



UNIVERSIDAD
DE GRANADA

Escuela de Posgrado

MÁSTER EN CIENCIA DE DATOS E INGENIERÍA DE
COMPUTADORES

TRABAJO DE FIN DE MÁSTER

**MODELOS GENERATIVOS
DE DIFUSIÓN APLICADOS
AL ÁMBITO DE LA
IMAGEN MÉDICA**

Presentado por:
Álvaro Zorrilla Carriquí

Curso académico 2023-2024

Modelos generativos de difusión aplicados al ámbito de la imagen médica

Álvaro Zorrilla Carriquí

Palabras clave: modelos de difusión, imagen médica, generación de imagen, métricas de calidad de generación.

Diffusion models applied to medical image field

Álvaro Zorrilla Carriquí

Keywords: diffusion models, medical image, image generation, generation quality metrics.

Álvaro Zorrilla Carriquí *Modelos generativos de difusión aplicados al ámbito de la imagen médica.*

Trabajo de Fin de Máster. Curso académico 2023-2024.

Responsables de tutorización

Ignacio Álvarez Illán
Instituto Andaluz Interuniversitario en Ciencia de Datos e Inteligencia Computacional (DaSCI). Departamento de Teoría de la Señal, Redes y Comunicaciones

Máster en Ciencia de datos e Ingeniería de Computadores
Escuela de Posgrado
Universidad de Granada

Fermín Segovia Román
Instituto Andaluz Interuniversitario en Ciencia de Datos e Inteligencia Computacional (DaSCI). Departamento de Teoría de la Señal, Redes y Comunicaciones

DECLARACIÓN DE ORIGINALIDAD

D. Álvaro Zorrilla Carriquí

Declaro explícitamente que el trabajo presentado como Trabajo de Fin de Máster (TFM), correspondiente al curso académico 2023-2024, es original, entendido esto en el sentido de que no he utilizado para la elaboración del trabajo fuentes sin citarlas debidamente.

En Granada a 8 de marzo de 2025

Fdo: Álvaro Zorrilla Carriquí

*A mi madre,
porque de verla a ella,
soñaba con ser médico de pequeño.*

Índice general

| | |
|---|-------------|
| Agradecimientos | xii |
| Summary | xiii |
| Introducción | xvii |
| 1 Modelos de difusión y la imagen médica | 1 |
| 1.1 ¿Qué son los modelos de difusión? | 1 |
| 1.2 Uso de los modelos de difusión en imagen médica | 3 |
| 2 Fundamento teórico de los modelos de difusión | 7 |
| 2.1 Perspectiva variacional | 7 |
| 2.1.1 Denoising Probabilistic Diffusion Models (<i>DDPMs</i>) | 7 |
| 2.2 Perspectiva de <i>score</i> | 15 |
| 2.2.1 Noise-Conditioned Score Networks (<i>NCSNs</i>) | 17 |
| 2.2.2 Score SDEs | 19 |
| 2.3 Conditional Diffusion Probabilistic Models (<i>CDPMs</i>) | 21 |
| 3 Métricas de calidad en la generación de imágenes | 23 |
| 3.1 Métricas de calidad cualitativas | 25 |
| 3.2 Métricas de calidad cuantitativas | 27 |
| 3.2.1 Inception Score (IS) | 27 |
| 3.2.2 Fréchet Inception Distance (FID) | 29 |
| 3.2.3 SSIM | 30 |
| 3.2.4 Proporción Máxima de Señal a Ruido (PSNR) | 32 |
| 3.2.5 Classification Accuracy Score (CAS) | 33 |
| 4 Extractores de características | 35 |
| 4.1 U-Net | 36 |
| 4.2 ResNet50 | 37 |
| 4.3 SkinLesNet | 39 |
| 5 Trabajo experimental y resultados | 41 |
| 5.1 Generación de imagen monocromática. Imágenes de <i>Hand</i> de MedNIST. Métricas de calidad | 42 |

Índice general

| | | |
|-------|--|-----------|
| 5.2 | Generación de imagen a color. Imágenes de lesiones de piel de ISIC | 46 |
| 5.2.1 | <i>Datasets</i> y preprocesamiento | 47 |
| 5.2.2 | En busca de la mejor estrategia | 48 |
| 5.2.3 | Modelos de difusión generativos incondicionales | 52 |
| 5.3 | Cálculo de las métricas de calidad para las imágenes a color | 54 |
| 5.3.1 | Entrenamiento de los extractores de características | 56 |
| 5.3.2 | Métricas FID y SSIM | 59 |
| 5.3.3 | Métrica CAS | 64 |
| | Conclusiones y trabajo futuro | 69 |
| | Bibliografía | 73 |

Índice de figuras

| | | |
|-----|---|----|
| 1.1 | Esquema flujo de trabajo GAN y VAE | 2 |
| 1.2 | Trilema de los modelos generativos | 3 |
| 1.3 | Número de publicaciones de los modelos de difusión aplicados a imagen médica según el año y relevancia | 4 |
| 1.4 | Proceso de difusión de un ejemplo de imagen médica | 5 |
| 2.1 | Esquema DDPM | 10 |
| 2.2 | Esquema red U-Net para el modelo de difusión | 16 |
| 2.3 | Esquema cálculo de la función de <i>score</i> | 18 |
| 2.4 | Esquema <i>Annealed Langevin Dynamics</i> | 19 |
| 2.5 | Esquema DDPM condicional | 22 |
| 3.1 | Funcionamiento IS | 28 |
| 4.1 | Esquema de la arquitectura de red U-Net para segmentación | 36 |
| 4.2 | Esquema de la arquitectura de la red ResNet50. | 38 |
| 4.3 | Esquema de arquitectura de red SkinLesNet para clasificación | 39 |
| 5.1 | Imágenes del conjunto de datos de manos de MedNIST | 43 |
| 5.2 | Imágenes de MedNIST alteradas para el entrenamiento | 44 |
| 5.3 | Evolución del entrenamiento del modelo generador de imágenes de manos de MedNIST | 45 |
| 5.4 | Algunas imágenes generadas durante el entrenamiento del modelo de difusión DDPM para observar la mejora cualitativa de las imágenes sintetizadas a lo largo de dicho proceso. Las imágenes están ordenadas según la época de su generación. | 46 |
| 5.5 | Muestra gráfica del proceso de eliminación de ruido (<i>denoising</i>) para la generación de imágenes. | 46 |
| 5.6 | Ejemplos de imágenes generadas por el modelo de difusión una vez finalizado el entrenamiento del mismo. | 47 |
| 5.7 | Muestra del conjunto de imágenes de entrenamiento de la clase <i>malignant</i> obtenidos de ISIC 2018, 2019 y 2020. | 49 |
| 5.8 | Muestra del conjunto de imágenes de entrenamiento de la clase <i>benign</i> obtenidos de ISIC 2018, 2019 y 2020. | 49 |
| 5.9 | Muestra del conjunto de imágenes de entrenamiento para la segmentación de las lesiones obtenidas de ISIC 2016. | 50 |

Índice de figuras

| | | |
|------|---|----|
| 5.10 | <i>Data Augmentation</i> sobre ISIC 2016 | 50 |
| 5.11 | <i>Samplings</i> durante el entrenamiento del DDPM | 51 |
| 5.12 | Evolución del entrenamiento del modelo generador de imágenes malignas | 53 |
| 5.13 | Evolución del entrenamiento del modelo generador de imágenes benignas | 54 |
| 5.14 | Imágenes generadas durante el entrenamiento del modelo de lesiones malignas dispuestas en orden cronológico de generación. | 55 |
| 5.15 | Imágenes generadas durante el entrenamiento del modelo de lesiones benignas o desconocidas dispuestas en orden cronológico de generación. | 56 |
| 5.16 | Evolución del entrenamiento de la red ResNet50 para extractor | 58 |
| 5.17 | Evolución del entrenamiento de la red SkinLesNet para extractor | 59 |
| 5.18 | Evolución del entrenamiento de la red U-Net para extractor | 60 |
| 5.19 | Ejemplos de segmentación de imágenes de test con la red U-Net entrenada | 61 |
| 5.20 | Muestra del conjunto de 500 imágenes generadas usando el modelo DDPM incondicional entrenado con imágenes de lesiones malignas. | 62 |
| 5.21 | Muestra del conjunto de 500 imágenes generadas usando el modelo DDPM incondicional entrenado con imágenes de lesiones benignas o de tipo desconocido. | 63 |

Índice de tablas

| | | |
|-----|---|----|
| 5.1 | Valores métricas de validación en test para SkinLesNet y ResNet50 como posteriores extractores de características | 59 |
| 5.2 | Valores del SSIM y FID para las imágenes generadas de lesiones malignas y benignas de piel | 64 |
| 5.3 | Métricas de clasificación en test de SkinLesNet para el CAS | 66 |

Agradecimientos

Quiero expresar mi más sincero agradecimiento a todas las personas que, de una u otra manera, han contribuido a la realización de este trabajo de fin de máster.

En primer lugar, agradezco profundamente a mis tutores de Trabajo de Fin de Máster, Ignacio Álvarez Illán y Fermín Segovia por su inestimable guía y paciencia a lo largo de este proyecto. Su experiencia y consejos han sido fundamentales para el desarrollo y la culminación de este trabajo. Especialmente, a Ignacio por ofrecerse a ser mi futuro director de tesis, en caso de conseguir alguna beca.

Agradezco también a los profesores y compañeros del máster, cuyas enseñanzas y colaboraciones han enriquecido mi formación académica y personal. Sus aportaciones han sido vitales para ampliar mis conocimientos y perspectivas en este campo.

A mis amigos y compañeros, gracias por su constante apoyo, motivación y por los momentos compartidos que han aligerado el peso del trabajo y el estudio.

Finalmente, quiero dedicar un especial agradecimiento a mi familia, por su amor incondicional, comprensión y apoyo durante todos estos años de estudio. Les debo la fuerza y la determinación para alcanzar mis sueños.

Summary

The rapid advancement of artificial intelligence (AI) has led to significant changes in numerous fields, with medical imaging emerging as a particularly benefited area. The integration of AI into medical imaging is revolutionizing diagnostic processes, improving disease detection accuracy, and optimizing patient outcomes. Among the AI techniques applied to this field, generative diffusion models have garnered considerable attention due to their ability to generate high-quality images and extract meaningful features, as well as enhance the quality of existing images and generate synthetic data to train other AI algorithms.

In the context of medical imaging, image quality and coherence are fundamental for effective diagnostics and treatments. This study explores the application of diffusion models in the generation and enhancement of medical images, focusing on the different methods available to evaluate the samples quality and structural adequacy. The research analyzes the theoretical foundations of diffusion models, the mathematical formulations of the main three different approaches, and operating principles, as well as their performance compared to other methods.

Denoising Diffusion Probabilistic Models are one of the sort of diffusion models used in literature to generate image samples. For such a reason, this is the approach used in this dissertation to train various generative medical image models, each one specialized in a kind of medical image.

The study employs image quality evaluation metrics, including Structural Similarity Index (SSIM), Fréchet Inception Distance (FID) or Classification Accuracy Score (CAS), to ensure that the generated images trained by the diffusion models meet necessary clinical standards. Additionally, the methodology for implementing diffusion models is detailed, describing the datasets used, preprocessing steps, specific architectures of the models, and experimental configurations necessary for reproducibility of the results.

The findings underscore the promising potential of diffusion models to improve the quality of medical images, facilitate more accurate and efficient diagnoses, and personalize medical treatments. The ability of these models to generate high-quality synthetic data is especially useful in areas with limited availability of labeled data. So, diffusion models may offer multiple benefits that can transform the field of medical imaging, promoting better health management and more favorable patient outcomes.

Resumen

El rápido avance de la inteligencia artificial (IA) ha llevado a cambios significativos en numerosos campos, siendo la imagen médica una área particularmente beneficiada. La integración de la IA en la imagen médica está revolucionando los procesos de diagnóstico, mejorando la precisión en la detección de enfermedades y optimizando los resultados para los pacientes. Entre las técnicas de IA aplicadas a este campo, los modelos de difusión generativa han ganado considerable atención debido a su capacidad para generar imágenes de alta calidad y extraer características significativas, así como mejorar la calidad de las imágenes existentes y generar datos sintéticos para entrenar otros algoritmos de IA.

En el contexto de la imagen médica, la calidad y coherencia de las imágenes son fundamentales para diagnósticos y tratamientos efectivos. Este estudio explora la aplicación de los modelos de difusión en la generación y mejora de imágenes médicas, enfocándose en los diferentes métodos disponibles para evaluar la calidad de las muestras y la adecuación estructural. La investigación analiza los fundamentos teóricos de los modelos de difusión, las formulaciones matemáticas de los tres enfoques principales y los principios operativos, así como su rendimiento en comparación con otros métodos.

Los Modelos de Difusión Probabilística de Reducción de Ruido son uno de los tipos de modelos de difusión utilizados en la literatura para generar muestras de imágenes. Por tal motivo, este es el enfoque utilizado en esta tesis para entrenar varios modelos generativos de imágenes médicas, cada uno especializado en un tipo de imagen médica.

El estudio emplea métricas de evaluación de la calidad de la imagen, incluyendo el Índice de Similitud Estructural (SSIM), la Distancia de *Inception* de Fréchet (FID) o el Score del Accuracy de Clasificación (CAS), para asegurar que las imágenes generadas entrenadas por los modelos de difusión cumplan con los estándares clínicos necesarios. Además, se detalla la metodología para implementar los modelos de difusión, describiendo los conjuntos de datos utilizados, los pasos de preprocesamiento, las arquitecturas específicas de los modelos y las configuraciones experimentales necesarias para la reproducibilidad de los resultados.

Los hallazgos subrayan el potencial prometedor de los modelos de difusión para mejorar la calidad de las imágenes médicas, facilitar diagnósticos más precisos y eficientes, y personalizar los tratamientos médicos. La capacidad de estos modelos para generar datos sintéticos de alta calidad es especialmente útil en áreas con disponibilidad limitada de datos etiquetados. Así, los modelos de difusión pueden ofrecer múltiples beneficios que pueden transformar el campo de la imagen médica, promoviendo una mejor gestión de la salud y resultados más favorables para los pacientes.

Introducción

El rápido avance de la inteligencia artificial (IA) en los últimos años ha traído cambios transformadores en numerosos campos, destacándose la imagen médica como un dominio particularmente impactado[19, 20]. La integración de la IA en la imagen médica está revolucionando los procesos diagnósticos, mejorando la precisión en la detección de enfermedades y los resultados de los pacientes. Entre las diversas técnicas de IA aplicadas a este campo, como son la segmentación, clasificación o la generación de imágenes, los modelos de difusión generativos han captado una atención significativa en la generación de imagen médica debido a su capacidad para generar imágenes de alta calidad y extraer características significativas de dichas imágenes en otro tipo de trabajos[26]. No obstante, no es este su único uso, ya que los modelos de difusión se aplican en una amplia gama de tareas dentro de la imagen médica, desde la mejora de la calidad de imágenes existentes hasta la generación de nuevos datos sintéticos que pueden ser usados para entrenar otros algoritmos de IA, lo que incrementa aún más su relevancia en este campo.

La imagen médica es una herramienta indispensable en la atención sanitaria moderna, que abarca varias modalidades de imagen complementarias entre sí como radiografías, tomografía computarizada (TC), resonancia magnética (RM), tomografía por emisión de positrones (PET), ultrasonido y fotografía, entre otros. Estas técnicas de imagen son importantes para diagnosticar y monitorear una amplia gama de condiciones de salud, desde fracturas óseas y tumores hasta enfermedades cardiovasculares y trastornos neurológicos. Sin embargo, la interpretación de las imágenes médicas a menudo requiere un alto nivel de experiencia y puede ser tanto lenta como propensa a errores humanos. La IA, especialmente a través del uso de modelos de aprendizaje automático y profundo, ofrece una solución a estos desafíos al automatizar y mejorar el proceso de análisis de imágenes.

Los modelos de difusión, un tipo de modelos generativos, han surgido como un área prometedora dentro de la investigación en IA. Estos modelos se basan en la idea de revertir un proceso de difusión (donde se añade gradualmente ruido a los datos), en que el modelo aprende a revertir este proceso para generar nuevas muestras de datos. A diferencia de los modelos generativos tradicionales como las Redes Generativas Adversarias (GAN) y los Autoencoders Variacionales (VAE), entre otros, los modelos de difusión utilizan un enfoque que ha demostrado un rendimiento superior en la generación de imágenes realistas y de alta fidelidad[15]. Esto los hace particularmente valiosos en la imagen médica, donde la claridad y precisión de las imágenes son fundamentales.

Introducción

Además, los modelos de difusión ofrecen un método efectivo para la síntesis de imágenes de pacientes que pueden ser utilizadas para el desarrollo y la evaluación de algoritmos de análisis de imágenes, permitiendo así la optimización y validación de nuevas técnicas diagnósticas antes de su implementación clínica. Esta capacidad de generar datos sintéticos de alta calidad también es fundamental para el entrenamiento y la evaluación de algoritmos de IA destinados a la detección temprana de enfermedades o a la personalización de tratamientos médicos.

Aparte de su capacidad para generar imágenes de alta calidad y mejorar la interpretación de datos médicos, los modelos de difusión también se destacan por su versatilidad en aplicaciones adicionales dentro del ámbito de la imagen médica. Estos modelos pueden emplearse en la restauración de imágenes deterioradas o de baja resolución [33]. Este uso es crucial en situaciones donde se necesite mejorar la claridad de las imágenes adquiridas, como en estudios con equipos médicos menos avanzados o en casos donde las condiciones de imagen originales sean subóptimas debido a factores técnicos o biológicos.

En este contexto, la calidad de las imágenes generadas es de suma importancia, ya que estas son utilizadas para tomar decisiones críticas sobre la salud del paciente. Por lo tanto, es esencial contar con métricas de evaluación de la calidad de imagen adecuadas que puedan medir con precisión la fidelidad y la utilidad clínica de las imágenes generadas. Entre las métricas más comunes se encuentran el Índice de Similitud Estructural (SSIM), la distancia de Fréchet Inception (FID) o el *score* del *accuracy* en clasificación (CAS). Estas métricas permiten evaluar la calidad de las imágenes generadas y compararlas con imágenes reales para asegurar que cumplan con los estándares necesarios para su uso en entornos clínicos.

El objetivo principal de esta disertación es explorar la aplicación de los modelos de difusión en el ámbito de la generación de imagen médica, y estudiar el uso de las distintas métricas de evaluación de calidad de imagen generada en este ámbito tan particular. Entre los objetivos específicos de este trabajo se incluyen proporcionar una explicación de los modelos de difusión, incluidas sus bases teóricas y avances recientes; entrenar redes neuronales convolucionales en el contexto de los modelos de difusión para la síntesis de nuevas imágenes, así como para la evaluación del rendimiento de dichos modelos y la mejora de imágenes médicas en varias modalidades, usando como referencia las métricas de evaluación de calidad.

La revisión de la literatura cubrirá la evolución de los modelos generativos, con un enfoque en el desarrollo y los avances de los modelos de difusión. Explorará las bases teóricas de estos modelos, incluidas sus formulaciones matemáticas y los principios subyacentes a su operación. La revisión también examinará las aplicaciones existentes de los modelos de difusión en varios campos, destacando su potencial para la imagen médica. Se mencionarán comparaciones con otros modelos generativos como las GAN y los VAE para subrayar las fortalezas y debilidades únicas de los modelos de difusión.

La metodología del trabajo experimental detallará el diseño e implementación de

los varios modelos generativos adaptados a la tarea de imagen médica. Se entrenarán algoritmos para la generación de imágenes en blanco y negro de radiografías de manos y para la generación de imágenes a color de lesiones de la piel. Además, se evaluará de forma cualitativa y cuantitativa la calidad y coherencia de las imágenes generadas. En cuanto al conjunto de datos empleado, se emplean los conjuntos de imágenes ofrecidos por MedNIST y la competición anual ISIC, y se explica los pasos de preprocesamiento y las arquitecturas específicas de los modelos de difusión utilizados en el estudio. La metodología también describirá las métricas y técnicas de evaluación empleadas para evaluar el rendimiento de estos modelos. Se describirán detalladamente los montajes experimentales, incluidas las configuraciones de hardware y software, para garantizar la reproducibilidad de los resultados.

El impacto potencial de los modelos de difusión en la imagen médica es significativo[29], ofreciendo numerosos beneficios que pueden avanzar en el campo de varias maneras. Al generar imágenes de alta calidad e interpretabilidad, estos modelos pueden ayudar a realizar diagnósticos más precisos y eficientes, lo que lleva a una mejor gestión y resultados para los pacientes. Además, su capacidad para generar datos sintéticos de alta calidad puede facilitar el entrenamiento de algoritmos de IA sin la necesidad de grandes volúmenes de datos reales, lo que es especialmente útil en áreas donde la disponibilidad de datos etiquetados es limitada. Los modelos de difusión también pueden contribuir a la personalización de tratamientos médicos mediante la generación de imágenes específicas del paciente, permitiendo así un enfoque más preciso y adaptado a las necesidades individuales de cada paciente.

1 Modelos de difusión y la imagen médica

1.1. ¿Qué son los modelos de difusión?

Los modelos generativos son un conjunto de métodos de *Deep Learning*, muchos de ellos desarrollados principalmente en la última década, con una gran variedad de aplicaciones como imagen, sonido, grafos, etc. Dentro de este mundo, destacan tres clases de modelos generativos: las redes generativas adversarias (GANs), los autoencoders variacionales (VAEs) y los modelos de difusión (DMs), aparte de otras técnicas como los modelos basados en energía (EBMs) o los *normalizing flows*.

Una GAN [22] es un tipo de red neuronal compuesta por dos modelos principales: el generador y el discriminador, que se entrena de manera competitiva[11]. El generador, como su nombre indica, aprende a generar nuevas instancias de datos que parezcan auténticas. El discriminador, por su parte, aprende a distinguir entre datos reales y datos generados por el generador, mejorando su capacidad para detectar falsificaciones mediante el entrenamiento. Durante el proceso de entrenamiento, ambos modelos compiten entre sí: el generador busca engañar al discriminador creando datos realistas, mientras que el discriminador se vuelve más preciso en la detección de falsificaciones. Este proceso continuo de competencia mejora la capacidad del generador para producir datos indistinguibles de los datos reales según el discriminador.

Los VAEs [32], que también están basados en *Deep Learning*, aprenden una representación latente de los datos para generarlos de nuevo a partir de esta representación[30]. Funcionan con un codificador que transforma la entrada en una distribución probabilística (media y desviación estándar) en el espacio latente, y un decodificador que reconstruye la entrada original a partir de una muestra de esta distribución. La función de pérdida de un VAE combina la pérdida de reconstrucción y la pérdida de regularización, que mide la diferencia entre la distribución latente aprendida y una distribución a priori. Esto asegura que la estructura del espacio latente sea regular y que las muestras generadas sean coherentes. Una vez entrenado, el VAE puede generar nuevos datos tomando muestras aleatorias en el espacio latente y pasándolas a través del decodificador. Los VAEs son útiles en la generación de datos sintéticos y en la reducción de la dimensionalidad de los datos mientras preservan la estructura probabilística de los mismos.

Los modelos generativos deben cumplir tres requisitos generales para que sean ampliamente aceptados y usados en problemas del mundo real [52]. Estos requisitos son (i) generación de ejemplos de alta calidad, (ii) diversidad de ejemplos generados y (iii) generación de *samples* rápida y computacionalmente poco costosa. En la generación de

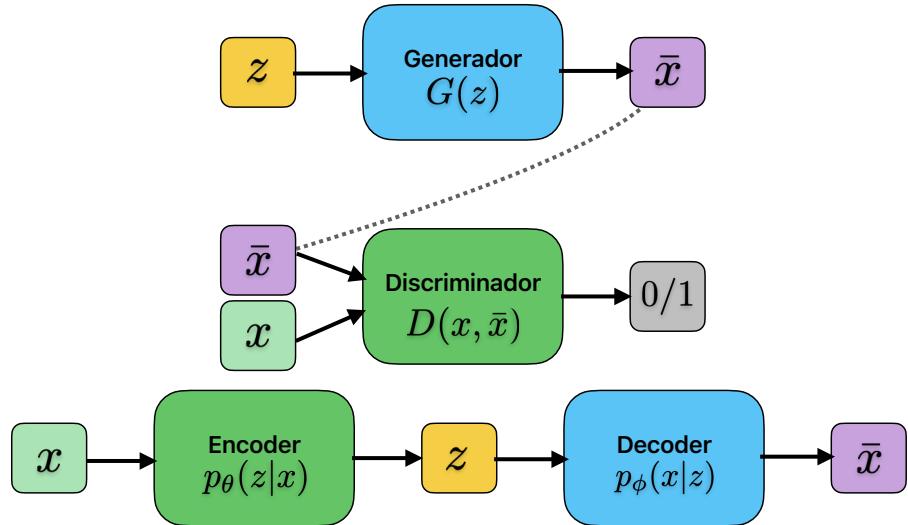


Figura 1.1: Esquema del flujo de trabajo de GAN (arriba) y de VAE (abajo).

imágenes se pone mucho interés en que las imágenes generadas tengan buena calidad; pero también es importante generar imágenes de clases minoritarias para que no haya un impacto negativo sobre ellas en la sociedad; o que el tiempo de *sampleado* no sea excesivo (esto último toma especial importancia en la generación de discurso a tiempo real). Este problema se conoce con el nombre del *trilema* de los modelos generativos.

Todas las clases de modelos generativos mencionados cumplen con dos de los tres requisitos del *trilema* [29]. En el caso de las GANs, falla la diversidad de ejemplos generados; los VAEs suelen generar registros de baja calidad; mientras que los modelos difusos son muy lentos en el proceso de generación. En la figura 1.2 se muestra un diagrama en el que se entiende mejor esta dicotomía.

Por todo ello, los modelos difusos se usan para evitar las principales desventajas de los otros dos modelos generativos, aunque debido a su naturaleza, requieren mucho coste de cómputo, por lo que son modelos caros que requieren una mejora [29].

Los modelos difusos son la última clase de modelos en incorporarse a los tipos de modelos generativos y que ha demostrado su efectividad aprendiendo distribuciones de datos complejas y su utilidad en varias aplicaciones, que van desde tareas de generación, como generación de imagen, superresolución o *inpainting*, hasta tareas de discriminación como la segmentación, clasificación y la detección de anomalías[29]. Su entrenamiento se compone de dos fases: un proceso “adelante” (*forward process*) y un proceso inverso (*reverse process*). El primero básicamente añade ruido a los datos paulatinamente hasta que se transforman en puro ruido gaussiano. El segundo, también conocido como *denoising process*, se aplica para recuperar la forma original de los datos. De este modo, se entrena un modelo generativo que puede generar distribuciones complejas de datos a partir de ruido aleatorio de forma precisa.

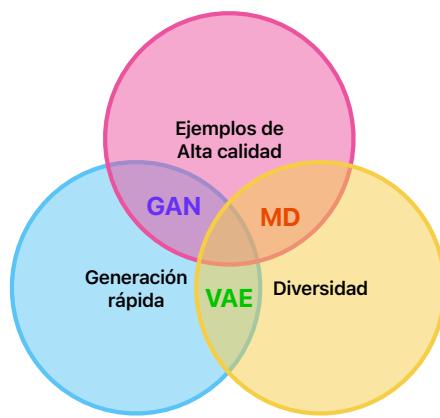


Figura 1.2: El trilema de los modelos generativos. En la región donde los tres círculos intersecan no aparece ningún modelo generativo.

Las investigaciones actuales sobre modelos difusos están basadas en dos formulaciones predominantes de estos modelos: *Denoising probabilistic diffusion models* (Modelos difusos probabilísticos de eliminación de ruido o DDPMs), *Score-based generative models* (Modelos generativos basados en un *score* o SGMs). Los primeros pueden clasificarse como aquellos que usan la inferencia variacional para entrenar el modelo; mientras que los otros, se consideran constituyentes de la perspectiva de *score*, ya que se basan en el cálculo del gradiente del logaritmo de la distribución de los datos (lo que comúnmente se denomina función de *score*) para su entrenamiento.

No obstante, existen muchas versiones y técnicas basadas en los enfoques mencionados y que los alteran o combinan con otras técnicas. Ese es el caso de los modelo de difusión latentes[42], los cuales combinan los espacios latentes de autoencoders preentrenados con los modelos de difusión en dichos espacios. De este modo, se puede conseguir un equilibrio entre la reducción de complejidad y la preservación de detalles, mejorando el coste computacional de los modelos de difusión tradicionales con una buena calidad y fidelidad de imagen.

1.2. Uso de los modelos de difusión en imagen médica

La imagen médica es un campo muy importante en el área de salud que utiliza técnicas especializadas para crear imágenes del cuerpo humano que se usan en el diagnóstico clínico y en investigación. Engloba varias modalidades de imagen como rayos X, tomografía computacional (CT), resonancia magnética (MRI) y ultrasonidos, entre otras.. El procesamiento y análisis de imágenes médica ha tomado una relevancia significativa en los últimos años, con técnicas de segmentación, resgitrado, fusión y

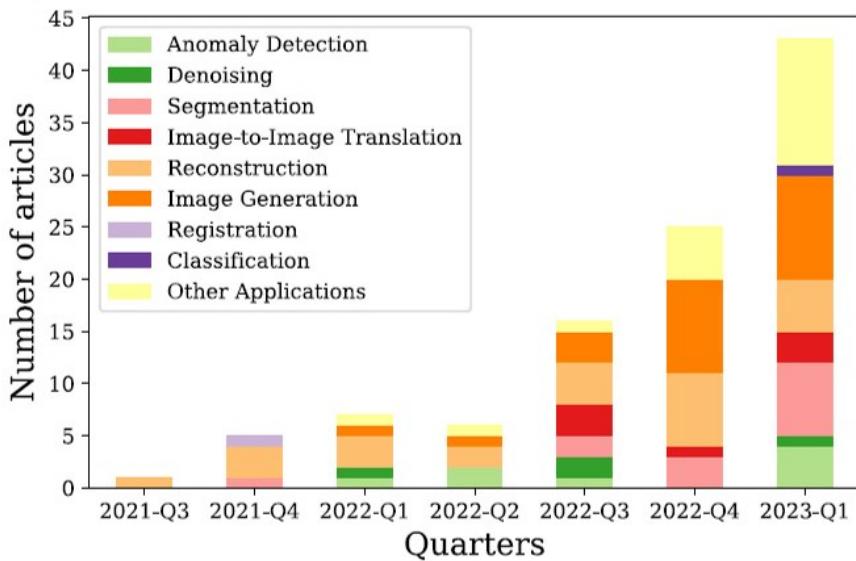


Figura 1.3: Número de artículos publicados (*number of articles*) en el ámbito de la imagen médica basados en modelos de difusión en función del año y del cuartil (*quarters*). Se observa una tendencia al alza en el número de trabajos sobre modelos de difusión en dicho ámbito. Imagen extraída de [29].

eliminación de ruido. Técnicas matemáticas y computacionales avanzadas, como el *deep learning*, se están empezando a usar[3] en tareas como la reconstrucción, la generación o la clasificación de imágenes, obteniendo buenos resultados.

Los modelos generativos están empezando a tener un gran impacto en la imagen médica, como puede observarse en la Figura 1.3, debido a que estos modelos solucionan gran parte de los problemas existentes. Es decir, la complejidad de los procedimientos de recolección de datos, la falta de información de manos de expertos en la materia y los problemas de privacidad de los pacientes son los principales escollos para el desarrollo de este campo. Es en este escenario donde los modelos generativos son verdaderamente útiles; más concretamente los modelos de difusión.

En medicina, muchos conjuntos de datos están fuertemente desbalanceados, debido a que ciertas enfermedades tienen un bajo impacto en la población, por lo que se tienen pocos registros de estas. Ello dificulta gravemente el exitoso entrenamiento de algoritmos de *Machine Learning* e IA, como clasificadores o herramientas de detección de anomalías, que se utilizarían como apoyo a profesionales médicos e investigadores en la detección y estudio de enfermedades. La propiedad de diversidad de los MD hace que estos puedan generar varias imágenes aparentemente realistas y diversas para balancear esos conjuntos. Además, esta característica repercute sobre la educación, ya que la generación de muchas imágenes diferentes de distintas modalidades favorece

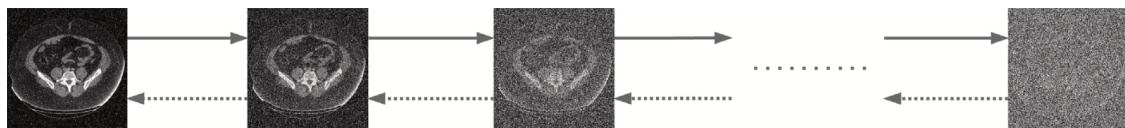


Figura 1.4: Proceso de generación de una imagen de un corte transversal de tórax usando modelos de difusión. A partir de una imagen de ruido aleatorio (derecha), se neutraliza dicho ruido, paso a paso, hasta obtener una imagen realista que siga la distribución de datos reales (izquierda).

en la formación y aprendizaje de futuros profesionales médicos.

Existen varios trabajos que tratan la generación de imagen médica sintética 2D/3D usando modelos de difusión. Por ejemplo, Kim et al.[31] propusieron el modelo de difusión deformable (DDM), basado en un DDPM, el cual genera una secuencia temporal de imágenes 3D, probándolo ellos con imágenes PET del corazón. Packhäuser et al.[40] utilizan un modelo de difusión latente (aplica DDPM sobre un espacio latente entrenado con VAEs) para producir imágenes de alta calidad de radiografías de pecho sin usar información privada del sujeto en el proceso. Incluso, los modelos de difusión se han usado en la generación de imágenes histopatológicas, capaces de mostrar propiedades morfológicas y genéticas[37].

Otra ventaja que permite el uso de modelos generativos en la generación de imagen médica, es que estas imágenes no tienen ningún problema en cuanto a la privacidad del usuario, ya que el único uso que se hace de imágenes de pacientes verdaderos es para entrenar los modelos generativos. Las imágenes generadas con esos modelos no corresponden a ningún paciente, por lo que su uso público no necesitaría ningún permiso ni violaría ningún derecho de privacidad, salvando el escollo de la falta de imágenes por cuestiones de privacidad, y los trámites burocráticos asociados a las leyes de protección de datos para el uso de imágenes de pacientes.

Generalmente, etiquetar las distintas imágenes médicas es un proceso lento y costoso que requiere de un experto. De esta manera, los modelos difusos alivian en gran parte el problema, ya que pueden entrenarse para su uso en tareas de clasificación y segmentación[2]. Esto repercute en la agilización del tiempo en la detección de enfermedades y en una mejora de la rapidez y eficacia del diagnóstico de graves problemas de salud. Yang et al. proponen *DiffMIC* como el “primer método basado en difusión en abordar la clasificación de imágenes médicas generales mediante la eliminación de ruidos y perturbaciones inesperadas y la captura sólida de la representación semántica”[56].

A pesar de que el uso de técnicas de generación de imágenes no ha sido aún gran objeto de estudio por parte de la comunidad de imagen médica, varios estudios han demostrado su gran utilidad en situaciones reales. Por ejemplo, experimentos llevados a cabo por [1] han demostrado que el uso de imágenes sintéticas generadas por

1 Modelos de difusión y la imagen médica

modelos difusos mejora la precisión para clasificadores de imágenes de melanoma de piel, y que los modelos que han sido entrenados con imágenes reales y sintéticas tienen un mejor desempeño que aquellos entrenados con una sola fuente, todo ello sin comprometer el problema de la privacidad del paciente.

También se han llevado a cabo trabajos en otras áreas de la imagen médica como modelos de difusión, como la reconstrucción, la segmentación y la detección de anomalías [29].

Es importante destacar de nuevo el estudio [37], en el que se hizo distinguir entre imágenes reales y sintéticas generadas por modelos difusos a dos patólogos con distintivo nivel de experiencia. Los resultados advirtieron que no eran capaces ninguno de advertir cuáles eran las imágenes reales de las creadas, recalando la gran calidad de las imágenes generadas usando este tipo de modelos generativos.

Respecto a los otros modelos generativos y su empleo en imagen médica, las GANs se utilizan en una variedad de aplicaciones, desde la mejora de resolución de imágenes hasta la síntesis de datos e imágenes médicas para *Data Augmentation*, ya que tienen muy buenas capacidades para la generación de imagen [11]. No obstante, los VAEs no se han utilizado tan comúnmente para *Data Augmentation* en comparación con los GANs debido a la naturaleza borrosa y nebulosa de las muestras generadas[30]. Existen enfoques, como VAE-GAN o VAE-condicionales, viables para la generación de imágenes médicas, cada uno ofreciendo diferentes beneficios y compensaciones en términos de diversidad, calidad y fidelidad de las muestras generadas.

Por tanto, con todo esto, puede decirse que los modelos difusos están probando ser una herramienta muy valorada en entornos clínicos que cubre una gran cantidad de problemas en imagen médica. Se espera que uso aumente en un futuro, proporcionando nuevas oportunidades en el campo de la imagen médica y en investigación[29].

2 Fundamento teórico de los modelos de difusión

En este capítulo se explica el fundamento teórico-matemático de los tres tipos base de modelos difusos, divididos en dos categorías según el método de calcular la distribución objetivo: perspectiva variacional o de *score*. Se describen cómo se plantean y formalizan los procesos *forward* y *reverse* y se obtiene la expresión para la función objetivo a minimizar en el entrenamiento de estos modelos. Para ello, nos basamos principalmente en los estudios y publicaciones sobre el tema de Yang L. et al. [55], Cao H. et al. [10], y Weng L. [51].

De entre los tipos mencionados, hacemos mayor hincapié en el detalle de explicación de los modelos de la categoría variacional porque son el tipo de modelo más empleado en literatura para entrenar modelos de difusión y va a ser el enfoque que usemos para generar imágenes en este trabajo ([Sección 5.2](#)).

2.1. Perspectiva variacional

La perspectiva variacional incluye modelos que usan la inferencia variacional para aproximar la distribución objetivo, generalmente minimizando la divergencia de Kullback-Leibler entre las distribuciones objetivo y aproximada, estando esta última en una familia de densidades sobre las variables latentes, parametrizadas por parámetros variacionales libres. Los modelos difusos probabilísticos de eliminación de ruido (*DDPMs* por sus siglas en inglés) ([46], [26]) son un ejemplo de este tipo de modelos, los cuales explicamos a continuación.

2.1.1. Denoising Probabilistic Diffusion Models (*DDPMs*)

2.1.1.1. Proceso de inyección de ruido

Dado un ejemplo x_0 obtenido a partir de una distribución original de los datos $q(x_0)$ ($x_0 \sim q(x_0)$), se define el proceso *forward* o “hacia delante” como una cadena de Markov en la que se añade una pequeña cantidad de ruido gaussiano al ejemplo durante T pasos sucesivos, para así obtener un conjunto de ejemplos ruidosos que denominamos por x_1, \dots, x_T .

Por cada iteración t , la distribución de los datos se transforma en una distribución nueva $q(x_t)$, analíticamente manejable, al aplicar un kernel difuso de Markov $T_q(x_t|x_{t-1}; \beta_t)$ de forma repetida sobre la distribución original. El parámetro β_t es la tasa de difusión del proceso. Es decir,

$$q(x_{0,\dots,T}) = q(x_0) \prod_{t=1}^T T_q(x_t|x_{t-1};\beta_t) = q(x_0) \prod_{t=1}^T q(x_t|x_{t-1}), \quad (2.1)$$

donde tomamos

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right), \forall t. \quad (2.2)$$

Definiendo $\alpha_t = 1 - \beta_t$ y $\bar{\alpha}_t = \prod_{s=0}^t \alpha_s$, y a partir de la Eq. (2.2), podemos escribir que:

$$\begin{aligned} x_t &= \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon_t \\ &= \sqrt{1-\beta_t}\sqrt{1-\beta_{t-1}}x_{t-2} + \sqrt{1-\beta_t}\sqrt{\beta_{t-1}}\epsilon_{t-1} + \sqrt{\beta_t}\epsilon_t \\ &\stackrel{*}{=} \sqrt{1-\beta_t}\sqrt{1-\beta_{t-1}}x_{t-2} + \sqrt{\beta_t + \beta_{t-1}(1-\beta_t)}\tilde{\epsilon}_{t-1} \\ &= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\tilde{\epsilon}_{t-1}, \end{aligned} \quad (2.3)$$

donde $\tilde{\epsilon}_{t-1} \sim \mathcal{N}(0, I)$, ya que se ha usado en * la propiedad $\mathcal{N}(0, \sigma_1^2 I) + \mathcal{N}(0, \sigma_2^2 I) = \mathcal{N}(0, (\sigma_1^2 + \sigma_2^2)I)$. Por tanto, puede afirmarse que

$$q(x_t|x_{t-2}) = \mathcal{N}\left(x_t | \sqrt{\alpha_t\alpha_{t-1}}x_{t-2}, (1 - \alpha_t\alpha_{t-1})I\right). \quad (2.4)$$

De forma recursiva, se tiene entonces:

$$q(x_t|x_0) = \mathcal{N}\left(x_t | \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)I\right), \quad (2.5)$$

y

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(0, I). \quad (2.6)$$

Dado que el ejemplo se vuelve cada vez más ruidoso, se suele considerar que $\beta_1 < \beta_2 < \dots < \beta_T$. De hecho, cuando $\bar{\alpha}_T \approx 0$, según Eq. (2.6), se obtiene que $q(x_T) = \int q(x_T|x_0)q(x_0)dx_0 \approx \mathcal{N}(0, I)$.

2.1.1.2. Proceso de eliminación de ruido

A partir de un *input* de ruido gaussiano $x_T \sim \mathcal{N}(0, I)$ y si somos capaces de calcular las distribuciones $q(x_{t-1}|x_t)$, entonces somos capaces de recrear ejemplos reales. No obstante, dichas distribuciones no son fáciles de estimar, ya que requieren de todo el conjunto de datos. Por ello, aprendemos un modelo p_θ que aproxima dichas distribuciones de probabilidad condicionada. Por tanto, podemos describir el proceso inverso de forma que

$$p_\theta(x_{0,\dots,T}) = p(x_T) \prod_{t=1}^T p_\theta(x_{t-1}|x_t), \quad (2.7)$$

donde sabemos que $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$ si β_t es lo suficientemente pequeño, con θ denotando los parámetros del modelo. Se ha elegido la distribución $p(x_T) = \mathcal{N}(0, I)$ porque el proceso hacia delante se ha construido conforme a que $q(x_T) \approx \mathcal{N}(0, I)$. La media $\mu_\theta(x_t, t)$ y la varianza $\Sigma_\theta(x_t, t)$ son parametrizadas mediante redes neuronales.

La siguiente proposición es muy útil para comprobar cómo se relacionan los parámetros de las distribuciones del proceso *forward* y del *reverse*.

Proposición 2.1. *La distribución de probabilidad inversa condicionada al ejemplo inicial, $q(x_{t-1}|x_t, x_0)$, se puede conocer y es:*

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I) \quad (2.8)$$

Demostración. Usando la regla de Bayes, se tiene que:

$$\begin{aligned} q(x_{t-1}|x_t, x_0) &= q(x_t|x_{t-1}, x_0) \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} \\ &\stackrel{\text{(Ecuación (2.2),(2.5))}}{\propto} \exp \left[-\frac{1}{2} \left(\frac{(x_t - \sqrt{\alpha_t}x_{t-1})^2}{\beta_t} + \frac{(x_{t-1} - \sqrt{\bar{\alpha}_{t-1}}x_0)^2}{1 - \bar{\alpha}_{t-1}} - \right. \right. \\ &\quad \left. \left. \frac{(x_t - \sqrt{\bar{\alpha}_t}x_0)^2}{1 - \bar{\alpha}_t} \right) \right] \\ &= \exp \left[-\frac{1}{2} \left(\left(\frac{\alpha_t}{\beta_t} + \frac{1}{1 - \bar{\alpha}_{t-1}} \right) x_{t-1}^2 - \left(\frac{2\sqrt{\alpha_t}}{\beta_t} x_t + \right. \right. \right. \\ &\quad \left. \left. \left. \frac{2\sqrt{\bar{\alpha}_{t-1}}}{1 - \bar{\alpha}_{t-1}} x_0 \right) x_{t-1} + C(x_t, x_0) \right) \right]. \end{aligned} \quad (2.9)$$

La función $C(x_t, x_0)$ no contiene términos que involucren a x_{t-1} . Relacionando esta expresión con la de la función de densidad de una normal estándar, se tiene que:

$$\tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t} x_0 \quad (2.10)$$

Usando la expresión de x_0 que se obtiene de despejar en Eq. (2.6), se obtiene:

$$\tilde{\mu}_t = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_t \right), \quad (2.11)$$

$$\tilde{\beta}_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t. \quad (2.12)$$

□

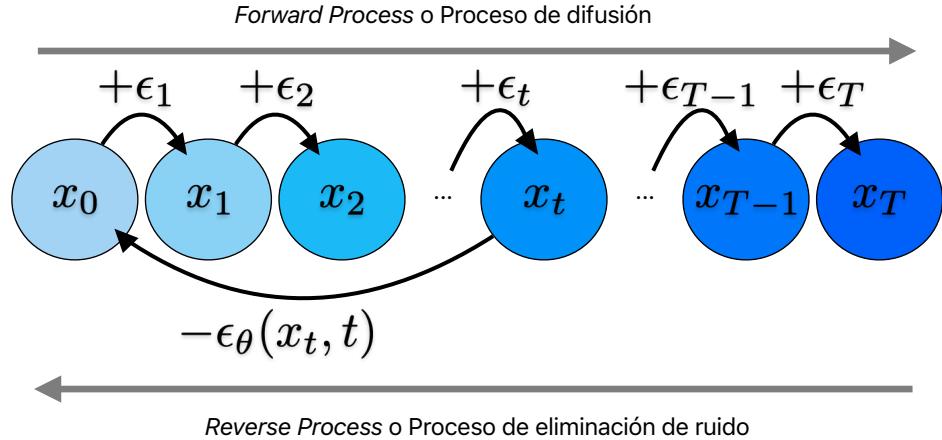


Figura 2.1: Esquema del proceso de difusión de la imagen y de eliminación de ruido para un modelo DDPM.

2.1.1.3. Función de pérdida

El objetivo del entrenamiento del modelo es obtener aquel valor del hiperparámetro θ de forma que las distribuciones $p_\theta(x_{1,\dots,T}|x_0)$ sean lo más parecidas o cercanas posible a las distribuciones reales $q(x_{1,\dots,T}|x_0)$. Esta distancia puede cuantificarse mediante el uso de la divergencia de Kullback-Leibler (KL), $\mathcal{D}_{KL}(q(x_{1,\dots,T}|x_0)||p_\theta(x_{1,\dots,T}|x_0))$, la cual mide cuánta información se pierde si la distribución $p_\theta(x_{1,\dots,T}|x_0)$ se usara para representar $q(x_{1,\dots,T}|x_0)$. Para este caso, lo que se desea es minimizar la divergencia respecto de θ .

Para dos distribuciones de probabilidad P_1 y P_2 de una variable aleatoria continua, su divergencia de Kullback-Leibler se define de la forma:

$$\mathcal{D}_{KL}(P_1||P_2) = \int_x p_1(x) \log \left(\frac{p_1(x)}{p_2(x)} \right) dx = \mathbb{E}_{z \sim p_1(x)} \left[\log \frac{p_1(x)}{p_2(x)} \right] \quad (2.13)$$

La divergencia KL no es una función de distancia simétrica, menos en el caso de que las distribuciones del argumento sean idénticas (salvo un conjunto numerable de puntos) [38]. De la Eq. (2.13), se observa que la función $\log \frac{p_1(x)}{p_2(x)}$ contribuye a la divergencia cuando $p_1(x) > 0$. En ese caso, $\lim_{p_2(x) \rightarrow 0} \log \frac{p_1(x)}{p_2(x)} = \infty$; lo que significa que la divergencia será alta siempre y cuando $p_2(x)$ no sea capaz de “cubrir” a $p_1(x)$. Por tanto, podemos asegurar que el valor de la divergencia se minimiza cuando $p_2(x) > 0$ siempre que $p_1(x) > 0$. A este valor se lo conoce como *forward-KL*.

Sin embargo, minimizar el valor de $\mathcal{D}_{KL}(P_2||P_1)$, que lo denominamos por *reverse-KL*,

tiene un efecto contrario, ya que si $p_1(x) = 0$, debemos conseguir que $p_2(x) = 0$ para que la divergencia no crezca. En resumen, minimizar el *forward-KL* "estira" la función de distribución P_2 para que cubra toda P_1 como una lona; mientras que minimizar *reverse-KL* estruja la función P_2 debajo de P_1 .

Desarrollamos la divergencia *forward-KL*:

$$\begin{aligned}
 \mathcal{D}_{KL}[q(x_{1,\dots,T}|x_0)||p_\theta(x_{1,\dots,T}|x_0)] &= \int q(x_{1,\dots,T}|x_0) \log \frac{q(x_{1,\dots,T}|x_0)}{p_\theta(x_{1,\dots,T}|x_0)} dx_1 \dots dx_T \\
 &= \int q(x_{1,\dots,T}|x_0) \left[\log p_\theta(x_0) + \log \frac{q(x_{1,\dots,T}|x_0)}{p_\theta(x_{0,\dots,T})} \right] dx_1 \dots dx_T \\
 &= \log p_\theta(x_0) + \int q(x_{1,\dots,T}|x_0) \log \frac{q(x_{1,\dots,T}|x_0)}{p_\theta(x_0|x_{1,\dots,T})p_\theta(x_{1,\dots,T})} dx_1 \dots dx_T \\
 &= \log p_\theta(x_0) + \mathbb{E}_{x_{1,\dots,T} \sim q(x_{1,\dots,T}|x_0)} \left[\log \frac{q(x_{1,\dots,T}|x_0)}{p_\theta(x_{1,\dots,T})} - \log p_\theta(x_0|x_{1,\dots,T}) \right] \\
 &= \log p_\theta(x_0) + \mathcal{D}_{KL}[q(x_{1,\dots,T}|x_0)||p_\theta(x_{1,\dots,T})] - \mathbb{E}_{x_{1,\dots,T} \sim q(x_{1,\dots,T}|x_0)} [\log p_\theta(x_0|x_{1,\dots,T})]
 \end{aligned} \tag{2.14}$$

Reordenando términos a izquierda y derecha de la igualdad, se consigue:

$$\begin{aligned}
 \log p_\theta(x_0) - \mathcal{D}_{KL}[q(x_{1,\dots,T}|x_0)||p_\theta(x_{1,\dots,T}|x_0)] \\
 = \mathbb{E}_{x_{1,\dots,T} \sim q(x_{1,\dots,T}|x_0)} [\log p_\theta(x_0|x_{1,\dots,T})] - \mathcal{D}_{KL}[q(x_{1,\dots,T}|x_0)||p_\theta(x_{1,\dots,T})]
 \end{aligned} \tag{2.15}$$

El lado de la izquierda en la Eq. (2.15) es exactamente lo que se quiere maximizar cuando se aprende las verdaderas distribuciones: maximizar el logaritmo de la verosimilitud cuando se generan muestras a partir de ruido aleatorio, y se quiere minimizar la diferencia entre las distribuciones reales y estimadas. A este término se le denomina cota inferior variacional (VLB), y si lo desarrollamos,

$$\begin{aligned}
 -\log p_\theta(x_0) &\leq -\log p_\theta(x_0) + \mathcal{D}_{KL}[q(x_{1,\dots,T}|x_0)||p_\theta(x_{1,\dots,T}|x_0)] \\
 &= -\log p_\theta(x_0) + \mathbb{E}_{x_{1,\dots,T} \sim q(x_{1,\dots,T}|x_0)} \left[\log \frac{q(x_{1,\dots,T}|x_0)}{p_\theta(x_{0,\dots,T})/p_\theta(x_0)} \right] \\
 &= -\log p_\theta(x_0) + \mathbb{E}_{x_{1,\dots,T} \sim q(x_{1,\dots,T}|x_0)} \left[\log \frac{q(x_{1,\dots,T}|x_0)}{p_\theta(x_{0,\dots,T})} + \log p_\theta(x_0) \right] \\
 &= \mathbb{E}_{x_{1,\dots,T} \sim q(x_{1,\dots,T}|x_0)} \left[\log \frac{q(x_{1,\dots,T}|x_0)}{p_\theta(x_{0,\dots,T})} \right]
 \end{aligned} \tag{2.16}$$

2 Fundamento teórico de los modelos de difusión

Por tanto definimos la función objetivo en el entrenamiento del modelo como:

$$\mathcal{L}_{\text{VLB}} = \mathbb{E}_{x_{0,\dots,T} \sim q(x_{0,\dots,T})} \left[\log \frac{q(x_{1,\dots,T}|x_0)}{p_\theta(x_{0,\dots,T})} \right] \geq -\mathbb{E}_{x_0 \sim q(x_0)} [\log p_\theta(x_0)] \quad (2.17)$$

Para poder computarla analíticamente, la función objetivo puede reescribirse como una combinación de varios términos de divergencia KL y de entropía [46].

$$\begin{aligned} \mathcal{L}_{\text{VLB}} &= \mathbb{E}_{x_{0,\dots,T} \sim q(x_{0,\dots,T})} [\mathcal{D}_{\text{KL}}(q(x_T|x_0) || p_\theta(x_T))] \\ &\quad + \sum_{t=2}^T \mathcal{D}_{\text{KL}}(q(x_{t-1}|x_t, x_0) || p_\theta(x_{t-1}|x_t)) - \log p_\theta(x_0|x_1) \end{aligned} \quad (2.18)$$

Nombramos los siguientes términos:

$$\mathcal{L}_T = \mathcal{D}_{\text{KL}}(q(x_T|x_0) || p_\theta(x_T)) \quad (2.19)$$

$$\mathcal{L}_{t-1} = \mathcal{D}_{\text{KL}}(q(x_{t-1}|x_t, x_0) || p_\theta(x_{t-1}|x_t)) \quad (2.20)$$

$$\mathcal{L}_0 = -\log p_\theta(x_0|x_1) \quad (2.21)$$

Puede verse que \mathcal{L}_T es constante y puede ignorarse durante el entrenamiento (ya que q no tiene parámetros para aprender, dado que los coeficientes β_t se fijan previamente, y x_T es ruido gaussiano). Cada uno de los términos \mathcal{L}_{t-1} se puede calcular en la forma cerrada (ver 2.1.1.3).

En [26], se calcula \mathcal{L}_0 usando un *decoder* discreto independiente derivado a partir de la distribución $\mathcal{N}(x_0; \mu_\theta(x_1, 1), \beta_1 I)$. Asumiendo que los valores de la imagen se han reescalado linealmente al intervalo $[-1, 1]$, se tiene que:

$$p_\theta(x_0|x_1) = \prod_{i=1}^D \int_{\delta_-(x_0^i)}^{\delta_+(x_0^i)} \mathcal{N}(x; \mu_\theta(x_1, 1)^i, \beta_1 I) dx \quad (2.22)$$

donde D es la dimensionalidad de los datos y el superíndice i indica la extracción de la coordenada i -ésima. Además,

$$\begin{aligned} \delta_+(x) &= \begin{cases} \infty & \text{si } x = 1 \\ x + 1/255 & \text{si } x < 1 \end{cases} \\ \delta_-(x) &= \begin{cases} -\infty & \text{si } x = -1 \\ x - 1/255 & \text{si } x > -1 \end{cases} \end{aligned}$$

No obstante, los términos verdaderamente importantes para la función de pérdida en el entrenamiento del modelo son los \mathcal{L}_t .

Parametrización de \mathcal{L}_{t-1} para la función de pérdida en entrenamiento. Considérese el siguiente resultado previo.

Observación 2.1. Sean dos distribuciones normales k -dimensionales $\mathcal{N}_0(\mu_0, \Sigma_0)$ y $\mathcal{N}_1(\mu_1, \Sigma_1)$. Se tiene que la divergencia de Kullback-Leibler para estas dos distribuciones es[17]:

$$\mathcal{D}_{\text{KL}}(\mathcal{N}_0 || \mathcal{N}_1) = \frac{1}{2} \left[\text{traza}(\Sigma_1^{-1} \Sigma_0) - k + (\mu_1 - \mu_0)^T \Sigma_1^{-1} (\mu_1 - \mu_0) + \log \left(\frac{\det \Sigma_1}{\det \Sigma_0} \right) \right] \quad (2.23)$$

Usando la descomposición de Cholesky $\Sigma_0 = L_0 L_0^T$ y $\Sigma_1 = L_1 L_1^T$, y definiendo $L_1 M = L_0$ y $L_1 y = \mu_1 - \mu_0$, se tiene que la divergencia KL anterior se puede escribir como:

$$\mathcal{D}_{\text{KL}}(\mathcal{N}_0 || \mathcal{N}_1) = \frac{1}{2} \left[\sum_{i,j=1}^k M_{ij}^2 - k + |y|^2 + 2 \sum_{i=1}^k \log \frac{(L_1)_{ii}}{(L_0)_{ii}} \right] \quad (2.24)$$

Concretamente,

$$\mathcal{D}_{\text{KL}}(\mathcal{N}((\mu_1, \dots, \mu_k)^T, \text{diag}(\sigma_1^2, \dots, \sigma_k^2)) || \mathcal{N}(0, I)) = \frac{1}{2} \sum_{i=1}^k (\sigma_i^2 + \mu_i^2 - 1 - \log \sigma_i^2) \quad (2.25)$$

Tras este breve paréntesis, recordemos que se va a entrenar una red neuronal para aproximar la distribución de probabilidad condicionada en el proceso inverso, de forma que buscamos aproximar $\mu_\theta(x_t, t)$ a $\tilde{\mu}_t = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_t \right)$. Dado que x_t está disponible como *input* durante el entrenamiento, se puede reparametrizar el término de ruido gaussiano ϵ_t para predecirlo a partir de x_t :

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right). \quad (2.26)$$

Entonces, se tendría que:

$$x_{t-1} \sim \mathcal{N} \left(x_{t-1}; \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right), \Sigma_\theta(x_t, t) \right). \quad (2.27)$$

En [26], Ho et al. proponen que $\Sigma_\theta(x_t, t) = \sigma_t^2 I$, donde $\sigma_t = \beta_t$ ó $= \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}} \beta_t$. Experimentalmente, comprobaron que se obtienen resultados similares con ambos valores.

2 Fundamento teórico de los modelos de difusión

Usando la expresión de la divergencia de Kullback-Leibler para distribuciones normales multivariantes, se obtiene que el término de pérdida \mathcal{L}_t está parametrizado para minimizar la diferencia con $\tilde{\mu}$:

$$\begin{aligned}\mathcal{L}_t &= \mathbb{E}_{x_0 \sim q(x_0), \epsilon_t \sim N(0, I)} \left[\frac{1}{2\|\Sigma_\theta(x_t, t)\|_2^2} \|\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)\|^2 \right] \\ &= \mathbb{E}_{x_0, \epsilon_t} \left[\frac{1}{2\|\Sigma_\theta(x_t, t)\|_2^2} \left\| \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \right) \right\|^2 \right] \\ &= \mathbb{E}_{x_0, \epsilon_t} \left[\frac{(1 - \alpha_t)^2}{2\alpha_t(1 - \bar{\alpha}_t)\|\Sigma_\theta\|_2^2} \|\epsilon_t - \epsilon_\theta(x_t, t)\|^2 \right] \\ &= \mathbb{E}_{x_0, \epsilon_t} \left[\frac{(1 - \alpha_t)^2}{2\alpha_t(1 - \bar{\alpha}_t)\|\Sigma_\theta\|_2^2} \|\epsilon_t - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon_t, t)\|^2 \right]\end{aligned}\tag{2.28}$$

Por tanto, el término \mathcal{L}_t queda de la forma:

$$\mathcal{L}_t = \mathbb{E}_{x_0, \epsilon_t} \left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1 - \bar{\alpha}_t)} \|\epsilon_t - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon_t, t)\|^2 \right] + \text{cte}\tag{2.29}$$

Simplificación. La función de pérdida, compuesta por los términos dados por las Eq. (2.29) y (2.22), es diferenciable respecto del conjunto de parámetros θ , por lo que puede usarse en el entrenamiento. No obstante, en [26] se propone simplificar esta función con el objetivo de mejorar el entrenamiento y generación de imágenes, además de hacer más sencilla la implementación:

$$\mathcal{L}_t^{\text{simple}}(\theta) = \mathbb{E}_{x_0 \sim q(x_0), \epsilon_t \sim N(0, I)} \left[\|\epsilon_t - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon_t, t)\|^2 \right]\tag{2.30}$$

El caso para $t = 1$ corresponde con \mathcal{L}_0 en donde se calcula la integral Ecuación 2.22 usando el método del trapecio. Para $t > 1$, se tiene una versión no ponderada de la expresión de \mathcal{L}_t Ecuación 2.29.

Para parametrizar los coeficientes β_t , podemos tomar una secuencia de constantes linealmente decreciente desde $\beta_1 = 10^{-4}$ a $\beta_T = 0.02$. Una variación que introduce mejoras fue la propuesta por Nichol & Dhariwal [39], la cual consiste en tomar $\beta_t = \text{clip}\left(1 - \frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}, 0.999\right)$, con $\bar{\alpha}_t = f(t)/f(0)$ y $f(t) = \cos\left(\frac{\pi}{2} \frac{t/(T+s)}{1+s}\right)^2$, siendo s un pequeño offset.

En el modelo desarrollado por Ho et al.[26], el algoritmo de generación de imágenes hace uso de la dinámica de Langevin donde ϵ_θ se considera una estimación del gradiente de la distribución de los datos de entrenamiento, lo cual se explica en mayor detalle en la Subsección 2.2.1.

Por último, mostramos el pseudocódigo para los algoritmos de entrenamiento y sampleado basados en el procedimiento teórico expuesto, y que se han extraído de [26]. En la Figura 2.2 se muestra el esquema de una red U-Net que representa la función ϵ_θ a calcular como estimación del ruido introducido a la imagen para el entrenamiento del modelo DDPM.

Algorithm 1 Entrenamiento del modelo DDPM

```

1: repeat
2:    $x_0 \sim q(x_0)$ 
3:    $t \sim \mathcal{U}[1, \dots, T]$ 
4:    $\epsilon \sim \mathcal{N}(0, I)$ 
5:   Tomar gradiente descendente según  $\nabla_\theta ||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon_t, t)||^2$ 
6: until convergencia

```

Algorithm 2 *Ancestral sampling*. Sampleado de imágenes con dinámica de Langevin (Párrafo 2.2.1).

```

1:  $x_T \sim \mathcal{N}(0, I)$ 
2: for  $\text{dot} = T, \dots, 1$ 
3:    $z \sim \mathcal{N}(0, I)$  si  $t > 1$ , en cambio  $z = 0$ 
4:    $x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( x_t - \frac{1 - \alpha_t}{\sqrt{1 - \alpha_t}} \epsilon_\theta(x_t, t) \right) + \sqrt{\beta_t} z$ 
5: end for
6: return  $x_0$ 

```

2.2. Perspectiva de score

Los modelos de difusión con una perspectiva de *score* se fundamentan en el concepto de función de *score*, la cual, para una determinada distribución de probabilidad $p(x)$, se define como el gradiente en la variable independiente del logaritmo de dicha distribución $\nabla_x \log p(x)$. Esta función se emplea para estimar los parámetros del proceso de difusión. Geométricamente, el *score* puede verse como un campo de vectores en el que en cada punto hay un vector que apunta en la dirección y el sentido en los que la función de densidad de probabilidad tiene la mayor tasa de crecimiento.

Existen dos modelos que caen dentro de esta fundamentación: *Noise-Conditioned Score Networks* (NCSNs) y *Stochastic Differential Equations* (Score SDEs). La diferencia entre ambos reside en la forma de producir nuevas imágenes, ya que los primeros usan dinámica de Langevin; mientras que los segundos, ecuaciones diferenciales estocásticas. Incluso hay métodos que los combinan a ambos [[48]]. De hecho, los Score SDEs se consideran como una generalización de los NCSNs y de los DDPMs.

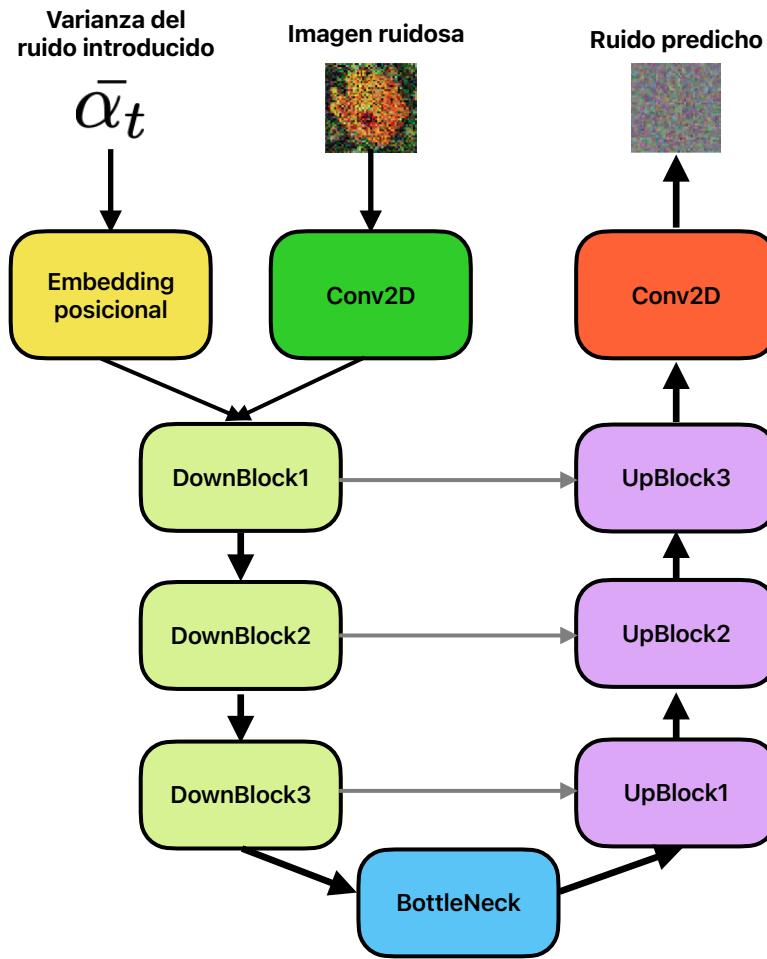


Figura 2.2: Esquema de una red U-Net empleada para estimar el error introducido a una imagen conociendo la imagen alterada y la cantidad de ruido introducido. En la mayoría de los casos se usan un *embedding* sinusoidal para el valor del paso temporal introducido, el cual se concatena con la salida del imagen a través de un red convolucional, formando juntos la entrada de la red U-Net. La salida de dicha red es el ruido predicho para el dato alterado. Para más detalles sobre la arquitectura U-Net, ver [Sección 4.1](#)

Además, el entrenamiento y el sampleado para esta perspectiva son dos procesos totalmente independientes en la perspectiva de *score*, por lo que pueden usarse muchas técnicas de generación de imágenes tras haber estimado el *score*.

2.2.1. Noise-Conditioned Score Networks (NCSNs)

Este modelo fue introducido por Song y Ermon[47], de forma que crearon un método generativo en el que los ejemplos se producían usando dinámica de Langevin mediante las estimaciones del gradiente de la distribución de probabilidad de los datos. Para aprender el campo vectorial asociado con el *score*, se entrena una red neuronal vía *score matching* a partir de los datos; es decir, se busca ajustar el modelo para que el gradiente de su *score* coincida con el de los datos observados. El “sampleado” se realiza moviendo un estado inicial aleatorio hacia zonas de alta densidad a través del campo vectorial de *scores* previamente estimado.

Debido al coste computacional de calcular $\nabla_x \log p(x)$, el método de *score matching* no es escalable a redes neuronales profundas ni a datos con alta dimensión; por lo que se usan otros procedimientos en el entrenamiento de las redes como *denoising score matching* o *sliced score matching*[29]. Además, el mayor problema al que se enfrentan para implementar este método es una consecuencia de la hipótesis de la variedad; es decir, el hecho de que los datos reales tienden a concentrarse en variedades de baja dimensión embebidas dentro de espacios de alta dimensión. A esto hay que añadir que las funciones de *score* no son precisas en regiones de baja densidad de datos. Para resolverlos, los creadores del método perturbaron los datos con ruido Gaussiano en distintas magnitudes, consiguiendo que la distribución resultante no colapsara a bajas dimensiones, y estimaron las funciones de *score* de todas las distribuciones de datos ruidosas entrenando una red neuronal condicionada a los niveles de ruido (NCSN).

Usando la misma notación que la que hemos empleado en la sección 2.1.1, consideremos $q(x)$ como la función de distribución de los datos reales y la secuencia $0 < \sigma_1 < \sigma_2 < \dots < \sigma_t < \dots < \sigma_T$ de niveles de ruido. Perturbamos una muestra x_0 a un valor x_t introduciendo un ruido Gaussiano, teniendo la distribución $q(x_t|x_0) = \mathcal{N}(x_t; x_0, \sigma_t^2 I)$. Así se obtiene un conjunto ordenado de distribuciones de datos ruidosos $\{q(x_t) = \int q(x_t|x_0)q(x_0)dx_0\}_{t=1,\dots,T}$. En este punto, entrenamos una red neuronal $s_\theta(x_t, t)$ para estimar el valor de la función de *score*, $\nabla_{x_t} \log q(x_t)$ para cada t . La función objetivo, en el *denoising score matching*, puede escribirse de la siguiente forma[47]:

$$\begin{aligned} \mathbb{E}_{\mathbf{x} \sim q(\mathbf{x})} \left[\|s_\theta(\mathbf{x}) - \nabla_{\mathbf{x}} \log q(\mathbf{x})\|_2^2 \right] &= \\ \mathbb{E}_{x_0 \sim q(x_0), \mathbf{x}_t \sim q(\mathbf{x}_t|x_0)} \left[\|s_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t|x_0)\|_2^2 \right] &= \\ \frac{1}{T} \sum_{t=1}^T \lambda(\sigma_t) \mathbb{E}_{x_0 \sim q(x_0), \mathbf{x}_t \sim \mathcal{N}(x_0, \sigma_t^2 I)} \left[\left\| s_\theta(\mathbf{x}_t, \sigma_t) + \frac{\mathbf{x}_t - \mathbf{x}_0}{\sigma_t} \right\|_2^2 \right] &= \\ \frac{1}{T} \sum_{t=1}^T \lambda(\sigma_t) \mathbb{E}_{x_0 \sim q(x_0), \epsilon_t \sim \mathcal{N}(0, I)} \left[\|\sigma_t s_\theta(\mathbf{x}_t, \sigma_t) + \epsilon_t\|_2^2 \right], \quad (2.31) \end{aligned}$$

donde $\lambda(\sigma_t)$ es una función de peso utilizada para mantener la dependencia temporal de la función perdida en la misma magnitud. En la última igualdad se ha tenido en cuenta que $x_t = x_0 + \sigma_t \epsilon$ con $\epsilon \sim \mathcal{N}(0, I)$. Comparando con la función objetivo de los DDPMs (Eq. (2.30)), se tiene que ambas son equivalentes si $\epsilon_\theta(x_t, t) = -\sigma_t s_\theta(x_t, t)$. Además, se puede generalizar los métodos de *score* para que tengan en cuenta las derivadas de orden superior, lo cual aporta información adicional local sobre la función de distribución de los datos. Una de las ventajas de este método, resaltada por sus creadores, es que este marco de trabajo permite un uso flexible de arquitecturas de redes neuronales para estimar la función de *score*, no requiere sampleado o métodos adversarios (como las GANs) durante el entrenamiento del modelo[47]. La Figura 2.3 es una representación esquemática del proceso de cálculo de la función de *score* siguiendo el proceso aquí explicado.

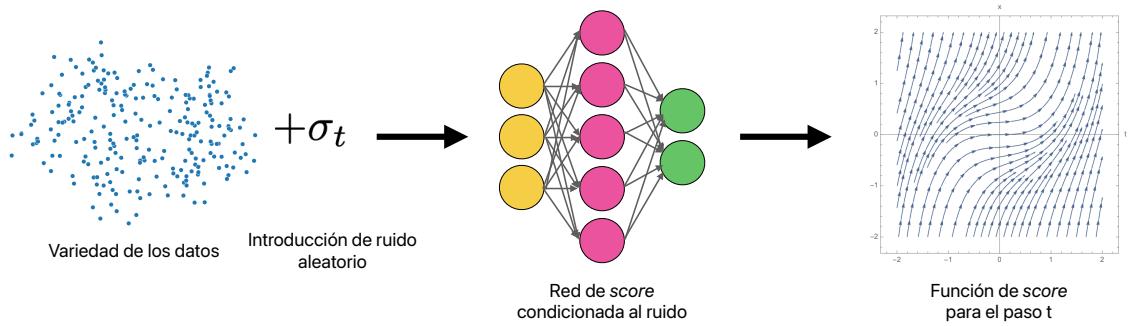


Figura 2.3: Esquema sobre el proceso de cálculo de la función de *score* con el enfoque *denoising score matching*. A los datos de entrenamiento se les introduce un ruido aleatorio en distintas cantidades, estimando la función de *score* para cada nivel de ruido usando una red neuronal.

Sampleado. Debido a la desvinculación entre el entrenamiento y la inferencia en los modelos generativos basados en *score*, existen varios métodos de sampleado de imágenes. Uno de ellos, es el *Annealed Langevin Dynamics* (ALD)[29]. Considérese N el número de iteraciones por cada momento temporal t , y δ_t el tamaño de salto en el momento t . Inicializamos el procedimiento con $x_T^{(N)} \sim \mathcal{N}(0, I)$, y en cada momento $0 \leq t < T$, tomamos $x_t^{(0)} = x_{t+1}^{(N)}$. Entonces, iteramos de acuerdo a la siguiente regla para $i = 0, \dots, N-1$, donde $\epsilon^{(i)} \sim \mathcal{N}(0, I)$:

$$x_t^{(i+1)} = x_t^{(i)} + \frac{1}{2} \delta_t s_\theta(x_t^{(i)}, t) + \sqrt{\delta_t} \epsilon^{(i)}. \quad (2.32)$$

Es decir, para cada $t = T, T-1, \dots, 1$, se aplica “Langevin Monte Carlo” de forma

sucesiva. Este método garantiza que $x_0^{(N)}$ sea una realización muestral válida para la distribución $q(x_0)$ cuando $\delta_t \rightarrow 0$ y $N \rightarrow \infty$. Por tanto, $x_0^{(N)}$ será el ejemplo que obtenemos. A continuación se muestra el algoritmo de sampleado propuesto por [47] y la Figura 2.4 representa un sencillo esquema del algoritmo.

Algorithm 3 ALD para sampleado.

Require: $\{\sigma_t\}_1^T, \eta, N$.

```

1: Inicializar  $x_T^{(0)}$ .
2: for  $\text{dot} = T, \dots, 1$ ,
3:    $\delta_t = \eta \sigma_t^2 / \sigma_T^2$ 
4:   for  $\text{doi} = 1, \dots, N$ ,
5:     Generar  $\epsilon^{(i)} \sim \mathcal{N}(0, I)$ 
6:      $x_t^{(i+1)} = x_t^{(i)} + \frac{1}{2} \delta_t s_\theta(x_t^{(i)}, t) + \sqrt{\delta_t} \epsilon^{(i)}$ .
7:   end for
8:    $x_{t-1}^{(0)} = x_t^{(N)}$ 
9: end for
10: return  $x_1^{(N)}$ 
```

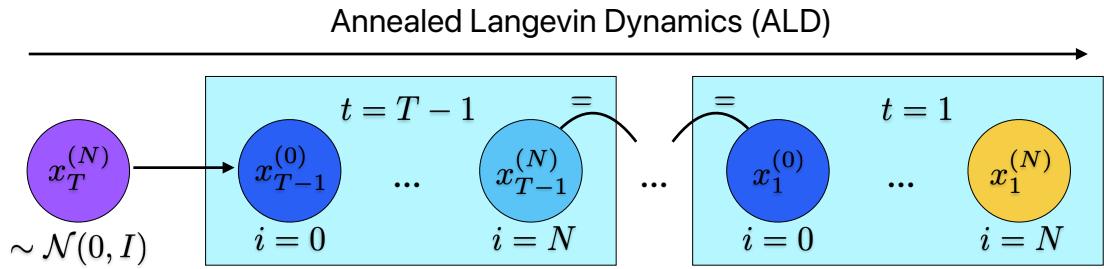


Figura 2.4: Esquema sobre el proceso de sampleado usando el método de *Annealed Langevin Dynamics*. Cada cuadro azul celeste corresponde con un paso temporal de introducción de ruido en los datos reales, dentro de los cuales, en cada paso, la imagen se mueve en la dirección de la función de score N veces.

2.2.2. Score SDEs

Los dos tipos anteriores de modelos de difusión, tanto DDPM como NCSN, pueden generalizarse al caso en que se consideren infinitos niveles o pasos de introducción de ruido ($T \rightarrow \infty$). De hecho, en el entrenamiento del DDPM se calculan implícitamente los *scores* en cada nivel de ruido[48]. Al considerar un continuo de ruido, los proce-

2 Fundamento teórico de los modelos de difusión

sos de perturbación y de sampleado por eliminación de ruido se modelan mediante ecuaciones diferenciales estocásticas (SDEs).

Considerando un continuo de distribuciones de ruido que evolucionan con el tiempo según el proceso de difusión, éstas se usan para perturbar datos en ruido aleatorio. Este proceso viene dado por una SDE independiente de los datos y sin parámetros que entrenar. El proceso inverso viene dado también por una SDE, la cual se aproxima entrenando una red neuronal dependiente del tiempo, que se encarga de estimar los *scores*, y mediante el uso de métodos numéricos de resolución de SDEs.

Este enfoque tiene como ventajas la posibilidad del cálculo exacto de la función de verosimilitud, el uso flexible de métodos para el sampleado, el control de la generación de datos (condicionando la información en el sampleado que no estaba disponible durante el entrenamiento) y un marco teórico único para todos los modelos generativos basados en la función de *score*[48].

Muchos procesos estocásticos, como los procesos *forward* de difusión, son solución de una SDE de la forma:

$$dx = f(x, t)dt + g(t)dw, \quad (2.33)$$

donde $f(x, t)$ es el coeficiente de *drift* de la SDE, $g(t)$ es el coeficiente de difusión, independiente de x y w representa el movimiento Browniano estándar. Por tanto, el modo de perturbar los datos $x_0 \sim p_0$ a ruido $x_T \sim p_T$ viene gobernado por dicha ecuación. Hay varias formas de diseñar la SDE de la Ecuación 2.33 de forma que altere los datos originales a una distribución prefijada, como una Gaussiana. Por ejemplo, tomando la siguiente ecuación:

$$dx = \sqrt{\frac{d[\sigma^2(t)]}{dt}} dw, \quad (2.34)$$

si se discretiza, se consigue el proceso de perturbación usado para el NCSN. No obstante, tomando esta otra SDE:

$$dx = -\frac{1}{2}\beta(t)xdt + \sqrt{\beta(t)}dw, \quad (2.35)$$

su discretización nos devuelve el proceso DDPM[48].

El proceso de generación de imágenes a partir de realizaciones muestrales de p_T está dado por la correspondiente SDE de tiempo inverso para el proceso de difusión:

$$dx = [f(x, t) - g(t)^2 \nabla_x \log q_t(x)]dt + g(t)d\bar{w}, \quad (2.36)$$

siendo dt el paso de tiempo infinitesimal y negativo, y \bar{w} el movimiento Browniano marcha atrás.

Las trayectorias de las soluciones de la SDE inversa comparten las mismas distribuciones marginales que las de la SDE para la difusión, salvo que dichas distribuciones evolucionan en sentido contrario. La forma de conocer la función de *score* es entrenando una red neuronal como se ha explicado anteriormente (2.2.1). Por este motivo, este método se denomina *Score SDE*, ya que combina las herramientas de cálculo de la función de *score* y las SDEs para llevar a cabo la generación de imágenes. Aplicando cualquier método numérico apropiado a la Eq. (2.36) se puede realizar el sampleado, aunque existe una técnica previa muy empleada para dicha tarea. Existe una ecuación diferencial ordinaria (ODE), denominada, la ODE de flujo de probabilidad, cuyas trayectorias tienen las mismas marginales que la SDE inversa, por lo que resolviendo la ecuación se puede también samplear imágenes a partir del mismo ruido. La expresión de esta ecuación diferencial para la Eq. (2.36) viene dada por la Eq. (2.37).

$$dx = \left[f(x, t) - \frac{1}{2}g(t)^2 \nabla_x \log q_t(x) \right] dt \quad (2.37)$$

2.3. Conditional Diffusion Probabilistic Models (CDPMs)

Hasta ahora, en la explicación sobre el fundamento teórico de los modelos de difusión no se ha tenido en cuenta la posible existencia de distintas clases y/o etiquetas dentro de los datos de entrenamiento. Por ejemplo, si entrenamos un modelo DDPM para que sea capaz de generar imágenes de lesiones de piel, no hemos tenido en cuenta la existencia de los distintos tipos de lesiones que pueden aparecer en la piel: lunares benignos, melanomas, etc. En ese caso, el modelo generará imágenes en las que no habrá distinción entre las distintas clases, llegando incluso a crear híbridos entre esas clases. A dichos modelos se los denomina modelos de difusión incondicionales, porque ni durante el entrenamiento, ni en el sampleado de imágenes, se hizo uso de una etiqueta para diferenciar las imágenes según la clase a la que pertenecían.

Los modelos de difusión condicionales extienden el marco de los incondicionales introduciendo información adicional (condiciones) sobre la clase o características de la imagen que guíen el proceso de entrenamiento y generación[10]. Las condiciones pueden ser de estilo muy diverso, desde etiquetas de clase, imágenes parciales, textos descriptivos, etc. El objetivo de estos modelos es la generación de datos que sean coherentes con las condiciones proporcionadas.

La condición, que denominaremos como c , se integra con la red neuronal ligada al modelo de difusión escogido durante el entrenamiento de la misma.

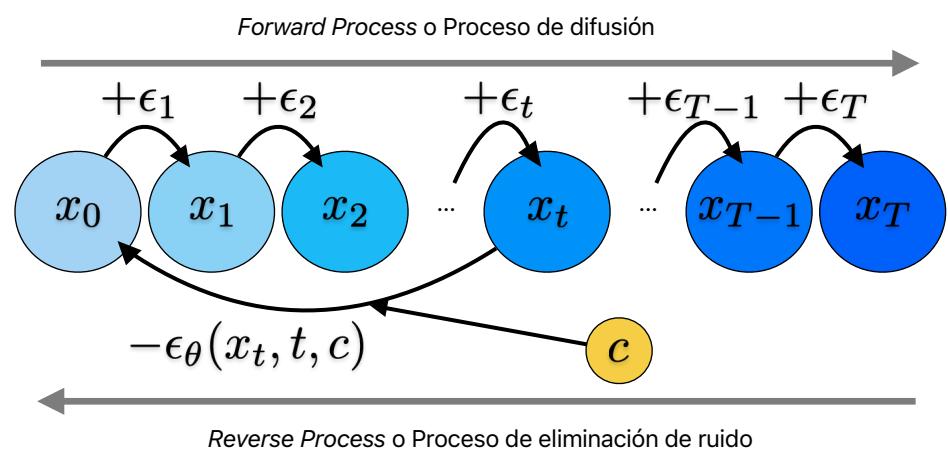


Figura 2.5: Esquema sobre la formulación de los DDPMs condicionales. En este caso, se usa la condición c en cada uno de los pasos del sampleado para conseguir una generación de datos controlada.

3 Métricas de calidad en la generación de imágenes

Introducción

La generación de imágenes médicas mediante modelos difusos ha emergido como una de las áreas de investigación más prometedoras en la intersección entre la inteligencia artificial y la medicina. Estos modelos han demostrado un potencial significativo para crear imágenes de alta calidad que pueden ser utilizadas en diversas aplicaciones clínicas, desde el diagnóstico hasta la planificación de tratamientos. Sin embargo, la evaluación de la calidad de estas imágenes generadas es un desafío crítico debido a la naturaleza compleja y multifacética de las imágenes médicas.

La evaluación precisa y confiable de las imágenes médicas generadas es realmente importante para asegurar que estos modelos no solo produzcan imágenes visualmente atractivas, sino también clínicamente coherentes, relevantes y útiles. A diferencia de otros dominios, las imágenes médicas deben cumplir con estrictos criterios de exactitud y detalle, ya que cualquier error o imprecisión puede tener consecuencias graves en el diagnóstico y tratamiento de los pacientes. Este desafío se agrava por la variabilidad inherente en las modalidades de imagen médica (como resonancia magnética, tomografía computarizada, y ultrasonido) y la necesidad de evaluar no solo la apariencia visual, sino también la fidelidad anatómica y la relevancia clínica. A todo ello se suma el problema de la falta de imágenes para entrenar estos modelos, lo cual hace más difícil aún el correcto aprendizaje de las distintas redes neuronales que conforman los modelos de difusión.

Es por todo ello que no solo se busca que en el entrenamiento de los modelos se minimice la función de pérdida correspondiente, lo cual indica que los parámetros son los adecuados para poder conseguir imágenes próximas a las imágenes de entrenamiento, sino que también se requiere de formas de cuantificar y medir cómo de realistas son las imágenes generadas a partir del modelo. Un valor de la función objetivo en el entrenamiento del modelo no aporta información útil sobre su capacidad para generar imágenes anatómicas de calidad y realistas. Sin embargo, combinando esa información junto con la aportada por distintas métricas de calidad, sí se puede tener una mejor idea de cuán bueno será el modelo obtenido.

En esta sección, revisaremos las distintas métricas de evaluación que se han propuesto y utilizado en la literatura para medir la calidad de las imágenes generadas por modelos difusos. Estas métricas pueden agruparse en dos categorías principales, según si son de carácter cualitativo o cuantitativo. Dentro de estos tipos se incluyen métricas como las basadas en la percepción humana, métricas clínicas específicas o métricas

3 Métricas de calidad en la generación de imágenes

derivadas de la teoría de la información y el aprendizaje automático. Cada una de estas categorías aborda diferentes aspectos de la calidad de la imagen, desde la claridad visual hasta la precisión anatómica y la utilidad clínica.

Discutiremos en detalle la importancia de cada categoría de métricas, su implementación, y sus ventajas y limitaciones en el contexto de la generación de imágenes médicas. Además, exploraremos estudios recientes que han aplicado estas métricas a conjuntos de datos reales, proporcionando una visión crítica sobre el estado actual de la evaluación de la calidad en este campo emergente.

3.1. Métricas de calidad cualitativas

Las técnicas cualitativas de evaluación de imágenes generadas se centran en analizar la calidad, realismo y coherencia de las salidas producidas por los modelos generativos, sin tener que recurrir a métricas numéricas que se calculan a partir de magnitudes cuantitativas de la imagen.

Por lo general, se usa un grupo de personas, al cual se le pide que evalúe las imágenes generadas. Esta técnica se basa en la percepción subjetiva de la calidad y se utiliza frecuentemente debido a la capacidad humana para detectar detalles y coherencias que las métricas automáticas podrían pasar por alto. Dicha evaluación se puede realizar de distintas formas; por ejemplo, usando cuestionarios en los que haya que puntuar ciertas características de la imagen, como su calidad, detalle, coherencia, etc. Entre las ventajas de estos métodos está la capacidad del ser humano de capturar detalles subjetivos y matices que las métricas automáticas pueden ignorar, así como de la coherencia de la imagen. Sin embargo, el estado mental del evaluador, prejuicios y sesgos individuales afectan a su capacidad de percepción, además de que es un método costoso y largo.

De hecho, para el caso de la generación de imágenes médicas, la evaluación de las mismas no las puede hacer un persona cualquiera, como sí podría en el caso de tener imágenes aleatorias y genéricas. Se requiere de un grupo de expertos médicos o profesionales con una sólida formación en anatomía humana los que califiquen dichos ejemplos. Esto encarece mucho más el método y hace más difícil su implantación.

La evaluación comparativa o *A/B testing*, consiste en mostrar pares de imágenes a los evaluadores y se les pide que elijan la que prefieren o la que consideran de mejor calidad. Esta técnica es útil para comparar directamente dos modelos de difusión, dos configuraciones diferentes de un mismo modelo o una imagen generada con una real. De este modo, la persona hace una comparación directa y sencilla y se reduce el sesgo al enfocarse en elecciones binarias. No obstante, puede llegar un punto en el que se desconozcan los aspectos específicos que influencian la preferencia, al llegar un punto en el que sea puramente subjetiva. Incluso puede ser influenciada por la presentación de los pares de imágenes.

El análisis de detalles y coherencia técnica implica una revisión exhaustiva de las imágenes para identificar detalles específicos como texturas, bordes, integridad de objetos y coherencia entre diferentes partes de la imagen. Los evaluadores buscan artefactos, inconsistencias y anomalías que podrían no ser evidentes a simple vista. Las ventajas que presenta esta evaluación es que proporciona una evaluación detallada y específica de las imágenes, y permite identificar áreas problemáticas para mejoras futuras del modelo. Pero, es una técnica intensiva en términos de tiempo y recursos, y requiere evaluadores con experiencia y conocimientos técnicos, tanto en imagen como, es este caso, de anatomía humana.

Las técnicas cualitativas de evaluación de imágenes generadas son importantes para comprender cómo los usuarios perciben la calidad y el realismo de estas imágenes. A

3 Métricas de calidad en la generación de imágenes

pesar de sus desafíos, estas técnicas son un complemento las métricas automáticas y ofrecen una visión más completa del rendimiento de los modelos de difusión, ya que proporcionan el sentido humano de coherencia de la imagen generada. La combinación de múltiples técnicas cualitativas puede proporcionar una evaluación más robusta y equilibrada, ayudando a mejorar continuamente la calidad de las imágenes generadas. Sin embargo, el mayor reto que presentan estas métricas en el campo de la generación de imagen médica es la necesidad de tener a personas con preparación y experiencia en medicina y anatomía humana para poder evaluar dichas imágenes.

3.2. Métricas de calidad cuantitativas

Este tipo de métricas permite medir de manera objetiva y consistente el rendimiento de los algoritmos, facilitando la comparación y mejora continua. Entre las más métricas comunes [6, 34] se encuentran la puntuación de Frechet Inception Distance (FID), que evalúa la similitud entre las distribuciones de características de las imágenes generadas y las reales; el Inception Score (IS), que mide tanto la calidad como la diversidad de las imágenes generadas; y el índice de similitud estructural (SSIM), usado para estudiar la similaridad entre las imágenes. Estas herramientas son esenciales para avanzar en el desarrollo de modelos de IA capaces de generar imágenes realistas y variadas.

3.2.1. Inception Score (IS)

El IS es una métrica popular empleada principalmente para juzgar el realismo de las imágenes sintetizadas por modelos generativos, como las GANs, ya que en palabras de sus autores “we find [the IS] to correlate well with human evaluation [of image quality]” (“encontraos que [el IS] correlaciona bien con la evaluación humana [de la calidad de imagen]”) [45]. Junto con el Fréchet Inception Distance (3.2.2), son las métricas más empleadas en la actualidad para este tipo de evaluaciones.

Esta métrica se calcula usando la salida de un modelo de clasificación de imágenes (usualmente *InceptionV3* entrenado con *ImageNet*), pre-entrenado de forma independiente, sobre un conjunto de imágenes generadas por el modelo a evaluar. De hecho, toma su nombre del clasificador *Inception*. El valor del IS se maximiza cuando se cumple que:

- La entropía de la distribución de *labels* predicha por el modelo *InceptionV3* para las imágenes generadas es mínima; es decir, el clasificador predice con seguridad una sola etiqueta para cada imagen.
- Las predicciones del clasificador se distribuyen uniformemente por el conjunto de todos los *labels* posibles.

Por tanto, cuando mayor sea el valor del *score*, mayor variedad de imágenes realista es capaz de generar el modelo.

Considérese Ω_X el espacio de imágenes, Ω_Y el espacio finito de N *labels* y p_{gen} la distribución de probabilidad del generador sobre Ω_X , la cual deseamos juzgar. Sea $p_{dis} : \Omega_X \rightarrow M(\Omega_Y)$, la función que a cada imagen le asigna una distribución de probabilidad de *labels* (esta función viene descrita por el clasificador pre-entrenado). De esta forma, definimos el Inception Score de p_{gen} relativo a p_{dist} como:

$$IS(p_{gen}, p_{dis}) := \exp \left(\mathcal{E}_{x \sim p_{gen}(x)} \left[\mathcal{D}_{KL} \left(p_{dist}(\cdot|x) \middle\| \int_{\Omega_X} p_{dist}(\cdot|x) p_{gen}(x) dx \right) \right] \right) \quad (3.1)$$

3 Métricas de calidad en la generación de imágenes

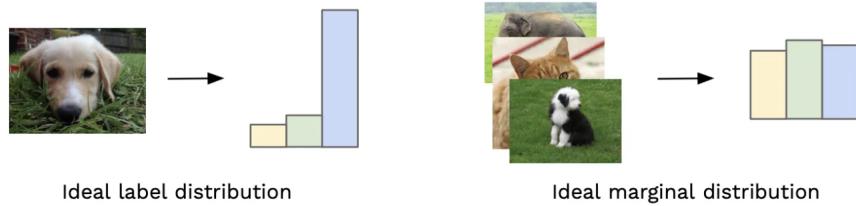


Figura 3.1: Imagen extraída de <https://medium.com/octavian-ai/a-simple-explanation-of-the-inception-score-372dff6a8c7a>. Vemos que el comportamiento ideal del generador es que para una imagen determinada, su etiqueta esté muy bien determinada, mientras que sea capaz de generar imágenes de todos los tipos.

Usando la desigualdad de Jensen se puede comprobar que $\ln IS \geq 0$. El motivo de la forma de la Eq. (3.1) reside en que lo que se busca es que la distribución marginal de *labels*, $\int_{\Omega_X} p_{\text{dist}}(\cdot|x)p_{\text{gen}}(x)dx$, difiera mucho de la distribución de etiquetas de cada imagen, $p_{\text{dist}}(\cdot|x)$, (porque para una imagen determinada, queremos que se muestre claramente una etiqueta; mientras que también deseamos que el generador sintetice imágenes de todas las etiquetas con la misma frecuencia, ver figura 3.1).

Puede probarse que (ver anexos) $\ln IS(p_{\text{gen}}, p_{\text{dis}}) \in [0, \ln N]$. Si es nulo, eso significa que $\int_{\Omega_X} p_{\text{dist}}(\cdot|x)p_{\text{gen}}(x)dx = p_{\text{dist}}(\cdot|x)$, por lo que para cualquier imagen x obtenida del generador, el clasificador devuelve exactamente la misma distribución en el espacio de etiquetas. No obstante, el mayor valor se alcanza cuando toda imagen es correctamente clasificada y $\mathcal{E}_{x \sim p_{\text{gen}}(x)} p_{\text{dist}}(\cdot|x) = 1/N$ (las imágenes generadas se distribuyen por igual en el conjunto de *labels*).

Sin embargo, esta métrica presenta ciertas limitaciones [7]:

- Las ejecuciones de entrenamiento de la red *Inception* en *ImageNet* generan diferentes pesos debido a la aleatoriedad, pero esto apenas afecta la precisión de clasificación. Sin embargo, pequeños cambios en los pesos pueden producir grandes diferencias en las puntuaciones del IS para el mismo conjunto de imágenes.
- La puntuación está limitada por lo que el clasificador *Inception* (u otra red) puede detectar, lo cual está ligado a los datos de entrenamiento. Como consecuencia, si se aprende a generar algo que no está presente en los datos de entrenamiento del clasificador, es posible que siempre se obtenga un IS bajo a pesar de generar imágenes de alta calidad. Además, si el clasificador no puede detectar características relevantes para el concepto de calidad de imagen, entonces las imágenes de mala calidad aún pueden obtener altas puntuaciones.
- No hay una medida de diversidad intra-clase, por lo que si se genera la misma

imagen para cada clase de forma repetida, se obtiene un alto valor del *score*.

- Si el generador sobreajusta a los datos de entrenamiento, también se puede obtener un alto valor del IS.

3.2.2. Fréchet Inception Distance (FID)

El FID es una métrica muy empleada para evaluar la calidad de las imágenes generadas por un modelo. A diferencia del IS, el cual evaluaba la distribución de las imágenes generadas, el FID compara dicha distribución con la de un conjunto de imágenes reales, por lo que éste puede verse como una mejora del primero. Fue introducida por Heusel et al. en 2018[25], y es una de las métricas más empleadas en el estudio de IA generativa.

Dadas dos distribuciones de probabilidad p y q sobre \mathbb{R}^n con media y varianza finitas, su distancia de Fréchet es la distancia de 2-Wasserstein sobre \mathbb{R}^n .

$$d_F(p, q) := \sqrt{\inf_{\mu \in \Gamma(p, q)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|^2 \mu(x, y) dx dy}, \quad (3.2)$$

donde $\Gamma(p, q)$ es el conjunto de todas las distribuciones de probabilidad en $\mathbb{R}^n \times \mathbb{R}^n$ cuyas marginales son p y q en el primer y segundo factor respectivamente.

En el caso de dos distribuciones normales multivariantes $\mathcal{N}(\mu_1, \Sigma_1)$ y $\mathcal{N}(\mu_2, \Sigma_2)$, su distancia de Fréchet sería: [16]

$$d_F(\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)) = \|\mu_1 - \mu_2\|_2^2 + \text{traza} \left(\Sigma_1 + \Sigma_2 - 2\sqrt{\Sigma_1 \Sigma_2} \right) \quad (3.3)$$

Para Ω_X el espacio de imágenes, consideramos S el subconjunto de imágenes generadas por el modelo y S' un subconjunto de imágenes reales. Dada $f : \Omega_X \rightarrow \mathbb{R}^n$ la función que representa de un modelo pre-entrenado con imágenes reales (tradicionalmente, el vector de activación 2048-dimensional de la última capa de *pooling* de un modelo *InceptionV3* entrenado con *ImageNet*), la forma práctica de operar consiste en obtener por f los vectores de características de las imágenes de S y S' , y ajustar los resultados a dos Gaussianas multivariantes (calcular la media y varianza de los datos). Finalmente, se calcula d_F para dichas distribuciones. El motivo de suponer la normal multivariante reside en que esta es la distribución con mayor entropía para un valor dado de media y varianza.

Cuanto menor sea el valor del FID, menor será la distancia entre ambas distribuciones en el espacio de funciones correspondiente, por lo que el rendimiento del modelo será mejor y las imágenes generadas tendrán mucha más calidad.

Sin embargo, existen importantes limitaciones para esta métrica, señaladas por Jayasumana et al.[28]. Por un lado, los vectores de 2048 características obtenidos de *Inception*

3 Métricas de calidad en la generación de imágenes

para conjuntos de imágenes típicos están lejos de seguir una distribución normal. Esto implica que distribuciones distintas pueden obtener un valor pequeño de la distancia de Fréchet. Por otro lado, estimar matrices de covarianzas 2048×2048 a partir de un conjunto pequeño puede conducir a graves errores de cálculo. Además, en ese mismo artículo, sus autores proponen una nueva métrica alternativa, CMMD, la cual “offers a more robust and reliable assessment of image quality”.

También es importante destacar que en [35] y [28] demuestran empíricamente mediante ejemplos que, a veces, el FID proporciona resultados inconsistentes con la evaluación humana. A pesar de ello, sigue siendo una métrica muy empleada por ser, de momento, lo mejor que se tiene y conoce. En el caso de [35], la mejora propuesta es una pequeña variación del FID, que denomina CAFD (promedio del FID para cada una de las clases de imágenes). Otros fallos de esta métrica se han encontrado y se propuesto variaciones de la misma para solventarlos [9] o para adaptarlas mejor al problema [8]. A pesar de todos estos contras, el FID sigue siendo una métrica empleada en la evaluación de la calidad de los modelos generativos, en específico en imagen médica [27].

Aunque para ambas métricas se haya mencionado la red *InceptionV3* pre-entrenada con *ImageNet*, por ser el modelo usado por los autores de estas métricas, pueden usarse otras arquitecturas de redes entrenadas con datos más acordes al problema de generación que se quiera resolver para así obtener mejores resultados.

3.2.3. SSIM

La métrica de similitud estructural (SSIM) se basa en la asunción de que el sistema visual humano está altamente adaptado para extraer información estructural de una imagen o escena, por lo que esta medida puede proporcionar una buena aproximación de lo que se considera calidad de imagen. Este índice viene mejorado al introducir un enfoque multi-escala del mismo (MS-SSIM) [50]. Esta variación se basa en que la percepción de los detalles de una imagen depende de la densidad de señal de la imagen, la distancia del observador a la imagen y la capacidad del sistema visual del observador. Por ello, el método multi-escala incorpora detalles de la imagen en diferentes resoluciones, para imitar los efectos mencionados.

Considérese $\{x_i, y_i : i = 1, \dots, N\}$ dos señales discretas no negativas y alineadas. Pueden considerarse μ_x y σ_x^2 , la media y la varianza del vector x como estimaciones de la luminosidad y el contraste de la señal x , mientras que σ_{xy} es una indicación de la similitud estructural de ambas.

Definiendo las medidas de comparación de luminancia, contraste y estructura como:

$$l(x, y) = \frac{2\mu_x\mu_y + (K_1 L)^2}{\mu_x^2 + \mu_y^2 + (K_1 L)^2} \quad (3.4)$$

$$c(x, y) = \frac{2\sigma_x\sigma_y + (K_2 L)^2}{\sigma_x^2 + \sigma_y^2 + (K_2 L)^2} \quad (3.5)$$

$$s(x, y) = \frac{\sigma_{xy} + (K_2 L)^2/4}{\sigma_x\sigma_y + (K_2 L)^2/4}, \quad (3.6)$$

donde L es el rango dinámico de los valores de los píxeles. Para evitar que los denominadores sean muy próximos a cero, suele tomarse como valores de las constantes $K_1 = 0.01$ y $K_2 = 0.03$. La forma general del índice de similitud estructural (SSIM) es:

$$\text{SSIM}(x, y) = |l(x, y)|^\alpha |c(x, y)|^\beta |s(x, y)|^\gamma, \quad (3.7)$$

donde suele tomarse que $\alpha = \beta = \gamma = 1$ (igualdad de importancia para las tres componentes). Este índice presenta las propiedades simetría y que está acotado por 1, el cual se alcanza si y solo si $x = y$.

Considerando ahora una imagen real y otra sintética a comparar como inputs, el sistema aplica de forma interativa M veces un filtro de paso bajo y submuestrea la imagen filtrada por un factor igual a 2, lo que se denomina una escala. Para cada escala j , se calcula el índice de contraste y de estructura. El índice de luminosidad se calcula en la escala final. Así se define el índice de similitud estructural multi-escala (MS-SSIM) como:

$$\text{MS-SSIM}(x, y) = |l_M(x, y)|^{\alpha_M} \prod_{j=1}^M |c_j(x, y)|^{\beta_j} |s_j(x, y)|^{\gamma_j}. \quad (3.8)$$

Cuanto más cercano a uno sea el valor del índice, más similares son las imágenes perceptualmente. A pesar de que esta métrica sea un buen método para comparar imágenes, a veces falla al considerar matices que sí perciben los humanos [6].

Existen variantes del SSIM más acordes al problema que se quiera resolver, como el SSIM multi-componente (3-SSIM) o la disimilitud estructural (DSSIM). Esta métrica se utiliza como medida de la pérdida por compresión de imagen, restauración de imagen y reconocimiento de patrones. También es muy empleada actualmente en imagen médica para evaluar la calidad y diversidad de imágenes generadas por distintos modelos [21, 4].

3.2.4. Proporción Máxima de Señal a Ruido (PSNR)

La métrica PSNR (Peak Signal-to-Noise Ratio) es ampliamente utilizada en el ámbito del procesamiento de imágenes y video para medir la calidad de la reconstrucción o compresión de una imagen o secuencia de video. Es una medida objetiva que compara la calidad de una imagen comprimida o modificada con respecto a la imagen original. Su cálculo se basa en el error cuadrático medio (MSE) entre ambas imágenes:

$$\text{MSE} = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i,j) - K(i,j)]^2, \quad (3.9)$$

donde $I(i,j)$ es el valor del píxel en la posición (i,j) de la imagen original, $K(i,j)$ es el valor del píxel en la posición (i,j) de la imagen comprimida o reconstruida y m y n son las dimensiones de la imagen. Para imágenes en formato RGB, el MSE se calcula como la media aritmética de los MSEs en los tres canales de colores. Así, el PSNR se define como

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right), \quad (3.10)$$

donde MAX_I es el valor máximo posible de un píxel en la imagen. Para imágenes de 8 bits por canal, $\text{MAX}_I = 2^8 - 1 = 255$.

El PSNR se mide en decibelios (dB) y valores más altos indican mejor calidad de la imagen reconstruida o comprimida, ya que el error entre la imagen original y la imagen procesada es menor. En general, un valor de PSNR superior a 30 dB se considera que indica buena calidad, aunque este umbral puede variar dependiendo de la aplicación específica.

Las ventajas que presenta es su facilidad de implantación y su amplio reconocimiento en el ámbito del procesamiento de imagen y vídeo. No obstante, no se correlaciona con la percepción humana y no es sensible a pequeñas distorsiones localizadas, que sí pueden ser significativas. Estas limitaciones son muy importantes en imagen médica, donde el detalle anatómico que se requiere en las imágenes generadas es bastante alto.

Debido a que las imágenes generadas por modelos de difusión pretenden ser nuevos ejemplos con características similares a las imágenes de entrenamiento, y no ser idéntica a ninguna imagen de referencia, el PSNR no parece ser una medida buena de la calidad de los samples. No obstante, sí que se emplea en algunos estudios de la calidad de ciertos modelos generativos para imagen médica[4]. En estos casos, se puede tomar el promedio del PSNR de una imagen sintética con cada una de las imágenes reales, ya que a pesar de las diferencias de cada individuo, existen similitudes anatómicas que deben compartir dichas imágenes y que medirá el PSNR. Sin embargo, en el trabajo expuesto en este estudio no se tendrá en cuenta esta métrica.

3.2.5. Classification Accuracy Score (CAS)

Esta métrica está basada en la sencilla idea de que si un modelo captura correctamente la distribución real de los datos, entonces su rendimiento en cualquier tarea posterior (*downstream task*) a su entrenamiento deberá ser similar usando los datos generados por este modelo a si se emplean los datos reales[41].

Por ejemplo, se podría entrenar un clasificador utilizando únicamente datos sintéticos generados por uno (si es un modelo generativo condicional) o varios (si se entrena uno por cada *label* del conjunto imágenes) modelos generativos, y se evalúa el rendimiento del clasificador sobre los datos reales, de los que conocemos su clase. Si este rendimiento se mide a través del *accuracy*, obtenemos una medida indirecta de la calidad de las imágenes generadas por el modelo, el *Classification Accuracy Score*.

Con esta misma idea, pueden definirse muchas otras métricas usando distintos *scores* como la precisión o el *recall*, entre otros muchos. La elección de la métrica base a usar en el *Classification Score* dependerá del problema objeto de estudio y de la distribución de las clases, entre otros.

Nótese que un buen valor del CAS no implica que el modelo generativo modele de forma precisa la distribución real de los datos. Entre los motivos teóricos, puede darse en caso en que el modelo “memorice” el conjunto de entrenamiento, consiguiendo el mismo CAS que para el conjunto de imágenes reales (esto también puede suceder con el IS o el FID). Aunque, es de esperar que las imágenes sintéticas difieran del conjunto real.

Una variante del CAS es el *Naive Augmentation Score* (NAS), el cual entrena el clasificador usando tanto los datos reales como los sintéticos.

A pesar de estos problemas, se encuentra que los modelos generativos tienen valores de CAS menores que los de los datos originales, indicando que estos fallan a la hora de conseguir replicar la distribución de los datos[41]. Otro hecho bastante importante es que parece que esta métrica no correlaciona con otras métricas tan importantes como el IS o el FID, lo que muestra la importancia de conocer el valor de distintas métricas a la hora de evaluar la calidad y variedad de las imágenes generadas por estos modelos como los modelos de difusión.

Como se puede ver, cada métrica de evaluación tiene sus aspectos positivos y negativos; por lo tanto, la comunidad investigadora sigue intentando encontrar una opción mejor. No existe una métrica perfecta para el estudio de imágenes generadas. La mejor opción depende de la aplicación específica y de qué aspectos de la calidad de la imagen son más importantes. A menudo, se usa una combinación de métricas para obtener una visión más completa de las fortalezas y debilidades de un modelo generativo.

Es importante considerar la evaluación cualitativa junto con las métricas cuantitativas. La inspección visual puede revelar problemas que las métricas podrían pasar

3 Métricas de calidad en la generación de imágenes

por alto, como artefactos, iluminación antinatural o inconsistencias anatómicas en las imágenes generadas. Por ejemplo, un puntaje FID podría indicar alta fidelidad, pero un observador humano podría notar artefactos extraños en el fondo que la métrica no capturaría. Al combinar la evaluación humana (naíf o experta) con métricas cuantitativas adecuadas, los desarrolladores pueden lograr una evaluación más equilibrada del rendimiento de los modelos de fusión e identificar áreas de mejora en sus modelos.

Esto es especialmente relevante en el campo de la imagen médica, donde la precisión y la calidad son críticas. Las imágenes generadas pueden ser utilizadas para la detección, diagnóstico y planificación del tratamiento. En este contexto, los artefactos o las inconsistencias anatómicas no solo afectan la calidad visual, sino que también pueden tener implicaciones serias para la salud del paciente. Por lo tanto, además de utilizar métricas tradicionales, es esencial realizar una evaluación cualitativa detallada por parte de profesionales médicos para asegurar que las imágenes generadas sean clínicamente útiles y seguras.

El campo de las métricas de evaluación está en constante evolución. Se están desarrollando nuevas métricas que buscan capturar mejor los matices de la percepción humana y la calidad de la imagen. En el contexto de la imagen médica, estas nuevas métricas también deben tener en cuenta los criterios clínicos y la capacidad de las imágenes generadas para soportar decisiones médicas precisas.

4 Extractores de características

Introducción

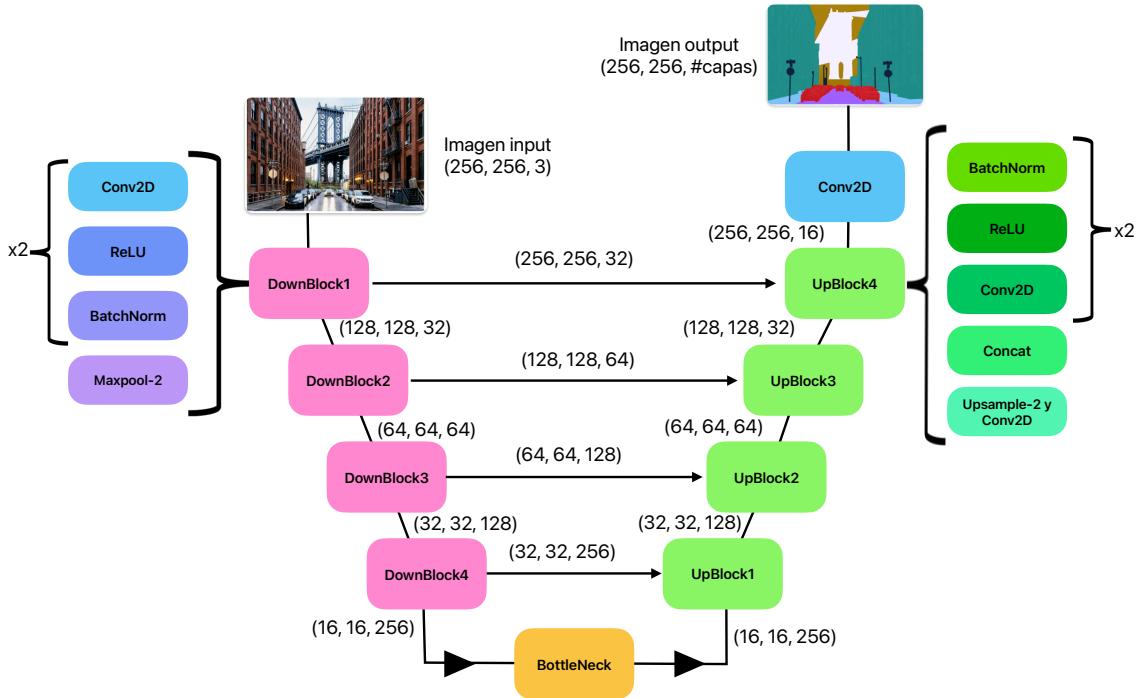
En la sección anterior ([Capítulo 3](#)), se mostraron varias de las métricas cuantitativas más empleadas en la actualidad para evaluar la calidad y diversidad de las imágenes generadas por modelos generativos como los modelos de difusión. Entre esas métricas podemos destacar el Fréchet Inception Distance (FID) ([Subsección 3.2.2](#)), el cual computaba la distancia de Fréchet entre las distribuciones de las características de imágenes reales e imágenes generadas. Cuanto menor es el valor del FID, más parecidas son las distribuciones; es decir, más similares son las imágenes sintéticas y las reales.

Las propiedades de una imagen pueden venir dadas como resultado de aplicar descriptores de características empleados en visión por computador, como Histogram of Oriented Gradients (HOG) o Local Binary Patterns (LBP), entre otros. Otra aproximación sería usando redes neuronales. La ventaja que supone el uso de redes neuronales sobre otros descriptores es que las características que éstas extraen de una imagen dependerán del problema en concreto que se les plantee, por lo que las redes neuronales serían descriptores “personalizados” a las imágenes que se empleen. Esta adaptación al problema lleva asociada un inconveniente ligado a la propia naturaleza de las redes neuronales, y es su, en parte, inexplicabilidad. Sería difícil tener conocimiento alguno sobre qué información considera la red importante de la imagen; lo cual se traduciría, en el caso del ámbito de imagen médica, en que no sabríamos qué elementos considera la red para distinguir entre un tumor benigno o maligno en una mamografía, o entre un melanoma o un lunar en una imagen dermoscópica.

Arquitecturas de redes neuronales como ResNet50^[24], SkinLesNet^[5] o U-Net^[43] se han consolidado como herramientas fundamentales en la extracción de características de imágenes médicas debido a su arquitectura eficiente y su capacidad para capturar representaciones ricas y discriminativas de los datos visuales. Esta sección explora el uso de estas redes en la extracción de características de imágenes, y su aplicación en el cálculo del FID para evaluar la calidad de imágenes generadas por modelos de difusión. Se discutirán los principios fundamentales de estos modelos, su implementación práctica y las ventajas que ofrecen en la mejora de la evaluación de la calidad de imágenes generadas.

4.1. U-Net

La U-Net es una red neuronal convolucional diseñada específicamente para tareas de segmentación de imágenes biomédicas. Fue desarrollada por Olaf Ronneberger, Philipp Fischer y Thomas Brox [43]. El desarrollo de la U-Net surge de la necesidad de superar las limitaciones de métodos anteriores. En la [Figura 4.1](#) se muestra un esquema de la estructura de la U-Net.



[Figura 4.1:](#) Esquema de la arquitectura de la primera U-Net. Esta arquitectura puede variarse según el interés del problema a resolver.

Diseñada con una arquitectura en forma de “U”, esta red está compuesta por un camino de contracción (*Encoder*) y un camino de expansión (*Decoder*). Esta estructura permite una segmentación precisa combinando la información de contexto global y la precisión de localización.

El Encoder sigue la arquitectura típica de una red convolucional, y su objetivo es capturar el contexto de la imagen a través de una serie de operaciones de convolución y pooling. Cada bloque consiste en dos convoluciones de 3×3 píxeles, cada una seguida por una activación ReLU. Estas convoluciones aumentan la profundidad del mapa de características mientras mantienen las dimensiones espaciales de la imagen. Después de cada bloque de convolución, se aplica una operación de *max pooling* 2×2 con un stride de 2. Esta operación reduce las dimensiones espaciales a la mitad, permitiendo

que la red aprenda características a diferentes escalas. Con cada operación de pooling, el número de canales de características se duplica, lo que permite a la red capturar más características complejas a medida que se profundiza en la red. La profundidad típica comienza en 64 canales y puede llegar hasta 1024 canales en las capas más profundas.

El Decoder tiene la tarea de reconstruir la imagen a su tamaño original, combinando la información de características extraída con precisión espacial. Cada bloque comienza con una operación de upsampling (normalmente mediante una convolución transpuesta) que aumenta las dimensiones espaciales al doble. Esta operación es seguida por una convolución 2x2 que reduce el número de canales a la mitad.

La característica clave de la U-Net es la concatenación (skip connections) de las características del Decoder correspondientes. Esto se hace uniendo las características del Encoder a las del Decoder para asegurar que la información detallada de la imagen original se mantenga, mejorando así la precisión de la segmentación. Después de la concatenación, se aplican dos convoluciones 3x3 seguidas por activaciones ReLU. Estas convoluciones refinan las características combinadas, permitiendo que la red aprenda a generar una salida segmentada precisa.

Al final del Decoder, una convolución 1x1 se aplica para reducir el número de canales de características a la cantidad de clases de segmentación deseadas (por ejemplo, 1 canal para una segmentación binaria o múltiples canales para segmentación multi-clase).

La U-Net ha sido probada en diversos desafíos y conjuntos de datos biomédicos, mostrando un rendimiento superior a los métodos anteriores. Por ejemplo, en el desafío de segmentación EM del ISBI 2012, superó significativamente a la red de ventanas deslizantes de Ciresan et al. En términos de error de deformación (warping error) y error Rand, la U-Net demostró ser superior sin necesidad de preprocesamiento o postprocesamiento adicional.

Como extractor de características para el cálculo del FID, se emplea solo el Encoder de la U-Net, ya que la salida del Encoder contiene las características más importantes y complejas de las imágenes.

4.2. ResNet50

ResNet[24], abreviatura de *Residual Network*, es una arquitectura de red neuronal profunda desarrollada para superar los problemas de degradación que se presentan al aumentar la profundidad de las redes neuronales. Este problema se caracteriza por una reducción en la precisión a medida que se agregan más capas a una red, lo cual fue abordado con éxito por Kaiming He y sus colegas mediante la introducción de bloques residuales.

ResNet50 (ver Figura 4.2) es una versión específica de la arquitectura ResNet con 50 capas, diseñada para tareas de reconocimiento de imágenes. La estructura general de

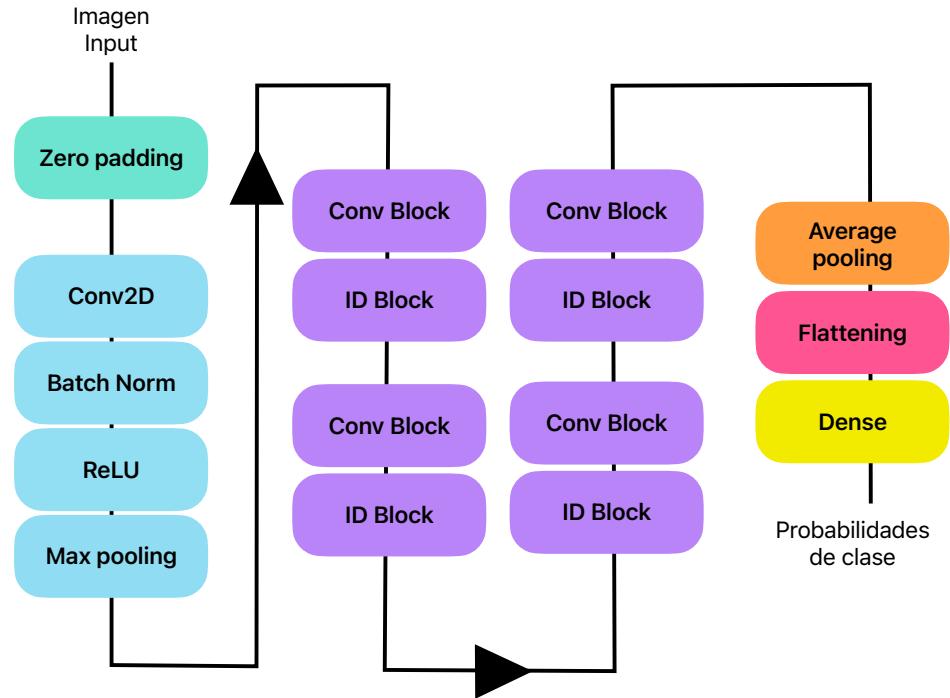


Figura 4.2: Esquema de la arquitectura de la red ResNet50.

ResNet50 se organiza en bloques residuales que contienen capas convolucionales.

La idea central de los bloques residuales es aprender una función residual que se añade a la entrada original mediante una conexión de atajo (shortcut connection). Esto facilita el aprendizaje de la identidad o transformaciones cercanas a la identidad. Cada bloque residual en ResNet50 contiene múltiples capas convolucionales y de normalización, seguidas de una suma con la entrada original y una función de activación ReLU.

La red comienza con una capa convolucional de 7×7 seguida de una capa de agrupación máxima (*max-pooling*) de 3×3 . A continuación, hay cuatro capas convolucionales, formadas respectivamente por: tres bloques residuales con tres capas convolucionales de 1×1 con 64 filtros, 3×3 con 64 filtros y 1×1 con 256; cuatro bloques residuales con tres capas convolucionales de 1×1 con 128 filtros, 3×3 con 128 filtros y 1×1 con 512; seis bloques residuales con capas convolucionales de 1×1 con 256 filtros, 3×3 con 256 filtros y 1×1 con 1024; y tres bloques residuales con capas convolucionales de 1×1 con 512 filtros, 3×3 con 512 filtros y 1×1 con 2048. Finalmente, una capa de agrupación promedio (*average pooling*) seguida de una capa completamente conectada (*fully connected*) para la clasificación [24].

En cuanto al entrenamiento de la red, en lugar de aprender directamente la función de mapeo deseada los bloques residuales aprenden la diferencia de dicha función con la identidad. Esto simplifica la tarea de ajuste de los parámetros y permite un entre-

namiento más eficiente de redes profundas. Además, el uso de conexiones de atajo facilita la propagación de gradientes, mejorando la convergencia durante el entrenamiento mediante técnicas de optimización como el descenso de gradiente estocástico (SGD). El conjunto de datos empleado es *ImageNet*, compuesto por más de 14 millones de imágenes etiquetadas entre 1000 clases.

Durante el proceso de inferencia, una imagen de entrada se procesa a través de las múltiples capas convolucionales y bloques residuales, acumulando las características aprendidas a lo largo de la red. Las capas finales realizan la clasificación basada en estas características acumuladas.

Esta arquitectura se utiliza ampliamente en tareas de visión por computadora, incluyendo clasificación de imágenes, detección de objetos y segmentación. Por lo tanto, la usaremos como extractor de características en el cálculo del FID.

4.3. SkinLesNet

La arquitectura de la red neuronal *SkinLesNet*^[5] fue desarrollada recientemente a finales del año 2023 con el objetivo exclusivo de mejorar el rendimiento de otras redes en el problema de clasificación multiclase de imágenes de lesiones de piel, así como el de tener un modelo de clasificación con una baja complejidad. Esta arquitectura de red fue escogida de entre otras varias posibilidades por ser la que mejor valores obtuvo en las métricas de evaluación de la clasificación. En la Figura 4.3 puede verse un esquema de la SkinLesNet.

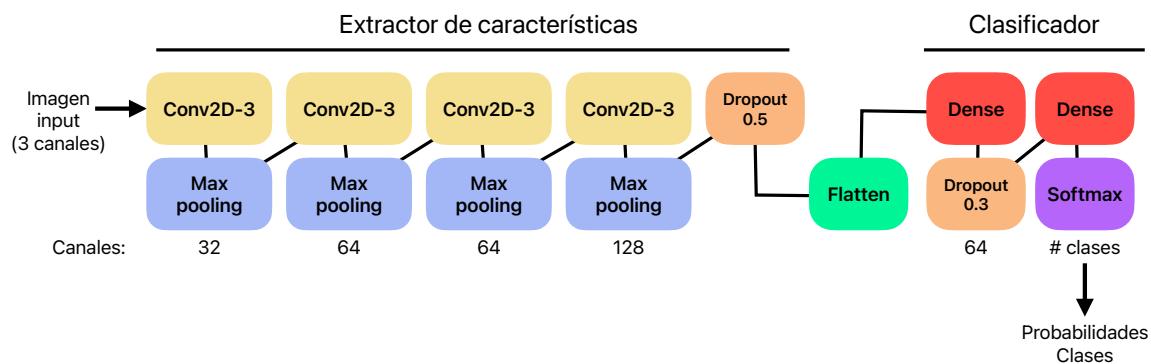


Figura 4.3: Gráfico de la arquitectura de la red neuronal *SkinLesNet*. Se colocan cuatro capas convolucionales bidimensionales con un tamaño de kernel igual a 3, cada una seguida de una capa *max-pooling*, de forma que aumentan los canales de la imagen mientras se reduce la dimensión espacial de la misma. Esta parte compone el extractor de características. Dos capas lineales disminuyen la dimensión para calcular las probabilidades de que el registro pertenezca a cada una de las clases consideradas.

4 Extractores de características

La elección de una red neuronal convolucional (CNN) multicapa se basó en su eficacia probada para aprender características jerárquicas de datos de imágenes complejas. La fortaleza de SkinLesNet radica en su arquitectura de cuatro capas, optimizada sistemáticamente mediante pruebas extensas, que actúan como extractores de características, capturando patrones intrincados en imágenes de lesiones cutáneas[5]. La inclusión de capas de *max-pooling* ayuda en la reducción de dimensiones espaciales mientras retiene características esenciales. La función de activación ReLU añade no linealidad al modelo, permitiéndole reconocer patrones complejos en los datos. De esta manera, se tiene que esta arquitectura es hábil en el procesamiento de diversas características, incluyendo señales de textura, aspectos geométricos y características de color, fundamentales para la clasificación precisa de lesiones cutáneas.

El conjunto de datos empleado para entrenar esta red fue PAD-UFES-20, al cual se aplicaron técnicas de *Data Augmentation* para intentar solventar los problemas derivados del desbalanceo entre las tres clases que conformaban el *dataset*. En cuanto a los resultados, obtuvieron que esta red superaba en todas las métricas, con cierta diferencia, a otros modelos de la literatura re-entrenados con estas imágenes, como ResNet50 o VGG16.

Estos motivos, junto con el uso de imágenes de lesiones cutáneas en nuestro proyecto (ver [Sección 5.2](#)), hacen que escojamos la SkinLesNet como uno de los extractores de características para el cálculo del coeficiente FID en los modelos de difusión generativos que entrenamos en este trabajo.

5 Trabajo experimental y resultados

Consideraciones generales

En esta parte de la disertación se exponen y explican tanto los procedimientos experimentales seguidos - el conjunto de datos empleado, técnicas de preprocesamiento aplicadas o modelos entrenados - como los resultados obtenidos sobre el rendimiento de los modelos.

La parte experimental puede dividirse en dos principales secciones. En primer lugar, como una primera toma de contacto, se entrenó un modelo de difusión incondicional para la generación de imágenes monocromáticas 2D usando el conjunto de datos de imágenes de radiografías de manos de MedNIST. Además, se incluyó el estudio de algunas métricas de calidad de las imágenes generadas con este modelo. En segundo lugar, se detalla el entrenamiento de dos modelos de difusión generativos incondicionales para la generación de imágenes a color 2D de lesiones de piel malignas y benignas respectivamente, usando como conjunto de datos aquellos proporcionados en varias ediciones de la competición ISIC. Para este último caso, se estudiaron también el valor de varias métricas de calidad, y de algunas variaciones de las mismas, para las imágenes generadas por estos modelos.

En todo momento, los experimentos se realizaron en un ordenador portátil MacBook Pro con Chip Apple Silicon M1 Pro y 16Gb de memoria RAM unificada. Todo los scripts de código empleado en esta disertación están escritos en lenguaje Python usando Pytorch y pueden consultarse en el siguiente repositorio de [GitHub](#). La implementación final en código que utilizamos de los DDPMs y de algunas métricas de calidad es la desarrollada por el [Proyecto MONAI](#), que es un conjunto de marcos de trabajo de código abierto y gratuitos, creados para ayudar a la investigación y la colaboración clínica en imagen médica. Este código, basado en el de la comunidad [Huggingface](#), puede encontrarse en su repositorio de [GitHub](#) dedicado a los modelos generativos. Las implementaciones en código de las redes usadas como extractores de características, así como del preprocesamiento de las imágenes, del entrenamiento de los modelos y del cálculo de los valores de las métricas de calidad cuantitativas es propio.

5.1. Generación de imagen monocromática. Imágenes de *Hand* de MedNIST. Métricas de calidad

En imagen médica, y de forma muy genérica, podemos distinguir entre dos tipos de imagen: aquellas imágenes en blanco y negro (monocromáticas) obtenidas a partir de radiografías, mamografías y otras pruebas similares; y las imágenes a color, obtenidas mediante cámaras fotográficas y que muestran regiones anatómicas observables con luz visible. Siguiendo esta clasificación, es por lo que distinguimos dos marcos dentro de la parte experimental, para así estudiar el comportamiento de los modelos de difusión según el número de canales de las imágenes a generar.

En primer lugar, se estudió la calidad de la generación de imágenes médicas monocromáticas a modo de primera aproximación, ya que es menos compleja en términos computacionales. Como conjunto de datos se emplearon las 8000 imágenes de radiografías de manos de MedNIST[53, 54], todas ellas imágenes en blanco y negro con una resolución de 64×64 píxeles (Figura 5.1).

Para aumentar la diversidad de los datos e intentar conseguir un modelo robusto que aprendiera mejor las características de las mismas, se llevó a cabo un enfoque de *Data Augmentation*, en el que, de forma aleatoria, se invertían horizontalmente, verticalmente y también se intercambiaban las zonas blancas y negras entre sí. No obstante, en las primeras pruebas no se aplicaron estas transformaciones, sino que vinieron impulsadas por los resultados de las distintas pruebas realizadas para la generación de imagen médica a color (Sección 5.2). En cambio, todas las imágenes siempre estuvieron normalizadas al intervalo $[0, 1]$, y se les hizo un recorte central (*CenterCrop*) de 45 píxeles y un reescalado (*Resize*) a 32 píxeles, para así poder eliminar fondo uniforme de las imágenes que no aporta información alguna al modelo, y reducir el tamaño de las imágenes a fin de poder entrenar el modelo sin saturar la memoria de la GPU disponible. Un ejemplo de cómo se ven las imágenes de entrenamiento tras estas transformaciones se muestra en la Figura 5.2.

Una vez completado todo el preprocesamiento de las imágenes, el conjunto se dividió en un 90 % para imágenes de entrenamiento y el 10 % restante para validación durante el aprendizaje del modelo. El conjunto de validación lo emplearemos para observar la evolución de la función de pérdida sobre estas imágenes durante el entrenamiento del modelo. Se cargaron ambos *sets* en *DataLoaders* con tamaño de *batch* igual a 32, y se entrenó el modelo durante 70 épocas, midiendo la pérdida para las imágenes de validación cada cinco épocas. También se puso un enfoque de reducción lineal por época de la tasa de aprendizaje con un valor inicial de 0.001 y final de 0.0002, con el fin de conseguir un buen ajuste del valor de los parámetros de la red U-Net asociada al modelo, para lo cual se empleó el optimizador Adam.

Dado que esta parte del experimento se consideró como un “calentamiento” previo al trabajo computacionalmente más costoso e interesante, y a los buenos resultados cualitativos obtenidos, se probó con una única arquitectura para la red neuronal U-Net

5.1 Generación de imagen monocromática. Imágenes de Hand de MedNIST. Métricas de calidad



Figura 5.1: Subconjunto de ocho imágenes de la clase *Hand* del *dataset* MedNIST. Se observa la cantidad de fondo negro innecesario para entrenar el modelo de difusión, por lo que se opta por eliminarlo con el recorte central.

ligada al modelo de difusión. Esta arquitectura ([Figura 2.2](#)) consta de tres bloques en el *Encoder* y el *Decoder* con número de canales 128, 256 y 256, añadiendo una capa de atención en los dos últimos. Se usó esta disposición basándonos en la escogida en varios ejemplos de DDPM del Proyecto MONAI. Además, el número de pasos temporales en el proceso de introducción y eliminación de ruido fue de 1000 en ambos casos. Así pues, se procedió al entrenamiento del modelo.

Las gráficas con la evolución de la función de pérdida y de la tasa de aprendizaje durante el entrenamiento se observa en la [Figura 5.3](#).

Durante el proceso de entrenamiento, cada cinco épocas, se generaba una imagen a partir de ruido aleatorio para observar la evolución del aprendizaje de la red de forma cualitativa. Algunas de estas imágenes se muestran, ordenadas según la época en que se generaron, en la [Figura 5.4](#). La primera imagen es un fondo completamente negro, sobre el cual, en los sucesivos *samples*, van apareciendo formas de manos radiografiadas en escalas de grises. En la segunda imagen no se muestra anatómicamente precisa la mano, mientras que, a partir de la tercera, es visualmente claro que aparece una mano. Hay imágenes, como la cuarta, en la que falta un dedo de la extremidad, que puede deberse a que, bien haya casos en el conjunto de entrenamiento de malformaciones o amputaciones que expliquen estos ejemplos, o bien que se deba a la falta de un mayor entrenamiento de la red. Añadir que en la [Figura 5.5](#) se muestra el proceso de generación de una imagen de radiografía de mano mediante la eliminación de ruido por la red que hemos entrenado.

A pesar de haber escogido 70 épocas, la evolución de la curva de la pérdida para

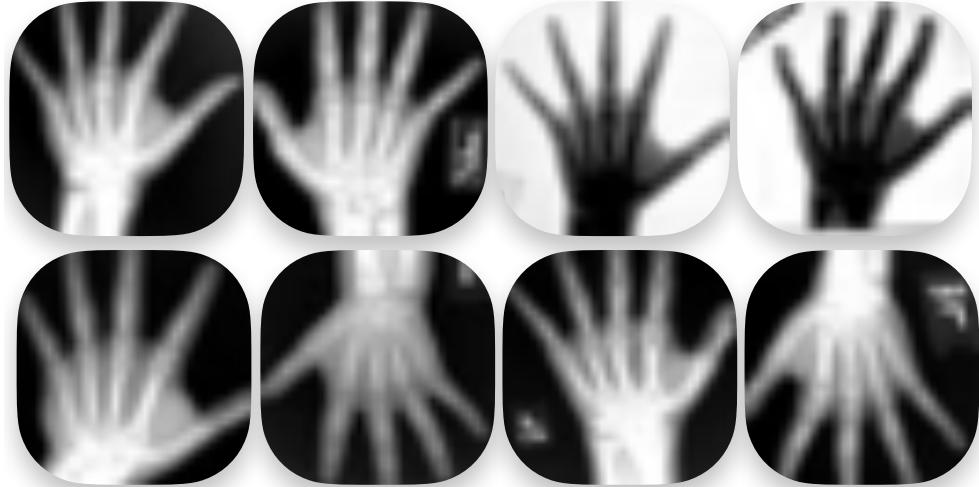


Figura 5.2: Subconjunto de ocho imágenes de la clase *Hand* del *dataset* MedNIST una vez han sufrido las transformaciones descritas. Se observa una mayor atención en los detalles anatómicos de la mano, a pesar del recorte y del rescalado, y una mayor diversidad en el conjunto de imágenes que posibilitará un mejor aprendizaje del modelo de difusión.

validación nos indica que podríamos haber aumentado el número de épocas en el entrenamiento, ya que esta curva parece no haber encontrado un punto de equilibrio en el que no sigue descendiendo. No obstante, como las imágenes que generaba el modelo eran cualitativamente buenas al final de entrenamiento, se optó por no continuarlo.

Para evaluar de forma cuantitativa la calidad de las imágenes generadas por el modelo DDPM una vez terminado el entrenamiento de éste, generamos un total de 100 imágenes sintéticas para las que calculamos los coeficientes FID, SSIM y MS-SSIM. A pesar de que se aconseja usar un total de 50k imágenes para calcular correctamente estos coeficientes, no se generó toda esa cantidad por motivos de falta de memoria principal para almacenarlas. Algunas de las imágenes generadas pueden verse en la [Figura 5.6](#), para las que podemos observar una gran diversidad de *samples*, tanto en relación a la proporción blanco/negro, como en la disposición de la mano.

En el cálculo del FID se empleó como extractor de características de las imágenes el extractor de una red ResNet50 previamente entrenada usando el conjunto de imágenes [Rad Image Net](#)^[36], la cual se descargó del siguiente repositorio de [GitHub](#).

El *dataset* de RadImageNet está conformado por imágenes radiológicas de distintas áreas anatómicas del cuerpo, por lo que la parte extractora de esta red en concreto, será capaz de obtener las particularidades que definen las radiografías de manos humanas. Como conjunto de imágenes reales se usó la partición de validación, ya que en el proceso de entrenamiento del modelo, estos datos no intervieron. No obstante, es una

5.1 Generación de imagen monocromática. Imágenes de Hand de MedNIST. Métricas de calidad

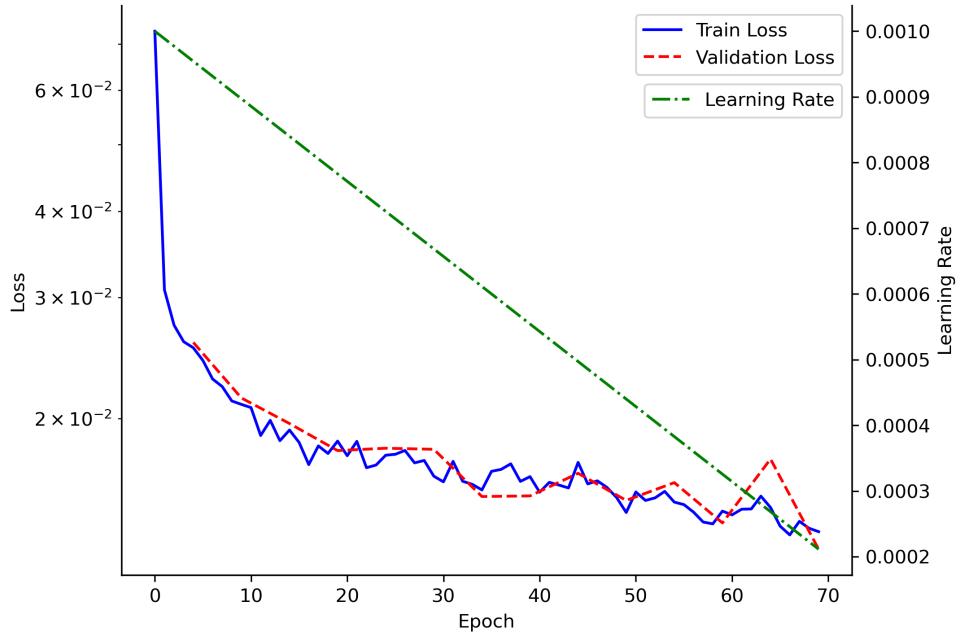


Figura 5.3: Evolución de la pérdida *loss* para el conjunto de entrenamiento (azul) y para el de validación (rojo) y de la tasa de aprendizaje (*learning rate* (verde) durante el entrenamiento de la red U-Net del modelo de difusión.

práctica común el uso del conjunto de entrenamiento para tal fin[26]. El valor del FID obtenido es de 5.9603, el cual es un muy buen valor, si se compara con los valores que aparecen en la literatura[6], lo cual indica que, en el caso de que las distribuciones de los datos fuera una Normal, ambas distribuciones serían muy similares.

En el caso del SSIM y del MS-SSIM, los valores obtenidos son, respectivamente, 0.1654 ± 0.1078 y 0.1784 ± 0.2409 , los cuales no son tan buenos como cabría esperar. Dada la naturaleza de este coeficiente, se consideraron únicamente las imágenes generadas con fondo negro para el cálculo del mismo, mejorando la cifra en 0.2049 ± 0.0836 y 0.2345 ± 0.2510 , no siendo aún suficientemente bueno. Este problema puede deberse a que el modelo genera las imágenes con distinta disposición y nivel de detalle de las manos, haciendo que estructuralmente las imágenes difieran y se obtengan un mal valor del coeficiente de similitud estructural.

Por tanto, mientras que el modelo DDPM demuestra una notable capacidad para generar imágenes radiológicas sintéticas en blanco y negro con una distribución de características muy similar a las imágenes reales, existe una discrepancia significativa en la similitud estructural. La diversidad en términos de luminancia y disposición de las manos generadas, aunque muestra la flexibilidad del modelo, también puede ser la causa de los bajos valores de SSIM y MS-SSIM. Estas observaciones indican que, aunque el modelo es efectivo en general, se podrían realizar ajustes adicionales para mejorar la consistencia estructural de las imágenes generadas.



Figura 5.4: Algunas imágenes generadas durante el entrenamiento del modelo de difusión DDPM para observar la mejora cualitativa de las imágenes sintetizadas a lo largo de dicho proceso. Las imágenes están ordenadas según la época de su generación.



Figura 5.5: Muestra gráfica del proceso de eliminación de ruido (*denoising*) para la generación de imágenes.

5.2. Generación de imagen a color. Imágenes de lesiones de piel de ISIC

Una vez abordado un caso de generación de imagen médica en blanco y negro, pasamos a evaluar la calidad de los modelos de difusión en la generación de imagen a color. Para este caso, se ha optado por centrar la atención en el estudio de las imágenes de lesiones cutáneas relacionadas con el cáncer de piel.

Las enfermedades crónicas de la piel, entre las que están el melanoma (un tipo de cáncer maligno de este órgano), están tomando cada vez mayor relevancia como un problema serio de salud pública debido al aumento progresivo de casos diagnosticados[18]. Este aumento de la incidencia de este tipo de enfermedades se atribuyen a la mayor exposición a la radiación ultravioleta debido al cambio en el estilo de vida de la población. Además, la mayor proporción de población envejecida, predisposiciones genéticas y la no toma de medidas de precaución para prevenir estos trastornos no hacen más que exacerbar el problema. La detección temprana y una prevención activa son esenciales para mitigar la incidencia y mortalidad del cáncer de piel.

5.2 Generación de imagen a color. Imágenes de lesiones de piel de ISIC

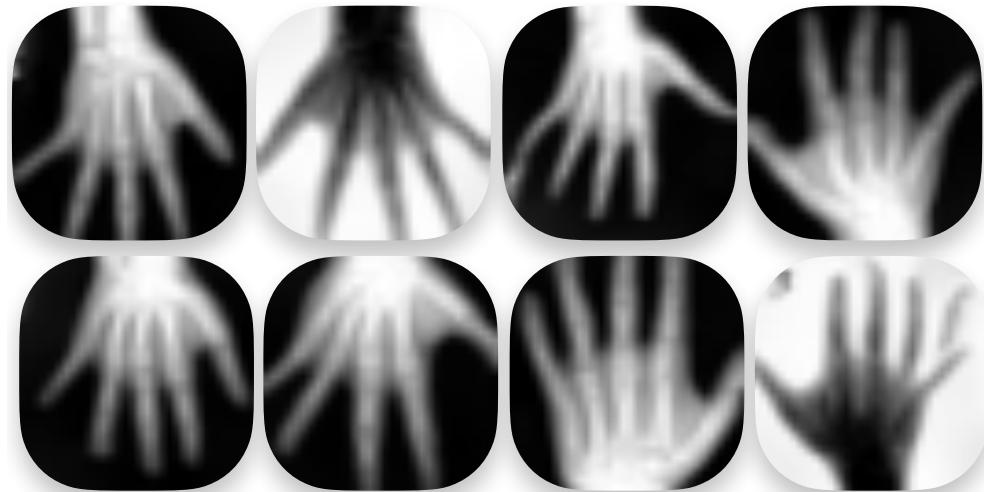


Figura 5.6: Ejemplos de imágenes generadas por el modelo de difusión una vez finalizado el entrenamiento del mismo.

Como lesiones pigmentadas que ocurren en la superficie de la piel, el melanoma es susceptible de detección temprana mediante inspección visual experta. También es susceptible de detección automatizada mediante análisis de imágenes. Dada la amplia disponibilidad de cámaras de alta resolución, los algoritmos que pueden mejorar nuestra capacidad para cribar y detectar lesiones problemáticas pueden ser de gran valor. Por tanto, en la detección temprana de este tipo de enfermedades es donde mayor uso de los algoritmos y las técnicas de *Machine Learning* se está produciendo, para así crear sistemas de diagnóstico por ordenador basados en IA, que puedan ser usados por doctores e investigadores para el diagnóstico y el estudio de varias enfermedades de la piel. Sin embargo, uno de los mayores escollos en el desarrollo de estos sistemas reside en la falta de conjuntos de datos voluminosos y ricos en variedad ([Sección 1.2](#)). La adquisición de esos conjuntos suele ser difícil, por lo que una vía para solventar esta limitación es la de intentar generar imágenes realistas de lesiones cutáneas, que puedan ser utilizadas. Es por este motivo, por el que nos centramos en estudiar si los modelos de difusión son válidos candidatos a generadores de este tipo de imágenes.

5.2.1. Datasets y preprocesamiento

Desde 2016, la Colaboración Internacional de Imágenes de la Piel (ISIC, por sus siglas en inglés), ha organizado desafíos anuales para la comunidad de ciencias de la computación en colaboración con conferencias de visión por computadora, creciendo en escala, complejidad y participación. Utilizan conjuntos de datos de imágenes y metadatos validados por humanos. Inicialmente, los desafíos se centraron en la precisión diagnóstica del melanoma, y para 2018, los algoritmos superaban consistentemente a

5 Trabajo experimental y resultados

los médicos. Los desafíos de 2019 y 2020 abordaron problemas de distribución fuera del conjunto de datos y el impacto del contexto clínico, respectivamente. En 2020, participaron 3,300 personas de todo el mundo. Además, ISIC organiza desafíos en vivo para evaluar algoritmos de manera continua.

ISIC, patrocinada por la Sociedad Internacional para la Imagen Digital de la Piel (ISDIS), busca mejorar el diagnóstico del melanoma. El Archivo ISIC contiene más de 13,000 imágenes dermoscópicas de calidad controlada, recolectadas de centros clínicos internacionales. Las imágenes se revisan para garantizar privacidad y calidad, y la mayoría incluye metadatos clínicos verificados por expertos en melanoma. Un subconjunto de estas imágenes ha sido anotado por expertos en cáncer de piel, incluyendo características dermoscópicas relevantes para el diagnóstico.

En este trabajo usamos varios de los conjuntos de datos ofrecidos en las ediciones de los desafíos organizados por ISIC de 2016, 2018, 2019 y 2020, según el tipo de datos que sean necesarios. Por ejemplo, para el entrenamiento de los modelos de difusión incondicionales, usamos las imágenes de entrenamiento de las ediciones de 2018[12, 49], 2019[49, 14, 13] y 2020[44]. Dado que las resoluciones y tamaños de las imágenes son variables entre ediciones, se emplea el mismo conjunto de imágenes pero ya preprocesado (con un *centre cropping* y un *resize* a 224 píxeles), disponible en [Kaggle](#). Mencionar que el conjunto de datos de ISIC2016 ([Figura 5.9](#)) está conformado por dos subconjuntos, uno de entrenamiento y otro de test, en el que cada imagen esta acompañada por su respectiva máscara de segmentación de la lesión; por lo que el conjunto de validación empleado en el entrenamiento es una pequeña partición de éste. Sin embargo, las imágenes descargadas de ISIC 2018, 2019 y 2020 ([Figuras 5.7](#) y [5.8](#)) no están divididas en conjunto de entrenamiento, validación y test, por lo que los dos últimos los tomaremos como divisiones extraídas del primero.

5.2.2. En busca de la mejor estrategia

Aunque en un principio, debido a las limitaciones computacionales, se quiso trabajar únicamente con los datos de la edición de 2016 e intentar usar técnicas para suplir la falta de imágenes de la clase maligna, esa misma carencia nos hizo recurrir a los *datasets* de otras ediciones más recientes. La idea de emplear únicamente ISIC 2016 vino motivada por los errores de falta de memoria durante el entrenamiento del modelo DDPM usando el *dataset* de 2020. Dado que no queríamos renunciar a poder generar imágenes con buena resolución (128 o 224 píxeles) para así justificar la propiedad de generación de imágenes de calidad de los modelos de difusión, se recurrió a reducir el número de imágenes de entrenamiento empleadas tomando ISIC 2016 como conjunto de entrenamiento.

La edición de 2016 ofrecía únicamente 171 imágenes de lesiones malignas de un total de 899 casos, un número insuficiente para entrenar correctamente un modelo de difusión. Por ello, se recurrió, probando primero con resoluciones de 64 píxeles de las imágenes, a técnicas de *Data Augmentation*, como rotaciones, reflexiones, zoom y

5.2 Generación de imagen a color. Imágenes de lesiones de piel de ISIC



Figura 5.7: Muestra del conjunto de imágenes de entrenamiento de la clase *malignant* obtenidos de ISIC 2018, 2019 y 2020.

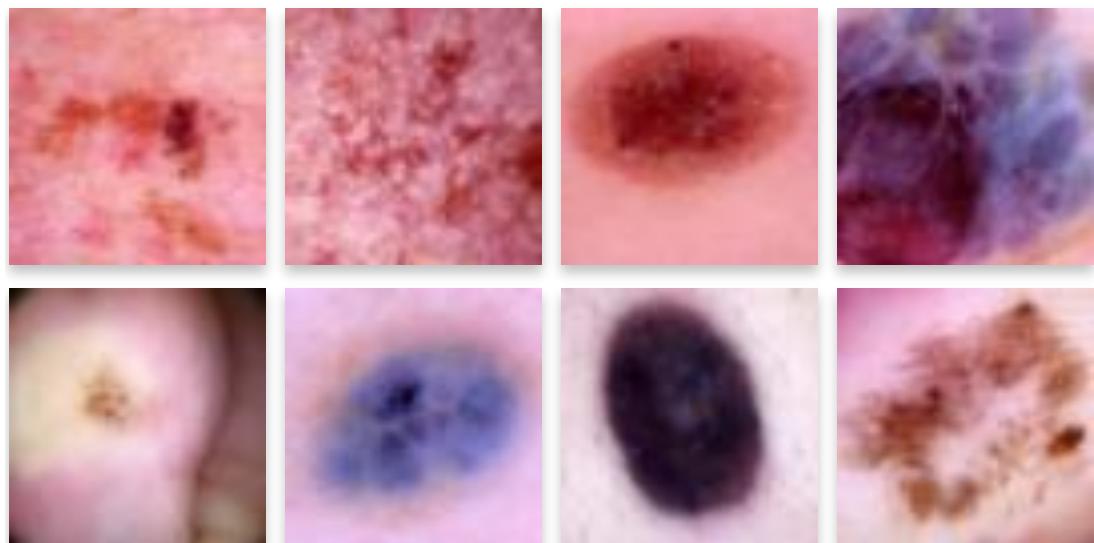


Figura 5.8: Muestra del conjunto de imágenes de entrenamiento de la clase *benign* obtenidos de ISIC 2018, 2019 y 2020.

5 Trabajo experimental y resultados

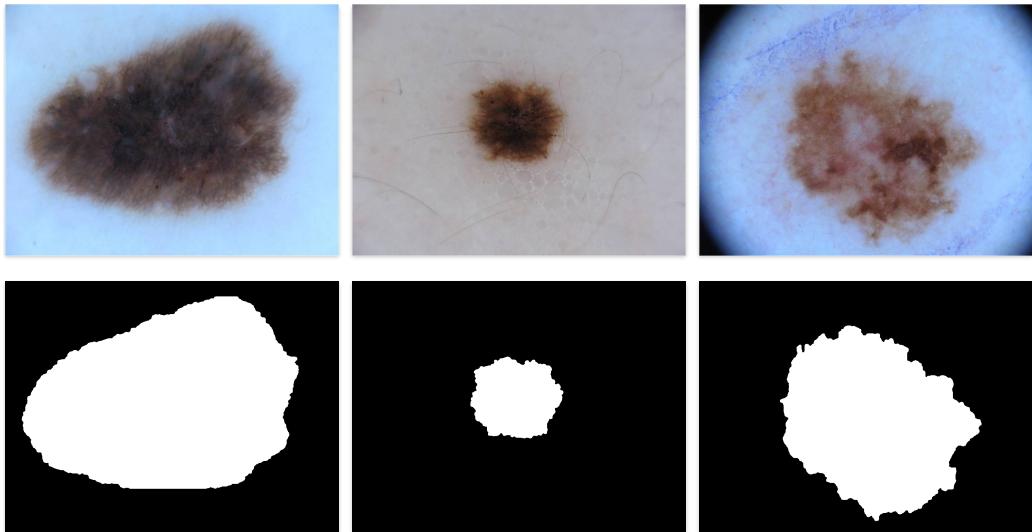


Figura 5.9: Muestra del conjunto de imágenes de entrenamiento para la segmentación de las lesiones obtenidas de ISIC 2016.

cambios en el brillo, contraste y un poco en tonalidad, para aumentar el número de casos de entrenamiento y conseguir un modelo generativo más robusto y de calidad ([Figura 5.10](#)). Sin embargo, las imágenes generadas por los distintos modelos entrenados con este *dataset* no tenían una calidad convincente ([Figura 5.11](#)). Esto, sumado a que, de por sí mismo el conjunto de datos de entrenamiento en otras ediciones es lo suficientemente grande como para mostrar variedad, y que *Data Augmentation* mostraba imágenes muy similares a las originales, haciendo que se pudiera llegar a sobreajuste, hizo descartar este enfoque.



Figura 5.10: Aplicación de las técnicas de *Data Augmentation* sobre el conjunto de imágenes malignas de ISIC 2016 para aumentar el número de ejemplos de entrenamiento.

Volver a usar conjunto de imágenes más grandes sin renunciar a la alta calidad de las mismas nos devolvía al problema de la falta de memoria. La estrategia que se buscó para intentar evitar esta limitación fue usar la carga diferida o *lazy loading*, la cual en lugar de cargar todos los recursos de una vez al inicio, se cargan en demanda, es decir,

5.2 Generación de imagen a color. Imágenes de lesiones de piel de ISIC

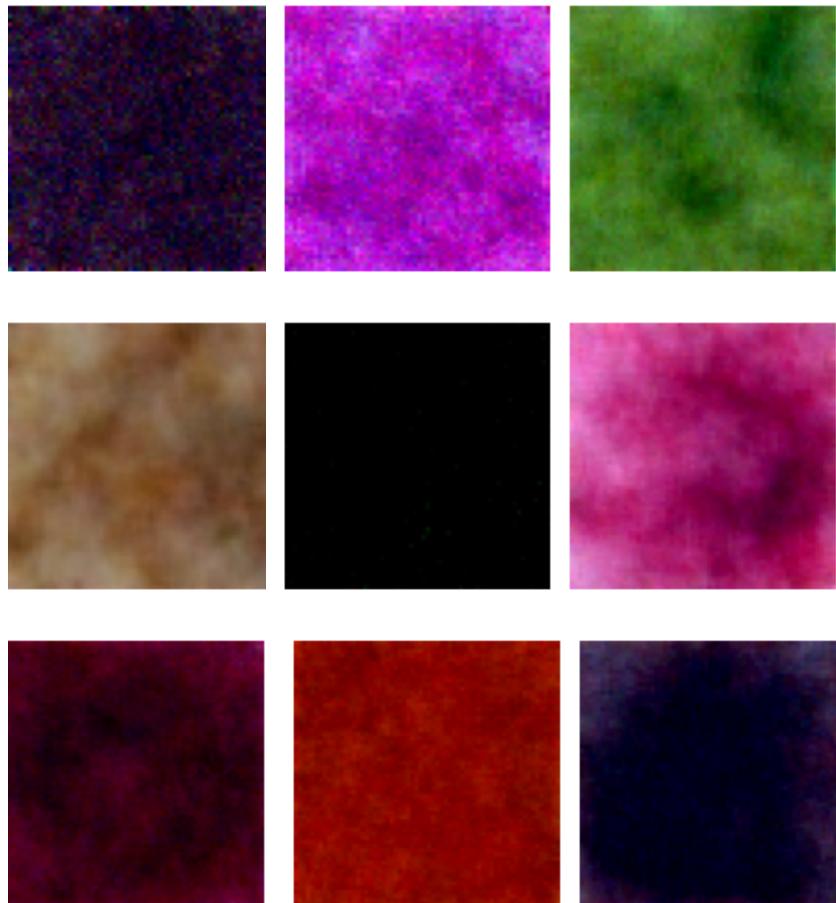


Figura 5.11: Ejemplos sampleados durante el entrenamiento del correspondiente modelos de difusión generativo con dichas imágenes de entrenamiento.

solo cuando el usuario los necesita o los solicita. Esta técnica reduce el tiempo de carga inicial y el consumo de recursos, especialmente en dispositivos con limitaciones. De esta manera, se tiene en memoria únicamente aquellas imágenes pertenecientes al batch empleado por el optimizador en cada iteración del entrenamiento. No obstante, a pesar de investigar en profundidad el motivo, esta carga diferida no funcionaban correctamente, ya que volvía a llenarse la memoria del sistema en pocas iteraciones del entrenamiento. Incluso tomando tamaños de batch pequeños, seguía colapsando la memoria.

Así pues, se optó finalmente por reducir el tamaño de las imágenes a 64 píxeles, y a usar tamaños de *batch* 4 imágenes con el fin de poder usar todas las imágenes de la clase maligna a nuestra disposición. Para evitar los inconvenientes de un tamaño de *batch* demasiado pequeño, como puede ser el riesgo de *overfitting*, la baja precisión en la estimación de los gradientes o la alta variabilidad de los pesos durante el entrenamiento, se usó la técnica de acumulación de gradientes, consistente en actualizar los

5 Trabajo experimental y resultados

pesos del modelo utilizando los gradientes acumulados y reiniciar estos gradientes a cero, repitiendo el proceso después de procesar suficientes mini-*batches* para alcanzar el tamaño de *batch* deseado. Esto permite entrenar eficientemente modelos grandes en hardware limitado, mejorando la estabilidad y precisión del entrenamiento sin requerir más memoria que la necesaria. Por tanto, usando un tamaño de 8 para la acumulación de gradientes, se tiene un tamaño efectivo del *batch* de 32.

La reducción del tamaño de la imagen por interpolación bilineal (método por defecto con el método *Resize* de Pytorch), finalmente, no parece destruir la información importante de la imagen, y que permite distinguirla correctamente en las distintas clases, como muestran los resultados del entrenamiento de clasificadores con estas imágenes procesadas ([Sección 5.3](#)).

5.2.3. Modelos de difusión generativos incondicionales

Dado que los *datasets* de ISIC hacen la distinción entre imágenes de lesiones benignas y malignas, nosotros usaremos esta diferenciación para entrenar dos modelos de difusión DDPM incondicionales para la generación de imágenes de estos dos tipos de daños cutáneos. Es importante reseñar que en algunas ediciones de la competición, la clase benigna incluye tanto imágenes de dicho tipo como imágenes catalogadas de tipo “desconocido”, lo cual puede influir negativamente en el entrenamiento del modelo de imágenes benignas, ya que pueden estar mezclados ambas clases de lesiones.

También es importante notar que en estos *datasets*, la clase *malignant* está muy desbalanceada respecto a la clase mayoritaria *benign*. Para el conjunto de ISIC 2018, 2019 y 2020, hay 67439 imágenes, de las cuales 5768 pertenecen a la clase minoritaria (menos del 9 % del total). Es por ello que, desde este momento y en lo sucesivo, para el entrenamiento de los modelos generativos y de clasificación, tomaremos todas las imágenes de la clase minoritaria, y tantas de la mayoritaria como para que el número de casos de ambas clases esté equilibrado (*undersampling*). Sin embargo, este desbalanceo no era de importancia para el conjunto de ISIC 2016, ya que su uso final fue entrenar a la red neuronal de segmentación, para lo cual no era necesario conocer el tipo de lesión.

Para cada uno de los modelos incondicionales, se entrenó una red U-Net, ϵ_θ , encargada de determinar el ruido introducido a una imagen, ϵ_t , conociendo la imagen alterada, x_t y el valor de t en el proceso *forward* ([Párrafo 2.1.1.3](#)). La arquitectura escogida, de entre varias que se probaron, para la red fue de 4 bloques, de 64, 128, 256 y 256 canales respectivamente. El proceso de inyección de ruido se ajustó a $T = 1000$ pasos con valores de $\beta_1 = 10^{-4}$ y $\beta_T = 0.02$, y el tamaño de las imágenes de entrenamiento se reescaló a 64 píxeles por los motivos mencionados anteriormente, junto con un tamaño de *batch* de 4 imágenes y un tamaño de acumulación de gradientes de 8 *batches*.

En consideración con la tasa de aprendizaje, se optó por emplear un enfoque de *Reduce On Plateau*, de tal manera que ésta redujese su valor, acorde a un factor, según si la función objetivo no mejoraba en una época. El valor inicial escogido para la tasa de

5.2 Generación de imagen a color. Imágenes de lesiones de piel de ISIC

aprendizaje fue de 0.001.

Finalmente, para el modelo de generación de lesiones malignas, se emplearon las 5768 imágenes disponibles de dicha clase con un entrenamiento de 200 épocas de duración; mientras que para el generador de lesiones benignas fueron usadas 6834 imágenes respectivas con 100 épocas para el *training*. La diferencia en el número de épocas escogido no responde a ningún motivo, sino al intento de que el modelo aprendiese lo mejor posible las características de las imágenes de tipo maligno. Con todos los parámetros establecidos, se entrenó el modelo.

En las Figuras 5.12 y 5.13 se muestran la evolución de la función de pérdida y de la tasa de aprendizaje durante el entrenamiento de ambos modelos. Se puede observar que hay un cierto sobreaprendizaje de las imágenes de entrenamiento, ya que la pérdida para el conjunto de validación al final está por encima en un orden de magnitud de la pérdida para el conjunto de entrenamiento. No obstante, hubiera bastado con 100 épocas de entrenamiento, o incluso menos, ya que la curva de pérdida para el conjunto de validación en ambos casos parece oscilar en torno a un valor medio desde muy temprano, no mejorando el modelo en las sucesivas iteraciones.

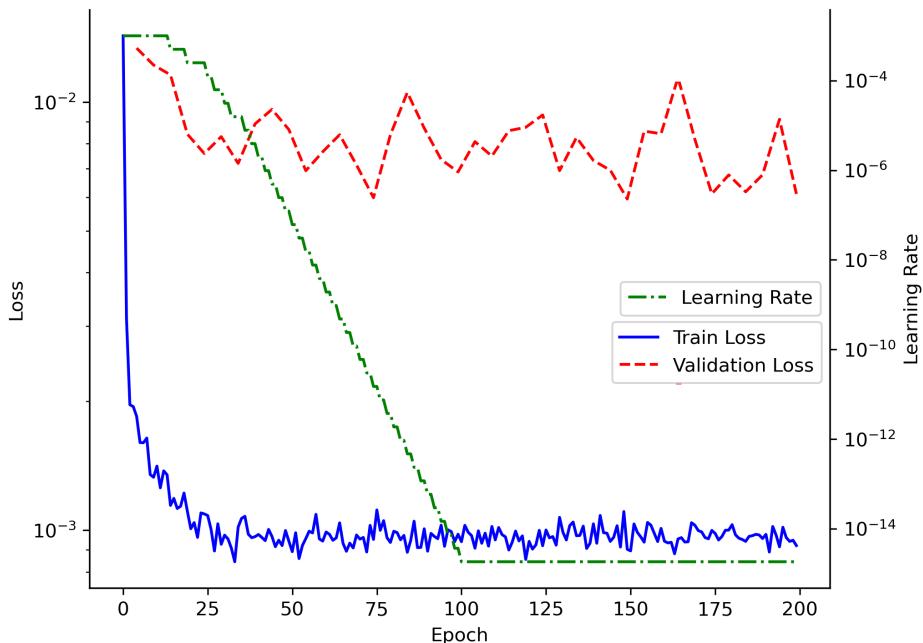


Figura 5.12: Evolución de las funciones de pérdida para el conjunto de imágenes de entrenamiento (azul) y de validación (rojo) y de la tasa de aprendizaje (verde), durante el entrenamiento del DDPM incondicional de las imágenes de lesiones malignas.

Las Figuras 5.14 y 5.15 representan algunos *samplings* realizados durante el proceso de entrenamiento de las redes. Un análisis cualitativo de la calidad de las imágenes

5 Trabajo experimental y resultados

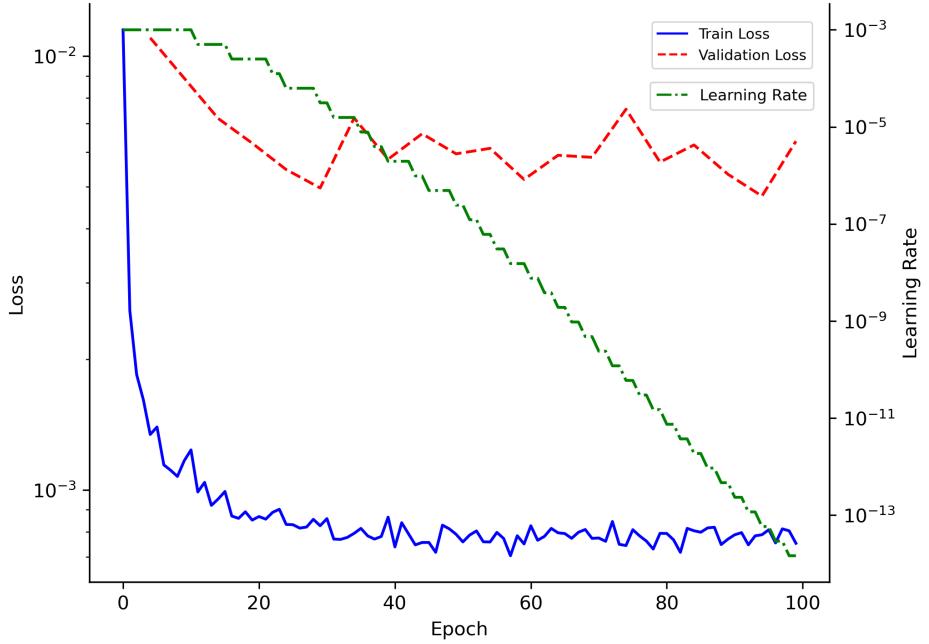


Figura 5.13: Evolución de las funciones de pérdida para el conjunto de imágenes de entrenamiento (azul) y de validación (rojo) y de la tasa de aprendizaje (verde), durante el entrenamiento del DDPM incondicional de las imágenes de lesiones benignas o desconocidas.

generadas con los modelos una vez entrenados se lleva a cabo en la [Sección 5.3](#). Destacar que durante el entrenamiento del modelo generador de imágenes de lesiones malignas, se observaba que el modelo generaba imágenes más coherentes y de mayor calidad conforme avanzaba el entrenamiento; mientras que para el otro modelo de lesiones benignas, los *samplings* realizados eran poco coherentes, de menor calidad y, a veces, incluso erráticos, como es el caso de las imágenes de color negro uniforme o aquellas con una tonalidad muy azulada, las cuales no aparecen en el conjunto de imágenes de entrenamiento. Puede ser que en esas imágenes “incorrectas” haya información que el modelo haya extraído de los datos reales y que no seamos capaces de entender; aunque de forma cualitativa estas imágenes no cumplen los mínimos de calidad y coherencia esperados.

5.3. Cálculo de las métricas de calidad para las imágenes a color

Una vez entrenados los DDPMs generativos incondicionales, pasamos a evaluar la calidad y diversidad de las imágenes sampleadas con dichos modelos. Para ello, nos basamos en los valores cuantitativos para tres de las métricas explicadas anteriormente: *Fréchet Inception Distance* (FID), Índice de similitud estructural (SSIM) y *Classification*

5.3 Cálculo de las métricas de calidad para las imágenes a color

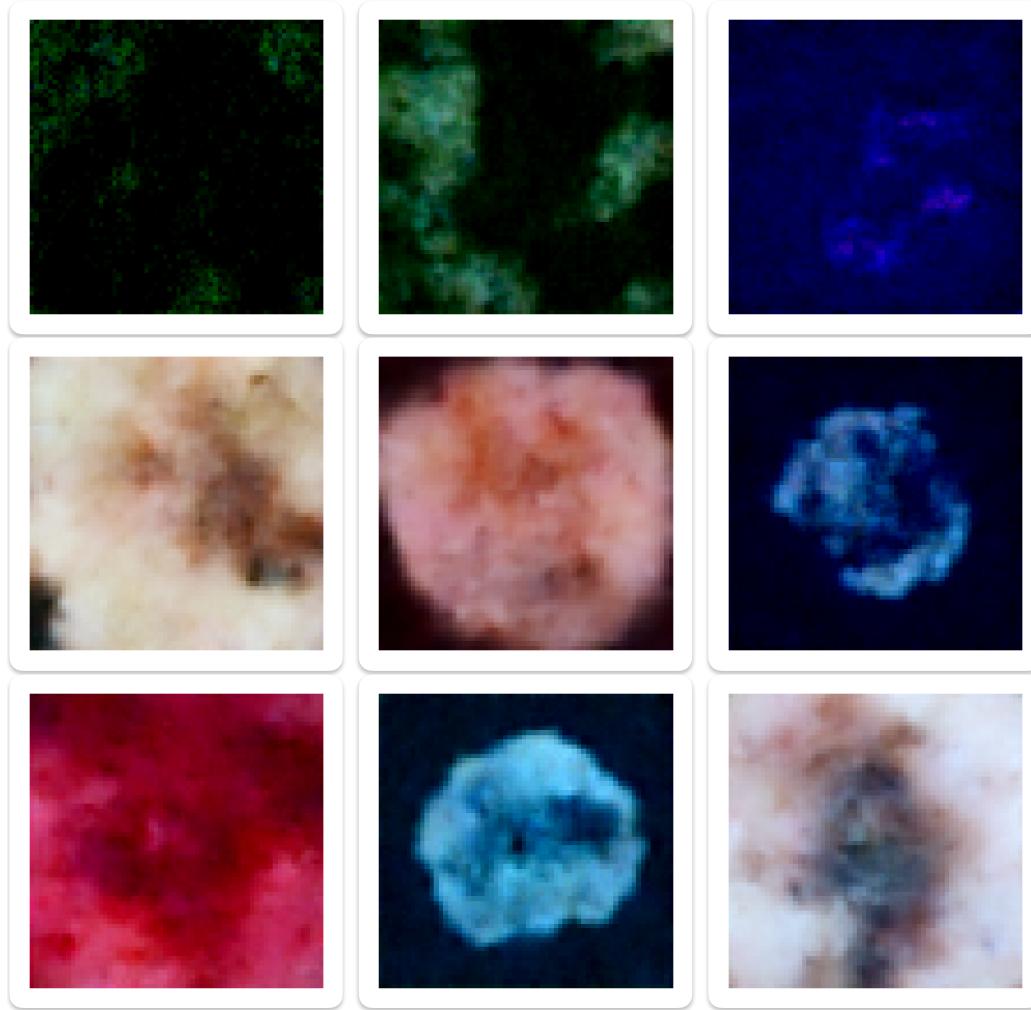


Figura 5.14: Imágenes generadas durante el entrenamiento del modelo de lesiones malignas dispuestas en orden cronológico de generación.

Accuracy Score (CAS). No se han considerado el *Inception Score*, dado que FID lo engloba en parte, y la Proporción máxima de señal a ruido (PSNR), dado que en el ámbito de la generación de imagen esta métrica no tiene demasiado sentido porque usa el error cuadrático medio entre la imagen generada y otra de referencia para calcular su valor.

Además, se llevará a cabo un breve estudio visual cualitativo de las imágenes para ambas clases, teniendo en cuenta que no somos profesionales médicos ni expertos en la materia, por lo que habrá limitaciones sobre lo que podamos comentar de las imágenes.

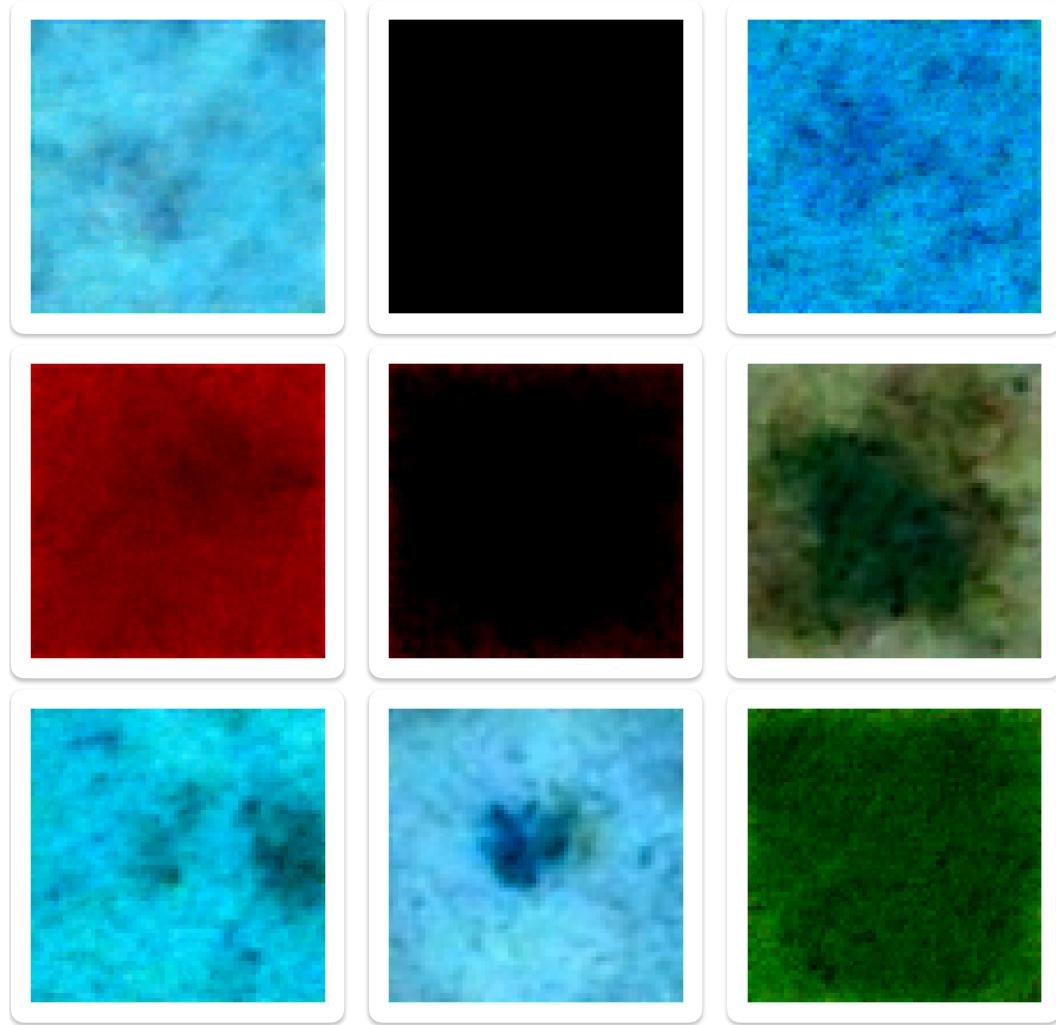


Figura 5.15: Imágenes generadas durante el entrenamiento del modelo de lesiones benignas o desconocidas dispuestas en orden cronológico de generación.

5.3.1. Entrenamiento de los extractores de características.

Para el cálculo del Fréchet Inception Distance (FID), es necesario el uso de unos extractores de características, de los cuales obtener una serie de propiedades concretas de las imágenes de melanoma de piel generadas, para poder compararlas con las características de las imágenes tomadas como reales, y así poder cuantificar el parecido entre ambas. De este modo, para el FID se tienen tantas variantes como extractores de características distintos se utilicen, teniendo cada uno sus ventajas e inconvenientes según las propiedades en las que se enfoque cada extractor.

En este estudio se han empleado tres extractores de características, extraídos de los tipos de redes explicados en la [Capítulo 4](#) y entrenados con las imágenes de ISIC. Estas

5.3 Cálculo de las métricas de calidad para las imágenes a color

redes son U-Net, entrenada para la segmentación de imágenes de lesiones de piel con los datos de entrenamiento de ISIC 2016[23], y ResNet50 y SkinLesNet, entrenadas para la clasificación binaria con las imágenes de ISIC 2018, 2019 y 2020, previamente mencionadas, distinguiendo entre dos clases: lesión benigna o maligna.

En el entrenamiento de la ResNet50, se alteró su arquitectura para adaptarla al problema de clasificación binaria en cuestión, de modo que se cambió la parte encargada de la clasificación por una serie de capas *fully connected* lineales con tamaños 1024, 512, 256, 128, 64 y 2. Además, solo para dicha red se llevó a cabo un enfoque de *transfer learning*, aprovechando los parámetros ya conocidos del modelo original, y solo calculando aquellos nuevos introducidos en la variante. Es por ello, que el extractor de características de la red ResNet50 será la red completa salvo la última capa, para así añadir los conocimientos nuevos de la red en el extractor.

De esta manera, tenemos tres extractores de características, cada uno de ellos con un modo de operar distinto: U-Net, la cual se entrena para realizar segmentación, se centrará más en propiedades de formas y contornos de las lesiones; mientras que SkinLesNet y ResNet50 tendrá en cuenta otras propiedades, como pueden ser el color y la textura.

Mencionar que, hasta que no se encontró el preprocesamiento y tratamiento de las imágenes para el entrenamiento de los modelos de difusión adecuados a los recursos disponibles, no se entrenó ninguna de estas redes neuronales, ya que era necesario emplear imágenes procesadas de la misma forma con el fin de que el extractor de características, *a posteriori*, funcionase correctamente sobre las imágenes generadas por el modelo.

De forma común a las tres redes, se empleó el optimizador Adam para ajustar el valor de los parámetros y un enfoque de *Reduce On Plateau* en la tasa de aprendizaje de nuevo. En cuanto a la mencionada función de pérdida, para la U-Net se empleó el error cuadrático medio (MSE) entre la máscara real y la predicha por la red; mientras que para las otras dos redes se empleó la entropía cruzada en las distribuciones de las predicciones. El número de épocas fue distinto en cada caso, no superando el centenar en ninguno; aunque el valor de las métricas para el conjunto de validación se calculaba cada cinco épocas para las tres arquitecturas.

En las Figuras 5.18, 5.17 y 5.16 se muestra la evolución de las distintas métricas consideradas apropiadas durante el entrenamiento de las redes, obteniendo valores finales para la función de pérdida y métricas consideradas en el entrenamiento buenos en los tres casos. Dado que la pérdida para el conjunto de validación parece alcanzar un valor estable en el entrenamiento, podemos asegurar que el número de épocas empleado en el mismo era suficiente para las tres redes.

La Tabla 5.1 muestra el valor de las métricas de rendimiento de la clasificación sobre el conjunto de test para las redes SkinLesNet y ResNet50. Ambas obtienen buenos valores, aunque SkinLesNet es ligeramente superior en la mayoría de las métricas, debido a que es una red que se creó buscando el rendimiento óptimo en la clasificación de lesiones

5 Trabajo experimental y resultados

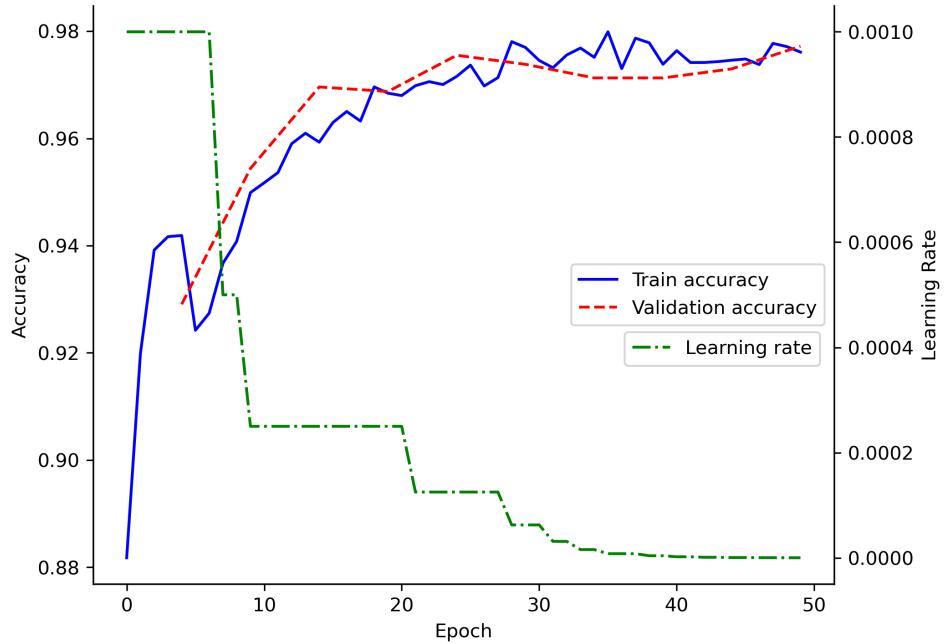


Figura 5.16: Evolución del *accuracy* y de la tasa de aprendizaje (verde) durante el entrenamiento de la ResNet50 para ambos conjuntos de entrenamiento (azul) y validación (rojo).

de piel, y que se ha entrenado desde cero usando imágenes de dicha temática. Por el contrario, el enfoque de *transfer learning* puede perjudicar al rendimiento de ResNet50. La diferencia más significativa se da en el *recall*, estando SkinLesNet dos puntos por encima de ResNet50. En este contexto de imagen médica, es especialmente importante un valor alto de dicha métrica, para tener la seguridad de que se encuentren todos los casos malignos y ninguno se pase por alto, por lo que SkinLesNet parece ser un buen clasificador en este caso.

En la [Figura 5.19](#) se muestran los resultados de aplicar la red U-Net para segmentación sobre algunas de las imágenes de test, las cuales son ciertamente buenas desde un punto de vista cualitativo. Las máscaras que genera la red U-Net no son binarias (0 ó 1) en el sentido estricto, sino que son mapas de colores que indican la probabilidad de que el píxel correspondiente esté dentro de la zona de la lesión. Esto se hizo para dotar de una mayor flexibilidad al modelo durante el aprendizaje de los parámetros y con lo que se han obtenido buenos resultados, ya que en la [Figura 5.19](#) las zonas con probabilidades intermedias son muy estrechas y la mayoría de la imagen está ocupada por zonas de baja probabilidad (morado) y alta probabilidad (amarillo). Si deseamos una máscara puramente binaria, se puede aplicar un filtro de paso alto ideal a la misma a partir de un determinado valor de probabilidad. Resaltar, que en la última imagen, la segmentación no es totalmente precisa con el área de la lesión aunque sí cubre la gran mayoría del área problemática.

5.3 Cálculo de las métricas de calidad para las imágenes a color

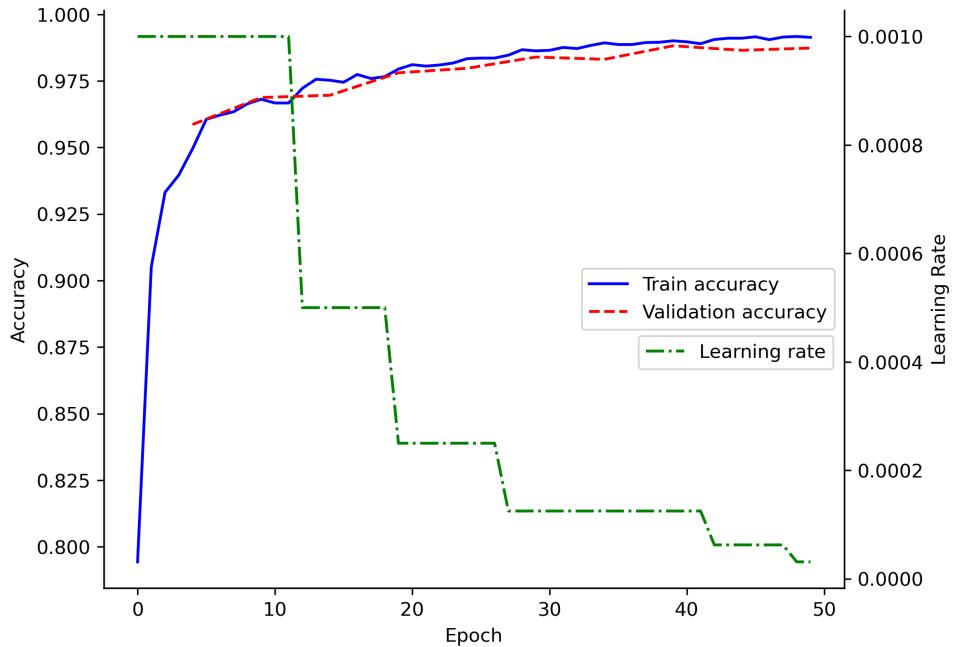


Figura 5.17: Evolución del *accuracy* y de la tasa de aprendizaje (verde) durante el entrenamiento de la SkinLesNet para los conjuntos de entrenamiento (azul) y validación (rojo).

Tabla 5.1: Valores de las métricas de rendimiento para la clasificación binaria sobre el conjunto de test de ISIC para las redes SkinLesNet y ResNet50 en su entrenamiento con las imágenes de lesiones de piel.

| | SkinLesNet | ResNet50 |
|----------------------------|---|---|
| Accuracy | 98.35 % | 98.09 % |
| Recall | 99.13 % | 97.41 % |
| Precision | 97.60 % | 98.50 % |
| F1-score | 98.35 % | 97.96 % |
| Matriz de confusión | $\begin{pmatrix} 568 & 14 \\ 5 & 567 \end{pmatrix}$ | $\begin{pmatrix} 605 & 8 \\ 14 & 527 \end{pmatrix}$ |

5.3.2. Métricas FID y SSIM

En la Tabla 5.2 se muestran los valores obtenidos para el FID y el SSIM a partir de un conjunto de 500 imágenes generadas (ver Figuras 5.20 y 5.21) usando cada uno de los dos modelos incondicionales entrenados, y tomando las imágenes de entrenamiento como *ground truth* para las métricas que así lo requerían, como es el caso del FID. En la generación de imágenes, al igual que en el entrenamiento de los modelos, también se emplearon 1000 pasos temporales con los mismos valores para el conjunto de $\{\beta_t\}_{t=1}^T$. De esta forma, el error introducido en la generación de la imagen por el proceso de

5 Trabajo experimental y resultados

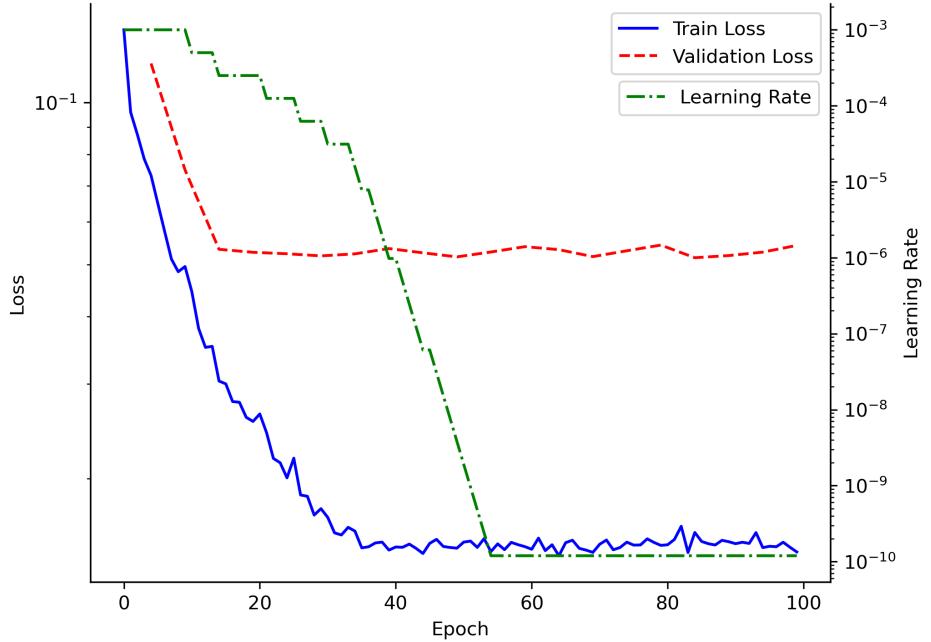


Figura 5.18: Evolución de la pérdida y de la tasa de aprendizaje (verde) para los conjuntos de entrenamiento (azul) y de validación (rojo) para la red U-Net.

eliminación de ruido se reduce, ya que empleamos el mismo esquema de ruido que en el entrenamiento.

Desde el punto de vista cualitativo, los modelos de difusión entrenados generan gran variedad de imágenes de lesiones de piel. Comparando con las imágenes de entrenamiento, hay *samples* que asemejan muy bien las características de las primeras, mientras que otros son algo confusos y de dudosa calidad. En el caso del modelo generador de lesiones benignas, estos casos, que podemos tildar de “indeseables”, son más comunes. Por ejemplo, en la Figura 5.21, hay imágenes sintéticas que son un fondo granulado de color verde, azul o rojo en el que no se observa o aprecia ninguna lesión cutánea ni semejanza con piel humana, lo cual no debería ser a causa de a una falta de mayor número de épocas de entrenamiento de la red U-Net asociada al modelo. Este no es el motivo dado que la función de pérdida para el conjunto de validación en el entrenamiento no mejora desde un valor temprano de épocas. No obstante, hay otras que sí son adecuadas visualmente al contexto. Para la generación de lesiones malignas, son menos frecuentes este tipo de *samples* “defectuosos”. Recalcar que este juicio cualitativo de la calidad de los *samples* no tiene en cuenta conocimientos en medicina y que se limita y basa en la comparación de imágenes con las de entrenamiento. Quizás un experto en la materia considere igualmente válidos esos ejemplos catalogados como “dudosos” en este trabajo, ya que puede ser que muestren características importantes de estas lesiones aisladas de otras que, en conjunto, nos recuerden a una imagen de una área de la piel humana.

5.3 Cálculo de las métricas de calidad para las imágenes a color

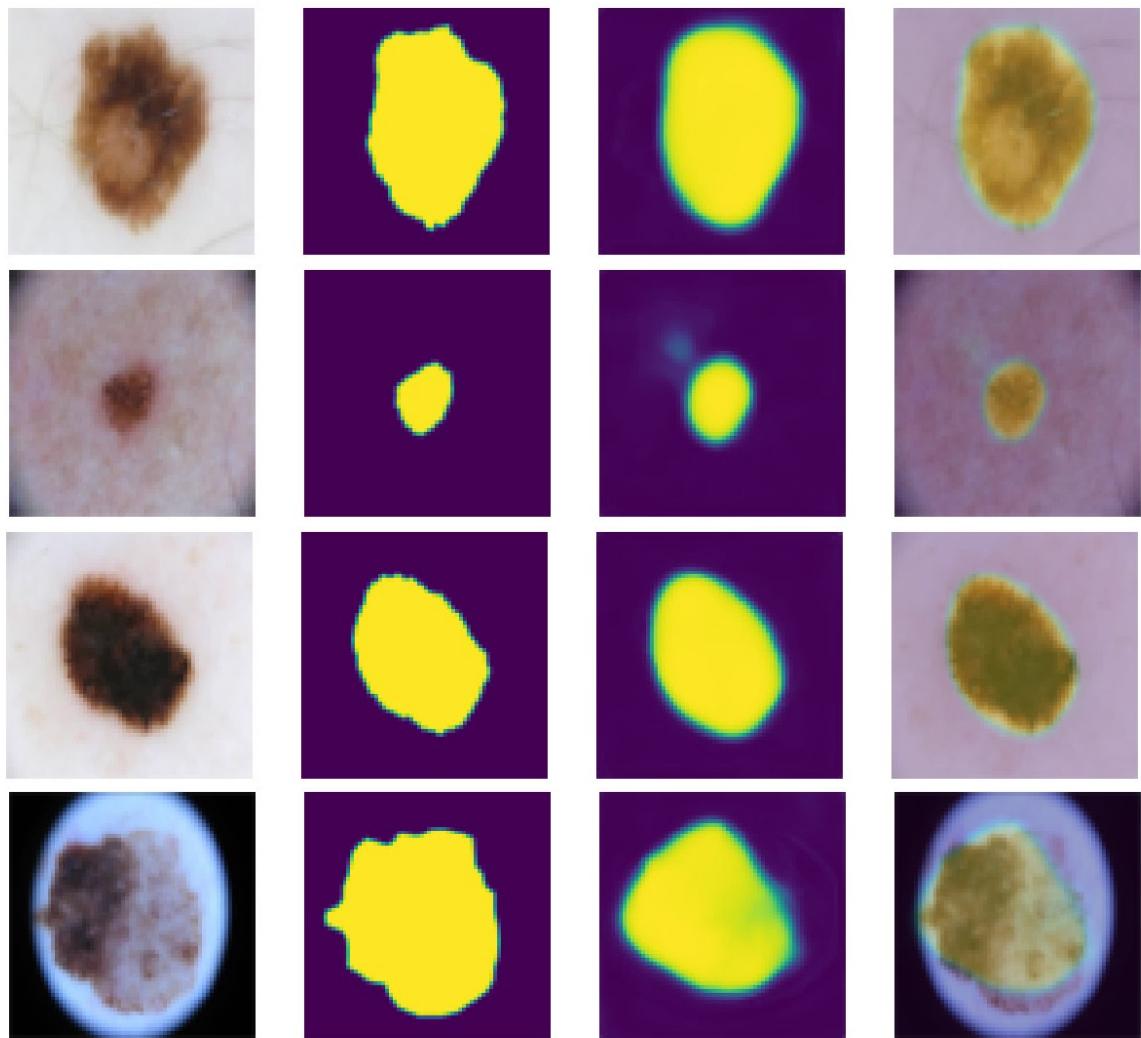


Figura 5.19: Muestra de aplicar la segmentación de la red U-Net entrenada sobre algunos ejemplos del conjunto de test De izquierda a derecha: imagen de test, máscara real correspondiente, máscara predicha por la red y superposición de la imagen real y la máscara predicha.

En primer lugar, los valores obtenidos de SSIM y su variante, el MS-SSIM, para ambos tipos de imagen generada son ciertamente bajos desde un punto de vista objetivo, ya que si las imágenes fueran perceptualmente similares, ese valor sería próximo a uno. Esto nos indica entonces hay imágenes en los conjuntos generados que diferan entre sí desde el punto de vista de la percepción. No obstante, este valor no es malo si se compara con los resultados de calcular estos coeficientes sobre los respectivos conjuntos de entrenamiento de los modelos, obteniendo un valor del SSIM y MS-SSIM,

5 Trabajo experimental y resultados

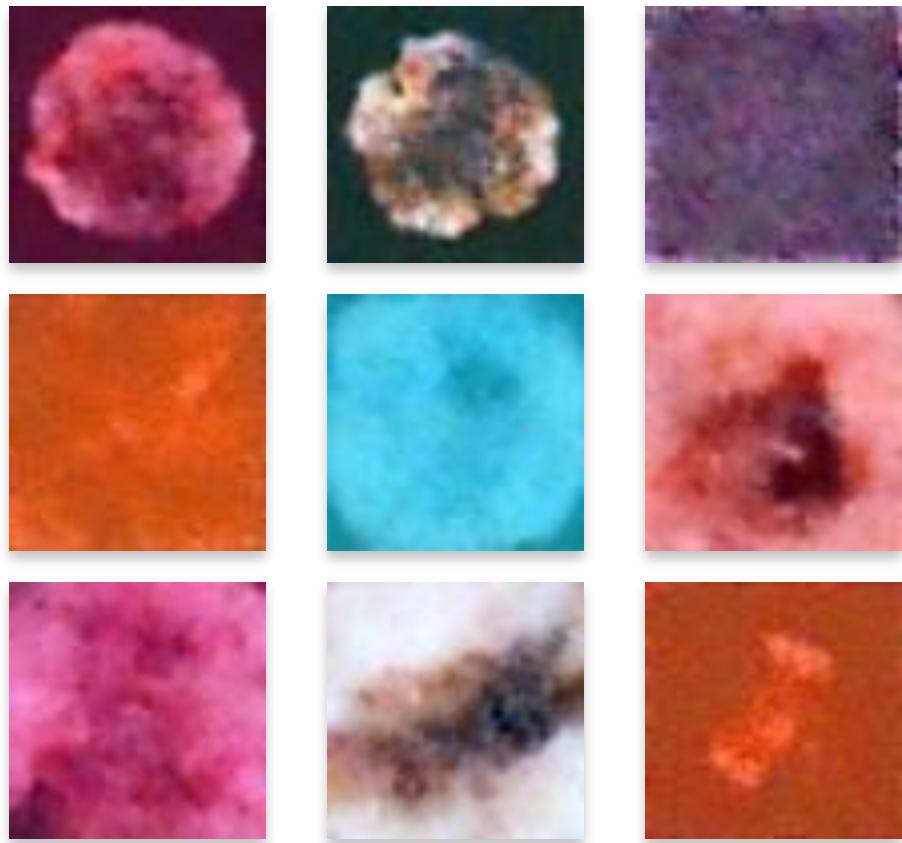


Figura 5.20: Muestra del conjunto de 500 imágenes generadas usando el modelo DDPM incondicional entrenado con imágenes de lesiones malignas.

respectivamente, de 0.42 ± 0.16 y 0.23 ± 0.19 para las imágenes de la clase maligna, y de 0.64 ± 0.09 y 0.43 ± 0.19 para la clase benigna. Es decir, la similitud estructural de las imágenes generadas está dentro de los márgenes que cabría esperar con los datos de entrenamiento de los que se ha dispuesto. Por lo tanto, para dicho coeficiente se obtiene un valor aceptable y esperable para los datos de entrenamiento de los que se dispone.

En cuanto al FID, se tiene una gran disparidad de valores según el extracto de características que se haya empleado. Primero, el valor del coeficiente para la U-Net como extracto de características es alto para ambas clases, lo cual indica alta diferencia entre las distribuciones de las propiedades, consideradas por la red, respecto de aquella de las imágenes de referencia. Dado que la U-Net está entrenada para la realización de segmentación de la lesión de piel, las características que extraiga el *Encoder* estarán relacionadas con parámetros de tamaño y forma de la lesión y el resto de la imagen. Así pues, la distribución de tamaños y formas de las lesiones generadas no parecer casar con la distribución en las imágenes reales (bajo el supuesto de normalidad teó-

5.3 Cálculo de las métricas de calidad para las imágenes a color

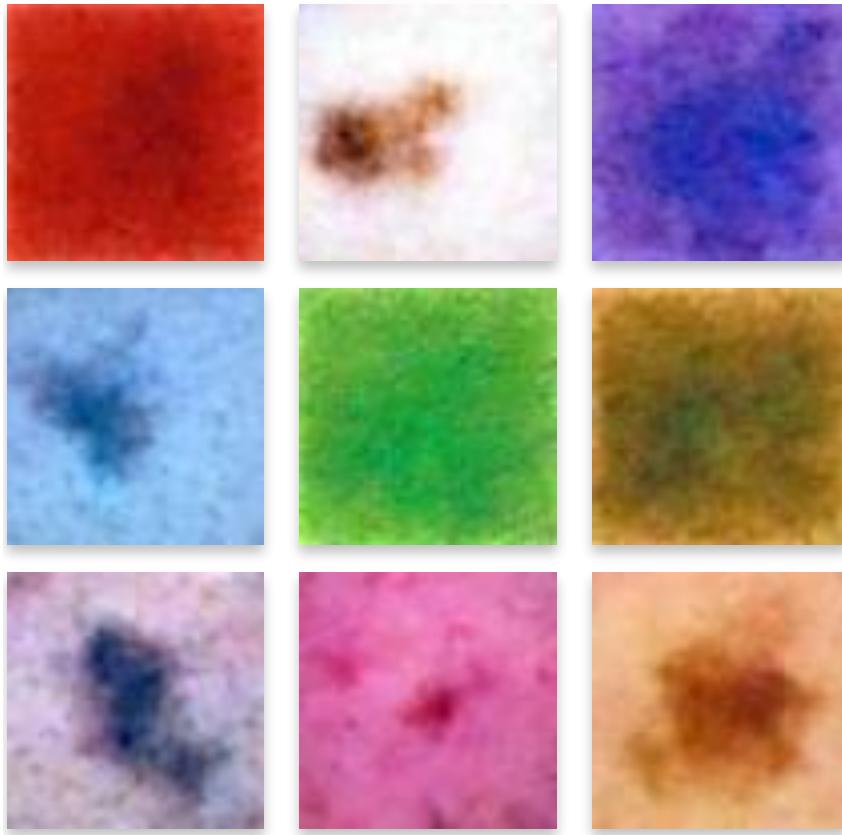


Figura 5.21: Muestra del conjunto de 500 imágenes generadas usando el modelo DDPM incondicional entrenado con imágenes de lesiones benignas o de tipo desconocido.

rico de dicha distribución). En el estudio visual cualitativo realizado anteriormente, se han observado imágenes en las que las lesiones sí parecen tener formas similares a las de las imágenes de entrenamiento, que, de hecho, coinciden con aquellas con una buena calidad; mientras que las imágenes que señalábamos ser inadecuadas no muestran una clara distinción entre la lesión y la zona sana. Estas imágenes pueden ser el motivo de que el FID sea tan alto en este caso.

Para la red SkinLesNet, ocurre algo similar a la U-Net, debido a que esta red también tiene en cuenta patrones de tamaño y forma para la clasificación de las imágenes. No obstante, el valor del FID en este caso es apreciablemente menor, ya que la SkinLesNet tiene en cuenta en la clasificación otros parámetros de la imagen, como, por ejemplo, el color, haciendo que la diferencia entre las distribuciones de características extraídas entre imágenes reales y generadas sea menor. Aun así, la diferencia es alta, probablemente debido a la existencia de aquellas imágenes indeseadas que se ha señalado en el caso anterior.

5 Trabajo experimental y resultados

Tabla 5.2: Valores obtenidos de las métricas de calidad de imagen generada para un conjunto de 500 imágenes generadas de cada clase con los modelos DDPM incondicionales.

| | SSIM | MS-SSIM | FID U-Net | FID SkinlesNet | FID ResNet50 |
|------------------------|-----------------|-----------------|------------------|-----------------------|---------------------|
| Clase malignant | 0.50 ± 0.11 | 0.34 ± 0.22 | 2350.8118 | 474.5047 | 0.4878 |
| Clase benign | 0.51 ± 0.11 | 0.51 ± 0.18 | 2677.1074 | 952.0612 | 3.5287 |

Por último, el valor del FID teniendo a la ResNet50 como extractor de características nos arroja un buen muy valor para ambos casos, mostrando una gran similitud entre las imágenes generadas y las reales, atendiendo, claro está, a las propiedades que dicha red use para la clasificación. Aunque este dato contradiga en parte los valores obtenidos con anterioridad, no debe tomarse como erróneo o restarle importancia, dado que la ResNet50 en su entrenamiento obtuvo unos muy buenos valores para las métricas de clasificación de imágenes de lesiones de piel. Por tanto, las características que toma de las imágenes la ResNet50 son, parte compartidas por SkinLesNet (dado que está obtiene valor más bajos para el FID que U-Net), y otras diferentes y que serían difíciles de conocer debido a la inexplicabilidad de las redes neuronales.

Aunque los valores de métricas como el FID pueden variar significativamente según el extractor de características utilizado, la ResNet50 muestra una consistente similitud entre las imágenes generadas y reales, basándose en las propiedades que considera relevantes para la clasificación. Esto sugiere que, aunque haya disparidades detectadas por otras métricas, la capacidad de generar imágenes similares a las reales existe y puede ser refinada con modelos que capturan características más adecuadas.

5.3.3. Métrica CAS

Si un modelo generativo produce *samples* de calidad, entonces éstos deberían ser tan útiles como las imágenes reales en tareas posteriores a su generación como el entrenamiento de algoritmos de clasificación o de segmentación. En este último apartado, estudiamos el valor las métricas de rendimiento de varios clasificadores entrenados con las imágenes sintéticas o con una combinación de éstas junto con las imágenes reales. Los resultados obtenidos proporcionarán un resultado a un tema actualmente en boga, que estudia si el uso de imágenes sintéticas junto con el de imágenes reales mejora el rendimiento de algoritmos de *Machine Learning* aplicados a imagen médica[18].

Por un lado, se entrenó un clasificador SVM usando como extractor de características de las imágenes el descriptor Histograma de Gradientes Orientados (HOG). El tamaño de las imágenes es de 64 píxeles, por lo que se empleó un tamaño de bloque de 8 píxeles, un tamaño de ventana de 4 píxeles y un paso de bloque de 2 píxeles. Se

5.3 Cálculo de las métricas de calidad para las imágenes a color

empleó un enfoque de *parameters grid*, con el valor de la constante de regularización *C* y el *kernel*, para encontrar los mejores parámetros de la SVM según el *accuracy*. Las imágenes de entrenamiento fueron las sintéticas, estando ambas clases balanceadas. Como conjunto de test, se usó un subconjunto de las imágenes reales.

Los valores de las métricas obtenidos fueron 0.49 en *accuracy*, 0.47 en *precision* y 0.70 en *recall*. A pesar de los malos resultados de este clasificador, llama poderosamente la atención el resultado para el *recall*, por lo que parece que el modelo distingue algo mejor las imágenes de la clase maligna. Aún así, descartamos este clasificador como útil para este conjunto de datos.

HOG es un descriptor que captura la distribución de los gradientes de intensidad en las imágenes, lo cual es útil para ciertos tipos de problemas, como la detección de objetos. Sin embargo, las lesiones de piel pueden tener patrones y texturas muy sutiles y variadas que HOG podría no capturar de manera efectiva. Las diferencias en las texturas y los detalles finos entre las imágenes sintéticas y las reales podrían ser demasiado complejas para que HOG las represente adecuadamente. Incluso la baja resolución de las imágenes y/o los parámetros del descriptor pueden no ser adecuados para capturar características importantes a dicha resolución. Por tanto, si HOG no proporciona características suficientemente distintivas, la SVM no podrá encontrar un hiperplano que separe eficazmente las clases, lo cual podría explicar el bajo valor de *accuracy* y precisión. El valor de *recall* indica que el modelo es relativamente bueno para identificar la mayoría de las lesiones malignas (positivos verdaderos), lo que sugiere que el modelo está más inclinado a identificar cualquier posible lesión maligna, aunque a costa de generar más falsos positivos, lo que se refleja en la baja precisión.

Por otro lado, y a causa de los malos resultados obtenidos con este clasificador, se recurrió a un enfoque de *deep learning*, más usado en la literatura en este tipo de problemas, para la clasificación de las imágenes, usando de nuevo la arquitectura de SkinLesNet. En primer lugar, se entrenó y se puso a prueba la red con el mismo conjunto de datos que se empleó para HOG+SVM. Posteriormente, se hizo lo mismo usando solo el conjunto de datos reales. Y, por último, se usaron ambos *datasets* como conjuntos de entrenamiento de la red neuronal. Los resultados obtenidos para el mismo conjunto de test al usado anteriormente se muestran en la [Tabla 5.3](#). En todos los casos, se entrenó el modelo durante 75 épocas, usando el optimizador Adam, con función de pérdida igual a la entropía cruzada, y la tasa de aprendizaje variando con un enfoque *Reduce On Plateau* desde un valor inicial de 0.001.

Como primera observación de los resultados cabe destacar que éstos son mucho mejores que los obtenidos con el descriptor HOG junto con la SVM. Entrenar la red usando únicamente las imágenes sintéticas arroja resultados ciertamente prometedores. Hay que tener en cuenta, que las imágenes de test son imágenes reales, y que las de entrenamiento usadas contienen casos en los que la calidad de la imagen es dudosa. Así pues, la mejora de la calidad de generación de imagen de los modelos generativos entrenados supondría valores más óptimos de las métricas de clasificación aquí expuestas. Además, llama la atención que, en este caso, el *recall* supere en siete puntos a

5 Trabajo experimental y resultados

Tabla 5.3: Valores de las métricas de rendimiento para el clasificador SkinLesNet usando distintos conjuntos de datos de entrenamiento.

| Datos conjunto de entrenamiento | Acuracy | Precisión | Recall | F1-score |
|---------------------------------|---------|-----------|--------|----------|
| Sintéticos | 0.71 | 0.67 | 0.74 | 0.71 |
| Reales | 0.86 | 0.87 | 0.816 | 0.84 |
| Sintéticos + Reales | 0.853 | 0.850 | 0.832 | 0.84 |

la precisión del modelo, lo cual pueda deberse a que la clase benigna está “contaminada” por imágenes de tipo desconocido y que sean de lesiones malignas; o bien a que el modelo, al estar entrenado con imágenes sintéticas, se vuelva más conservador en su predicción y tienda a clasificar más lesiones benignas como malignas, en un esfuerzo por no pasar por alto ninguna lesión maligna.

En cuanto al uso combinado de imágenes reales y sintéticas, que es la casuística más interesante, para entrenar este clasificador, cabe destacar que el rendimiento de la red disminuye un poco cuando se usan ambos tipos de datos de entrenamiento frente a usar únicamente datos reales, salvo en el caso del *recall*, el cual aumenta.

El *recall* es la capacidad de un modelo para identificar correctamente todos los casos positivos (en este contexto, lesiones malignas de piel). Al integrar imágenes sintéticas en el conjunto de datos de entrenamiento, es posible que el modelo aprenda características adicionales o más variadas de las lesiones malignas. Esto puede llevar a una mejora en la sensibilidad del modelo. En otras palabras, el modelo es más propenso a detectar correctamente las lesiones malignas, reduciendo así los falsos negativos (casos donde no se detecta una lesión maligna cuando sí está presente). Como el objetivo principal de estos algoritmos de aprendizaje es detectar las lesiones malignas en la piel, las cuales son minoría frente al número de lesiones benignas que se registran, es muy importante que el valor de esta métrica sea lo más elevado posible. Además, es preferible equivocarse en detectar como maligna una lesión inofensiva, como puede ser un lunar, frente al caso contrario, ya que estos métodos constituyen en muchos casos una forma de cribado de pacientes, quedando las imágenes catalogadas como malignas bajo la confirmación de un especialista médico. Por tanto, se puede concluir que anexar al conjunto de imágenes reales, otras imágenes sintéticas de calidad, ayuda a mejorar la tasa de acierto en la detección de lesiones malignas de piel en clasificadores como SkinLesNet.

Sin embargo, al mismo tiempo que mejora el *recall*, la precisión puede disminuir. La precisión se refiere a la proporción de resultados positivos que son verdaderamente positivos. Cuando se incluyen imágenes sintéticas, el modelo puede volverse más propenso a identificar falsos positivos (casos donde identifica una lesión maligna donde no la hay), lo cual afecta la precisión. Finalmente, el F1-score es una medida que combina tanto la precisión como el *recall* en una sola métrica. Como el F1-score se mantiene constante, esto indica que hay una compensación entre ambas métricas y que

5.3 Cálculo de las métricas de calidad para las imágenes a color

el rendimiento general del modelo no se ve afectado negativamente por la inclusión de imágenes sintéticas, aunque el *accuracy* muestra que sí es posible que se haya visto un poco afectado.

Incluso, para todos los clasificadores aquí expuestos, hay que tener en cuenta que, a pesar de que las clases están balanceadas en el conjunto de entrenamiento sintético, esto no garantiza que la distribución de características en las imágenes reales de prueba sea similar. Las imágenes sintéticas pueden tener diferencias sutiles pero significativas en comparación con las reales, lo que afecta a la capacidad del clasificador para generalizar. De hecho, muestra de ello dan los valores altos del FID para algunos de los extractores de características empleados.

Pueden usarse otros clasificadores para el cálculo del CAS, como otras arquitecturas de redes neuronales más actuales y conocidas, nnU-Net, *Segment Anything Model* (SAM), InceptionV3..., así como otros extractores de características para el FID.

Conclusiones y trabajo futuro

En este trabajo se ha llevado a cabo un análisis exhaustivo de los modelos de difusión, especialmente de su aplicación a la generación de imágenes médicas, con un enfoque particular en las lesiones cutáneas. A lo largo de los experimentos y análisis realizados, se han obtenido resultados que destacan el posible potencial de estos modelos para mejorar la precisión y eficacia en el ámbito de la detección y clasificación de enfermedades.

Las imágenes generadas mediante los modelos de difusión entrenados han demostrado tener cierto nivel de calidad, exceptuando ciertos ejemplos donde el estudio cualitativo la descartaba, evaluada a través de métricas como FID, SSIM o CAS y alcanzando niveles comparables a las imágenes reales para las dos últimas. En el caso del FID, dependiendo del extractor explotado, los resultados son bastante variables, lo cual puede deberse a las características que destaque cada red extractora de la imagen. Respecto a los ejemplos “indeseables”, las redes SkinLesNet y ResNet50 aprendían a clasificarlos correctamente, por lo que deben mostrar características importantes de la clase a la que pertenecen. La capacidad de generar imágenes de alta calidad es fundamental para su aplicación en tareas médicas, donde la precisión visual es crucial para un diagnóstico correcto. La resolución y tamaño de las imágenes usadas en el entrenamiento y de aquellas que el modelo generase podría aumentarse usando un hardware más potente que soportara dicha carga de trabajo de forma razonable, lo que aumentaría la calidad de las imágenes generadas (más allá del aumento de la resolución) y podría mejorar el valor de las métricas.

En cuanto a las imágenes generadas cualitativamente ‘indeseables’, como era el caso de las imágenes en negro, o de aquellas con un tono azul general, puede tomarse otros espacios de color, en vez de RGB, para intentar evitar que la red aprenda a generar dichos ejemplos. Otra alternativa a probar sería entrenar un modelo generativo para cada uno de los canales del espacio de color, en este caso RGB, y observar los resultados obtenidos.

La integración de imágenes sintéticas en el conjunto de datos de entrenamiento de clasificadores para la detección de lesiones de piel malignas ha mostrado un rendimiento muy similar de los clasificadores como SkinLesNet, por lo que dichas imágenes no lo perjudican. De hecho, se ha observado una mayor tasa de acierto, especialmente del *recall*, en la identificación de lesiones malignas, lo cual es esencial, particularmente en el contexto clínico, donde es preferible minimizar los falsos negativos, y así garantizar la detección temprana y el tratamiento efectivo de condiciones potencialmente mortales. No obstante, ese aumento del *recall* se produce en detrimento de la precisión,

Conclusiones y trabajo futuro

quedando así el F1-score invariante entre el rendimiento del clasificador entrenado solo con imágenes reales y entrenado con imágenes sintéticas y reales. Este aumento puede deberse a que al introducir imágenes sintéticas en el conjunto de datos de entrenamiento, es posible que el modelo aprenda características adicionales o más variadas de las lesiones malignas, y puede volverse más propenso a identificar falsos positivos (lo cual afecta a la precisión).

Los modelos de difusión han demostrado ser adaptables a diversos tipos de imágenes médicas, lo que sugiere un alto potencial para su aplicación en diferentes áreas de la medicina. Esta adaptabilidad permite la posibilidad de extender su uso más allá de los que se dan en este estudio, incluyendo otras modalidades de imagen médica como son radiografías, tomografías y resonancias magnéticas. Los modelos de difusión también pueden utilizarse para la generación de imágenes 3D y de vídeo (tanto en su versión 2D como 3D). Esto abre un amplio abanico de nuevas modalidades donde estudiar su aplicabilidad y rendimiento.

No obstante, la calidad de las imágenes generadas, sobre todo el caso de las lesiones de piel, podría mejorarse con un preprocesamiento más exhaustivo de las imágenes de entrenamiento de los modelos. Por ejemplo, muchas de dichas imágenes contenían vello corporal o parches de colores, problemáticas las cuales pueden entorpecer el proceso de aprendizaje del modelo. Además, el rescalado de las imágenes por limitaciones de memoria podía causar la pérdida de características importantes en las imágenes, haciendo que los *samples* generados no tuvieran el nivel de detalle deseado, aunque esto no parece haber sucedido en nuestro caso. Aun así, es deseable tener ejemplos de alta resolución para así poder ver detalles y tener más información de la imagen.

El uso de técnicas de segmentación, como la red U-Net entrenada para la extracción de características en este proyecto, podría segmentar las imágenes de entrenamiento, usando esta información para la eliminación de la parte innecesaria de la imagen y así centrar la atención de las redes neuronales en la parte importante de la imagen.

Para una imagen de una lesión maligna en la piel, esta puede corresponderse con algunos de los distintos tipos de cáncer de piel existentes, como carcinoma, melanoma, linfoma, etc. Además, de los diferentes subtipos dentro de estos, cada uno con unas determinadas características que los diferencian del resto. Por ejemplo, en las imágenes de los conjuntos de entrenamiento utilizados, se observa una gran variedad de imágenes dentro de la misma clase, existiendo diferencias significativas entre ellas. Es por ello que se podrían clasificar las imágenes de los conjuntos maligno y benigno en las distintas categorías de las lesiones, y entrenar a los modelos generativos y de clasificación atendiendo a dicho etiqueta más concreto. A pesar de que ISIC no suele proporcionar información sobre las imágenes más allá de si pertenece a una lesión de tipo maligno, benigno o desconocido, podrían emplearse técnicas de *clustering* para hacer una agrupación entre imágenes con características similares y usarla en el entrenamiento de modelos posteriores. Al final, como el principal problema a resolver es la distinción entre lesión maligna y benigna, consideramos estas clases más generales en el análisis y presentación de los resultados.

Otro aspecto en el que se puede mejorar es el uso de las otros tipos de modelos de difusión explicados en [Capítulo 2](#), como los modelos de difusión condicionales, o en la utilización de hibridaciones de éstos con otras formas de métodos generativos como los VAEs o GANs. También, explorar más arquitecturas de la redes convolucionales asociadas a estos métodos generativos puede ayudar a incrementar la calidad de las imágenes sintéticas. Todas estas son propuestas de mejora que se recogen como trabajo a futuro para conseguir mejorar el rendimiento de estos métodos generativos.

La investigación actual ha establecido una base sólida para la futura integración de modelos generativos en prácticas clínicas, ofreciendo una herramienta poderosa para la generación de datos sintéticos que pueden complementar y enriquecer los datos existentes. Se vislumbra un futuro donde estas tecnologías pueden facilitar diagnósticos más rápidos y precisos, optimizando los recursos médicos y mejorando la atención al paciente. Los resultados presentados pueden servir como precedente para futuras investigaciones y aplicaciones prácticas que podrían revolucionar el campo de la imagenología médica, beneficiando tanto a los profesionales de la salud como a los pacientes

Bibliografía

- [1] Mohamed Akrout, Bálint Gyepesi, Péter Holló, Adrienn Poór, Blága Kincső, Stephen Solis, Katrina Cirone, Jeremy Kawahara, Dekker Slade, Latif Abid, Máté Kovács, y István Fazekas. Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images, 2023.
- [2] Tomer Amit, Eliya Nachmani, Tal Shaharbany, y Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *ArXiv*, abs/2112.00390, 2021.
- [3] Syed Muhammad Anwar, Muhammad Majid, Adnan Qayyum, Muhammad Awais, Majdi Alnowami, y Muhammad Khurram Khan. Medical image analysis using convolutional neural networks: A review. *Journal of Medical Systems*, 42(11), October 2018.
- [4] Tehreem Awan y Khan Bahadar Khan. Investigating the impact of novel xraygan in feature extraction for thoracic disease detection in chest radiographs: lung cancer. *Signal, Image and Video Processing*, 18(5):3957–3972, May 2024.
- [5] Muhammad Azeem, Kaveh Kiani, Taha Mansouri, y Nathan Topping. Skinlesnet: Classification of skin lesions and detection of melanoma cancer using a novel multi-layer deep convolutional neural network. *Cancers*, 16(1), 2024.
- [6] Samah Saeed Baraheem, Trung-Nghia Le, y Tam V. Nguyen. Image synthesis: a review of methods, datasets, evaluation metrics, and future outlook. *Artificial Intelligence Review*, 56(10):10813–10865, February 2023.
- [7] Shane Barratt y Rishi Sharma. A note on the inception score, 2018.
- [8] Yaniv Benny, Tomer Galanti, Sagie Benaim, y Lior Wolf. Evaluation metrics for conditional image generation. *International Journal of Computer Vision*, 129(5):1712–1731, March 2021.
- [9] Ali Borji. Pros and cons of gan evaluation measures: New developments, 2021.
- [10] Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, y Stan Z. Li. A survey on generative diffusion model, 2023.
- [11] Yizhou Chen, Xu-Hua Yang, Zihan Wei, Ali Asghar Heidari, Nenggan Zheng, Zhicheng Li, Huiling Chen, Haigen Hu, Qianwei Zhou, y Qiu Guan. Generative adversarial networks in medical image augmentation: A review. *Computers in Biology and Medicine*, 144:105382, May 2022.
- [12] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, y Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), 2019.
- [13] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, y

Bibliografía

- Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic), 2017.
- [14] Marc Combalia, Noel C. F. Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C. Halpern, Susana Puig, y Josep Malvehy. Bcn20000: Dermoscopic lesions in the wild, 2019.
 - [15] Prafulla Dhariwal y Alex Nichol. Diffusion models beat gans on image synthesis, 2021.
 - [16] D.C Dowson y B.V Landau. The fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982.
 - [17] John Duchi. *Derivations for Linear Algebra and Optimization*.
 - [18] Muhammad Ali Farooq, Wang Yao, Michael Schukat, Mark A Little, y Peter Corcoran. Derm-t2im: Harnessing synthetic skin lesion data via stable diffusion models for enhanced skin disease classification using vit and cnn, 2024.
 - [19] Hugo Francisco. News-medical, May 2024.
 - [20] Tami Freeman. Deep transfer learning detects six different cancers on pet/ct scans – physics world, Jun 2024.
 - [21] Changfei Gong, Yuling Huang, Mingming Luo, Shunxiang Cao, Xiaochang Gong, Shenggou Ding, Xingxing Yuan, Wenheng Zheng, y Yun Zhang. Channel-wise attention enhanced and structural similarity constrained cyclegan for effective synthetic ct generation from head and neck mri images. *Radiation Oncology*, 19(1), March 2024.
 - [22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, y Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, October 2020.
 - [23] David Gutman, Noel C. F. Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, y Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic), 2016.
 - [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, y Jian Sun. Deep residual learning for image recognition, 2015.
 - [25] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, y Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018.
 - [26] Jonathan Ho, Ajay Jain, y Pieter Abbeel. Denoising diffusion probabilistic models, 2020.
 - [27] Siwon Hwang y Jitae Shin. Prognosis prediction of alzheimer's disease based on multi-modal diffusion model. En *2024 18th International Conference on Ubiquitous Information Management and Communication (IMCOM)*. IEEE, January 2024.
 - [28] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, y Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation, 2024.

- [29] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, y Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88:102846, 2023.
- [30] Aghiles Kebaili, Jérôme Lapuyade-Lahorgue, y Su Ruan. Deep learning approaches for data augmentation in medical imaging: A review. *Journal of Imaging*, 9(4):81, April 2023.
- [31] Boah Kim y Jong Chul Ye. *Diffusion Deformable Model for 4D Temporal Medical Image Generation*, página 539–548. Springer Nature Switzerland, 2022.
- [32] Diederik P. Kingma y Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.
- [33] Xin Li, Yulin Ren, Xin Jin, Cuiling Lan, Xingrui Wang, Wenjun Zeng, Xinchao Wang, y Zhibo Chen. Diffusion models for image restoration and enhancement – a comprehensive survey, 2023.
- [34] Fangjian Liao, Xingxing Zou, y Wai Keung Wong. Attentional pixel-wise deformation for pose-based human image generation. *Expert Systems with Applications*, 246:123073, July 2024.
- [35] Shaohui Liu, Yi Wei, Jiwen Lu, y Jie Zhou. An improved evaluation framework for generative adversarial networks, 2018.
- [36] Xueyan Mei, Zelong Liu, Philip M. Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E. Link, Thomas Yang, Ying Wang, Hayit Greenspan, Timothy Deyer, Zahi A. Fayad, y Yang Yang. Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 4(5), September 2022.
- [37] Puria Azadi Moghadam, Sanne Van Dalen, Karina C. Martin, Jochen Lennerz, Stephen Yip, Hossein Farahani, y Ali Bashashati. A morphology focused diffusion probabilistic model for synthesis of histopathology images, 2022.
- [38] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [39] Alex Nichol y Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021.
- [40] Kai Packhäuser, Lukas Folle, Florian Thamm, y Andreas Maier. Generation of anonymous chest radiographs using latent diffusion models for training thoracic abnormality classification systems, 2022.
- [41] Suman Ravuri y Oriol Vinyals. Classification accuracy score for conditional generative models, 2019.
- [42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, y Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.
- [43] Olaf Ronneberger, Philipp Fischer, y Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.
- [44] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, Allan Halpern, Brian Helba, Harald Kittler, Kivanc Kose, Steve Langer, Konstantinos

Bibliografía

- Lioprys, Josep Malvehy, Shenara Musthaq, Jabpani Nanda, Ofer Reiter, George Shih, Alexander Stratigos, Philipp Tschandl, Jochen Weber, y H. Peter Soyer. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data*, 8(1), January 2021.
- [45] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, y Xi Chen. Improved techniques for training gans, 2016.
- [46] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, y Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015.
- [47] Yang Song y Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *CoRR*, abs/1907.05600, 2019.
- [48] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, y Ben Poole. Score-based generative modeling through stochastic differential equations, 2021.
- [49] Philipp Tschandl, Cliff Rosendahl, y Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1), August 2018.
- [50] Z. Wang, E.P. Simoncelli, y A.C. Bovik. Multiscale structural similarity for image quality assessment. En *The Thirly-Seventh Asilomar Conference on Signals, Systems Computers*, 2003, volume 2, páginas 1398–1402 Vol.2, 2003.
- [51] Lilian Weng. What are diffusion models? *lilianweng.github.io*, Jul 2021.
- [52] Zhisheng Xiao, Karsten Kreis, y Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion GANs. En *International Conference on Learning Representations*, 2022.
- [53] Jiancheng Yang, Rui Shi, y Bingbing Ni. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. En *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, páginas 191–195, 2021.
- [54] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, y Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.
- [55] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, y Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications, 2024.
- [56] Yijun Yang, Huazhu Fu, Angelica I. Aviles-Rivero, Carola-Bibiane Schönlieb, y Lei Zhu. Diffmic: Dual-guidance diffusion network for medical image classification, 2023.