Nombre del tesista: Halan Alexander Lucas Villarroel Dueñas

Fecha de inscripción: 1 octubre 2024

Nombre de tesis: Aumentación de datos usando modelos de difusión y su aplicación en medicina

Institución: Universidad Andres Bello

Profesor: Billy Mark Peralta Márquez

Facultad de Ingeniería

Magíster en Ciencias de la Computación

Aumentación de datos usando modelos de difusión y su aplicación en medicina

Autor: Halan Alexander Lucas Villarroel Dueñas

Profesor guía: Billy Mark Peralta Márquez

Tesis para optar al Grado de

Magíster en Ciencias de la Computación

Viña del Mar - Chile

2024

Índice general

Agradecimientos

Deseo expresar mi agradecimiento al profesor guía Billy Mark Peralta Márquez por su paciencia, enseñanza y constante tiempo durante esta investigación.

También deseo agradecer enormemente a mi familia, mi madre, padre, hermana, madrina y a mi novia Noelia por los constantes apoyos y ánimos que fueron de suma importancia durante este proceso.

Por último, agradecer a Dios por darme la perseverancia y sabiduría necesaria para lograr concluir esta investigación.

Abstract

En esta tesis se propone analizar y evaluar el uso de modelos de difusión como una herramienta avanzada para la aumentación de datos en el ámbito médico, un área donde la disponibilidad de conjuntos de datos de calidad suele estar restringida debido a limitaciones éticas, logísticas y económicas. La aumentación de datos, como estrategia en el aprendizaje automático, busca enriquecer los conjuntos disponibles mediante la generación de ejemplos sintéticos que complementen la información original, mejorando la representación de las características inherentes a los datos. Este enfoque es especialmente relevante en escenarios médicos, donde el desbalance o la escasez de datos puede afectar de manera crítica el rendimiento y la fiabilidad de los modelos predictivos.El trabajo de investigación se enfoca en implementar un modelo de difusión para generar características adicionales, ofreciendo una solución innovadora frente a los métodos tradicionales de aumentación. Este modelo busca destacar por su capacidad para preservar la fidelidad y diversidad de los datos, asegurando que los ejemplos sintéticos reflejen de manera precisa las propiedades clínicas esenciales. Además, se busca comparar esta técnica con enfoques tradicionales de aumentación aplicados a las características, evaluando su impacto en tareas específicas.

Capítulo 1

Introducción

1. Contextualización del problema

1.1. La tecnología en el campo médico. El avance de la tecnología ha transformado significativamente el ámbito médico, comenzando desde la introducción de dispositivos de diagnóstico hasta el uso de inteligencia artificial (IA) para el análisis de datos complejos. Entre las nuevas tecnologías se encuentran los modelos generativos y las redes neuronales, que están posicionados para revolucionar el tratamiento de problemas asociados con la limitada disponibilidad de datos médicos de calidad.

1.2. Modelos generativos en medicina. Los modelos generativos, como las GANs y los modelos de difusión, han encontrado diversidad de aplicaciones en el sector médico, incluyendo la síntesis de imágenes, la aumentación de datos y la generación de características sintéticas. Estas técnicas apoyan a superar desafíos tradicionales como el desbalance de clases y la falta de datos etiquetados en diagnósticos específicos.

1.3. La aumentación de datos y sus aplicaciones. La aumentación de datos ha demostrado ser una solución clave para mejorar la representación de características en conjuntos de datos desbalanceados o reducidos. En el ámbito médico, se utiliza para aumentar la diversidad de datos cruciales para entrenar modelos que sean capaces de generalizar mejor en tareas como la clasificación de enfermedades y la predicción de riesgos clínicos.

1.4. Importancia de los datos. La calidad y cantidad de datos disponibles son determinantes para el éxito de los modelos de aprendizaje automático. Datos limitados o desequilibrados pueden llevar a modelos poco fiables, especialmente en aplicaciones críticas como el diagnóstico de enfermedades. La capacidad de generar datos sintéticos de alta calidad es, por lo tanto, una necesidad urgente.

1.5. Barreras y desafíos en el campo. Un gran desafío en la aumentación de datos médicos es la falta de enfoques que combinen la fidelidad clínica y la diversidad de los datos

generados. Además, la dependencia de conjuntos de datos etiquetados manualmente sigue siendo una barrera debido al alto costo y tiempo requerido para su creación. Este problema se agrava en casos de enfermedades raras o condiciones específicas con datos limitados.

1.6. Impacto de Enfermedades Críticas en la Investigación Médica. En la investigación médica, algunas enfermedades se caracterizan por tener una alta mortalidad a pesar de su baja incidencia. Un ejemplo es el melanoma, que representa aproximadamente el 1 % de todos los casos de cáncer de piel [**1**]. Sin embargo, cuando se consideran únicamente los cánceres de piel más agresivos y clínicamente diagnosticados, el melanoma es responsable de alrededor del 62.1 % de las muertes [**2**], debido a su alta agresividad y capacidad de propagación.

Este marcado desbalance entre incidencia y mortalidad plantea un desafío significativo para los sistemas de salud y la investigación médica. Aunque el melanoma es menos frecuente que otros tipos de cáncer de piel, su elevada tasa de mortalidad resalta la necesidad de diagnósticos tempranos y tratamientos eficaces.

1.7. Propuesta innovadora. En este estudio, se propone el uso de modelos de difusión como una solución innovadora para la generación de características sintéticas en datos médicos, motivado por el aprovechamiento de su capacidad para producir imágenes altamente realistas. Estos modelos no solo generan datos de alta calidad, sino que también preservan las características esenciales de los datos originales, superando así las limitaciones de las técnicas tradicionales de aumentación de datos. El enfoque planteado promete mejorar la representación de las clases minoritarias, reducir el desbalance en los conjuntos de datos y optimizar el rendimiento de los modelos predictivos en aplicaciones médicas críticas.

2. Definición del problema

Hasta la fecha de esta investigación, se han desarrollado diversas técnicas de aumentación de datos que han demostrado cierto potencial. Sin embargo, hay un amplio margen de mejora y opciones para investigar, como el aumento de las características de los datos. Estas técnicas enfrentan desafíos importantes en el ámbito médico, donde los datos disponibles suelen ser escasos, desbalanceados y difíciles de etiquetar de manera precisa. Dos problemas principales se identifican en este contexto:

- Variabilidad intrínseca en los datos médicos: Las imágenes médicas presentan una alta variabilidad debido a factores como la calidad del equipo, condiciones de adquisición y diferencias entre pacientes, lo que dificulta la consistencia en el entrenamiento de modelos.
- Necesidad de cumplir con regulaciones éticas: El uso de datos médicos está sujeto a estrictas regulaciones y políticas de privacidad, lo que limita el acceso a grandes volúmenes de datos y dificulta el desarrollo de modelos robustos.

En el ámbito médico, la recopilación de datos etiquetados enfrenta múltiples desafíos. La necesidad de contar con expertos para etiquetar imágenes de manera precisa eleva significativamente el costo y tiempo de este proceso. Además, las imágenes médicas suelen presentar alta variabilidad debido a factores como las condiciones de adquisición, diferencias entre pacientes y calidad del equipo utilizado, lo que complica aún más la generación de conjuntos de datos consistentes. Esto resalta la importancia de desarrollar métodos capaces de aprender eficazmente con datos escasos y minimizar el esfuerzo de etiquetado humano, como los modelos de difusión combinados con aumentación de datos.

Capítulo 2

Justificación de la investigación

El aprendizaje automático en el ámbito médico enfrenta retos significativos debido a la naturaleza compleja y limitada de los datos disponibles. La variabilidad en las imágenes médicas, ocasionada por factores como las condiciones de adquisición, calidad de los equipos y diferencias entre pacientes, sumado a otros factores como el cumplimiento de regulaciones éticas y políticas de privacidad, restringe el acceso a grandes volúmenes de datos, lo que dificulta aún más la capacidad de desarrollar sistemas robustos y fiables.

El etiquetado de datos agrega otra capa de complejidad, ya que este proceso requiere la intervención de expertos, lo que incrementa considerablemente los costos y el tiempo necesarios para generar conjuntos adecuados. Adicionalmente, el desbalance de clases, donde ciertas condiciones o patologías están subrepresentadas, limita la capacidad de los modelos para aprender de manera eficiente, incrementando los riesgos de sesgo y reduciendo la precisión en escenarios críticos como lo es el diagnóstico médico.

En este contexto, la presente investigación busca abordar estas limitaciones a través del uso de modelos de difusión combinados con técnicas de aumentación de datos. Estos modelos permiten generar características sintéticas que enriquecen los conjuntos de datos existentes, mitigando la escasez y el desbalance. Este enfoque no solo optimiza el uso de los datos disponibles, sino que también minimiza la necesidad de etiquetado manual, reduciendo la dependencia de recursos especializados.

La relevancia de esta investigación radica en que, al mejorar el rendimiento y la adaptabilidad de los modelos de aprendizaje automático bajo condiciones de datos limitados, se contribuye al avance de aplicaciones médicas críticas como la detección temprana de enfermedades y el desarrollo de sistemas de diagnóstico más accesibles y precisos. Además, la exploración de nuevas técnicas para la generación de datos sintéticos tiene el potencial de establecer una base para futuras investigaciones en escenarios donde los datos sean un recurso escaso o costoso de obtener.

Capítulo 3

Preguntas de investigación e hipótesis

¿Cómo se puede implementar una solución innovadora en el contexto médico, en el cual se puede encontrar una cantidad desequilibrada y reducida de datos?

1. Hipótesis

El uso de modelos de difusión para la generación de datos sintéticos en el campo médico mejora significativamente el rendimiento de los modelos de aprendizaje automático en tareas de clasificación y predicción en comparación con los métodos tradicionales de aumentación, especialmente en escenarios con conjuntos limitados o desbalanceados.

2. Objetivo general

Desarrollar y evaluar un modelo basado en redes de difusión para la generación de características sintéticas en datos médicos con el fin de mejorar el rendimiento y la precisión de los modelos de aprendizaje automático en escenarios con datos limitados y/o desbalanceados.

3. Objetivos específicos

- Seleccionar un conjunto de datos para comenzar la implementación de la solución propuesta.
- Implementar un pipeline de preprocesamiento de datos, incluyendo la extracción de características con redes neuronales preentrenadas, garantizando la calidad y consistencia de los datos iniciales.
- Entrenar y optimizar un modelo de difusión para generar imágenes de características sintéticas que preserven la diversidad y la fidelidad de las originales.
- Comparar el modelo de difusión propuesto con técnicas tradicionales de aumentación de datos en términos de métricas como precisión, F1-score y recall.

- Validar la capacidad del modelo en escenarios con desbalance de clases, evaluando su impacto en tareas de diagnóstico médico con conjuntos de datos limitados.

Capítulo 4

Marco Teórico

El aprendizaje automático (Machine Learning, ML) ha transformado la manera en la cual se abordan problemas dentro del ámbito médico. Su capacidad para procesar grandes volúmenes de datos y extraer patrones complejos ha resultado clave en aplicaciones como el diagnóstico asistido por computadora, la segmentación de imágenes médicas y la predicción de resultados clínicos. Tecnologías que prometen mejorar la precisión, rapidez y eficiencia en los procesos médicos, impactando positivamente en la calidad de la atención sanitaria.

En el campo médico un área importante es la generación y el procesamiento de datos médicos para escenarios con limitaciones de datos, como en imágenes dermatológicas. Se han investigado diversos enfoques para abordar estos desafios, destacándose las siguientes técnicas:

1. Redes Neuronales

- **Red Neuronal Convolucional (CNN):** Las CNN son estructuras avanzadas de redes neuronales diseñadas para tareas supervisadas. Se inspiran en la organización de la corteza visual humana, lo que les permite extraer características relevantes a partir de imágenes en una jerarquía de capas. Las capas iniciales detectan patrones básicos como bordes y texturas, mientras que las capas más profundas identifican características más abstractas. En este trabajo, se emplea una ResNet18 preentrenada para la extracción de características de imágenes dermatológicas, destacando su capacidad para generar representaciones robustas.
- **Modelos de Difusión:** La arquitectura U-Net fue diseñada específicamente para la segmentación de imágenes biomédicas. Su estructura en forma de U consta de una fase de contracción, donde se extraen características de bajo nivel, y una fase expansiva, donde se generan predicciones pixel a pixel. En este proyecto, la U-Net se integra con un modelo de difusión para la generación de datos sintéticos de alta calidad.

2. Métodos de validación

- **Enfoque holdout:** Es una técnica que divide un set de datos en dos o más subconjuntos. Se utilizó una división 80-20 del conjunto de datos de entrenamiento, separándolo en subconjuntos de entrenamiento y validación. El conjunto de entrenamiento se empleó para ajustar los parámetros del modelo, mientras que el conjunto de validación permitió evaluar su desempeño durante el proceso de entrenamiento, ayudando a prevenir problemas.
- **Justificación con Métricas y Evaluación:** En los proyectos que involucran modelos de aprendizaje automático, la justificación del rendimiento del modelo se fundamenta en métricas de evaluación que cuantifican la precisión, capacidad de generalización y eficiencia del modelo. Estas métricas no solo permiten comparar modelos y técnicas, sino también asegurar que el sistema desarrollado cumple con los objetivos planteados. En esta investigación se usaron varias métricas de evaluación para el modelo, entre ellas:
  - **Exactitud (Accuracy):** La proporción de predicciones correctas respecto al total de predicciones realizadas.
  - **Precisión (Precision):** La proporción de verdaderos positivos entre todas las predicciones positivas realizadas. Indicador clave para minimizar falsos positivos.
  - **Sensibilidad o Recall:** Mide la capacidad del modelo para identificar correctamente todas las instancias positivas (minimizar falsos negativos).
  - **F1-Score:** La media armónica entre precisión y sensibilidad, utilizada especialmente en conjuntos de datos desbalanceados.
  - **Matriz de Confusión:** Una representación tabular que permite analizar en detalle los errores cometidos por el modelo (falsos positivos, falsos negativos, verdaderos positivos y verdaderos negativos).

Capítulo 5

Revisión bibliográfica

1. Revisión de métodos

1.1. Wang et al. [3]. proponen emular la capacidad humana de visualizar objetos desde diferentes perspectivas para mejorar la visión artificial. Introducen un método que combina un meta-learner con un generador de ejemplos (*alucinador*) para optimizar la clasificación utilizando pocos ejemplos de entrenamiento. Durante las etapas de *meta-training* y *meta-testing*, generan ejemplos adicionales mediante una función de alucinación, lo que mejora significativamente el *few-shot learning*. Los experimentos demostraron que este enfoque es independiente del algoritmo de *meta-learning* utilizado y ofrece una mejora notable en la precisión de clasificación.

1.2. Ye y Zang [4]. abordan los retos de la segmentación semántica de imágenes médicas (*Few-Shot Segmentation, FSS*) en un contexto donde obtener etiquetas es costoso y complejo. Proponen un marco dinámico basado en autoentrenamiento y generación de etiquetas pseudo para mejorar la segmentación de imágenes médicas, con un enfoque especial en flujos de datos continuos. Los experimentos con imágenes de tomografía computarizada (CT) y resonancia magnética (MRI) mostraron mejoras significativas en comparación con otros métodos, destacando la efectividad del enfoque para evitar el colapso del modelo y mejorar la consistencia.

1.3. Kim et al. [5]. evalúan el uso de aprendizaje con pocos ejemplos para el diagnóstico temprano del glaucoma mediante imágenes *FUNDUS*. Proponen una arquitectura que combina redes de coincidencia con redes convolucionales de alta resolución, preservando la calidad de las imágenes originales y logrando resultados prometedores. Sus experimentos indicaron que este enfoque es adecuado para conjuntos de datos médicos reducidos y tiene potencial para aplicarse a otros tipos de enfermedades y análisis basados en texto.

1.4. Ho et al. [6]. proponen un modelo que se centra en la generación de datos en espacios continuos de alta dimensión utilizando un proceso de difusión estocástica. Durante cada paso del proceso introduce ruido, y un desruidizador guía la difusión hacia la distribución de datos objetivo. Los resultados preliminares mostraron que este enfoque supera a otros métodos generativos en términos de calidad y diversidad de muestras generadas, posicionándolo como una alternativa sólida en el campo de los modelos generativos.

1.5. Dhariwal y Nichol [7]. su estudio demuestra que los modelos de difusión superan a las GANs en la síntesis de imágenes. Se implementa la técnica de "guía del clasificador"para equilibrar la fidelidad y la diversidad en la síntesis condicional. Los experimentos realizados en *ImageNet* lograron métricas FID competitivas, destacando el potencial de estos modelos como una alternativa más robusta frente a las GANs.

1.6. Rombach et al. [8]. presentan un nuevo enfoque llamado Modelos de Difusión Latente (LDM), diseñado para superar las limitaciones de las GANs y de los modelos de difusión tradicionales en la generación de imágenes de alta resolución. Los LDM introducen variables latentes en cada paso del proceso de difusión, lo que mejora la calidad y la eficiencia. Además, se utiliza una técnica piramidal para manejar imágenes de alta resolución, logrando resultados excepcionales en términos de calidad de síntesis y eficiencia computacional.

1.7. Yao et al. [9]. en este estudio abordan el problema del desbalance de clases y la limitada disponibilidad de datos médicos para el diagnóstico de melanoma. Dado que el melanoma es responsable de la gran mayoría de las muertes por cáncer de piel a pesar de representar una fracción mínima de los casos, su detección temprana es crucial. Utilizando técnicas avanzadas de visión por computadora y aprendizaje profundo, el trabajo analiza métodos de clasificación y segmentación aplicados a imágenes de lesiones cutáneas del conjunto de datos ISIC 2016-2020.

Los investigadores implementaron técnicas como redes neuronales convolucionales preentrenadas, aprendizaje por transferencia y clasificadores avanzados para mejorar el rendimiento del modelo. El estudio también aborda el desbalance de clases mediante estrategias de preprocesamiento y generación de datos sintéticos, mitigando la falta de muestras positivas.

Capítulo 6

Método propuesto

1. Innovación Metodológica

1.0.1. Modelo de Difusión. El modelo seleccionado se basa en **redes de difusión**, siguiendo la arquitectura propuesta por Ho et al. [**6**]. Estos modelos generativos utilizan un proceso iterativo de adición y eliminación de ruido, permitiendo generar imágenes sintéticas de alta calidad al preservar detalles esenciales de los datos originales.

A diferencia de otros enfoques generativos, como las redes adversarias generativas (GANs), los modelos de difusión logran un equilibrio superior entre fidelidad y diversidad de las muestras generadas. Esta característica es especialmente relevante en el ámbito médico, donde la precisión y el realismo de los datos sintéticos son cruciales para tareas de diagnóstico.

El proceso de difusión se define mediante una cadena de Markov que agrega ruido gaussiano a los datos originales. La distribución conjunta de este proceso es:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{1}$$

donde cada paso se define como:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{2}$$

donde β_t es la varianza de ruido en el paso t, ajustada según una programación predefinida.

El modelo aprende una cadena de reversión para reconstruir la imagen original desde el ruido, definida como:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{3}$$

donde la media μ_θ y la covarianza Σ_θ son parámetros aprendidos por una red neuronal.

Para entrenar el modelo, se minimiza una cota variacional del logaritmo negativo de verosimilitud:

$$L = \mathbb{E}_q \left[D_{KL}(q(x_T|x_0) \,\|\, p(x_T)) + \sum_{t=2}^{T} D_{KL}(q(x_{t-1}|x_t, x_0) \,\|\, p_\theta(x_{t-1}|x_t)) \right] \tag{4}$$

La implementación del modelo se llevó a cabo utilizando el código fuente proporcionado por Wang en el repositorio de *lucidrains* [**10**], adaptado específicamente a los requerimientos del problema. La arquitectura empleada fue una **U-Net** configurada para generar características sintéticas a partir de imágenes médicas, asegurando así la fidelidad y calidad necesarias para el análisis y la evaluación posteriores.

1.1. Métricas de Evaluación. Se emplearon métricas específicas para evaluar la calidad y eficacia de los modelos de difusión. Entre estas métricas se incluyen la precisión (*accuracy*), *F1-score* y el *recall*, todas ellas ajustadas al contexto médico. Estas métricas aseguran que los datos generados sean relevantes y útiles para mejorar el rendimiento de los modelos predictivos.

1.2. Estrategia de Entrenamiento. Durante el entrenamiento del modelo de difusión, se utilizó un optimizador avanzado. Además, se implementó una estrategia de monitoreo continuo del rendimiento en un conjunto de validación, lo que facilitó el control de la calidad de las imágenes generadas y aseguró su capacidad de generalización.

1.3. Preprocesamiento de Datos. Dado que el trabajo se centra en imágenes médicas, se implementaron técnicas de normalización para garantizar que todas las imágenes tuvieran la misma escala de intensidad. Los datos se organizaron en *DataLoaders*, y se emplearon modelos preentrenados como *ResNet18* para extraer las características. Posteriormente se redimensionaron para utilizar de manera efectiva la capacidad del modelo de difusión.

1.4. Validación y Pruebas. Se utilizó un enfoque de *hold-out* para separar los datos en conjuntos de entrenamiento, validación y prueba. Esta división permitió entrenar el modelo en un conjunto de datos, validar su rendimiento de manera continua, y evaluar finalmente su capacidad predictiva en un conjunto de datos completamente independiente.

1.5. Diseño del Algoritmo. El algoritmo desarrollado tiene como objetivo principal la extracción de características (*features*) a partir de datos médicos, para posteriormente, realizar el procesamiento y la generación de características sintéticas, mejorando la representación de clases minoritarias y maximizando la eficacia de los modelos predictivos. Este enfoque trabaja a nivel de *features*, preservando las propiedades esenciales de los datos clínicos y reduciendo significativamente la complejidad computacional.

1.5.1. Pilares del Algoritmo. El enfoque del algoritmo se fundamenta en tres pilares principales:

- **Fidelidad:** Las características generadas son de alta calidad, capturando con precisión propiedades diagnósticas clave. Esto garantiza que los modelos entrenados sean clínicamente relevantes y confiables.
- **Diversidad:** El algoritmo genera un conjunto amplio y variado de características sintéticas, reflejando la heterogeneidad encontrada en las condiciones médicas reales. Esto asegura su aplicabilidad en múltiples diagnósticos y escenarios clínicos.
- **Eficiencia de Datos:** Diseñado para operar eficazmente con conjuntos de datos desbalanceados y reducidos, un escenario común en el ámbito médico debido a restricciones éticas, económicas y logísticas.

1.6. Perspectivas y Conclusión. El diseño del algoritmo prioriza la adaptabilidad y la escalabilidad, permitiendo su integración con distintos flujos de trabajo. Este enfoque promete avances significativos en la generación y procesamiento de características sintéticas a partir de datos médicos limitados, estableciendo una herramienta versátil para impulsar el desarrollo en el campo de la medicina. Estas innovaciones no solo facilitan el manejo de datos desbalanceados, sino que también contribuyen a la creación de modelos predictivos robustos y precisos, capaces de abordar desafíos clínicos complejos.

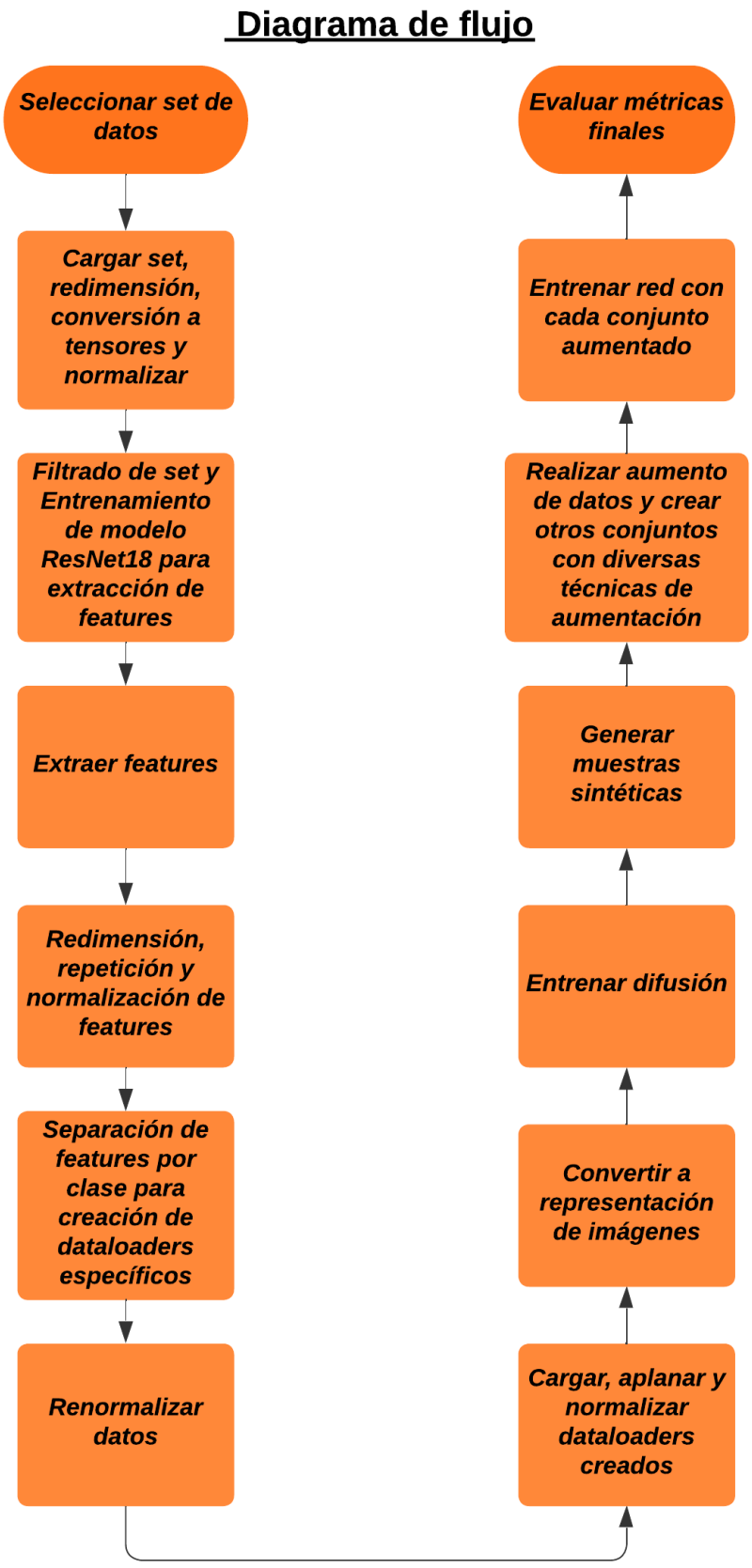

FIGURA 1. Diagrama de flujo del algoritmo completo

Algorithm 1 Preprocesamiento de Datos

function PREPROCESARDATOS(*datosOriginales*)
    Redimensionar imágenes
    Convertir imágenes a tensores
    Normalizar tensores
    Filtrar y crear DataLoaders
    return *datosProcesados*
end function

Algorithm 2 Extracción de Características

function EXTRAERCARACTERISTICAS(*datosProcesados*)
    Cargar modelo ResNet18
    Extraer características
    return *caracteristicas*
end function

Algorithm 3 Redimensionamiento de Características

function REDIMENSIONARCARACTERISTICAS(*caracteristicas*)
    Repetir características
    Redimensionar características para adaptarlas al modelo de difusión
    Normalizar características al rango adecuado
    return *caracteristicasListas*
end function

Algorithm 4 Separación de Clases

function SEPARARCLASES(*caracteristicasListas*)
    Separar features por clases
    Crear dataloaders
    return *caracteristicasSeparadasListas*
end function

Algorithm 5 Entrenamiento del Modelo de Difusión

function EntrenarModeloDifusion(*caracteristicasSeparadasListas*)
    Inicializar y configurar modelo de difusión
    Entrenar modelo usando las características separadas por clase
    return *modeloEntrenado*
end function

Algorithm 6 Generación de Muestras Sintéticas

function GenerarMuestrasSinteticas(*modeloEntrenado*)
    Utilizar el modelo entrenado para generar nuevas muestras sintéticas
    return *muestrasSinteticas*
end function

Algorithm 7 Realización del Aumento de Datos

function RealizarAumento(*muestrasSinteticas*)
    Realizar el aumento de los datos con las características sintéticas generadas
    Realizar otros aumentos con diversas técnicas comunes
    return *conjuntosAumentado*
end function

Algorithm 8 Evaluación del Modelo

function EvaluarModelo(*conjuntosAumentados*, *datosProcesados*)
    Comparar el conjunto aumentado contra otras técnicas de aumentación usando diversas métricas de evaluación y los datos originales
    return *resultadosEvaluacion*
end function

Algorithm 9 Algoritmo Principal: Generación y Evaluación de Datos Sintéticos

1: $datosOriginales \leftarrow$ Cargar imágenes originales
2: $datosProcesados \leftarrow$ PreprocesarDatos($datosOriginales$)
3: $caracteristicas \leftarrow$ ExtraerCaracteristicas($datosProcesados$)
4: $caracteristicasListas \leftarrow$ RedimensionarCaracteristicas($caracteristicas$)
5: $caracteristicasSeparadasListas \leftarrow$ SepararClases($caracteristicasListas$)
6: $modeloEntrenado \leftarrow$ EntrenarModeloDifusion($caracteristicasSeparadasListas$)
7: $muestrasSinteticas \leftarrow$ GenerarMuestrasSinteticas($modeloEntrenado$)
8: $conjuntosAumentado \leftarrow$ RealizarAumento($muestrasSinteticas$)
9: $resultadosEvaluacion \leftarrow$ EvaluarModelo($conjuntosAumentado, datosProcesados$)
10: **return** $resultadosEvaluacion$

Capítulo 7

Diseño de experimentos

1. Datos del estudio

Para el desarrollo del modelo basado en redes de difusión para la aumentación de características (features), se utilizó principalmente el conjunto de datos *Skin-data-Cancer*.[**11**] Este conjunto contiene imágenes categorizadas en dos clases principales: *Cancer* y *Non-Cancer*.

2. Análisis de los datos

Previo a la implementación del modelo, se llevó a cabo un análisis detallado de las características de los datos:

2.1. Distribución de etiquetas. El análisis reveló un claro desequilibrio en las clases, con 162 imágenes correspondientes a *Non-Cancer* y 42 imágenes a *Cancer*. Este desequilibrio destaca la importancia de utilizar técnicas avanzadas de aumentación para garantizar un balance adecuado en los datos de entrenamiento.

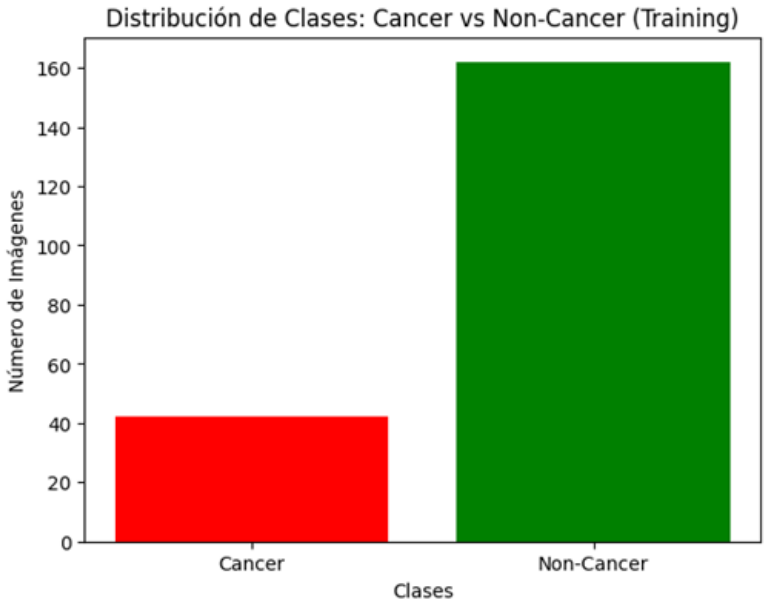

FIGURA 1. Gráfico de la distribución de las clases en Skin-data-Cancer

2.2. Estadísticas descriptivas. Se calcularon estadísticas básicas de las intensidades de los píxeles en las imágenes:

- **Media:** 103.60, indicando que las imágenes presentan, en promedio, una luminosidad moderada-baja.
- **Mediana:** 100.0, confirmando una tendencia hacia tonos oscuros.
- **Desviación estándar:** 49.79, mostrando una variabilidad moderada en las intensidades.

```
Media de Intensidad de Píxeles (Train): 103.60306932138725
Mediana de Intensidad de Píxeles (Train): 100.0
Desviación Estándar de Intensidad de Píxeles (Train): 49.79454652187412
Valor Mínimo de Intensidad de Píxeles (Train): 2
Valor Máximo de Intensidad de Píxeles (Train): 255
```

FIGURA 2. Estadísticas básicas de intensidades de píxeles

2.3. Detección de valores atípicos. Se identificaron 6,325 valores atípicos en las intensidades de los píxeles, lo cual es esperable en imágenes médicas debido a su alta variabilidad. Estos valores, aunque limitados, no se consideraron como un impedimento significativo para el rendimiento del modelo.

```
Número de valores atípicos detectados: 6325
```

FIGURA 3. Número de valores atípicos detectados

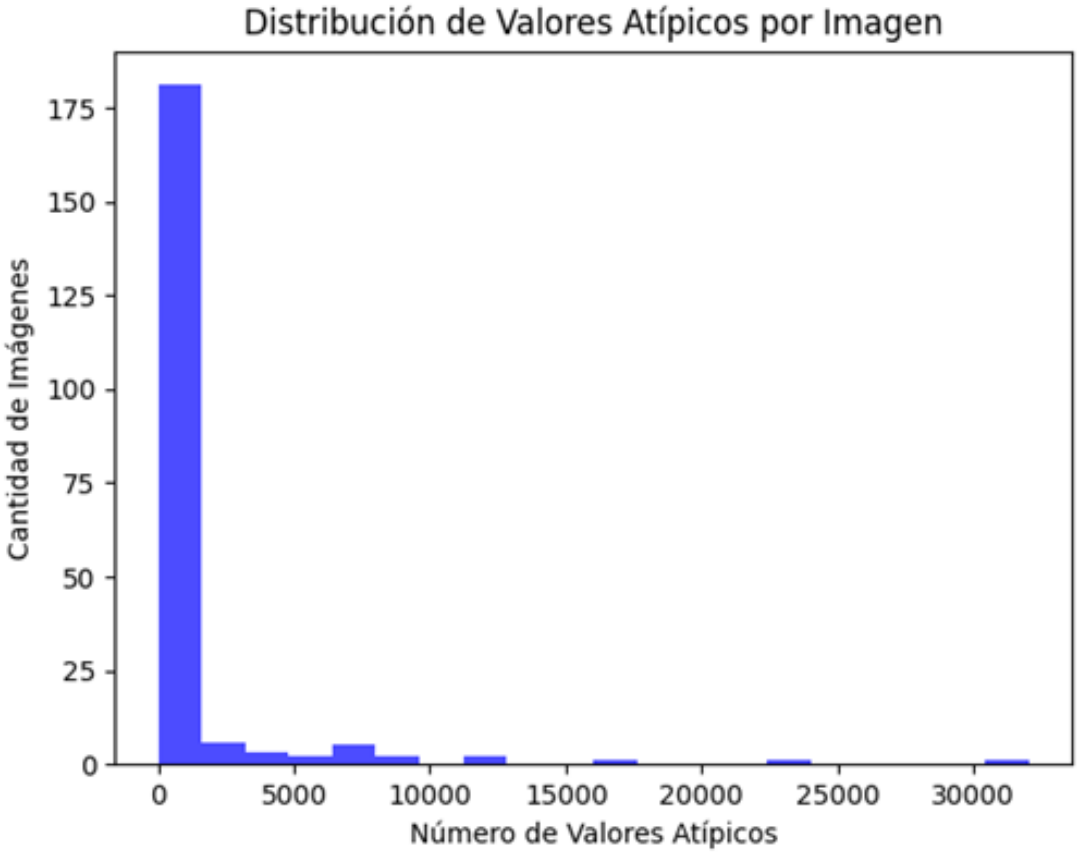

FIGURA 4. Gráfico de la distribución de valores atípicos por imagen

2.4. Visualización y análisis de asimetría. El histograma de intensidades mostró una concentración predominante en valores entre 0 y 150, reflejando una ligera asimetría hacia la derecha y una distribución menos centralizada, como lo evidencia la kurtosis negativa. Esto destaca la necesidad de una normalización adecuada antes del entrenamiento del modelo.

FIGURA 5. Histograma de intensidades de píxeles

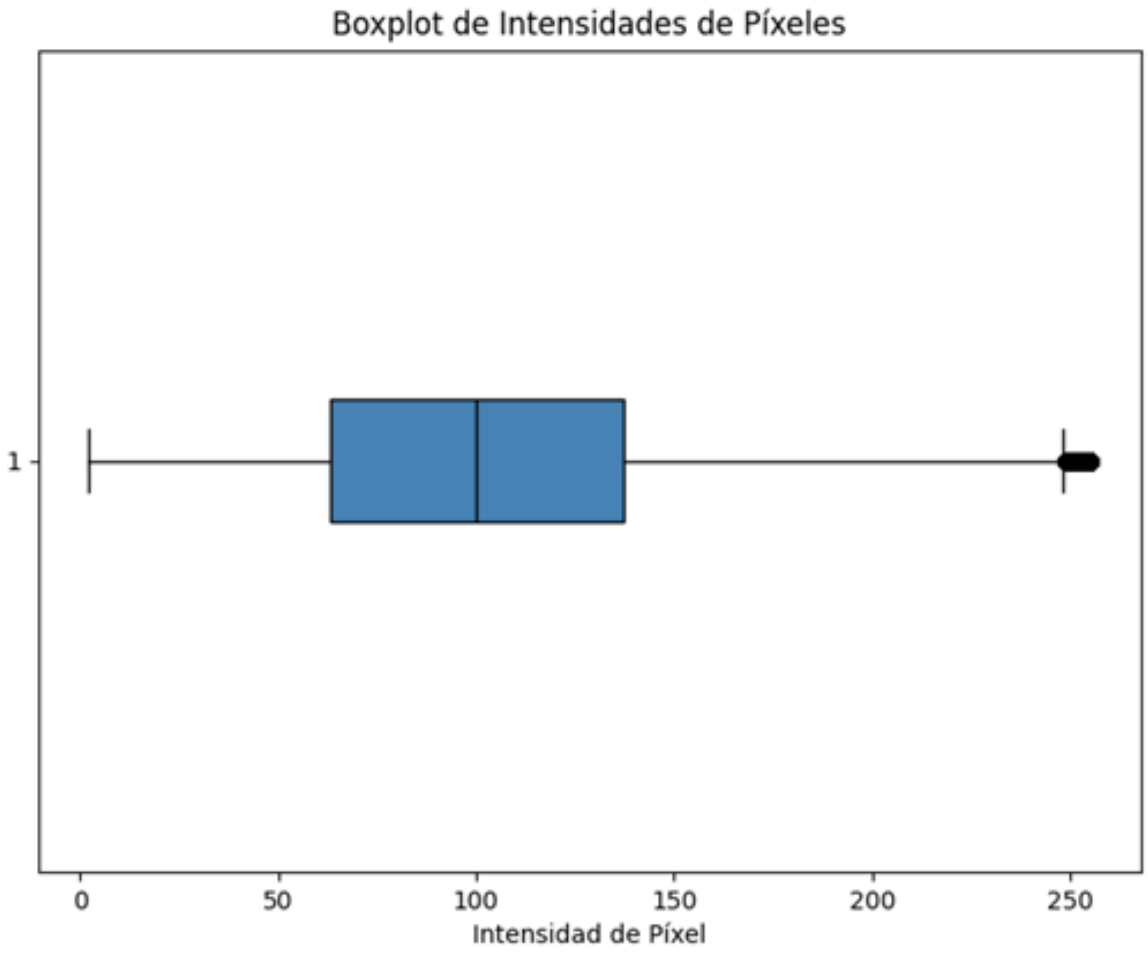

FIGURA 6. Boxplot de intensidades de píxeles

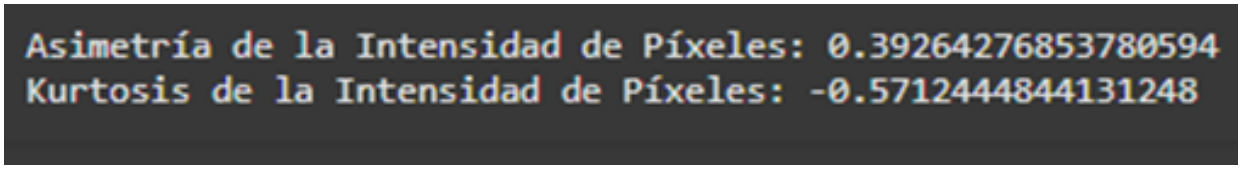

FIGURA 7. Análisis de asimetría y de kurtosis

3. Diseño experimental

El diseño experimental se estructuró de la siguiente manera:

- **Extracción de características:** Se utilizó un modelo *ResNet18* preentrenado, ajustando su arquitectura al eliminar la última capa para obtener las features de las imágenes. Estas se normalizaron y convirtieron en tensores para garantizar su compatibilidad con los algoritmos posteriores. El modelo fue entrenado durante 30 épocas para asegurar una extracción robusta.
- **Preprocesamiento de los datos:** Las imágenes se redimensionaron y normalizaron, manteniendo su integridad. Adicionalmente, se realizaron pruebas para verificar la correcta extracción y distribución de las características generadas.
- **Entrenamiento del modelo de difusión:** Se entrenó un modelo de difusión para generar imágenes de características sintéticas de alta calidad, con un total de 400 iteraciones por clase. El modelo fue monitoreado continuamente para garantizar una convergencia estable y eficiente.
- **Aumentación de características:** Las features generadas para la clase minoritaria (*Cancer*) se utilizaron para balancear el conjunto de datos. Paralelamente, se aplicaron técnicas de aumentación controlada a las características originales, evaluando el impacto en el aprendizaje y comparando los resultados con el modelo propuesto.
- **Validación del modelo:** Se implementó una técnica de *hold-out* para dividir los datos en conjuntos de entrenamiento, validación y prueba. Esto permitió evaluar el rendimiento del modelo en datos no vistos y asegurar su capacidad de generalización.

Capítulo 8

Experimentos

1. Configuración de Parámetros

1.1. Extracción de características. El proceso de extracción de características se llevó a cabo utilizando el modelo ResNet18 preentrenado. A continuación, se detalla el ajuste realizado:

- Se eliminó la última capa del modelo para obtener las características (*features*) de las imágenes.
- Las características fueron normalizadas y convertidas en tensores para garantizar su compatibilidad con los algoritmos posteriores.
- El número de épocas utilizado para el entrenamiento fue de 30, determinado tras pruebas preliminares que mostraron que este valor proporcionaba una extracción robusta sin sobreajuste.

1.2. Preprocesamiento. El preprocesamiento de las imágenes incluyó los siguientes pasos:

- Redimensionamiento de las imágenes a un tamaño uniforme, asegurando que las proporciones se mantuvieran constantes.
- Normalización de los valores de píxeles para mejorar la estabilidad durante el entrenamiento.

1.3. Visualización de datos. Se revisan las imagenes para asegurar la efectiva carga y preprocesamiento de los datos, en ellas se puede observar la complejidad del problema al buscar intentar diferenciar entre clases.

FIGURA 1. Conjunto de imágenes de ambas clases en Skin-data-Cancer.

FIGURA 2. Conjunto de imágenes de ambas clases en Skin-data-Cancer.

2. Entrenamiento del Modelo de Difusión

El entrenamiento del modelo de difusión se llevó a cabo utilizando un esquema basado en la arquitectura UNet. A continuación, se describen los principales componentes y configuraciones:

- **Arquitectura:**
  - Se empleó un modelo UNet, ampliamente utilizado en tareas de generación de imágenes.
  - El modelo incluyó capas convolucionales y bloques de atención para garantizar la calidad de las imágenes generadas.
  - La configuración del modelo fue la siguiente:
    - Dimensión base (`dim`): 64.
    - Multiplicadores de dimensión (`dim_mults`): (1, 2, 4, 8).
    - Tamaño de las imágenes generadas: 16x16.
    - Número de pasos de difusión (`timesteps`): 100.

- **Hiperparámetros de Entrenamiento:**
  - Tasa de aprendizaje: 1×10^{-4}.
  - Tamaño del lote: 32.
  - Número de iteraciones: 400 por clase.

- **Evaluación de la Convergencia:**
  - Se monitoreó la pérdida en cada iteración para asegurar la estabilidad y calidad del modelo.

3. Aumentación de Características

Para abordar el desbalance de clases en el conjunto de datos, se implementó la técnica de aumentación propuesta, asegurando un balance en las clases y evaluando su impacto en el modelo final. Se volvieron a convertir las imágenes de features obtenidas por la difusión y los conjuntos de cada clase se combinaron. Posteriormente, se realizaron aumentos con otros métodos para evaluar su rendimiento en comparación a la técnica propuesta en la investigación. A continuación, se describen las metodologías aplicadas:

3.1. Generación de Características Sintéticas con Modelos de Difusión. Se empleó un modelo de difusión para generar imágenes de características sintéticas, utilizando las siguientes especificaciones y procedimientos:

- **Arquitectura del modelo:** Se utilizó una arquitectura `UNet` con capas convolucionales y bloques de atención, configurada para procesar imágenes de tamaño $16 \times 16 \times 3$. Estas imágenes se generaron a partir de las características originales del conjunto de datos.

- **Entrenamiento del modelo:** Posteriormente el modelo fue entrenado con los siguientes parámetros anteriormente expuestos:
  - Tasa de aprendizaje: 1×10^{-4}.
  - Número de iteraciones: 400.
  - Tamaño del lote: 32.

- **Generación de imágenes sintéticas:**
  - El modelo de difusión generó imágenes sintéticas, siguiendo un proceso iterativo de 100 pasos para alcanzar convergencia en cada muestra.

- **Conversión de imágenes a características:**
  - Las imágenes generadas por difusión se convirtieron para llevarlas a su forma original.
  - Cada imagen generada fue reconvertida a un vector de características de dimensión 512, utilizando un procedimiento de aplanamiento que preservó la estructura y propiedades de las características originales.

- **Integración al conjunto de datos:**
  - Las características sintéticas generadas fueron combinadas con las características originales de la clase minoritaria, incrementando el número total de muestras en dicha clase.
  - El conjunto resultante fue mezclado y utilizado para entrenar un modelo final, garantizando un balance entre las clases.

3.2. Aumentación con Ruido Gaussiano. Adicionalmente, dentro de los modelos a comparar, se implementó la adición de ruido gaussiano a las características originales de la clase minoritaria:

- Se generaron perturbaciones controladas con media $\mu = 0{,}0$ y desviación estándar $\sigma = 0{,}1$.
- Estas perturbaciones se aplicaron para aumentar la variabilidad en los datos de la clase minoritaria, reduciendo la dependencia en las muestras originales.

3.3. Replicación de Muestras. Para evaluar el impacto de técnicas más simples, se realizó replicación directa de las características de la clase minoritaria:

- Las muestras existentes fueron duplicadas hasta igualar el número de muestras de la clase mayoritaria.
- Esta técnica permitió un balance rápido del conjunto de datos sin agregar ruido adicional.

3.4. Aumentación con Ruido Gaussiano Escalado. Se utilizó ruido gaussiano escalado para generar características sintéticas de la clase minoritaria:

- Las características originales de la clase minoritaria fueron replicadas y perturbadas añadiendo ruido gaussiano con desviación estándar de 0,01.

- Se generaron suficientes muestras para igualar la cantidad de la clase mayoritaria y las características aumentadas se combinaron con las originales, creando un conjunto balanceado.

4. Validación del Modelo

La validación del modelo se realizó utilizando una técnica de *hold-out*, con las siguientes proporciones:

- 80 % para el entrenamiento (del conjunto de train).
- 20 % para la validación (del conjunto de train).
- 100 % para la prueba (del conjunto de test).

5. Evaluación y Resultados, 1er set de datos

Todas las técnicas de aumentación fueron evaluadas en el conjunto de pruebas original mediante las siguientes métricas:

- Accuracy
- Precisión
- Recall
- F1 Score
- Matriz de Confusión

Los resultados comparativos entre los datos originales y los datos aumentados se presentan en la Tabla 1.

Tabla 1. Comparación de desempeño entre datos originales y técnicas de aumentación de set 1

Dataset	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Original - Modelo Original	0.599	0.728	0.599	0.532
Original - Datos de Difusión	0.747	0.766	0.747	0.740
Original - Replicación	0.643	0.746	0.643	0.586
Original - Ruido Escalado	0.677	0.748	0.677	0.642
Original - Ruido Gaussiano	0.728	0.766	0.728	0.712

5.1. Análisis Comparativo. La Tabla 1 muestra los resultados promedio obtenidos tras evaluar en 10 ocasiones las diferentes técnicas de aumentación de datos aplicadas al conjunto de datos original. Para evaluar el impacto del modelo de difusión, se analizarán las cuatro métricas clave, cada una con un significado relevante para la clasificación:

- **Accuracy (Exactitud):** Representa el porcentaje de predicciones correctas realizadas por el modelo. En este caso, la técnica de difusión logró una *Accuracy* de 0.747, indicando que casi el 75 % de las muestras fueron correctamente clasificadas.
- **Precision (Precisión):** Indica la proporción de muestras correctamente clasificadas como positivas frente a todas las muestras clasificadas como positivas. Una precisión alta significa pocos falsos positivos. El modelo de difusión obtuvo una precisión de 0.766, significativamente mejor que el modelo original (0.728), mostrando mayor confianza en sus predicciones positivas.
- **Recall (Sensibilidad):** Indica la capacidad del modelo para identificar correctamente todas las muestras positivas. El modelo basado en difusión logró un *Recall* de 0.747, lo que implica una adecuada identificación de muestras relevantes, especialmente respecto al original de 0.599.
- **F1 Score (Puntaje F1):** Combina precisión y recall en un solo valor. Es especialmente útil cuando existe un desbalance en las clases, ya que evita sesgos hacia una sola métrica. El modelo de difusión alcanzó un *F1 Score* de 0.740, un aumento notable frente al modelo original (0.532), mostrando una mejora equilibrada en ambas métricas.

El uso del modelo de difusión demostró ser una técnica efectiva para la generación de datos sintéticos, superando a otros métodos tradicionales de aumento de datos. El análisis de las métricas muestra que este enfoque es capaz de generar características sintéticas de alta calidad, mejorando significativamente el rendimiento del modelo de clasificación en términos de exactitud, precisión, recall y F1 Score. Estos resultados validan el potencial de los modelos de difusión para aplicaciones en el ámbito médico, donde la disponibilidad de datos suele ser limitada.

6. Evaluación y Resultados, 2do set de datos

Se realizó el mismo procedimiento de evaluación en 10 ocasiones de las métricas con los modelos en otro set de datos de mayor volumen llamado *Skin Cancer*.[**12**] Este set de datos igualmente contiene imágenes de melanomas, dando un total de 4.094 imágenes , en donde se clasifican como *benign* y *malignant*.

Tabla 2. Comparación de desempeño entre datos originales y técnicas de aumentación de set 2.

Dataset	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Original - Modelo Original	0.744	0.863	0.744	0.768
Original - Datos de Difusión	0.791	0.867	0.791	0.809
Original - Replicación	0.749	0.864	0.749	0.773
Original - Ruido Escalado	0.735	0.863	0.735	0.761
Original - Ruido Gaussiano	0.754	0.864	0.754	0.777

6.1. Análisis Comparativo. La Tabla 2 muestra los resultados promedio obtenidos tras aplicar diferentes técnicas de aumentación de datos al modelo base. Se evaluaron las mismas cuatro métricas esenciales para verificar el rendimiento del modelo: *Accuracy*, *Precision*, *Recall* y *F1 Score*.

- **Accuracy (Exactitud):** El mejor rendimiento se observó con los datos generados mediante técnicas de difusión, logrando una *Accuracy* de 0.791, mejorando notablemente respecto al modelo original (0.744), con un altísimo porcentaje de predicciones correctas realizadas.
- **Precision (Precisión):** El modelo con datos de difusión alcanzó una precisión de 0.867, superando ligeramente los otros modelos. Esto que sugiere una reducción de falsos positivos.
- **Recall (Sensibilidad):** El modelo con datos de difusión logró un *Recall* de 0.791, superando el 0.744 del modelo original, indicando una mejor detección de muestras positivas.
- **F1 Score (Puntaje F1):** El *F1 Score* alcanzó 0.809, frente a 0.768 del modelo original, mostrando un mejor equilibrio entre precisión y sensibilidad.

En esta tabla, el aumento de los datos mediante difusión demostró ser efectiva para mejorar el rendimiento del modelo. Las métricas clave resaltaron su capacidad para generar características sintéticas de alta calidad, esto sugiere su potencial y efectiva aplicabilidad en tareas de clasificación en contextos médicos con datos limitados.

7. Evaluación y Resultados, 3er set de datos

De igual manera, se realiza el mismo procedimiento de evaluación en 10 ocasiones de las métricas con los modelos en otro set de datos, siendo este el de mayor volumen, llamado *Skin Cancer Test.*[**13**] Este set de datos igual a los dos anteriores contiene imágenes de melanomas, dando un total de 10.605 imágenes , en donde se clasifican como *benign* y *malignant.*

Tabla 3. Comparación de desempeño entre datos originales y técnicas de aumentación de set 3.

Dataset	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Original - Modelo Original	0.853	0.856	0.853	0.852
Original - Datos de Difusión	0.854	0.856	0.854	0.854
Original - Replicación	0.852	0.855	0.852	0.851
Original - Ruido Escalado	0.852	0.855	0.852	0.851
Original - Ruido Gaussiano	0.853	0.855	0.853	0.852

7.1. Análisis Comparativo. La Tabla 3 muestró resultados los promedio obtenidos tras aplicar las diferentes técnicas de aumentación de datos al modelo base. Se evaluaron las mismas cuatro métricas esenciales para verificar el rendimiento del modelo: *Accuracy*, *Precision*, *Recall* y *F1 Score.*

- **Accuracy (Exactitud):** El mejor rendimiento se observó con los datos generados mediante técnicas de difusión, logrando una *Accuracy* de 0.854, superando ligeramente los otros modelos.
- **Precision (Precisión):** Tanto el modelo original como el modelo con datos de difusión alcanzaron una precisión de 0.856, mostrando que ambas configuraciones lograron minimizar los falsos positivos de manera efectiva.
- **Recall (Sensibilidad):** El modelo con datos de difusión también logró un *Recall* de 0.854, una ligera mejora frente al modelo original (0.853), indicando una mejora en la detección de muestras positivas.
- **F1 Score (Puntaje F1):** El *F1 Score* más alto también se obtuvo con los datos de difusión (0.854), superando el 0.852 del modelo original. Esto evidencia un equilibrio ligeramente superior entre precisión y sensibilidad en este caso.

En esta tabla, el aumento de los datos mediante difusión demostró ser efectivo para mantener e incluso mejorar ligeramente el rendimiento del modelo. Las métricas clave mostraron que las características sintéticas generadas por esta técnica lograron un equilibrio sólido entre precisión y sensibilidad, lo que establece su potencial aplicabilidad en tareas de clasificación.

8. Imágenes de features

Se obtuvieron las representaciones en imágenes de las features de cada modelo y se compararon con respecto a las originales.

Se observaron claras diferencias entre clases, esto se repitió en cada modelo, sin demasiadas variaciones.

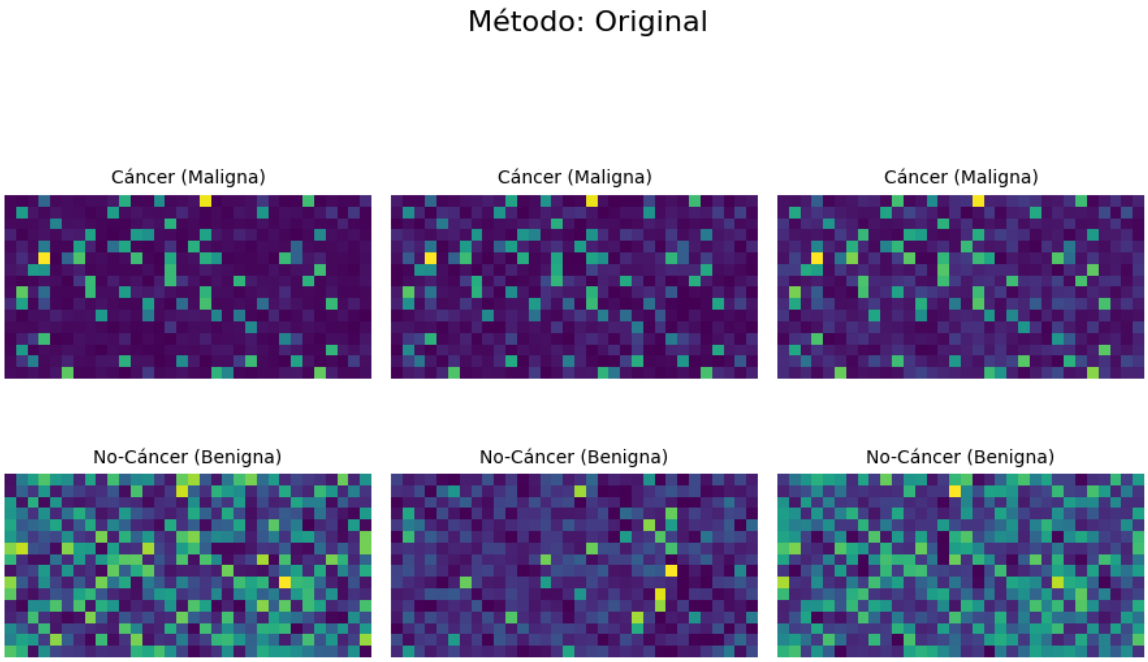

FIGURA 3. Visualización de features. Método Original.

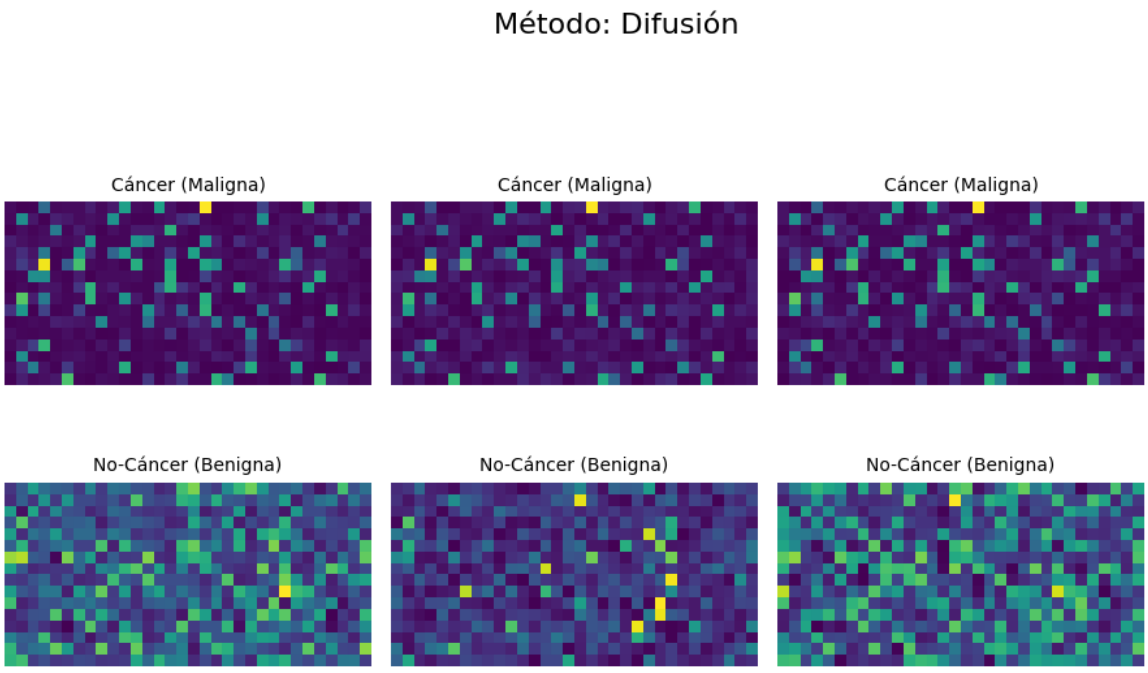

FIGURA 4. Visualización de features. Método Difusión.

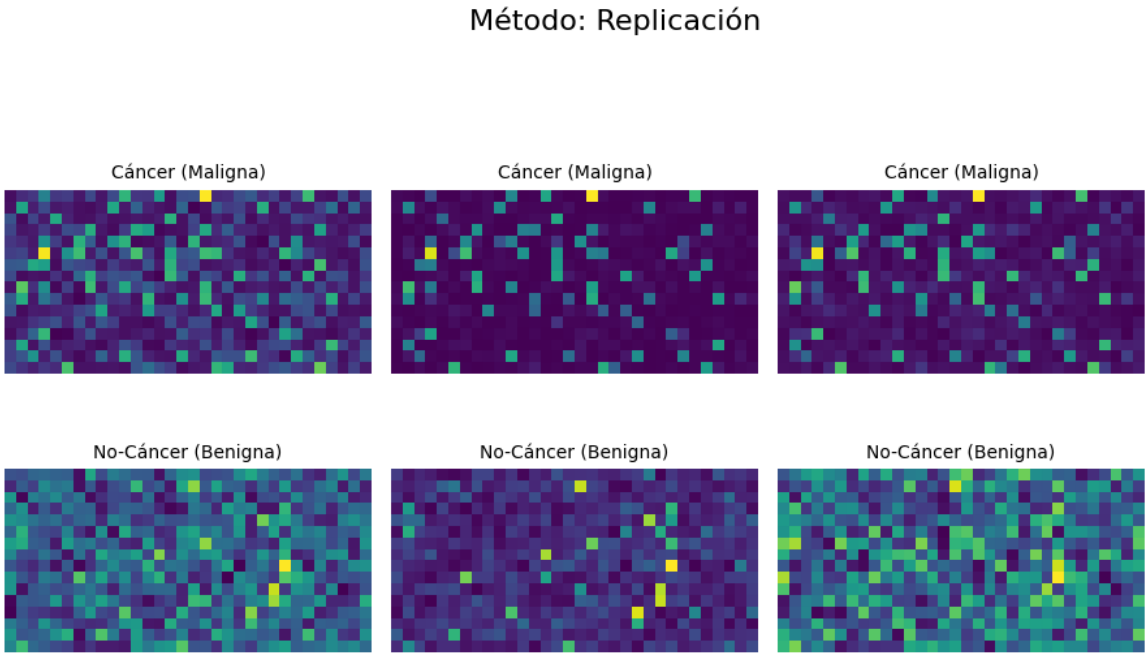

FIGURA 5. Visualización de features. Método Replicación

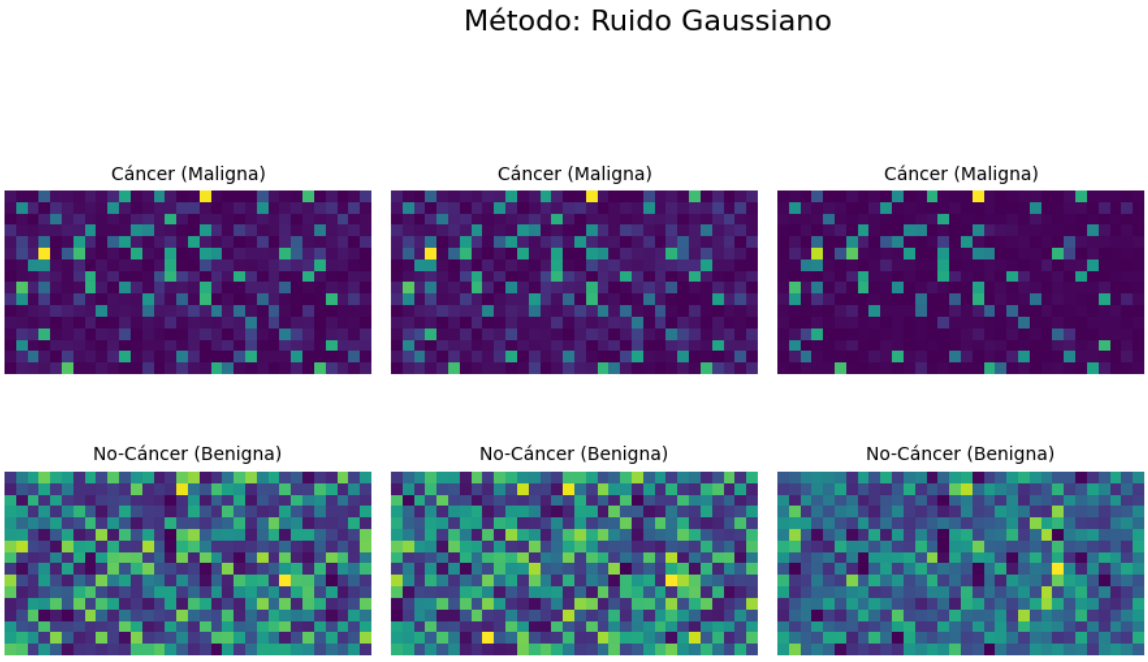

FIGURA 6. Visualización de features. Método Ruido Gaussiano.

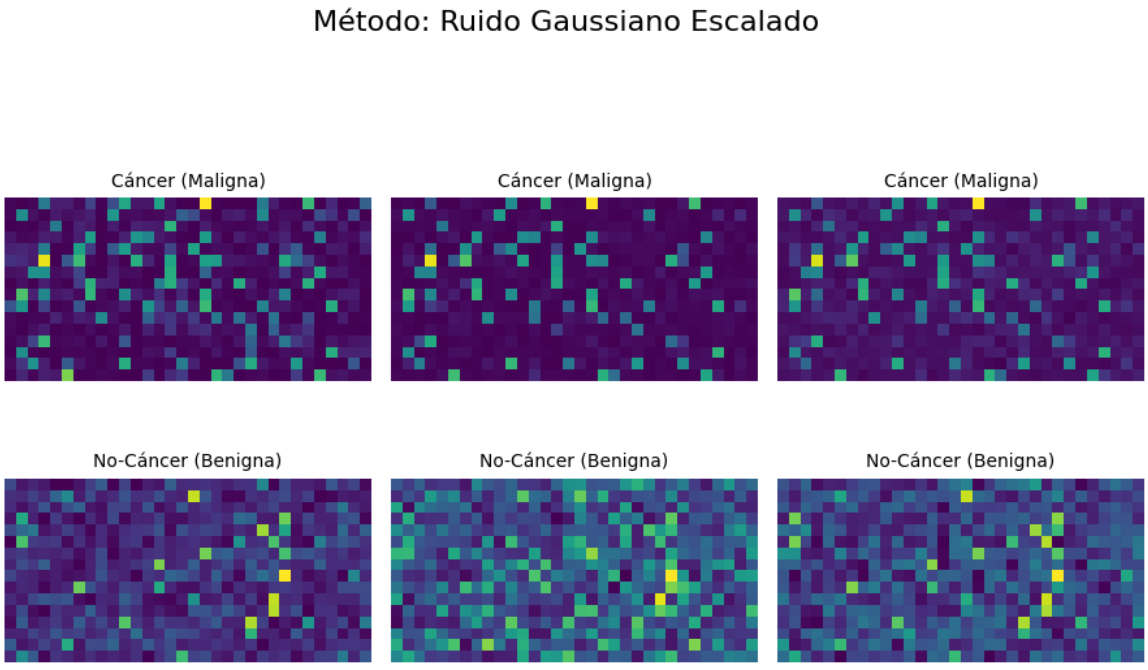

FIGURA 7. Visualización de features. Método Ruido Gaussiano Escalado.

Capítulo 9

Conclusiones

A lo largo de toda la investigación, experimentación y análisis de los resultados, se pudieron destacar varios aspectos en diversos puntos:

- **Importancia de la extracción de características robustas:** Utilizar un modelo ResNet18 preentrenado permitió extraer características relevantes de las imágenes del conjunto de datos.

- **Eficiencia del modelo de difusión para generación de datos sintéticos:** El modelo de difusión basado en una arquitectura `UNet` demostró ser eficaz para generar características sintéticas. Estas características se integraron con éxito al conjunto de datos, mejorando la representación de las clases minoritarias y, en consecuencia, el desempeño del modelo de clasificación.

- **Validación exhaustiva del modelo:** El uso de métricas como *accuracy*, precisión, *recall* y F1 Score evidenció mejoras en el rendimiento al incorporar datos sintéticos. En particular, los datos generados mediante difusión alcanzaron un balance óptimo entre estas métricas, destacándose como una técnica efectiva para el balance de datos en problemas de clasificación desbalanceada.

- **Aportaciones al estado del arte:** Este trabajo contribuye al área de aprendizaje profundo, específicamente en el ámbito de las técnicas de aumentación de datos a nivel de características. La integración de modelos de difusión para generar características sintéticas ofrecen una solución innovadora para abordar el desbalance de clases en conjuntos de datos. Los resultados obtenidos demuestran que estas técnicas no solo mejoran sustancialmente el desempeño en tareas de clasificación, sino que también amplían las posibilidades de aplicación en problemas complejos, donde el acceso a datos balanceados y representativos es limitado.

- **Limitaciones identificadas y posibles extensiones:**
  - **Limitaciones:** El modelo se probó por ahora lo suficiente solo en dos conjuntos de datos. Aunque los resultados son prometedores, es necesario validar la generalización del método en otros dominios.
  - **Extensiones futuras:** Se pueden explorar varios enfoques, tales como explorar modelos de difusión más avanzados, probar nuevos ajustes en los parámetros y extender los experimentos a conjuntos más diversos.

En conclusión, la investigación demuestra que integrar modelos de difusión para la generación y aumentación de características sintéticas es una estrategia efectiva para abordar problemas de desbalance de clases, mejorando la precisión y robustez de los modelos de clasificación.

Con el propósito de promover la replicabilidad y facilitar el desarrollo de investigaciones futuras, el código fuente y los experimentos realizados durante este trabajo se encuentran disponibles públicamente en el siguiente repositorio de GitHub: Repositorio de la Tesis MCC.

Bibliografía

[1] Neel Kanwal, Roger Amundsen, Helga Hardardottir, Luca Tomasetti, Erling Sandoy Undersrud, Emiel A.M. Janssen, and Kjersti Engan. Detection and localization of melanoma skin cancer in histopathological whole slide images. *arXiv preprint arXiv:2302.03014*, 2023. Accepted at EUSIPCO 23.

[2] Mohamad Taghizadeh and Karim Mohammadi. The fast and accurate approach to detection and segmentation of melanoma skin cancer using fine-tuned yolov3 and segnet based on deep transfer learning. *arXiv preprint arXiv:2210.05167*, 2023. Last revised 11 Jan 2023, v2.

[3] Y.-X. Wang, R. Girshick, M. Hebert, and B. Hariharan. Low-shot learning from imaginary data. In *Facebook AI Research (FAIR), Carnegie Mellon University, Cornell University*, 2018.

[4] Z. Ye and W. Zhang. A dynamic few-shot learning framework for medical image stream mining based on self-training. *N/A*, 2023.

[5] M. Kim, J. Zuallaert, and W. De Neve. Few-shot learning using a small-sized dataset of high-resolution fundus images for glaucoma diagnosis. *N/A*, 2017.

[6] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *N/A*, 2020.

[7] P. Dhariwal and A. Nichol. Diffusion models beat gans. *N/A*, 2021.

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. *N/A*, 2022.

[9] Chengdong Yao. A comprehensive evaluation study on risk level classification of melanoma by computer vision on isic 2016-2020 datasets. *arXiv preprint*, arXiv:2302.09528, 2023. 9 pages, 12 figures, 11 tables.

[10] Phil Wang. Denoising diffusion pytorch implementation. `https://github.com/lucidrains/denoising-diffusion-pytorch/tree/main`, 2024. Accessed: 16-Dec-2024.

[11] Fares Abbasai. Skin data cancer. https://www.kaggle.com/datasets/faresabbasai2022/skin-data-cancer, 2022. Accessed: 2024-11-29.

[12] malicks111. Skin cancer dataset. `https://www.kaggle.com/datasets/malicks111/skin-cancer`. Accessed: 16-Dec-2024.

[13] sherifabdelzaher. Skin cancer test. `https://www.kaggle.com/datasets/sherifabdelzaher/skin-cancer-test/data`. Accessed: 16-Dec-2024.