Resumen

Este Trabajo de Fin de Grado aborda el problema del desequilibrio de clases en conjuntos de datos de imágenes médicas, tomando como caso de estudio el diagnóstico de patologías oculares a partir del *dataset* ODIR-5K. Se propone un diseño experimental comparativo en el que se analizan diferentes estrategias de aumento de datos para mejorar la representatividad de las clases minoritarias.

Se evaluaron técnicas tradicionales (rotaciones, *flips*, brillo), avanzadas (MixUp y CutMix) y generativas, mediante redes adversarias (WGAN-GP) y modelos de difusión (DDPM). A partir de cada enfoque, se generaron versiones equilibradas del conjunto de datos, sobre las cuales se entrenaron modelos de clasificación con *embeddings* extraídos mediante RETFound y EfficientNetB3.

La evaluación incluyó métricas cuantitativas (precisión, *recall*, *F1-score*, AUC), análisis cualitativo con Grad-CAM y evaluación de la calidad de las imágenes generadas (FID, IS). Los resultados muestran que el uso combinado de estas técnicas mejora significativamente el rendimiento del modelo sobre las clases minoritarias, ofreciendo un enfoque práctico y eficaz en escenarios clínicos con escasez de datos.

Palabras clave: Desequilibrio de clases, imágenes médicas, aumento de datos, GANs, modelos de difusión, clasificación, Grad-CAM.

Abstract

This Final Degree Project addresses the issue of class imbalance in medical image datasets, using ocular disease classification on the ODIR-5K dataset as a case study. A structured experimental design is proposed to compare different data augmentation strategies aimed at improving the representation of underrepresented classes.

The study explores traditional augmentation techniques (rotations, flips, brightness adjustment), advanced strategies (MixUp and CutMix), and generative approaches using adversarial networks (WGAN-GP) and diffusion models (DDPM). From each technique, balanced versions of the dataset were generated and used to train classification models based on features extracted with RETFound and EfficientNetB3.

The evaluation includes quantitative metrics such as precision, recall, F1-score, and AUC; qualitative interpretation with Grad-CAM; and image quality assessment using FID and IS scores. The results demonstrate that combining data augmentation strategies significantly improves classification performance on minority classes, offering a practical and effective approach for clinical scenarios with limited data availability.

Keywords: Class imbalance, medical imaging, data augmentation, GANs, diffusion models, classification, Grad-CAM.

Índice

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

El desarrollo de sistemas de inteligencia artificial en el ámbito médico ha abierto nuevas posibilidades para mejorar la precisión y eficiencia de los diagnósticos clínicos (Wang, Weibin; Liang, Dong; Chen, Qingquing; Iwamoto, Yutaro; Han, Xian-Hua; Zhang, Qiaowei; Hu, Hongjie; Lin, Lanfen; Chen, Yen-Wei, 2019). Sin embargo, uno de los retos más persistentes en este campo es el desequilibrio de clases en los conjuntos de datos, especialmente cuando se trabaja con imágenes médicas (Shorten & Khoshgoftaar , 2019). Este Trabajo de Fin de Grado se centra en ese problema, tomando como caso de estudio un *dataset* de imágenes oculares. La escasa representación de algunas patologías dificulta que los modelos de aprendizaje profundo aprendan correctamente a identificarlas (Galdran, Carneiro, & González Ballester, 2021), lo que puede derivar en errores de diagnóstico.

Ante esta situación, se propone un enfoque basado en técnicas de *data augmentation* (Perez & Wang, 2017), que combina métodos tradicionales, avanzados y generativos para ampliar y equilibrar el conjunto de datos. El objetivo principal es mejorar el rendimiento de modelos de clasificación de imágenes médicas, evaluando el impacto de cada técnica tanto desde una perspectiva cuantitativa como cualitativa.

Se plantea la hipótesis de que el uso de imágenes sintéticas, junto con técnicas de *data augmentation* avanzadas, mejora el rendimiento en la clasificación de enfermedades oculares respecto al entrenamiento con el *dataset* original desequilibrado.

Para comprobarlo, se entrenarán clasificadores utilizando diferentes versiones del *dataset*: el original desequilibrado, y tres versiones equilibradas mediante técnicas tradicionales (Nanni, Paci, Brahnam, & Lumini, 2021), generación por StyleGAN (Dash & Swarnkar, 2025) y generación por modelos de difusión (Müller-Franzes, Gustav; Moritz Niehues, Jan; Khader, Firas; Tayebi Arasteh, Soroosh; Haarburger, Christoph; Kuhl, Christiane; Wang, Tianci; Han, Tianyu; Nebelung, Sven; Nikolas Kather, Jakob; Truhn, Daniel, 2022). Adicionalmente, se evaluará el efecto de aplicar las técnicas Mixup (Zhang, Cisse, Dauphin, & Lopez-Paz, 2018) y CutMix (Liu, Fan, Schwarz, & Maier, 2024) durante el entrenamiento sobre cada uno de estos *datasets*.

1.1 Motivación

El uso de inteligencia artificial (IA) en el diagnóstico médico ha experimentado un crecimiento notable en los últimos años (Wang, Weibin; Liang, Dong; Chen, Qingquing; Iwamoto, Yutaro; Han, Xian-Hua; Zhang, Qiaowei; Hu, Hongjie; Lin, Lanfen; Chen, Yen-Wei, 2019). La posibilidad de contar con sistemas automáticos que apoyen al personal sanitario en la detección temprana de enfermedades representa un avance importante, tanto en términos de

calidad asistencial como de eficiencia (Goceri, 2023). Sin embargo, en la práctica, el entrenamiento de estos sistemas se ve limitado por la falta de datos suficientes y equilibrados, sobre todo en patologías poco frecuentes (Shorten & Khoshgoftaar , 2019).

En el caso de las enfermedades oculares, este problema es especialmente visible. Muchas condiciones graves no cuentan con un número adecuado de muestras que permita a los modelos de IA aprender a identificarlas correctamente. A nivel personal, este proyecto surge del interés por encontrar soluciones técnicas que ayuden a mejorar la calidad de estos sistemas, haciendo que sean más robustos y accesibles, incluso en contextos con recursos limitados.

Aplicar técnicas de aumento de datos y generación de imágenes sintéticas no solo permite equilibrar los *datasets*, sino que también abre la puerta a nuevas formas de entrenamiento más eficientes. Explorar la combinación de estas técnicas y evaluar su impacto real en modelos clínicos es una oportunidad para aportar valor tanto desde el ámbito técnico como desde el punto de vista de la aplicación práctica en medicina.

1.2 Justificación

Desde una perspectiva técnica, este trabajo aborda uno de los desafíos más relevantes en la aplicación del aprendizaje profundo a la medicina: cómo mejorar el rendimiento de los modelos cuando no se dispone de datos suficientes o equilibrados (Nanni, Paci, Brahnam, & Lumini, 2021). Las técnicas de aumento de datos permiten multiplicar el número de ejemplos disponibles, y la incorporación de imágenes sintéticas generadas por modelos como redes adversarias generativas (GANs) o modelos de difusión ofrece un enfoque innovador con gran potencial (Dash & Swarnkar, 2025) (Müller-Franzes, Gustav; Moritz Niehues, Jan; Khader, Firas; Tayebi Arasteh, Soroosh; Haarburger, Christoph; Kuhl, Christiane; Wang, Tianci; Han, Tianyu; Nebelung, Sven; Nikolas Kather, Jakob; Truhn, Daniel, 2022).

En el plano clínico, los beneficios también son evidentes. Un sistema de diagnóstico basado en IA que sea capaz de identificar correctamente enfermedades minoritarias puede tener un gran impacto, especialmente en contextos donde no siempre se dispone de especialistas. Además, reducir el sesgo presente en los modelos al entrenarlos con datos más diversos contribuye a ofrecer una atención médica más equitativa.

Este proyecto también incluirá una evaluación exhaustiva de la calidad de las imágenes generadas, combinando métricas objetivas como FID (*Fréchet Inception Distance*) e IS (*Inception Score*) (Müller-Franzes, Gustav; Moritz Niehues, Jan; Khader, Firas; Tayebi Arasteh, Soroosh; Haarburger, Christoph; Kuhl, Christiane; Wang, Tianci; Han, Tianyu; Nebelung, Sven; Nikolas Kather, Jakob; Truhn, Daniel, 2022). El objetivo es asegurar que las imágenes no solo sean técnicamente válidas, sino que también tengan utilidad real en escenarios clínicos.

1.3 Objetivos

El proyecto cuenta con los siguientes objetivos:

Objetivo general:

El objetivo principal de TFG es desarrollar un estudio comparativo que evalúe el impacto de distintas técnicas de aumento de datos —tradicionales, avanzadas y generativas— en la mejora del rendimiento de modelos de clasificación aplicados a un conjunto de datos desequilibrado de imágenes médicas oculares.

Para ello se pretenden conseguir los siguientes **objetivos específicos**:

1. Implementar transformaciones tradicionales (rotaciones, cambios de brillo, escalado, ruido, etc.) para aumentar de forma controlada la diversidad de las clases minoritarias en el *dataset.*

2. Aplicar técnicas avanzadas como MixUp y CutMix, que permiten generar ejemplos más representativos combinando imágenes y etiquetas, y evaluar su efecto sobre los *embeddings* extraídos.
3. Generar imágenes sintéticas mediante redes generativas adversarias (GANs) y modelos de difusión (DDPM), centradas en mejorar la representación de las clases menos frecuentes.
4. Evaluar el impacto de cada técnica de aumento sobre el rendimiento de modelos de clasificación, utilizando arquitecturas como RETFound (Zhang, Juzhao; Lin, Senlin; Cheng, Tianhao; Xu, Yi; Lu, Lina; He, Jiangnan; Yu, Tao; Peng, Yajun; Zou, Haidong; Ma, Yingyan, 2024) y EfficientNetB3 (Tan & Le, 2019), y métricas como precisión, sensibilidad, *F1-score* y AUC-ROC.
5. Analizar la calidad visual y estadística de las imágenes sintéticas generadas mediante métricas cuantitativas como FID (*Fréchet Inception Distance*) e IS (*Inception Score*).

1.4 Impacto esperado

Este Trabajo de Fin de Grado se circunscribe al estudio del desequilibrio de clases en conjuntos de datos médicos, tomando como caso de uso el diagnóstico de patologías oculares mediante imágenes de fondo de ojo. El alcance del proyecto incluye el diseño y evaluación de un *pipeline* completo de clasificación, desde el preprocesamiento del *dataset* hasta la comparación de modelos entrenados con distintas estrategias de aumento de datos: tradicionales, avanzadas (MixUp y CutMix) (Zhang, Cisse, Dauphin, & Lopez-Paz, 2018) (Liu, Fan, Schwarz, & Maier, 2024) y generativas (GANs y modelos de difusión) (Dash & Swarnkar, 2025) (Müller-Franzes, Gustav; Moritz Niehues, Jan; Khader, Firas; Tayebi Arasteh, Soroosh; Haarburger, Christoph; Kuhl, Christiane; Wang, Tianci; Han, Tianyu; Nebelung, Sven; Nikolas Kather, Jakob; Truhn, Daniel, 2022). Asimismo, se contempla tanto la evaluación cuantitativa como cualitativa de los resultados, integrando técnicas de interpretabilidad como Grad-CAM (Selvaraju, y otros, 2017).

No obstante, el proyecto presenta una serie de limitaciones. En primer lugar, los modelos generativos fueron entrenados con recursos computacionales limitados, lo que condicionó la resolución y calidad de las imágenes sintéticas, especialmente en el caso de los modelos de difusión. En segundo lugar, la validación clínica de las imágenes generadas no fue posible, por

lo que su utilidad diagnóstica se ha evaluado desde un punto de vista computacional y no experto. Por último, el estudio se restringe a un enfoque de clasificación *single-label* (Galdran, Carneiro, & González Ballester, 2021), pese a que en la práctica clínica es común encontrar pacientes con múltiples patologías oculares concurrentes.

A pesar de estas limitaciones, se espera que los resultados obtenidos tengan un impacto positivo tanto en el plano técnico como en el clínico. Desde el punto de vista de la ingeniería de datos, se demuestra que la combinación estratégica de técnicas de aumento puede mejorar el rendimiento de modelos de clasificación en escenarios con escasez de datos. Desde una perspectiva aplicada, este trabajo contribuye a avanzar hacia soluciones más robustas y equitativas en inteligencia artificial médica, facilitando el entrenamiento de modelos más sensibles a patologías poco representadas sin necesidad de grandes volúmenes de datos reales.

1.5 Estructura de la memoria

Este Trabajo de Fin de Grado se estructura en siete capítulos que reflejan el desarrollo progresivo del proyecto:

- **Capítulo 1. Introducción**: Se presenta el contexto clínico del problema, la motivación personal y técnica, la hipótesis de trabajo, los objetivos del proyecto y la organización general del documento.
- **Capítulo 2. Estado del arte**: Se realiza una revisión crítica de las principales técnicas de aumento de datos (tradicionales, avanzadas y generativas) aplicadas en el ámbito de la imagen médica, con especial énfasis en oftalmología. También se revisan modelos de clasificación comunes y su integración en flujos clínicos asistidos por IA.
- **Capítulo 3. Metodología**: Se describe el análisis exploratorio del *dataset* ODIR-5K, el preprocesamiento y la creación de versiones equilibradas mediante técnicas tradicionales, GANs y modelos de difusión. También se detallan los modelos de clasificación desarrollados, las herramientas utilizadas y las métricas de evaluación aplicadas.
- **Capítulo 4. Resultados, evaluación y discusión**: Se presentan los resultados obtenidos en los distintos bloques experimentales (*dataset* desequilibrado, aumento tradicional, imágenes GAN, imágenes de difusión), se comparan sus métricas de rendimiento, y se discuten los hallazgos más relevantes, incluyendo el efecto de MixUp y CutMix sobre los *embeddings*. Además, se muestran visualizaciones de activaciones obtenidas con Grad-CAM sobre los modelos mejor entrenados en cada escenario, destacando patrones de atención clínica y errores de clasificación.
- **Capítulo 5. Conclusiones y trabajo futuro**: Se resumen los principales hallazgos del estudio, se valora el cumplimiento de los objetivos, se discuten las limitaciones técnicas encontradas y se proponen futuras líneas de investigación y mejora.

Capítulo 2

Estado del arte

El objetivo de este estado del arte es situar esta investigación dentro del panorama actual de la generación de imágenes sintéticas y las técnicas de *data augmentation* aplicadas a imágenes médicas, con especial atención al campo de la oftalmología. Se trata de entender cómo se han desarrollado estas metodologías y qué papel juegan a la hora de mejorar el rendimiento de los sistemas de diagnóstico basados en inteligencia artificial.

Para ello, se revisarán tres líneas principales. Primero, las técnicas clásicas de aumento de datos, como las rotaciones, los cambios de brillo o la adición de ruido, ampliamente utilizadas para enriquecer la variedad de imágenes disponibles. Luego, se abordarán enfoques más recientes como Mixup y CutMix, que proponen maneras innovadoras de combinar imágenes para mejorar la capacidad de generalización de los modelos. Finalmente, se explorará el impacto de los modelos generativos —como las GANs y los modelos de difusión— en la creación de imágenes sintéticas realistas, una solución prometedora frente al problema del desequilibrio de clases en *datasets* médicos.

Estos tres enfoques nos ofrecen una visión completa sobre el estado actual de las técnicas de aumento de datos y su relevancia en el contexto clínico.

2.1 Generalidades sobre aprendizaje automático y el problema del desequilibrio de clases en medicina

Antes de entrar en detalle sobre la metodología utilizada, resulta fundamental situar esta investigación dentro de un marco teórico y metodológico que ayude a comprender mejor su enfoque. Este capítulo sirve como base conceptual para entender por qué es necesario abordar el problema del desequilibrio en los conjuntos de datos médicos y cómo las técnicas de aumento de datos y generación de imágenes sintéticas pueden ofrecer soluciones innovadoras, especialmente en el contexto de la oftalmología.

En los últimos años, el aprendizaje automático —y en particular el aprendizaje profundo— se ha consolidado como una herramienta poderosa en el ámbito de la medicina (Wang, Weibin;

Liang, Dong; Chen, Qingquing; Iwamoto, Yutaro; Han, Xian-Hua; Zhang, Qiaowei; Hu, Hongjie; Lin, Lanfen; Chen, Yen-Wei, 2019). Su capacidad para identificar patrones complejos en grandes volúmenes de datos ha permitido avances significativos en tareas como la clasificación de imágenes médicas, el diagnóstico precoz o el seguimiento de enfermedades (Wang, Weibin; Liang, Dong; Chen, Qingquing; Iwamoto, Yutaro; Han, Xian-Hua; Zhang, Qiaowei; Hu, Hongjie; Lin, Lanfen; Chen, Yen-Wei, 2019).

Sin embargo, la eficacia de estos modelos depende en gran medida de la calidad y cantidad de los datos con los que se entrenan. Cuando los datos son escasos o están desequilibrados, como ocurre frecuentemente en medicina, el rendimiento del modelo puede verse comprometido, sobre todo a la hora de reconocer condiciones menos comunes (Wang, y otros, 2019).

Una de las limitaciones más habituales al trabajar con imágenes médicas es la falta de equilibrio entre clases. Algunas enfermedades aparecen con mucha menos frecuencia que otras, lo que se traduce en conjuntos de datos donde ciertos diagnósticos están ampliamente representados mientras que otros apenas cuentan con ejemplos. Esta desproporción complica la tarea del modelo, que tiende a aprender patrones de las clases mayoritarias y pasa por alto las minoritarias. El riesgo, por tanto, no es solo técnico, sino clínico: se podrían dejar de detectar condiciones relevantes.

Por ello, distintas investigaciones han subrayado la necesidad de aplicar técnicas que ayuden a compensar esta desigualdad en los datos (Shorten & Khoshgoftaar , 2019).

2.2 Técnicas tradicionales

Las técnicas de *data augmentation* más convencionales han sido durante años una solución eficaz para aumentar la diversidad de los datos médicos disponibles y, con ello, mejorar el rendimiento de los modelos de aprendizaje profundo (Shorten & Khoshgoftaar , 2019). Estas técnicas consisten en aplicar transformaciones simples, tanto geométricas como fotométricas, sobre las imágenes originales, simulando así distintas condiciones reales.

Por ejemplo, rotar o voltear imágenes permite entrenar modelos que sean más robustos frente a variaciones en la orientación de los órganos o tejidos analizados. En el caso de las imágenes de fondo de ojo, esto ha demostrado ser útil para detectar mejor las enfermedades como la retinopatía diabética (Zhang, Xie, Xing, McGough, & Yang, 2017). Ajustes de brillo y contraste también son habituales, ya que simulan distintas condiciones de iluminación, algo muy relevante en imágenes como radiografías o fotografías del fondo ocular (Shorten & Khoshgoftaar , 2019).

Otras técnicas incluyen la introducción de ruido —como el gaussiano— para emular imperfecciones propias de dispositivos de menor calidad, o el escalado y recorte de imágenes para generar variaciones en tamaño y posición de las estructuras anatómicas (Perez & Wang, 2017). Además, en imágenes a color como las oftalmológicas, se realizan ajustes de saturación y tono para simular diferencias en las características del tejido. Estas transformaciones, aunque sencillas, siguen siendo un pilar importante para mejorar la capacidad de los modelos de adaptarse a nuevas condiciones.

2.3 Técnicas avanzadas

Con la evolución de los modelos de aprendizaje profundo, han surgido técnicas de *data augmentation* más sofisticadas que van más allá de las transformaciones simples. Mixup y CutMix son dos ejemplos destacados, especialmente útiles cuando se trabaja con *datasets* limitados o con un fuerte desequilibrio entre clases.

Mixup, propuesto en (Zhang, Cisse, Dauphin, & Lopez-Paz, 2018), combina dos imágenes y sus etiquetas correspondientes para generar una nueva imagen "intermedia". En el ámbito médico, esta técnica ha sido adaptada por (Galdran, Carneiro, & González Ballester, 2021) en su versión **Balanced-Mixup**, pensada específicamente para dar mayor visibilidad a las clases menos representadas. Los resultados, aplicados a imágenes de retina y gastroenterología, mostraron mejoras claras en la clasificación de enfermedades poco frecuentes y una reducción del sobreajuste.

CutMix, por su parte, funciona de una manera diferente: reemplaza una región concreta de una imagen con un fragmento de otra, ajustando la etiqueta en función del área modificada (Dong & Yang, 2019). Esta técnica ha sido evaluada en tareas como la segmentación de órganos, mostrando mejoras notables, incluso cuando las imágenes resultantes no eran visualmente coherentes (Liu, Fan, Schwarz, & Maier, 2024) y recientemente en problemas de clasificación *multilabel* en medicina, donde ha mejorado tanto la precisión como la capacidad de localización anatómica (Liu, Fan, Schwarz, & Maier, 2024). Lo importante aquí no es tanto la estética, como la diversidad de información que se aporta al modelo durante el entrenamiento.

Estudios recientes han confirmado que técnicas de combinación como Mixup y CutMix resultan efectivas incluso en conjuntos de datos médicos pequeños o con etiquetado *multilabel,* permitiendo un mejor aprendizaje de características relevantes en imágenes complejas (Galdran, Carneiro, & González Ballester, 2021) (Liu, Fan, Schwarz, & Maier, 2024) (Chen, Yu, Feng, Chen, & Wu, 2021).

No obstante, presentan ciertos retos: Mixup puede dificultar la interpretación visual de las imágenes generadas, mientras que CutMix puede introducir elementos irreales que compliquen el aprendizaje de detalles anatómicos específicos. La Tabla 1 resume las ventajas e inconvenientes de estas dos técnicas.

Técnica	Ventajas	Inconvenientes
Mixup	Regulariza el modelo, mejora la generalización	Imágenes difíciles de interpretar
CutMix	Aumenta diversidad local de características	Introduce incoherencias anatómicas

Tabla 1. *Resumen de técnicas avanzadas de aumento de datos.*

2.4 Generación de imágenes sintéticas

Una línea de investigación cada vez más relevante es el uso de modelos generativos para crear imágenes médicas sintéticas de alta calidad. Estos modelos han demostrado un gran potencial para ampliar *datasets,* especialmente cuando los datos reales son escasos o desequilibrados.

Las **Redes Generativas Antagónicas (GANs)**, introducidas en (Goodfellow, y otros, 2014), funcionan mediante un juego entre dos redes: una que genera imágenes y otra que intenta distinguirlas de las reales. En el campo médico, las GANs se han aplicado con éxito a tareas como la segmentación de retina o la detección de nódulos pulmonares (Dash & Swarnkar, 2025), ayudando a crear conjuntos de datos más equilibrados y diversos.

Más recientemente, los **modelos de difusión**, como los DDPMs, han ganado popularidad por su capacidad de generar imágenes aún más fieles y variadas mediante un proceso de eliminación progresiva de ruido. (Müller-Franzes, Gustav; Moritz Niehues, Jan; Khader, Firas; Tayebi Arasteh, Soroosh; Haarburger, Christoph; Kuhl, Christiane; Wang, Tianci; Han, Tianyu; Nebelung, Sven; Nikolas Kather, Jakob; Truhn, Daniel, 2022) presentaron el modelo Medfusion, que superó a las GANs en múltiples métricas —como el FID— al generar imágenes de fondo de ojo y radiografías más realistas.

Para medir la calidad de estas imágenes sintéticas, se utilizan métricas como:

- **FID (Frechet Inception Distance)** (Heusel, Ramsauer, Unterthiner, Nessler, & Hochreiter, 2018): mide la similitud estadística entre las imágenes reales y las sintéticas. Cuanto más bajo, mayor similitud con las imágenes reales.
- **IS (Inception Score)** (Salimans, y otros, 2016): evalúa la diversidad y fidelidad de las imágenes generadas. Cuanto más alto, mayor calidad y diversidad.

Los modelos de difusión, según los estudios revisados, ofrecen resultados prometedores en ambas métricas, consolidándose como una alternativa sólida a las GANs. La Tabla 2 resume las ventajas e inconvenientes de estas dos técnicas.

Modelo	Ventajas	Desventajas
GANs	Rápidos de entrenar, generación eficiente	Artefactos, riesgo de *mode collapse*
Modelos de difusión	Alta calidad y realismo en imágenes	Entrenamiento mucho más costoso

Tabla 2. *Comparativa crítica entre GANs y modelos de difusión.*

2.5 Combinación de técnicas

Una tendencia emergente es la combinación de diferentes enfoques de *data augmentation* para aprovechar lo mejor de cada uno (Dash & Swarnkar, 2025). Integrar transformaciones

tradicionales, técnicas avanzadas como Mixup o CutMix, y modelos generativos como GANs o modelos de difusión puede generar *datasets* más robustos y variados.

En este sentido, (Dash & Swarnkar, 2025) demostraron que combinar rotaciones, Mixup y GANs mejoró notablemente la precisión en tareas de clasificación de imágenes médicas. Por su parte, (Goceri, 2023) probó diversas combinaciones en imágenes de resonancia magnética, tomografías y mamografías, confirmando que esta estrategia no solo incrementa la diversidad del *dataset*, sino también la resiliencia de los modelos frente a ruido o desequilibrio.

También evidenciaron que estas combinaciones, aplicadas a modelos como ResNet50 (Nanni, Paci, Brahnam, & Lumini, 2021), aumentan la precisión hasta un 20% y reducen significativamente el sobreajuste, especialmente en conjuntos de datos pequeños.

2.6 Limitaciones y desafíos

A pesar de todos estos avances, todavía existen importantes retos por resolver en el uso de técnicas de *data augmentation* y generación de imágenes sintéticas en medicina.

Uno de los principales desafíos es la **calidad de las imágenes generadas**. Las GANs, por ejemplo, pueden producir artefactos visuales, mientras que los modelos de difusión, aunque más precisos, son computacionalmente más costosos (Dash & Swarnkar, 2025). Esto nos lleva al segundo reto: los **altos requerimientos computacionales**, que pueden ser una barrera en entornos clínicos donde el acceso a GPUs potentes es limitado (Goceri, 2023).

También hay que tener en cuenta la **generalización de los modelos**: las imágenes generadas pueden no cubrir adecuadamente toda la diversidad de los casos reales, y su **validación clínica** aún es escasa (Nanni, Paci, Brahnam, & Lumini, 2021). Además, el **problema del desequilibrio** de clases sigue sin resolverse del todo, ya que las técnicas actuales no siempre logran generar suficientes ejemplos representativos para las clases minoritarias.

2.7 Síntesis del estado del arte

En resumen, las técnicas de *data augmentation* —desde las más tradicionales hasta las más avanzadas y generativas— han evolucionado rápidamente y han demostrado su utilidad en el campo de la clasificación de imágenes médicas. Sin embargo, persisten limitaciones que dificultan su adopción plena en entornos clínicos.

Este Trabajo de Fin de Grado propone abordar esas limitaciones mediante un enfoque combinado, que integre lo mejor de cada técnica y explore configuraciones más eficientes para mejorar la clasificación en *datasets* médicos desequilibrados. La evaluación rigurosa de la calidad de las imágenes generadas y la búsqueda de soluciones adaptables a contextos con recursos limitados son pasos clave para avanzar hacia un diagnóstico asistido por inteligencia artificial más robusto, accesible y fiable.

Capítulo 3

Metodología

Este capítulo presenta de forma estructurada la metodología empleada para abordar el problema del desequilibrio de clases en un conjunto de imágenes médicas oculares. Se comienza con un análisis exploratorio del *dataset* ODIR-5K, donde se examina tanto la distribución de clases como las características demográficas de los pacientes. A continuación, se detalla el proceso de creación de distintos *datasets*: el original, una versión aumentada con técnicas tradicionales, y dos versiones equilibradas mediante generación de imágenes sintéticas utilizando GANs (Goodfellow, y otros, 2014) y modelos de difusión (Müller-Franzes, Gustav; Moritz Niehues, Jan; Khader, Firas; Tayebi Arasteh, Soroosh; Haarburger, Christoph; Kuhl, Christiane; Wang, Tianci; Han, Tianyu; Nebelung, Sven; Nikolas Kather, Jakob; Truhn, Daniel, 2022). Posteriormente, se describen los modelos de clasificación desarrollados para cada conjunto de datos, incluyendo variantes con Mixup (Zhang, Cisse, Dauphin, & Lopez-Paz, 2018) y CutMix (Liu, Fan, Schwarz, & Maier, 2024), así como las herramientas utilizadas para su implementación. Finalmente, se explican las métricas y métodos de evaluación aplicados, tanto cuantitativos como cualitativos, que permiten analizar el impacto de cada técnica en el rendimiento del modelo.

El punto de partida ha sido el *dataset* **ODIR-5K**, que incluye imágenes del fondo de ojo de pacientes con distintas patologías oculares. En total, hay unas **7000 imágenes**, divididas a partes iguales entre ojos izquierdos y derechos. Cada imagen va asociada, en teoría, a cierta información diagnóstica que se recoge en un archivo CSV original '`data.xlsx`'.

Este archivo contiene 3500 **registros**; cada registro corresponde a una persona con su ojo izquierdo y derecho con anotaciones completas.

Cada imagen puede estar etiquetada con una o varias de las siguientes enfermedades oculares, las cuales se explican brevemente para facilitar su interpretación a lo largo del estudio:

- **Normal (N)**: ojo sano.

- **Diabetes (D)**: la retinopatía diabética daña los vasos de la retina. Detectable en imágenes por signos como micro aneurismas o hemorragias.
- **Glaucoma (G):** daño al nervio óptico, muchas veces sin síntomas. Detectable por cambios en el nervio visible en el fondo de ojo.
- **Cataract (C)**: las cataratas provocan opacidad del cristalino que reduce la visión. Puede reflejarse indirectamente en imágenes retinianas.
- **AMD (A)**: la degeneración macular asociada a la edad afecta la mácula y causa pérdida de visión central. Visible por depósitos y alteraciones en la retina.
- **Hypertension (H)**: la hipertensión ocular puede alterar los vasos sanguíneos del ojo. Detectable por cambios en la vasculatura retiniana.
- **Myopia (M)**: la miopía afecta la visión lejana. En casos altos, puede provocar cambios estructurales en la retina visibles en imágenes.
- **Other (O)**: otras patologías no especificadas.

Una de las primeras cosas que llamó la atención al analizar el *dataset* fue que hay **un desequilibrio bastante claro entre clases**. Las categorías "Normal" y "Diabetes" son muy abundantes, mientras que otras, como "Hypertension" o "Myopia", tienen muchas menos imágenes. Además, existen **unos 600 casos con múltiples etiquetas**, lo que implica que algunos pacientes presentan más de una patología a la vez. Esto complica un poco el problema, ya que si mantenemos a los pacientes con múltiples patologías obligaría a tratarlo como una clasificación *multilabel*.

3.1 Análisis exploratorio

Antes de aplicar cualquier técnica de procesamiento o entrenamiento de modelos, se realizó un análisis exploratorio del conjunto de datos con el objetivo de comprender su estructura, distribución y posibles sesgos. Este análisis permite identificar desequilibrios en las clases, variabilidad en las características demográficas y otros factores relevantes que pueden influir en el rendimiento del modelo.

3.1.1 Distribución de clases

Uno de los primeros aspectos analizados fue la distribución de las diversas clases que representan diferentes patologías oculares. Como se muestra en la Figura 1, el conjunto de datos presenta un notable desequilibrio: las clases "Normal" (N) y "Diabetes" (D) son, con diferencia, las más abundantes, superando ambas las 1900 imágenes. En contraste, categorías como "Hipertensión" (H), "Degeneración macular relacionada con la edad" (A) y "Miopía" (M) cuentan con menos de 280 imágenes, lo que evidencia una menor representación de estas condiciones en el conjunto.

Es importante destacar que el gráfico refleja un total de **7327 muestras**, ya que algunas imágenes están asociadas a múltiples etiquetas de diagnóstico. Esto implica que una misma

imagen puede aparecer contabilizada en varias clases distintas, una característica típica en problemas de **clasificación *multilabel***.

Este desequilibrio que se muestra en la Figura 1 podría llevar a que un modelo de clasificación se enfoque en las clases mayoritarias, afectando negativamente su rendimiento en el reconocimiento de enfermedades menos comunes. Por lo tanto, esta razón subraya la importancia de incorporar técnicas de aumento de datos y generación de datos sintéticos dentro de la metodología, con el fin de mejorar la representatividad y robustez del modelo.

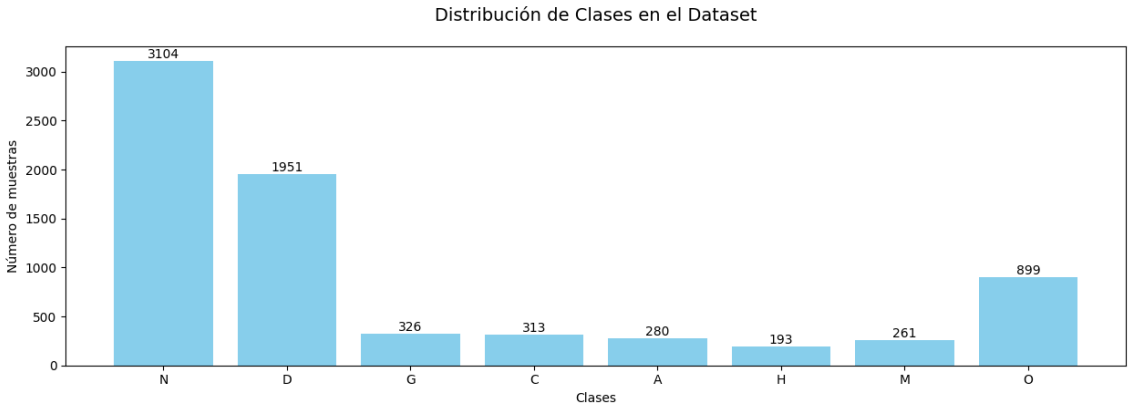

Figura 1. *Distribución de clases en el conjunto de datos.*

.

Para acompañar el análisis de las clases, se muestran en la Figura 2 algunas imágenes representativas de cada categoría diagnóstica del *dataset*. Estos ejemplos visuales ayudan a entender mejor las diferencias entre unas patologías y otras, y dan una idea más clara de los retos que supone para el modelo aprender a distinguirlas correctamente.

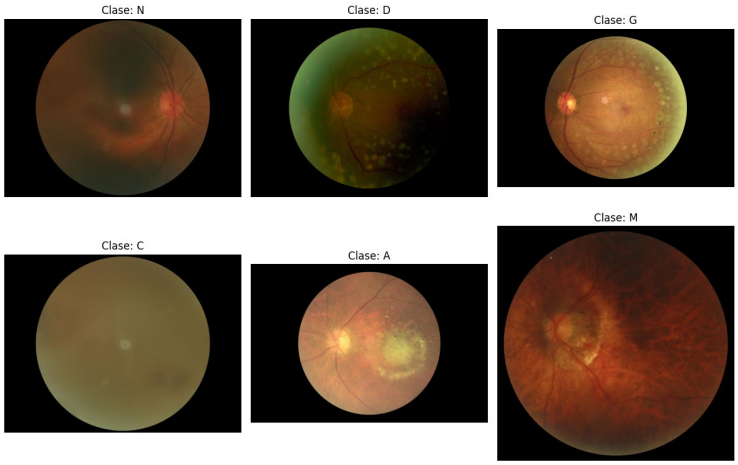

Figura 2. *Ejemplos visuales de cada una de las enfermedades.*

3.1.2 Características demográficas

Además de la distribución por clases, se exploraron dos variables demográficas: la edad y el sexo de los pacientes.

- **Edad**: La edad media de los pacientes ronda los 57.9 años, con valores que oscilan entre 1 y 91 años. El histograma correspondiente (Figura 3) muestra una mayor concentración en el rango de 50 a 70 años, lo cual coincide con el perfil habitual de aparición de muchas de las patologías oculares estudiadas.

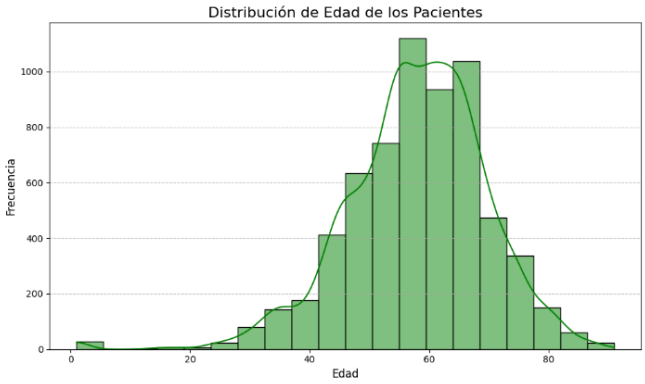

Figura 3. *Distribución por edad de los pacientes.*

- **Sexo**: En lo que respecta al sexo, el *dataset* se encuentra relativamente equilibrado, un 53.6% de pacientes son hombres y un 46.4% son mujeres. Esta proporción, reflejada en la Figura 4, sugiere que no existe un sesgo de género significativo en la muestra analizada.

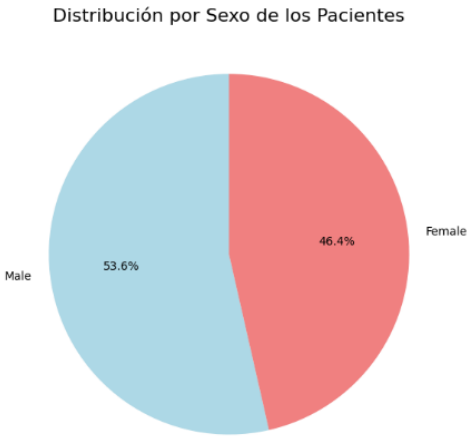

Figura 4. *Distribución por sexo de los pacientes.*

3.2 Creación de *datasets*

Antes de aplicar las distintas técnicas de aumento de datos y generación de imágenes sintéticas, fue necesario preparar cuidadosamente el conjunto de datos original. Esta sección

describe tanto las modificaciones realizadas sobre el *dataset* original para su mejor utilización, como la generación posterior de nuevas versiones equilibradas del mismo.

El objetivo es establecer una base sólida y coherente desde la cual desarrollar de forma justa todos los experimentos comparativos planteados en este proyecto.

3.2.1 *Dataset* original

El *dataset* original de ODIR-5K presentaba ciertas características que dificultaban su uso directo en los experimentos, como la presencia de entradas *multilabel* o una organización por paciente en lugar de por imagen.

Por este motivo, se llevó a cabo un proceso de preprocesamiento progresivo, cuyo objetivo fue adaptar el conjunto de datos a un formato más adecuado para la tarea de clasificación planteada.

A continuación, se describen las principales transformaciones realizadas hasta obtener el *dataset* final utilizado como punto de partida en todos los análisis.

El preprocesamiento ha sido una etapa clave para asegurar la calidad y coherencia del conjunto de datos empleado en este trabajo. A continuación, se detalla, de forma progresiva, cómo se ha transformado el *dataset* original hasta llegar al conjunto final utilizado en los experimentos:

1. **Adición de rutas completas a las imágenes**: Partiendo del *dataset* inicial (`data.xlsx`), se incorporó en las columnas `left-fundus` y `right-fundus` la ruta completa de los archivos de imagen correspondientes. Esta modificación permitió un acceso más directo y automatizado a las imágenes durante las siguientes fases de procesamiento y entrenamiento.

2. **Separación de imágenes por ojo**: En la siguiente fase, se reorganizó el *dataset* para que cada fila representase una única imagen de un solo ojo, en lugar de agrupar ambos ojos de un paciente en una misma entrada. Así, cada paciente pasó a estar representado por dos filas, una para el ojo izquierdo y otra para el derecho. Esta reestructuración no solo simplificó la estructura del conjunto de datos, sino que también duplicó el número de instancias disponibles.

3. **Eliminación de imágenes *multilabel***: Con el objetivo de plantear el problema como una clasificación de una única etiqueta por imagen (*single-label classification*), se eliminaron todas las muestras que presentaban más de un diagnóstico simultáneo. Con esta medida se buscó evitar ambigüedades en la asignación de clases, asegurando que cada imagen correspondiera claramente a un único diagnóstico. Esta decisión se tomó para simplificar el problema a una clasificación de una única patología por imagen, facilitando la comparación controlada entre los diferentes métodos de equilibrado. Tras eliminar las muestras *multi-label* quedaron 6699 muestras.

4. **Exclusión de la categoría hipertensión**: Durante el análisis exploratorio se detectó que todas las imágenes etiquetadas como hipertensión (H) aparecían siempre acompañadas de otras patologías. Al aplicar el filtrado, se comprobó que no quedaban imágenes asociadas únicamente a la categoría de hipertensión. Por este motivo, se optó por eliminarla como clase independiente, garantizando así la coherencia metodológica y evitando posibles inconsistencias en los experimentos posteriores.

5. **Exclusión de la categoría otras enfermedades**: Durante el análisis exploratorio, se detectó que la categoría "otras enfermedades" (O) agrupaba una serie de patologías que resultan heterogéneas y que además clínicamente no estaban especificadas. De este modo, resultaba difícil su interpretación y análisis en el marco de este trabajo, cuyo objetivo es evaluar el impacto del aumento de datos en clases bien definidas desde el punto de vista diagnóstico. Además, con las técnicas de aumento de datos se iban a aumentar el número de muestras de las enfermedades minoritarias correctamente definidas, entonces esta categoría se iba a quedar con un número de muestras muy reducido convirtiéndose en clase minoritaria. Por estas razones, se decidió excluir esta categoría del conjunto de datos final. Esta decisión permitió centrar el análisis en patologías claramente identificables y clínicamente relevantes, asegurando mayor consistencia metodológica y facilitando la evaluación comparativa de las técnicas de aumento de datos utilizadas. Tras todos los procesos descritos previamente la distribución resultante de las clases es la que se muestra en la Figura 5.

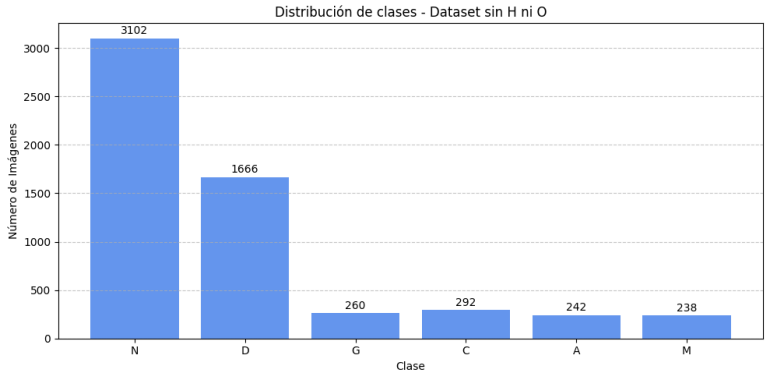

Figura 5. *Distribución de clases tras eliminar muestras con múltiples etiquetas, muestras de la categoría hipertensión y la categoría otras enfermedades.*

El conjunto de datos final tras el preprocesamiento consta de 5800 imágenes e incluye únicamente imágenes pertenecientes a las siguientes categorías:

- Normal
- Diabetes
- Glaucoma
- Cataratas
- Degeneración macular asociada a la edad (AMD)
- Miopía

Fase	**Descripción**
Adición de rutas	Inclusión de la ruta completa de las imágenes

Separación de ojos	Una fila por cada imagen de ojo (izquierdo o derecho)
Eliminación *multilabel*	Exclusión de imágenes con múltiples etiquetas
Eliminación 'hipertensión'	Supresión de la categoría H tras el filtrado
Eliminación de 'otras enfermedades'	Supresión de la categoría O tras el filtrado

***Tabla 3**. Resumen flujo de preprocesamiento.*

Este proceso de preprocesamiento resumido en la Tabla 3 ha sido esencial para garantizar la validez y robustez de los experimentos posteriores. Así, el conjunto de datos final está completamente preparado para aplicar las distintas estrategias de equilibrado y entrenamiento que se detallarán a continuación.

3.2.2 *Dataset* con técnicas tradicionales

Para abordar el problema del desequilibrio de clases en nuestro conjunto de datos, primero se implementó una estrategia de aumento de datos utilizando técnicas tradicionales de transformación de imágenes. El objetivo principal fue incrementar el número de muestras en las clases minoritarias (Glaucoma, Cataratas, AMD y Miopía) hasta alcanzar 2000 imágenes por clase, manteniendo la validez clínica de las imágenes generadas.

Proceso de aumento de datos

Se desarrolló un *pipeline* de aumento de datos que aplica múltiples transformaciones de manera aleatoria y controlada a las imágenes originales. Las transformaciones implementadas incluyen:

1. **Rotaciones geométricas**: Se aplicaron rotaciones de 90, 180 y 270 grados para crear diferentes perspectivas de la misma imagen, manteniendo la información diagnóstica relevante.
2. **Volteos (*flips*)**: Se realizaron reflexiones tanto horizontales como verticales de manera aleatoria, generando variaciones anatómicamente plausibles de las estructuras oculares.
3. **Ajustes de brillo y contraste**: Se modificaron los niveles de brillo (β entre -30 y +30) y contraste (α entre 0.8 y 1.2) para simular diferentes condiciones de iluminación que pueden presentarse en entornos clínicos reales.
4. **Traslaciones**: Se implementaron desplazamientos aleatorios en los ejes X e Y (±50 píxeles) para variar la posición del fondo ocular en la imagen.
5. **Zoom**: Se aplicaron transformaciones de escala (entre 0.8x y 1.2x) para simular diferentes niveles de acercamiento en la captura de las imágenes.
6. **Ruido gaussiano**: Se añadió ruido gaussiano de manera controlada para aumentar la robustez del modelo frente a variaciones en la calidad de imagen.

Resultados del aumento de datos

El proceso de aumento de datos generó los siguientes resultados mostrados en la Figura 6 para cada clase minoritaria:

- **Glaucoma (G)**: de 260 imágenes originales a 2000 (1740 generadas)
- **Cataratas (C)**: de 292 a 2000 (1708 generadas)
- **Degeneración macular asociada a la edad (A)**: de 242 a 2000 (1758 generadas)
- **Miopía (M)**: de 238 a 2000 (1762 generadas)

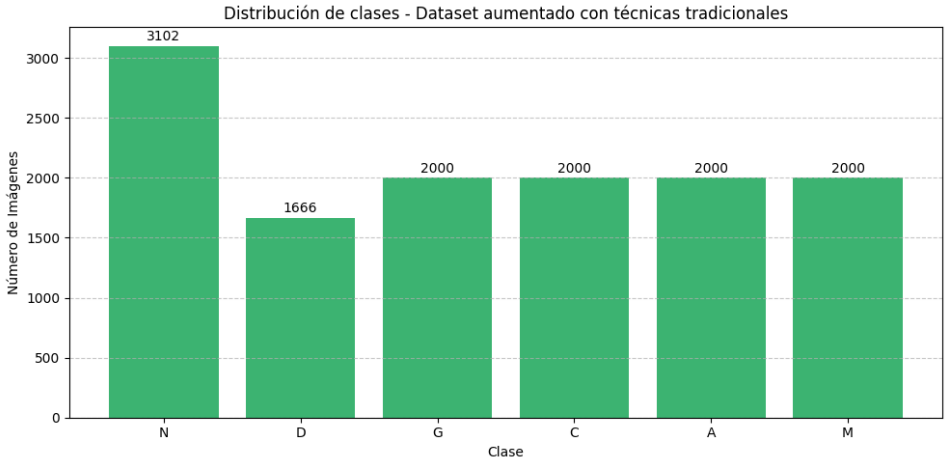

Figura 6. *Distribución de clases tras el aumento de las minoritarias con técnicas tradicionales.*

Consideraciones técnicas

Las transformaciones se aplicaron de manera combinada y aleatoria para maximizar la variabilidad de las imágenes generadas, evitando la creación de duplicados exactos. Cada imagen generada mantiene las etiquetas de diagnóstico de su imagen original, garantizando la consistencia en la información clínica.

Para preservar la calidad y utilidad diagnóstica de las imágenes, los parámetros de las transformaciones se ajustaron cuidadosamente dentro de rangos que mantienen las características patológicas relevantes. Todas las imágenes generadas se almacenaron con la misma resolución y formato que las originales, asegurando la compatibilidad con el *pipeline* de procesamiento existente.

Este proceso de aumento de datos ha permitido crear un conjunto de datos más equilibrado, manteniendo la integridad de la información médica y proporcionando una base más robusta para el entrenamiento del modelo de clasificación.

3.2.3 *Dataset* con GAN

Con el objetivo de generar una versión alternativa y más realista del *dataset* desequilibrado, se entrenaron modelos generativos de tipo WGAN-GP (*Wasserstein GAN con penalización de gradiente*) (Gulrajani, Arjovsky, Ahmed, Dumoulin, & Courville, 2017) para sintetizar imágenes de fondo de ojo correspondientes a las clases minoritarias: Glaucoma (G), Cataratas (C), Degeneración Macular Asociada a la Edad (A) y Miopía (M). Este enfoque busca no solo incrementar la cantidad de muestras disponibles, sino también mejorar la calidad clínica de las imágenes generadas mediante redes generativas adversarias.

Entrenamiento de modelos GAN especializados por clase

Para maximizar el rendimiento en cada clase, se entrenó un modelo GAN independiente por categoría minoritaria. La arquitectura del generador se diseñó con bloques residuales (*ResBlocks*) y una entrada latente de dimensión 256, generando imágenes RGB de resolución 128×128. El discriminador se construyó a partir del extractor de características del RETFound con *backbone* ResNet-18 (He, Zhang, Ren, & Sun, 2015), aplicado en régimen de *fine-tuning* parcial sobre las últimas capas. Esta configuración permitió evaluar las imágenes generadas desde una perspectiva clínica, fomentando la generación de imágenes con estructuras anatómicamente coherentes.

Durante el entrenamiento, se monitorizaron dos métricas fundamentales para evaluar la calidad de las imágenes generadas FID e IS.

Para estabilizar el proceso de entrenamiento y mejorar los resultados, se utilizó la técnica **EMA (*Exponential Moving Average*)**. Esta técnica mantiene una versión suavizada de los parámetros del generador principal, actualizando sus pesos de forma progresiva con una media exponencial. De esta forma, se mitigan oscilaciones y se obtiene un modelo más robusto y estable.

En este trabajo, se mantuvo una copia EMA del generador (`ema_G`) durante todo el entrenamiento. Esta versión fue utilizada para la generación final de imágenes, ya que obtuvo los mejores resultados en términos de FID e IS.

Generación del *dataset* aumentado con imágenes sintéticas

Una vez entrenados los modelos generativos, se utilizó el mejor *checkpoint* de cada clase para sintetizar el número exacto de imágenes necesarias hasta alcanzar un total de 2000 imágenes por clase minoritaria, de forma que la distribución de clases quedaría igual que en la Figura 6. El número de imágenes generadas por clase fue:

- Glaucoma (G): 1740 imágenes sintéticas
- Cataratas (C): 1708 imágenes sintéticas
- AMD (A): 1758 imágenes sintéticas
- Miopía (M): 1762 imágenes sintéticas

Integración con el *dataset* original

El nuevo *dataset* equilibrado se construyó combinando estas imágenes sintéticas con las imágenes reales del *dataset* desequilibrado. Esto dio lugar a un conjunto de datos más equilibrado que respeta la distribución de clases deseada y que, al mismo tiempo, incorpora variabilidad adicional generada por los modelos GAN.

Este enfoque de aumento de datos sintéticos con GANs permite al modelo de clasificación aprender a partir de un conjunto más representativo y diverso, especialmente en aquellas patologías con escasa representación en el conjunto original. La calidad de las imágenes fue controlada durante la generación mediante las métricas FID e IS, asegurando que solo se utilizaran los mejores generadores para la producción de las imágenes finales.

Limitaciones y análisis crítico

Aunque el uso de GANs ha permitido generar un volumen significativo de imágenes sintéticas para las clases minoritarias, se identificaron importantes limitaciones a nivel de calidad visual y utilidad clínica. En concreto, las imágenes generadas por los modelos entrenados para las clases Cataratas (C), presentada en la Figura 7, y AMD (A), mostrada en la Figura 8, presentaban ciertas estructuras que, de forma general, podían asemejarse a fondos de ojo reales. Sin embargo, en el caso de las clases Glaucoma (G), mostrada en la Figura 9, y Miopía (M), observable en la Figura 10, los resultados fueron considerablemente más deficientes, con imágenes poco definidas o con artefactos que comprometen su interpretación médica.

Figura 7. *Imagen de cataratas generada con GAN.*

Figura 8*. Iamgen de AMD generada con GAN.*

Figura 9. *Imagen de glaucoma generada con GAN.*

Figura 10. *Imagen de miopía generada con GAN.*

Estas deficiencias pueden atribuirse, en gran medida, a las restricciones computacionales del entorno de trabajo. La falta de memoria de GPU y la imposibilidad de utilizar arquitecturas más complejas o de entrenar durante más épocas limitaron notablemente el rendimiento alcanzable por los modelos GAN.

No obstante, pese a que las imágenes no presentan una validez clínica adecuada, su inclusión en el *dataset* permitió enriquecerlo y comprobar el potencial que podrían tener las GAN respecto a otros modelos.

3.2.4 *Dataset* con modelos de difusión

Con el objetivo de explorar el potencial de los modelos de difusión como herramienta para el aumento de datos en imágenes médicas, se desarrolló un *pipeline* específico basado en la arquitectura *Denoising Diffusion Probabilistic Models* (DDPM) (Ho, Jain, & Abbeel, 2020). Este enfoque se utilizó para generar imágenes sintéticas pertenecientes a las clases minoritarias del conjunto de datos: Glaucoma (G), Cataratas (C), Degeneración Macular Asociada a la Edad (A) y Miopía (M).

Entrenamiento del modelo de difusión DDPM

Para cada clase minoritaria, se entrenó un modelo de difusión independiente. La arquitectura propuesta combina un codificador basado en ResNet-18, que actúa como extractor de características visuales, y un decodificador convolucional transpuesto encargado de reconstruir la imagen a partir del ruido progresivamente reducido.

El entrenamiento se realizó mediante un proceso iterativo de 100 épocas por clase, aplicando una programación lineal de ruido con 100 pasos de difusión (*timesteps*). A cada imagen real se le añadió ruido gaussiano de forma controlada y el modelo fue entrenado para predecir este ruido en cada paso. Una vez finalizado el entrenamiento, el modelo aprendió a reconstruir imágenes partiendo de ruido puro.

Durante el entrenamiento, se utilizaron las métricas ***Fréchet Inception Distance*** **(FID)** e ***Inception Score*** **(IS)** para evaluar la calidad y diversidad de las imágenes generadas. Estas métricas se calcularon a partir de activaciones del modelo InceptionV3 (Szegedy, Vanhoucke, Ioffe, & Shlens, Rethinking the Inception Architecture for Computer Vision, 2015) sobre imágenes reales y sintéticas. El mejor modelo de cada clase se seleccionó en base al menor valor de FID alcanzado y al mayor IS alcanzado.

Generación del *dataset* aumentado con imágenes de difusión

Una vez entrenado el modelo por clase, se utilizó su mejor versión (con menor FID y mayor IS) para generar el número exacto de imágenes necesarias hasta alcanzar las 2000 muestras por clase minoritaria, la distribución de clases quedaría igual que en la Figura 6. El número de imágenes generadas fue el siguiente:

- Glaucoma (G): 1740 imágenes sintéticas
- Cataratas (C): 1708 imágenes sintéticas
- AMD (A): 1758 imágenes sintéticas
- Miopía (M): 1762 imágenes sintéticas

Cada imagen fue generada desde un tensor de ruido inicial que se fue refinando paso a paso mediante el modelo entrenado. El proceso de muestreo inverso implicó la aplicación de 100 pasos de *denoising* controlado para reconstruir progresivamente una imagen coherente desde el ruido gaussiano inicial.

Integración con el *dataset* original

Al igual que en las versiones anteriores, el nuevo *dataset* equilibrado con difusión se construyó combinando las imágenes sintéticas con las imágenes reales del *dataset*

desequilibrado. De este modo, se garantizó una distribución uniforme de clases sin perder representatividad clínica.

Limitaciones y análisis crítico

A pesar del potencial teórico de los modelos de difusión para generar imágenes de alta fidelidad, en este proyecto se observaron resultados visualmente insatisfactorios. Las imágenes generadas para todas las clases minoritarias cataratas (Figura 11), AMD (Figura 12), glaucoma (Figura 13) y miopía (Figura 14), carecían de detalles anatómicos reconocibles y presentaban estructuras borrosas o incoherentes desde una perspectiva clínica.

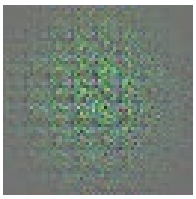

Figura 11. *Imagen de cataratas generada por modelos de difusión.*

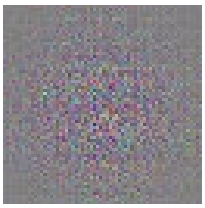

Figura 12. *Imagen de AMD generada por modelos de difusión.*

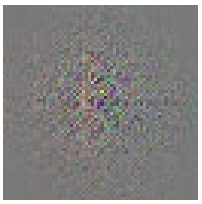

Figura 13. *Imagen de glaucoma generada por modelos de difusión.*

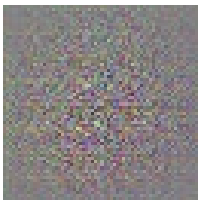

Figura 14. *Imagen de miopía generada por modelos de difusión.*

Al igual que con los modelos GAN, una de las principales causas identificadas fue la **limitación de recursos computacionales**. El entrenamiento de modelos de difusión requiere una gran cantidad de VRAM y tiempo de cómputo, lo que en este caso no se pudo satisfacer debido al entorno disponible para el TFG. En particular, el tamaño reducido de las imágenes (64×64 px) y la arquitectura simplificada utilizada fueron compromisos necesarios para poder ejecutar el proceso completo en una GPU con memoria limitada.

Al igual que para las GAN las imágenes baja calidad visual sirvieron para enriquecer el *dataset* y ver cuál podría ser el potencial de los modelos de difusión utilizados para el aumento de datos mediante imágenes sintéticas.

3.3 Herramientas utilizadas

Entorno de desarrollo

El desarrollo del proyecto se realizó íntegramente en **Python 3.11.9**[1], utilizando un entorno virtual gestionado con **Conda**[2] para asegurar la compatibilidad entre librerías y facilitar la replicabilidad del experimento. La programación y ejecución de los experimentos se llevó a cabo a través de **Visual Studio Code**[3], utilizando ***notebooks*** **de Jupyter**[4].

[1] https://www.python.org/downloads/release/python-3119/

[2] https://docs.conda.io/en/latest/

[3] https://code.visualstudio.com/

Librerías y *frameworks*

A lo largo del trabajo se emplearon diversas librerías especializadas para el procesamiento de datos, el entrenamiento de modelos y la evaluación de resultados:

- **NumPy, Pandas, Matplotlib, Seaborn**: para la manipulación de datos, análisis exploratorio y visualización de resultados.
- **Scikit-learn**: para métricas de evaluación y técnicas de sobremuestreo como SMOTE.
- **Imbalanced-learn**: específicamente para el uso de SMOTE y el manejo de *datasets* desbalanceados.
- **TensorFlow y Keras**: para la implementación y entrenamiento de los modelos de clasificación, así como para funciones de pérdida personalizadas y optimización.
- **PyTorch**: utilizado principalmente para la carga de modelos preentrenados, como RETFound, y para el entrenamiento de modelos generativos (WGAN, modelos de difusión).
- **Torchvision**: para transformaciones de imágenes y evaluación con métricas como FID e Inception Score.
- **Torchmetrics**: para el cálculo de métricas de evaluación de imágenes generadas (FID e IS).
- **TQDM**: para seguimiento del progreso durante el entrenamiento.
- **OpenCV (cv2)** y **PIL**: para procesamiento y lectura de imágenes.
- **TensorFlow Addons**: para el uso de *Focal Loss* y otras funciones avanzadas.

Recursos hardware

El proyecto se desarrolló en un entorno hardware principal consistente en un **ordenador personal** con procesador Intel Core i7, 16 GB de RAM y GPU **NVIDIA Quadro T2000** con 4 GB de memoria dedicada.

Control de versiones y documentación

Para asegurar una organización clara del trabajo y facilitar su trazabilidad, se emplearon los siguientes recursos:

- **Git**: control de versiones local para el seguimiento de *scripts* y evolución del código.
- **Markdown y Jupyter Notebooks**: para documentación de pruebas, hipótesis y observaciones durante el desarrollo.
- **Microsoft Word**: para la redacción final de la memoria académica.

[4] https://jupyter.org/

3.4 Arquitecturas utilizadas

Para las tareas de clasificación de imágenes oculares en este trabajo, se han empleado dos arquitecturas principales como extractores de características, ambas ampliamente validadas en el ámbito de la imagen médica:

- **RETFound** (Zhou, y otros, 2023): Un modelo basado en *Vision Transformers*, preentrenado específicamente sobre grandes volúmenes de imágenes de fondo de ojo. RETFound ha demostrado una gran capacidad para capturar representaciones visuales relevantes en tareas clínicas, por lo que se ha utilizado como extractor de características en combinación con clasificadores propios entrenados sobre esas representaciones.

- **EfficientNetB3** (Tan & Le, 2019): Una arquitectura convolucional eficiente que equilibra profundidad, anchura y resolución de entrada para maximizar la precisión manteniendo una complejidad computacional razonable. En este proyecto, EfficientNetB3 se ha empleado también como extractor de características complementario a RETFound.

En cuanto a la generación de imágenes sintéticas, se han utilizado arquitecturas adicionales adaptadas a este fin:

- **WGAN-GP (Wasserstein GAN con penalización de gradiente)** (Gulrajani, Arjovsky, Ahmed, Dumoulin, & Courville, 2017): Una variante robusta de las GANs que mejora la estabilidad del entrenamiento y reduce el riesgo de colapso del modo. Esta arquitectura se ha utilizado para generar imágenes sintéticas de clases minoritarias en el *dataset* ocular.

- **Modelo de difusión DDPM con codificador ResNet-18** (Ho, Jain, & Abbeel, 2020): Se ha implementado un modelo de *denoising diffusion probabilistic model* (DDPM) basado en un *autoencoder*, donde el codificador corresponde a una arquitectura ResNet-18. Este modelo ha sido entrenado para generar imágenes sintéticas de alta fidelidad, especialmente de clases minoritarias, y se ha evaluado mediante métricas FID e IS.

- **ResNet-18** (He, Zhang, Ren, & Sun, 2015): Red convolucional residual de 18 capas, utilizada en este proyecto como codificador dentro del modelo de difusión (DDPM). Esta arquitectura actúa como extractor de características visuales a partir de imágenes ruidosas en el proceso de reconstrucción, ofreciendo una alternativa computacionalmente eficiente frente a modelos más complejos como RETFound.

Estas arquitecturas han sido integradas en el *pipeline* del proyecto de forma modular, permitiendo evaluar su impacto tanto en la generación como en la clasificación de imágenes oculares en escenarios con datos desbalanceados.

3.5 Métricas y métodos de evaluación

Con el fin de analizar rigurosamente el comportamiento de los modelos desarrollados en este trabajo, se emplearon distintos métodos de evaluación, abarcando tanto métricas de rendimiento en tareas de clasificación como indicadores de calidad para modelos generativos y técnicas de interpretabilidad visual. Esta estrategia integral permite valorar no solo la precisión del sistema, sino también la fiabilidad de las imágenes sintéticas y la coherencia semántica de las predicciones.

Evaluación del rendimiento en clasificación

Para evaluar la eficacia de los modelos clasificadores, se utilizaron las siguientes métricas:

- **Matriz de confusión**: la Tabla 4 permite una visualización detallada de los aciertos y errores de clasificación por clase, aportando información relevante sobre patrones de confusión.

	Pred. Clase A	**Pred. Clase B**
Real Clase A	TP	FN
Real Clase B	FP	TP

***Tabla 4**. Matriz de confusion.*

  - **TP:** Verdaderos positivos
  - **TN:** Verdaderos negativos
  - **FP:** Falsos positivos
  - **FN:** Falsos negativos

- ***Accuracy***: mide la proporción de predicciones correctas sobre el total de muestras. Es una métrica intuitiva, pero puede ser **engañosa en contextos desbalanceados**, ya que un modelo puede obtener alta exactitud simplemente prediciendo siempre la clase mayoritaria.

  - **Fórmula**:

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN}$$

- ***Precision***: representa la proporción de verdaderos positivos entre todas las instancias que el modelo ha predicho como positivas. Es especialmente importante en medicina para **reducir falsos positivos**, lo cual evita alarmas innecesarias o tratamientos erróneos.

- **Fórmula**:

$$Precision = \frac{TP}{FP + TP}$$

- Una alta *precisión* indica que, cuando el modelo predice una clase, suele acertar.

- ***Recall* (Sensibilidad)**: mide la capacidad del modelo para **detectar correctamente todas las instancias reales** de una clase. En el contexto médico, es crítica: un bajo *recall* en una clase patológica significa que se están **omitiendo diagnósticos importantes**.
  - **Fórmula:**

$$Recall = \frac{TP}{FN + TP}$$

  - Un alto *recall* significa que el modelo no se "olvida" de las clases relevantes.
- **F1-*Score***: proporciona un equilibrio entre *precision* y *recall*. Es especialmente útil cuando se necesita **considerar ambas métricas al mismo tiempo**, como en escenarios desequilibrados donde maximizar solo una puede perjudicar a la otra.
  - **Fórmula:**

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

  - Penaliza duramente los casos en los que una de las dos métricas es baja.
- **AUC-ROC**: Área bajo la curva ROC para cada patología, evaluando la capacidad de discriminación entre clases. Mide la capacidad discriminativa global del modelo, considerando distintos umbrales de decisión.
  - La **curva ROC (*Receiver Operating Characteristic*)** representa gráficamente la relación entre la **tasa de verdaderos positivos (TPR)** y la **tasa de falsos positivos (FPR)** para distintos umbrales de clasificación.
  - $TPR\ (Recall) = \frac{TP}{TP+FN}$
  - $FPR = \frac{FP}{FP+TN}$
  - El **AUC (*Area Under Curve*)** mide el área bajo esa curva. Cuanto más cerca esté de 1, mejor será la capacidad del modelo para **distinguir entre clases**.

En todos los experimentos se empleó una estrategia de evaluación basada en **partición fija estratificada**, seleccionando un 10% de imágenes reales de cada clase para formar el conjunto de validación. Estas imágenes fueron extraídas exclusivamente del conjunto original (sin aumentos) y se eliminaron explícitamente de los *datasets* de entrenamiento, incluso en sus versiones aumentadas, para asegurar que el modelo no tuviera acceso a ellas durante el aprendizaje.

Este enfoque garantiza que la evaluación se realice siempre sobre imágenes no vistas por el modelo, preservando la distribución original de clases en ambos subconjuntos. Al mantener constante el conjunto de validación en todos los escenarios experimentales, se facilita además una comparación equitativa del rendimiento entre los distintos modelos desarrollados.

Evaluación de la calidad de imágenes generadas

En los experimentos relacionados con la generación de imágenes sintéticas —tanto mediante GANs como mediante modelos de difusión—, se emplearon dos métricas ampliamente aceptadas el FID (Heusel, Ramsauer, Unterthiner, Nessler, & Hochreiter, 2018) y el IS (Salimans, y otros, 2016) explicadas en el capítulo de estado del arte.

Ambas métricas permiten cuantificar de forma objetiva la fidelidad visual y la coherencia semántica de las muestras generadas, aspectos fundamentales en el contexto médico.

Evaluación de interpretabilidad con Grad-CAM

Con el objetivo de interpretar el comportamiento del modelo y verificar la coherencia clínica de las predicciones, se aplicó la técnica **Grad-CAM (*Gradient-weighted Class Activation Mapping*)** sobre imágenes seleccionadas. Esta técnica permite visualizar las regiones de una imagen que más influyen en la decisión del modelo, facilitando la identificación de patrones visuales relevantes y la validación por parte de expertos clínicos.

3.6 Modelos

Con el objetivo de mejorar la capacidad de los modelos de clasificación para detectar patologías oculares, se ha utilizado como punto de partida una arquitectura previamente entrenada con imágenes retinianas: RETFound (Zhou, y otros, 2023). En concreto, se han empleado los pesos disponibles en el archivo `RETFound_cfp_weights.pth` [5], correspondientes a un modelo *autoencoder* basado en Vision Transformers (ViT) (Dosovitskiy, y otros, 2021), entrenado de forma auto-supervisada sobre una amplia colección de imágenes clínicas del fondo de ojo.

El uso de estos pesos preentrenados se justifica por varias razones. Por un lado, han sido optimizados específicamente para extraer representaciones visuales relevantes en el ámbito oftalmológico, permitiendo aprovechar el conocimiento adquirido a partir de grandes volúmenes

[5] https://github.com/openmedlab/RETFound_MAE

de datos médicos. Por otro lado, su aplicación como *backbon*e facilita una mejor generalización, incluso en escenarios con conjuntos de datos limitados o desbalanceados, como es el caso del presente estudio.

Todos los modelos desarrollados en este proyecto, independientemente de la técnica de aumento de datos empleada, se han basado en esta arquitectura con inicialización `RETFound_cfp_weights.pth`, asegurando así una base metodológica común que permita realizar comparaciones justas entre enfoques.

Además, la evaluación de los modelos se realizó definiendo un conjunto de **validación fijo** y exclusivo para cada escenario experimental. Para ello, se reservó el **10% de las imágenes originales por clase** como conjunto de validación, garantizando que **ninguna de esas imágenes fuera utilizada durante el entrenamiento**. Este proceso se repitió de forma coherente en los cuatro casos: *dataset* original, *dataset* aumentado con técnicas tradicionales, *dataset* generado mediante GANs y *dataset* generado con modelos de difusión. En todos los casos, las imágenes sintéticas se utilizaron únicamente para el entrenamiento, mientras que la validación se realizó siempre con ejemplos reales, permitiendo una evaluación justa y comparable del rendimiento alcanzado por los modelos.

3.3.1 Modelos de clasificación con el *dataset* desequilibrado

Dado que el conjunto de datos original presenta un fuerte desequilibrio entre clases, se diseñaron cuatro modelos de clasificación progresivamente más sofisticados para evaluar estrategias que mejoraran su rendimiento. Todos los modelos se construyeron utilizando como base las representaciones extraídas mediante el modelo preentrenado **RETFound** (Zhou, y otros, 2023), y en algunos casos se integraron *embeddings* adicionales de **EfficientNetB3** (Tan & Le, 2019) mediante *ensembles*. A continuación, se describen las características clave de cada uno.

Modelo 1: RETFound + Focal Loss

Este modelo actúa como modelo de referencia inicial. Se entrenó un clasificador denso utilizando exclusivamente las características extraídas con RETFound (Zhou, y otros, 2023) sobre el 90% de las imágenes originales. Para abordar el desequilibrio de clases sin aumentar los datos, se utilizó la ***Focal Loss*** (Lin, Goyal, Girshick, He, & Dollar, 2018), una función de pérdida robusta frente a clases minoritarias. La arquitectura del clasificador incorpora varias capas densas con normalización y *dropout.* La evaluación se llevó a cabo sobre un conjunto de validación fijo formado por el 10% de cada clase original, y se aplicó **Grad-CAM** (Selvaraju, y otros, 2017) para interpretar visualmente las predicciones correctas y erróneas.

Modelo 2: Ensemble RETFound + EfficientNetB3 + Ajuste de Umbral

Este segundo modelo mejora el anterior integrando un sistema de ***ensemble*** **ponderado** (Goceri, 2023) entre dos clasificadores independientes: uno entrenado con RETFound (Zhou, y

otros, 2023) y otro con *embeddings* extraídos mediante EfficientNetB3 (Tan & Le, 2019). Ambos modelos fueron entrenados con ***Focal Loss*** (Lin, Goyal, Girshick, He, & Dollar, 2018). Para maximizar el rendimiento, se realizó un **ajuste por clase de los umbrales de decisión** en la salida *softmax* (Szegedy, Vanhoucke, Ioffe, Jonathon Shlens, & Wojna, Rethinking the Inception Architecture for Computer Vision, 2015). Esta estrategia permitió optimizar la sensibilidad específica por patología. También se aplicó Grad-CAM (Selvaraju, y otros, 2017) con RETFound (Zhou, y otros, 2023) para validar visualmente los resultados del modelo.

Modelo 3: Ensemble + MixUp

En este caso se reutilizó la misma estructura del modelo anterior, pero se introdujo una técnica de ***data augmentation*** **sobre los** ***embeddings*** mediante **MixUp** (Zhang, Cisse, Dauphin, & Lopez-Paz, 2018), aplicada exclusivamente al espacio de características de RETFound (Zhou, y otros, 2023). Esta técnica permite crear ejemplos intermedios entre clases, mejorando la capacidad de generalización del modelo. Se mantuvo el *ensemble* y el ajuste de umbrales, lo que permite evaluar el efecto directo de MixUp en el rendimiento sin modificar el resto del *pipeline*.

Modelo 4: Ensemble + SMOTE + CutMix

El cuarto modelo incorpora una combinación más ambiciosa de técnicas. Primero, se utilizó **SMOTE** (Chawla, Bowyer, Hall , & Philip Kegelmeyer, 2002) para aumentar de forma sintética las clases minoritarias directamente sobre los *embeddings* de RETFound. Posteriormente, se aplicó **CutMix** (Yun, y otros, 2019) como técnica adicional de mezcla entre representaciones. Ambas transformaciones buscan generar mayor diversidad en los datos de entrada al clasificador. Como en modelos anteriores, se utilizó *ensemble* con EfficientNetB3 (Tan & Le, 2019), *Focal Loss* (Lin, Goyal, Girshick, He, & Dollar, 2018) y ajuste de umbrales. Esta versión representa la configuración más avanzada dentro del bloque de modelos sobre el *dataset* original.

La Tabla 5 resume las principales características y diferencias entre los modelos desarrollados sobre el *dataset* desequilibrado.

Modelo	***Embeddings* utilizados**	**Estrategia de equilibrio**	**Regularización y mejoras**	**Técnica adicional**
Modelo 1	RETFound	Ninguna	*Dropout*, L2, LayerNorm	*Focal Loss*
Modelo 2	RETFound + EfficientNetB3	*Ensemble* ponderado	*Dropout*, L2, ajuste de umbral	*Focal Loss* + combinación de modelos
Modelo 3	RETFound + EfficientNetB3	MixUp + *ensemble*	*Label smoothing*, ajuste de umbral	*Focal Loss* + regularización avanzada

Modelo	*Embeddings* utilizados	Estrategia de equilibrio	Regularización y mejoras	Técnica adicional
Modelo 4	RETFound + EfficientNetB3	SMOTE + CutMix + *ensemble*	*Label smoothing*, ajuste de umbral, *class weights*	*Focal Loss* + mejoras acumuladas

***Tabla 5**. Comparación de modelos del dataset desequilibrado.*

Esta serie de modelos progresivos proporciona una base metodológica sólida y consistente para comparar, en igualdad de condiciones, el impacto de las distintas técnicas de aumento de datos utilizadas a lo largo del proyecto. Gracias a la aplicación controlada de estrategias como la combinación de *embeddings*, técnicas de regularización y métodos de equilibrio avanzados, es posible evaluar de forma precisa cómo evoluciona el rendimiento en función de cada intervención aplicada.

Tras explorar diversas aproximaciones sobre el conjunto de datos original desequilibrado, el siguiente bloque de modelos se centra en analizar escenarios en los que el *dataset* ha sido previamente equilibrado mediante transformaciones tradicionales de *data augmentation*.

3.3.2 Modelos de clasificación con el *dataset* equilibrado con técnicas tradicionales

Con el objetivo de analizar el impacto de un conjunto de datos previamente equilibrados mediante transformaciones tradicionales —como rotaciones, volteos, traslaciones y ajustes de brillo o contraste—, se desarrollaron cuatro modelos de clasificación de complejidad creciente. Todos ellos comparten una arquitectura base fundamentada en *embeddings* extraídos con el modelo preentrenado **RETFound** (Zhou, y otros, 2023), y algunos incorporan adicionalmente representaciones generadas con **EfficientNetB3** (Tan & Le, 2019) mediante técnicas de *ensemble*.

A diferencia del escenario con datos desequilibrados, este bloque parte de un *dataset* ya equilibrado, lo que permite evaluar más directamente el efecto de diferentes estrategias de entrenamiento, sin necesidad de aplicar técnicas de equilibrado explícito durante el proceso de aprendizaje. La evaluación se llevó a cabo sobre un conjunto de validación fijo e independiente, reservado desde el inicio y compartido por todos los modelos del bloque.

Aunque las arquitecturas se diseñaron de forma incremental, no todas las modificaciones introducidas condujeron a mejoras sustanciales en el rendimiento. Sin embargo, la exploración sistemática de estas variantes resulta clave para ilustrar el proceso metodológico seguido y comparar, en igualdad de condiciones, los efectos de técnicas como **MixUp** (Zhang, Cisse, Dauphin, & Lopez-Paz, 2018) y **CutMix** (Yun, y otros, 2019) cuando se parte de un *dataset* tradicionalmente equilibrado.

Modelo 1: RETFound + Crossentropy

Este primer modelo se basa exclusivamente en las características extraídas por el modelo preentrenado **RETFound** (Zhou, y otros, 2023). Se entrenó un clasificador multicapa utilizando una función de pérdida estándar (*Categorical Crossentropy*) (Goodfellow, Bengio, & Courville, Deep Learning, 2016), con clases ponderadas proporcionalmente para mantener la equidad del aprendizaje. Se aplicaron técnicas de regularización mediante *dropout* y el entrenamiento se llevó a cabo sobre un conjunto de datos previamente dividido en entrenamiento y validación. Este modelo actúa como ***baseline*** para los escenarios equilibrados.

Modelo 2: Ensemble RETFound + EfficientNetB3 + Focal Loss

Este modelo incorpora una estrategia de ***ensemble*** **ponderado** entre los *embeddings* de RETFound (Zhou, y otros, 2023) y **EfficientNetB3** (Tan & Le, 2019). Ambos clasificadores fueron entrenados de forma independiente y sus predicciones combinadas según una proporción ajustada. Para mejorar la sensibilidad frente a clases difíciles, se utilizó ***Focal Loss*** (Lin, Goyal, Girshick, He, & Dollar, 2018) como función de pérdida. Se aplicó además un ajuste de umbral específico por clase y se empleó Grad-CAM (Selvaraju, y otros, 2017) para interpretar visualmente las decisiones del modelo.

Modelo 3: Ensemble + MixUp

Sobre la misma arquitectura de *ensemble* que en el modelo anterior, este tercer modelo introduce **MixUp** (Zhang, Cisse, Dauphin, & Lopez-Paz, 2018), aplicado directamente a los *embeddings* como técnica de regularización. Esta estrategia mejora la generalización del modelo al exponerlo a interpolaciones entre ejemplos reales. El resto del *pipeline* —clasificadores independientes, *Focal Loss* (Lin, Goyal, Girshick, He, & Dollar, 2018) y ajuste de umbrales— se mantuvo sin cambios. Esta configuración permitió aislar el impacto de MixUp sobre el rendimiento del sistema.

Modelo 4: Ensemble + CutMix

El cuarto modelo implementa **CutMix** (Yun, y otros, 2019) como técnica alternativa de combinación entre ejemplos en el espacio de *embeddings*. Al igual que MixUp, esta estrategia se aplicó por separado sobre los dos conjuntos de características (RETFound y EfficientNetB3) antes del entrenamiento. El modelo mantuvo el uso de *Focal Loss*, ajuste de umbrales por clase y evaluación mediante Grad-CAM. Esta versión representa la configuración más avanzada del bloque basado en técnicas tradicionales.

La Tabla 6 resume las principales características y diferencias entre los modelos desarrollados sobre el *dataset* equilibrado mediante técnicas tradicionales.

Modelo	***Embeddings* utilizados**	**Estrategia de equilibrio**	**Regularización y mejoras**	**Técnica adicional**
Modelo 1	RETFound	*Dataset* previamente equilibrado	*Dropout, class weights*	*Categorical Crossentropy*
Modelo 2	RETFound + EfficientNetB3	*Dataset* equilibrado + *ensemble*	BatchNorm, *Dropout*, ajuste de umbral	*Focal Loss* + combinación de modelos
Modelo 3	RETFound + EfficientNetB3	*Dataset* equilibrado + MixUp + *ensemble*	MixUp, BatchNorm, ajuste de umbral	*Focal Loss* + aumento emb.
Modelo 4	RETFound + EfficientNetB3	*Dataset* equilibrado + CutMix + *ensemble*	CutMix, BatchNorm, ajuste de umbral	*Focal Loss* + aumento emb. + combinación de modelos

Tabla 6. *Comparación de modelos entrenados sobre el dataset equilibrado mediante técnicas tradicionales.*

A través de estos modelos, se ha podido analizar el efecto real de distintas técnicas de entrenamiento y regularización sobre un *dataset* ya equilibrado. La inclusión de estrategias como MixUp, CutMix y *ensembles* permitió observar su impacto de forma controlada, destacando la importancia de evaluar cada intervención en contextos sin sesgo de clases. Esta comparación servirá como base para valorar la utilidad de técnicas más avanzadas de generación en los siguientes bloques.

El siguiente bloque de modelos se centra en analizar escenarios en los que el *dataset* ha sido equilibrado mediante técnicas más avanzadas de *data augmentation* como redes generativas adversarias (GANs).

3.3.3 Modelos de clasificación con el *dataset* equilibrado con GANs

Con el objetivo de evaluar el impacto de la generación de imágenes sintéticas mediante GANs (Goodfellow, y otros, 2014) para equilibrar las clases minoritarias, se diseñaron tres

modelos de clasificación progresivos. Todos comparten una arquitectura basada en *embeddings* extraídos con RETFound (Zhou, y otros, 2023) y EfficientNetB3 (Tan & Le, 2019), y utilizan técnicas como *ensembles* ponderados, *Focal Loss* (Lin, Goyal, Girshick, He, & Dollar, 2018), y ajuste de umbrales por clase. Además, la evaluación se realizó sobre un conjunto de validación fijo, compuesto por el 10% de las imágenes originales por clase, no utilizado durante el entrenamiento. Esta división se mantuvo constante en todos los bloques experimentales, asegurando una comparación justa entre modelos entrenados con datos reales o sintéticos. A continuación, se describen sus configuraciones.

Modelo 1: Ensemble sin MixUp ni CutMix

Este modelo actúa como referencia inicial sobre el *dataset* generado con GANs. Se utilizaron las representaciones extraídas por RETFound y EfficientNetB3, entrenando clasificadores independientes con *Focal Loss*. Las predicciones se combinaron mediante un *ensemble* ponderado, y se aplicó un ajuste de umbral por clase. Esta configuración proporciona una primera evaluación del rendimiento alcanzado con datos sintéticos sin aplicar técnicas adicionales de regularización.

Modelo 2: Ensemble + MixUp

Partiendo de la arquitectura anterior, se introdujo MixUp (Zhang, Cisse, Dauphin, & Lopez-Paz, 2018) como técnica de regularización aplicada sobre los *embeddings* de entrenamiento. Esta estrategia mejora la generalización del modelo al interpolar entre ejemplos, generando representaciones intermedias. El resto del *pipeline* se mantuvo inalterado: *ensemble* ponderado, *Focal Loss* y ajuste de umbrales.

Modelo 3: Ensemble + CutMix

En este tercer modelo se sustituyó MixUp por CutMix (Yun, y otros, 2019), utilizando un generador específico que mezcla pares de *embeddings* durante el entrenamiento. La técnica busca aumentar la diversidad de los datos sin alterar su distribución original. Como en los modelos anteriores, se utilizaron clasificadores independientes entrenados con *Focal Loss*, combinados mediante *ensemble*, y validados con ajuste de umbrales y Grad-CAM (Selvaraju, y otros, 2017).

La Tabla 7 resume las principales características y diferencias entre los modelos desarrollados sobre el *dataset* equilibrado mediante redes adversarias generativas.

Modelo	***Embeddings* utilizados**	**Estrategia de equilibrio**	**Regularización y mejoras**	**Técnica adicional**
Modelo 1	RETFound + EfficientNetB3	*Dataset* generado con GANs +	*Label smoothing*, ajuste de umbral, *Dropout*	*Focal Loss* + *ensemble*

Modelo	*Embeddings* utilizados	Estrategia de equilibrio	Regularización y mejoras	Técnica adicional
		ensemble		ponderado
Modelo 2	RETFound + EfficientNetB3	GANs + MixUp + *ensemble*	MixUp, ajuste de umbral, *Dropout*	*Focal Loss* + mezcla de *embeddings*
Modelo 3	RETFound + EfficientNetB3	GANs + CutMix + *ensemble*	CutMix, ajuste de umbral, *Dropout*	*Focal Loss* + generador CutMix

Tabla 7. *Comparación de modelos entrenados sobre el dataset equilibrado mediante redes adversarias generativas.*

Este bloque de modelos ha permitido evaluar, en condiciones controladas, el impacto de técnicas como MixUp, CutMix y *ensembles* sobre un *dataset* equilibrado mediante GANs. La comparación entre configuraciones progresivas evidencia cómo cada intervención puede influir en la capacidad de generalización del modelo. Estos resultados sirven de base para valorar la utilidad real de las imágenes sintéticas generadas por redes generativas adversarias frente a otras estrategias de aumento.

El siguiente bloque de modelos se centra en analizar escenarios en los que el *dataset* ha sido equilibrado mediante técnicas más avanzadas de *data augmentation* como modelos de difusión.

3.3.4 Modelos de clasificación con el *dataset* equilibrado con modelos de difusión

Con el objetivo de analizar el impacto de la generación de imágenes sintéticas mediante modelos de difusión para abordar el desbalance de clases, se diseñaron tres modelos de clasificación progresivos. Al igual que en los bloques anteriores, todos comparten una arquitectura basada en *embeddings* extraídos con RETFound (Zhou, y otros, 2023) y EfficientNetB3 (Tan & Le, 2019), y utilizan técnicas comunes como *ensembles* ponderados, *Focal Loss* (Lin, Goyal, Girshick, He, & Dollar, 2018) y ajuste de umbrales por clase. Además, la evaluación se realizó sobre un conjunto de validación fijo, compuesto por el 10% de las imágenes originales por clase, que no se utilizó durante el entrenamiento. Esta estrategia fue coherente en todos los experimentos, asegurando una comparación justa entre modelos entrenados con datos reales o sintéticos. A continuación, se describen las configuraciones aplicadas.

Modelo 1: Ensemble sin MixUp ni CutMix

Este modelo establece una referencia inicial para evaluar el rendimiento alcanzado únicamente con imágenes sintéticas generadas por modelos de difusión. Se utilizaron *embeddings* de RETFound y EfficientNetB3 para entrenar clasificadores independientes,

aplicando *Focal Loss* y combinando sus predicciones mediante un *ensemble* ponderado. También se realizó un ajuste de umbral por clase para optimizar las predicciones finales.

Modelo 2: Ensemble + MixUp

En este modelo se aplicó MixUp (Zhang, Cisse, Dauphin, & Lopez-Paz, 2018) sobre los *embeddings* de entrenamiento, como estrategia de regularización. Esta técnica permite generar ejemplos intermedios entre muestras reales, mejorando la capacidad de generalización del modelo. La arquitectura, el *ensemble* ponderado y el ajuste de umbrales se mantuvieron sin cambios respecto al modelo anterior.

Modelo 3: Ensemble + CutMix

Finalmente, el tercer modelo reemplazó MixUp por la técnica CutMix (Yun, y otros, 2019), utilizando un generador específico para combinar pares de *embeddings* durante el entrenamiento. Esta aproximación introduce una mayor variabilidad sin alterar la distribución general del *dataset.* El resto del *pipeline* fue idéntico: clasificadores independientes entrenados con *Focal Loss*, combinación mediante *ensemble* y evaluación cualitativa con Grad-CAM (Selvaraju, y otros, 2017).

La Tabla 8 resume las principales características de los tres modelos entrenados sobre el *dataset* equilibrado con modelos de difusión.

Modelo	***Embeddings* utilizados**	**Estrategia de equilibrio**	**Regularización y mejoras**	**Técnica adicional**
Modelo 1	RETFound + EfficientNetB3	*Dataset* generado con difusión + *ensemble*	*Label smoothing*, ajuste de umbral, *Dropout*	*Focal Loss* + *ensemble* ponderado
Modelo 2	RETFound + EfficientNetB3	Difusión + MixUp + *ensemble*	MixUp, ajuste de umbral, *Dropout*	*Focal Loss* + mezcla de *embeddings*
Modelo 3	RETFound + EfficientNetB3	Difusión + CutMix + *ensemble*	CutMix, ajuste de umbral, *Dropout*	*Focal Loss* + generador CutMix

Tabla 8. *Comparación de modelos entrenados sobre el dataset equilibrado mediante modelos de difusión.*

Este último bloque ha permitido evaluar de manera controlada el efecto de distintas técnicas de regularización sobre un *dataset* generado mediante modelos de difusión. La comparación progresiva entre configuraciones ha ofrecido una visión clara del valor añadido que pueden aportar estrategias como MixUp y CutMix. Estos resultados completan el análisis comparativo entre las distintas aproximaciones de aumento sintético abordadas en este trabajo.

Capítulo 4

Resultados

Una vez definidos los *datasets* y entrenados los modelos correspondientes, este capítulo presenta los resultados obtenidos a lo largo de los distintos bloques experimentales. El objetivo es evaluar cómo afectan las diferentes estrategias de equilibrio —tradicionales, generativas y combinadas— al rendimiento de los clasificadores entrenados sobre imágenes del fondo de ojo.

Los experimentos se estructuraron en cuatro bloques, cada uno asociado a un tipo de *dataset*: el conjunto original desequilibrado, una versión equilibrada mediante transformaciones tradicionales, otra utilizando imágenes sintéticas generadas por GANs y, por último, una versión equilibrada con imágenes generadas mediante modelos de difusión. En cada bloque se evaluaron distintos modelos con técnicas progresivas, como MixUp (Zhang, Cisse, Dauphin, & Lopez-Paz, 2018), CutMix (Yun, y otros, 2019), SMOTE (Chawla, Bowyer, Hall , & Philip Kegelmeyer, 2002) o *ensembles*, siempre partiendo de *embeddings* extraídos con RETFound (Zhou, y otros, 2023) y, en algunos casos, combinados con EfficientNetB3 (Tan & Le, 2019).

Los resultados se analizan de forma cuantitativa mediante métricas como *accuracy*, F1 macro, *recall* macro y AUC-ROC por clase, y de forma cualitativa mediante Grad-CAM (Selvaraju, y otros, 2017), con el fin de valorar no solo el rendimiento de los modelos, sino también la coherencia clínica de sus decisiones. Cada subsección recoge los hallazgos clave de cada enfoque de equilibrado, identifica el mejor modelo dentro de su bloque y discute sus fortalezas y limitaciones.

4.1 Resultados por tipo de *dataset*

4.1.1 *Dataset* original desequilibrado

El conjunto de datos original presenta un desequilibrio marcado entre clases. Las categorías mayoritarias —**Normal** (N) y **Diabetes** (D)— concentran más del 70% de las imágenes disponibles, mientras que las clases minoritarias como **Glaucoma** (G), **Cataratas** (C), **Degeneración Macular** (A) y **Miopía** (M) apenas superan unas pocas decenas de ejemplos. Esta situación refleja una distribución común en entornos clínicos reales, pero supone un reto significativo para los modelos de clasificación, que tienden a favorecer las clases predominantes.

Con el objetivo de mitigar este efecto y explorar distintas estrategias de aprendizaje, se diseñaron cuatro modelos progresivos que incorporan técnicas como ***Focal Loss*** (Lin, Goyal, Girshick, He, & Dollar, 2018), ***ensembles***, **MixUp** (Zhang, Cisse, Dauphin, & Lopez-Paz, 2018), **SMOTE** (Chawla, Bowyer, Hall , & Philip Kegelmeyer, 2002) y **CutMix** (Yun, y otros, 2019) **x**, aplicadas sobre los *embeddings* extraídos con **RETFound** (Zhou, y otros, 2023) y, en algunos casos, combinados con **EfficientNetB3** (Tan & Le, 2019). La evaluación se realizó de forma consistente sobre un conjunto de validación exclusivo compuesto por el 10% de las imágenes reales por clase.

Resultados cuantitativos

Los resultados se presentan en las Tablas 9 y 10. La Tabla 9 recoge las métricas globales de cada modelo, mientras que la Tabla 10 muestra el F1-*score* por clase para facilitar la comparación del rendimiento específico en cada categoría.

Modelo	**Técnica adicional**	***Accuracy***	**F1 macro**	***Recall* macro**
Modelo 1	*Focal Loss*	0.58	0.43	0.47
Modelo 2	*Focal Loss* + *Ensemble*	0.68	0.64	0.63
Modelo 3	*Focal Loss* + MixUp	0.60	0.51	0.50
Modelo 4	*Focal Loss* + CutMix + SMOTE	**0.67**	**0.71**	**0.70**

Tabla 9. *Métricas generales por modelo para el dataset desequilibrado.*

Clase	**Modelo 1**	**Modelo 2**	**Modelo 3**	**Modelo 4**
N	0.67	0.75	0.68	**0.71**
D	0.44	0.46	0.45	**0.56**
G	0.00	0.51	0.14	**0.56**

Clase	Modelo 1	Modelo 2	Modelo 3	Modelo 4
C	0.71	0.84	0.78	**0.85**
A	0.00	0.44	0.15	**0.67**
M	0.77	0.87	0.85	**0.90**

Tabla 10. *F1-score por clase para el dataset desequilibrado.*

Estos resultados muestran una clara mejora progresiva del rendimiento a medida que se incorporan técnicas más sofisticadas. El Modelo 1, que actúa como referencia inicial con un clasificador denso y *Focal Loss*, evidencia dificultades importantes en las clases minoritarias como G (Glaucoma) o A (AMD), con F1-*scores* nulos. A partir del Modelo 2, la incorporación de un *ensemble* ponderado mejora notablemente el desempeño general, especialmente en C, M y G.

El Modelo 3 introduce **MixUp** sobre los *embeddings*, pero su efecto es limitado, posiblemente por la alta variabilidad generada al interpolar características en un contexto altamente desequilibrado. En cambio, el **Modelo 4**, que combina **SMOTE**, **CutMix** y un *ensemble* ponderado, consigue el mejor balance global, alcanzando el mayor F1 macro (0.71) y mejorando el desempeño en todas las clases, especialmente en A (0.67), G (0.56) y M (0.90).

Mejor modelo

El **Modelo 4** es el que ofrece el rendimiento más equilibrado y sólido en el escenario desequilibrado. A continuación, se presentan sus visualizaciones clave en las Figuras 15 y 16.

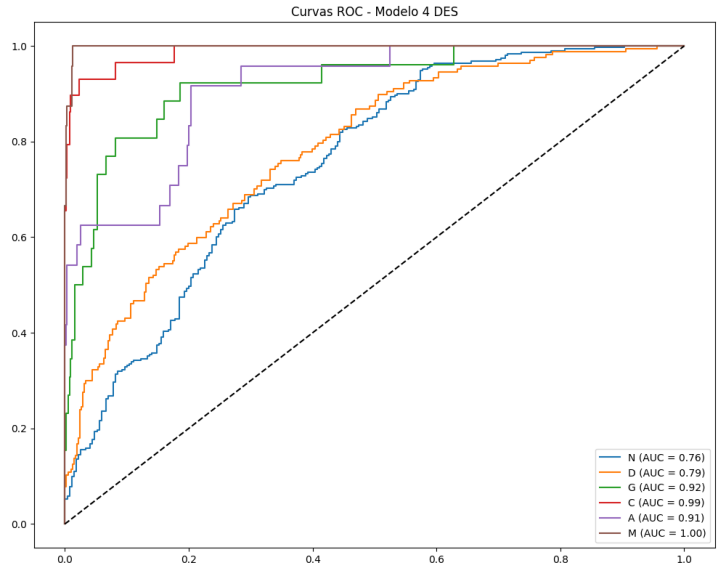

Figura 15*. Curva ROC por clase del modelo 4 para el dataset desequilibrado.*

La Figura 15 muestra un desempeño especialmente alto en las clases **M** (Miopía, AUC = 1.00) y **C** (Cataratas, AUC = 0.99), y buenos resultados en **G** (0.92) y **A** (0.91). Esto indica que el modelo logra identificar de forma precisa incluso las clases menos representadas.

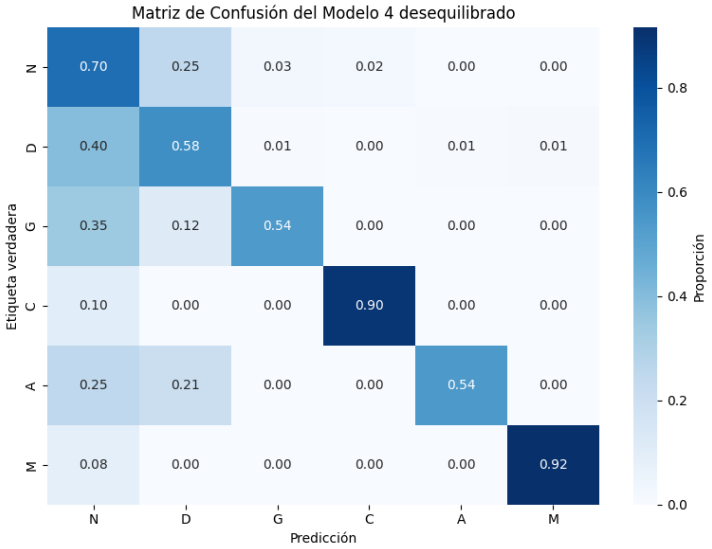

Figura 16. *Matriz de confusión modelo 4 para el dataset desequilibrado.*

La matriz de confusión de la Figura 16 refuerza estas conclusiones: se observan **aciertos muy consistentes en M (0'92), C (0'90)** y un equilibrio razonable en N y D, a pesar de la confusión esperada entre ambas clases por su prevalencia.

En resumen, este bloque de experimentación evidencia cómo la incorporación de técnicas de generación y mezcla sobre el espacio latente de los *embeddings*, junto con funciones de pérdida adaptadas y combinación de clasificadores, **permite mitigar en gran medida los efectos del desequilibrio** en tareas de clasificación médica. El Modelo 4 constituye la solución más robusta dentro de este escenario.

4.1.2 *Dataset* equilibrado con técnicas tradicionales

El uso de transformaciones tradicionales para equilibrar el conjunto de datos permitió entrenar modelos sin recurrir a técnicas de equilibrado adicionales durante la fase de aprendizaje. Se evaluaron cuatro modelos con arquitecturas y mejoras progresivas, desde un clasificador simple basado en RETFound (Zhou, y otros, 2023) hasta esquemas de *ensemble* con regularización mediante MixUp (Zhang, Cisse, Dauphin, & Lopez-Paz, 2018) y CutMix (Yun, y otros, 2019). Todos los modelos fueron evaluados sobre el mismo conjunto de validación fijo e independiente, garantizando la comparabilidad de los resultados entre configuraciones.

Resultados cuantitativos

Las métricas generales de rendimiento se resumen en la Tabla 11. Se observa una mejora sostenida desde el Modelo 1 hasta el Modelo 3, con una ligera estabilización en el Modelo 4.

Modelo	**Técnica adicional**	***Accuracy***	**F1 macro**	***Recall* macro**
Modelo 1	RETFound + *Categorical Crossentropy*	0.54	0.55	0.60
Modelo 2	*Ensemble* + *Focal Loss*	0.66	0.61	0.58
Modelo 3	MixUp + *Ensemble*	**0.69**	**0.67**	**0.64**
Modelo 4	CutMix + *Ensemble*	0.66	0.64	0.62

Tabla 11. *Métricas generales de los modelos con el dataset equilibrado por transformaciones tradicionales.*

La Tabla 12 muestra el F1-*score* por clase. El Modelo 3, que incorpora MixUp sobre los *embeddings*, consigue los mejores resultados promedio y un desempeño equilibrado en la mayoría de las categorías.

Modelo	**N**	**D**	**G**	**C**	**A**	**M**
Modelo 1	0.54	0.51	0.28	0.85	0.27	0.87
Modelo 2	0.74	0.48	0.37	0.81	0.45	0.82
Modelo 3	**0.76**	**0.49**	**0.51**	**0.85**	**0.55**	**0.88**
Modelo 4	0.73	0.48	0.44	0.83	0.53	0.85

Tabla 12*. F1-score por clase en modelos con el dataset equilibrado por transformaciones tradicionales.*

Estos resultados indican que el uso de MixUp en el Modelo 3 tuvo un impacto positivo en la generalización, especialmente en las clases minoritarias G, A y M. El Modelo 4, aunque también emplea una técnica avanzada (CutMix), no logra mejorar los resultados obtenidos con MixUp, manteniéndose ligeramente por debajo en la media global.

Mejor modelo

A continuación, se presentan los gráficos correspondientes al Modelo 3, que alcanzó el mejor rendimiento global en este bloque.

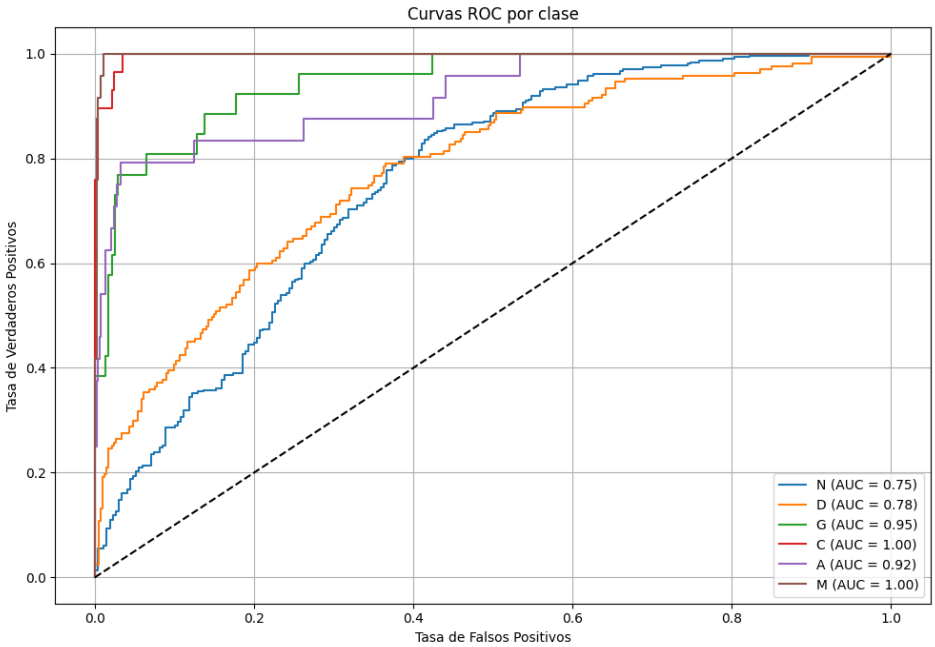

Figura 17. *Curva ROC por clase Modelo 3 para el dataset equilibrado con tranformaciones tradicionales.*

La Figura 17 revela un excelente desempeño en las clases M (AUC = 1.00), C (1.00) y G (0.95), con valores de AUC superiores a 0.85 incluso en clases menos representadas como A (0.92). Estos resultados evidencian la eficacia del *ensemble* combinado con MixUp para mejorar la sensibilidad del modelo en escenarios equilibrados artificialmente.

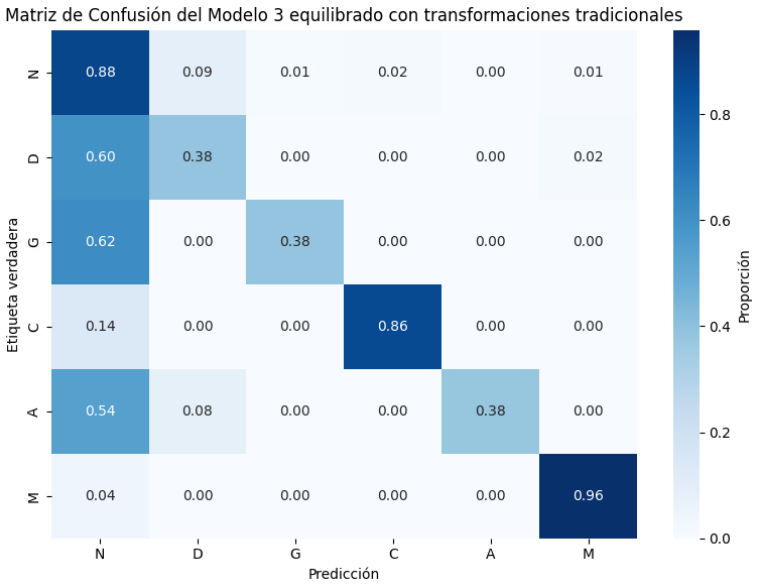

Figura 18. *Matriz de confusión modelo 3 para el dataset equilibrado mediante transformaciones tradicionales.*

En la Figura 18 se observa una tasa de aciertos elevada en la mayoría de las clases, especialmente en M (0.96) N (0.88) y C (0.86), mientras que las clases D, G y A continúan presentando cierta confusión con N, probablemente debido a la similitud visual en algunos casos clínicos. Aun así, el rendimiento general es robusto.

El equilibrio previo del *dataset* mediante transformaciones tradicionales permitió mejorar el rendimiento sin técnicas adicionales de balanceo. El modelo con MixUp fue el más eficaz, especialmente en clases minoritarias, destacando la utilidad de regularizaciones simples en escenarios equilibrados.

4.1.3 *Dataset* equilibrado con GANs

La generación de imágenes sintéticas con redes generativas adversarias (GANs) (Goodfellow, y otros, 2014) permitió equilibrar el conjunto de datos y entrenar modelos de clasificación sin necesidad de aplicar técnicas adicionales de balanceo durante el entrenamiento. En este bloque, se evaluaron tres modelos con configuraciones progresivas, partiendo de un *ensemble* simple hasta incorporar técnicas de regularización como MixUp (Zhang, Cisse, Dauphin, & Lopez-Paz, 2018) y CutMix (Yun, y otros, 2019). La validación se realizó con un conjunto fijo y exclusivo de imágenes reales, manteniendo la coherencia con los bloques anteriores y garantizando la comparabilidad.

Resultados cuantitativos

La Tabla 13 resume las métricas globales obtenidas por cada modelo. El Modelo 1 y el Modelo 2 alcanzan el mismo nivel de precisión global, con una ligera ventaja del Modelo 1 en métricas promedio. El Modelo 3, aunque introduce CutMix, no logra superar el rendimiento de sus predecesores.

Modelo	**Técnica adicional**	***Accuracy***	**F1 macro**	***Recall* macro**
Modelo 1	*Focal Loss* + *Ensemble*	**0.69**	**0.68**	**0.67**
Modelo 2	MixUp + *Ensemble*	0.69	0.63	0.59
Modelo 3	CutMix + *Ensemble*	0.66	0.58	0.56

Tabla 13. *Métricas generales de los modelos entrenados con los datos sintéticos generados por GANs.*

La Tabla 14 presenta el **F1-*score* por clase**, permitiendo evaluar la eficacia del modelo en cada categoría. El Modelo 1 obtiene el mejor balance global, con resultados especialmente sólidos en N, C y M. MixUp mejora algunas clases específicas (como C), pero reduce el rendimiento en otras como A. Por su parte, CutMix presenta mayor inestabilidad.

Modelo	**N**	**D**	**G**	**C**	**A**	**M**
Modelo 1	**0.75**	**0.55**	**0.59**	**0.83**	**0.54**	**0.85**
Modelo 2	0.77	0.48	0.47	0.87	0.32	0.84

Modelo	N	D	G	C	A	M
Modelo 3	0.74	0.53	0.33	0.83	0.22	0.79

Tabla 14. *F1-score por clase de los modelos con el dataset equilibrado mediante GANs.*

Mejor modelo

El **Modelo 1**, que combina RETFound y EfficientNetB3 con *Focal Loss* y *ensemble* ponderado, destaca como el más equilibrado de este bloque. A continuación, se muestran sus resultados visuales.

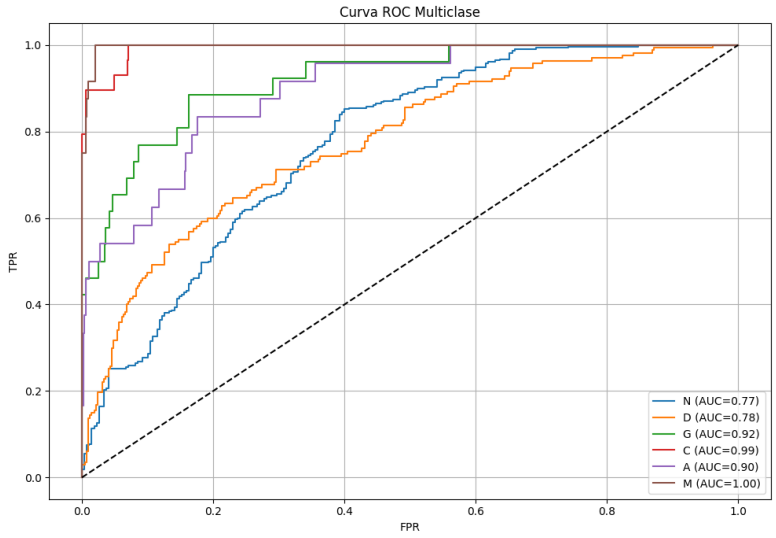

Figura 19. *Curva ROC por clase del Modelo 1 entrenado con el dataset equilibrado mediante GANs.*

La Figura 19 muestra la curva ROC; refleja una alta capacidad discriminativa en casi todas las clases, especialmente M (AUC = 1.00), C (0.99) y G (0.92). Incluso en clases más desafiantes como A y D se obtienen AUC superiores a 0.75, lo que confirma una generalización adecuada.

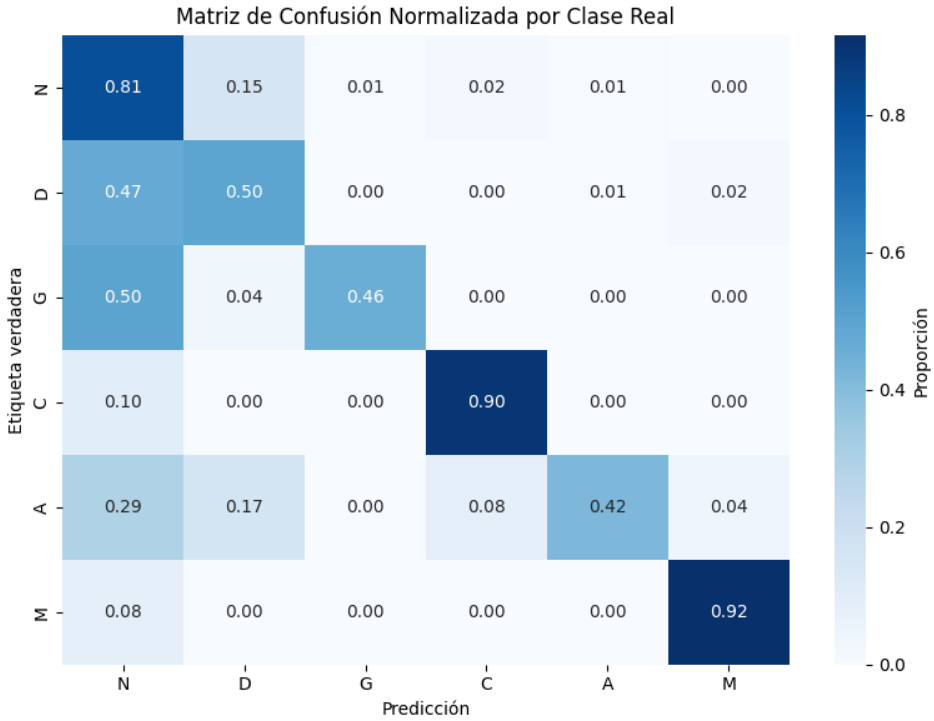

Figura 20. *Matriz de confusión del Modelo 1 con el dataset generado por GANs.*

La Figura 20 muestra la matriz de confusión indicando una clasificación correcta predominante en clases como C, M y N. Las clases minoritarias presentan algunas confusiones, pero se mantienen dentro de un margen razonable, reflejando la efectividad del modelo para detectar patrones relevantes incluso en ejemplos generados artificialmente.

Este bloque evaluó modelos entrenados con datos equilibrados mediante GANs, manteniendo una validación fija con imágenes reales. El Modelo 1, sin técnicas adicionales de mezcla, obtuvo el mejor rendimiento global. MixUp y CutMix aportaron mejoras puntuales en algunas clases, pero no superaron en promedio la estabilidad y eficacia del modelo base.

4.1.4 *Dataset* equilibrado con modelos de difusión

La generación de imágenes sintéticas mediante modelos de difusión permitió construir un conjunto de datos equilibrado para entrenar modelos de clasificación sin recurrir a técnicas adicionales de equilibrado. En este bloque se evaluaron tres configuraciones progresivas, todas basadas en un esquema de *ensemble* de *embeddings* extraídos con RETFound (Zhou, y otros, 2023) y EfficientNetB3 (Tan & Le, 2019). La validación se realizó sobre un conjunto fijo del 10% de imágenes reales por clase, garantizando la coherencia con los bloques anteriores y permitiendo una comparación justa del rendimiento.

Resultados cuantitativos

La Tabla 15 resume las métricas globales obtenidas por cada modelo. El Modelo 1 alcanza el mejor F1 macro, mientras que los Modelos 2 y 3 presentan resultados ligeramente inferiores, especialmente en *recall* medio:

Modelo	**Técnica adicional**	***Accuracy***	**F1 macro**	***Recall* macro**
Modelo 1	*Focal Loss + Ensemble*	**0.67**	**0.65**	**0.62**
Modelo 2	MixUp + *Ensemble*	0.66	0.60	0.58
Modelo 3	CutMix + *Ensemble*	0.65	0.64	0.61

Tabla 15. *Métricas generales de los modelos entrenados con el dataset equilibrado por modelos de difusión.*

La Tabla 16 muestra el F1-*score* por clase. El Modelo 1 consigue el mejor equilibrio general, con buenos resultados en clases complejas como A (0.42) y G (0.50). El Modelo 2, aunque mejora en precisión en algunas clases como G y A, pierde rendimiento en *recall*. El Modelo 3 ofrece valores intermedios, sin superar al Modelo 1 en media.

Modelo	**N**	**D**	**G**	**C**	**A**	**M**
Modelo 1	**0.74**	**0.52**	**0.50**	**0.87**	**0.42**	**0.82**
Modelo 2	0.75	0.44	0.50	0.80	0.29	0.84
Modelo 3	0.70	0.55	0.47	0.88	0.42	0.80

Tabla 16. *F1-score por clase en los modelos entrenados con el dataset equilibrado mediante modelos de difusión.*

Mejor modelo

El Modelo 1 es el que mejor equilibrio global alcanza, tanto en métricas promedio como por clase. A continuación, se presentan sus resultados visuales.

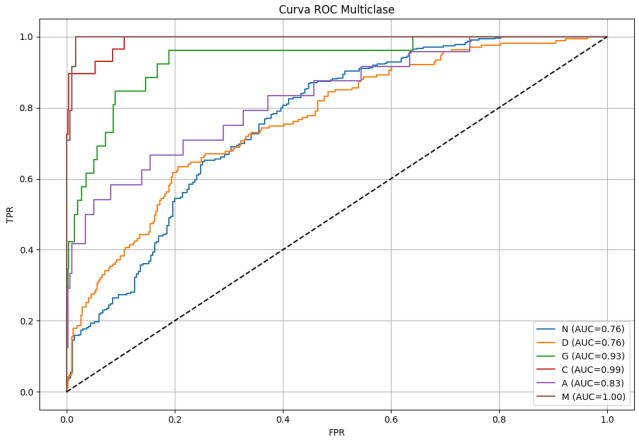

Figura 21*.Curva ROC por clase del Modelo 1 con el dataset equilibrado por modelos de difusión.*

La Figura 21 muestra una gran capacidad de discriminación en clases como M (AUC = 1.00), C (0.99) y G (0.93), junto con valores aceptables en N, D y A (superiores a 0.75), reflejando una buena generalización del modelo en escenarios clínicos realistas.

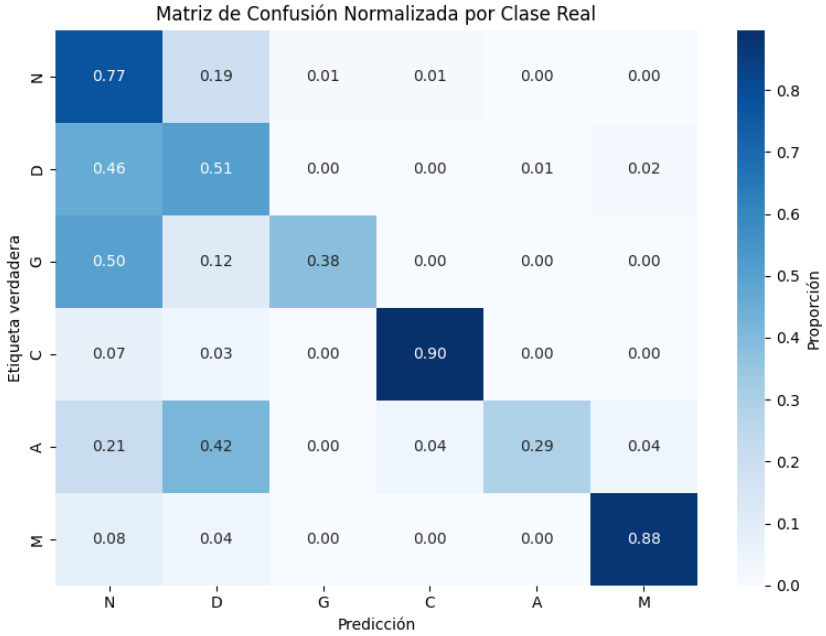

***Figura 22**. Matriz de confusión del Modelo 1 con el dataset equilibrado por modelos de difusión.*

La Figura 22 de la matriz de confusión evidencia una buena tasa de acierto en la mayoría de las clases, especialmente en C, M y N. Aunque persiste cierta confusión entre las clases D y N, el rendimiento sobre clases minoritarias se mantiene alto, validando la utilidad del enfoque basado en modelos de difusión para generar datos sintéticos clínicamente útiles.

Los resultados obtenidos con el *dataset* equilibrado mediante modelos de difusión confirman la efectividad de este enfoque para generar datos sintéticos útiles en tareas de clasificación médica. El Modelo 1, sin técnicas adicionales de regularización, alcanzó el mejor rendimiento global, destacando en precisión y estabilidad entre clases. Aunque MixUp y CutMix introdujeron variabilidad, no lograron mejorar los resultados del modelo base. En conjunto, este bloque demuestra que los modelos de difusión son una herramienta prometedora para afrontar el desequilibrio de clases sin comprometer la calidad del entrenamiento.

4.2 Comparativa entre los mejores modelos de cada *dataset*

En esta sección se comparan los mejores modelos de cada uno de los cuatro bloques de experimentación, con el objetivo de determinar qué enfoque de equilibrio ofrece un mayor rendimiento clasificatorio y una mejor capacidad de generalización. Para garantizar una evaluación justa, todos los modelos comparten una arquitectura común basada en *embeddings* extraídos con **RETFound** (Zhou, y otros, 2023) y **EfficientNetB3** (Tan & Le, 2019), entrenados con *Focal Loss* (Lin, Goyal, Girshick, He, & Dollar, 2018) y validados sobre un conjunto fijo del 10% de imágenes reales por clase.

Los modelos seleccionados en la Tabla 17 son aquellos que obtuvieron el **mayor F1-*score* macro** dentro de su bloque:

Dataset	**Técnica de equilibrado**	***Accuracy***	**F1 macro**	***Recall* macro**	**AUC promedio**
Original desequilibrado	CutMix + SMOTE sobre *embeddings*	0.67	0.71	0.70	~0.91
Equilibrado con transformaciones tradicionales	MixUp sobre *embeddings*	0.69	0.67	0.64	~0.95
Equilibrado con imágenes GAN	Sin MixUp/CutMix	0.69	0.68	0.67	~0.96
Equilibrado con imágenes de difusión	Sin MixUp/CutMix	0.67	0.65	0.62	~0.96

Tabla 17. *Comparativa entre los mejores modelos de cada tipo de dataset.*

Los resultados revelan que, contra lo esperado, el mejor modelo en términos de F1 macro se obtuvo entrenando directamente sobre el *dataset* original desequilibrado, utilizando una combinación de SMOTE y CutMix sobre los *embeddings*. Este modelo no solo superó en rendimiento a las configuraciones basadas en *datasets* equilibrados con imágenes sintéticas, sino que también mantuvo un buen compromiso entre precisión y *recall*, incluso en clases minoritarias.

Por otro lado, los *datasets* equilibrados mediante imágenes generadas con GANs o modelos de difusión mostraron un rendimiento más estable y altos valores de AUC promedio (~0.96), lo que indica una gran capacidad discriminativa. No obstante, estos modelos no lograron superar el F1 macro del modelo entrenado sobre datos reales desequilibrados con técnicas avanzadas de mezcla y sobremuestreo.

En resumen, aunque las imágenes sintéticas ofrecen una vía prometedora para abordar el desequilibrio de clases, en este caso el modelo más eficaz fue aquel que combinó directamente estrategias de equilibrio sobre el *dataset* original. Este hallazgo destaca la importancia de aprovechar al máximo los datos reales disponibles y refuerza el valor de enfoques simples, pero bien optimizados en contextos clínicos.

4.3 Evaluación cualitativa mediante Grad-CAM

Además de la evaluación cuantitativa mediante métricas tradicionales, resulta fundamental analizar si las decisiones de los modelos de clasificación están justificadas desde el punto de

vista clínico. Para ello, se ha utilizado la técnica de *Gradient-weighted Class Activation Mapping* (Grad-CAM) (Selvaraju, y otros, 2017), que permite visualizar las regiones de una imagen que han influido más en la predicción final del modelo.

Este enfoque cualitativo aporta información clave sobre el comportamiento interno de las redes neuronales, revelando si se apoyan en estructuras anatómicamente relevantes o si, por el contrario, tienden a fijarse en artefactos o regiones irrelevantes. Se seleccionaron los mejores modelos de cada uno de los cuatro escenarios experimentales (*dataset* desequilibrado, equilibrado con transformaciones tradicionales, equilibrado con GANs y equilibrado con difusión) y se aplicó Grad-CAM sobre dos imágenes por modelo: una correctamente clasificada y otra erróneamente clasificada.

El objetivo es analizar si las activaciones observadas en los mapas de calor son coherentes con lo esperado desde el punto de vista médico, y si existen patrones comunes de error que puedan atribuirse a las limitaciones de cada enfoque de equilibrio de datos.

4.3.1 *Dataset* original desequilibrado

Con el objetivo de interpretar visualmente las decisiones del modelo entrenado sobre el conjunto de datos original desequilibrado, se aplicó la técnica de Grad-CAM al mejor modelo de este bloque experimental (Modelo 4, basado en CutMix sobre *embeddings* y *ensemble* ponderado).

Esta técnica permite visualizar qué regiones de la imagen han contribuido más a la predicción final del modelo, facilitando así una evaluación cualitativa de su comportamiento. Se seleccionaron dos ejemplos representativos: uno correctamente clasificado y otro mal clasificado, con el fin de analizar la coherencia clínica de las activaciones.

Ejemplo 1: Clasificación correcta (Clase A – Degeneración Macular)

En este caso, en la Figura 23, el modelo identificó correctamente una imagen como perteneciente a la clase A. El mapa de activación destaca principalmente la zona central del fondo de ojo, coherente con la localización habitual de las lesiones características de la degeneración macular asociada a la edad. La atención del modelo se concentra en regiones anatómicamente relevantes, lo que respalda la fiabilidad de la predicción.

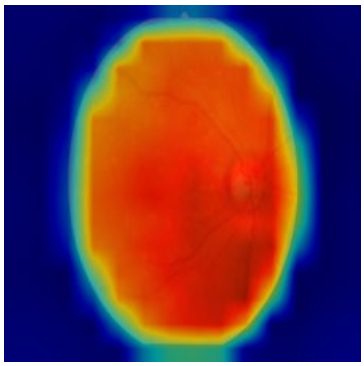

***Figura 23**. Grad-CAM – Clasificación correcta clase A (Modelo desequilibrado).*

Ejemplo 2: Clasificación incorrecta (Clase A predicha como N)

En el segundo ejemplo, la Figura 24 correspondiente a la clase A fue erróneamente clasificada como Normal (N). El mapa de activación muestra una atención dispersa, centrada en regiones periféricas sin una clara relevancia clínica. Esta distribución sugiere una falta de sensibilidad del modelo para detectar patrones sutiles asociados a la degeneración macular, posiblemente derivada del desequilibrio en el conjunto de entrenamiento, que dificulta el aprendizaje de características representativas de las clases minoritarias.

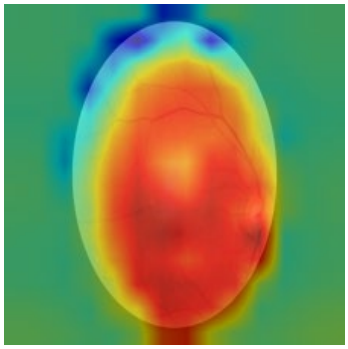

***Figura** 24. Grad-CAM – Clasificación incorrecta clase A predicha como N (Modelo desequilibrado).*

Estos resultados cualitativos coinciden con los hallazgos cuantitativos: aunque el modelo alcanza un rendimiento general competitivo, sigue mostrando dificultades al clasificar correctamente clases poco representadas. La aplicación de Grad-CAM permite comprobar que, en algunos casos, el modelo se basa en regiones visualmente coherentes, mientras que en otros puede desviarse hacia zonas irrelevantes, reflejando las limitaciones de aprender a partir de un *dataset* originalmente desequilibrado.

4.3.2 *Dataset* equilibrado con técnicas tradicionales

Para evaluar la coherencia clínica del mejor modelo entrenado sobre el *dataset* equilibrado mediante transformaciones tradicionales, se aplicó la técnica de Grad-CAM sobre dos imágenes representativas: una correctamente clasificada y otra mal clasificada por el Modelo 3 (basado en *ensemble* y regularización mediante MixUp (Zhang, Cisse, Dauphin, & Lopez-Paz, 2018)).

Este análisis cualitativo permite identificar si las decisiones del modelo se apoyan en estructuras relevantes desde el punto de vista médico o si, por el contrario, reflejan patrones espurios o artefactos.

Ejemplo 1: Clasificación correcta (Clase A)

En la Figura 25, correspondiente a un caso de Degeneración Macular Asociada a la Edad (A), el modelo emitió una predicción correcta. El mapa de activación generado por Grad-CAM revela una atención intensa en la región central del fondo de ojo, especialmente en torno a la mácula, zona anatómica donde suelen manifestarse los signos clínicos de esta patología. Esta focalización es coherente desde el punto de vista oftalmológico y respalda la fiabilidad del modelo.

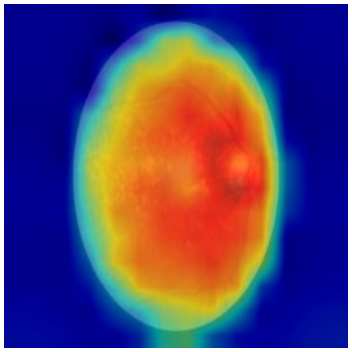

Figura 25. *Grad-CAM – Clasificación correcta, clase A (Modelo con transformaciones tradicionales).*

Ejemplo 2: Clasificación incorrecta (Clase A predicha como N)

En este segundo ejemplo, la Figura 26 etiquetada como A fue clasificada erróneamente como Normal (N). El mapa de calor muestra una atención más difusa, sin una localización clara en la región macular, lo que sugiere que el modelo no identificó correctamente los patrones sutiles asociados a la enfermedad. Esta confusión podría explicarse por la apariencia aparentemente regular de algunas imágenes de AMD en fases iniciales, así como por una variabilidad en la expresión visual de la patología.

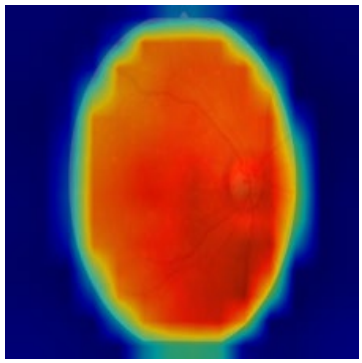

Figura 26. *Grad-CAM – Clasificación incorrecta, clase A predicha como N (Modelo con transformaciones tradicionales MixUp).*

En conjunto, estos resultados muestran que el modelo basado en MixUp tiende a centrarse en regiones clínicas relevantes cuando acierta, aunque sigue mostrando cierta debilidad al distinguir patrones más sutiles de enfermedades menos representadas. Grad-CAM permite evidenciar estas diferencias de comportamiento y refuerza la utilidad de las técnicas cualitativas en la evaluación de modelos de clasificación médica.

4.3.3 *Dataset* equilibrado con GANs

Para el análisis cualitativo del mejor modelo entrenado con el *dataset* equilibrado mediante imágenes sintéticas generadas por GANs (Modelo 1, sin MixUp ni CutMix), se aplicó la técnica Grad-CAM con el objetivo de examinar la coherencia clínica de las regiones activadas por la red durante la predicción.

Ejemplo 1: Clasificación correcta (Clase A)

La Figura 27 muestra un mapa de calor correspondiente a una imagen correctamente clasificada como AMD (A). El modelo focaliza su atención en la zona macular, lo que es clínicamente relevante dado que las manifestaciones de degeneración macular suelen localizarse en esa región. Esta activación indica que el modelo es capaz de identificar patrones visuales anatómicamente consistentes, incluso cuando ha sido entrenado con imágenes generadas artificialmente.

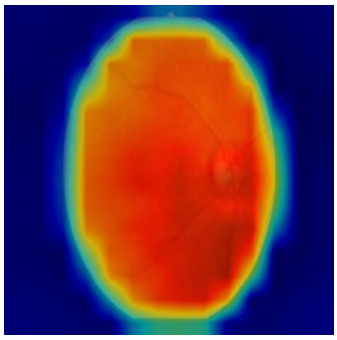

Figura 27. *Grad-CAM para una predicción correcta en clase A (Modelo con GAN sin MixUp).*

Ejemplo 2: Clasificación incorrecta (Clase A predicha como D)

En la Figura 28 se observa una imagen de la clase A que fue erróneamente clasificada como Diabetes (D). El mapa de activación se concentra en áreas periféricas donde podrían existir leves irregularidades en el patrón vascular. Esta dispersión de atención sugiere que el modelo podría estar confundiendo artefactos o estructuras no patológicas con signos típicos de retinopatía diabética, como microhemorragias. Este tipo de error puede deberse a la variabilidad en la calidad visual de las imágenes sintéticas generadas, así como a similitudes texturales entre clases.

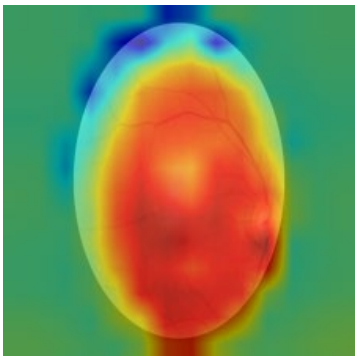

***Figura 28**. Grad-CAM para una predicción incorrecta: clase A clasificada como D (Modelo con GAN sin MixUp).*

Los mapas generados muestran que el modelo tiende a enfocar su atención en zonas relevantes en los casos correctos, validando la utilidad de las imágenes generadas con GANs. No obstante, en los errores de clasificación persisten ambigüedades, especialmente entre clases con solapamiento visual, lo que sugiere la necesidad de estrategias adicionales para mejorar la especificidad del modelo.

4.3.4 *Dataset* equilibrado con modelos de difusión

Para evaluar la coherencia clínica del modelo entrenado sobre el *dataset* equilibrado con imágenes generadas mediante modelos de difusión, se aplicó Grad-CAM al mejor clasificador de este bloque (Modelo 1, sin MixUp ni CutMix). Este modelo, basado en un *ensemble* de *embeddings* de RETFound y EfficientNetB3, fue validado con un conjunto de imágenes reales independientes. A continuación, se presentan dos ejemplos ilustrativos.

Ejemplo 1: Clasificación correcta (Clase A)

La primera imagen presentada en la Figura 29 corresponde a un caso correctamente clasificado como Degeneración Macular (A). El mapa de calor muestra una activación centrada en la región macular del fondo de ojo, con una atención distribuida de forma anatómicamente plausible. Esto sugiere que el modelo ha aprendido a focalizarse en las zonas donde suelen manifestarse los signos clínicos asociados a esta patología, como alteraciones pigmentarias o drusas, lo que refuerza la fiabilidad de su predicción.

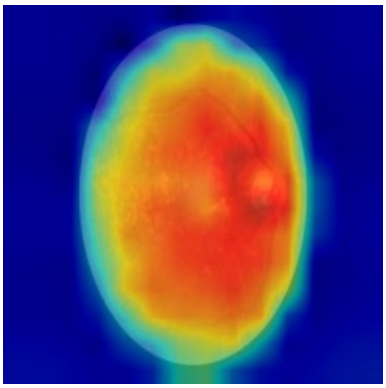

***Figura 29.** Grad-CAM de una imagen correctamente clasificada como A. Modelo con modelos de difusión.*

Ejemplo 2: Clasificación incorrecta (Clase A predicha como N)

En este segundo caso mostrado en la Figura 30, una imagen de la clase A fue erróneamente clasificada como Normal (N). El Grad-CAM revela una activación difusa, más extendida por la periferia del fondo de ojo, sin un foco claro en la región macular. Esta falta de atención dirigida podría haber llevado al modelo a ignorar señales relevantes de la enfermedad, lo cual evidencia una limitación en su sensibilidad ante ciertos patrones sutiles, incluso cuando el *dataset* está equilibrado mediante imágenes generadas artificialmente.

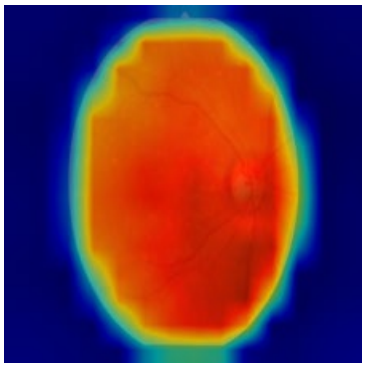

Figura 30. *Grad-CAM de una imagen de clase A mal clasificada como N. Modelo con modelos de difusión.*

Estos resultados cualitativos muestran que, en general, el modelo entrenado con datos de difusión tiende a centrarse en zonas relevantes, aunque aún puede fallar en detectar alteraciones menos evidentes. No obstante, la activación en estructuras clínicas significativas en las predicciones correctas confirma que este enfoque de generación sintética no distorsiona los patrones anatómicos esenciales para la tarea de clasificación.

4.3.5 Discusión de hallazgos y limitaciones

La evaluación cualitativa mediante Grad-CAM ha permitido analizar si las predicciones de los modelos se fundamentan en regiones anatómicamente coherentes, especialmente en la detección de patologías minoritarias como la Degeneración Macular Asociada a la Edad (A). En general, los mejores modelos de cada bloque muestran activaciones centradas en la región macular en los casos correctamente clasificados, lo que valida parcialmente la interpretación clínica de sus decisiones.

Sin embargo, en las clasificaciones erróneas se observan patrones comunes de atención difusa o desplazada hacia zonas periféricas, lo que sugiere una limitada sensibilidad ante manifestaciones sutiles o una posible confusión con estructuras similares entre clases (por ejemplo, N y D). Estas limitaciones fueron especialmente evidentes en los modelos entrenados con el *dataset* original desequilibrado, donde la escasa representación de clases minoritarias dificultó el aprendizaje de características discriminativas.

Aunque los modelos entrenados con datos generados por GANs y difusión demostraron una mayor estabilidad visual y mejores activaciones en predicciones correctas, todavía presentaron errores atribuibles a la ambigüedad visual de ciertas clases o a la variabilidad introducida por las imágenes sintéticas.

En conjunto, Grad-CAM ha servido como herramienta complementaria para evaluar la interpretabilidad de los modelos y ha permitido detectar posibles sesgos de atención. No obstante, su interpretación sigue siendo cualitativa y dependiente del contexto clínico, por lo que se recomienda su uso junto con otras técnicas de explicabilidad y una validación experta para su aplicación en entornos médicos reales.

Capítulo 5

Conclusiones

El presente Trabajo de Fin de Grado tuvo como punto de partida la hipótesis de que ampliar un conjunto de datos desequilibrado mediante distintas técnicas de aumento y generación sintética permitiría entrenar modelos de clasificación más eficaces, con un mejor rendimiento al enfrentarse a imágenes reales. Sin embargo, los resultados experimentales obtenidos han revelado una realidad más matizada: no siempre un mayor volumen de datos sintéticos conduce a una mejora clasificatoria sustancial, y, en algunos casos, estrategias aplicadas directamente sobre el *dataset* original desequilibrado —como CutMix y SMOTE— ofrecieron un rendimiento superior al de los modelos entrenados sobre *datasets* equilibrados mediante generación de imágenes.

Esta observación no debilita el valor del trabajo, sino que refuerza su rigor científico, ya que demuestra una evaluación crítica de la hipótesis planteada y una apertura a conclusiones basadas en la evidencia. Una posible explicación de estos resultados puede residir en las **limitaciones computacionales disponibles durante el desarrollo del proyecto**, que condicionaron la resolución, cantidad y fidelidad de las imágenes sintéticas generadas. Es razonable pensar que, **con mayores recursos y modelos generativos más potentes, la calidad de las imágenes sintéticas podría aumentar considerablemente, impactando positivamente en el rendimiento de los clasificadores entrenados con ellas**.

El análisis exhaustivo de distintas estrategias de equilibrio, combinadas con técnicas modernas como *Focal Loss*, MixUp y Grad-CAM, ha permitido no solo comparar el impacto de cada enfoque desde un punto de vista cuantitativo, sino también explorar la coherencia cualitativa de las predicciones a nivel clínico. En conjunto, los resultados de este TFG aportan información útil para futuras investigaciones en clasificación de imágenes médicas, y destacan la importancia de no asumir que más datos generados implican automáticamente mejores modelos, sino de analizar cuidadosamente la calidad, la técnica empleada y el contexto clínico de aplicación.

5.1 Cumplimiento de los objetivos

Los objetivos planteados al inicio del trabajo han sido plenamente alcanzados:

- **Objetivo general**: Se ha desarrollado un estudio comparativo que evalúa el impacto de diversas técnicas de aumento de datos —tradicionales, avanzadas y generativas— sobre el rendimiento de modelos de clasificación en un conjunto de imágenes médicas oculares desequilibrado.

- **Objetivos específicos:**

1. Se han implementado transformaciones tradicionales (rotaciones, cambios de brillo, escalado, traslaciones, ruido gaussiano, etc.) para aumentar la diversidad de las clases minoritarias y generar una versión equilibrada del *dataset* original.
2. Se han aplicado técnicas avanzadas como MixUp y CutMix sobre *embeddings* previamente extraídos, y se ha analizado su efecto en distintos bloques experimentales.
3. Se han entrenado modelos generativos por clase minoritaria, concretamente una WGAN-GP y un modelo de difusión DDPM, logrando generar imágenes sintéticas para clases poco representadas.
4. Se ha evaluado el rendimiento de varios modelos de clasificación basados en RETFound y EfficientNetB3 sobre los distintos *datasets* equilibrados, utilizando métricas como precisión, sensibilidad, F1-*score*, AUC-ROC y curvas ROC por clase.
5. Se ha analizado la calidad de las imágenes generadas mediante métricas cuantitativas como FID e IS, y se ha reforzado esta evaluación con interpretaciones visuales mediante Grad-CAM sobre modelos entrenados con distintos tipos de datos.

En conjunto, los resultados obtenidos muestran que la aplicación estratégica de técnicas de aumento y generación de datos puede contribuir a mejorar la capacidad de los modelos para reconocer ciertas patologías oculares minoritarias. No obstante, esta mejora no es uniforme ni garantizada: el impacto depende en gran medida de la calidad de las imágenes sintéticas, del enfoque de equilibrado empleado y del contexto del modelo. Aun así, el trabajo aporta una visión crítica y valiosa sobre cómo diferentes estrategias afectan al rendimiento clasificatorio, lo cual representa una aportación útil en entornos clínicos donde la escasez de datos es una limitación habitual.

5.2 Limitaciones del trabajo

A pesar de los logros obtenidos, este trabajo también ha enfrentado diversas limitaciones técnicas y operativas:

- **Restricciones de memoria RAM y GPU**: El entrenamiento de modelos generativos, especialmente los modelos de difusión, ha estado condicionado por la disponibilidad limitada de recursos computacionales. Esto obligó a reducir la resolución de las imágenes, simplificar arquitecturas y ajustar el *batch size* para evitar errores de ejecución.
- **Tiempo de entrenamiento**: El coste computacional de entrenar modelos de difusión y modelos generativos ha sido elevado, lo que limitó la posibilidad de realizar pruebas más amplias, ajustes finos de hiperparámetros y validaciones cruzadas más extensas.
- **Limitaciones clínicas**: Aunque se ha cuidado que las imágenes generadas y las transformaciones aplicadas mantuvieran su coherencia anatómica, a pesar de que no se consiguió la calidad deseada, no se contó con validación por parte de profesionales clínicos, lo cual impide asegurar totalmente su utilidad diagnóstica en la práctica.

5.3 Trabajo futuro

A partir de los resultados obtenidos y las limitaciones encontradas, se identifican varias líneas de trabajo que podrían desarrollarse en el futuro:

- **Validación clínica de las imágenes sintéticas**: Un siguiente paso importante sería evaluar las imágenes generadas mediante GANs y modelos de difusión con la ayuda de especialistas en oftalmología, para comprobar si son clínicamente válidas y útiles como soporte en el diagnóstico.
- **Clasificación *multilabel***: Aunque este trabajo se centró en una versión simplificada del problema (clasificación *single-label*), una línea futura interesante sería abordar la clasificación *multilabel*, que se ajusta mejor a la realidad clínica donde un mismo paciente puede presentar múltiples patologías.
- **Aumento de resolución y calidad**: Explorar la generación de imágenes a mayor resolución, especialmente en los modelos de difusión, podría mejorar aún más el rendimiento de los clasificadores y aumentar la fidelidad de los detalles anatómicos.
- **Optimización de eficiencia computacional**: Dado el elevado coste de entrenamiento de algunos modelos, como los DDPM, sería interesante investigar técnicas de aceleración o versiones ligeras para hacer viable su implementación en entornos clínicos con recursos limitados.
- **Generalización a otros dominios médicos**: Finalmente, sería relevante evaluar si el *pipeline* desarrollado es aplicable o adaptable a otras especialidades médicas donde también exista desequilibrio de clases, como dermatología o radiología.

5.4 Relación con los estudios cursados

Este Trabajo de Fin de Grado se apoya de manera transversal en los conocimientos adquiridos a lo largo del Grado en Ciencia de Datos, integrando competencias técnicas, metodológicas y analíticas que han sido fundamentales para su desarrollo.

En primer lugar, asignaturas del bloque de *aprendizaje automático* como **Modelos descriptivos y predictivos I y II**, **Técnicas escalables en aprendizaje automático** y **Evaluación, despliegue y monitorización de modelos** han aportado las bases teóricas y prácticas necesarias para entrenar y evaluar clasificadores con métricas como precisión, *recall*, F1-*score* y AUC, además de implementar funciones de pérdida especializadas como *Focal Loss* y estrategias de *ensemble*.

La parte experimental del trabajo, centrada en la generación de imágenes sintéticas mediante GANs y modelos de difusión, encuentra su fundamento directo en la optativa **Análisis de imágenes y vídeos**, que ha proporcionado los conocimientos necesarios para entender las particularidades de las imágenes médicas y aplicar modelos generativos en el contexto clínico. Esta asignatura también favoreció el desarrollo de competencias transversales como la

capacidad de analizar visualmente resultados mediante Grad-CAM y detectar patrones anatómicamente relevantes.

En cuanto al diseño y manipulación de los datos, asignaturas como **Adquisición y transmisión de datos**, **Gestión de datos** y **Bases de datos** han sido clave para llevar a cabo el preprocesamiento, limpieza y organización del *dataset* ODIR-5K, incluyendo tareas como la separación de imágenes por ojo, la eliminación de muestras *multilabel* o la integración de rutas y etiquetas.

El trabajo se ha apoyado además en competencias desarrolladas en **Infraestructura para el procesamiento de datos** y **Optimización**, especialmente para gestionar entornos computacionales limitados, reducir tiempos de entrenamiento y ajustar parámetros en contextos de baja disponibilidad de recursos.

Desde el punto de vista metodológico, las asignaturas del bloque de proyectos —**Proyecto I, II y III**— han sido esenciales para estructurar el flujo de trabajo, establecer hipótesis contrastables, documentar resultados y mantener una coherencia metodológica a lo largo del estudio. Por su parte, **Gestión de proyectos** ha contribuido a planificar las fases del trabajo y priorizar tareas de forma eficiente.

Finalmente, la dimensión ética y profesional del TFG ha sido reforzada por la asignatura **Marco profesional, legal y deontológico**, que subraya la responsabilidad en el uso de datos clínicos y el impacto potencial de los modelos de inteligencia artificial en el ámbito sanitario, especialmente en lo que respecta a sesgos, explicabilidad y equidad en el diagnóstico.

En conjunto, este trabajo representa una síntesis práctica de los aprendizajes adquiridos a lo largo del grado, aplicados a un problema real y complejo, con una clara orientación tanto investigadora como profesional.

Bibliografía

Chawla, N., Bowyer, K., Hall , L., & Philip Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *arXiv*, 37. Obtenido de https://arxiv.org/pdf/1106.1813

Chen, J., Yu, H., Feng, R., Chen, D., & Wu, J. (2021). Flow-Mixup: Classifying Multi-labeled Medical Images with Corrupted Labels. *arXiv*, 8. Obtenido de https://arxiv.org/pdf/2102.08148

Dash, A., & Swarnkar, T. (2025). Data-GAN Augmentation Techniques in Medical Image Analysis: A Deep Survey. *SN Computer Science, Vol 6*(348), 21 . Obtenido de https://link.springer.com/article/10.1007/s42979-025-03867-9

Dong, X., & Yang, Y. (2019). Teacher Supervises Students How to Learn From Partially Labeled Images for Facial Landmark Detection. *arXiv*, 10 . Obtenido de https://arxiv.org/pdf/1908.02116

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., . . . Houlsby, N. (2021). AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE. *arXiv*, 22. Obtenido de https://arxiv.org/pdf/2010.11929

Galdran, A., Carneiro, G., & González Ballester, M. (2021). Balanced-MixUp for Highly Imbalanced Medical Image Classification. *arXiv*, 12. Obtenido de https://arxiv.org/pdf/2109.09850

Goceri, E. (2023). Medical image data augmentation: techniques, comparisons and interpretations. *Artificial Intelligence Review, 56*, pp 12561-12605. Obtenido de https://link.springer.com/article/10.1007/s10462-023-10453-z

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* Cambridge, Massachusetts: MIT Press. Obtenido de https://www.deeplearningbook.org/

Goodfellow, I., Pouguet-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). Generative Adversarial Networks. *arXiv*, 9. Obtenido de https://arxiv.org/pdf/1406.2661

Gulrajani, I., Arjovsky, M., Ahmed, F., Dumoulin, V., & Courville, A. (2017). Improved Training of Wasserstein GANs. *arXiv*, 20. Obtenido de https://arxiv.org/pdf/1704.00028

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *arXiv*, 12. Obtenido de https://arxiv.org/pdf/1512.03385

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2018). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *arXiv*, 38 . Obtenido de https://arxiv.org/pdf/1706.08500

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. *arXiv*, 25. Obtenido de https://arxiv.org/pdf/2006.11239

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2018). Focal Loss for Dense Object Detection. *arXiv*, 10. Obtenido de https://arxiv.org/pdf/1708.02002

Liu, C., Fan, F., Schwarz, A., & Maier, A. (2024). Cut to the Mix: Simple Data Augmentation Outperforms Elaborate Ones in Limited Organ Segmentation Datasets. *MICCAI 2024. LNCS, vol 15008*, pp 145-154.

Müller-Franzes, Gustav; Moritz Niehues, Jan; Khader, Firas; Tayebi Arasteh, Soroosh; Haarburger, Christoph; Kuhl, Christiane; Wang, Tianci; Han, Tianyu; Nebelung, Sven; Nikolas Kather, Jakob; Truhn, Daniel. (2022). DIFFUSION PROBABILISTIC MODELS BEAT GAN ON MEDICAL 2D IMAGES. *arXiv*, 13. Obtenido de https://arxiv.org/pdf/2212.07501

Nanni, L., Paci, M., Brahnam, S., & Lumini, A. (2021). Comparison of Different Image Data Augmentation Approaches. *Journal of imaging, 7*(12), 13 . Obtenido de https://doi.org/10.3390/jimaging7120254

Perez, L., & Wang, J. (2017). The Effectiveness of Data Augmentation in Image Classification using Deep Learning. *arXiv*, 8. Obtenido de https://arxiv.org/pdf/1712.04621

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved Techniques for Training GANs. *arXiv*, 10 . Obtenido de https://arxiv.org/pdf/1606.03498

Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *arXiv*, 9 . Obtenido de https://arxiv.org/pdf/1610.02391

Shorten, C., & Khoshgoftaar , T. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*(60), 6 . Obtenido de https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0197-0

Szegedy, C., Vanhoucke, V., Ioffe, S., & Shlens, J. (2015). Rethinking the Inception Architecture for Computer Vision. *arXiv*, 10. Obtenido de https://arxiv.org/pdf/1512.00567

Szegedy, C., Vanhoucke, V., Ioffe, S., Jonathon Shlens, J., & Wojna, Z. (2015). Rethinking the Inception Architecture for Computer Vision. *arXiv*, 10. Obtenido de https://arxiv.org/pdf/1512.00567

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv*, 11 . Obtenido de https://arxiv.org/pdf/1905.11946

Wang, Weibin; Liang, Dong; Chen, Qingquing; Iwamoto, Yutaro; Han, Xian-Hua; Zhang, Qiaowei; Hu, Hongjie; Lin, Lanfen; Chen, Yen-Wei. (2019). Medical Image Classification Using Deep Learning. *Deep Learning in Healthcare, vol 171*, pp 33-51. Obtenido de https://link.springer.com/chapter/10.1007/978-3-030-32606-7_3

Yun, S., Han, D., Joon Oh, S., Chu, S., Choe, J., & Yoo, Y. (2019). CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. *arXiv*, 14. Obtenido de https://arxiv.org/pdf/1905.04899

Zhang, H., Cisse, M., Dauphin, Y., & Lopez-Paz, D. (2018). mixup: BEYOND EMPIRICAL RISK MINIMIZATION. *arXiv*, 13. Obtenido de https://arxiv.org/pdf/1710.09412

Zhang, Juzhao; Lin, Senlin; Cheng, Tianhao; Xu, Yi; Lu, Lina; He, Jiangnan; Yu, Tao; Peng, Yajun; Zou, Haidong; Ma, Yingyan. (2024). RETFound-enhanced community-based fundus disease screening: real-world evidence and decision curve analysis. *npj digital medicine*(108), 7 . Obtenido de https://www.nature.com/articles/s41746-024-01109-5

Zhang, Z., Xie, Y., Xing, F., McGough, M., & Yang, L. (2017). MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network. *arXiv*, 9. Obtenido de https://arxiv.org/pdf/1707.02485

Zhou, Y., Chia, M. A., Wagner, S. K., Ayhan, M. S., Williamson, D. J., Struyven, R. R., . . . Keane, P. A. (2023). A foundation model for generalizable disease detection from retinal images. *Nature*(622), pp 156-163. Obtenido de https://doi.org/10.1038/s41586-023-06555-x

Anexos

ANEXO 1: OBJETIVOS DE DESARROLLO SOSTENIBLE

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenible	**Alto**	**Medio**	**Bajo**	**No procede**
ODS 1. Fin de la pobreza			x	
ODS 2. Hambre cero			x	
ODS 3. Salud y bienestar	x			
ODS 4. Educación de calidad	x			
ODS 5. Igualdad de género			x	
ODS 6. Agua limpia y saneamiento			x	
ODS 7. Energía asequible y no contaminante			x	
ODS 8. Trabajo decente y crecimiento económico			x	
ODS 9. Industria, innovación e infraestructuras	x			
ODS 10. Reducción de las desigualdades			x	
ODS 11. Ciudades y comunidades sostenibles			x	
ODS 12. Producción y consumo responsables			x	
ODS 13. Acción por el clima			x	
ODS 14. Vida submarina			x	
ODS 15. Vida de ecosistemas terrestres			x	
ODS 16. Paz, justicia e instituciones sólidas			x	
ODS 17. Alianzas para lograr los objetivos			x	

En relación con los ODS, este Trabajo de Fin de Grado se vincula especialmente con los objetivos 3, 4 y 9:

- **ODS 3. Salud y bienestar:** Este trabajo contribuye al desarrollo de modelos de inteligencia artificial orientados a mejorar la clasificación de enfermedades oculares, lo cual puede facilitar diagnósticos más precisos y tempranos, beneficiando a pacientes en entornos con escasos recursos clínicos.

- **ODS 4. Educación de calidad:** La realización del proyecto ha implicado una fuerte componente formativa en áreas como *deep learning*, visión por computador y generación de imágenes sintéticas, con potencial de ser reutilizado como material de aprendizaje y experimentación en contextos educativos.

- **ODS 9. Industria, innovación e infraestructuras:** La aplicación de tecnologías innovadoras como GANs y modelos de difusión en el ámbito de la salud digital demuestra cómo la investigación puede impulsar soluciones efectivas, incluso con infraestructuras limitadas. Este enfoque promueve una innovación responsable y accesible en medicina asistida por IA.

ANEXO 2: GitHub al código y a los datos

Link: https://github.com/eoroneoron/TFG_Imagenes