arquitectura Vision Transformer en la tarea más compleja de detección de múltiples patologías. Además, es notable que varios modelos ViT logran esta superioridad mientras mantienen desviaciones estándar menores o comparables a sus contrapartes CNN, indicando una mayor robustez en sus predicciones.

4.2.4.2. Análisis inferencial para determinar diferencias significativas entre los modelos

Para validar la significancia estadística de las diferencias observadas en el rendimiento para la detección de múltiples patologías, se realizó un análisis inferencial exhaustivo. Como paso preliminar, se verificaron los supuestos fundamentales de normalidad y homocedasticidad.

Tabla 22

Supuestos de normalidad y homocedasticidad para la comparación entre modelos CNN y Vision Transformers en el caso de detección de patología múltiple

CNN vs. ViT	Shapiro-Wilk CNN (p-value)	Shapiro-Wilk ViT (p-value)	Levene (p-value)
VGG16 vs. ViT-S/16	0.12406636	0.751630001	0.57938474
VGG19 vs. ViT-R26-S32	0.63291633	0.62401617	0.34119207
ResNet50 vs. ViT-B/32	0.70776796	0.40457714	0.79975382
ResNet101 vs. ViT-R50-L32	0.79725402	0.55375177	0.09384379

Fuente: Elaboración propia

El análisis de los supuestos estadísticos demuestra un sólido cumplimiento de las condiciones necesarias para la aplicación de pruebas paramétricas. La prueba de Shapiro-Wilk para la normalidad en la comparación VGG16 vs. ViT-S/16 muestra valores p de 0.124 y 0.752 respectivamente, superando ampliamente el nivel crítico de 0.05. Este patrón se mantiene consistente en las demás comparaciones, con valores p que oscilan entre 0.405 y 0.797. De manera similar, la prueba de Levene para la homocedasticidad revela valores p entre 0.094 y 0.800, confirmando la homogeneidad de varianzas en todos los pares analizados. El cumplimiento robusto de ambos supuestos fundamenta sólidamente la aplicación de la prueba t de Student.

Tabla 23

Resultados de la prueba t de Student para la comparación entre modelos CNN y Vision Transformers en el caso de detección de patología múltiple

CNN vs ViT	T-Statistic	P-Value (one-tailed)
VGG16 vs. ViT-S/16	3.79044276	0.00265424
VGG19 vs. ViT-R26-S32	7.50153519	3.46E-05
ResNet50 vs. ViT-B/32	4.29605235	0.00131494
ResNet101 vs. ViT-R50-L32	2.06094318	0.03665263

Fuente: Elaboración propia

Para evaluar formalmente la superioridad de los Vision Transformers, se establecieron las siguientes hipótesis estadísticas:

- H_0: $\mu ViT \leq \mu CNN$ (El accuracy medio de los Vision Transformers no es mayor que el de las CNN)

- H_1: μViT > μCNN (El accuracy medio de los Vision Transformers es mayor que el de las CNN)

Los resultados de la prueba t de Student unilateral proporcionan evidencia contundente para rechazar la hipótesis nula en todas las comparaciones realizadas. La comparación VGG19 vs. ViT-R26-S32 exhibe la evidencia más robusta ($t = 7.502$, $p = 3.46 \times 10^{-5}$), seguida por ResNet50 vs. ViT-B/32 ($t = 4.296$, $p = 0.001$) y VGG16 vs. ViT-S/16 ($t = 3.790$, $p = 0.003$). Incluso en la comparación ResNet101 vs. ViT-R50-L32, que muestra el estadístico t más modesto ($t = 2.061$), se obtiene un valor p (0.037) que mantiene la significancia estadística al nivel $\alpha = 0.05$.

La consistencia de estos resultados, reflejada en valores p sistemáticamente inferiores al nivel de significancia establecido, proporciona respaldo estadístico sólido a la superioridad observada en el rendimiento de los Vision Transformers sobre las CNN en la tarea más compleja de detección de múltiples patologías. Esta evidencia estadística robusta valida las diferencias identificadas en el análisis descriptivo previo y refuerza la conclusión sobre la mayor capacidad de los Vision Transformers para abordar tareas de clasificación multiclase en el contexto de imágenes radiográficas.

4.3. DISCUSIÓN

La presente investigación ha proporcionado evidencia empírica sustancial sobre la superioridad de los Vision Transformers (ViT) en comparación con las Redes Neuronales Convolucionales (CNN) en tareas de diagnóstico automatizado mediante imágenes radiográficas. Los resultados obtenidos no solo validan las tendencias

emergentes en la literatura científica reciente, sino que también aportan nuevas perspectivas sobre la eficacia relativa de estas arquitecturas.

En primera instancia, en el contexto de detección de patología única, los hallazgos demuestran que el modelo ViT-S/16 alcanzó un accuracy medio de 0.9132 (± 0.0144), superando notablemente al mejor modelo CNN (VGG16, 0.8717 ± 0.0206). Esta diferencia significativa encuentra respaldo en el exhaustivo análisis de Takahashi et al. (2024), quienes, a través de una revisión sistemática de 36 estudios, evidenciaron la predominancia de los modelos basados en atención sobre las arquitecturas convolucionales tradicionales. De manera similar, Sarmadi et al. (2024) corroboran esta tendencia en su investigación sobre detección de osteoporosis, atribuyendo la superioridad de los ViT a su capacidad inherente para capturar relaciones de largo alcance en las imágenes médicas.

Particularmente revelador resulta el análisis de la detección de múltiples patologías, donde el estudio evidenció un rendimiento aún más destacado del ViT-S/16, alcanzando un accuracy medio de 0.9313 (± 0.0281) en comparación con el VGG16 (0.8622 ± 0.0295). Esta brecha de rendimiento se alinea con los hallazgos de Arshed et al. (2023) en la clasificación multiclase de cáncer de piel. Adicionalmente, estos resultados encuentran respaldo en el trabajo de Hwang et al. (2023), quienes demostraron la superioridad de los ViT en la detección de neuropatía óptica glaucomatosa a través de seis bases de datos independientes, destacando especialmente su eficacia en conjuntos de datos heterogéneos.

La robustez de los resultados se ve significativamente respaldada por el riguroso análisis inferencial realizado, el cual reveló diferencias estadísticamente significativas ($p < 0.05$) en todas las comparaciones por pares. Esta consistencia en el rendimiento superior

de los ViT se ve reforzada por los hallazgos de Garcia-Martin y Sanchez-Reillo (2023), quienes reportaron tasas de identificación excepcionalmente altas utilizando arquitecturas ViT. Además, H. E. Kim et al. (2023) proporcionan evidencia adicional en su estudio sobre clasificación bacteriana, donde los modelos ViT demostraron no solo mayor precisión sino también mejor eficiencia en términos de tiempo de entrenamiento.

La validación cruzada realizada proporciona evidencia convincente sobre la estabilidad y reproducibilidad del rendimiento de los ViT, con el modelo ViT-S/16 exhibiendo una variabilidad notablemente reducida en sus predicciones tanto en patología única (± 0.0144) como en múltiple (± 0.0281). Esta estabilidad encuentra respaldo en el trabajo de Nafisah et al. (2023), quienes demostraron rendimientos excepcionales y consistentes en la detección de COVID-19 mediante arquitecturas basadas en transformers.

Los hallazgos presentados no solo validan la creciente adopción de arquitecturas ViT en el diagnóstico médico automatizado, sino que también sugieren su potencial para establecerse como el estándar de facto en aplicaciones clínicas donde la precisión y consistencia son imperativas.

V. CONCLUSIONES

PRIMERA: En la detección de patología única por región anatómica, los Vision Transformers, específicamente el modelo ViT-S/16, han demostrado una superioridad significativa sobre las Redes Neuronales Convolucionales, alcanzando un accuracy medio de 0.9132 (± 0.0144) en validación cruzada y métricas sobresalientes en el conjunto de prueba: accuracy de 0.9011, precision de 0.9045, recall de 0.9011, F1-Score de 0.9015 y ROC AUC de 0.9195. En comparación, el mejor modelo CNN (VGG16) logró un accuracy medio de 0.8717 (± 0.0206) y valores inferiores en todas las métricas evaluadas. Esta diferencia ha sido validada estadísticamente mediante pruebas t de Student.

SEGUNDA: Para la detección de múltiples patologías por región anatómica, el modelo ViT-S/16 ha exhibido un rendimiento aún más destacado, con un accuracy medio de 0.9313 (± 0.0281) en validación cruzada y métricas superiores en el conjunto de prueba: accuracy de 0.9145, precision de 0.9203, recall de 0.9145, F1-Score de 0.9158 y ROC AUC de 0.9525. Estos resultados superan significativamente al VGG16, que alcanzó un accuracy medio de 0.8622 (± 0.0295), con una diferencia estadísticamente significativa respaldada por la prueba t de Student.

TERCERA: El análisis comparativo por pares entre modelos de complejidad similar ha revelado una superioridad consistente de los Vision Transformers, con diferencias estadísticamente significativas en todas las métricas evaluadas, tanto para patología única como múltiple. Esta consistencia se mantiene a

través de diferentes particiones de datos en la validación cruzada, demostrando la robustez de los resultados.

CUARTA: Los Vision Transformers han demostrado una mayor estabilidad en sus predicciones, evidenciada por menores desviaciones estándar en la validación cruzada (± 0.0144 y ± 0.0281 para patología única y múltiple respectivamente) en comparación con las CNN (± 0.0206 y ± 0.0295). Esta característica, junto con los valores superiores en ROC AUC (0.9195 y 0.9525), indica una mayor confiabilidad y capacidad discriminativa de los modelos ViT.

QUINTA: En respuesta al problema general de investigación, se concluye que los Vision Transformers constituyen la arquitectura con mejor desempeño predictivo para el diagnóstico automatizado mediante imágenes radiográficas, tanto en la detección de patología única (artrosis de rodilla) como en la clasificación de múltiples patologías (neumonía y tuberculosis), superando consistentemente a las CNN en todas las métricas de evaluación empleadas.

VI. RECOMENDACIONES

PRIMERA: Se recomienda extender la evaluación comparativa realizada a otros tipos de imágenes médicas más allá de las radiográficas, como resonancias magnéticas, tomografías computarizadas y ultrasonidos.

SEGUNDA: Se aconseja realizar estudios de validación externa utilizando datos de múltiples instituciones médicas, con el fin de evaluar la robustez y generalización de los modelos Vision Transformer frente a variaciones en protocolos de adquisición de imágenes, equipamiento y poblaciones de pacientes.

TERCERA: Se recomienda investigar la eficiencia computacional y energética de los Vision Transformers en comparación con las CNN, con el objetivo de optimizar su implementación en dispositivos médicos y entornos con recursos limitados como la realidad tecnológica de muchos centros de salud en la región Puno.

CUARTA: Es fundamental promover la colaboración interdisciplinaria entre profesionales de la salud e investigadores en inteligencia artificial para el desarrollo y validación de sistemas basados en Vision Transformers, asegurando que las soluciones desarrolladas respondan efectivamente a las necesidades clínicas reales.

VII. REFERENCIAS BIBLIOGRÁFIAS

Abou Ali, M., Dornaika, F., y Arganda-Carreras, I. (2023). White Blood Cell Classification: Convolutional Neural Network (CNN) and Vision Transformer (ViT) under Medical Microscope. *Algorithms 2023, Vol. 16, Page 525*, *16*(11), 525. https://doi.org/10.3390/A16110525

Acenjo, B. X. T., Pariona, M. A. T., y Cárdenas, E. J. E. (2023). Comparativa entre RESNET-50, VGG-16, Vision Transformer y Swin Transformer para el reconocimiento facial con oclusión de una mascarilla. *Interfases*, *017*, 56–78. https://doi.org/10.26439/INTERFASES2023.N017.6361

Ahsan, M. M., y Siddique, Z. (2022). Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine*, *128*, 102289. https://doi.org/10.1016/J.ARTMED.2022.102289

Akwo, J. D., Trieu, P. D. (Yun), Barron, M. L., Reynolds, T., y Lewis, S. J. (2024). Access to prior screening mammograms affects the specificity but not sensitivity of radiologists' performance. *Clinical Radiology*. https://doi.org/10.1016/J.CRAD.2024.09.007

Arias, F. G. (2012). *El proyecto de investigación: Introducción a la metodología científica* (6a ed.). Editorial Episteme.

Arshed, M. A., Mumtaz, S., Ibrahim, M., Ahmed, S., Tahir, M., y Shafi, M. (2023). Multi-Class Skin Cancer Classification Using Vision Transformer Networks and Convolutional Neural Network-Based Pre-Trained Models. *Information 2023, Vol. 14, Page 415*, *14*(7), 415. https://doi.org/10.3390/INFO14070415

Bressem, K. K., Adams, L. C., Erxleben, C., Hamm, B., Niehues, S. M., y Vahldiek, J. L. (2020). Comparing Different Deep Learning Architectures for Classification of Chest Radiographs. *Scientific Reports*. https://doi.org/10.1038/s41598-020-70479-z

Cabrejos Yalán, V. M. (2022). *Las Redes Neuronales Convolucionales y la mejora en el diagnóstico de Neumonía – área de Radiología*. Universidad Ricardo Palma. https://hdl.handle.net/20.500.14138/5267

Chen, P. (2018). *Knee Osteoarthritis Severity Grading Dataset*. Mendeley Data. https://doi.org/10.17632/56rmx5bjcr.1

Chicco, D., y Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 1–13. https://doi.org/10.1186/S12864-019-6413-7/TABLES/5

Dang, K. M., Zhang, Y. J., Zhang, T., Wang, C., Sinner, A., Coronica, P., y Poon, J. K. S. (2024). NeuroQuantify – An image analysis software for detection and quantification of neuron cells and neurite lengths using deep learning. *Journal of Neuroscience Methods*, *411*, 110273. https://doi.org/10.1016/J.JNEUMETH.2024.110273

Dierickx, P., Van Damme, A., Dupuis, N., y Delaby, O. (2023). Comparison Between CNN, ViT and CCT for Channel Frequency Response Interpretation and Application to G.Fast. *IEEE Access*, *11*, 24039–24052. https://doi.org/10.1109/ACCESS.2023.3247877

Etikan, I., Musa, S. A., y Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. American Journal of Theoretical and Applied Statistics, 5(1), 1-4. https://doi.org/10.11648/j.ajtas.20160501.11

Ferdousi, J., Lincoln, S. I., Alom, M. K., y Foysal, M. (2024). A deep learning approach for white blood cells image generation and classification using SRGAN and VGG19. *Telematics and Informatics Reports*, *16*, 100163. https://doi.org/10.1016/J.TELER.2024.100163

Garcia-Martin, R., y Sanchez-Reillo, R. (2023). Vision Transformers for Vein Biometric Recognition. *IEEE Access*, *11*, 22060–22080. https://doi.org/10.1109/ACCESS.2023.3252009

Ghaffari Laleh, N., Truhn, D., Veldhuizen, G. P., Han, T., van Treeck, M., Buelow, R. D., Langer, R., Dislich, B., Boor, P., Schulz, V., y Kather, J. N. (2022). Adversarial attacks and adversarial robustness in computational pathology. *Nature Communications 2022 13:1*, *13*(1), 1–10. https://doi.org/10.1038/s41467-022-33266-0

Grandini, M., Bagli, E., y Visani, G. (2020). *Metrics for Multi-Class Classification: an Overview*. https://arxiv.org/abs/2008.05756v1

Grandizio, L. C., Ozdag, Y., Mettler, A. W., Garcia, V. C., Manzar, S., Akoon, A., Dwyer, C. L., y Klena, J. C. (2024). Sensitivity, Specificity, and Reliability of the CTS-6 for Carpal Tunnel Syndrome Administered by Medical Assistants. *The Journal of Hand Surgery*, *49*(7), 656–662. https://doi.org/10.1016/J.JHSA.2024.04.001

Gupta, S., y Gupta, A. (2019). Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review. *Procedia Computer Science*, *161*, 466–474. https://doi.org/10.1016/J.PROCS.2019.11.146

Heinrich, A., Güttler, F., Wendt, S., Schenkl, S., Hubig, M., Wagner, R., Mall, G., y Teichgräber, U. (2018). Forensic Odontology: Automatic Identification of Persons

Comparing Antemortem and Postmortem Panoramic Radiographs Using Computer Vision. *Röfo - Fortschritte Auf Dem Gebiet Der Röntgenstrahlen Und Der Bildgebenden Verfahren*. https://doi.org/10.1055/a-0632-4744

Huanco Ramos, F. (2023). *Modelo de identificación de Covid 19 usando técnicas de deep learning a partir de imágenes de Rayos X de Torax de los pulmones de los pacientes*. https://repositorio.unap.edu.pe/handle/20.500.14082/21522

Hwang, E. E., Chen, D., Han, Y., Jia, L., y Shan, J. (2023). Multi-Dataset Comparison of Vision Transformers and Convolutional Neural Networks for Detecting Glaucomatous Optic Neuropathy from Fundus Photographs. *Bioengineering 2023, Vol. 10, Page 1266*, *10*(11), 1266. https://doi.org/10.3390/BIOENGINEERING10111266

Ilesanmi, A. E., Ilesanmi, T., y Gbotoso, G. A. (2023). A systematic review of retinal fundus image segmentation and classification methods using convolutional neural networks. *Healthcare Analytics*, *4*, 100261. https://doi.org/10.1016/J.HEALTH.2023.100261

Jiang, W., y Zhang, L. (2019). Geospatial Data to Images: A Deep-Learning Framework for Traffic Forecasting. *Tsinghua Science y Technology*. https://doi.org/10.26599/tst.2018.9010033

Kanuri, N., Abdelkarim, A. Z., y Rathore, S. (2022). Trainable WEKA (Waikato Environment for Knowledge Analysis) Segmentation Tool: Machine-Learning-Enabled Segmentation on Features of Panoramic Radiographs. *Cureus*. https://doi.org/10.7759/cureus.21777

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M. K., Pei, J., Ting, M., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., … Zhang, K. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, *172*(5), 1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010

Kim, C., Kim, B. Y., y Paeng, D. G. (2024). Bottlenose dolphin identification using synthetic image-based transfer learning. *Ecological Informatics*, *84*, 102909. https://doi.org/10.1016/J.ECOINF.2024.102909

Kim, H. E., Maros, M. E., Miethke, T., Kittel, M., Siegel, F., y Ganslandt, T. (2023). Lightweight Visual Transformers Outperform Convolutional Neural Networks for Gram-Stained Image Classification: An Empirical Study. *Biomedicines 2023, Vol. 11, Page 1333*, *11*(5), 1333. https://doi.org/10.3390/BIOMEDICINES11051333

Kök, H., Acilar, A. M., y İzgi, M. S. (2019). Usage and Comparison of Artificial Intelligence Algorithms for Determination of Growth and Development by Cervical Vertebrae Stages in Orthodontics. *Progress in Orthodontics*. https://doi.org/10.1186/s40510-019-0295-8

Krizhevsky, A., Sutskever, I., y Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Lecun, Y., Bengio, Y., y Hinton, G. (2015). Deep learning. *Nature 2015 521:7553*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Leite, D., Brito, A., y Faccioli, G. (2024). Advancements and outlooks in utilizing Convolutional Neural Networks for plant disease severity assessment: A

comprehensive review. *Smart Agricultural Technology*, *9*, 100573. https://doi.org/10.1016/J.ATECH.2024.100573

Leoni, L., BahooToroody, A., Abaei, M. M., Cantini, A., BahooToroody, F., y De Carlo, F. (2024). Machine learning and deep learning for safety applications: Investigating the intellectual structure and the temporal evolution. *Safety Science*, *170*, 106363. https://doi.org/10.1016/J.SSCI.2023.106363

Madhur Jain, Mayank Singh Bora, Sameer Chandnani, Sanidhay Grover, y Shivank Sadwal. (2023). Comparison of VGG-16, VGG-19, and ResNet-101 CNN Models for the purpose of Suspicious Activity Detection. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 121–130. https://doi.org/10.32628/CSEIT2390124

Matsuda, S., Miyamoto, T., Yoshimura, H., y Hasegawa, T. (2020). Personal Identification With Orthopantomography Using Simple Convolutional Neural Networks: A Preliminary Study. *Scientific Reports*. https://doi.org/10.1038/s41598-020-70474-4

Miralles Linares, F., Martín Quirós, A., y Jaén Cañadas, M. (2024). Triaje en urgencias basado en inteligencia artificial: una herramienta prometedora. *Medicina Clínica*. https://doi.org/10.1016/J.MEDCLI.2024.10.003

Nafisah, S. I., Muhammad, G., Hossain, M. S., y AlQahtani, S. A. (2023). A Comparative Evaluation between Convolutional Neural Networks and Vision Transformers for COVID-19 Detection. *Mathematics 2023, Vol. 11, Page 1489*, *11*(6), 1489. https://doi.org/10.3390/MATH11061489

Norman, B., Pedoia, V., Noworolski, A., Link, T. M., y Majumdar, S. (2018). Applying Densely Connected Convolutional Neural Networks for Staging Osteoarthritis Severity From Plain Radiographs. *Journal of Digital Imaging*. https://doi.org/10.1007/s10278-018-0098-3

Patel, P., Hussain, H., y Fahey, J. (2019). Delayed Diagnosis of Ankylosing Spondylitis: A Missed Opportunity? *Cureus*. https://doi.org/10.7759/cureus.5723

Pauly, G., y Ashok, N. (2018). The Marvel of 3D Imaging: A Case of an Undetectable Condylar Split Fracture. *Modern Research in Dentistry*. https://doi.org/10.31031/mrd.2018.03.000561

Qin, C., Yao, D., Shi, Y., y Song, Z. (2018). Computer-Aided Detection in Chest Radiography Based on Artificial Intelligence: A Survey. *Biomedical Engineering Online*. https://doi.org/10.1186/s12938-018-0544-y

Rahman, T., Khandakar, A., Kadir, M. A., Islam, K. R., Islam, K. F., Mazhar, R., Hamid, T., Islam, M. T., Kashem, S., Mahbub, Z. Bin, Ayari, M. A., y Chowdhury, M. E. H. (2020). Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization. *IEEE Access*, *8*, 191586–191601. https://doi.org/10.1109/ACCESS.2020.3031384

Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D. Y., Bagul, A., Langlotz, C. P., Patel, B. N., Yeom, K. W., Shpanskaya, K., Blankenberg, F. G., Seekins, J., Amrhein, T. J., Mong, D. A., Halabi, S., Zucker, E. J., … Lungren, M. P. (2018). Deep Learning for Chest Radiograph Diagnosis: A Retrospective Comparison of the CheXNeXt Algorithm to Practicing Radiologists. *Plos Medicine*. https://doi.org/10.1371/journal.pmed.1002686

Rao, K. N., Khalaf, O. I., Krishnasree, V., Kumar, A. S., Alsekait, D. M., Priyanka, S. S., Alattas, A. S., y AbdElminaam, D. S. (2024). An efficient brain tumor detection and classification using pre-trained convolutional neural network models. *Heliyon*, *10*(17), e36773. https://doi.org/10.1016/J.HELIYON.2024.E36773

Saha, S., Kumar, A., y Nandi, D. (2024). ViT-ILD: A Vision Transformer-based Neural Network for Detection of Interstitial Lung Disease from CT Images. *Procedia Computer Science*, *235*, 779–788. https://doi.org/10.1016/J.PROCS.2024.04.074

Hernández-Sampieri, R., Fernández-Collado, C., y Baptista-Lucio, M. del P. (2014). Metodología de la investigación (6ª ed.). McGraw-Hill.

Hernández-Sampieri, R., y Mendoza-Torres, C. P. (2018). Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta. McGraw-Hill.

Sarmadi, A., Razavi, Z. S., van Wijnen, A. J., y Soltani, M. (2024). Comparative analysis of vision transformers and convolutional neural networks in osteoporosis detection from X-ray images. *Scientific reports*, *14*(1). https://doi.org/10.1038/S41598-024-69119-7

Shapiro, M. C., Zubkov, K., y Landau, R. (2020). Diagnosis of Stress Fractures in Military Trainees: A Large-Scale Cohort. *BMJ Military Health*. https://doi.org/10.1136/bmjmilitary-2020-001406

Slimi, H., Abid, S., y Sayadi, M. (2024). Advanced deep learning strategies for breast cancer image analysis. *Journal of Radiation Research and Applied Sciences*, *17*(4), 101136. https://doi.org/10.1016/J.JRRAS.2024.101136

Takahashi, S., Sakaguchi, Y., Kouno, N., Takasawa, K., Ishizu, K., Akagi, Y., Aoyama, R., Teraya, N., Bolatkan, A., Shinkai, N., Machino, H., Kobayashi, K., Asada, K.,

Komatsu, M., Kaneko, S., Sugiyama, M., y Hamamoto, R. (2024). Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review. *Journal of medical systems*, *48*(1). https://doi.org/10.1007/S10916-024-02105-8

Targonski, C., Bender, M. R., Shealy, B. T., Husain, B., Paseman, B., Smith, M. C., y Feltus, F. A. (2020). Cellular State Transformations Using Deep Learning for Precision Medicine Applications. *Patterns*, *1*(6), 100087. https://doi.org/10.1016/J.PATTER.2020.100087

Tian, D., Zhou, C., Wang, Y., Zhang, R., y Yao, Y. (2024). NB-TCM-CHM: Image dataset of the Chinese herbal medicine fruits and its application in classification through deep learning. *Data in Brief*, *54*, 110405. https://doi.org/10.1016/J.DIB.2024.110405

Tseng, Y. H., y Jiang, F. J. (2024). Learning the phase transitions of two-dimensional Potts model with a pre-trained one-dimensional neural network. *Results in Physics*, *56*, 107264. https://doi.org/10.1016/J.RINP.2023.107264

Wang, G., Luo, X., Gu, R., Yang, S., Qu, Y., Zhai, S., Zhao, Q., Li, K., y Zhang, S. (2023). PyMIC: A deep learning toolkit for annotation-efficient medical image segmentation. *Computer Methods and Programs in Biomedicine*, *231*, 107398. https://doi.org/10.1016/J.CMPB.2023.107398

Yan An Montoya, C., y Sofía Alejandra Cornejo, G. (2022). Detección de COVID-19 a partir de imágenes radiográficas utilizando redes neuronales convolucionales: una revisión bibliográfica. *INGENIERÍA INVESTIGA*, *4*. https://doi.org/10.47796/ING.V4I0.626

Yang, H.-C., Jo, E. A., Kim, H. J., Cha, I., Jung, Y.-S., Nam, W., Kim, J., Kim, Y. H., Oh, T. G., Han, S.-S., Kim, H., y Kim, D.-W. (2020). Deep Learning for Automated Detection of Cyst and Tumors of the Jaw in Panoramic Radiographs. *Journal of Clinical Medicine*. https://doi.org/10.3390/jcm9061839

Yang, L., Finnerty, P., y Ohta, C. (2024). Applications of cluster-based transfer learning in image and localization tasks. *Machine Learning with Applications*, *18*, 100601. https://doi.org/10.1016/J.MLWA.2024.100601

Yoo, J.-H., Yeom, H.-G., Shin, W., Yun, J. P., Lee, J. H., Jeong, S., Lim, H. J., Lee, J., y Kim, B.-C. (2021). Deep Learning Based Prediction of Extraction Difficulty for Mandibular Third Molars. *Scientific Reports*. https://doi.org/10.1038/s41598-021-81449-4

Zhang, Y., Liu, Y. L., Nie, K., Zhou, J., Chen, Z., Chen, J. H., Wang, X., Kim, B., Parajuli, R., Mehta, R. S., Wang, M., y Su, M. Y. (2023). Deep Learning-based Automatic Diagnosis of Breast Cancer on MRI Using Mask R-CNN for Detection Followed by ResNet50 for Classification. *Academic Radiology*, *30*, S161–S171. https://doi.org/10.1016/J.ACRA.2022.12.038

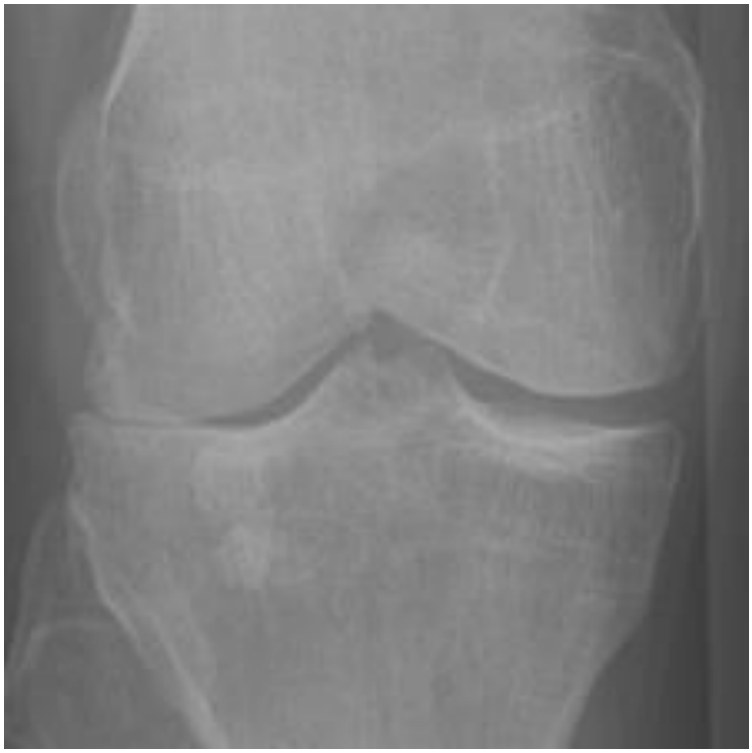
Zhao, Y. F., Sun, X. L., y Yang, J. X. (2023). Automatic recognition of surface defects of hot rolled strip steel based on deep parallel attention convolution neural network. *Materials Letters*, *353*, 135313. https://doi.org/10.1016/J.MATLET.2023.135313

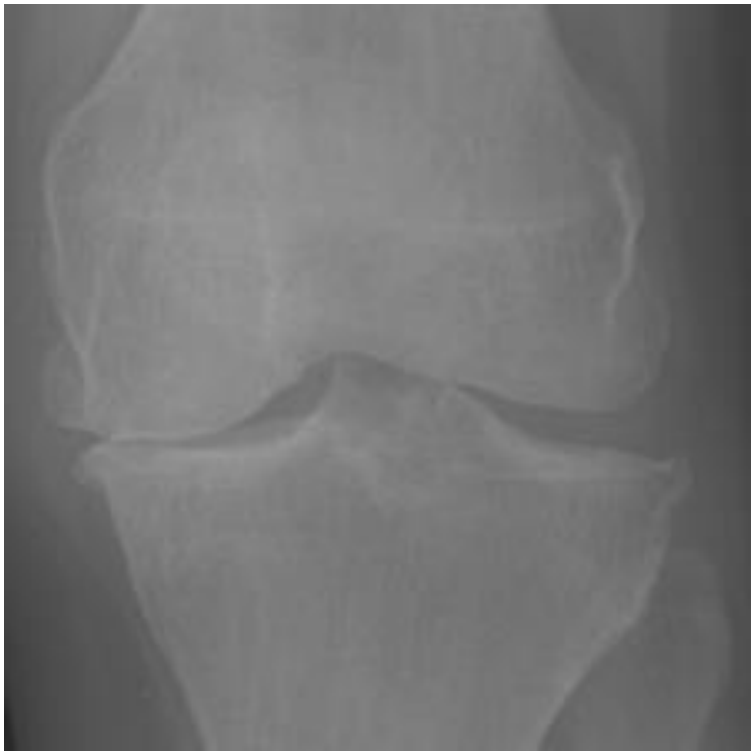
ANEXOS

ANEXO 1: Matriz de consistencia

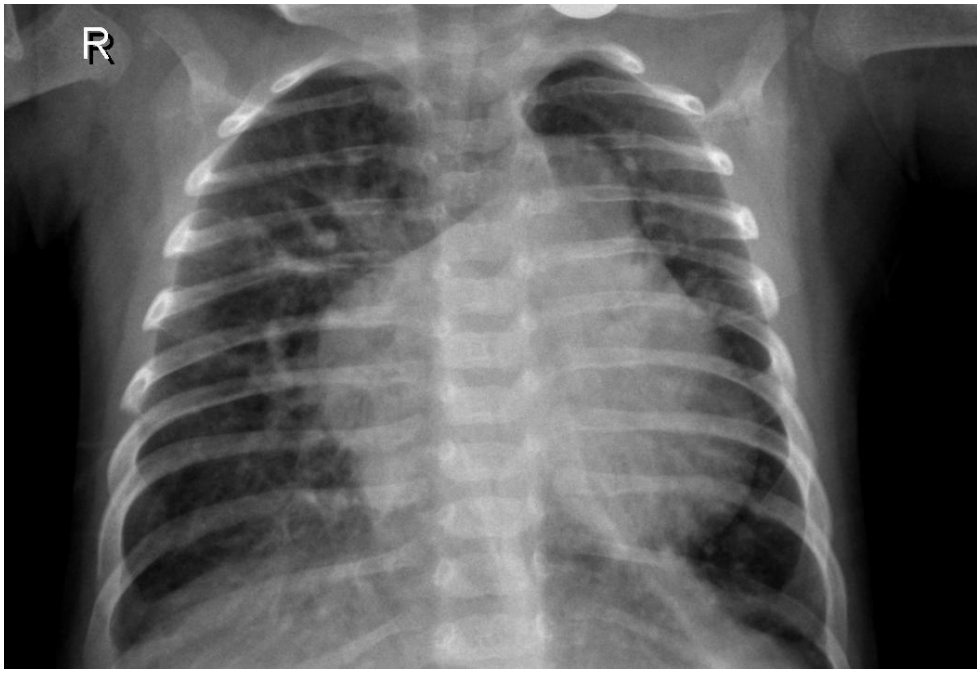
PROBLEMA	OBJETIVO	HIPÓTESIS	VARIABLES	METODOLOGÍA
General			**V. Independiente:** Arquitecturas de deep learning **Dimensiones:** - CNN: VGG16, VGG19, ResNet50, ResNet101 - ViT: ViT-S/16, ViT-R26-S32, ViT-B/32, ViT-R50-L32 **V. Dependiente:** Desempeño predictivo **Indicadores:** - Accuracy - Precision - Recall - F1-Score - ROC AUC	**Tipo:** Cuantitativa **Diseño:** No experimental **Nivel:** Comparativo **Población:** 15,834 imágenes radiográficas **Muestra:** - 5,778 (osteoporosis) - 5,863 (neumonía) - 4,193 (tuberculosis) **Técnicas:** - Validación cruzada - Pruebas estadísticas - Análisis comparativo
¿Cuál es la arquitectura con mejor desempeño predictivo entre las Redes Neuronales Convolucionales (CNN) y los Vision Transformers (ViT) en el diagnóstico automatizado de imágenes radiográficas para la detección de patologías únicas y múltiples por región anatómica?	Evaluar y comparar el desempeño predictivo de las Redes Neuronales Convolucionales (CNN) y los Vision Transformers (ViT) en el diagnóstico automatizado de imágenes radiográficas para la detección de patologías únicas y múltiples por región anatómica.	Los Vision Transformers (ViT) presentan un mejor desempeño predictivo que las Redes Neuronales Convolucionales (CNN) en el diagnóstico automatizado de imágenes radiográficas para la detección de patologías únicas y múltiples por región anatómica.		
Específicos				
¿Cuál es la arquitectura con mejor desempeño predictivo entre las Redes Neuronales Convolucionales (CNN) y los Vision Transformers (ViT) en el diagnóstico de osteoporosis de rodilla, como caso de estudio para la detección de una patología única en una región anatómica específica?	Evaluar y comparar el desempeño predictivo de las Redes Neuronales Convolucionales (CNN) y los Vision Transformers (ViT) en el diagnóstico de osteoporosis de rodilla, como caso de estudio de detección de patología única por región anatómica.	Los Vision Transformers (ViT) superan a las Redes Neuronales Convolucionales (CNN) en el desempeño predictivo para el diagnóstico de osteoporosis en imágenes radiográficas de rodilla, como caso de estudio de detección de una patología única en una región anatómica específica.		
¿Cuál es la arquitectura con mejor desempeño predictivo entre las Redes Neuronales Convolucionales (CNN) y los Vision Transformers (ViT) en el diagnóstico de neumonía y tuberculosis, como caso de estudio para la detección de múltiples patologías en una misma región anatómica?	Evaluar y comparar el desempeño predictivo de las Redes Neuronales Convolucionales (CNN) y los Vision Transformers (ViT) en el diagnóstico de neumonía y tuberculosis como caso de estudio de detección de patologías múltiples por región anatómica.	Los Vision Transformers (ViT) presentan un mejor desempeño predictivo que las Redes Neuronales Convolucionales (CNN) en el diagnóstico de neumonía y tuberculosis a partir de imágenes radiográficas, como caso de estudio de detección de múltiples patologías en una misma región anatómica.		

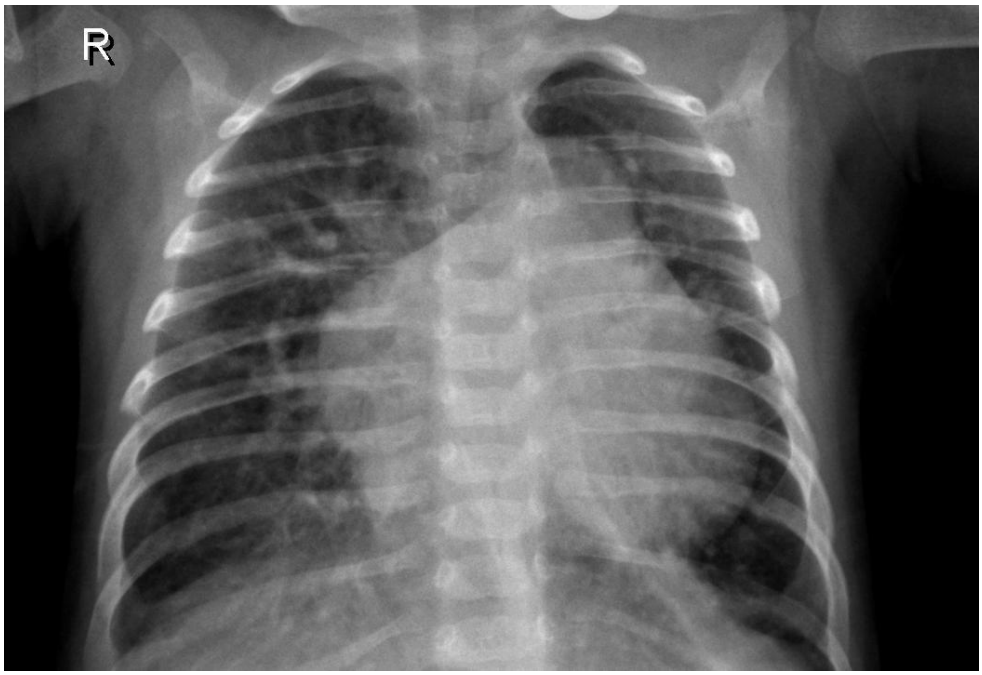
ANEXO 2: Imágenes radiográficas con artrosis de rodilla



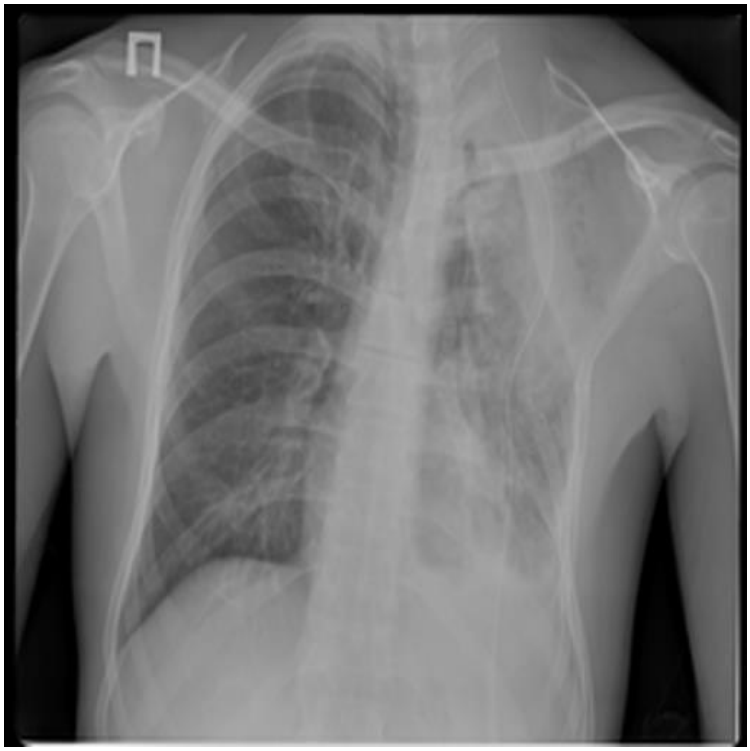


ANEXO 3: Imágenes radiográficas con neumonía





ANEXO 4: Imágenes radiográficas con tuberculosis



ANEXO 5: Código para la normalización y preparación de datos

```
def prepare_data(x_knee, y_knee, test_size=0.2, validation_split=0.2):
    """
    Prepara los datos dividiendo en conjuntos de entrenamiento,
validación y prueba
    """
    # Convertir a float32 y normalizar
    x_knee = x_knee.astype('float32') / 255.0

    # Expandir dimensiones para imagenes en escala de grises
    if len(x_knee.shape) == 3:
        x_knee = np.expand_dims(x_knee, axis=-1)
    if x_knee.shape[-1] == 1:
        x_knee = np.concatenate([x_knee] * 3, axis=-1)

    # Dividir en train+val y test
    X_train_val, X_test, y_train_val, y_test = train_test_split(
        x_knee, y_knee,
        test_size=test_size,
        random_state=42,
        stratify=y_knee
    )

    # Dividir train+val en train y val
    X_train, X_val, y_train, y_val = train_test_split(
        X_train_val, y_train_val,
        test_size=validation_split,
        random_state=42,
        stratify=y_train_val
    )

    return X_train, X_val, X_test, y_train, y_val, y_test

def create_data_generators(X_train, X_val, X_test, y_train, y_val,
y_test, batch_size=32):
    """
    Crea generadores de datos con aumentación para entrenamiento
    """
    train_datagen = ImageDataGenerator(
        rescale=1./255,
        rotation_range=20,
        zoom_range=0.1,
        brightness_range=[0.8, 1.2],
        width_shift_range=0.2,
        height_shift_range=0.2,
        horizontal_flip=True,
        fill_mode='nearest'
```

```
)

# Generadores para validación y prueba (solo rescaling)
val_datagen = ImageDataGenerator(rescale=1./255)
test_datagen = ImageDataGenerator(rescale=1./255)

train_generator = train_datagen.flow(
    X_train, y_train,
    batch_size=batch_size
)

val_generator = val_datagen.flow(
    X_val, y_val,
    batch_size=batch_size
)

test_generator = test_datagen.flow(
    X_test, y_test,
    batch_size=batch_size
)

return train_generator, val_generator, test_generator
```

ANEXO 6: Código para el entrenamiento de los modelos CNN

```
def create_model(base_model_name, num_classes, input_shape=(224, 224,
3)):
    """
    Crea el modelo CNN según la arquitectura especificada
    """
    if base_model_name == 'VGG16':
        base_model = VGG16(weights='imagenet', include_top=False,
input_shape=input_shape)
    elif base_model_name == 'VGG19':
        base_model = VGG19(weights='imagenet', include_top=False,
input_shape=input_shape)
    elif base_model_name == 'ResNet50':
        base_model = ResNet50(weights='imagenet', include_top=False,
input_shape=input_shape)
    elif base_model_name == 'ResNet101':
        base_model = ResNet101(weights='imagenet', include_top=False,
input_shape=input_shape)
    else:
        raise ValueError(f"Modelo {base_model_name} no soportado")

    # Congelar las capas del modelo base
    base_model.trainable = False

    x = base_model.output
    x = GlobalAveragePooling2D()(x)
    x = Dense(1024, activation='relu')(x)
    x = Dropout(0.5)(x)
    predictions = Dense(num_classes, activation='softmax')(x)

    # Crear el modelo final
    model = Model(inputs=base_model.input, outputs=predictions)

    return model

def train_model(model, train_generator, val_generator, epochs=50,
model_name="model"):
    """
    Entrena el modelo con los generadores proporcionados
    """
    # Compilar el modelo
    model.compile(
        optimizer='adam',
        loss='sparse_categorical_crossentropy',
        metrics=['accuracy']
    )
```

```
    # Callbacks
    checkpoint = ModelCheckpoint(
        f'best_{model_name}.h5',
        monitor='val_accuracy',
        save_best_only=True,
        mode='max',
        verbose=1
    )

    early_stopping = EarlyStopping(
        monitor='val_accuracy',
        patience=10,
        restore_best_weights=True
    )

    # Entrenar el modelo
    history = model.fit(
        train_generator,
        validation_data=val_generator,
        epochs=epochs,
        callbacks=[checkpoint, early_stopping]
    )

    return history

def evaluate_model(model, test_generator):
    """
    Evalúa el modelo en el conjunto de prueba y calcula métricas
adicionales.
    """
    # Evaluar la pérdida y la precisión del modelo
    test_loss, test_accuracy = model.evaluate(test_generator)

    # Obtener predicciones y etiquetas verdaderas
    y_true = []
    y_pred = []
    y_probs = []

    for batch in test_generator:
        X, y = batch
        predictions = model.predict(X)
        y_probs.extend(predictions)
        y_pred.extend(np.argmax(predictions, axis=1))
        y_true.extend(np.argmax(y, axis=1))

    # Convertir a arrays
    y_true = np.array(y_true)
    y_pred = np.array(y_pred)
    y_probs = np.array(y_probs)
```

```
    # Calcular métricas adicionales
    report = classification_report(y_true, y_pred, output_dict=True,
zero_division=0)
    accuracy = test_accuracy
    precision = report["macro avg"]["precision"]
    recall = report["macro avg"]["recall"]
    f1_score = report["macro avg"]["f1-score"]

    # Calcular ROC AUC (One-vs-One)
    y_true_bin = label_binarize(y_true, classes=np.unique(y_true))
    auc = roc_auc_score(y_true_bin, y_probs, average="macro",
multi_class="ovo")

    # Retornar todas las métricas en un diccionario
    return {
        "Test Loss": test_loss,
        "Test Accuracy": accuracy,
        "Precision (Macro Avg)": precision,
        "Recall (Macro Avg)": recall,
        "F1-Score (Macro Avg)": f1_score,
        "ROC AUC": auc
    }

# Función principal para entrenar todos los modelos
def train_all_models(x_knee, y_knee, num_classes):
    # Preparar los datos
    X_train, X_val, X_test, y_train, y_val, y_test = prepare_data(x_knee,
y_knee)

    # Crear generadores de datos
    train_generator, val_generator, test_generator =
create_data_generators(
        X_train, X_val, X_test, y_train, y_val, y_test
    )

    # Lista de modelos a entrenar
    models_to_train = ['VGG16', 'VGG19', 'ResNet50', 'ResNet101']
    results = {}

    # Entrenar cada modelo
    for model_name in models_to_train:
        print(f"\nEntrenando {model_name}...")

        # Crear y entrenar el modelo
        model = create_model(model_name, num_classes)
        history = train_model(model, train_generator, val_generator,
model_name=model_name)
```

```
        # Evaluar el modelo
        test_loss, test_accuracy = evaluate_model(model, test_generator)

        # Guardar resultados
        results[model_name] = {
            'history': history.history,
            'test_loss': test_loss,
            'test_accuracy': test_accuracy
        }

        # Mostrar gráficas
        plot_training_history(history, model_name)

        print(f"\nResultados de {model_name}:")
        print(f"Test Loss: {test_loss:.4f}")
        print(f"Test Accuracy: {test_accuracy:.4f}")

    return results
```

ANEXO 7: Código para el entrenamiento de los modelos ViT

```
def get_vit_model(model_name, num_classes):
    """
    Crea el modelo ViT según la arquitectura especificada
    """
    vit_models = {
        'vit_s16_fe': "https://www.kaggle.com/models/spsayakpaul/vision-
transformer/TensorFlow2/vit-s16-fe/1",
        'vit_r26_s32_medaug_fe':
"https://www.kaggle.com/models/spsayakpaul/vision-
transformer/TensorFlow2/vit-r26-s32-medaug-fe/1",
        'vit_b32_fe': "https://www.kaggle.com/models/spsayakpaul/vision-
transformer/TensorFlow2/vit-b32-fe/1",
        'vit_r50_l32_fe':
"https://www.kaggle.com/models/spsayakpaul/vision-
transformer/TensorFlow2/vit-r50-l32-fe/1"
    }

    if model_name not in vit_models:
        raise ValueError(f"Modelo {model_name} no soportado")

    # Crear modelo secuencial
    model = tf.keras.Sequential([
        hub.KerasLayer(vit_models[model_name], trainable=True),
        tf.keras.layers.Dense(num_classes, activation='softmax')
    ])

    # Compilar modelo
    model.compile(
        optimizer=tfa.optimizers.AdamW(learning_rate=3e-4,
weight_decay=0.0001),
        loss='sparse_categorical_crossentropy',
        metrics=['accuracy']
    )

    return model

def train_vit_model(model, train_generator, val_generator, epochs=50,
model_name="model"):
    """
    Entrena el modelo ViT
    """
    # Callbacks
    checkpoint = ModelCheckpoint(
        f'best_{model_name}.h5',
        monitor='val_accuracy',
        save_best_only=True,
```

```
        mode='max',
        verbose=1
    )

    early_stopping = EarlyStopping(
        monitor='val_accuracy',
        patience=10,
        restore_best_weights=True
    )

    # Entrenar modelo
    history = model.fit(
        train_generator,
        validation_data=val_generator,
        epochs=epochs,
        callbacks=[checkpoint, early_stopping]
    )

    return history

def evaluate_model(model, test_generator):
    """
    Evalúa el modelo en el conjunto de prueba y calcula métricas
adicionales.
    """
    # Evaluar la pérdida y la precisión del modelo
    test_loss, test_accuracy = model.evaluate(test_generator)

    # Obtener predicciones y etiquetas verdaderas
    y_true = []
    y_pred = []
    y_probs = []

    for batch in test_generator:
        X, y = batch
        predictions = model.predict(X)
        y_probs.extend(predictions)
        y_pred.extend(np.argmax(predictions, axis=1))
        y_true.extend(np.argmax(y, axis=1))

    # Convertir a arrays numpy
    y_true = np.array(y_true)
    y_pred = np.array(y_pred)
    y_probs = np.array(y_probs)

    # Calcular métricas adicionales
    report = classification_report(y_true, y_pred, output_dict=True,
zero_division=0)
    accuracy = test_accuracy
```

```
    precision = report["macro avg"]["precision"]
    recall = report["macro avg"]["recall"]
    f1_score = report["macro avg"]["f1-score"]

    # Calcular ROC AUC (One-vs-One)
    y_true_bin = label_binarize(y_true, classes=np.unique(y_true))
    auc = roc_auc_score(y_true_bin, y_probs, average="macro",
multi_class="ovo")

    # Retornar todas las métricas en un diccionario
    return {
        "Test Loss": test_loss,
        "Test Accuracy": accuracy,
        "Precision (Macro Avg)": precision,
        "Recall (Macro Avg)": recall,
        "F1-Score (Macro Avg)": f1_score,
        "ROC AUC": auc
    }

def train_all_vit_models(x_knee, y_knee, num_classes):
    """
    Entrena todos los modelos ViT
    """
    # Preparar datos
    X_train, X_val, X_test, y_train, y_val, y_test = prepare_data(x_knee,
y_knee)

    # Crear generadores de datos
    train_generator, val_generator, test_generator =
create_data_generators(
        X_train, X_val, X_test, y_train, y_val, y_test
    )

    # Lista de modelos a entrenar
    vit_models = [
        'vit_s16_fe',
        'vit_r26_s32_medaug_fe',
        'vit_b32_fe',
        'vit_r50_l32_fe'
    ]

    results = {}

    # Entrenar cada modelo
    for model_name in vit_models:
        print(f"\nEntrenando {model_name}...")

        # Crear y entrenar modelo
        model = get_vit_model(model_name, num_classes)
```

```
        # Mostrar resumen del modelo
        print("\nArquitectura del modelo:")
        model.summary()

        history = train_vit_model(model, train_generator, val_generator,
model_name=model_name)

        # Evaluar modelo
        test_loss, test_accuracy = evaluate_model(model, test_generator)

        # Guardar resultados
        results[model_name] = {
            'history': history.history,
            'test_loss': test_loss,
            'test_accuracy': test_accuracy
        }

        # Mostrar gráficas
        plot_training_history(history, model_name)

        print(f"\nResultados de {model_name}:")
        print(f"Test Loss: {test_loss:.4f}")
        print(f"Test Accuracy: {test_accuracy:.4f}")

        # Guardar métricas en archivo
        with open(f'{model_name}_metrics.txt', 'w') as f:
            f.write(f"Test Loss: {test_loss:.4f}\n")
            f.write(f"Test Accuracy: {test_accuracy:.4f}\n")

        # Limpiar memoria
        tf.keras.backend.clear_session()

    return results
```

ANEXO 8: Declaración jurada de autenticidad de tesis

Universidad Nacional del Altiplano Puno | VRI Vicerrectorado de Investigación | Repositorio Institucional

DECLARACIÓN JURADA DE AUTENTICIDAD DE TESIS

Por el presente documento, Yo YEFER ANDERSSON MAMANI CHAMBI,
identificado con DNI 75449896 en mi condición de egresado de:

☒ **Escuela Profesional,** ☐ **Programa de Segunda Especialidad,** ☐ **Programa de Maestría o Doctorado**

INGENIERÍA ESTADÍSTICA E INFORMÁTICA,

informo que he elaborado el/la ☒ **Tesis** o ☐ **Trabajo de Investigación** denominada:
" ANÁLISIS COMPARATIVO DE REDES NEURONALES CONVOLUCIONALES Y VISION TRANSFORMERS PARA EL DIAGNÓSTICO AUTOMATIZADO EN IMÁGENES RADIOGRÁFICAS"

Es un tema original.

Declaro que el presente trabajo de tesis es elaborado por mi persona y **no existe plagio/copia** de ninguna naturaleza, en especial de otro documento de investigación (tesis, revista, texto, congreso, o similar) presentado por persona natural o jurídica alguna ante instituciones académicas, profesionales, de investigación o similares, en el país o en el extranjero.

Dejo constancia que las citas de otros autores han sido debidamente identificadas en el trabajo de investigación, por lo que no asumiré como suyas las opiniones vertidas por terceros, ya sea de fuentes encontradas en medios escritos, digitales o Internet.

Asimismo, ratifico que soy plenamente consciente de todo el contenido de la tesis y asumo la responsabilidad de cualquier error u omisión en el documento, así como de las connotaciones éticas y legales involucradas.

En caso de incumplimiento de esta declaración, me someto a las disposiciones legales vigentes y a las sanciones correspondientes de igual forma me someto a las sanciones establecidas en las Directivas y otras normas internas, así como las que me alcancen del Código Civil y Normas Legales conexas por el incumplimiento del presente compromiso

Puno 16 de diciembre del 2024

FIRMA (obligatoria) Huella

ANEXO 9: Autorización para el depósito de tesis en el Repositorio Institucional

Universidad Nacional del Altiplano Puno | VRI Vicerrectorado de Investigación | Repositorio Institucional

AUTORIZACIÓN PARA EL DEPÓSITO DE TESIS O TRABAJO DE INVESTIGACIÓN EN EL REPOSITORIO INSTITUCIONAL

Por el presente documento, Yo YEFER ANDERSSON MAMANI CHAMBI,
identificado con DNI 75449896 en mi condición de egresado de:

☒ **Escuela Profesional,** ☐ **Programa de Segunda Especialidad,** ☐ **Programa de Maestría o Doctorado**

INGENIERÍA ESTADÍSTICA E INFORMÁTICA,
informo que he elaborado el/la ☒ **Tesis** o ☐ **Trabajo de Investigación** denominada:

" ANÁLISIS COMPARATIVO DE REDES NEURONALES CONVOLUCIONALES Y VISION TRANSFORMERS PARA EL DIAGNÓSTICO AUTOMATIZADO EN IMÁGENES RADIOGRÁFICAS "

para la obtención de ☐**Grado,** ☒ **Título Profesional** o ☐ **Segunda Especialidad.**

Por medio del presente documento, afirmo y garantizo ser el legítimo, único y exclusivo titular de todos los derechos de propiedad intelectual sobre los documentos arriba mencionados, las obras, los contenidos, los productos y/o las creaciones en general (en adelante, los "Contenidos") que serán incluidos en el repositorio institucional de la Universidad Nacional del Altiplano de Puno.

También, doy seguridad de que los contenidos entregados se encuentran libres de toda contraseña, restricción o medida tecnológica de protección, con la finalidad de permitir que se puedan leer, descargar, reproducir, distribuir, imprimir, buscar y enlazar los textos completos, sin limitación alguna.

Autorizo a la Universidad Nacional del Altiplano de Puno a publicar los Contenidos en el Repositorio Institucional y, en consecuencia, en el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto, sobre la base de lo establecido en la Ley N° 30035, sus normas reglamentarias, modificatorias, sustitutorias y conexas, y de acuerdo con las políticas de acceso abierto que la Universidad aplique en relación con sus Repositorios Institucionales. Autorizo expresamente toda consulta y uso de los Contenidos, por parte de cualquier persona, por el tiempo de duración de los derechos patrimoniales de autor y derechos conexos, a título gratuito y a nivel mundial.

En consecuencia, la Universidad tendrá la posibilidad de divulgar y difundir los Contenidos, de manera total o parcial, sin limitación alguna y sin derecho a pago de contraprestación, remuneración ni regalía alguna a favor mío; en los medios, canales y plataformas que la Universidad y/o el Estado de la República del Perú determinen, a nivel mundial, sin restricción geográfica alguna y de manera indefinida, pudiendo crear y/o extraer los metadatos sobre los Contenidos, e incluir los Contenidos en los índices y buscadores que estimen necesarios para promover su difusión.

Autorizo que los Contenidos sean puestos a disposición del público a través de la siguiente licencia:

Creative Commons Reconocimiento-NoComercial-CompartirIgual 4.0 Internacional. Para ver una copia de esta licencia, visita: https://creativecommons.org/licenses/by-nc-sa/4.0/

En señal de conformidad, suscribo el presente documento.

Puno 16 de diciembre del 2024

FIRMA (obligatoria)

Huella