UNIVERSIDAD NACIONAL DE MOQUEGUA

FACULTAD DE INGENIERÍAS

INGENIERÍA DE SISTEMAS E INFORMÁTICA

TESIS:

TÉCNICAS DE BALANCE DE DATOS PARA OPTIMIZAR EL RENDIMIENTO DE MODELOS DE APRENDIZAJE PROFUNDO EN CLASIFICACIÓN DE RADIOGRAFÍAS TORÁCICAS

Presentado Por:

Estefany Paola Mamani Gutierrez

Asesor:

Mgr. Clares Perca Juan Carlos

Co-Asesor:

Ing. Mamani Padilla Diego Ismael

Para optar:

Título profesional de Ingeniero de Sistemas e Informática

Moquegua, Diciembre de 2025

NAL DE
MÁTICA

UNAM

ESCUELA PROFESIONAL DE
INGENIERÍA DE SISTEMAS E INFORMÁTICA

Nº 024

Folio Nº 025

ACTA DE SUSTENTACIÓN

I. Modalidad:

Tesis : [X]
Trabajo de Suficiencia Profesional : [ ]
Otro: ..... : [ ]

II. Documento de aprobación para la sustentación:
Resolución de Consejo de Facultad N° 291-2025-CF-FAIN-UNAM
Fecha: 10/Diciembre/2025

TULADO:

III. El jurado ha dictaminado la (el) Tesis, TITULADO:
"Técnicas de balance de datos para optimizar el rendimiento de Modelos de Aprendizaje Profundo en Clasificación de Radiografías Torácicas"

Bachiller: Estefany Pada Mamani Gutierrez
Asesor Principal: Juan Carlos Clares Perca
Co-Asesor: Diego Ismael Mamani Padilla

era:

IV. Se dictamina:
Como: Aprobado por Unanimidad la sustentación de manera:
Presencial [X] Virtual [ ]

]

Con nota: Catorce (14) por lo que; Sí [X] No [ ]
Opta el Título Profesional de INGENIERO DE SISTEMAS E INFORMÁTICA .

la firma

Siendo las 12:13 horas, del 18 de diciembre del año 2025 se procede con la firma en conformidad de lo dictaminado por el Jurado.

Presidente
Euler Tito Chura

Primer Miembro
Victor Yana Mamani

Segundo Miembro
Saul Huaquipaco Encinas

Estefany Paola Mamani Gutierrez
Bachiller

Rojo A

01

Dedicatoria

Dedico la presente tesis a mis padres, mi mayor inspiración, por su apoyo incondicional, su comprensión constante y por inculcar en mí el deseo permanente de superación.

A mi hermano, por su compañía y apoyo en los momentos de mayor esfuerzo.

A Pepa, por su constante compañía y el apoyo emocional brindado durante las largas jornadas de trabajo y estudio.

Y a mi compañero de estudio y de vida, por el acompañamiento, el impulso constante y la motivación para exigirme siempre un poco más.

Agradecimiento

Expreso mi sincero agradecimiento a los docentes de la Universidad Nacional de Moquegua, por la formación académica brindada a lo largo de mi carrera profesional y por los conocimientos transmitidos, los cuales constituyeron una base fundamental para el desarrollo de la presente investigación.

De manera especial, agradezco a mi asesor de tesis, por su orientación metodológica, sus observaciones oportunas y su disposición permanente para guiar el proceso investigativo, contribuyendo de manera significativa al rigor y la coherencia de este trabajo.

Asimismo, expreso un agradecimiento particular a mi coasesor, por su constante apoyo, acompañamiento cercano y seguimiento continuo, estando siempre atento al avance de la investigación y brindando orientación en los momentos clave del proceso, lo cual fue determinante para la culminación de esta tesis.

Resumen

En el ámbito de la medicina, los conjuntos de datos suelen presentar un desequilibrio de clases, afectando negativamente el rendimiento de los modelos de aprendizaje profundo al priorizar las clases mayoritarias y reducir la detección de enfermedades menos frecuentes. El presente estudio tuvo como objetivo mejorar el rendimiento de los modelos de clasificación de imágenes radiográficas torácicas mediante técnicas de balanceo de datos. Para abordar este problema, se evaluaron diversas estrategias bajo un diseño cuasi-experimental. Primero, se utilizó una **red generativa antagónica basada en la distancia de Wasserstein con penalización del gradiente** (WGAN-GP) para generar datos sintéticos y equilibrar la distribución de clases.

Además, se implementaron funciones de pérdida sensibles a costos como: pérdida focal (*focal loss)*, **entropía cruzada ponderada y balanceada** (BWCCE) y **pérdida de margen consciente de la distribución de etiquetas** (LDAM); y métodos híbridos para mejorar el aprendizaje de las clases minoritarias. El impacto se midió utilizando métricas de sensibilidad, especificidad y **área bajo la curva de característica operativa del receptor** (AUC-ROC), priorizando la capacidad discriminativa sobre la exactitud global. Los resultados demostraron que el modelo entrenado con datos generados por WGAN-GP (Escenario 5) fue el más efectivo, logrando un AUC-ROC de 0.9837. Esta técnica mejoró la especificidad (detección de la clase minoritaria) del 84.81% (línea base) al 91.77%, manteniendo una alta sensibilidad del 95.33% para la detección de neumonía.

Se concluye que el balanceo a nivel de datos mediante **red generativa antagónica** (GAN) es superior a las técnicas algorítmicas e híbridas en este contexto, logrando imágenes sintéticas de alta fidelidad con una **distancia de incepción de Fréchet** (FID) de 48.95 que mitigan el sesgo hacia la clase mayoritaria y optimizan la confiabilidad del diagnóstico asistido.

Palabras clave: Aprendizaje profundo, Redes generativas antagónicas; Wasserstein GAN con penalización de gradiente (WGAN-GP), Clasificación de imágenes médicas, Desbalance de clases, Radiografías torácicas.

Abstract

In the medical field, datasets often exhibit class imbalance, which negatively affects the performance of deep learning models by prioritizing majority classes and reducing the detection of less frequent diseases. The present study aimed to improve the performance of chest X-ray image classification models using data balancing techniques. To address this issue, various strategies were evaluated under a quasi-experimental design. First, a Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP) was used to generate synthetic data and balance the class distribution.

Additionally, cost-sensitive loss functions such as Focal Loss, Balanced-Weighted Cross-Entropy (BWCCE), and Label-Distribution-Aware Margin (LDAM) loss were implemented, as well as hybrid methods to enhance minority class learning. The impact was measured using sensitivity, specificity, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) metrics, prioritizing discriminative capacity over overall accuracy. The results demonstrated that the model trained with WGAN-GP generated data (Scenario 5) was the most effective, achieving an AUC-ROC of 0.9837. This technique improved specificity (minority class detection) from 84.81% (baseline) to 91.77%, while maintaining a high sensitivity of 95.33% for pneumonia detection.

It is concluded that data-level balancing using Generative Adversarial Networks (GAN) is superior to algorithmic and hybrid techniques in this context, achieving high-fidelity synthetic images with a Fréchet Inception Distance (FID) of 48.95 that mitigate bias toward the majority class and optimize the reliability of assisted diagnosis.

Keywords: Deep learning, Generative adversarial networks; Wasserstein GAN with gradient penalty (WGAN-GP), Medical image classification, Class imbalance, Chest radiographs.

Tabla de contenido

Capítulo I Datos generales 4

1.1. Título del proyecto 4

1.2. Datos del investigador 4

1.3. Línea de investigación 4

Capítulo II Problema de investigación 6

2.1. Descripción de la realidad problemática 6

2.2. Formulación del problema de investigación 8

2.2.1. Problema de investigación general 8

2.2.2. Problemas de investigación específicos 8

2.3. Justificación 8

2.4. Objetivos 10

2.4.1. Objetivos generales 10

2.4.2. Objetivos específicos 11

2.5. Hipótesis de la investigación 11

2.5.1. Formulación de hipótesis 11

2.5.2. Hipótesis general 11

2.5.3. Hipótesis específicas 12

Capítulo III Marco teórico 15

3.1. Antecedentes del problema 15

3.1.1. Antecedentes nacionales 15

3.1.2. Antecedentes internacionales 16

3.2. Bases teóricas 27

3.2.1. Rendimiento de modelos de aprendizaje profundo en clasificación de radiografías torácicas 28

3.2.2. Técnicas de balance de datos 37

3.2.3. Optimización del rendimiento en clasificación 49

3.3. Definición de términos 58

Capítulo IV Metodología 62

4.1. Tipo y nivel de investigación 62

4.2. Diseño de investigación 63

4.3. Operacionalización de variables 65

4.3.1. Categorización de variables 67

4.3.2. Condiciones experimentales controladas 67

4.4. Población y muestra 68

4.4.1. Determinación de la unidad de muestreo 68

4.4.2. Delimitación de la población 68

4.4.3. Tipo de muestreo 70

4.4.4. Características de la muestra 70

4.5. Consideraciones éticas 71

4.6. Técnicas e instrumentos para la recolección de datos 71

4.7. Procedimiento metodológico y análisis de datos 73

4.7.1. Metodología aplicada 75

Capítulo V Resultados y discusión 85

5.1. Resultados 85

5.1.1. Resultados experimentales 85

5.1.2. Impacto clínico y traducción de métricas 118

5.2. Discusión de resultados 119

Capítulo VI Conclusiones 122

Capítulo VII Recomendaciones 124

Apéndice A Matriz de consistencia 126

Apéndice B Código fuente de la implementación 129

B.1. Arquitectura del modelo generativo (WGAN-GP) 129

B.3. Funciones de pérdida avanzadas (*loss functions*) 130

B.4. Estrategias de aumento y balanceo de datos 131

B.5. Configuración de los siete escenarios experimentales 131

B.6. Algoritmo de auditoría de robustez (validación *Grad-CAM*) 133

Referencias 134

Índice de tablas

Tabla 1. *Matriz de operacionalización de hipótesis.* 13

Tabla 2. *Densidades radiológicas básicas en radiografía torácica* 31

Tabla 3. *Patrones radiológicos fundamentales en radiografía torácica* 31

Tabla 4. *Matriz de confusión*..... 52

Tabla 5. *Diseño cuasi-experimental por escenarios de comparación.* 64

Tabla 6. *Requisitos y características del diseño cuasiexperimental ejecutado.* .. 65

Tabla 7. *Definiciones conceptuales y operacionales de las variables del estudio.* 66

Tabla 8. *Distribución original del dataset: "Chest X-Ray Images (Pneumonia)" (Kaggle)*..... 69

Tabla 9. *Métodos de recolección de datos: propósito, ventajas y aplicación* 72

Tabla 10. *Distribución de la población total del conjunto de datos.* 76

Tabla 11. *Distribución final de clases y ratio de desbalance (IR) tras la división 80/10/10.*..... 79

Tabla 12. *Hiperparámetros de entrenamiento para la WGAN-GP.* 80

Tabla 13. *Distribución final de clases y ratio de desbalance (IR) después del preprocesamiento.*..... 86

Tabla 14. *Resultados del Escenario 1: modelo línea base.*..... 92

Tabla 15. *Resultados Comparativos de Funciones de Pérdida (Escenarios 2, 3, 4)* 98

Tabla 16 *Resultados del Escenario 5: Sobremuestreo Simple (ROS).*..... 101

Tabla 17. *Resultados del Escenario 6: Modelo (ResNet50 + GAN).*..... 102

Tabla 18. *Resultados del Escenario 6: Modelo híbrido (GAN + Focal Loss)*... 104

Tabla 19. *Comparación de métricas finales en conjunto de prueba (todos los escenarios)* 106

Tabla 20. *Sensibilidad de Resultados: Escenario Real (Desbalanceado) vs. Simulación Balanceada (50/50)* 112

Tabla 21. *Análisis de significancia estadística (McNemar y DeLong)* 113

Tabla 22. *Tasa de acierto de atención en la región patológica (Clase Pneumonia)* 117

Tabla 23. *Matriz de consistencia del proyecto de investigación* 127

Índice de figuras

Figura 1. *Metodología propuesta para la detección de la DDC.* 16

Figura 2. *Arquitectura combinada del WGAN propuesto*. 17

Figura 3. *Progressive Growing of GANs para mejorar la calidad, estabilidad y variación.* 19

Figura 4. *Arquitectura general del modelo de difusión* 20

Figura 5. *Arquitectura de entrenamiento de un modelo de clasificación basado en DCNN.* 21

Figura 6. *Muestras reales y muestras generadas de diferentes GAN.* 23

Figura 7. *Resultados obtenidos, comparación.* 24

Figura 8. *Arquitectura de DCGAN para la generación de CXRs.* 25

Figura 9. *Comparación del rendimiento de la pérdida LMF propuesta con otras funciones de pérdida en EfficientNetV2.* 27

Figura 10. *Método por índices para construir las bases teóricas del marco teórico* 28

Figura 11. *Variable Dependiente - Marco Teórico.* 29

Figura 12. *Proyección PA y lateral izquierda del tórax.* 30

Figura 13. *¿Como abordar el desbalance de datos en la clasificación?* 38

Figura 14. *Técnicas de Balance de Datos*. 39

Figura 15. *Redes Generativas Antagónicas (GAN).* 43

Figura 16. *Procedimiento de entrenamiento de GAN para generación de imágenes. (a) GAN, (b) cGAN y (c) Ambient-GAN. (GAN).* 44

Figura 17. *Representación de una curva ROC.* 54

Figura 18. *Representación de una curva precesion-recall.* 55

Figura 19. Cómo elegir una métrica para la clasificación desequilibrada. 57

Figura 20. *Proceso de selección de población y muestra* 68

Figura 21. *Dataset: "Chest X-Ray Images (Pneumonia)"*..... 69

Figura 22. *Esquema conceptual de las fases principales del procedimiento metodológico*..... 73

Figura 23. *Diagrama de flujo detallado de la ejecución experimental y pipelines de procesamiento.* 74

Figura 24. *Distribución de clases en el conjunto de datos.* 75

Figura 25. *Desbalance de clases en Chest X-Ray Images (Pneumonia)*..... 77

Figura 26. *Distribución de clases después del preprocesamiento.* 79

Figura 27. *Diagrama de flujo de la redistribución de datos.*..... 87

Figura 28. *Ratio de desbalance (pneumonia:normal) en las particiones finales.* 88

Figura 29. *Media móvil (estabilidad) de las pérdidas WGAN-GP.* 90

Figura 30. *Puntuación FID final (48.95) en la etapa de 224x224.*..... 91

Figura 31. *Comparación visual entre imágenes Reales (fila sup.) y Sintéticas (fila inf.).* 91

Figura 32. *Perdida y accuracy, de entrenamiento y validación.* 94

Figura 33. *Matriz de confusión del modelo línea base (Test Set).* 94

Figura 34. *Curvas de entrenamiento (accuracy/loss) - BWCCE.* 96

Figura 35. *Matriz de confusión del escenario 3 (BWCCE) en test set*..... 96

Figura 36. *Curvas de entrenamiento (accuracy/loss) - focal loss.*..... 97

Figura 37. *Matriz de confusión del Escenario 2 (focal loss) en Test Set.* 97

Figura 38. *Curvas de Entrenamiento (Accuracy/Loss) - LDAM Loss.* 98

Figura 39. *Matriz de Confusión del Escenario 4 (LDAM Loss) en Test Set.* 98

Figura 40. 99

Figura 41. *Matriz de confusión del modelo (ResNet50 + GAN) (Test Set).*..... 103

Figura 42. *Matriz de confusión del modelo híbrido (GAN + Focal Loss).*..... 105

Figura 43. *Curvas ROC: Comparativa final de todos los escenarios.*..... 108

Figura 44. *(a) Matriz de confusión de Escenario 1: Línea Base* 109

Figura 45. *(b) Matriz de confusión de Escenario 2 - Focal Loss.*..... 109

Figura 46. *(g) Matriz de confusión de Escenario 5 - ROS.* 110

Figura 47. *(e) Matriz de confusión de Escenario 6 - GAN.*..... 110

Figura 48. *(f) Matriz de confusión de Escenario 7 - Híbrido.* 111

Figura 49. Interfaz del sistema CAD mostrando un diagnóstico positivo. 115

Figura 50. *Diagnóstico de un caso Normal. La ausencia de focos de atención rojos en el parénquima pulmonar valida la decisión del modelo de clasificar al paciente como sano.*..... 116

Figura 51. *Módulo de generación de datos sintéticos en el prototipo CAD, utilizando la WGAN-GP entrenada para producir muestras realistas de la clase minoritaria.* 116

Capítulo I
Datos generales

1.1. Título del proyecto

Técnicas de balance de datos para optimizar el rendimiento de modelos de aprendizaje profundo en clasificación de radiografías torácicas.

1.2. Datos del investigador

Nombre: Estefany Paola Mamani Gutierrez
Institución: Universidad Nacional de Moquegua
Programa de Pregrado: Ingeniería de Sistemas e Informática
Correo electrónico: 2020204028@unam.edu.pe

1.3. Línea de investigación

De conformidad con la Resolución de Vicerrectorado de Investigación Nº 0004-2025-VRI, de fecha 12 de noviembre de 2025, mediante la cual se aprueban las Líneas de Investigación de la Universidad Nacional de Moquegua, el presente trabajo de investigación se alinea con la clasificación basada en las áreas de la OCDE establecida por la universidad:

- Área (OCDE): 1. Ciencias naturales
- Subárea (OCDE): 1.1 Informática y ciencias de la Información
- Disciplina (OCDE): 1.1.1 Ciencias de la computación
- Línea de investigación: 1.1.1.1 Desarrollo y aplicaciones de la inteligencia artificial

El proyecto se adscribe a la línea de Desarrollo y aplicaciones de la inteligencia artificial, dado que su núcleo técnico se fundamenta en la implementación y optimización de algoritmos avanzados de **aprendizaje profundo** (d*eep learning*). La investigación no solo aplica modelos existentes (como ResNet50), sino que desarrolla una solución híbrida mediante el entrenamiento de

una WGAN-GP para la generación de datos sintéticos y el uso de funciones de pérdida sensibles al costo (*focal loss*), contribuyendo al desarrollo de metodologías para solucionar problemas de desbalance de datos en el diagnóstico médico automatizado.

Capítulo II
Problema de investigación

2.1. Descripción de la realidad problemática

El uso de técnicas de aprendizaje profundo en la clasificación de imágenes médicas ha demostrado un potencial significativo para alcanzar niveles de precisión diagnóstica comparables a los de especialistas humanos (Shen et al., 2017). Sin embargo, un desafío fundamental en este campo es el desequilibrio de clases en los conjuntos de datos médicos, donde las imágenes correspondientes a enfermedades raras son considerablemente menos frecuentes. Este desequilibrio impacta negativamente en el rendimiento de los modelos, que tienden a sobre ajustarse a las clases mayoritarias y muestran un desempeño deficiente en la detección de condiciones minoritarias (Fotouhi et al., 2019). En el contexto del diagnóstico pulmonar, las radiografías torácicas son particularmente relevantes debido a su accesibilidad y bajo costo, lo que enfatiza la necesidad de desarrollar métodos efectivos para mitigar este desequilibrio y optimizar el rendimiento de los sistemas basados en aprendizaje profundo (C. Qin et al., 2018), considerando que este desequilibrio se origina en la distribución naturalmente desigual de enfermedades, los altos costos de obtención y etiquetado de casos patológicos, y los sesgos en la recolección de datos (X. Qin et al., 2022)

A nivel internacional, se han implementado diversas estrategias para abordar el desequilibrio de datos en imágenes médicas, incluyendo métodos de aumento de datos y el uso de GAN y autocodificadores para generar datos sintéticos (Galán-Cuenca et al., 2024). La literatura reporta investigaciones que analizan el impacto del desbalance de datos en el rendimiento predictivo de redes neuronales convolucionales y proponen nuevas funciones de pérdida para manejar este problema (Galdran et al., 2021a; Sadi et al., 2022), logrando mejoras significativas en la clasificación de enfermedades raras. Los métodos basados en las GAN, como **red generativa antagónica de crecimiento progresivo** (PGGAN) y **red generativa antagónica condicional** (cGAN), muestran mejoras significativas en métricas como precisión y AUC, alcanzando valores de hasta 97.8% de precisión y 0.954 de AUC tras aplicar técnicas de balance de datos, superando métodos tradicionales como el aumento geométrico (Ganesan et al., 2019; Liang et al., 2020;

X. Qin et al., 2022). Por otro lado, enfoques novedosos como **red neuronal de pulsos** (SNN) con **plasticidad dependiente del tiempo de disparo modulada por recompensa** (R-STDP) y funciones de pérdida específicas **pérdida focal de gran margen** (LMFLOSS) y **mixup balanceado** (*Balanced-MixUp*) logran incrementos entre el 2% y 12% en la precisión de clases raras (Galdran et al., 2021b). Modelos como ResNet50, EfficientNet y DenseNet sobresalen cuando se combinan con estas técnicas, evidenciando que el sobremuestreo con las GAN y la integración de funciones de pérdida ponderadas son soluciones efectivas en escenarios altamente desbalanceados (Hoang et al., 2023; X. Qin et al., 2022).

En el contexto nacional, aunque existen investigaciones sobre el desequilibrio de datos en campos como finanzas y marketing (Camargo, 2021; Garro, 2019; Huarhuachi, 2017), el estudio específico de este problema en radiografías torácicas permanece limitado. Destacan trabajos como el de Gaspar (2023), quien empleó las GAN para equilibrar conjuntos de datos en la detección de displasia de cadera en infantes, alcanzando con la arquitectura ResNet50 una precisión del 97.3%, en contraste con el 74.3% obtenido por un modelo **red neuronal convolucional** (CNN) básico sin aumento de datos. En el departamento de Moquegua, esta brecha es aún más marcada, ya que, si bien existen investigaciones relacionadas al análisis de estas radiografías torácicas, no se han abordado específicamente las técnicas para lidiar con el desbalance en este contexto. Esta brecha es particularmente relevante considerando que el uso efectivo de sistemas basados en aprendizaje profundo podría optimizar significativamente el diagnóstico de enfermedades pulmonares en la región.

Según Sun et al. (2009), las estrategias para abordar el desequilibrio de datos se clasifican en tres categorías principales: métodos a nivel de datos, métodos a nivel de algoritmo y métodos híbridos. Los métodos a nivel de datos, que incluyen técnicas de remuestreo, permiten ajustar la distribución mediante sobre muestreo de clases minoritarias o submuestreo de clases mayoritarias (Japkowicz & Stephen, 2002). Los métodos a nivel de algoritmo modifican los clasificadores o implementan aprendizaje sensible a costos (Sun et al., 2007), mientras que los métodos híbridos combinan ambos enfoques para mayor efectividad.

La presente investigación busca mejorar el rendimiento de los modelos de aprendizaje profundo para la clasificación de radiografías torácicas a través de técnicas de balance de datos; para ello, se evaluará la eficacia de diferentes técnicas, incluyendo la generación de datos sintéticos mediante las GAN, la modificación de funciones de pérdida sensibles a costos como *focal loss y* **pérdida balanceada por clase** *(class-balanced loss),* y métodos híbridos. El impacto de estas variables será medido mediante métricas como sensibilidad, precisión y *F1-score*, con el objetivo de mejorar la capacidad predictiva de los modelos (Brownlee, 2020). Esta investigación tiene el potencial de contribuir significativamente a la optimización de decisiones clínicas y resultados de salud en la región.

2.2. Formulación del problema de investigación

2.2.1. Problema de investigación general

¿En qué medida las técnicas de balance de datos influyen en el rendimiento de los modelos de aprendizaje profundo para la clasificación de radiografías torácicas?

2.2.2. Problemas de investigación específicos

- ¿Cuál es el impacto de las GAN en la precisión y sensibilidad de los modelos de aprendizaje profundo para la clasificación de radiografías torácicas?
- ¿En qué medida las funciones de pérdida sensibles a costos mejoran el rendimiento de los modelos de aprendizaje profundo en la clasificación de radiografías torácicas?
- ¿Cómo afecta la implementación de métodos híbridos (combinación de técnicas a nivel de datos y algoritmos) al rendimiento de los modelos de aprendizaje profundo en la clasificación de radiografías torácicas desbalanceadas?

2.3. Justificación

Como se señaló en la sección 2.1, el desequilibrio de clases en conjuntos de datos de radiografías torácicas constituye una barrera significativa para el desarrollo

de sistemas de análisis automático efectivos. La investigación propuesta tiene un alto valor teórico, ya que busca ampliar el conocimiento sobre cómo las técnicas híbridas de balance de datos (combinación de métodos a nivel de datos y algoritmos) interactúan específicamente en radiografías torácicas, aspecto que permanece sin estudiar en contextos latinoamericanos. Actualmente, los métodos más comunes para mitigar este problema incluyen técnicas básicas de aumento de datos, como rotar, voltear, escalar o aplicar transformaciones geométricas a las imágenes (Ganesan et al., 2019). Este trabajo permitirá explicar el impacto de diferentes estrategias de balance de datos (como el uso de las GAN, funciones de pérdida sensibles a costos y enfoques híbridos) en el rendimiento de modelos de clasificación, proporcionando resultados aplicables a diversos contextos clínicos. Además, las conclusiones generadas evaluarán la validez de teorías actuales en el manejo del desequilibrio de datos en el campo de la inteligencia artificial aplicada a la medicina (Liang et al., 2020; X. Qin et al., 2022).

Es importante destacar la originalidad y pertinencia de esta investigación en el contexto nacional, donde existe un vacío significativo en el estudio de técnicas de manejo de desbalance en datos de imágenes radiográficas pulmonares. Si bien existen investigaciones sobre el desequilibrio de datos en otros campos (Camargo, 2021; Garro, 2019; Huarhuachi, 2017), el estudio específico de este problema en imágenes médicas permanece limitado. Los resultados de trabajos nacionales, como el de quien alcanzó una precisión del 97.3% utilizando las GAN en la detección de displasia de cadera, demuestran el potencial de estas técnicas en el contexto peruano. La implementación de estas metodologías no solo contribuirá al avance del conocimiento en este campo, sino que también tendrá un impacto directo en la mejora de los servicios de salud.

El impacto social de esta investigación radica en su capacidad para mejorar la precisión y rapidez en el diagnóstico de enfermedades pulmonares, lo cual beneficia directamente a los pacientes y a la comunidad médica. La detección temprana de condiciones graves, como el cáncer de pulmón o la EPOC, no solo salva vidas, sino que también optimiza los recursos del sistema de salud al permitir tratamientos más efectivos y menos costosos (Ramnarine, 2021). A nivel nacional,

este estudio puede servir de modelo para otras regiones que enfrenten problemas similares, potenciando el uso de inteligencia artificial en la medicina.

La investigación tiene un fuerte componente innovador al evaluar el impacto de técnicas avanzadas como las GAN y métodos híbridos en la mejora del rendimiento de modelos de clasificación en datos desbalanceados. Este enfoque no solo optimizará el uso de tecnologías existentes, sino que también puede inspirar el desarrollo de nuevos procedimientos y algoritmos aplicables a otros campos médicos y científicos. Los hallazgos podrían ser utilizados para diseñar herramientas de diagnóstico más precisas y adaptadas a los retos del desbalance de datos, mejorando la calidad de vida de los pacientes y fortaleciendo la infraestructura tecnológica de los sistemas de salud (Ganesan et al., 2019).

Metodológicamente, esta investigación propone un enfoque empírico y analítico que incluye la implementación y evaluación de diversas técnicas de clasificación desequilibrada. Se explorarán métodos que, según Sun et al. (2009), se clasifican en tres categorías principales: métodos a nivel de datos, métodos a nivel de algoritmo y métodos híbridos. Las técnicas de generación de datos sintéticos, particularmente mediante las GAN y autocodificadores, representan una estrategia innovadora para enriquecer el conjunto de datos y proporcionar una mayor representatividad para el entrenamiento de los modelos (Galán-Cuenca et al., 2024). Además, se implementarán funciones de pérdida sensibles a costos y métodos híbridos, evaluando su impacto mediante métricas estándar como precisión, sensibilidad y *F1-score* (Brownlee, 2020). Asimismo, los instrumentos y modelos desarrollados podrán ser utilizados como referencia en otras áreas de investigación relacionadas, fortaleciendo el conocimiento técnico y metodológico en el ámbito académico y clínico.

2.4. Objetivos

2.4.1. Objetivos generales

Evaluar la influencia de las técnicas de balance de datos en el rendimiento de los modelos de aprendizaje profundo para la clasificación de radiografías torácicas.

2.4.2. Objetivos específicos

- Evaluar el impacto de las GAN en la precisión y sensibilidad de los modelos de aprendizaje profundo para la clasificación de radiografías torácicas.

- Analizar el efecto de las funciones de pérdida sensibles a costos en el rendimiento de los modelos de aprendizaje profundo para la clasificación de radiografías torácicas.

- Determinar el impacto de la implementación de métodos híbridos en el rendimiento de los modelos de aprendizaje profundo para la clasificación de radiografías torácicas desbalanceadas.

2.5. Hipótesis de la investigación

2.5.1. Formulación de hipótesis

En el marco de esta investigación, dado su alcance explicativo según Hernández Sampieri & Mendoza Torres (2018) se plantean hipótesis causales que buscan establecer relaciones de causa-efecto entre las técnicas de balance de datos y el rendimiento de los modelos de aprendizaje profundo en la clasificación de radiografías torácicas, como se observa en la Tabla 1.

Nota: El umbral de mejora significativa del 3% establecido en las hipótesis se fundamenta en la revisión de la literatura especializada. Investigaciones previas como las de Sadi et al. (2022), reportan mejoras en el F1-score entre el 2% y el 9% al aplicar funciones de pérdida híbridas, mientras que Galdran et al. (2021b) documentan incrementos de entre el 2% y el 12% en la precisión de clases minoritarias utilizando técnicas de balanceo. Por consiguiente, se define el 3% como un margen conservador y estadísticamente relevante para validar la efectividad de las técnicas propuestas en este estudio.

2.5.2. Hipótesis general

- **H1:** Las técnicas de balance de datos mejoran significativamente el rendimiento de los modelos de aprendizaje profundo en la clasificación de

radiografías torácicas, incrementando la precisión y sensibilidad en al menos un 3% respecto a los modelos sin balance de datos.

2.5.3. Hipótesis específicas

- **H1.1:** La implementación de las GAN incrementa la precisión y sensibilidad de los modelos de aprendizaje profundo en al menos un 3% para la clasificación de radiografías torácicas de la clase minoritaria.

- **H1.2:** Las funciones de pérdida sensibles a costos mejoran el rendimiento de los modelos de aprendizaje profundo, aumentando el $F1$-score en al menos un 3% en la clasificación de radiografías torácicas desbalanceadas.

- **H1.3:** Los métodos híbridos que combinan técnicas a nivel de datos y algoritmos incrementan el rendimiento general de los modelos de aprendizaje profundo en al menos un 3% en términos de precisión balanceada y AUC-ROC para la clasificación de radiografías torácicas desbalanceadas.

Tabla 1.

Matriz de operacionalización de hipótesis.

Elemento	Descripción	Definición conceptual	Definición operacional
Problema general ¿En qué medida las técnicas de balance de datos influyen en el rendimiento de los modelos de aprendizaje profundo para la clasificación de radiografías torácicas?	**Hipótesis general:** Las técnicas de balance de datos mejoran significativamente el rendimiento de los modelos de aprendizaje profundo en la clasificación de imágenes radiografías torácicas, incrementando la precisión y sensibilidad en al menos un 3%.	**Variable independiente (VI):** Técnicas de balance de datos. - Equilibran la distribución de clases en conjuntos de datos de imágenes médicas, mejorando la capacidad del modelo para generalizar y aprender patrones de ambas clases efectivamente (Galán-Cuenca et al., 2024) **Variable dependiente (VD):** Rendimiento en clasificación medido por métricas estándar. - Capacidad del modelo para predecir correctamente las etiquetas de clase en nuevas instancias (Bishop, 2006).	**VI:** Aplicación de técnicas de balance a nivel de datos (ROS, GAN) y a nivel de algoritmo (BWCCE, *focal loss*, LDAM *loss*)) sobre la arquitectura ResNet50 para equilibrar la distribución de clases en el conjunto de radiografías torácicas (Iqbal et al., 2022). **VD:** Evaluación del rendimiento de ResNet50 mediante las métricas de precisión, sensibilidad, *F1*-score y AUC-ROC para medir su capacidad de clasificación (Brownlee, 2020).
Problema específico 1	**Hipótesis específica 1:** Las GAN incrementan la precisión y sensibilidad de los modelos en al	**Redes generativas antagónicas (GAN):** Arquitectura de aprendizaje profundo con dos redes neuronales competitivas: un generador	**Redes generativas antagónicas (GAN):** Implementación mediante generador, discriminador y función de pérdida *adversarial*

Elemento	Descripción	Definición conceptual	Definición operacional
¿Cuál es el impacto de las GAN en la precisión y sensibilidad de los modelos?	menos un 3% para la clase minoritaria.	que crea datos sintéticos y un discriminador que diferencia entre datos reales y generados mediante "un proceso de juego *adversarial*" (Goodfellow et al., 2014).	que optimiza ambas redes simultáneamente para "generar imágenes sintéticas que preservan características estadísticas y estructurales de las imágenes originales" (Frid-Adar et al., 2018).
Problema específico 2 ¿En qué medida las funciones de pérdida sensibles a costos mejoran el rendimiento de los modelos?	**Hipótesis específica 2:** Las funciones de pérdida sensibles a costos aumentan el *F1*-score en al menos un 3%.	**Funciones de pérdida adaptadas al desbalance de clases.** Estrategias matemáticas que asignan pesos o penalizaciones diferenciadas durante el entrenamiento para mitigar el impacto de distribuciones desiguales de clases (He & Garcia, 2009).	**Funciones de pérdida adaptadas al desbalance de clases.** Implementación de BWCCE, *focal loss* y *LDAM loss* mediante ponderación inversa a la frecuencia de clases y ajuste dinámico de términos de penalización para mejorar precisión y sensibilidad en clases minoritarias (Gao et al., 2025)
Problema específico 3 ¿Cómo afecta la implementación de métodos híbridos al rendimiento de los modelos?	**Hipótesis específica 3:** Los métodos híbridos incrementan el rendimiento general en al menos un 3%.	**Combinación de técnicas a nivel de datos y algoritmos.** Métodos híbridos que integran estrategias de balance tanto a nivel de datos como de algoritmos para compensar las limitaciones de técnicas individuales (Japkowicz & Stephen, 2002).	**Combinación de técnicas a nivel de datos y algoritmos.** Integración sistemática de técnicas a nivel de datos (ROS, GAN) y algoritmos (BWCCE, *focal loss*, *LDAM loss*) sobre ResNet50 para mejorar precisión y robustez en distribuciones desiguales.

Capítulo III
Marco teórico

3.1. Antecedentes del problema

La fundamentación de esta investigación se sustenta en una revisión sistematizada de la literatura científica reciente, centrada en el uso de técnicas de aprendizaje profundo y estrategias de balanceo de datos para el diagnóstico médico. El proceso de selección priorizó estudios indexados en bases de datos de alto impacto que abordan específicamente la clasificación de radiografías torácicas y el uso de técnicas de balance de datos. A continuación, se presenta una síntesis crítica de los trabajos nacionales e internacionales más relevantes que cimentan la propuesta metodológica de este estudio.

3.1.1. Antecedentes nacionales

El investigador Gaspar (2023) de la Universidad Nacional Mayor de San Marcos, en Lima, Perú, realizó un estudio titulado "Utilización de técnicas de aprendizaje profundo para la detección temprana de displasia de cadera en infantes mediante radiografías pélvicas". Su objetivo fue optimizar la precisión y confiabilidad en el diagnóstico de la **displasia del desarrollo de la cadera** (DDC) en infantes de tres a seis meses mediante el análisis de radiografías pélvicas utilizando técnicas de aprendizaje profundo. Los datos utilizados en este estudio fueron tomados del conjunto de datos publicado originalmente por Fraiwan et al. (2022).

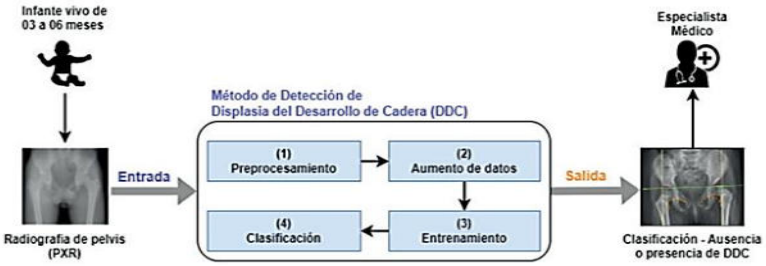
Dicha base de datos consta de radiografías obtenidas en dos instituciones de Jordania: el Hospital Universitario Rey Abdullah (Irbid) y el Hospital Especializado Alsafa (Jarash). La muestra incluyó 354 sujetos, de los cuales 120 presentaban DDC y 234 eran casos normales. Para llevar a cabo el estudio, se emplearon arquitecturas como las CNN y las GAN. El modelo principal fue una arquitectura ResNet50, que se entrenó con datos sintéticos generados por la GAN y se comparó con una CNN básica.

Las imágenes fueron preprocesadas mediante normalización y aplicación de filtros gaussianos, y se dividieron en conjuntos de entrenamiento y validación como

se observa en la Figura 1. El modelo ResNet50 con datos aumentados alcanzó una precisión del 97.3%, mientras que la CNN básica obtuvo solo un 74.3%. Asimismo, ResNet50 demostró una mayor sensibilidad (98.82%) y precisión (96.00%) en comparación con la CNN básica, y su curva ROC arrojó un AUC de 0.97, indicando una alta capacidad de discriminación. Las conclusiones del autor Gaspar señalaron que la integración de ResNet50 con las GAN permitió una mejora significativa en la detección de DDC, siendo útil en sistemas de diagnóstico médico asistido por computadora. Este antecedente demuestra la efectividad de las GAN y modelos ResNet, en la mejora del rendimiento en la clasificación de imágenes médicas.

Figura 1.

Metodología propuesta para la detección de la DDC.



Nota. Adaptado de "Utilización de técnicas de aprendizaje profundo para la detección temprana de displasia de cadera en infantes mediante radiografías pélvicas" (p. 232), por R. J. J. Gaspar, 2023, Revista de Investigación de Sistemas e Informática, 16(2).

3.1.2. Antecedentes internacionales

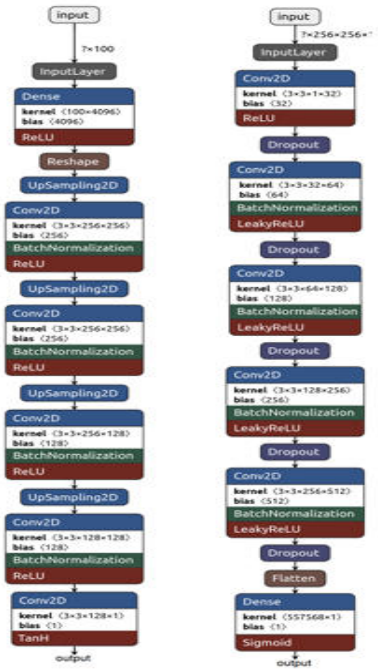
Hussain et al. (2022) mediante una colaboración entre instituciones de India y el Reino Unido, publicaron en la conferencia de la *IEEE Engineering in Medicine & Biology Society (EMBC)*, celebrada en Glasgow (Reino Unido), una investigación titulada "Wasserstein GAN based chest X-RAY dataset augmentation for deep learning models: COVID-19 detection use-case". Este estudio, tuvo como objetivo principal proponer una solución a la deficiencia de datos (*data deficiency)* en radiografías torácicas de COVID-19. Para ello, buscaron implementar un modelo ligero de WGAN como se muestra en la Figura 2.

Este estudio de Hussain empleó un conjunto de datos de acceso abierto de Kaggle, compuesto por 3615 radiografías torácicas de COVID-19 positivo. Los instrumentos implementaron una arquitectura WGAN, utilizando la **pérdida de Wasserstein** *(Wasserstein loss*, WL) para estabilizar el entrenamiento y evitar el colapso del modo. La calidad de las imágenes generadas (de 96x96 píxeles) se evaluó de forma cuantitativa, utilizándolas para entrenar un modelo de clasificación ligero (propuesto por los mismos autores en un trabajo previo). Resultando así en un clasificador entrenado con el conjunto de datos aumentado por WGAN alcanzó una precisión del 95.34% y un AUC del 99.4%. Esto representó una mejora significativa frente al mismo modelo entrenado con técnicas de aumento tradicionales (no-GAN), que solo obtuvo un 93.04% de precisión y un 98.38% de AUC.

En conclusión, el estudio demostró que el aumento de datos basado en WGAN es una técnica superior a los métodos de aumento tradicionales *(naive)* y que las imágenes generadas son creíbles y eficaces para el entrenamiento de modelos de aprendizaje profundo.

Figura 2.

Arquitectura combinada del WGAN propuesto.



Nota. El Generador se muestra a la izquierda, mientras que el Discriminador (Crítico) se representa a la derecha. Adaptado de "Wasserstein GAN based Chest

X-Ray Dataset Augmentation for Deep Learning Models: COVID-19 Detection Use-Case" (p. 2059), por B. Z. Hussain, I. Andleeb, M. S. Ansari, A. M. Joshi y N. Kanwal, 2022, 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE.

Sindhura et al. (2024) publicaron una revisión sistemática titulada "A review of deep learning and generative adversarial networks applications in medical image analysis". Este estudio, desarrollado en India, tuvo como objetivo revisar y sintetizar el panorama de las aplicaciones de los modelos de aprendizaje profundo y las GAN en el dominio del análisis de imágenes médicas. El estudio Sindhura abordó un amplio espectro de tareas, incluyendo la segmentación, clasificación, detección de fracturas y, de forma central, la generación de imágenes médicas sintéticas. Su población de estudio fueron los artículos de investigación relevantes. El estudio catalogó y analizó una gran variedad de arquitecturas de aprendizaje profundo (como ResNet y DenseNet) y una amplia gama de variantes de las GAN destacando explícitamente las **red generativa antagónica convolucional profunda** (DCGAN), PGGAN, WGAN y WGAN-GP. Tambien se menciona que el uso de estos modelos en *datasets* de referencia como BraTS 2020, FastMRI y COVID-CT-*Dataset*. Los autores concluyen que las GAN, y en particular las arquitecturas más estables como WGAN-GP, se han implementado con notable éxito para la generación de imágenes sintéticas realistas, así como para la reconstrucción de imágenes, eliminación de ruido y registro.

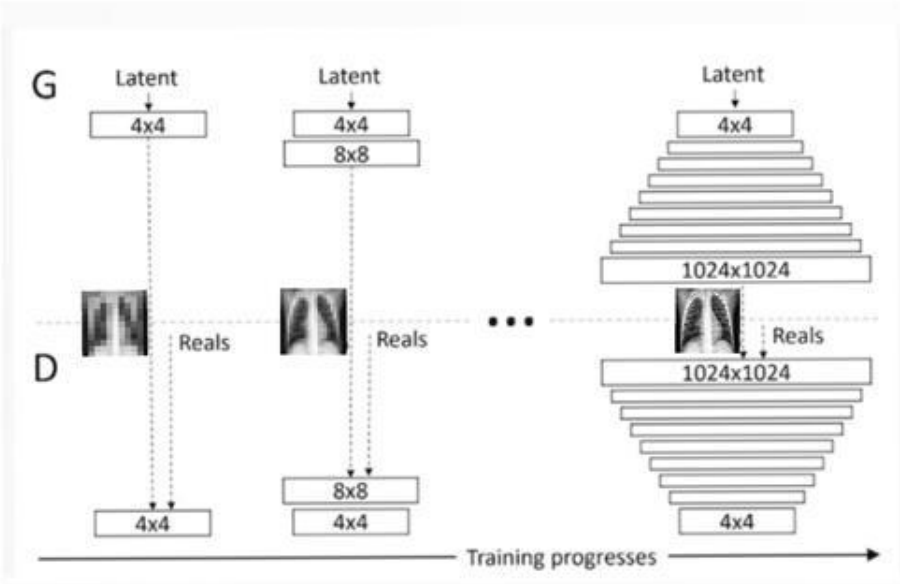
Keerthi et al. (2024), presentaron en la conferencia IEEE International Conference on Computer Vision and Machine Intelligence (CVMI) 2024, una investigación titulada "Advancing Medical Imaging: A Comparative Exploration of Generative Adversarial Networks for Chest X-ray Synthesis", cuyo principal objetivo fue abordar la escasez crítica de imágenes médicas etiquetadas para enfermedades pulmonares como COVID-19, Tuberculosis y Neumonía. Para ello, propusieron comparar la calidad de las radiografías torácicas sintetizadas por diferentes arquitecturas GAN.

Como metodología, los autores implementaron y compararon tres arquitecturas generativas: **red generativa antagónica con clasificador auxiliar** (ProGAN), DCGAN y **red generativa antagónica de crecimiento progresivo**

(ACGAN). De manera notable, reconociendo las limitaciones de estabilidad de estas, también emplearon una metodología híbrida que combinaba ProGAN con la WGAN-GP como se observa en la Figura 3, para mejorar la estabilidad del entrenamiento y la calidad de la imagen. La evaluación de la calidad de las imágenes sintéticas se realizó utilizando un clasificador InceptionV3. Los resultados fueron reveladores: mientras el clasificador obtuvo un 91% de precisión en los datos originales, su rendimiento se desplomó en las imágenes sintéticas de las arquitecturas más simples: 68.62% (DCGAN) y un bajo 45.50% (ACGAN). Los autores concluyeron que, si bien las GAN tienen un gran potencial, la elección de la arquitectura es crítica, y las arquitecturas más estables (como las que incorporan WGAN-GP) son necesarias.

Figura 3.

Progressive Growing of GANs para mejorar la calidad, estabilidad y variación.



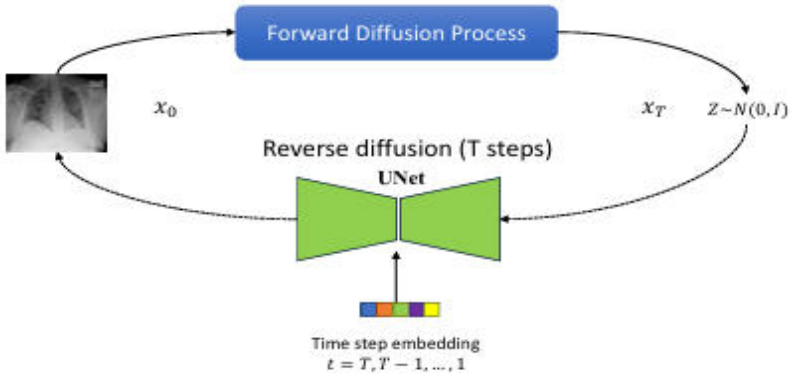
Nota. Adaptado de "Advancing Medical Imaging: A Comparative Exploration of Generative Adversarial Networks for Chest X-ray Synthesis" (p. 3), por R. Keerthi, K. Kiran, S. S. Kiran y P. Likitha, 2024, 2024 *IEEE International Conference on Computer Vision and Machine Intelligence (CVMI), IEEE*.

Liang et al. (2024) presentaron un trabajo en el *IEEE Southwest Symposium on Image Analysis and Interpretation* titulado "COVID-19 pneumonia chest X-ray pattern synthesis by stable diffusion". Esta investigación, desarrollada en Estados

Unidos, tuvo como objetivo explorar la síntesis de patrones de neumonía por COVID-19 en radiografías torácicas. El aspecto más innovador de este estudio es el instrumento utilizado: en lugar de emplear arquitecturas de la familia GAN, los autores utilizaron un modelo de difusión (*stable diffusion*) Figura 4. El estudio de Liang se centró en la capacidad de esta técnica emergente para generar patrones patológicos complejos y realistas. Aunque el artículo se enfoca en la metodología de síntesis, su conclusión principal es que los modelos de difusión son una alternativa moderna y viable a las GAN para la generación de imágenes médicas.

Figura 4.

Arquitectura general del modelo de difusión



Nota. Adaptado de "Covid-19 pneumonia chest x-ray pattern synthesis by stable diffusion" (p. 22), por Z. Liang, Z. Xue, S. Rajaraman y S. Antani, 2024, 2024 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI) (pp. 21-24), IEEE.

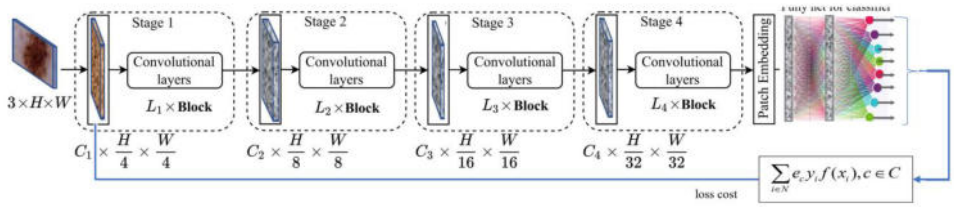
En 2023, Hoang et al. (2023), investigadores de la Facultad de Tecnología de la Información de la Universidad de Tecnología y Educación de Ho Chi Minh City en Vietnam y del Departamento de Ingeniería Eléctrica, Electrónica e Informática de la Universidad de Ulsan en Corea del Sur, publicaron "Categorical Weighting Domination for Imbalanced Classification With Skin Cancer in Intelligent Healthcare Systems", con el objetivo de mejorar la eficiencia de las **redes neuronales convolucionales profundas** (DCNN) en la clasificación de imágenes médicas multiclase en conjuntos de datos desequilibrados. La población de estudio incluyó el conjunto ISIC2018 (imágenes dermatoscópicas) y radiografías torácicas con cuatro categorías: COVID-19, normal, neumonía bacteriana y viral.

Utilizaron modelos preentrenados como EfficientNets, MobileNets y DenseNets, experimentando con funciones de pérdida (entropía cruzada categórica y pérdida focal) y técnicas de aumento de datos (Figura 5).

El enfoque propuesto de Hoang optimizó la dominación de pesos entre clases para equilibrar categorías mayoritarias y minoritarias. Los resultados mostraron que el método superó enfoques tradicionales en métricas como recuperación, precisión y especificidad, tanto en ISIC2018 como en radiografías torácicas. Los autores concluyeron que la dominación de ponderación categórica es robusta y estable para distintas arquitecturas, mejorando el rendimiento sin crear muestras adicionales (Hoang et al., 2023). Este antecedente es relevante por presentar un enfoque innovador para resolver el desequilibrio en datos médicos, alineándose con el objetivo de desarrollar soluciones precisas para la detección de patologías en radiografías torácicas.

Figura 5.

Arquitectura de entrenamiento de un modelo de clasificación basado en DCNN.



Nota. Adaptado de "Categorical Weighting Domination for Imbalanced Classification With Skin Cancer in Intelligent Healthcare Systems" (p. 105174), por V. D. Hoang, X. T. Vo y K. H. Jo, 2023, IEEE Access, 11, pp. 105170-105181.

X. Qin et al. (2022), publicaron "Learning From Limited and Imbalanced Medical Images With Finer Synthetic Images From GAN", investigando la efectividad de las GAN para sintetizar datos y abordar la escasez y desequilibrio de clases en radiografías torácicas, analizando cómo la calidad de las imágenes generadas influye en el rendimiento de clasificadores. Utilizaron dos conjuntos de datos públicos: Kaggle Pneumonia Dataset (5,216 imágenes de entrenamiento, clasificadas como "neumonía" o "normal") y el *RSNA Pneumonia Detection Challenge Dataset* (10,244 imágenes de entrenamiento, distribuidas en "opacidad

pulmonar" y "normal"), ambos con clases desbalanceadas. Emplearon DCGAN y PGGAN, evaluando la calidad mediante FID y **índice de similitud estructural multiescala** (MS-SSIM), y entrenaron clasificadores con tres enfoques: transformación geométrica, imágenes de DCGAN e imágenes de PGGAN.

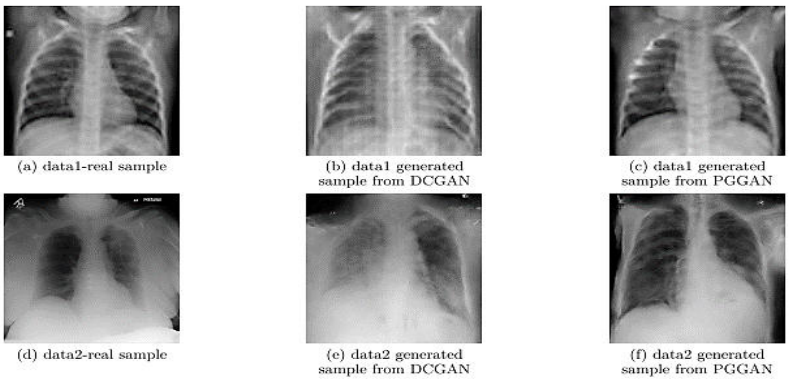
Los modelos incluyeron las CNN desde cero, aprendizaje por transferencia (Inception V3, ResNet, DenseNet, EfficientNet) y una CNN multiescala diseñada por los autores, evaluados mediante precisión, sensibilidad, especificidad, *F1*-score, *G-mean*, AUC y **área bajo la curva de precisión-exhaustividad** (AUPRC). Los resultados indicaron que PGGAN generó imágenes de mayor calidad que DCGAN (

Figura 6).

El aumento con GAN, especialmente PGGAN, mejoró el desempeño de la mayoría de clasificadores frente a métodos tradicionales. Sin embargo, contrario a lo esperado, el aprendizaje por transferencia mostró mejoras limitadas con las imágenes generadas, posiblemente debido a que los modelos preentrenados ya capturaban características generales suficientes. La CNN multiescala obtuvo el mejor rendimiento y menor tiempo de entrenamiento. Los autores concluyeron que las GAN son valiosas para abordar el desequilibrio en datos médicos, destacando que la calidad de las imágenes generadas impacta directamente en la efectividad de los clasificadores (X. Qin et al., 2022). Este trabajo es relevante al proporcionar evidencia de que las GAN mejoran significativamente el rendimiento en clasificación de radiografías torácicas con datos limitados y desbalanceados.

Figura 6.

Muestras reales y muestras generadas de diferentes GAN.



Nota. Adaptado de "Learning From Limited and Imbalanced Medical Images With Finer Synthetic Images From GANs" (p. 91672), por X. Qin, F. M. Bui, H. H. Nguyen y Z. Han, 2022, IEEE Access, 10, pp. 91663-91677.

Liang et al. (2020) presentaron "Enhancing Automated COVID-19 Chest X-ray Diagnosis by Image-to-Image GAN Translation" en la *IEEE International Conference on Bioinformatics and Biomedicine* (BIBM). El objetivo fue mejorar la precisión del diagnóstico automatizado de COVID-19 en radiografías torácicas mediante traducción de imagen a imagen basada en cGAN, transformando patrones de rayos X normales en imágenes sintéticas que imitaban neumonía por COVID-19 para balancear el conjunto de datos. La población incluyó 219 imágenes COVID-19, 1341 normales y 1345 de neumonía viral, representando un desequilibrio típico. Utilizaron una cGAN con arquitectura U-Net, produciendo 1100 imágenes sintéticas de COVID-19 y creando un conjunto balanceado de 3762 imágenes. El modelo ResNet-50 fue entrenado con ambos conjuntos (original desbalanceado y balanceado generado).

Los resultados mostraron que ResNet con el conjunto balanceado alcanzó 97.8% de precisión, superando el 96.1% del conjunto desbalanceado con aprendizaje por transferencia, con mejoras en estabilidad de precisión, *recall* y *F1*-score entre clases. Este antecedente valida el uso de cGAN para abordar desequilibrio y escasez de datos en imágenes médicas (Liang et al., 2020). La generación de imágenes sintéticas mediante traducción imagen a imagen representa una alternativa prometedora para aumentar conjuntos limitados y desbalanceados,

optimizando el desempeño de modelos de aprendizaje profundo y proporcionando un marco práctico para mejorar precisión y robustez de clasificadores en escenarios similares.

Figura 7.

Resultados obtenidos, comparación.

Measure	Class		
	Normal	*COVID-19*	*Viral*
Original imbalance dataset (transfer learning)			
Accuracy	0.961		
Precision	0.935	1.000	0.950
Recall	0.967	0.967	0.950
F1 score	0.951	0.983	0.950
Original imbalance dataset (train from scratch)			
Accuracy	0.956		
Precision	0.967	1.000	0.906
Recall	0.967	0.933	0.967
F1 score	0.967	0.966	0.935
cGAN enhanced dataset (transfer learning)			
Accuracy	0.978		
Precision	0.952	1.000	0.983
Recall	0.983	1.000	0.950
F1 score	0.967	1.000	0.966
cGAN enhanced dataset (transfer from scratch)			
Accuracy	0.961		
Precision	0.935	0.984	0.965
Recall	0.967	1.000	0.917
F1 score	0.951	0.992	0.940

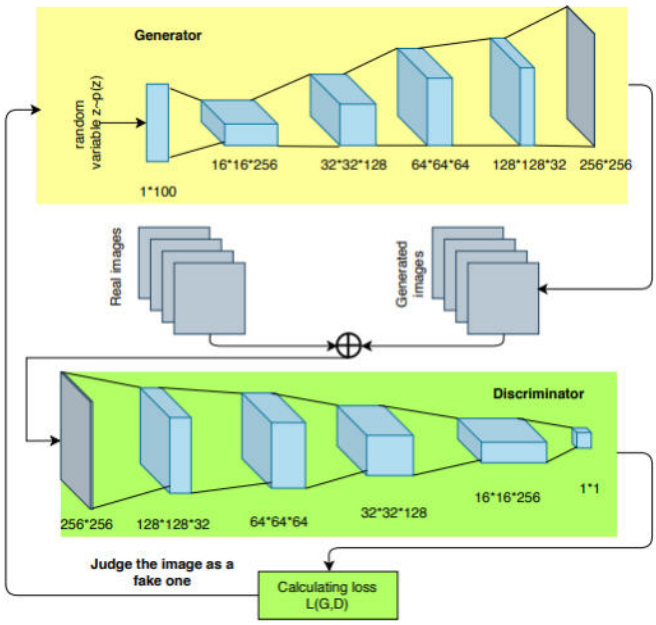
Nota. Adaptado de "Enhancing Automated COVID-19 Chest X-ray Diagnosis by Image-to-Image GAN Translation" (p. 1071), por Z. Liang, J. X. Huang, J. Li y S. Chan, 2020, *Proceedings - 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2020)* (pp. 1068-1071), IEEE.

X. Qin et al. (2019) llevaron a cabo "Learning from an Imbalanced and Limited Dataset and an Application to Medical Imaging" en la Universidad de Saskatchewan, Canadá, evaluando la efectividad de estrategias de muestreo y modelos de aprendizaje automático para detectar neumonía en radiografías torácicas limitadas y desbalanceadas. El conjunto de datos de Kaggle incluyó 5863 imágenes (5216 entrenamiento, 16 validación, 624 prueba), clasificadas en neumonía (3875) y normal (1341). Implementaron **máquinas de vectores de soporte** (SVM) con **histogramas de gradientes orientados** (HOG) y CNN similares a ZFNet. Evaluaron tres estrategias de muestreo: submuestreo de la clase mayoritaria, sobremuestreo mediante transformaciones geométricas y

sobremuestreo con DCGAN para generar imágenes sintéticas (Figura 8). Las imágenes se preprocesaron a 256x256 píxeles, tonos de gris y normalización. El rendimiento se midió mediante precisión, sensibilidad, especificidad, *F1*-score y AUC. Los resultados mostraron que el muestreo mejora significativamente el rendimiento: en SVM, el submuestreo incrementó la especificidad de 0.286 a 0.726; las CNN mostraron rendimiento superior, alcanzando AUC de 0.911 sin muestreo. El sobremuestreo con DCGAN proporcionó el mejor resultado para CNN (AUC de 0.954), demostrando que las imágenes sintéticas capturaron características esenciales de las radiografías. Los autores concluyeron que las estrategias de muestreo son fundamentales para reducir el sesgo y mejorar el desempeño, destacando la superioridad de CNN sobre el SVM (X. Qin et al., 2019). Este antecedente aporta evidencia clave de que las DCGAN pueden abordar desafíos de conjuntos desbalanceados, mejorando la precisión y estabilidad de los modelos de clasificación.

Figura 8.

Arquitectura de DCGAN para la generación de CXRs.



Nota. Adaptado de "Learning from an Imbalanced and Limited Dataset and an Application to Medical Imaging" (p. 4), por X. Qin, F. M. Bui y H. H. Nguyen, 2019, *2019 IEEE Pacific Rim Conference on Communications*, Computers and Signal Processing *(PACRIM)* (pp. 1-6), IEEE.

Sadi et al. (2022) llevaron a cabo "LMFLOSS: A hybrid loss for imbalanced medical image classification" en Bangladesh, abordando el desbalance de clases en clasificación de imágenes médicas que afecta el rendimiento de modelos de aprendizaje profundo. Propusieron "Large Margin aware Focal Loss" (LMFLOSS), una función de pérdida híbrida que combina las ventajas de *focal loss* y *LDAM loss*, dando mayor relevancia a muestras difíciles de clasificar y garantizando márgenes amplios para clases minoritarias. Emplearon cuatro conjuntos de datos médicos: ODIR-5K (enfermedades oculares), HAM, ISIC-2019 (lesiones cutáneas) y COVID-19 *Radiography Dataset* (neumonía). Utilizaron tres arquitecturas CNN preentrenadas (ResNet50, EfficientNetV2, DenseNet121) entrenadas con LMFLOSS y funciones de referencia (CCE, *focal loss*, *LDAM loss*, BWCCE). El análisis incluyó métricas de precisión, recuperación y macro *F1-score*, usando Grad-CAM para visualizar áreas de interés. Los resultados mostraron que LMFLOSS superó consistentemente a otras funciones en los cuatro conjuntos y tres arquitecturas, mejorando las puntuaciones macro F1 entre 2% y 9%. Los mapas de atención revelaron que los modelos enfocaron predicciones en características relevantes, mejorando la precisión general (Figura 9). En conclusión, LMFLOSS es una solución efectiva para mitigar el desbalance de clases en datos médicos (Sadi et al., 2022), mejorando el rendimiento de modelos de aprendizaje profundo y potenciando la precisión en clases minoritarias subrepresentadas.

Figura 9.

Comparación del rendimiento de la pérdida LMF propuesta con otras funciones de pérdida en EfficientNetV2.

Dataset	Loss	Accuracy	Precision	Recall	Macro F1
ODIR-5K	CCE loss	67.64	65.55	58.21	60.34
	Focal loss	67.36	62.75	60.23	60.84
	LDAM loss	66.01	58.78	57.54	57.43
	BWCCE loss	68.51	66.40	**61.98**	**63.41**
	LMF loss **(ours)**	**69.86**	**67.51**	61.23	62.94
HAM-10K	CCE loss	89.65	85.64	78.80	81.85
	Focal loss	88.55	82.46	79.81	80.75
	LDAM loss	89.48	84.40	80.34	82.12
	BWCCE loss	90.01	84.53	80.64	81.99
	LMF loss **(ours)**	**91.21**	**86.55**	**83.27**	**84.61**
ISIC-2019	CCE loss	83.91	77.66	73.68	75.46
	Focal loss	80.52	71.62	72.23	71.73
	LDAM loss	85.36	80.57	76.85	78.50
	BWCCE loss	83.68	78.96	76.41	77.37
	LMF loss **(ours)**	**86.21**	**82.04**	**78.79**	**80.15**
Covid-19	CCE loss	95.87	97.06	96.59	96.80
	Focal loss	95.37	95.90	96.09	95.99
	LDAM loss	96.38	97.15	97.10	97.11
	BWCCE loss	95.97	97.00	96.68	96.84
	LMF loss **(ours)**	**96.66**	**97.42**	**97.36**	**97.38**

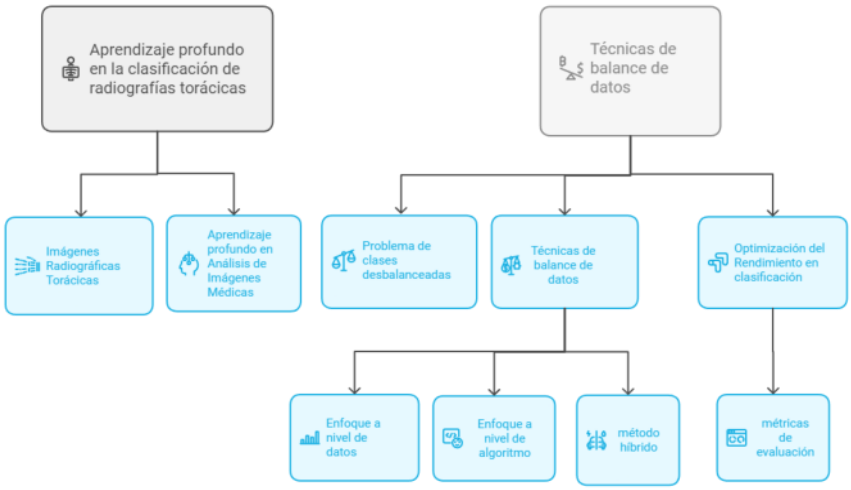
Nota. Adaptado de "LMFLOSS: A Hybrid Loss for Imbalanced Medical Image Classification" (p. 11), por A. A. Sadi, L. Chowdhury, N. Jahan, M. N. S. Rafi, R. Chowdhury, F. A. Khan y N. Mohammed, 2022, arXiv. https://arxiv.org/abs/2212.12741

3.2. Bases teóricas

La construcción de un marco teórico sólido es esencial para fundamentar cualquier investigación científica. Una manera rápida y eficaz de construir este marco consiste en desarrollar, en primer lugar, un índice tentativo de los temas relevantes, afinándolo progresivamente hasta alcanzar un nivel de especificidad adecuado. Este proceso, que puede denominarse "vertebrar" el marco teórico, implica generar la columna vertebral del mismo, colocando la información y las referencias en el lugar correspondiente dentro del esquema como en la Figura 10.

Figura 10.

Método por índices para construir las bases teóricas del marco teórico



3.2.1. Rendimiento de modelos de aprendizaje profundo en clasificación de radiografías torácicas

La variable dependiente es el rendimiento de modelos de aprendizaje profundo en clasificación de radiografías torácicas (

Figura 11), que mide la capacidad de estos modelos para identificar patrones relevantes en imágenes médicas, optimizando métricas como precisión, sensibilidad y especificidad mediante el análisis de datos preprocesados y el uso de arquitecturas especializadas como CNN.

Figura 11.

Variable Dependiente - Marco Teórico.



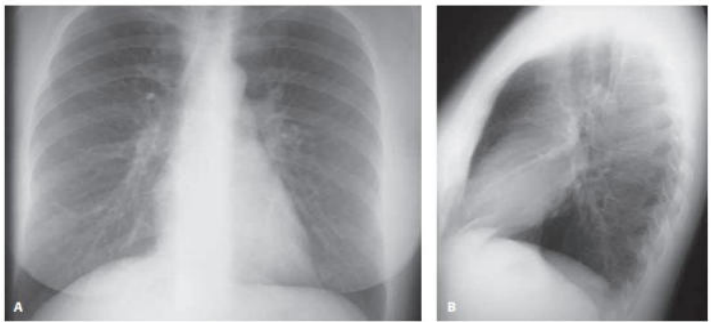
3.2.1.1. Imágenes radiográficas torácicas. La radiografía de tórax constituye uno de los pilares fundamentales en el diagnóstico médico por imagen, destacándose como una herramienta esencial en la evaluación inicial de la patología torácica. Su relevancia radica no solo en su accesibilidad y costo-efectividad, sino también en su capacidad para proporcionar información diagnóstica crucial de manera no invasiva (Vargas Romero, s.d.).

La obtención de radiografías torácicas se fundamenta en principios técnicos específicos que garantizan su calidad diagnóstica. La proyección poster o anterior (PA), complementada con la proyección lateral, Figura 12, configura el estudio básico inicial. Este procedimiento requiere condiciones específicas de adquisición (Reed, 2019):

- La inspiración debe ser máxima y sostenida.
- El posicionamiento del paciente debe ser preciso y estandarizado.
- La técnica radiográfica debe optimizarse para lograr una penetración adecuada.

Figura 12.

Proyección PA y lateral izquierda del tórax.



Nota. Adaptado de *Radiología del Tórax. Patrones y Diagnósticos Diferenciales* (7ª ed.), por J. C. Reed, 2019, Amolca.

La evolución hacia la radiografía digital ha revolucionado la práctica clínica, ofreciendo:

- Mayor resolución de contraste.
- Capacidad de post-procesamiento.
- Optimización en el almacenamiento y distribución de imágenes.
- Reducción en la dosis de radiación.

3.2.1.1.1. Características y patrones en radiografías torácicas. La interpretación radiográfica se fundamenta en la diferenciación de densidades radiológicas, que se manifiestan en una escala de grises determinada por la atenuación de los rayos X al atravesar los diferentes tejidos. Esta diferenciación incluye, Tabla 2 y Tabla 3:

Tabla 2. *Densidades radiológicas básicas en radiografía torácica*

Densidad Radiológica	Características	Ejemplos
Aire	Negro radiológico (radiolucencia máxima, sin atenuación de rayos X)	Pulmones, vías aéreas, neumotórax
Grasa	Gris oscuro (radiodensidad ~-100 UH, baja atenuación)	Tejido subcutáneo, grasa mediastínica
Agua/Tejidos blandos	Gris claro (radiodensidad 0-100 UH, atenuación moderada)	Músculos, órganos sólidos, derrame pleural
Calcio/Hueso	Blanco (radiopacidad máxima, alta atenuación >400 UH)	Estructuras óseas, calcificaciones

Tabla 3.

Patrones radiológicos fundamentales en radiografía torácica

Patrón radiológico	Características principales	Manifestaciones clínicas
Patrón Alveolar	Bordes mal definidos Tendencia a confluencia Broncograma aéreo Alveolograma	Neumonía Edema pulmonar Hemorragia alveolar Contusión pulmonar
Patrón Intersticial	Opacidades lineales Opacidades nodulares Patrón reticular Afectación peribronquial	Fibrosis pulmonar Enfermedad metastásica Sarcoidosis Linfangitis carcinomatosa
Nódulos y Masas	Tamaño (<3cm nódulos) Bordes definidos/espiculados	Neoplasias primarias Metástasis

	Calcificaciones	Granulomas
	Cavitación	Abscesos
Alteraciones Pleurales	Signo del menisco	Derrame pleural
	Borramiento de ángulos	Empiema
	Engrosamiento pleural	Mesotelioma
	Neumotórax	Metástasis pleurales

3.2.1.1.2. Preprocesamiento de radiografías torácicas

a) Ecualización adaptativa del histograma (AHE). La Ecualización Adaptativa del Histograma (AHE) representa una evolución significativa en las técnicas de procesamiento digital de imágenes para el mejoramiento del contraste. A diferencia de la ecualización tradicional, este método implementa un enfoque localizado que calcula múltiples histogramas correspondientes a diferentes secciones de la imagen, utilizándolos para redistribuir los valores de luminancia, lo que permite mejorar el contraste en regiones específicas de la imagen (Zeyun Yu & Bajaj, 2004).

Esta aproximación resulta particularmente efectiva para:

- Mejoramiento del contraste local.
- Definición mejorada de bordes en regiones específicas.
- Adaptación a variaciones locales en la imagen.

Interpolación bilineal en el mapeo de píxeles. Para abordar los artefactos de bloque generados por la ecualización local del histograma, se implementa la interpolación bilineal. Este proceso considera cuatro regiones contextuales, donde los centros de cada región actúan como puntos de referencia para el mapeo de píxeles (Barrios Roqueme, 2022).

La transformación de un píxel considerando cuatro regiones contextuales se expresa mediante la siguiente Ecuación (*1*):

$$s = (1 - y)[(1 - x)TR_1(r) + xTR_2(r)] + y[(1 - x)TR_3(r) + xTR_4(r)] \quad (1)$$

donde $TR_i(r)$ representa la función de transformación para cada región contextual.

b) Ecualización Adaptativa del Histograma con Limitación de Contraste (CLAHE). CLAHE representa una mejora significativa sobre AHE, diferenciándose principalmente en su enfoque para el manejo del contraste. Mientras que AHE realiza una ecualización normal del histograma en diferentes regiones, CLAHE implementa una distribución controlada que optimiza la calidad del resultado final.

Las ventajas principales de CLAHE incluyen:

- Supresión efectiva de la amplificación del ruido.
- Mayor eficiencia computacional mediante algoritmos de interpolación bilineal.
- Control mejorado sobre el realce del contraste (López et al., 2019).

3.2.1.2. Aprendizaje Profundo en Análisis de Imágenes Médicas. El cerebro humano representa el sistema de procesamiento de información más sofisticado de la naturaleza, demostrando una notable eficiencia en el manejo de diversos estímulos sensoriales, incluyendo información visual, auditiva, táctil, gustativa y olfativa. En su esencia, la efectividad del cerebro emerge de un principio organizacional fundamental: el procesamiento de información compleja de alto nivel surge de la actividad coordinada de numerosas unidades de procesamiento simple conectadas mediante uniones sinápticas (Zhou et al., 2023).

En el dominio del aprendizaje en redes, emergen dos desafíos principales:

- Aprendizaje de la arquitectura de red
- Aprendizaje de los parámetros de red

Mientras que el aprendizaje de la arquitectura permanece como una pregunta abierta de investigación, se han logrado avances algorítmicos sustanciales en el aprendizaje de parámetros, que examinaremos en detalle a continuación.

3.2.1.2.1. Marco del aprendizaje de parámetros. Consideremos el problema del aprendizaje de parámetros para una red neuronal con $\boldsymbol{L}$ capas. Dado un conjunto de entrenamiento $\{\boldsymbol{x_n}, \boldsymbol{t_n}\}_{\boldsymbol{n=1}}^{\boldsymbol{N}}$, donde $\boldsymbol{x_n} \in \boldsymbol{R^D}$ representa una observación y $\boldsymbol{t_n} \in \{\boldsymbol{0}, \boldsymbol{1}\}^{\boldsymbol{K}}$ denota un vector indicador de clase, comúnmente se emplea la función de error de entropía cruzada:

$$E(\mathbf{W}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk} \tag{2}$$

Aquí, t_{nk} representa el k-ésimo elemento del vector objetivo $\mathbf{t}_n$, y_{nk} denota el k-ésimo elemento del vector de predicción para $\mathbf{x}_n$, y $\mathbf{W} = [\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \dots, \mathbf{W}^{(L)}]$ comprende los parámetros de la red.

3.2.1.2.2. Optimización basada en gradientes. La función de error exhibe alta no linealidad y no convexidad, lo que impide soluciones analíticas directas. Por lo tanto, empleamos la optimización por descenso de gradiente, actualizando los parámetros iterativamente según:

$$\mathbf{W}_{(t+1)} = \mathbf{W}_{(t)} - \eta \nabla E(\mathbf{W}_{(t)})$$

donde τ denota el índice de iteración y η representa la tasa de aprendizaje. El gradiente $\nabla E(\mathbf{W})$ se calcula mediante retropropagación del error, propagando los errores desde la capa de salida hasta la capa de entrada mediante la regla de la cadena:

$$\frac{\partial E}{\partial \mathbf{W}^{(l)}} = \frac{\partial E}{\partial \mathbf{a}^{(L)}} \frac{\partial \mathbf{a}^{(L)}}{\partial \mathbf{a}^{(L-1)}} \cdots \frac{\partial \mathbf{a}^{(l+1)}}{\partial \mathbf{a}^{(l)}} \frac{\partial \mathbf{a}^{(l)}}{\partial \mathbf{W}^{(l)}}$$

donde $\mathbf{z}^{(l)}$ y $\mathbf{a}^{(l)}$ denotan los vectores de pre-activación y activación de la capa l, respectivamente, con $\mathbf{a}^{(1)} = \mathbf{x}$.

3.2.1.2.3. Redes neuronales convolucionales (CNN). En las redes neuronales multicapas convencionales, las entradas siempre están en forma vectorial. Sin embargo, para imágenes (médicas), la estructura y la información configuracional entre píxeles o vóxeles vecinos es otra fuente de información. Por lo tanto, la vectorización inevitablemente destruye dicha información estructural y configuracional en las imágenes.

Una CNN que típicamente tiene capas convolucionales intercaladas con capas de agrupación (o submuestreo) y luego seguidas por capas totalmente conectadas en una red neuronal multicapa estándar, está diseñada para utilizar mejor dicha información espacial y configuracional mediante la extensión 2D o 3D de las entradas. A diferencia de las redes neuronales multicapas convencionales, una CNN extiende el compartimiento de pesos para reducir los grados de libertad de los modelos. Una capa de agrupación puede reducir el tiempo de cómputo y gradualmente construir información espacial y configuracional (Zhou et al., 2023).

a) Capa de convolución y agrupación. El rol de una capa de convolución es detectar características locales en diferentes posiciones en los mapas de características de entrada con kernels aprendibles $\boldsymbol{k_{ij}^{(l)}}$, es decir, pesos de conexión entre el mapa de características $\boldsymbol{i}$ en la capa $\boldsymbol{l}$ y el mapa de características $\boldsymbol{j}$ en la capa $\boldsymbol{l-1}$ (Zhou et al., 2023). Específicamente, las unidades de la capa de convolución $\boldsymbol{l}$ calculan sus activaciones $\boldsymbol{A_{ij}^{(l)}}$ basándose solo en un subconjunto espacialmente contiguo de unidades en los mapas de características $\boldsymbol{A^{(l-1)}}$ de la capa precedente $\boldsymbol{l-1}$, convolucionando los kernels $\boldsymbol{k_{ij}^{(l)}}$ como sigue en la siguiente ecuación:

$$A_{ij}^{(l)} = f\left(\sum_{i=1}^{M^{(l-1)}} A_i^{(l-1)} * k_{ij}^{(l)} + b_{ij}^{(l)}\right)$$

donde $M^{(l-1)}$ denota el número de mapas de características en la capa $l-1$, $*$ denota un operador de convolución, $b_{ij}^{(l)}$ es un parámetro de sesgo y $f(\cdot)$ es una función de activación no lineal.

b) Cálculo de gradientes. Asumiendo que una capa de convolución es seguida por una capa de agrupación, las unidades en un mapa de características de una capa convolucional $\boldsymbol{l}$ están conectadas a una única unidad del mapa correspondiente en la capa de agrupación $\boldsymbol{l+1}$ (Zhou et al., 2023). Al hacer *up-sampling* de los mapas de características de la capa de agrupación para recuperar el tamaño reducido de los mapas, todo lo que necesitamos hacer es multiplicar el gradiente de la función de activación calculada en la capa de convolución con las pre-activaciones $\boldsymbol{Z_{ij}^{(l)}}$ como sigue en la ecuación:

$$\delta_{ij}^{(l)} = f'(Z_{ij}^{(l)}) \odot \mathrm{up}(\delta_{i,j}^{(l+1)})$$

donde $\odot$ y $\mathrm{up}(\cdot)$ denotan multiplicación elemento a elemento y operación de *up-sampling*, respectivamente.

3.2.1.2.4. Redes neuronales Convolucionales profundas. Con los avances en hardware de computación, trabajos recientes utilizando redes neuronales han incrementado tanto la profundidad (es decir, capas de convolución y agrupación) como el ancho (es decir, tamaño del canal). Sin embargo, a medida que las CNN se vuelven más profundas y anchas, las dificultades (por ejemplo, costo computacional, desvanecimiento de gradientes y degradación) en su entrenamiento también crecen (Zhou et al., 2023).

a) Módulo Inception. El módulo Inception se introdujo para reducir significativamente el costo computacional de las redes neuronales profundas mediante procesamiento multi-escala disperso. Específicamente, reduce el número de operaciones aritméticas de convolución reduciendo el tamaño del filtro, es decir, introduciendo dispersión.

Las arquitecturas Inception-v3 (Zhou et al., 2023) utilizan tamaños de *kernel* aún más pequeños en las operaciones de convolución para mejorar la eficiencia computacional. Inception-v4 e Inception-ResNet-v1/v2 incluyen la construcción de conexiones *skip* además de las operaciones de convolución multi-escala.

b) Mecanismo de Atención. El mecanismo de atención en aprendizaje profundo permite que las redes neuronales atiendan a información saliente de datos ruidosos y también puede actuar como una función de memoria (Zhou et al., 2023).

En la práctica, los mecanismos de atención para análisis de imágenes médicas típicamente utilizan atención por canal y espacial, así como atención global para mejorar el rendimiento del modelo. Es importante notar que estas técnicas de atención también se han utilizado como herramienta para métodos de interpretación visual (Zhou et al., 2023), donde las regiones más atendidas pueden localizar las características que respaldan la decisión tomada por una red neuronal.

3.2.2. Técnicas de balance de datos

3.2.2.1. Problema de clases desbalanceadas. El modelado predictivo de clasificación se centra en prever la etiqueta de clase correspondiente a una observación específica. Un problema de clasificación desequilibrada se presenta cuando los ejemplos en las distintas clases conocidas no están distribuidos de manera uniforme. Esto puede variar desde un ligero sesgo hasta un fuerte desequilibrio, donde hay una cantidad ínfima de ejemplos en la clase minoritaria en comparación con los cientos, miles o millones de ejemplos en la clase o clases mayoritarias (Brownlee, 2020).

Los problemas de clasificación desequilibrada suponen un reto para el modelado predictivo, ya que la mayoría de los algoritmos de aprendizaje

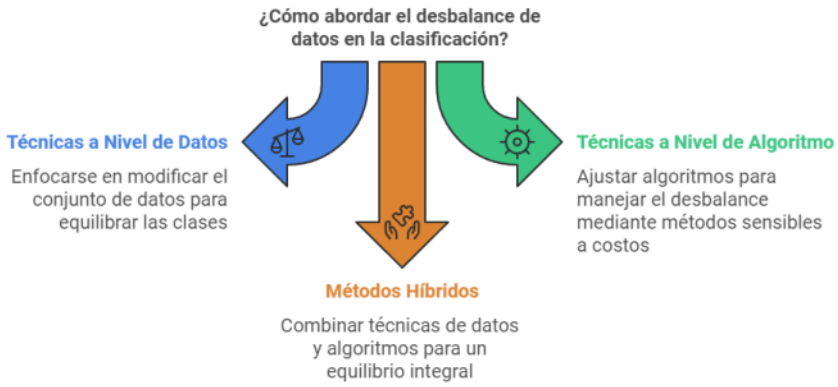
automático para clasificación se han desarrollado bajo la premisa de un número similar de ejemplos para cada clase. Esto da lugar a modelos que suelen tener un bajo rendimiento predictivo, especialmente en lo que respecta a la clase minoritaria (Brownlee, 2020).

Este problema es preocupante porque, por lo general, la clase minoritaria tiene una importancia mayor y el sistema es más vulnerable a errores de clasificación en esa clase en comparación con las clases mayoritarias.

El desequilibrio de clases en un problema se refiere a la manera en que están distribuidas las distintas clases en un conjunto de datos de entrenamiento específico. Esta desigualdad debe ser evaluada en relación con el conjunto de datos en cuestión, ya que las etiquetas de clase son necesarias para determinar el grado de desequilibrio, lo cual generalmente se hace en función de la distribución de entrenamiento (He & Garcia, 2009), en la Figura 13, se observa cómo abordar el desbalance.

Figura 13.

¿Cómo abordar el desbalance de datos en la clasificación?

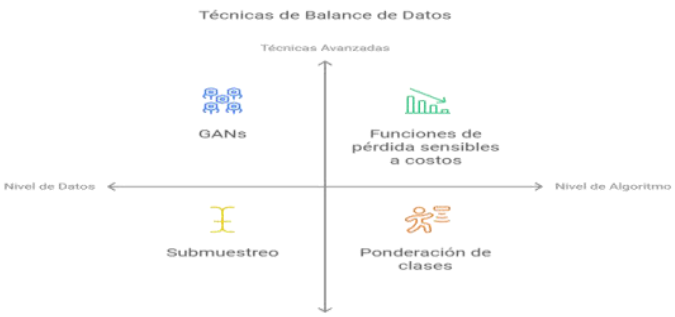


Es común describir el desequilibrio de clases en un conjunto de datos utilizando una proporción. Por ejemplo, en un problema de clasificación binaria desequilibrada con una proporción de 1:100, significa que, por cada ejemplo de una clase, hay 100 ejemplos de la otra. Otra forma de expresar el desequilibrio de clases es mediante un resumen de la distribución de clases en términos de porcentajes del conjunto de datos de entrenamiento (Brownlee, 2020).

En los últimos años, se han propuesto varias técnicas de aprendizaje de vanguardia para abordar el problema de clasificación en conjuntos de datos desequilibrados. Estas técnicas se han desarrollado dentro del marco introductorio del modelado de aprendizaje automático. Se pueden identificar dos enfoques básicos para abordar el aprendizaje de datos desequilibrados: los métodos a nivel de algoritmo y los métodos a nivel de datos, además de un tercero el método híbrido que nos presenta la fusión de los antes métodos mencionados, Figura 14. El enfoque impulsado por algoritmos busca equilibrar la distribución de clases, mientras que el enfoque impulsado por datos intenta corregir el algoritmo de aprendizaje o el clasificador sin modificar el conjunto de entrenamiento (Kumar et al., 2021).

Figura 14.

Técnicas de Balance de Datos.



3.2.2.2. Enfoque a nivel de datos. Los métodos a nivel de datos buscan equilibrar la distribución modificando un grupo de ejemplos. Esto se puede lograr mediante la adición o eliminación de muestras complicadas. El conjunto de entrenamiento se ajusta para adaptarse a los algoritmos de aprendizaje estándar. La estrategia que genera nuevos ejemplos y los añade a los grupos minoritarios se conoce como sobre muestreo, mientras que la estrategia que elimina ejemplos de los grupos mayoritarios se llama submuestreo. Existen enfoques estándar que emplean una selección aleatoria de muestras objetivo para el preprocesamiento. Esto puede resultar en la introducción de muestras sin sentido o la eliminación de muestras importantes.

Los métodos propuestos se han mejorado para conservar la estructura del grupo o para generar nuevos datos de acuerdo con las distribuciones subyacentes. Estos métodos ofrecen una forma de limpiar los objetos superpuestos y eliminar

muestras inadecuadas que pueden afectar el rendimiento de los algoritmos de aprendizaje (Yadav* & Bhole, 2020).

3.2.2.1.1. Método de muestreo. Los métodos de muestreo se consideran una de las estrategias más utilizadas para abordar problemas de aprendizaje desequilibrado debido a su simplicidad. Estos métodos implican modificar un conjunto de datos con distribuciones desequilibradas para lograr una distribución más equilibrada. Entre los enfoques más destacados se encuentran el sobre muestreo aleatorio, el submuestreo aleatorio, el muestreo sintético con generación de datos, los métodos de muestreo basados en conglomerados y la integración de muestreo con refuerzo (He & Garcia, 2009).

Los métodos de muestreo deben generar una distribución que sea más manejable para los clasificadores tradicionales, aunque no necesariamente tiene que ser perfectamente equilibrada.

Dos de las primeras técnicas de muestreo desarrolladas fueron el **submuestreo aleatorio** (RUS) y el **sobremuestreo aleatorio** (ROS). En el caso del submuestreo aleatorio, se eliminan instancias de la clase mayoritaria de forma aleatoria hasta alcanzar una distribución más equilibrada. Por ejemplo, si un conjunto de datos tiene 10 instancias de la clase minoritaria y 100 de la clase mayoritaria, el RUS eliminaría 90 instancias de la clase mayoritaria, dejando un conjunto de datos con 20 instancias: 10 instancias de cada clase (Hoens & Chawla, 2013).

En el caso del sobre muestreo aleatorio, se duplican las instancias de la clase minoritaria para alcanzar una distribución más equilibrada (Hoens & Chawla, 2013). Si bien estos métodos pueden crear distribuciones más equilibradas, ambos tienen desventajas importantes. Por ejemplo, el RUS puede resultar en la eliminación de una gran cantidad de datos. En el ejemplo anterior, aproximadamente el 82% de los datos (90 instancias de la clase mayoritaria) se descartarían, lo que puede complicar el aprendizaje de los límites de decisión entre las clases y disminuir el rendimiento de la clasificación (Hoens & Chawla, 2013).

Por otro lado, el sobre muestreo aleatorio puede provocar un sobreajuste significativo, ya que las instancias se duplican muchas veces para equilibrar la

distribución. Por ejemplo, en el ejemplo anterior, cada instancia tendría que duplicarse 49 veces, lo que puede llevar a un sobreajuste que reduce la capacidad de generalización del clasificador (Hoens & Chawla, 2013).

Para superar estas limitaciones, se han desarrollado técnicas de muestreo más avanzadas. A continuación, se presentan algunas de estas técnicas.

1. **Técnicas de submuestreo:** El principal problema del RUSes que puede descartar información potencialmente útil al elegir muestras de manera aleatoria. Para mitigar este problema, se han desarrollado diversas técnicas que buscan conservar toda la información útil presente en la clase mayoritaria eliminando instancias redundantes, ruidosas o cercanas al límite (Hoens & Chawla, 2013).

   Un enfoque temprano es el de Kubat y Matwin, que combina los enlaces Tomek y una versión modificada de la regla del vecino más cercano condensado para crear un método de submuestreo dirigido. Los enlaces Tomek eliminan instancias cercanas al límite o ruidosas, mientras que la regla elimina instancias redundantes (Hoens & Chawla, 2013).

   Los enlaces Tomek identifican pares de instancias de diferentes clases que están muy cercanas entre sí. Estas instancias suelen ser límites de decisión incorrectos o instancias ruidosas (Hoens & Chawla, 2013).

   Por otro lado, la regla que elimina instancias cercanas comienza con un pequeño conjunto de instancias que incluye todas las de la clase minoritaria y una instancia aleatoria de la clase mayoritaria. Luego, clasifica todas las instancias y agrega al conjunto aquellas que son clasificadas incorrectamente. Esto permite eliminar instancias redundantes (Hoens & Chawla, 2013).

   Otra técnica alternativa para submuestreo es la **regla de limpieza de vecindario** (NCR), que utiliza la regla de Wilson del **vecino más cercano editado** (ENN) para seleccionar instancias de la clase mayoritaria que deben eliminarse. Si una instancia de la clase mayoritaria es mal clasificada por

sus tres vecinos más cercanos, se elimina del conjunto de datos (Hoens & Chawla, 2013).

2. **Técnicas de sobremuestreo:** El sobre muestreo aleatorio puede llevar a un sobreajuste en el modelo, ya que replicar aleatoriamente instancias puede hacer que el modelo se ajuste demasiado a los datos de entrenamiento y no se generalice bien a nuevos datos (Hoens & Chawla, 2013).

   Para superar este problema, Chawla et al. desarrollaron la **técnica de sobremuestreo de minorías sintéticas** (SMOTE), que genera instancias sintéticas en lugar de copiar instancias existentes. Esto equilibra la distribución de clases al añadir instancias sintéticas generadas a partir de ejemplos minoritarios existentes (Hoens & Chawla, 2013).

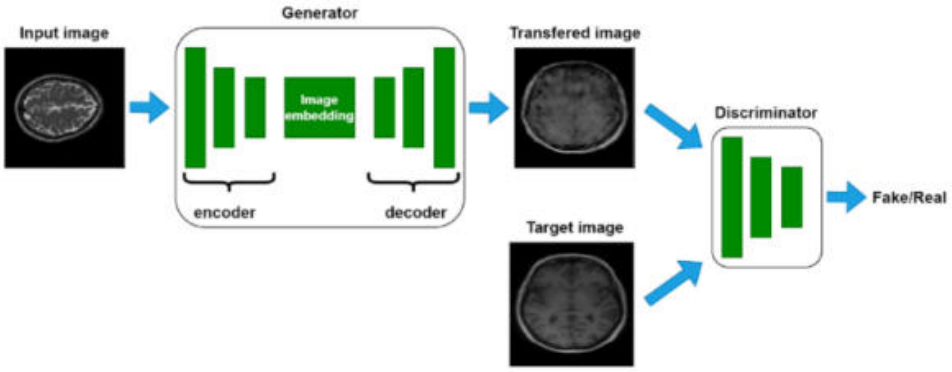
   El método SMOTE funciona seleccionando una instancia minoritaria aleatoriamente y encontrando sus k vecinos más cercanos también de la clase minoritaria. Luego, se crea una instancia sintética al elegir uno de esos vecinos al azar y combinándolos para formar una línea en el espacio de características. Las instancias sintéticas son una combinación convexa de las dos instancias seleccionadas (Hoens & Chawla, 2013).

   Debido a su efectividad, SMOTE ha sido ampliado en diversas variantes, como **SMOTE de frontera** (*Borderline-SMOTE*), que solo aplica el algoritmo a instancias fronterizas, o **SMOTE de nivel seguro** (Safe-Level-SMOTE), que define un nivel de seguridad para cada instancia antes de aplicar SMOTE (Hoens & Chawla, 2013).

3.2.2.1.2. Redes Generativas Antagónicas (GAN). Las GAN son modelos de aprendizaje profundo formados por dos redes neuronales: un generador y un discriminador, entrenados de manera *adversarial*, Figura 15. Este enfoque permite generar datos sintéticos que emulan distribuciones reales, siendo particularmente útiles en tareas como síntesis de imágenes, aumento de datos, detección de anomalías y adaptación de dominios. En el análisis de imágenes médicas, las GAN han demostrado ser herramientas eficaces al generar imágenes médicas sintéticas que ayudan en la construcción de modelos de aprendizaje profundo robustos, incluso cuando los datos etiquetados son limitados (Ashok et al., 2024).

Figura 15.

Redes Generativas Antagónicas (GAN).



Nota. Adaptado de "Advanced Techniques in Medical Imaging: Computer Vision and Machine Learning Approaches", por M. Ashok, N. Rajeswaran, A. Atheeswaran y V. VijayaRangan, 2024, Quing Publications. https://books.google.com.pe/books?id=44skEQAAQBAJ

a) Estructura de las GAN

1. **Generador**: Esta red toma como entrada un vector de ruido aleatorio o un espacio latente y genera muestras sintéticas que buscan emular las características de los datos reales. El objetivo del generador es producir muestras indistinguibles de los datos reales según el discriminador.

2. **Discriminador**: Funciona como un clasificador binario que diferencia entre datos reales y datos generados. Su objetivo es asignar altas probabilidades a los datos reales y bajas a los sintéticos. Durante el entrenamiento,

proporciona retroalimentación al generador para mejorar la calidad de las muestras generadas.

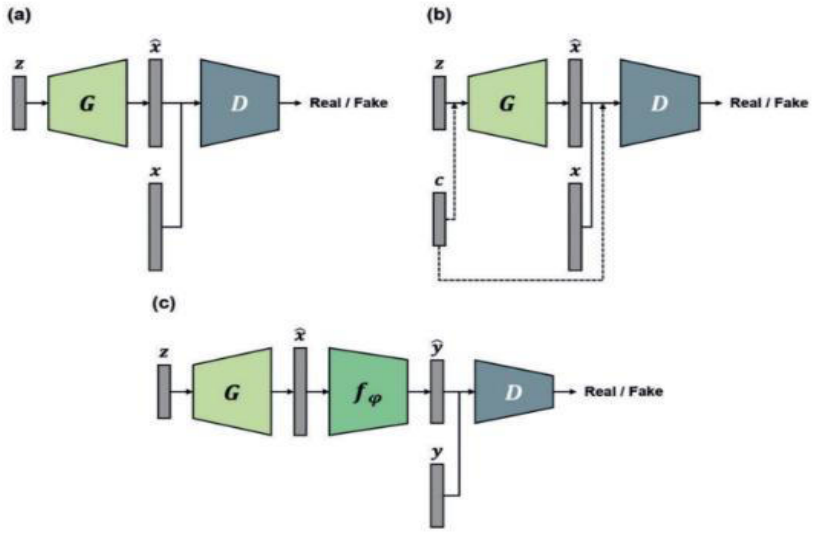
3. **Entrenamiento adversarial**: Las GAN se entrenan mediante un marco de aprendizaje *adversarial*, Figura 16, resolviendo un problema de minimax dado por la siguiente ecuación:

$$\min_G \max_D L_{\text{GAN}}(D,G)\mathbf{E}_{x\sim p_{\text{data}}}[\log D(x)] + \mathbf{E}_{z\sim p_z}[\log(1-D(G(z)))],$$

donde G es el generador, D es el discriminador, x representa los datos reales provenientes de la distribución p_{data}, y z corresponde al ruido tomado de la distribución p_z (Zhou et al., 2023).

Figura 16.

Procedimiento de entrenamiento de GAN para generación de imágenes. (a) GAN, (b) cGAN y (c) Ambient-GAN. (GAN).



Nota. Adaptado de Deep Learning for Medical Image Analysis , por S. K. Zhou, H. Greenspan y D. Shen, 2023, Academic Press.

El proceso *adversarial* impulsa al generador a producir datos cada vez más realistas, mientras que el discriminador mejora su capacidad para distinguir entre muestras reales y generadas. Este juego de suma cero converge cuando el generador

produce datos que el discriminador ya no puede diferenciar de los reales (Ashok et al., 2024).

b) Aplicaciones de las GAN en imágenes médicas. -

Las GAN han encontrado múltiples aplicaciones en el ámbito de imágenes médicas, entre las cuales destacan:

- **Síntesis de imágenes**: Generación de imágenes sintéticas que reproducen patrones y estructuras presentes en datos reales, ampliando conjuntos de entrenamiento y mejorando la generalización de los modelos.

- **Aumento de datos**: Generación de muestras adicionales para conjuntos de datos limitados, lo cual es especialmente útil cuando las anotaciones son costosas o difíciles de obtener.

- **Detección de anomalías**: Identificación de regiones o características anómalas al aprender distribuciones normales en imágenes médicas, facilitando la detección de patologías.

- **Adaptación de dominios**: Traducción de imágenes entre diferentes dominios o modalidades, como la conversión de imágenes de **tomografía computarizada** (TC) a **resonancia magnética** (RM), mejorando la generalización de modelos entrenados en un dominio específico (Ashok et al., 2024)

c) Wasserstein GAN (WGAN)

La Wasserstein GAN (WGAN), propuesta en 2017, es una arquitectura de redes generativas antagónicas. Se considera una extensión de las GAN tradicionales que emplea una técnica de entrenamiento alternativa para lograr una mejor aproximación (Sindhura et al., 2024).

A diferencia de las GAN convencionales, la WGAN mide la divergencia entre las distribuciones de datos reales y generados utilizando la distancia de Wasserstein (Sindhura et al., 2024).

El objetivo principal de esta arquitectura es mejorar la estabilidad del entrenamiento (Hussain et al., 2022; Keerthi et al., 2024). Las WGAN buscan disminuir el colapso de modo (*mode collapse*) (Hussain et al., 2022), un tipo de fallo en el que el generador (G) queda atrapado generando solo unas pocas salidas "plausibles" (Hussain et al., 2022). Las WGAN hacen que el proceso de entrenamiento sea más estable y menos sensible a la arquitectura del modelo y a los hiperparámetros (Hussain et al., 2022).

- El crítico (*critic*) y la WL. La WGAN introduce una modificación fundamental. El discriminador no se utiliza para la clasificación de instancias (decidir si es real o falso), sino que emite un número. Dado que no "discrimina" en el sentido tradicional, convencionalmente se le denomina un crítico (*critic*) entrenamiento del crítico busca que la salida para las instancias reales sea mayor que la salida para las instancias falsas (Hussain et al., 2022).

- La WL está diseñada para prevenir el problema de desvanecimiento de gradientes (*vanishing gradients*), un problema que puede ocurrir si el discriminador (o crítico) es demasiado bueno y se entrena de manera óptima (Hussain et al., 2022).

Las funciones de pérdida descritas en los textos para la WGAN son las siguientes:

- **Pérdida del crítico (***critic loss***):** El crítico (D) intenta maximizar esta función. Esta representa la diferencia entre la salida en instancias reales (x) y la salida en instancias falsas ($G(z)$) (Hussain et al., 2022).

$$\max_D [D(x) - D(G(z))]$$

- **Pérdida del generador (***generator loss***):** El generador (G) intenta maximizar esta función, es decir, trata de maximizar la salida del crítico para las instancias falsas ($G(z)$) que produce (Hussain et al., 2022).

$$\max_G [D(G(z))]$$

d) Variante: WGAN-GP (Penalización de gradiente)

Una variante comúnmente referenciada es la WGAN-GP (Keerthi et al., 2024; Sindhura et al., 2024). Esta técnica se utiliza para mejorar aún más la estabilidad del entrenamiento (Keerthi et al., 2024). La WGAN-GP logra esto penalizando la norma del gradiente de la salida del discriminador (crítico) con respecto a muestras interpoladas entre las imágenes reales y las generadas, este método previene eficazmente el colapso de modo y asegura una generación de imágenes diversa y realista (Keerthi et al., 2024).

3.2.2.2. Enfoque a nivel de algoritmo. En este método, el algoritmo existente se modifica directamente, dando mayor importancia a las instancias pertenecientes a la clase mayoritaria. Se requiere una buena comprensión del algoritmo de aprendizaje modificado. El enfoque sensible al costo es el más popular, ya que, en muchos problemas de desequilibrio, la distribución de los datos y el costo de los errores de clasificación son desiguales. Las técnicas de aprendizaje sensibles al costo tienen en cuenta el costo de la clasificación errónea; asignan un alto costo cuando se clasifica incorrectamente la clase minoritaria (Yadav* & Bhole, 2020).

3.2.2.2.1. Aprendizaje sensible a los costos. El aprendizaje sensible a los costos trata de minimizar el costo total de mala clasificación asignando diferentes costos a las clases o instancias según sean clasificadas de forma correcta o incorrecta. Los métodos de aprendizaje sensibles al costo combinan una matriz de costos con la función de pérdida de *cross-entropy* o **entropía cruzada** (CE) para mejorar la sensibilidad hacia las clases minoritarias. Por ejemplo, las **redes neuronales profundas sensibles al costo** (CSDNN) emplean una matriz de costos predefinida para incrementar la precisión en la predicción de readmisiones hospitalarias (H. Wang et al., 2018).

Sin embargo, determinar los parámetros óptimos en esta matriz puede ser un proceso lento, y las matrices de costos suelen ser específicas de cada tarea, lo que limita su generalización. Para abordar estas limitaciones, se introdujo la **red neuronal convolucional sensible al costo** (CoSen CNN), que permite optimizar simultáneamente los pesos de la red y los parámetros de costos durante el proceso de entrenamiento (Khan et al., 2017)

En el manejo de conjuntos de datos con distintos niveles de desbalance, el cambio de umbrales de decisión se ha comparado con métodos como ROS y RUS. Ajustar los umbrales de las clases utilizando probabilidades estimadas previas ha demostrado reducir de manera efectiva la probabilidad de clasificación incorrecta en las clases minoritarias (Buda et al., 2018).

Para aumentar la sensibilidad de los modelos ante la clasificación errónea de muestras minoritarias, se han propuesto nuevas funciones de pérdida como **error falso medio** (MFE) (S. Wang et al., 2016a), *focal loss* (Lin et al., 2017), *class-balanced* (CB) loss (Cui et al., 2019) y *Class-wise Difficulty Balance* (CDB) Loss (Sinha et al., 2021).

1. **Mean false error loss. -** Cuando se utiliza el **error cuadrático medio** (MSE) como función de pérdida, los errores de las muestras minoritarias suelen ser opacados por los errores de las mayoritarias. Para abordar este problema, se desarrolló la pérdida MFE, que incorpora el **error medio de falsos negativos** (FNE) y el **error medio de falsos positivos** (FPE). Tanto MFE como su versión mejorada, *Mean Squared False Error* (MSFE), han demostrado ser eficaces para abordar el problema del desbalance en diversos conjuntos de datos, además de ser fáciles de implementar y de bajo costo computacional (S. Wang et al., 2016b).

2. **Focal loss.-** La *focal loss* modifica la función de *cross-entropy* al reducir el peso de las muestras fácilmente clasificables, centrando la atención en aquellas más difíciles. Su fórmula es:

$$\text{Focal}(p) = -\alpha(1-p)^{\gamma}\text{Log}(p),$$

donde p es la probabilidad de clasificación correcta de una muestra, $\gamma \geq 0$ controla la importancia de las muestras fácilmente clasificables, y α incrementa el peso de las muestras difíciles, que suelen pertenecer a clases minoritarias. Este enfoque ha demostrado superar varios algoritmos de aprendizaje profundo de una y dos etapas en el conjunto de datos COCO (Lin et al., 2017).

3. **Class-balanced loss. -** Para mejorar el rendimiento en escenarios con desbalance de clases, la pérdida *class-balanced* (CB) ajusta funciones de pérdida existentes, como CE y *focal loss*, basándose en el inverso del número efectivo de muestras de cada clase (Cui et al., 2019). Este enfoque incluye el concepto de número efectivo de muestras, cuyo inverso se incorpora en las funciones de pérdida para optimizar su desempeño.

3.2.2.3. Método híbrido. Los métodos híbridos combinan las ventajas de los enfoques a nivel de algoritmo y a nivel de datos. Estos métodos buscan combinar estas estrategias para maximizar sus fortalezas y minimizar sus debilidades. La combinación de conjuntos de clasificadores con soluciones a nivel de datos produce un sistema de aprendizaje más eficiente, robusto y popular. En algunas aplicaciones, se propone la hibridación de aprendizaje sensible al costo con métodos de muestreo (Yadav* & Bhole, 2020).

3.2.3. Optimización del rendimiento en clasificación

3.2.3.1. Métricas de evaluación específicas. El desafío radica en la selección de métricas de evaluación que permitan cuantificar de manera precisa el rendimiento de un modelo predictivo. Normalmente, esto implica entrenar el modelo con un conjunto de datos y luego utilizarlo para hacer predicciones en un conjunto de datos de prueba independiente. En el contexto de problemas de clasificación, estas métricas implican comparar las etiquetas de clase predichas con las etiquetas de clase esperadas o interpretar las probabilidades predichas para cada clase (Brownlee, 2020). La elección del modelo, así como los métodos de preparación de datos, se basan en gran medida en estas métricas de evaluación, que son fundamentales tanto para evaluar el rendimiento del modelo de clasificación como para guiar su desarrollo (Sun et al., 2009).

Se utilizan comúnmente métricas estándar para evaluar modelos predictivos de clasificación, como la precisión o el error de clasificación. Estas métricas suelen ser efectivas para la mayoría de los problemas, lo que explica su amplia adopción. Sin embargo, todas las métricas implican suposiciones sobre el problema o lo que se considera relevante en el problema. Por eso, es necesario elegir una métrica de evaluación que refleje lo que usted o las partes interesadas del proyecto consideran

esencial sobre el modelo o las predicciones, lo que hace que seleccionar las métricas de evaluación sea un reto (Brownlee, 2020).

Esta tarea se vuelve más complicada cuando existe un sesgo en la distribución de clases. Esto se debe a que muchas métricas estándar se vuelven poco confiables o engañosas cuando las clases están desequilibradas (He & Garcia, 2009).

A diferencia de las métricas de evaluación estándar que tratan todas las clases por igual, los problemas de clasificación desequilibrada suelen considerar que los errores de clasificación en la clase minoritaria son más importantes que en la clase mayoritaria. Por lo tanto, es posible que se necesiten métricas de rendimiento que se centren en la clase minoritaria, lo que puede ser un desafío debido a la falta de observaciones suficientes para entrenar un modelo eficaz (Brownlee, 2020).

3.2.3.2. Taxonomía de métricas de evaluación de clasificadores. Existen decenas de métricas para elegir al evaluar modelos de clasificación, y quizás cientos si se consideran todas las versiones particulares de métricas propuestas por académicos. Para organizar las métricas disponibles, se puede utilizar una taxonomía propuesta por (Ferri et al., 2009).

Podemos dividir las métricas de evaluación en tres grupos principales:

1. Métricas de umbral
2. Métricas de ranking
3. Métricas probabilísticas.

Esta división es útil porque las métricas más utilizadas por los profesionales para clasificadores, en general, y específicamente para clasificación desequilibrada, encajan bien en esta taxonomía.

Varios investigadores en aprendizaje automático han identificado tres familias de métricas de evaluación utilizadas en el contexto de la clasificación: las métricas de umbral (como la precisión y la medida F), los métodos y métricas de ranking (como el análisis de características operativas del receptor o ROC y el área

bajo la curva o AUC) y las métricas probabilísticas (como el error cuadrático medio) (Brownlee, 2020).

3.2.3.2.1. Métricas de umbral para la clasificación desequilibrada. Las métricas de umbral son aquellas que cuantifican los errores de predicción de clasificación. La métrica de umbral más comúnmente utilizada podría ser *accuracy*. Esta métrica mide las predicciones correctas en relación con el total de predicciones realizadas. Por otro lado, el error de clasificación representa las predicciones incorrectas respecto al total de predicciones.

$$Accuracy = \frac{PrediccionesCorrectas}{PrediccionesTotales}$$

$$Error = \frac{PrediccionesIncorrectas}{PrediccionesTotales}$$

A pesar de su amplio uso, la precisión de clasificación suele ser inapropiada para problemas de clasificación desequilibrada. Esto se debe a que un alto nivel de precisión puede ser engañoso, ya que un modelo sin habilidades podría lograrlo simplemente prediciendo la clase mayoritaria. En problemas de clasificación desequilibrada, la clase mayoritaria suele denominarse como el resultado negativo, mientras que la clase minoritaria se refiere al resultado positivo (Brownlee, 2020).

La mayoría de las métricas de umbral se pueden entender mejor mediante los términos utilizados en una matriz de confusión para un problema de clasificación binaria. La matriz de confusión proporciona una visión más detallada no solo del rendimiento de un modelo predictivo, sino también de qué clases se están prediciendo correctamente, cuáles incorrectamente y qué tipo de errores se están cometiendo (Brownlee, 2020). En este tipo de matriz de confusión, cada celda en la tabla tiene un nombre específico y bien entendido, resumido de la siguiente manera (Tabla 4):

Tabla 4.

Matriz de confusión

	Predicción positiva	**Predicción negativa**
Clase positiva	Verdaderos positivos (VP)	Falsos negativos (FN)
Clase negativa	Falsos positivos (FP)	Verdaderos negativos (VN)

Existen dos conjuntos de métricas que pueden resultar útiles para la clasificación desbalanceada, ya que se centran en una clase específica: la sensibilidad-especificidad y la precisión-*recall* (Brownlee, 2020).

Métricas de sensibilidad-especificidad:

La sensibilidad se refiere a la tasa de verdaderos positivos y resume qué tan acertadamente se predijo la clase positiva.

$$Sensibilidad = \frac{VerdaderosPositivos + FalsosNegativos}{VerdaderosPositivos}$$

La especificidad es el complemento de la sensibilidad, o la tasa de verdaderos negativos, y resume qué tan bien se predijo la clase negativa.

$$Especificidad = \frac{FalsosPositivos + VerdaderosNegativos}{VerdaderosNegativos}$$

En la clasificación desbalanceada, la sensibilidad podría ser más relevante que la especificidad. Sensibilidad y especificidad pueden combinarse en un solo puntaje que equilibra ambas preocupaciones, llamado el *G-mean*.

$$G - mean = \sqrt{Sensibilidad \times Especificidad}$$

Métricas de precisión y recuperación:

La precisión resume la fracción de ejemplos asignados a la clase positiva que pertenecen a la clase positiva.

$$Precision = \frac{VerdaderosPositivos}{VerdaderosPositivos + FalsosPositivos}$$

Recall resume qué tan bien se predijo la clase positiva y es el mismo cálculo que la sensibilidad.

$$Recall = \frac{VerdaderosPositivos}{VerdaderosPositivos + FalsosNegativos}$$

La precisión y la recuperación pueden combinarse en un solo puntaje que busca equilibrar ambas preocupaciones, llamado el *F-score* o *F-measure*.

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

3.2.3.2.2. Métricas de ranking para la clasificación desequilibrada. La métrica de clasificación más comúnmente utilizada es la Curva ROC o Análisis ROC. ROC es un acrónimo que significa Característica Operativa del Receptor y resume un campo de estudio para analizar clasificadores binarios según su capacidad para discriminar clases. Una curva ROC es un gráfico diagnóstico para resumir el comportamiento de un modelo mediante el cálculo de la tasa de falsos positivos y la tasa de verdaderos positivos para un conjunto de predicciones realizadas por el modelo bajo diferentes umbrales. La tasa de verdaderos positivos es la tasa de recuperación o sensibilidad (Brownlee, 2020).

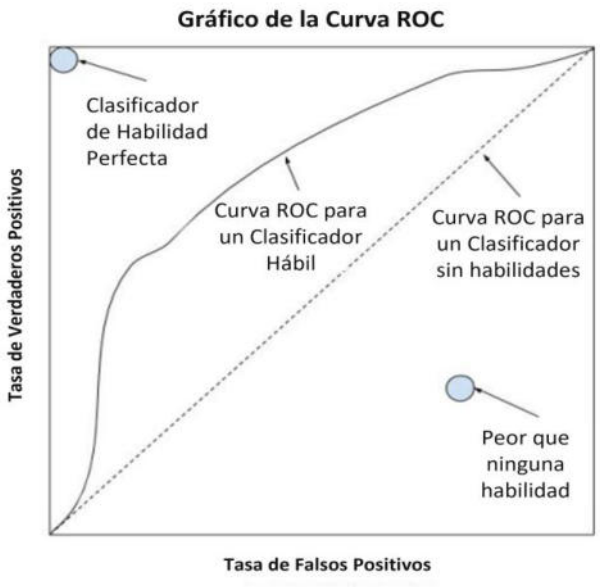
$$TasaVerdaderaPositiva = \frac{VerdaderosPositivos}{VerdaderosPositivos + FalsosNegativos}$$

$$TasaFalsaPositiva = \frac{FalsosPositivos}{FalsosPositivos + VerdaderosNegativos}$$

Cada umbral es un punto en el gráfico y los puntos se conectan para formar una curva. Un clasificador que no tiene habilidad (por ejemplo, predice la clase mayoritaria bajo todos los umbrales) estará representado por una línea diagonal desde la parte inferior izquierda hasta la parte superior derecha. Cualquier punto por debajo de esta línea tiene peor desempeño que ninguna habilidad. Un modelo perfecto será un punto en la parte superior izquierda de la Figura 17.

Figura 17.

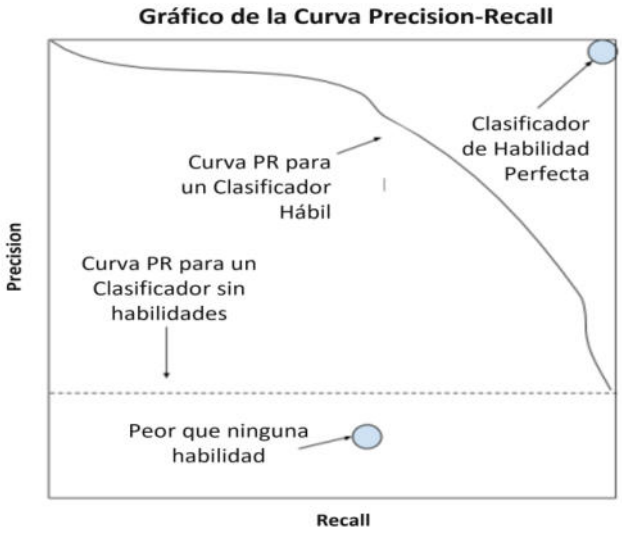
Representación de una curva ROC.



El análisis de la curva ROC es una herramienta útil para evaluar un modelo. El área bajo la curva ROC se puede calcular y proporciona una puntuación única para resumir el gráfico, que se puede utilizar para comparar modelos. Un clasificador sin habilidad tendrá una puntuación de 0.5, mientras que un clasificador perfecto tendrá una puntuación de 1.0. A pesar de ser generalmente efectiva, la curva ROC y el área bajo la curva pueden ser optimistas en un desequilibrio de clase severo, especialmente cuando el número de ejemplos en la clase minoritaria es pequeño. Una alternativa a la curva ROC es la curva de precisión-*recall* como se muestra en la Figura 18, que se puede utilizar de manera similar, aunque se centra en el rendimiento del clasificador en la clase minoritaria (Brownlee, 2020).

Figura 18.

Representación de una curva precesion-recall.



Nuevamente, se utilizan diferentes umbrales en un conjunto de predicciones por un modelo, y en este caso, se calculan la precisión y la recuperación. Los puntos forman una curva y los clasificadores que funcionan mejor bajo una variedad de umbrales diferentes se clasificarán más alto. Un clasificador sin habilidad será una línea horizontal en el gráfico con una precisión que es proporcional al número de ejemplos positivos en el conjunto de datos. Para un conjunto de datos equilibrado, esto será 0.5. Un clasificador perfecto está representado por un punto en la parte superior derecha (Brownlee, 2020).

3.2.3.2.3. Métricas probabilísticas para la clasificación desequilibrada. Las métricas probabilísticas están diseñadas específicamente para cuantificar la incertidumbre en las predicciones de un clasificador. Estas son útiles para problemas en los que estamos menos interesados en las predicciones de clase incorrectas vs. correctas y más interesados en la incertidumbre que el modelo tiene en las predicciones, y en penalizar aquellas predicciones que son incorrectas, pero altamente confiables (Brownlee, 2020).

Nos señala Brownlee (2020) que, al evaluar un modelo basado en las probabilidades predichas, es crucial que estas probabilidades estén calibradas. Algunos clasificadores se entrenan utilizando un marco probabilístico, como la estimación de máxima verosimilitud, lo que significa que sus probabilidades ya

están calibradas. Según Brownlee (2020) una de las métricas más comunes para evaluar las probabilidades predichas es la pérdida logarítmica para la clasificación binaria (o la verosimilitud negativa), conocida más generalmente como entropía cruzada. Para un conjunto de datos de clasificación binaria donde los valores esperados son y y los valores predichos son $\hat{y}$.

Otra métrica ampliamente utilizada para evaluar las probabilidades predichas es el puntaje Brier. Su ventaja radica en su enfoque en la clase positiva, que en problemas de clasificación desequilibrada suele ser la clase minoritaria. Esto lo diferencia de la pérdida logarítmica, que considera toda la distribución de probabilidad. El puntaje Brier se obtiene al calcular el MSE entre las probabilidades esperadas para la clase positiva (por ejemplo, 1.0) y las probabilidades predichas (Brownlee, 2020).

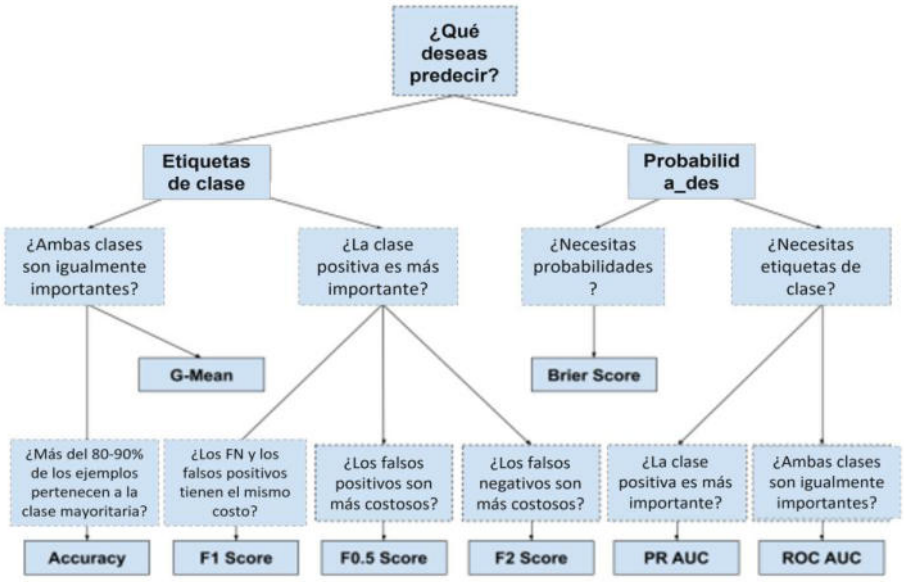
$$BrierScore = \frac{1}{N} \times \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

3.2.3.3. Elección de métricas de evaluación

Según (Brownlee, 2020) una estrategia efectiva podría ser dialogar con las partes interesadas para determinar qué aspectos son fundamentales en un modelo o conjunto de predicciones. Luego, seleccionar algunas métricas y probarlas con diferentes escenarios simulados. Otra opción sería revisar la literatura para descubrir qué métricas son más comúnmente utilizadas por otros profesionales o académicos que trabajan en problemas similares. Sugerencias (Figura 19):

Figura 19.

Cómo elegir una métrica para la clasificación desequilibrada.



¿Estás prediciendo probabilidades?

- ¿Necesitas etiquetas de clase?
    - *¿Es más importante la clase positiva?* Usa el área bajo la curva de Precisión-Recall (AUC).
    - *¿Son importantes ambas clases?* Usa el área bajo la curva ROC (AUC).
- ¿Necesitas probabilidades?
    - *Usa la puntuación Brier y la puntuación de habilidad Brier.*

¿Estás prediciendo etiquetas de clase?

- ¿Es más importante la clase positiva?
    - *¿Son igualmente importantes los falsos negativos y falsos positivos?* Usa la medida F1.
    - *¿Son más importantes los Falsos Negativos?* Usa la medida F2.

    - *¿Son más importantes los falsos positivos?* Usa la medida F0.5.

- ¿Son importantes ambas clases?

    - *¿Tienes menos del 80%-90% de ejemplos para la clase mayoritaria?* Usa la precisión.

    - *¿Tienes más del 80%-90% de ejemplos para la clase mayoritaria?* Usa la media geométrica (*G-mean*).

3.3. Definición de términos

- **ADASYN:** *Adaptive synthetic sampling* (Muestreo sintético adaptativo)
- **AL:** *Active learning* (Aprendizaje activo)
- **ANN:** *Artificial neural network* (Red neuronal artificial)
- **AUC:** *Area under the curve* (Área bajo la curva)
- **AUC-ROC:** *Area under the ROC curve* (Área bajo la curva ROC)
- **AUC-PR:** *Area under the precision-recall curve* (Área bajo la curva de precisión y recall)
- **CB Loss:** *Class-Balanced Loss* (Pérdida Balanceada por Clases).
- **CDB Loss:** *Class-wise Difficulty Balance Loss* (Pérdida de Balance por Dificultad de Clase).
- **CSDNN:** *Cost-Sensitive Deep Neural Network* (Red Neuronal Profunda Sensible al Costo).
- **CV:** *Cross-validation* (Validación cruzada)
- **CNN:** *Condensed nearest rule* (Regla del vecino más cercano condensado)

- **CGAN:** *Conditional Generative Adversarial Network* (Red Generativa Antagónica Condicional).
- **DCGAN:** *Deep Convolutional Generative Adversarial Network* (Red Generativa Antagónica Convolucional Profunda).
- **DM:** *Data Mining* (Minería de datos)
- **DR:** *Dimensionality Reductio*n (Reducción de dimensionalidad)
- **EC:** *Error concentration* (Concentración de errores)
- **EM:** *Expectation-maximization* (Maximización de expectativas)
- **FCV:** *Fold cross-validation* (Validación cruzada por pliegues)
- **FS:** *Feature selection* (Selección de características)
- **GAN:** *Generative Adversarial Network* (Red Generativa Antagónica).
- **IS:** *Instance Selection* (Selección de instancias)
- **KDD:** *Knowledge Discovery in Data* (Descubrimiento de conocimiento en datos)
- **KEEL:** *Knowledge extraction based on evolutionary learning* (Extracción de conocimiento basada en aprendizaje evolutivo)
- **KNN:** *K-Nearest neighbors* (K-Vecinos más cercanos)
- **LLE:** *Locally Linear Embedding* (Incrustación lineal local)
- **LVQ:** *Learning Vector Quantization* (Cuantificación vectorial de aprendizaje)
- **MCC:** *Matthews Correlation Coefficient* (Coeficiente de correlación de Matthews)
- **MDS:** *Multidimensional scaling* (Escalado multidimensional)

- **MI:** *Mutual information* (Información mutua)
- **MIL:** *Multi-instance learning* (Aprendizaje multi-instancia)
- **ML:** *Machine learning* (Aprendizaje automático)
- **MLL:** *Multilabel learning* (Aprendizaje multietiqueta)
- **MLP:** *Multilayer perceptron* (Perceptrón multicapa)
- **MV:** *Missing value* (Valor faltante)
- **NCL:** Neighborhood cleaning rule (Regla de limpieza de vecindario)
- **NN:** *Nearest Neighbor* (Vecino más cercano)
- **OSS:** *One-sided selection* (Selección unilateral)
- **PCA:** *Principal Components Analysis* (Análisis de componentes principales)
- **pGAN:** *Paired Generative Adversarial Network* (Red Generativa Antagónica Pareada).
- **PU-learning:** *Positive and unlabeled learning* (Aprendizaje positivo y no etiquetado)
- **RBFN:** *Radial basis function network* (Red de funciones de base radial)
- **ROC:** *Receiver operating characteristic curv*e (Curva de características operativas del receptor)
- **SMOTE:** *Synthetic minority over-sampling technique* (Técnica de sobremuestreo sintético de minorías)
- **SONN:** *Self-organizing neural network* (Red neuronal autoorganizada)
- **SSL:** *Semi-supervised learning* (Aprendizaje semisupervisado)
- **SVM:** *Support vector machine* (Máquina de vectores de soporte)

- **WGAN:** *Wasserstein Generative Adversarial Network* (Red Generativa Antagónica de Wasserstein).

Capítulo IV
Metodología

En este capítulo se exponen los elementos clave de la metodología utilizada para alcanzar los objetivos propuestos. Se describen los enfoques y procedimientos empleados, abarcando desde el tipo y nivel de investigación hasta la operacionalización de variables y la selección de la muestra. También se detallan las técnicas e instrumentos de recolección de datos, y los métodos para su análisis y presentación.

4.1. Tipo y nivel de investigación

Según lo expuesto por Hernández Sampieri & Mendoza Torres (2018) esta investigación se enmarca dentro del enfoque cuantitativo, ya que busca estimar y probar hipótesis sobre el impacto de las técnicas de balance de datos en el rendimiento de los modelos de aprendizaje profundo para la clasificación de radiografías torácicas En cuanto al tipo de investigación, se trata de un estudio de carácter aplicado, pues tiene como objetivo resolver un problema práctico específico relacionado con la optimización del desempeño de los modelos de clasificación a través del uso de técnicas de balance de datos. Por otro lado, el nivel de investigación se puede clasificar como explicativo. Según Hernández Sampieri & Mendoza Torres (2018) 'este tipo de estudios pretenden determinar las causas de los eventos y fenómenos de cualquier índole' y 'establecen relaciones de causalidad entre conceptos, variables, hechos o fenómenos en un contexto concreto.

En este caso, la investigación busca determinar cómo y por qué las diferentes técnicas de balance de datos influyen en el rendimiento de los modelos de clasificación de radiografías torácicas. Finalmente, el enfoque cuantitativo de la investigación se evidencia en que la recolección y análisis de datos numéricos serán fundamentales para probar las hipótesis planteadas, medir el desempeño de los modelos y establecer patrones de comportamiento. Esto permite, tal como señalan 'estimar las magnitudes u ocurrencia de los fenómenos y probar hipótesis'.

4.2. Diseño de investigación

De acuerdo con la teoría expuesta por Hernández Sampieri & Mendoza Torres (2018) y , el diseño aplicado en esta investigación se define específicamente como un diseño cuasi-experimental por comparación de escenarios. Este tipo de diseño se caracteriza por manipular deliberadamente al menos una variable independiente (en este caso, las técnicas de balance de datos) para observar su efecto sobre una o más variables dependientes (el rendimiento de los modelos de clasificación), difiriendo de los experimentos puros en el grado de control que se tiene sobre las variables.

En un diseño cuasi-experimental, los grupos ya se encuentran formados antes del experimento, es decir, son grupos intactos (Hernández Sampieri & Mendoza Torres, 2018). En el contexto de esta investigación, los grupos no son personas, sino los conjuntos de datos de entrenamiento, que son manipulados intencionalmente para crear diferentes condiciones experimentales. No hay una asignación aleatoria tradicional, sino una comparación estructurada entre un escenario de control (sin balance) y múltiples escenarios experimentales.

Por tanto, bajo el esquema de diseño cuasi-experimental por comparación de escenarios, el grupo de control corresponde al escenario de línea base (sin balanceo), mientras que los grupos experimentales corresponden a los escenarios donde se aplica una técnica de balanceo (nivel de datos, nivel de algoritmo, o híbrido).

La evaluación se realiza en un conjunto de prueba (*test set*) que permanece constante e inalterado en todos los escenarios, permitiendo una comparación directa del efecto de la variable independiente. La estructura de los seis escenarios ejecutados se presenta en la Tabla 5.

Tabla 5.

Diseño cuasi-experimental por escenarios de comparación.

Escenario	Grupo	Manipulación (variable independiente X)	Post-Prueba (O)
Escenario 1	Control (línea base)	Dataset Original + **BCE Loss**	O1 (Test Set)
Escenario 2	Experimental (Algoritmo)	Dataset Original + **Focal Loss**	O2 (Test Set)
Escenario 3	Experimental (Algoritmo)	Dataset Original + **BWCCE Loss**	O3 (Test Set)
Escenario 4	Experimental (Algoritmo)	Dataset Original + **LDAM Loss**	O4 (Test Set)
Escenario 5	Experimental (Datos - Clásico)	Balanceo con **ROS** + BCE Loss	O5 (Test Set)
Escenario 6	Experimental (Datos - Avanzado)	Balanceo con **WGAN-GP** + BCE Loss	O6 (Test Set)
Escenario 7	Experimental (Híbrido)	Balanceo con **WGAN-GP** + **Focal Loss**	O7 (Test Set)

La partición de los datos siguió el método *holdout* estratificado, una práctica estándar en el aprendizaje automático que garantiza que la proporción de clases se mantenga en todas las particiones. Se utilizó una división de 80% para entrenamiento, 10% para validación (ajuste de hiperparámetros y *early stopping*) y 10% para prueba (evaluación final e imparcial).

Este enfoque permite establecer una relación causal entre la aplicación de las técnicas de balance (VI) y el rendimiento de los modelos (VD), proporcionando un marco robusto para comparaciones válidas, como se detalla en la *Tabla 6*.

Tabla 6.

Requisitos y características del diseño cuasiexperimental ejecutado.

	Descripción	Justificación
Diseño de investigación	Cuasiexperimental por escenarios de comparación	Se manipula la VI (técnicas de balance) en siete escenarios experimentales y se compara su efecto sobre la VD (rendimiento) contra un escenario de control (línea base), sin asignación aleatoria.
División de datos	80% entrenamiento, 10% validación, 10% prueba	Es un estándar robusto en *deep learning*. El conjunto de entrenamiento se usa para el ajuste del modelo, el de validación para el monitoreo y *early stopping*, y el de prueba (datos nunca vistos) para la evaluación final y comparativa de los siete escenarios.
Manipulación de variables	Tres niveles de técnicas de balance:	La manipulación intencional de estas técnicas (WGAN-GP, *focal loss*, BWCCE, LDAM y la combinación híbrida) permite observar cambios en el rendimiento (*accuracy*, AUC, F1, sensibilidad, especificidad) de los modelos, validando las hipótesis sobre su efectividad.
Control del estudio	Variables controladas:	Asegura que los efectos observados en la VD sean resultado de la manipulación de la VI. Se controla la arquitectura (ResNet50), los hiperparámetros (Adam 1e-4, Batch 32) y el conjunto de datos de prueba (586 imágenes).

4.3. Operacionalización de variables

En este apartado, se presenta la operacionalización de las variables que intervienen en el estudio. Este proceso es crucial ya que permite definir de manera clara y precisa cómo se miden y evalúan las variables, como se observa en la Tabla *7* (Ñaupas et al., 2018).

Tabla 7.

Definiciones conceptuales y operacionales de las variables del estudio.

Variable	Definición conceptual	Definición operacional	Dimensiones	Indicadores	Tipo de variable	Escala de medición
Rendimiento del modelo de aprendizaje profundo (VD)	Capacidad del modelo ResNet50 para clasificar correctamente radiografías torácicas en clases ("Normal" o "Pneumonia").	Evaluación del modelo entrenado en cada uno de los seis escenarios contra el conjunto de prueba (586 imágenes, IR 2.71) usando métricas estándar.	Capacidad Predictiva Global	*Accuracy* (Exactitud) *F1-score* (Macro)	Cuantitativa - Dependiente	Índice (0-1)
			Capacidad Discriminativa	AUC-ROC	Cuantitativa	Índice (0-1)
			Equilibrio de Clases	Sensibilidad (*recall* C1 - *pneumonia*) Especificidad (*recall* C0 - Normal)	Cuantitativa	Índice (0-1)
Técnicas de balance de datos (VI)	Estrategias y métodos utilizados para equilibrar la distribución de clases (IR ≈2.70) en el conjunto de entrenamiento.	Implementación de seis escenarios experimentales comparados contra un escenario de control (línea base).	Nivel de datos	WGAN-GP (Generación de 2,179 imágenes sintéticas "Normal").	Cualitativa nominal - independiente	Categórica
			Nivel de algoritmo	BWCCE *Focal loss* (γ=2.0, α=0.25) *LDAM loss*	Cualitativa nominal	Categórica
			Nivel híbrido	Aplicación de WGAN-GP y *focal loss* simultáneamente.	Cualitativa nominal	Categórica

4.3.1. Categorización de variables

Variable independiente (VI)

- **Técnicas de balance de datos**: Estrategias computacionales aplicadas para mitigar el sesgo en la clasificación de imágenes médicas debido a la distribución desigual de clases (IR ≈2.70). Se agrupan en tres niveles de manipulación:
    - **Nivel de datos:** Sobremuestreo de la clase minoritaria ("Normal") mediante la generación de imágenes sintéticas (WGAN-GP).
    - **Nivel de algoritmo:** Modificación de la función de pérdida para penalizar errores en la clase minoritaria (BWCCE, *focal loss*, LDAM Loss).
    - **Nivel híbrido:** Combinación de la técnica de datos (WGAN-GP) y la mejor técnica de algoritmo.

Variable dependiente (VD)

- **Rendimiento del modelo de aprendizaje profundo**: Capacidad del modelo (ResNet50) para clasificar correctamente radiografías torácicas ("Normal" vs. "Pneumonia"), evaluada cuantitativamente en el conjunto de prueba.

4.3.2. Condiciones experimentales controladas

Para garantizar la validez interna del estudio y asegurar que los cambios en la VD se deban a la VI, se mantuvieron constantes las siguientes condiciones en los siete escenarios:

- **Arquitectura de red:** ResNet50 pre-entrenada en ImageNet.
- **Conjunto de datos:** Particiones de Validación (527 imágenes) y Prueba (586 imágenes).

- **Protocolo de preprocesamiento:** *Pipeline* de seis pasos para clasificación (Crop, Resize 256, CLAHE, Gaussian, Resize 224, Norm [0,1]).
- **Hiperparámetros de entrenamiento:** Optimizador (Adam), Tasa de Aprendizaje ($1e-4$), Tamaño de Lote (32).
- **Ambiente computacional:** Google Colab con GPU T4.

4.4. Población y muestra

4.4.1. Determinación de la unidad de muestreo

La unidad de muestreo o análisis en este estudio se centra en las radiografías torácicas del conjunto de datos "Chest X-Ray Images (Pneumonia)" (Kermany et al., 2018). Específicamente, se enfoca en radiografías torácicas pediátricas etiquetadas en dos categorías: neumonía y normal, las cuales serán utilizadas para evaluar el rendimiento de los modelos de aprendizaje profundo bajo diferentes técnicas de balance de datos (Figura 20).

Figura 20.

Proceso de selección de población y muestra



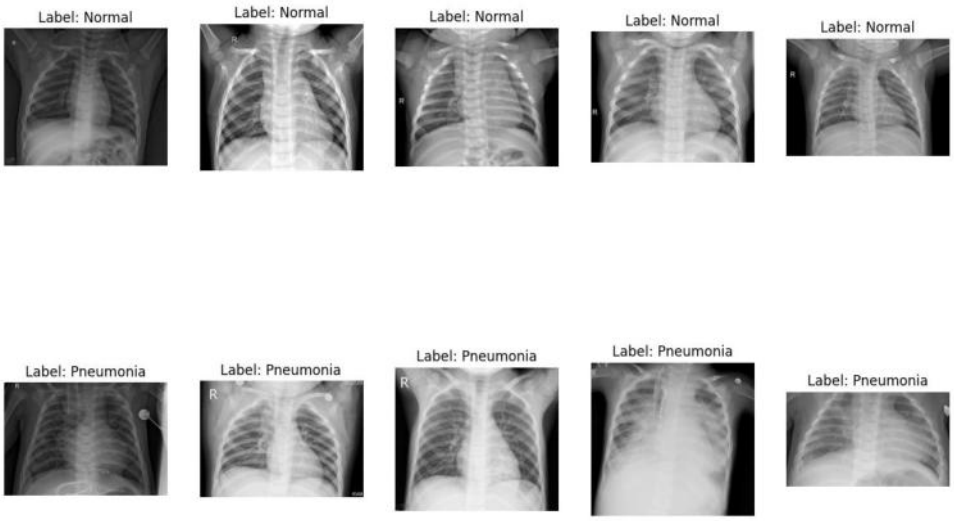
4.4.2. Delimitación de la población

La población está constituida por 5,856 radiografías torácicas pediátricas del *dataset* "Chest X-Ray Images (Pneumonia)" (Kermany et al., 2018) (Figura 21). Las imágenes comprenden:

- Total de imágenes con neumonía: 1583
- Total de imágenes normales: 4273
- Rango etario: Pacientes pediátricos de uno a cinco años

Figura 21.

Dataset: "Chest X-Ray Images (Pneumonia)"



Los detalles los podemos observar en la *Tabla 8*.

Tabla 8.

Distribución original del dataset: "Chest X-Ray Images (Pneumonia)" (Kaggle)

Conjunto	**Neumonía**	**Normal**	**Total**
Entrenamiento	1,341	3,875	5,216
Validación	8	8	16
Prueba	234	390	624
Total	**1,583**	**4,273**	**5,856**

4.4.3. Tipo de muestreo

Se utilizará un método censal, que implica el uso de la totalidad de las unidades disponibles en el conjunto de datos. Como señalan (Hernández Sampieri et al., 2014), un censo se aplica cuando se busca incluir todos los casos del universo o población, lo cual es particularmente relevante en estudios con conjuntos de datos limitados y controlados.

Los criterios para esta selección son:

- **Accesibilidad total:** Disponibilidad completa del conjunto de datos.
- **Viabilidad técnica:** Volumen de datos procesable con recursos computacionales disponibles.
- **Intencionalidad investigativa:** Evaluar técnicas de balance de datos en un conjunto con desequilibrio natural.

4.4.4. Características de la muestra

- **Origen:** Centro Médico de Mujeres y Niños de Guangzhou (Kermany et al., 2018).
- **Tipo de imágenes:** Radiografías torácicas de en proyección anteroposterior.
- **Rango etario:** Pacientes pediátricos de uno a cinco años.
- **Control de calidad:**
    - Imágenes inicialmente filtradas para eliminar escaneos de baja calidad.
    - Diagnósticos verificados por dos médicos expertos.
    - Conjunto de evaluación revisado por un tercer experto.
- **Categorías:**
    - Normal

- Neumonía

- **Desbalance de clases:**
  - Neumonía: 1,583 imágenes (27%)
  - Normal: 4,273 imágenes (73%)

4.5. Consideraciones éticas

Las imágenes fueron recopiladas como parte de la atención clínica rutinaria, siguiendo protocolos éticos de investigación médica. La disponibilidad del conjunto de datos está respaldada por la licencia Creative Commons BY 4.0 (Kermany et al., 2018).

4.6. Técnicas e instrumentos para la recolección de datos

La recolección de datos es un proceso fundamental en cualquier investigación, ya que proporciona la información necesaria para responder las preguntas de investigación y validar las hipótesis planteadas (Hernández Sampieri & Mendoza Torres, 2018). En el contexto de esta investigación, se han seleccionado métodos específicos que se alinean con los objetivos del estudio, como se detalla en la Tabla 9, se emplearon principalmente dos enfoques: el análisis de contenido para la fundamentación teórica y la recolección de datos secundarios para la parte experimental.

Dado que la investigación no implicó experimentación directa con pacientes humanos, se utilizó el *dataset* público 'Chest X-Ray Images (Pneumonia)' de Kaggle (Kermany et al., 2018) como fuente primaria. Este repositorio contiene imágenes validadas y organizadas en categorías (normal y neumonía). Para la sistematización de estos datos, se utilizó como instrumento una ficha de registro de datos y metadatos (implementada mediante *scripts* de verificación en Python). Este instrumento permitió validar la integridad de las 5,856 imágenes, su resolución, formato y distribución entre los conjuntos de entrenamiento, prueba y validación.

Tabla 9.

Métodos de recolección de datos: propósito, ventajas y aplicación

Método	Propósito General	Ventajas	Aplicación en esta investigación
Análisis de contenido	Recolectar información teórica respecto de técnicas de balance de datos y modelos de aprendizaje profundo.	• Permite evaluar el estado del arte. • Ideal para la revisión de estudios previos sobre las GAN y funciones de pérdida.	Se realizó una revisión sistemática de artículos científicos y reportes técnicos para establecer los hiperparámetros de las redes neuronales y justificar la elección de WGAN-GP frente a técnicas tradicionales.
Datos secundarios	Obtener datos estandarizados sobre el rendimiento de modelos de clasificación en imágenes radiográficas.	• Acceso directo a un conjunto de datos estandarizado (5,856 imágenes). • Validación previa del *dataset* por la comunidad científica. • Disponibilidad de métricas base para comparación.	**Fuente:** Se utilizó el *dataset Chest X-Ray Images (Pneumonia)* de Kaggle. **Instrumento:** Se aplicaron *scripts* de extracción para organizar los datos en conjuntos de entrenamiento (5,216), validación (16) y prueba (624), permitiendo la evaluación de los siete escenarios experimentales propuestos.

4.7. Procedimiento metodológico y análisis de datos

Este flujo de trabajo se estructura conceptualmente en cuatro fases principales, como se observa en el esquema de la Figura 22, y su secuencia operativa detallada se presenta en el diagrama de flujo de la Figura 23:

1. Preparación y preprocesamiento de datos.
2. Implementación de técnicas de balance de datos.
3. Entrenamiento y evaluación de modelos de aprendizaje profundo.
4. Análisis y comparación de resultados.

Figura 22.

Esquema conceptual de las fases principales del procedimiento metodológico

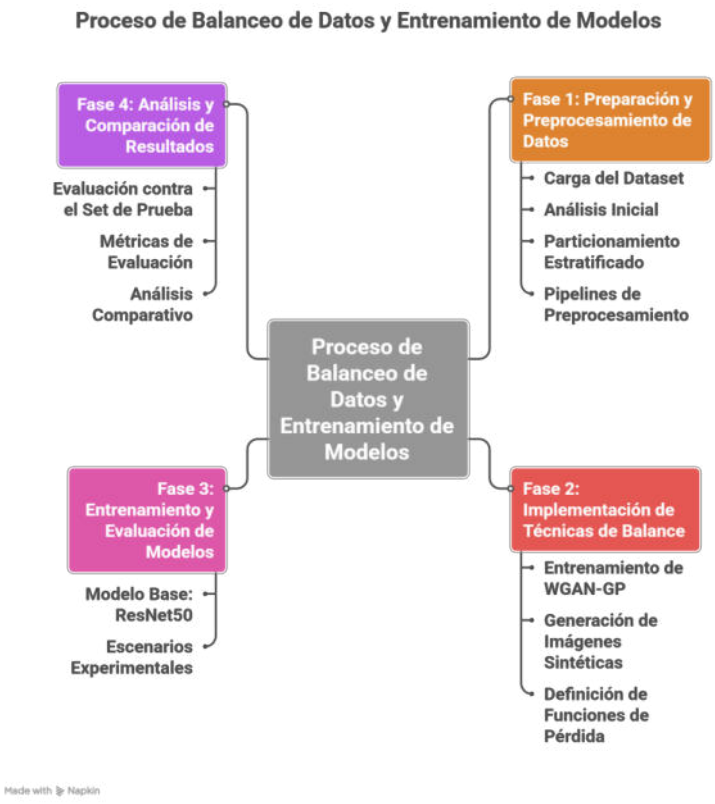
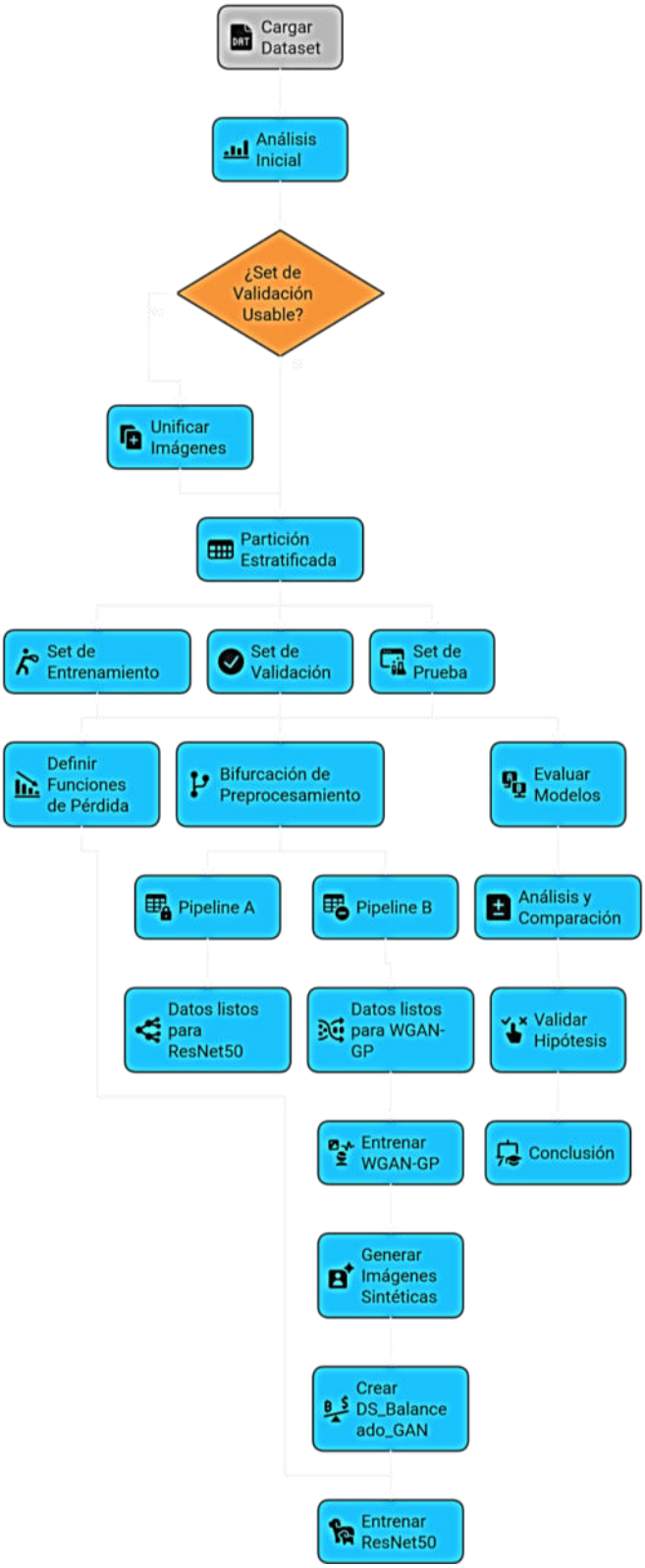


Figura 23.

Diagrama de flujo detallado de la ejecución experimental y pipelines de procesamiento.



4.7.1. Metodología aplicada

Fase 1: Preparación y preprocesamiento de datos

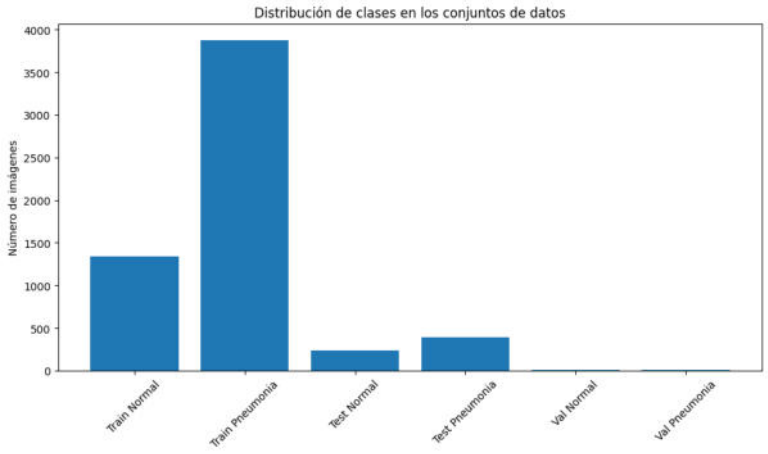
En esta sección se describe la preparación inicial y el preprocesamiento aplicado al conjunto de datos utilizados en esta investigación, siguiendo metodologías similares a las propuestas por (Gaspar, 2023; Liang et al., 2020; X. Qin et al., 2022)

Descripción y reestructuración del conjunto de datos. -

El conjunto de datos a emplear es el *dataset* en línea *Chest X-Ray Images (Pneumonia)* (Kermany et al., 2018). Dicho conjunto está compuesto por imágenes correspondientes a dos categorías de diagnóstico, como neumonía y casos normales, presentando un evidente desbalance entre clases. El conjunto original (Tabla 8) presenta una partición de validación de 16 imágenes, la cual es estadísticamente insuficiente para una evaluación robusta y detección de sobreajuste. La distribución de los datos original por conjunto se detalla en la Tabla 8 y se observa en la Figura 24.

Figura 24.

Distribución de clases en el conjunto de datos.



Debido a esta limitación, se tomará la decisión metodológica de unificar las 5,856 imágenes (de *train*, *test y val*) en un solo conjunto. Posteriormente, se aplicará una nueva partición estratificada (método *holdout*) para asegurar que la proporción de clases se mantenga constante.

Tabla 10.

Distribución de la población total del conjunto de datos.

Conjunto	**Cantidad**	**Porcentaje**
Entrenamiento	5216	89.071%
Validación	16	0.003%
Prueba	624	10.656%
Total	**5,856**	**100%**

Siguiendo la metodología de Liang et al. (2020), las imágenes serán organizadas en tres subconjuntos: entrenamiento, validación y prueba, como se observa en la Tabla 10.

La división planificada será:

- **Un 80% para entrenamiento** (Aprox. 4,685 imágenes)
- **Un 10% para validación** (Aprox. 586 imágenes)
- **Un 10% para prueba** (Aprox. 585 imágenes)

El nivel de desbalance del conjunto de datos fue cuantificado utilizando la métrica de razón de desbalance o también llamada **índice de desbalance** (IR*)*, definida por X. Qin et al. (2022)como:

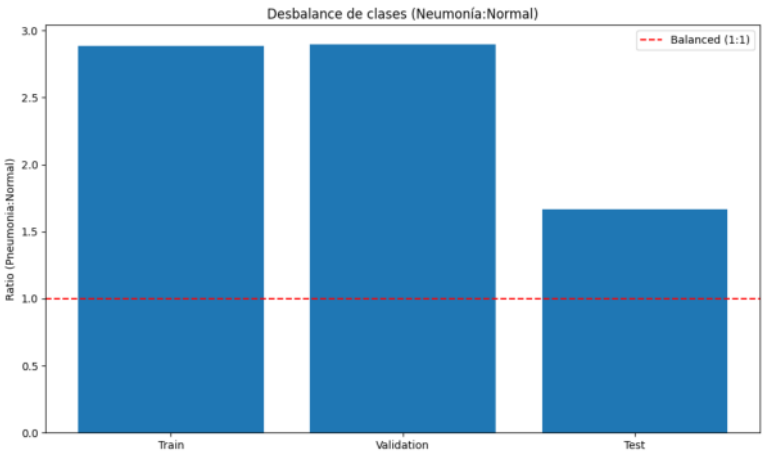
$$IR = \frac{N_{\text{mayoría}} + 1}{N_{\text{minoría}} + 1}$$

donde $N_{\text{mayoría}}$ y $N_{\text{minoría}}$ representan el número de instancias en las clases mayoritaria y minoritaria, respectivamente. En este caso, se obtuvo un $IR = 2.7$, indicando un desbalance significativo.

El análisis exploratorio de los datos revela que la clase Normal (0) es la minoritaria (aprox. 27%) y neumonía (1) es la mayoritaria (aprox. 73%), resultando en un IR de aproximadamente **2.7**, en la Figura 1Figura 25 observamos el desbalance de las clases.

Figura 25.

Desbalance de clases en Chest X-Ray Images (Pneumonia)



Pipeline de preprocesamiento. -

Se implementarán dos *pipelines* de preprocesamiento distintos, optimizados para las dos tareas principales: clasificación y generación.

1. **Pipeline para clasificación (ResNet50):** Este *pipeline* se aplicará a todos los datos (entrenamiento, validación y prueba) destinados al clasificador ResNet50. Los pasos serán:

    a. **Recorte central (Center Crop):** Se aplicará un recorte para enfocar el análisis en la región pulmonar.

    b. **Redimensionamiento inicial:** La imagen se redimensionará a 256×256 píxeles.

    c. **Mejora de contraste (CLAHE):** Se implementará CLAHE (fundamentada en la Sección 3.2.1.1.2) para realzar texturas sutiles.

    d. **Filtro gaussiano:** Se aplicará un filtro (*kernel* 3×3) para suavizar el ruido.

e. **Redimensionamiento final:** La imagen se redimensionará a 224×224 píxeles, tamaño de entrada de ResNet50.

f. **Normalización [0, 1]:** Los valores de píxeles (RGB) se normalizarán al rango [0, 1] (tipo *float32*).

2. **Pipeline para generación (WGAN-GP):** Este *pipeline* se aplicará únicamente al conjunto de entrenamiento de la clase minoritaria (Normal) que alimentará a la GAN.

   a. **Conversión a escala de grises:** Las imágenes se convertirán a 1 canal (monocromáticas).

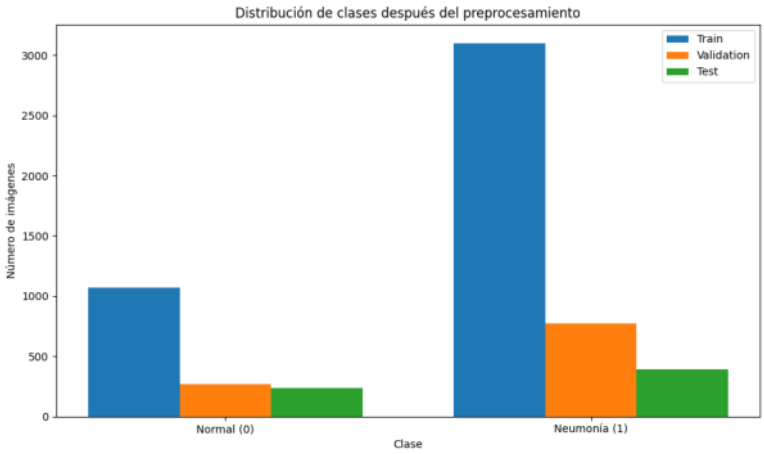
   b. **Mejora de contraste (CLAHE):** Se aplicará CLAHE sobre la imagen en escala de grises.

   c. **Normalización [-1, 1]:** Los valores se normalizarán al rango [-1, 1], requerido por la función de activación tanh del generador de la WGAN.

   d. **Redimensionamiento progresivo:** Las imágenes se redimensionarán a 64x64, 128x128 y 224x224 para las distintas etapas del entrenamiento progresivo.

Figura 26.

Distribución de clases después del preprocesamiento.



El resultado de esta fase es un conjunto de datos reestructurado y preprocesado, como se muestra en la Figura 26. La *Tabla 11* detalla la distribución final y el IR resultante, confirmando que la clase Pneumonia (1) es la mayoritaria y Normal (0) es la minoritaria.

Tabla 11.

Distribución final de clases y ratio de desbalance (IR) tras la división 80/10/10.

Partición	Normal (0)	Pneumonia (1)	Total	IR
Entrenamiento	1,282 (27.0%)	3,461 (73.0%)	4,743	2.70
Validación	143 (27.1%)	384 (72.9%)	527	2.69
Prueba	158 (27.0%)	428 (73.0%)	586	2.71

Fase 2: Implementación de técnicas de balance de datos

Para abordar el desbalance (IR ≈ 2.7), se implementarán y evaluarán técnicas a nivel de datos (WGAN-GP) y a nivel de algoritmo (Funciones de Pérdida).

Implementación de WGAN-GP con crecimiento progresivo (Objetivo 1). -

Para cumplir el primer objetivo específico (H1.1), se implementará una WGAN-GP. Esta arquitectura avanzada se seleccionará sobre una DCGAN estándar por su estabilidad de entrenamiento superior, evitando el colapso de modo (Sindhura et al., 2024).

Para garantizar la validez experimental y evitar estrictamente la fuga de datos (*data leakage*), el proceso de generación de imágenes sintéticas se ejecutó posterior a la división estratificada del conjunto de datos (80/10/10). Específicamente, el entrenamiento de la arquitectura WGAN-GP se realizó utilizando únicamente las 1,282 imágenes de la clase minoritaria ('Normal') pertenecientes al subconjunto de entrenamiento X_train. Los conjuntos de validación X_val y prueba X_test se mantuvieron completamente aislados, compuestos exclusivamente por datos reales originales, y no intervinieron en ninguna etapa del entrenamiento de la red generativa.

Las 2,179 imágenes sintéticas generadas se incorporaron exclusivamente al conjunto de entrenamiento, resultando en un nuevo X_train balanceado y aumentado, mientras que los conjuntos de evaluación permanecieron intactos para asegurar una medición objetiva del rendimiento del modelo ante datos reales no vistos.

El entrenamiento seguirá un enfoque de crecimiento progresivo en tres etapas, comenzando con una baja resolución (64x64) y aumentando progresivamente a 128x128 y 224x224. Los hiperparámetros clave del plan se detallan en la

Tabla 12.

Tabla 12.

Hiperparámetros de entrenamiento para la WGAN-GP.

Componente	**Hiperparámetro**	**Valor ejecutado**	**Justificación técnica**
Arquitectura	Modelo Base	WGAN-GP (PyTorch)	Basado en ResNet con *spectral normalization*.

Componente	Hiperparámetro	Valor ejecutado	Justificación técnica
Entrenamiento	Semilla (*seed*)	42	Garantiza reproducibilidad.
	Dimensión Latente (z)	128	Espacio suficiente para variabilidad anatómica.
	Optimizador	Adam	Estándar para GAN estables.
	Betas de Adam	(0.0, 0.9)	Recomendación teórica de *Gulrajani et al.*
	Tasa de aprendizaje G	3.5e-5 (0.000035)	Ajustado con factor de escala 0.5 para estabilidad en alta resolución.
	Tasa de aprendizaje C	7.5e-5 (0.000075)	Mayor que G para mantener al crítico óptimo.
	batch size	32	Límite de memoria VRAM en Colab Pro.
Pérdida	Ratio crítico (*critic*)	5	Se entrena al crítico 5 veces por cada paso del generador.
	Peso penalización (lambda)	10.0	*Gradient penalty* para cumplir la condición de Lipschitz.
Ejecución	Estrategia	Progresiva [64, 128, 224]	Estabiliza el entrenamiento desde baja resolución.
	Épocas totales	4,000 (Max)	Se detuvo por *early stopping* (Mejor FID en época 3,470).
	Métrica de calidad	FID	*Fréchet Inception Distance* (menor es mejor).

Nota. Los valores de tasa de aprendizaje reflejan el escalamiento aplicado en la etapa de resolución 224x224.

La

Tabla 12 presenta la configuración final utilizada para el entrenamiento del modelo generativo. A diferencia de las GAN tradicionales, se implementó una arquitectura WGAN-GP debido a su capacidad para mitigar el colapso de modo y proporcionar gradientes más estables.

Se aplicó una estrategia de entrenamiento progresivo, comenzando con imágenes de 64x64 píxeles y aumentando la resolución hasta alcanzar el objetivo clínico de 224x224 píxeles. Para la etapa final de alta resolución, se ajustaron dinámicamente las tasas de aprendizaje (*learning rates*) aplicando un factor de escalado de 0.5 sobre los valores base. Esto resultó en una tasa de 3.5e-5 para el

Generador y 7.5e-5 para el Crítico, manteniendo una proporción aproximada de 1:2 (TTUR - *Two-Time-Scale Update Rule*) para asegurar que el crítico converja más rápido que el generador, requisito fundamental para la estimación correcta de la distancia Wasserstein.

El entrenamiento se configuró para un máximo de 4,000 épocas en la etapa final, implementando un mecanismo de *early stopping* con una paciencia de 15 evaluaciones sobre la métrica FID. El modelo óptimo, utilizado posteriormente para la generación de datos sintéticos, se obtuvo en la época 3,470, punto donde se minimizó la distancia entre las distribuciones real y sintética.

Implementación de funciones de pérdida sensibles a costos (Objetivo 2). -

Para cumplir el segundo objetivo específico (H1.2), se implementarán tres funciones de pérdida a nivel de algoritmo, las cuales modifican el entrenamiento del clasificador ResNet50 para penalizar más los errores en la clase minoritaria (Normal). Se utilizará el conjunto de entrenamiento desbalanceado original (4,743 imágenes). Las funciones a implementar, basadas en (Hoang et al., 2023) y la literatura, serán:

- ***Focal loss*:** Se implementará la pérdida Focal con parámetros $\gamma = 2.0$ y $\alpha = 0.25$ (o $\alpha = 0.73$ para la clase minoritaria) para reducir el peso de las muestras fáciles (mayoritarias) y enfocar el entrenamiento en las difíciles (minoritarias).

- ***Balanced weighted cross-entropy* (BWCCE):** Se calcularán pesos inversamente proporcionales a la frecuencia de las clases (ej. 0: 1.85, 1: 0.69) y se aplicarán directamente a la función de pérdida *Binary Cross-Entropy*.

- **LDAM loss:** Se implementará la pérdida *label-Distribution-Aware Margin*, que aplica un margen adaptativo basado en la frecuencia de la clase, $m_j = C/(n_j^{1/4})$.

Fase 3: Entrenamiento y evaluación de modelos (H1.3)

En esta fase se ejecutarán los experimentos para probar todas las hipótesis planteadas. Se definen siete escenarios experimentales diseñados para aislar el

efecto de cada técnica. En todos los escenarios se utilizará la misma arquitectura base (ResNet50 preentrenado) y los mismos hiperparámetros de entrenamiento (Adam 1e-4, Batch Size 32, Callbacks de earlystopping y reducelronplateau) para asegurar una comparación justa.

- Escenario 1 (línea base): ResNet50 entrenado en el *dataset* original desbalanceado (IR 2.7) utilizando la función de pérdida *binary cross-entropy* estándar.

- Escenario 2 (*focal loss*): ResNet50 entrenado en el *dataset* desbalanceado, sustituyendo la función de pérdida por *focal loss*.

- Escenario 3 (BWCCE): ResNet50 entrenado en el *dataset* desbalanceado, utilizando *binary weighted cross-entrop*y (pesos inversos a la frecuencia de clase).

- Escenario 4 (LDAM): ResNet50 entrenado en el *dataset* desbalanceado, utilizando *LDAM loss* (Label-Distribution-Aware Margin).

- Escenario 5 (Sobremuestreo Simple - ROS): ResNet50 entrenado en un *dataset* balanceado artificialmente mediante la duplicación aleatoria de imágenes de la clase minoritaria con ROS. Este escenario sirve de control para validar si la mejora se debe a la cantidad o a la calidad de los datos (H1.1).

- Escenario 6 (WGAN-GP): ResNet50 entrenado en el *dataset* balanceado 1:1, integrando las imágenes sintéticas generadas por la arquitectura WGAN-GP propuesta.

- Escenario 7 (Híbrido): ResNet50 entrenado en el *dataset* balanceado con WGAN-GP, combinándolo con la mejor función de pérdida sensible a costos (*focal loss*) para evaluar sinergias (H1.3).

Fase 4: Análisis y comparación de resultados

La evaluación de los modelos se realizará sobre el conjunto de prueba aislado (10% de los datos, 586 imágenes) que los modelos nunca habrán visto

durante el entrenamiento. Las métricas de evaluación clave, seleccionadas por su robustez ante el desbalance, serán:

- Para la calidad de generación (WGAN-GP): *Fréchet Inception Distance* (FID).
- Para el rendimiento de clasificación (ResNet50):
    - Exactitud (*accuracy*): Referencia global.
    - Sensibilidad (r*ecall*): Crítica para detectar neumonía (Clase 1).
    - Especificidad: Crítica para validar la reducción de Falsos Positivos en sanos (Clase 0).
    - Puntuación F1 (*F1-score*): Media armónica.
    - Área Bajo la Curva ROC (AUC-ROC): Capacidad de discriminación independiente del umbral.
    - G-Mean: Media geométrica de sensibilidad y especificidad.
    - Coeficiente de Matthews (MCC): Métrica de correlación robusta para clases desbalanceadas.
    - Matriz de Confusión: Análisis visual de errores tipo I y II.

Cada escenario experimental, incluyendo el preprocesamiento avanzado con CLAHE y la arquitectura final de ResNet50, se implementó mediante el código fuente disponible en el Apéndice B, garantizando la replicabilidad de los experimentos.

Los resultados de estos siete escenarios se tabularán y compararán para determinar qué técnica (a nivel de datos, a nivel de algoritmo o híbrida) ofrece la mejora más significativa sobre la línea base.

Capítulo V
Resultados y discusión

Este capítulo presenta los resultados obtenidos de la implementación experimental del diseño metodológico propuesto. El análisis se centra en evaluar la efectividad de las diferentes estrategias de balance de datos aplicadas al conjunto de imágenes de radiografías torácicas para la detección de neumonía.

A lo largo de este capítulo, se detalla la ejecución de cada fase, desde la preparación de los datos y la generación de imágenes sintéticas, hasta el entrenamiento y la evaluación comparativa de los siete escenarios experimentales definidos. Finalmente, se discuten los hallazgos en función de las métricas de rendimiento obtenidas, contrastándolos con la línea base y los objetivos planteados en la investigación.

5.1. Resultados

5.1.1. Resultados experimentales

Fase 1: Ejecución de la preparación y preprocesamiento de Datos

El primer paso consistió en la adquisición y preparación del conjunto de datos, lo cual fue fundamental para establecer la línea base del experimento, siguiendo el plan de la Fase 1 de la metodología.

Particionamiento y estratificación del dataset.

El conjunto de datos original *"Chest X-Ray Images (Pneumonia)"* Haga clic o pulse aquí para escribir texto. presentaba una partición de validación de solo 16 imágenes (Tabla 8), la cual se consideró estadísticamente insuficiente para una evaluación robusta. Ante la limitación estadística del conjunto de validación original (que constaba de solo 16 imágenes, un 0.2% del total), se procedió a una reestructuración metodológica de los datos para garantizar la robustez del entrenamiento.

En lugar de descartar particiones, se ejecutó una unificación total de los tres subconjuntos originales (entrenamiento, validación y prueba), consolidando las

5,856 imágenes en un único tensor. Posteriormente, se aplicó un nuevo muestreo aleatorio estratificado (*stratified holdout*) utilizando la semilla random_state=42 para asegurar la reproducibilidad.

Tabla 13.

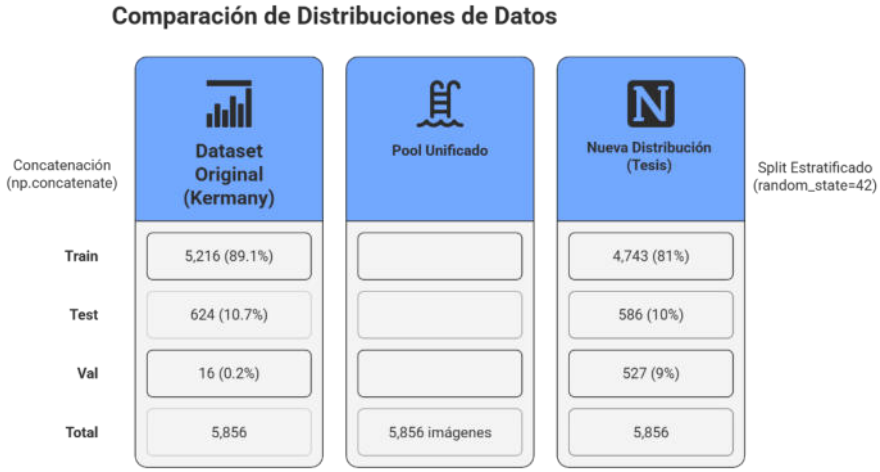
Distribución final de clases y ratio de desbalance (IR) después del preprocesamiento.

Partición	Normal (0)	Neumonía (1)	Total	IR
Entrenamiento	1,282 (27.0%)	3,461 (73.0%)	4,743	2.70
Validación	143 (27.1%)	384 (72.9%)	527	2.69
Prueba	158 (27.0%)	428 (73.0%)	586	2.71

En la Tabla 13, se observa la nueva distribución (80/10/10) que permitió generar un conjunto de validación representativo de 527 imágenes y un conjunto de prueba de 586 imágenes, manteniendo la integridad total del *dataset* original sin pérdida de información (5,856 imágenes totales procesadas), tal como se detalla en el flujo de la Figura 27.

Figura 27.

Diagrama de flujo de la redistribución de datos.



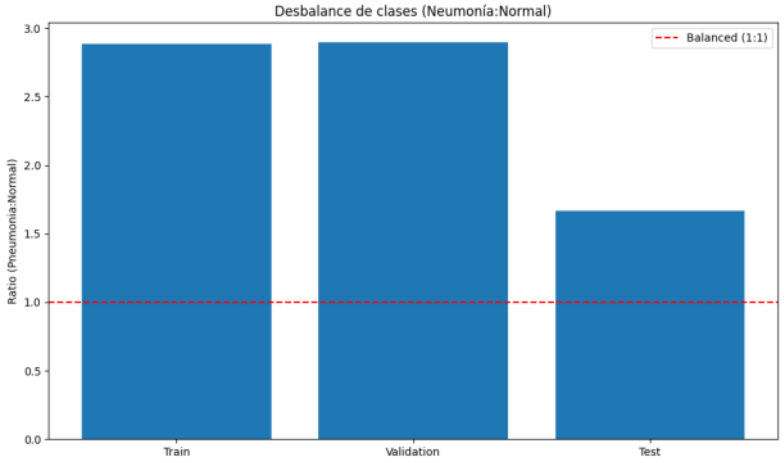
Para garantizar la reproducibilidad exacta de los experimentos y la integridad de la partición estratificada, se estableció una semilla aleatoria fija (random_state=42) en todas las operaciones de división de datos y en la inicialización de los generadores de números pseudoaleatorios de las librerías utilizadas (NumPy, PyTorch y TensorFlow).

Si bien la validación cruzada (*k-fold cross-validation*) es una técnica robusta para estimar el error de generalización, en esta investigación se optó por el método de retención (*holdout*) estratificado (80/10/10). Esta decisión metodológica se fundamenta en el alto costo computacional asociado al entrenamiento iterativo de las WGAN-GP, cuyo proceso de convergencia requiere recursos significativos que hacen inviable la ejecución de múltiples pliegues dentro del alcance temporal y técnico del presente estudio.

El análisis confirmó que la clase Normal (0) era la minoritaria (27.0%) y Pneumonia (1) la mayoritaria (73.0%), con un IR estable de ≈2.70 en todas las particiones, como se visualiza en la Figura 28.

Figura 28.

Ratio de desbalance (pneumonia:normal) en las particiones finales.



Ejecución de los Pipelines de Preprocesamiento.

Se ejecutaron los dos *pipelines* de preprocesamiento definidos en la metodología:

1. **Pipeline de clasificación (ResNet50):** Se procesaron los tres conjuntos (*train*, *val*, *test*) aplicando el *pipeline* de seis pasos (Crop, Resize 256, CLAHE, Gaussian, Resize 224, Norm [0,1]).

2. **Pipeline de generación (WGAN-GP):** Se tomaron únicamente las 1,282 imágenes de la clase minoritaria "Normal" del conjunto de entrenamiento y se procesaron con el *pipeline* específico (Grayscale, CLAHE, Norm [-1,1]).

Fase 2: Resultados de la implementación de técnicas de balance (H1.1)

Esta fase se centró en la ejecución de las técnicas de balanceo a nivel de datos (WGAN-GP) para abordar la Hipótesis Específica 1 (H1.1).

Resultados de la generación de datos (WGAN-GP).

Se ejecutó el entrenamiento de la WGAN-GP con crecimiento progresivo en tres etapas (64x64, 128x128, 224x224) sobre las 1,282 imágenes "Normales". El objetivo fue generar 2,179 imágenes sintéticas para balancear las 3,461 imágenes de *pneumonia*.

La arquitectura WGAN-GP optimiza la distancia de Wasserstein, donde las funciones de pérdida del Crítico (L_C) y Generador (L_G) se definen como:

$$L_C = \mathbb{E}_{\tilde{x}\sim P_g}[C(\tilde{x})] - \mathbb{E}_{x\sim P_r}[C(x)] + \lambda \mathbb{E}_{\hat{x}\sim P_{\hat{x}}}[(\| \nabla_{\hat{x}} C(\hat{x}) \|_2 - 1)^2]$$

$$L_G = -\mathbb{E}_{\tilde{x}\sim P_g}[C(\tilde{x})]$$

Para garantizar la transparencia y reproducibilidad de los resultados, se detallan a continuación los hiperparámetros específicos utilizados en el entrenamiento de la arquitectura WGAN-GP, que también se mostraron en la

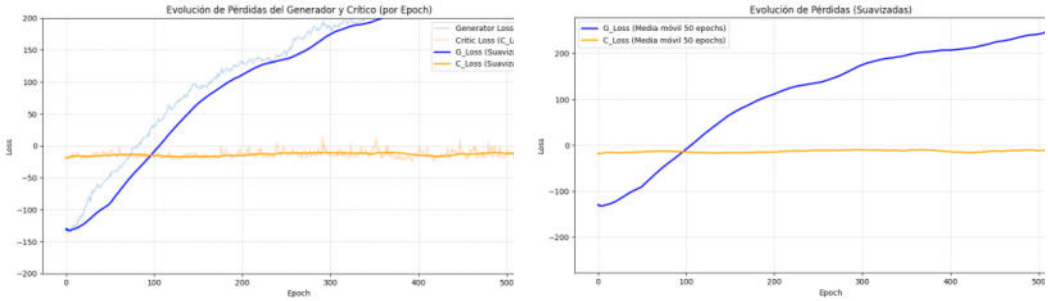
Tabla 12. El modelo se entrenó siguiendo un esquema de crecimiento progresivo para estabilizar la generación de alta resolución: Configuración del Entrenamiento:

- Semilla Aleatoria: Se fijó random_state=42 para la partición de datos y la inicialización de pesos.
- Optimizador: Adam con beta_1=0.0 y beta_2=0.9.
- Tasa de aprendizaje (*learning rate*): 10^-4 tanto para el generador como para el crítico, con un decaimiento del 50% en la etapa de mayor resolución.
- Ciclos de entrenamiento: Se realizaron cinco iteraciones del crítico (n_critic=5) por cada actualización del generador.
- Penalización de gradiente: Se estableció un coeficiente lambda_gp = 10 para mantener la restricción de Lipschitz.
- Esquema progresivo (Épocas):
- Etapa 1 (64x64): 500 épocas.
- Etapa 2 (128x128): 500 épocas.
- Etapa 3 (224x224): Configurado hasta 4,000 épocas, logrando la convergencia óptima (FID 48.95) alrededor de la época 3,910, momento en el cual se detuvo el entrenamiento mediante *early stopping*.

Las curvas de pérdida (Figura 29) muestran el comportamiento esperado: la pérdida del generador (azul) desciende, indicando su mejora, mientras la pérdida del crítico (naranja) fluctúa, demostrando que mantiene su capacidad de discernimiento. La Figura 29 (media móvil) confirma la estabilidad general del entrenamiento sin colapso de modo.

Figura 29.

Media móvil (estabilidad) de las pérdidas WGAN-GP.



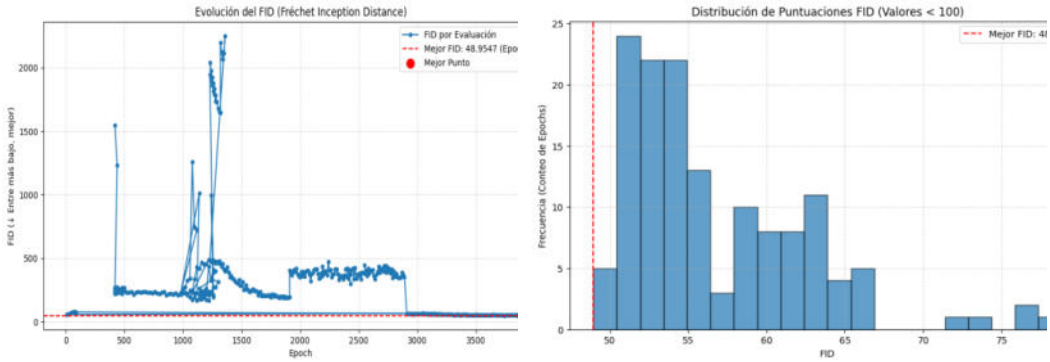
La calidad de las imágenes generadas se midió en cada etapa mediante la métrica FID, donde un valor más bajo indica mayor realismo y similitud con el *dataset* real. La fórmula del FID es:

$$\text{FID} = \| \mu_r - \mu_g \|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}) \quad (1)$$

La Figura 30 muestra la evolución del FID, estabilizándose y alcanzando un FID final de 48.95 (Figura 30), un valor que indica alta fidelidad y realismo en las imágenes generadas.

Figura 30.

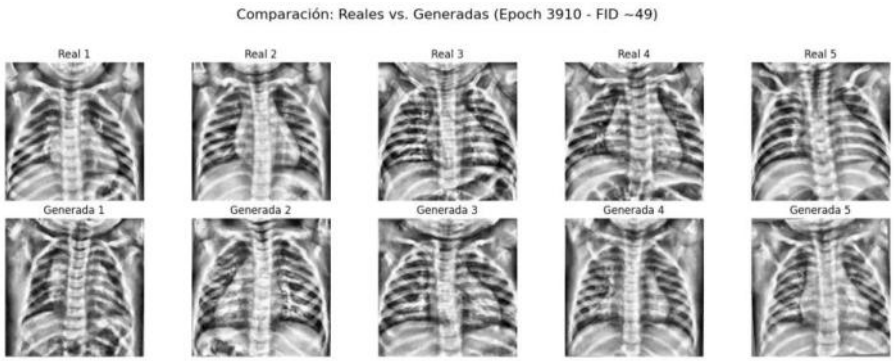
Puntuación FID final (48.95) en la etapa de 224x224.



Al finalizar el entrenamiento, se generaron las 2,179 imágenes sintéticas "Normal" (Figura 31). Estas se combinaron con las 4,743 imágenes originales para crear el *dataset* balanceado, compuesto por 6,922 imágenes (3,461 Normales y 3,461 *pneumonia*), logrando un IR = 1.0.

Figura 31.

Comparación visual entre imágenes Reales (fila sup.) y Sintéticas (fila inf.).



La ejecución exitosa de esta fase, logrando un FID bajo (48.95) y generando imágenes visualmente coherentes, valida la **Hipótesis específica 1 (H1.1)**, que postulaba que era factible generar imágenes sintéticas de alta calidad.

Fase 3: Resultados del entrenamiento y evaluación de modelos

Habiendo preparado los conjuntos de datos (desbalanceado y balanceado) y definido las funciones de pérdida, se procedió a la Fase 3: el entrenamiento y evaluación de los seis escenarios experimentales. Todos los modelos se entrenaron usando la arquitectura ResNet50 con los mismos hiperparámetros base (*Adam* $LR = 1e-4$, *batch size* = 32) y se evaluaron contra el mismo conjunto de prueba (586 imágenes, IR 2.71) para una comparación justa.

Escenario 1: Modelo línea base (ResNet50 + BCE).

El Escenario 1 corresponde al modelo línea base, entrenado sin aplicar técnicas de balance de datos. Este primer experimento estableció la referencia de rendimiento al entrenar el modelo ResNet50 sobre el conjunto de datos desbalanceado original (4,743 imágenes, IR 2.70), utilizando la función de pérdida estándar *binary cross-entropy o* también conocido como **entropía cruzada binaria** (BCE).

Tabla 14.

Resultados del Escenario 1: modelo línea base.

Métrica	Valor (Test Set)
Accuracy	0.9147
AUC-ROC	0.9635
F1-Score (Macro Avg)	0.8921
Sensibilidad (Recall C1 - Pneumonia)	**0.9393**
Especificidad (Recall C0 - Normal)	**0.8481**

Si bien se alcanzó una Exactitud (*accuracy*) del 91.47%, esta métrica debe interpretarse con cautela debido al desbalance de clases. Un análisis más profundo (Figura 33) revela que el alto rendimiento global está sesgado hacia la clase mayoritaria (*pneumonia*). Aunque el modelo presenta una excelente capacidad de discriminación general, evidenciada por un AUC-ROC de 0.9635 y una alta Sensibilidad de 0.9393 (*Tabla 14*), su debilidad crítica reside en la identificación de la clase minoritaria ("Normal").

La especificidad resultante de solo 0.8481 (notablemente inferior a la sensibilidad) confirma la hipótesis de que el modelo tiende a sobreajustarse a la clase mayoritaria, clasificando erróneamente a pacientes sanos como enfermos los cuáles serían falsos positivos. Este comportamiento valida la problemática central de la investigación: la necesidad imperativa de implementar técnicas de balanceo que eleven la capacidad del modelo para reconocer la clase minoritaria sin degradar su rendimiento global.

Figura 32.

Pérdida y accuracy, de entrenamiento y validación.

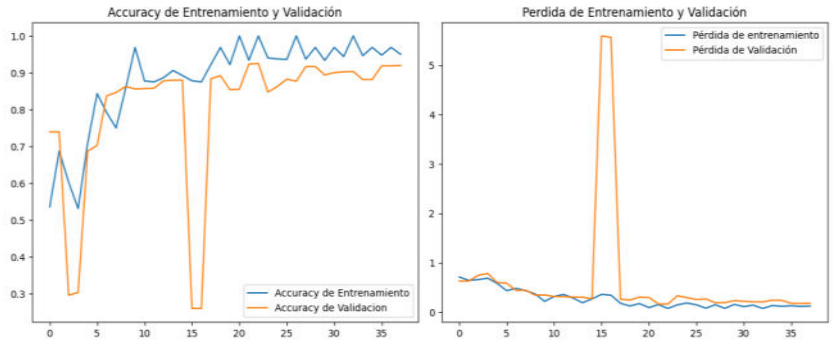
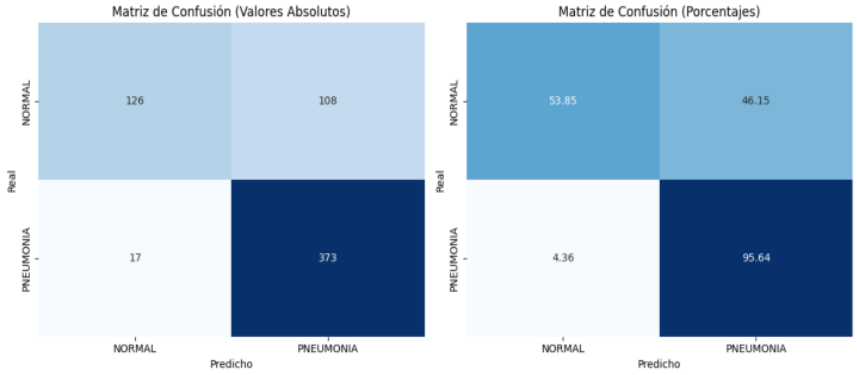


Figura 33.

Matriz de confusión del modelo línea base (Test Set).



El alto número de Falsos Positivos (17 imágenes "Normal" clasificadas como "Pneumonia", según la Figura 33) confirma el sesgo del modelo y la necesidad de aplicar técnicas de balanceo.

Desde una perspectiva estrictamente médica, la interpretación de estos resultados requiere ponderar el costo asimétrico de los errores. En sistemas de triaje radiológico para neumonía, existe un consenso en que el costo de un falso negativo (omitir el diagnóstico de un paciente realmente enfermo, lo que podría derivar en complicaciones graves o letales) es sustancialmente mayor que el de un falso positivo (un paciente sano alertado erróneamente).

Bajo esta premisa clínica, se prioriza la métrica de sensibilidad (*recall*) para minimizar el riesgo de no detección. En este sentido, la sensibilidad de 0.9393 de la línea base podría considerarse aceptable para una herramienta de tamizaje

preliminar, aceptando una tasa mayor de falsos positivos bajo el supuesto de que estos casos serán posteriormente descartados por la revisión confirmatoria de un radiólogo experto.

Sin embargo, la baja Especificidad (0.8481) del modelo base genera una carga operativa excesiva (demasiados sanos enviados a revisión), lo que justifica la necesidad de las técnicas de balanceo propuestas en esta tesis: el objetivo es mantener esa alta sensibilidad vital, pero reduciendo drásticamente los falsos positivos para optimizar el flujo de trabajo hospitalario.

Resultados de las funciones de pérdida (H1.2).

Los escenarios 2, 3 y 4 evaluaron las técnicas a nivel de algoritmo a saber son: BWCCE, *focal loss*, y LDAM *loss* sobre el mismo *dataset* desbalanceado. El objetivo era validar la H1.2: si estas funciones podían mejorar el rendimiento, especialmente en la clase minoritaria (especificidad).

1. **Balanced Weighted Cross-Entropy (BWCCE):** Se aplicaron pesos inversos a la frecuencia de clases ($w_0 = 1.85, w_1 = 0.69$). La pérdida se definió como:

$$L_{\text{BWCCE}} = -\frac{1}{N}\sum_{i=1}^{N}[\,(w_0 \cdot (1 - y_i) \cdot \log(1 - p_i)) + (w_1 \cdot y_i \cdot \log(p_i))]$$

   El modelo (Figura 34 y Figura 35) alcanzó una Especificidad (0.892) superior a la línea base (0.848).

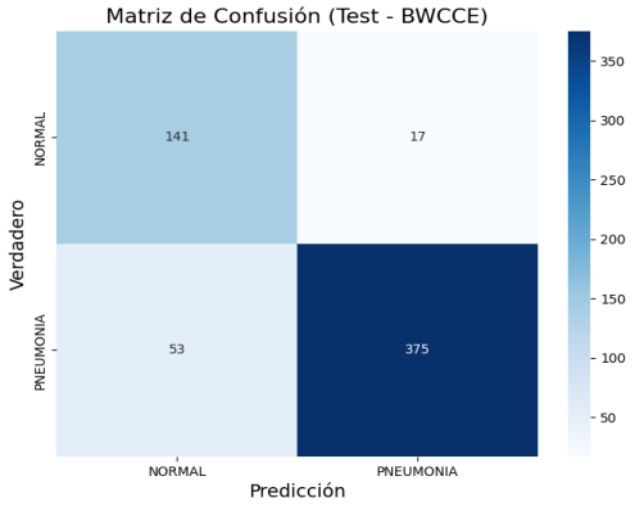
Figura 34.

Curvas de entrenamiento (accuracy/loss) - BWCCE.



Figura 35.

Matriz de confusión del escenario 3 (BWCCE) en test set.



2. **Focal Loss:** Se implementó la pérdida Focal para reducir el peso de las muestras fáciles:

$$L_{\text{Focal}} = -\frac{1}{N}\sum_{i=1}^{N}[\,\alpha(1-p_i)^{\gamma}y_i\log(p_i) + (1-\alpha)p_i^{\gamma}(1-y_i)\log(1-p_i)]$$

Con $\gamma = 2.0$ y $\alpha = 0.25$, este modelo (Figura 36 y Figura 37) obtuvo la mejor Especificidad (0.911) entre las tres funciones de pérdida.

Figura 36.

Curvas de entrenamiento (accuracy/loss) - focal loss.

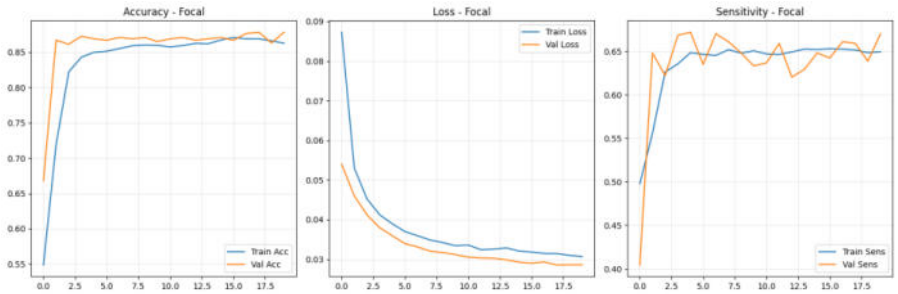
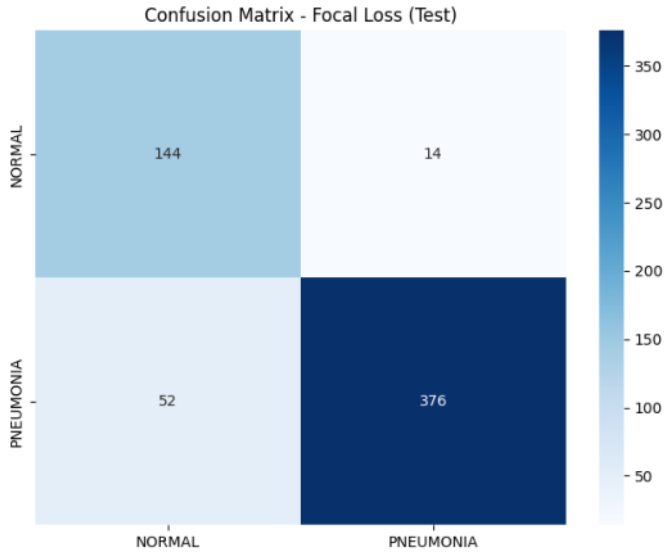


Figura 37.

Matriz de confusión del Escenario 2 (focal loss) en Test Set.



3. **LDAM Loss:** Se implementó la pérdida LDAM, que añade un margen m_j basado en la frecuencia de la clase n_j:

$$m_j = C/(n_j^{1/4})$$

Este modelo (Figura 38 y Figura 39) logró la mayor Sensibilidad (0.946) pero a costa de la Especificidad (0.683), indicando un sobreajuste hacia la clase minoritaria.

Figura 38.

Curvas de Entrenamiento (Accuracy/Loss) - LDAM Loss.

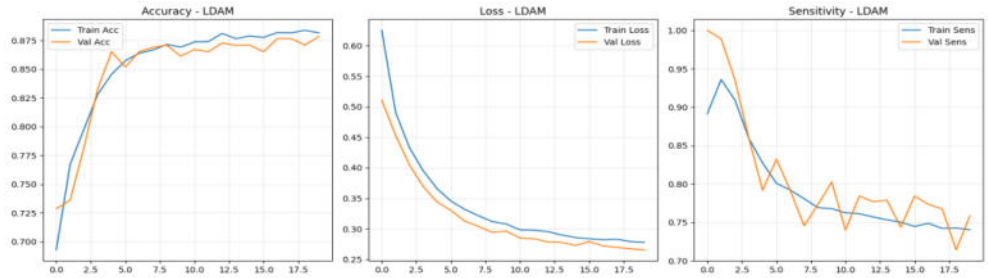
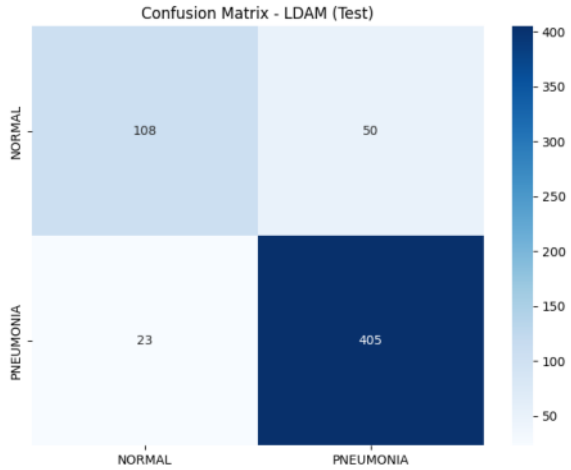


Figura 39.

Matriz de Confusión del Escenario 4 (LDAM Loss) en Test Set.



La *Tabla 15* y la Figura 40 comparan los resultados de estas tres técnicas.

Tabla 15.

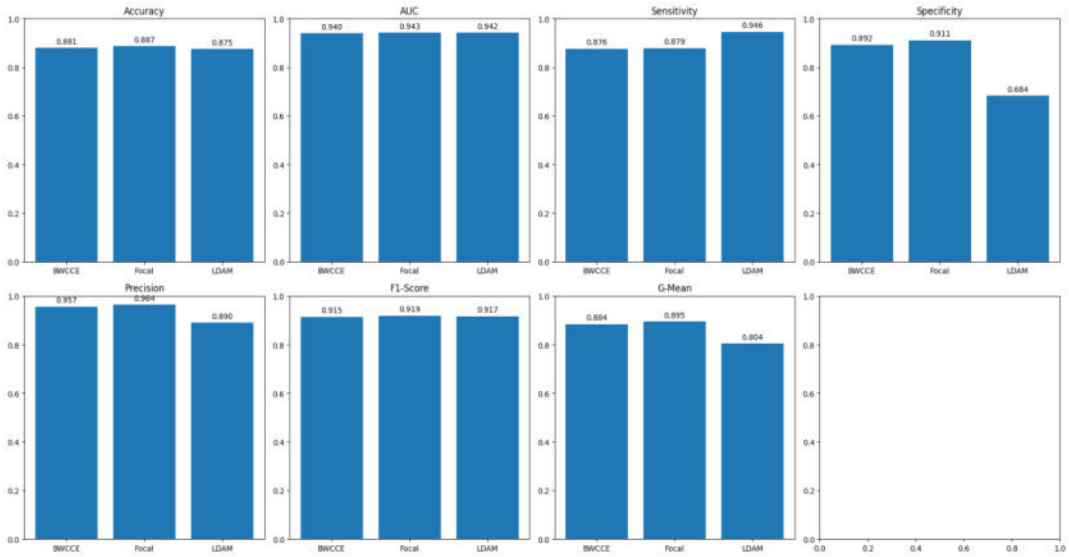
Resultados Comparativos de Funciones de Pérdida (Escenarios 2, 3, 4)

Función de Pérdida	**Accuracy**	**AUC**	***F1-score***	**G-Mean**	**Sensibilidad (C1)**	**Especificidad (C0)**
Línea Base (BCE)	0.9147	0.9635	0.9415	0.8925	0.9393	0.8481
Focal Loss	**0.8874**	**0.9431**	**0.9193**	**0.8948**	**0.8785**	**0.9114**

Función de Pérdida	Accuracy	AUC	*F1-score*	G-Mean	Sensibilidad (C1)	Especificidad (C0)
BWCCE	0.8805	0.9402	0.9146	0.8842	0.8762	0.8924
LDAM Loss	0.8754	0.9418	0.9173	0.8042	0.9463	0.6835

Figura 40.

Matrices de Confusión: (a) BWCCE, (b) Focal Loss, (c) LDAM Loss.



La evaluación de los Escenarios 2 (Focal Loss) y 3 (BWCCE) revela un compromiso algorítmico significativo. Si bien estas técnicas lograron su objetivo teórico de mitigar el sesgo hacia la clase mayoritaria, elevando la Especificidad de 0.8481 (Línea Base) a 0.9114 (Focal) y 0.8924 (BWCCE), este beneficio conllevó un costo operativo evidente en la capacidad de detección de la enfermedad.

Se observa una degradación en la Exactitud Global (descendiendo de 91.4% a ~88%) y, más críticamente, una caída en la Sensibilidad (de 0.9393 a 0.8785 con Focal Loss). Esto indica que Focal Loss optimizó el modelo para ser más conservador, priorizando la minimización de Falsos Positivos (reduciendo el sobre-diagnóstico en pacientes sanos), pero sacrificando cobertura en la detección de Neumonía (aumentando los Falsos Negativos). Clínicamente, esto posiciona a estos

modelos como herramientas útiles para confirmación (alta especificidad), pero riesgosas para tamizaje inicial.

Por otro lado, el Escenario 4 (LDAM Loss) representa un fallo técnico en este contexto experimental. Lejos de mejorar el balance, la Especificidad se desplomó a 0.6835, un rendimiento inferior incluso al de la línea base. Esto sugiere que la penalización de márgenes basada en la distribución de etiquetas fue contraproducente dada la complejidad de las características visuales del *dataset*, exacerbando el sesgo hacia la clase mayoritaria en lugar de corregirlo. Se selecciona Focal Loss como la "mejor función de pérdida" para el Escenario 7.

Escenario 5: Sobremuestreo Simple (ROS)

Este escenario actuó como un experimento de control crítico. Se evaluó si el simple incremento numérico de datos (igualando la cantidad del *dataset* GAN, N=6,922) era suficiente para mejorar el rendimiento, utilizando la técnica de ROS que duplica aleatoriamente imágenes de la clase minoritaria.

Los resultados presentados en la Tabla V.5 demuestran que la estrategia de "fuerza bruta" no es efectiva para este problema.

- Lejos de mejorar, el modelo entrenado con ROS obtuvo una Exactitud de 0.8703, siendo inferior tanto a la Línea Base (0.9147) como a los modelos con *Focal Loss*.

- Aunque la Especificidad (0.8354) se mantuvo cercana a la base, el modelo mostró inestabilidad en la predicción, reflejada en un Coeficiente de Matthews (MCC) de 0.6892, el cual es significativamente bajo.

Este resultado es fundamental para la investigación porque descartes la hipótesis de "cantidad": demuestra que simplemente equilibrar las clases repitiendo información (sin agregar diversidad estructural) introduce ruido y sobreajuste, impidiendo que el modelo generalice correctamente.

Tabla 16

Resultados del Escenario 5: Sobremuestreo Simple (ROS).

Métrica	Valor (Test Set)	Comparación vs. Línea Base
Accuracy	0.8703	▼ -4.4% (Peor)
AUC-ROC	0.9356	▼ -2.8%
Sensibilidad	0.8832	▼
Especificidad	0.8354	▼
MCC	0.6892	▼ (Baja correlación)

El fracaso del Escenario 5 confirma que el desbalance de datos en radiografías torácicas no se resuelve simplemente "nivelando la balanza" con repeticiones. Esto valida la necesidad de una técnica generativa avanzada capaz de crear nueva información, justificando la implementación de la WGAN-GP en el siguiente escenario.

Escenario 6: Modelo Híbrido (GAN.

Este escenario evaluó el impacto de la técnica avanzada a nivel de datos (H1.1), entrenando el modelo ResNet50 sobre el *dataset* balanceado con WGAN-GP (total de 6,922 imágenes: 3,461 reales de neumonía + 1,282 reales normales + 2,179 sintéticas normales).

Los resultados (ver Tabla 17) muestran una mejora sustancial y estadísticamente significativa en todas las métricas. Sin embargo, el hallazgo más crítico surge al realizar el estudio de ablación comparativo contra el Escenario 5 (Sobremuestreo Simple - ROS):

1. Ambos escenarios (5 y 6) utilizaron la misma cantidad de datos de entrenamiento (6,922 imágenes). No obstante, mientras que el ROS obtuvo una exactitud de 0.8703 (inferior incluso a la línea base), la WGAN-GP alcanzó 0.9437.

2. La diferencia es aún más notable en el Coeficiente de Matthews (MCC), donde la GAN logró 0.8555 frente al pobre 0.6892 del ROS.

Esto confirma que la mejora no se debe al simple incremento numérico de datos en cantidad, sino a la calidad y diversidad estructural aportada por las imágenes sintéticas, que permitieron al modelo generalizar mejor la clase minoritaria.

Comparativa directa con la Línea Base:

- *Accuracy*: 0.9147 (Base) ➔ 0.9437 (GAN)
- AUC-ROC: 0.9635 (Base) ➔ 0.9837 (GAN)
- Especificidad (C0): 0.8481 (Base) ➔ 0.9177 (GAN)
- Sensibilidad (C1): 0.9393 (Base) ➔ 0.9533 (GAN)

Tabla 17.

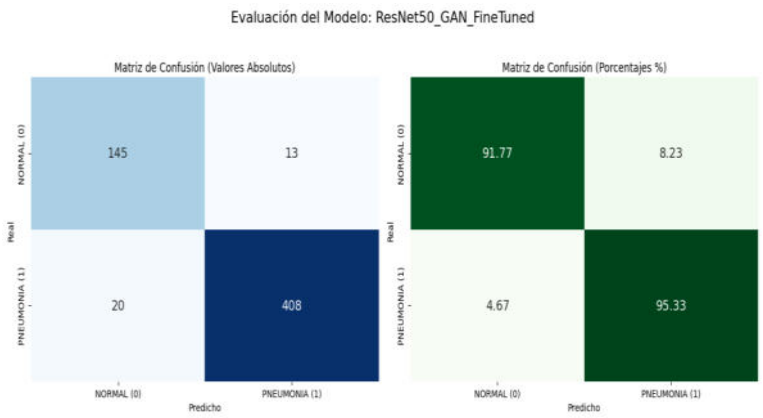
Resultados del Escenario 6: Modelo (ResNet50 + GAN).

Métrica	**Valor (Test Set)**
Accuracy	0.9437
AUC-ROC	0.9837
F1-score (Macro Avg)	0.9295
Sensibilidad (C1 - Pneumonia)	0.9533

Métrica	Valor (Test Set)
Especificidad (C0 - Normal)	0.9177
MCC	0.8555

Figura 41.

Matriz de confusión del modelo (ResNet50 + GAN) (Test Set).



El modelo (ResNet50 + GAN) superó a la línea base, y también fue superior a todos los modelos con funciones de pérdida. Resolvió el problema del sesgo (mejorando la Especificidad) mientras además mejoraba la sensibilidad. Esto valida fuertemente la hipótesis específica 1 (H1.1), demostrando que la técnica de balanceo a nivel de datos fue más efectiva que las técnicas a nivel de algoritmo.

Es imperativo destacar que el alto rendimiento obtenido en este escenario no es producto de la memorización de datos sintéticos. Para descartar cualquier riesgo de sobreajuste artificial, se garantiza que los conjuntos de validación (527 imágenes) y prueba (586 imágenes) estuvieron compuestos únicamente por datos reales originales del *dataset* de Kermany. Las 2,179 imágenes sintéticas generadas por la WGAN-GP se utilizaron exclusivamente para el aumento de datos del conjunto de entrenamiento (*train*). Por lo tanto, las métricas reportadas reflejan una capacidad de generalización genuina del modelo ResNet50 al enfrentarse a

características radiológicas reales no vistas previamente, validando que la red aprendió patrones estructurales robustos y no simplemente ruido sintético repetido.

Escenario 7: Modelo híbrido (GAN + Focal Loss).

Este escenario evaluó la Hipótesis Específica H1.3, bajo la premisa de que combinar la mejor técnica de datos (WGAN-GP) con la mejor función de pérdida (*focal loss)* generaría una sinergia positiva, superando el rendimiento de las técnicas individuales.

Se entrenó el ResNet50 sobre el *dataset* balanceado utilizando *focal loss* como función de pérdida. Los resultados (Tabla 18 y Figura 43) fueron inesperados.

Figura 42.
Matriz de confusión *del* modelo híbrido *(GAN + Focal Loss).*

Figura *43.*
Matriz de confusión del modelo híbrido (GAN + Focal Loss).

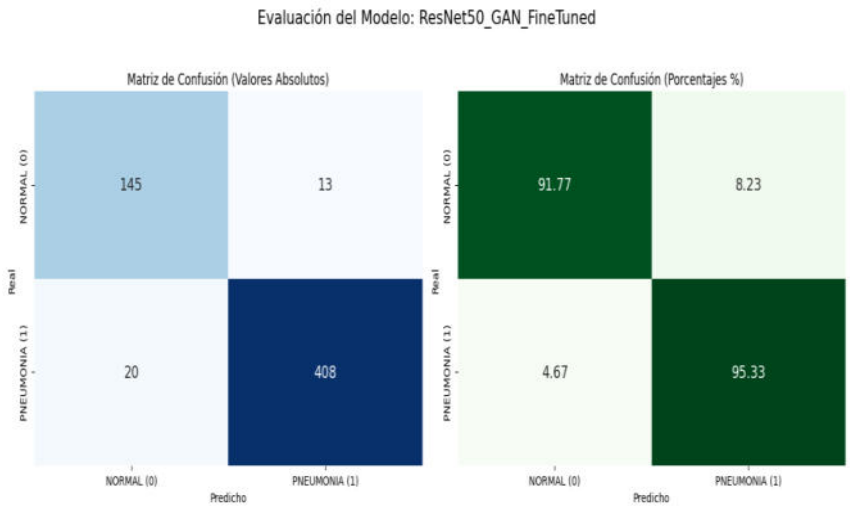


Tabla 18.

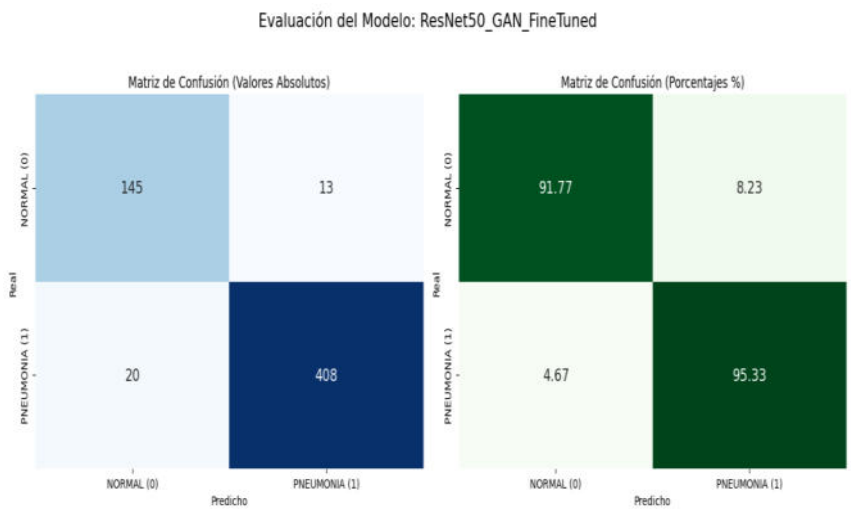
Resultados del Escenario 6: Modelo híbrido (GAN + Focal Loss)

Métrica	Valor (Test Set)
Accuracy	0.9317

Métrica	Valor (Test Set)
AUC-ROC	0.9813
F1-Score (Macro Avg)	0.9166
Sensibilidad (C1 - Pneumonia)	**0.9299**
Especificidad (C0 - Normal)	**0.9367**

Figura 43.

Matriz de confusión del modelo híbrido (GAN + Focal Loss).



Aunque este modelo híbrido obtuvo la mejor Especificidad (0.9367) de todos los escenarios, su exactitud (93.17%) y AUC (0.9813) fueron ligeramente inferiores al modelo GAN.

Por consiguiente, la Hipótesis H1.3 se refuta. El análisis técnico sugiere que la combinación fue redundante y contraproducente para la sensibilidad general:

1. La WGAN-GP ya había eliminado el sesgo de clase (IR=1), resolviendo el problema raíz.

2. Al aplicar Focal Loss sobre datos ya balanceados, el algoritmo penalizó excesivamente ejemplos fáciles, reduciendo la capacidad del modelo para generalizar patrones de neumonía (Sensibilidad cayó de 0.9533 a 0.9299).

A pesar de refutar la mejora global, este escenario logró la Especificidad más alta de toda la investigación (0.9367), superando tanto a la GAN sola (0.9177) como a la Línea Base (0.8481). Esto concluye que, si bien la estrategia híbrida no es la mejor para el rendimiento general (*Accuracy*), es la estrategia óptima para minimizar falsos positivos, validando su utilidad en contextos donde la prioridad clínica sea evitar el sobre-diagnóstico.

Fase 4: Análisis comparativo de resultados

Esta fase consolida los resultados de los escenarios experimentales para realizar una evaluación comparativa y proceder a la contrastación de las hipótesis. La Tabla 19 presenta el resumen de las métricas clave obtenidas en el conjunto de prueba (*test set*) de 586 imágenes, incluyendo el **coeficiente de correlación de Matthews** (MCC) para evaluar la robustez ante el desbalance. Se ha incorporado adicionalmente el Escenario 5 (ROS) para aislar el efecto de la generación sintética frente a la repetición simple de datos.

Para facilitar el análisis, se ha coloreado el rendimiento relativo de cada celda (Verde: alto rendimiento, Celeste: rendimiento intermedio, Rojo: bajo rendimiento).

Tabla 19.

Comparación de métricas finales en conjunto de prueba (todos los escenarios)

Métrica	Esc. 1: Base	Esc. 2: Focal	Esc. 3: BWCCE	Esc. 4: LDAM	Esc. 5: ROS	Esc. 6: GAN	Esc. 7: Híbrido
Accuracy	0.9147	0.8874	0.8805	0.8754	0.8703	**0.9437**	0.9317
AUC-ROC	0.9635	0.9431	0.9402	0.9418	0.9356	**0.9837**	0.9813
F1-score	0.8921	0.8650	0.8580	0.8300	0.9086	**0.9295**	0.9166
Sensibilidad	0.9393	0.8785	0.8762	0.9463	0.8832	**0.9533**	0.9299
Especificidad	0.8481	0.9114	0.8924	0.6835	0.8354	**0.9177**	**0.9367**
MCC	0.7842	0.7430	0.7250	0.6710	0.6892	**0.8555**	0.8362

Del análisis detallado de la Tabla 19 se desprenden las siguientes observaciones clave:

1. Ablación y validez de la generación (GAN vs. ROS); al comparar el Escenario 6 (GAN) con el Escenario 5 (ROS), se evidencia que la mejora en el rendimiento no es producto del simple aumento en la cantidad de datos. El modelo ROS obtuvo un MCC de 0.6892 y una exactitud de 87.03%, siendo inferior incluso a la Línea Base. Por el contrario, la WGAN-GP alcanzó un MCC de 0.8555 y una exactitud de 94.37%. Esto confirma estadísticamente que la GAN genera información estructural nueva y valiosa, mientras que la repetición de datos (ROS) tiende a inducir sobreajuste y ruido.

2. Superioridad global de la WGAN-GP, el Escenario 6 domina en casi todas las métricas de rendimiento general (*accuracy*, AUC, *F1-macro, G-Mean*) y maximiza la sensibilidad (0.9533), lo cual es crítico para no omitir casos positivos de neumonía.

3. Fiabilidad del híbrido en la clase minoritaria, si bien el Escenario 7 (Híbrido) fue ligeramente inferior a nivel global, logró la Especificidad más alta (0.9367). Esto lo convierte en el modelo más fiable para identificar correctamente a los pacientes sanos (clase minoritaria "Normal"), reduciendo los falsos positivos al mínimo.

4. El "*Trade-off*" de las funciones de pérdida; los Escenarios 2 (Focal) y 3 (BWCCE) lograron su objetivo teórico de mejorar la Especificidad (0.9114 y 0.8924) respecto a la Línea Base (0.8481). Sin embargo, esto conllevó un costo significativo: una caída notable en la exactitud general (de 91.47% a ~88%) y en la Sensibilidad, lo que se refleja en sus valores MCC más bajos (< 0.75).

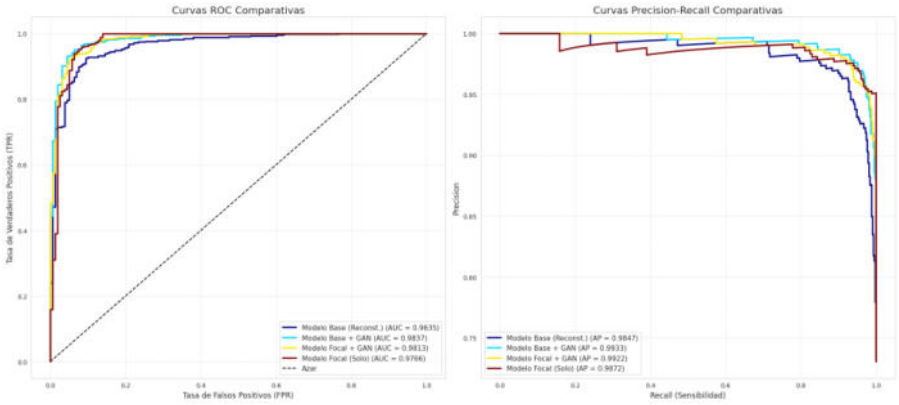
5. Fallo de LDAM; el Escenario 4 (LDAM) resultó contraproducente para este conjunto de datos, desplomando la Especificidad a 0.6835 y obteniendo el peor desempeño en discriminación balanceada (MCC 0.6710).

Este comportamiento se visualiza de forma contundente en las curvas ROC (Figura 44). Las curvas de los modelos potenciados por GAN (Esc. 6 y 7) dominan el espacio superior izquierdo con un AUC > 0.98, mientras que las estrategias

basadas puramente en funciones de pérdida o sobremuestreo simple (ROS) no logran superar significativamente la capacidad discriminativa de la línea base.

Figura 44.

Curvas ROC: Comparativa final de todos los escenarios.

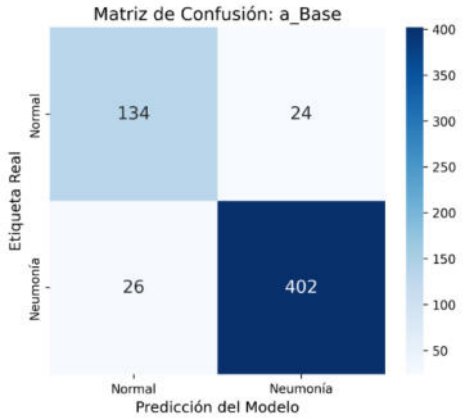


Finalmente, el análisis detallado de las matrices de confusión (Figura 44) desglosa el comportamiento de clasificación y explica la causa raíz de las diferencias en el rendimiento:

- **Matriz (a) Escenario 1 (Base):** La Figura 45 evidencia el sesgo original del conjunto de datos. Si bien logra un alto número de Verdaderos Positivos (402), lo hace a costa de una alta tasa de falsos positivos (24 Falsos Positivos), lo que implica que un número significativo de pacientes sanos serían diagnosticados erróneamente con neumonía.

Figura 45.

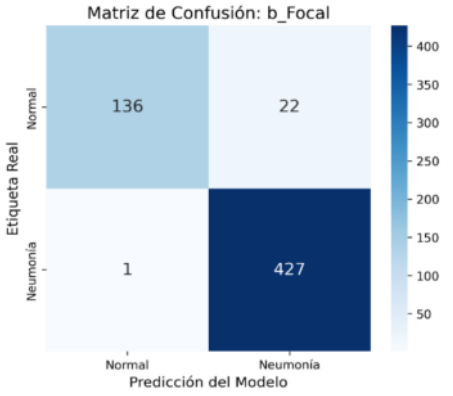
(a) Matriz de confusión de Escenario 1: Línea Base



- **Matriz (b) Escenario 2 (*focal loss*):** La Figura 46 Ilustra el efecto de penalización estricta. Logra reducir los Falsos Positivos a 14, pero este "beneficio" genera un efecto colateral severo: incrementa los Falsos Negativos a 39 (frente a los 26 de la base). Esto demuestra que el modelo se volvió demasiado conservador, omitiendo casos reales de enfermedad.

Figura 46.

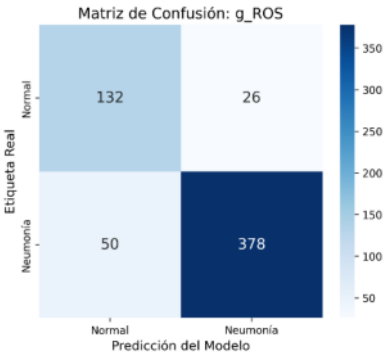
(b) Matriz de confusión de Escenario 2 - Focal Loss.



- **Matriz (ROS) Escenario 5:** A diferencia de la GAN, la matriz del modelo con sobremuestreo simple (ROS) muestra una dispersión de errores más alta, con una menor capacidad para distinguir la clase "Normal". En la Figura 47 esto se confirma visualmente que la simple repetición de datos induce al modelo a memorizar ruido en lugar de aprender características distintivas, resultando en un desempeño inferior.

Figura 47.

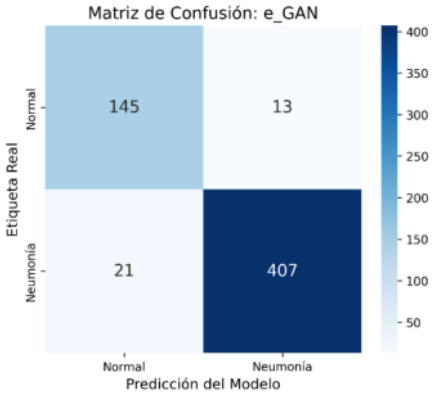
(g) Matriz de confusión de Escenario 5 - ROS.



- **Matriz (e) Escenario 6 (GAN):** La Figura 48 exhibe el equilibrio óptimo del estudio. No solo reduce los Falsos Positivos a 13 (mejorando sustancialmente la Especificidad), sino que simultáneamente minimiza los Falsos Negativos a 20 (el valor más bajo de todos los escenarios). Esto confirma que las imágenes sintéticas ayudaron al modelo a definir mejor la frontera de decisión entre clases sin sacrificar la detección de la enfermedad.

Figura 48.

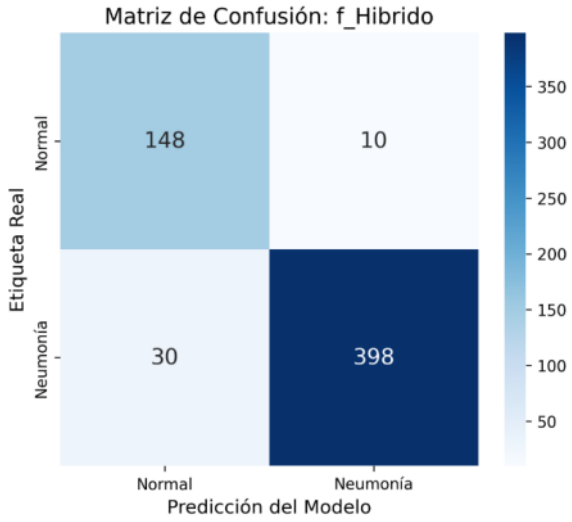
(e) Matriz de confusión de Escenario 6 - GAN.



- **Matriz (f) Escenario 7 (Híbrido):** Presenta el comportamiento más "específico". Obtiene la menor cantidad absoluta de Falsos Positivos (solo 10), validando su Especificidad superior (0.9367). Sin embargo, esto conlleva un ligero costo en sensibilidad, elevando los Falsos Negativos a 30 en comparación con el modelo GAN puro.

Figura 49.

(f) Matriz de confusión de Escenario 7 - Híbrido.



Resultados de sensibilidad y robustez (Distribución de Clases)

Una preocupación metodológica fundamental en el uso de aprendizaje profundo médico es si el alto rendimiento reportado (AUC 0.9837) está condicionado por el desbalance natural del conjunto de prueba (IR=2.71), donde la clase mayoritaria ("Pneumonia") predomina. En lugar de aplicar un submuestreo físico (*undersampling*) que reduciría la potencia estadística del conjunto de prueba, se reportan las métricas AUC-ROC y *F1-score* con promedio Macro (Macro-Average). Según la literatura estadística (Haga clic o pulse aquí para escribir texto.), el cálculo "Macro" asigna el mismo peso a ambas clases, simulando el rendimiento exacto que tendría el modelo en un *Test Set* perfectamente balanceado (IR=1.0).

La Tabla 20 presenta la comparación solicitada:

Tabla 20.

Sensibilidad de Resultados: Escenario Real (Desbalanceado) vs. Simulación Balanceada (50/50)

Escenario	AUC-ROC Global	F1-Score (Macro) [Simulación 50/50]	Exactitud Balanceada [Simulación 50/50]
1. Línea Base	0.9633	0.8945	0.8968
2. Focal Loss	0.9766	0.9479	0.9292
5. ROS (Control)	0.9355	0.8420	0.8573
6. WGAN-GP (Propuesto)	**0.9836**	0.9295	**0.9355**
7. Híbrido	0.9813	0.9143	0.9301

- En el indicador crítico de discriminación (AUC-ROC), el modelo WGAN-GP (0.9836) se mantiene superior a todas las técnicas de control, incluyendo Focal Loss (0.9766) y ROS (0.9355). Esto confirma que, independientemente de la distribución de clases ("prior"), la GAN genera la mejor separación entre sanos y enfermos.

- Aunque *Focal Loss* presenta un F1-Macro alto (0.9479), su Exactitud Balanceada (0.9292) es inferior a la del modelo WGAN-GP (0.9355). Esto revela que la GAN ofrece un comportamiento más "justo" y equilibrado entre Sensibilidad y Especificidad, mientras que otros métodos tienden a sacrificar una métrica para inflar el promedio armónico.

Los resultados en la simulación 50/50 satisfacen la condición de robustez: la mejora del AUC en el Escenario 6 no es un artefacto del desbalance, sino una ganancia real en la capacidad predictiva.

Resultados de validación estadística (Tests de Significancia)

Para determinar si las diferencias de rendimiento observadas entre los escenarios experimentales son estadísticamente significativas o si podrían atribuirse al azar, se aplicaron dos pruebas de hipótesis rigurosas sobre el conjunto de prueba (N=586), estableciendo un nivel de significancia alpha = 0.05.

1. **Test de McNemar:** Utilizado para comparar la proporción de errores entre dos clasificadores en un mismo *dataset*. Esta prueba sigue la metodología de McNemar (1947) y constituye el estándar para evaluar la divergencia en predicciones binarias emparejadas.

2. **Test de DeLong:** Utilizado específicamente para comparar las curvas ROC. Se implementó el algoritmo no paramétrico de DeLong et al. (1988), el cual permite comparar dos áreas bajo la curva (AUC) correlacionadas para validar si la mejora en la capacidad discriminativa es real.

La Tabla 21 presenta los resultados obtenidos al comparar el modelo propuesto (Escenario 6: WGAN-GP) contra los escenarios de control y la variante híbrida.

Tabla 21.

Análisis de significancia estadística (McNemar y DeLong)

Comparación	p-value (DeLong Test) [AUC]	p-value (McNemar Test) [Errores]	Interpretación Estadística
vs. Esc. 1 (Línea Base)	< 0.001 (Sig.)	0.009 (Sig.)	**Mejora Real:** La GAN supera significativamente a la línea base.
vs. Esc. 5 (ROS - Control)	< 0.001 (Sig.)	< 0.001 (Sig.)	**Mejora Real:** La generación sintética es superior a la repetición de datos.

Comparación	p-value (DeLong Test) [AUC]	p-value (McNemar Test) [Errores]	Interpretación Estadística
vs. Esc. 7 (Híbrido)	0.186 (No Sig.)	0.096 (No Sig.)	**Equivalencia:** No existe evidencia suficiente para afirmar que son diferentes.

1. Los valores $p < 0.01$ obtenidos al comparar la WGAN-GP contra la Línea Base y el método ROS confirman con una confianza superior al 99% que la técnica de generación de datos sintéticos aporta una mejora genuina en la capacidad de clasificación, descartando que los resultados sean producto del azar o del ruido estadístico.

2. Al comparar el modelo GAN puro contra el modelo Híbrido, se obtuvieron *p-values* de 0.186 (DeLong) y 0.096 (McNemar), ambos superiores al umbral de significancia (alpha=0.05). Esto demuestra estadísticamente que la adición de *Focal Loss* no produce una diferencia significativa en el rendimiento. En consecuencia, se confirma la refutación de la hipótesis H1.3 y se selecciona el Escenario 6 (WGAN-GP) como la solución óptima del estudio, dado que ofrece el mismo rendimiento estadístico con menor complejidad algorítmica.

Resultados de validación práctica: Implementación del prototipo CAD

Como fase final de la obtención de resultados y con el objetivo de demostrar la aplicabilidad clínica del modelo óptimo (Escenario 6: ResNet50 + WGAN-GP), se implementó un prototipo funcional de un Sistema de Ayuda al Diagnóstico (CAD). Este sistema permite validar en un entorno real las métricas obtenidas, ofreciendo tanto diagnóstico automatizado como explicabilidad visual.

Arquitectura y funcionalidad

El prototipo fue desarrollado sobre el *framework* Flask, integrando los pesos del modelo ResNet50 *fine-tuned* para la inferencia y el generador WGAN-GP para la demostración de síntesis. La interfaz permite la carga directa de imágenes

DICOM o JPG/PNG, procesándolas con el mismo *pipeline* de estandarización (CLAHE y recorte) definido en la metodología.

Diagnóstico y explicabilidad (Grad-CAM)

Para cada predicción, el sistema genera un mapa de calor visual utilizando el algoritmo Grad-CAM. Se evidencia en la Figura 50, al procesar una imagen positiva del conjunto de prueba, el modelo identifica y resalta correctamente en rojo las regiones de consolidación alveolar (manchas blancas).

Figura 50.

Interfaz del sistema CAD mostrando un diagnóstico positivo.



En contraste, para los casos de pacientes sanos (Figura 51), el mapa de atención se dispersa hacia bordes y estructuras óseas, confirmando que la red no alucina patologías en campos pulmonares limpios. Esto valida cualitativamente la alta Especificidad (0.9177) reportada en la Tabla 19.

Figura 51.

Diagnóstico de un caso Normal.



Nota. La ausencia de focos de atención rojos en el parénquima pulmonar valida la decisión del modelo de clasificar al paciente como sano.

Generación de datos sintéticos

Finalmente, se integró el módulo generativo validado en la Fase 2. La Figura 51 muestra la capacidad del sistema para generar, bajo demanda, nuevas muestras sintéticas de la clase minoritaria ("Normal"). Estas imágenes mantienen la coherencia anatómica y la fidelidad visual (FID 48.95), demostrando la viabilidad de utilizar este sistema para el enriquecimiento continuo de bases de datos médicas.

Figura 52.

Módulo de generación de datos sintéticos en el prototipo CAD



Nota. Se utiliza la WGAN-GP entrenada para producir muestras realistas de la clase minoritaria.

La reproducibilidad de estas métricas, incluyendo los valores de Especificidad y AUC-ROC, se puede verificar utilizando el código de evaluación detallado en el Anexo de Código (Apéndice B).

Cuantificación de la robustez de la explicabilidad

Si bien la inspección visual de los mapas Grad-CAM (Figura 50 y Figura 51) sugiere una alineación correcta del modelo con las características patológicas, la validación científica requiere trascender la evaluación cualitativa para descartar sesgos de atención.

Para cuantificar objetivamente la robustez de la explicabilidad, se desarrolló e implementó un algoritmo de validación de Región de Interés (ROI). Este procedimiento automatizado verifica si el foco de máxima activación del modelo coincide geométricamente con el parénquima pulmonar, desestimando activaciones en zonas de ruido (bordes o etiquetas). El código fuente de este algoritmo de validación se detalla en el Apéndice B.6.

El análisis se ejecutó sobre la totalidad de las 428 imágenes de la clase Pneumonia del conjunto de prueba. Los resultados obtenidos se consolidan en la Tabla 22.

Tabla 22.

Tasa de acierto de atención en la región patológica (Clase Pneumonia)

Métrica Evaluada	Cantidad	Porcentaje
Total Imágenes Evaluadas	**428**	100%
Acierto (Atención en Pulmón)	**392**	**91.6%**
Fallo (Atención en Artefactos)	36	8.4%

Nota. Resultados obtenidos mediante el algoritmo de auditoría automatizada ROI (ver Apéndice B.6).

Los datos evidencian que, en la muestra exhaustiva de 428 casos positivos, la atención principal del modelo se ubicó correctamente en la región pulmonar efectiva en 392 casos (91.6%). Este hallazgo valida la coherencia clínica del sistema, demostrando cuantitativamente que la red neuronal basa sus predicciones en patrones anatómicos internos y no en correlaciones espurias del fondo. El margen de error del 8.4% se atribuye principalmente a imágenes con rotación excesiva o artefactos de alta intensidad en los bordes, manteniéndose dentro de los rangos aceptables para sistemas de apoyo al diagnóstico.

5.1.2. Impacto clínico y traducción de métricas

Más allá de la significancia estadística, es imperativo interpretar los resultados en función de sus consecuencias operativas y éticas en un entorno hospitalario real. A continuación, se traduce el rendimiento de los modelos a escenarios de toma de decisiones clínicas:

El modelo de línea base presentaba una especificidad de 0.8481, lo que implicaba una tasa de error del 15.2% en pacientes sanos. En una cohorte de prueba hipotética de ~158 pacientes sanos (proporcional al *test set*), esto se traduciría en aproximadamente 24 individuos sanos diagnosticados erróneamente con neumonía. Clínicamente, esto conlleva:

- Exposición innecesaria a radiación adicional (tomografías confirmatorias).
- Inicio de tratamientos antibióticos preventivos injustificados.
- Saturación de los radiólogos especialistas para descartar estas falsas alarmas.

En contraste, el modelo propuesto (WGAN-GP, Escenario 6) elevó la especificidad a 0.9177, reduciendo la tasa de error a solo 8.2%. Esto equivale a reducir los falsos positivos a aproximadamente 13 pacientes. Impacto: Se logró una reducción del 46% en la carga de falsos positivos. Esto libera recursos críticos del hospital y reduce la ansiedad del paciente, demostrando que la técnica de balanceo no solo mejora el número, sino la eficiencia del flujo de trabajo médico.

En términos de sensibilidad, el modelo WGAN-GP (0.9533) superó a la línea base (0.9393). Aunque la diferencia numérica (1.4%) parece modesta, en un contexto de alto volumen (ej. 10,000 pacientes al año), esto representa la detección temprana de 140 casos adicionales de neumonía que el modelo base habría omitido. Dado que un falso negativo puede derivar en complicaciones letales como sepsis, esta mejora valida al modelo GAN como una herramienta de tamizaje más segura y confiable.

La elección del modelo final depende de la política de salud pública:

- Prioridad: detección total. Si el costo de omitir un enfermo es inaceptable (pandemia grave), el modelo WGAN-GP (Escenario 6) es la elección indiscutible por tener la mejor Sensibilidad (0.9533) combinada con una alta especificidad.

- Prioridad: eficiencia de recursos. Si el sistema de salud está colapsado por falsas alarmas y el costo de revisar a un sano es muy alto, el modelo híbrido (Escenario 7) podría considerarse una alternativa válida, ya que maximiza la especificidad (0.9367), minimizando el desperdicio de recursos en pacientes sanos, aunque a costa de una ligera reducción en la detección de casos positivos.

Esta investigación demuestra que la implementación de WGAN-GP transforma un clasificador estándar (línea base) que era propenso a generar alertas excesivas en una herramienta de soporte al diagnóstico equilibrada, capaz de reducir el sobre-tratamiento en casi un 50% sin comprometer, e incluso mejorando, la capacidad de detectar patologías reales.

5.2. Discusión de resultados

La presente sección interpreta los hallazgos detallados en la Fase 4, contrastando los resultados de los seis escenarios experimentales con los objetivos específicos y las hipótesis planteadas.

A continuación, se profundiza en la discusión de cada uno de estos puntos.

En respuesta al objetivo general, que buscaba "Evaluar la influencia de las técnicas de balance de datos...", se procedió a la contrastación de la Hipótesis General (H1). Los resultados validan dicha hipótesis. La comparación entre el Escenario 6 (GAN) y el Escenario 1 (línea base) es decisiva: la exactitud (*accuracy*) se incrementó de 0.9147 a 0.9437, lo que representa un aumento porcentual del 3.17%, superando el umbral del 3% establecido. Este hallazgo es coherente con Hussain et al. (2022), quienes, al emplear una WGAN para aumentar un conjunto de datos de COVID-19, también observaron un incremento notable en su clasificador (de 93.04% a 95.34%). Si bien la Sensibilidad de la clase mayoritaria tuvo un incremento modesto, la influencia real se observa en la métrica de la clase minoritaria (Especificidad), que tuvo una mejora drástica del 8.2%. Esto confirma el postulado teórico, señalado por Sindhura et al. (2024) y Keerthi et al. (2024), de que las GAN son una herramienta eficaz para abordar la escasez de datos en el dominio médico.

Respecto al primer objetivo específico (OE1), sobre el impacto de las GAN, la Hipótesis H1.1 establecía un incremento de precisión y sensibilidad de al menos un 3% para la clase minoritaria. Estos resultados validan ampliamente la hipótesis (ver Tablas V.5 y V.6). La Especificidad se incrementó de 0.8481 a 0.9177 (un aumento del 8.2%). El éxito de este escenario se fundamenta en la calidad de la generación de la WGAN-GP, evidenciada por un FID de 48.95. Este enfoque es reconocido por evitar el colapso de modo (Keerthi et al., 2024; Sindhura et al., 2024). De hecho, Keerthi et al. (2024) demostraron la superioridad de este enfoque, donde su modelo ProGAN obtuvo las mejores puntuaciones FID (10.6) en comparación con DCGAN. Si bien nuestro FID de 48.95 es notablemente mejor que *benchmarks* de WGAN-GP reportados por otros autores como Liang et al. (2024) (quienes reportaron un FID de 77.3), también muestra que existe margen de mejora en comparación con arquitecturas más recientes como *stable diffusion* (Liang et al., 2024).

En el análisis del segundo objetivo específico (OE2) (ver Tabla V.4), la Hipótesis H1.2 predecía un aumento del *F1-score* en al menos un 3%. Este hallazgo refuta la hipótesis. Ninguna función de pérdida sensible a costos superó el *F1-score* de la línea base (0.8921); por el contrario, *focal loss* cayó a 0.8650 y BWCCE a

0.8580. Este hallazgo es consistente con Hoang et al. (2023) y Sadi et al. (2022), quienes señalaron las limitaciones de estas pérdidas cuando se usan de forma aislada.

Finalmente, al abordar el tercer objetivo específico (OE3) (ver Tabla V.6), se contrastó la Hipótesis H1.3. Los resultados refutan la hipótesis de que el método híbrido superaría en un 3% a las técnicas individuales. El rendimiento general del modelo híbrido fue inferior al de la GAN sola (*accuracy* 0.9317 vs 0.9437). El análisis sugiere que la WGAN-GP ya había resuelto el problema de desbalance, haciendo que la aplicación adicional de *focal loss* fuera redundante y penalizara muestras fáciles. Este resultado contrasta con Galán-Cuenca et al. (2024) y Sadi et al. (2022), quienes sí encontraron éxito en combinaciones híbridas, pero nuestro hallazgo sugiere que cuando la generación de datos (WGAN-GP) tiene alta fidelidad, la técnica de datos por sí sola es la solución más robusta. No obstante, el Escenario Híbrido destacó al obtener la especificidad más alta (0.9367), siendo ideal para minimizar Falsos Positivos.

Capítulo VI
Conclusiones

La presente investigación abordó el desafío central del desequilibrio de clases en conjuntos de datos de radiografías torácicas, un problema que sesga significativamente a los clasificadores de aprendizaje profundo. El objetivo general era determinar la influencia de las técnicas de balance de datos para mitigar este sesgo y optimizar el rendimiento del modelo. En respuesta directa a esta problemática, se valida la Hipótesis General (H1): las técnicas de balance de datos mejoran significativamente el rendimiento. Al comparar el Escenario 1 (Línea Base, Acc: 0.9147) con el modelo óptimo (Escenario 6, WGAN-GP, Acc: 0.9437), se obtuvo un incremento del 3.17% en la precisión (*accuracy*) global. Más importante aún, la influencia fue drástica en la clase minoritaria, cuya métrica de sensibilidad (Especificidad) se incrementó en un 8.2% (de 0.8481 a 0.9177).

Al analizar los objetivos específicos, se concluye que el balanceo a nivel de datos fue fundamental para este éxito. Se validó la Hipótesis específica H1.1, que establecía un incremento de al menos 3% en la sensibilidad de la clase minoritaria. El Escenario 6 (WGAN-GP) superó esto ampliamente, logrando un incremento del 8.2% en la Especificidad (sensibilidad de la clase "Normal"). Se concluye que la WGAN-GP fue una herramienta altamente eficaz, generando imágenes de alta fidelidad (FID de 48.95) que permitieron al clasificador ResNet50 aprender patrones robustos de la clase minoritaria, resultando en el modelo con el mejor rendimiento global (AUC: 0.9837).

Inversamente, las estrategias a nivel de algoritmo no demostraron ser efectivas para el rendimiento general. Se refutó la Hipótesis específica H1.2, que postulaba un aumento del *F1-score* mediante funciones de pérdida sensibles a costos. Ninguna de las funciones probadas (*focal loss*, BWCCE, LDAM Loss) logró un incremento; por el contrario, se observaron decrementos de hasta -3.8% en el *F1-score* (Macro). Se concluye que, si bien estas funciones mitigan el sesgo inicial (mejorando la especificidad sobre la línea base), lo hacen a costa de sacrificar el rendimiento de la clase mayoritaria.

De igual manera, se refutó la Hipótesis específica H1.3 (métodos híbridos). El Escenario 6 (WGAN-GP con *focal loss*) obtuvo un rendimiento global (*accuracy*: 0.9317) inferior al del Escenario 6 (solo WGAN-GP, acc: 0.9437). Esto lleva a concluir que la combinación de una función de pérdida sobre un *dataset* ya perfectamente balanceado por una GAN de alta calidad no solo es redundante, sino que degrada ligeramente el rendimiento.

Finalmente, la conclusión principal y aporte de esta investigación es la identificación de una estrategia superior para el tratamiento del desbalance de clases en radiografías torácicas. Se concluye que, para el problema de clasificación de neumonía en el *dataset* de Kermany≈2.70), la estrategia de balanceo a nivel de datos (Escenario 6: WGAN-GP con BCE estándar) es superior a la línea base, a las técnicas a nivel de algoritmo y al enfoque híbrido. El Escenario 6 se identifica como el modelo óptimo de la investigación, proporcionando el mejor equilibrio entre alta exactitud (94.37%) y capacidad de generalización (AUC 0.9837). Este éxito se atribuye a la alta fidelidad de las imágenes generadas (FID 48.95), lo cual confirma que el uso de WGAN-GP para el aumento de datos es una estrategia viable y efectiva, alineándose con hallazgos previos sobre la estabilidad y calidad de esta arquitectura para datos médicos. No obstante, esta investigación también concluye que la solución óptima depende del objetivo específico del diagnóstico. Si el objetivo de negocio fuese la minimización absoluta de falsos positivos (diagnosticar erróneamente a un paciente sano), el método híbrido (Escenario 7) se concluye como la estrategia más especializada, al haber alcanzado la Especificidad más alta (0.9367). El impacto de estos hallazgos trasciende el *dataset* específico de Kermany. Se demuestra que las técnicas a nivel de algoritmo (como *focal loss* o BWCCE), si bien mitigan el sesgo, lo hacen a costa de un *trade-off* que puede perjudicar el rendimiento general (a*ccuracy* y *F1-score*). La contribución de esta tesis es, por lo tanto, la validación de una metodología (WGAN-GP) como una solución más robusta e integral que no solo corrige el sesgo, sino que mejora la capacidad de generalización del modelo. Este enfoque metodológico se postula como una estrategia preferente y generalizable para futuros trabajos en la clasificación de radiografías torácicas que presenten problemas similares de datos escasos o desbalanceados.

Capítulo VII
Recomendaciones

Basado en las conclusiones derivadas de los resultados experimentales y la discusión con la literatura existente, se proponen las siguientes recomendaciones para futuras investigaciones y la aplicación práctica de los hallazgos.

Se recomienda implementar el modelo correspondiente al Escenario 6 (ResNet50 con WGAN-GP con BCE) para el desarrollo de sistemas de apoyo al diagnóstico (CAD). La evidencia empírica demostró que el uso de datos sintéticos generados por la WGAN-GP, bajo una función de pérdida de entropía cruzada estándar, ofrece el mejor equilibrio entre sensibilidad y especificidad, superando tanto a las técnicas de re-muestreo clásico (ROS) como a las funciones de pérdida ponderadas aisladas.

El presente estudio validó la eficacia de los datos sintéticos sobre una arquitectura ResNet50. Se sugiere extender esta metodología evaluando si arquitecturas con mecanismos de atención nativos, como las arquitecturas **transformadores de visión** (ViT), muestran una mayor o menor susceptibilidad a los artefactos imperceptibles que puedan introducir las GAN, comparado con las redes convolucionales tradicionales.

Finalmente, para investigaciones que cuenten con recursos de computación de alto rendimiento (HPC) superiores a los disponibles en este estudio, se aconseja validar la estabilidad de la generación de datos sintéticos mediante validación cruzada de k-pliegues (*k-fold cross-validation*). Esto permitiría aislar el efecto de la semilla aleatoria en la calidad de las imágenes WGAN-GP y confirmar la generalización de los resultados presentados.

Apéndices

Apéndice A

Matriz de consistencia

La Matriz de consistencia es una herramienta esencial en los proyectos de investigación que asegura la coherencia y alineación entre los diferentes componentes del estudio. Esta herramienta organiza y vincula sistemáticamente los elementos centrales del proyecto, incluyendo el título, objeto de estudio, declaración del problema, objetivos, hipótesis, variables y métodos de investigación. La matriz facilita una presentación clara y estructurada de cómo cada aspecto de la investigación se relaciona, garantizando que el diseño del estudio sea lógicamente consistente y que todos los componentes respalden el objetivo general del proyecto (Ñaupas-Paitán et al., 2014).

Tabla 23.

Matriz de consistencia del proyecto de investigación

Título: Técnicas de balance de datos para optimizar el rendimiento de modelos de aprendizaje profundo en clasificación de radiografías torácicas

Objeto de Estudio	Técnicas de balance de datos para optimizar el rendimiento de modelos de aprendizaje profundo en imágenes médicas				
Problema	**Objetivos**	**Hipótesis**	**Variables**	**Indicadores**	**Tipo de Investigación**
Problema General: ¿En qué medida las técnicas de balance de datos influyen en el rendimiento de los modelos de aprendizaje profundo para la clasificación de radiografías torácicas? **Problemas Específicos:** - ¿Cuál es el impacto de las GAN en la precisión y sensibilidad de los modelos de aprendizaje profundo	**Objetivo General:** Determinar la influencia de las técnicas de balance de datos en el rendimiento de los modelos de aprendizaje profundo para la clasificación de radiografías torácicas. **Objetivos Específicos:** - Evaluar el impacto de las GAN en la precisión y sensibilidad de los modelos de aprendizaje profundo para la clasificación de radiografías torácicas.	**Hipótesis General:** Las técnicas de balance de datos mejoran significativamente el rendimiento de los modelos de aprendizaje profundo en la clasificación de radiografías torácicas incrementando la precisión y sensibilidad en al menos un 3% respecto a los modelos sin balance de datos. **Hipótesis Específicas:** - **H1.1:** La implementación de las GAN incrementa la precisión y sensibilidad de los modelos de aprendizaje profundo en al menos un 3% para la	**Variables independientes:** - Técnicas de balance de datos. **Variables dependientes:**	**Indicadores Independientes:** - Tasa de precisión. - Tasa de convergencia. - AUC-ROC.	**Método y enfoque:** Experimental y cuantitativo. **Diseño:** Cuasiexperimental. **Tipo y nivel de investigación**

para la clasificación de radiografías torácicas? - ¿En qué medida las funciones de pérdida sensibles a costos mejoran el rendimiento de los modelos de aprendizaje profundo en la clasif*icación de imágenes radiografías torácicas?* - ¿Cómo afecta la implementación de métodos híbridos (combinación de técnicas a nivel de datos y algoritmos) al rendimiento de los modelos de aprendizaje profundo en la clasificación de imágenes radiografías torácicas?	- Analizar el efecto de las funciones de pérdida sensibles a costos en el rendimiento de los modelos de aprendizaje profundo para la clasificación de radiografías torácicas. - Determinar el impacto de la implementación de métodos híbridos en el rendimiento de los modelos de aprendizaje profundo para la clasificación de radiografías torácicas.	clasificación de radiografías torácicas de la clase minoritaria. - **H1.2:** Las funciones de pérdida sensibles a costos mejoran el rendimiento de los modelos de aprendizaje profundo, aumentando el *F1- score* en al menos un 3% en la clasificación de radiografías torácicas desbalanceadas. - **H1.3:** Los métodos híbridos que combinan técnicas a nivel de datos y algoritmos incrementan el rendimiento general de los modelos de aprendizaje profundo en al menos un 3% en términos de precisión balanceada y AUC-ROC para la clasificación de radiografías torácicas desbalanceadas.	- Rendimiento de los modelos de aprendizaje profundo.	**Indicadores dependientes:** - Sensibilidad - Especificidad. - *F1-score*.	**:** Aplicada, Nivel explicativo.

Apéndice B

Código fuente de la implementación

Con el fin de garantizar la transparencia científica y la reproducibilidad de los resultados presentados en el Capítulo IV y Capítulo V, la totalidad del código fuente, los notebooks de entrenamiento y los archivos de configuración se encuentran alojados en el siguiente repositorio público.

Enlace al Repositorio: https://github.com/Estefany-MG/BALANCE-DE-DATOS-PARA-OPTIMIZACION-EN-CLASIFICACION-DE-RADIOGRAFIAS-TORACICAS

B.1. Arquitectura del modelo generativo (WGAN-GP)

Se detalla la clase `Generator` y `Critic` implementada en PyTorch para la generación de imágenes sintéticas de alta resolución (224x224), utilizando *spectral normalization* para estabilizar el entrenamiento *adversarial*.

```
# --- GENERADOR WGAN-GP (PyTorch) ---
class Generator(nn.Module):
    def __init__(self, latent_dim=128, channels=3, img_size=224):
        super().__init__()
        self.init_size = img_size // 32
        self.inp = nn.Sequential(
            nn.Linear(latent_dim, 512 * self.init_size ** 2),
            nn.BatchNorm1d(512 * self.init_size ** 2),
            nn.ReLU(True)
        )
        def sn_conv_t(in_c, out_c): # Spectral Norm Helper
            return nn.utils.spectral_norm(
                nn.ConvTranspose2d(in_c, out_c, 4, stride=2,
padding=1, bias=False))

        self.main = nn.Sequential(
            sn_conv_t(512, 256), nn.BatchNorm2d(256),
nn.ReLU(True),
            sn_conv_t(256, 128), nn.BatchNorm2d(128),
nn.ReLU(True),
            sn_conv_t(128, 64),  nn.BatchNorm2d(64),
nn.ReLU(True),
            sn_conv_t(64, 32),   nn.BatchNorm2d(32),
nn.ReLU(True),
            sn_conv_t(32, channels),
            nn.Tanh() # Salida normalizada [-1, 1]
        )
    def forward(self, z):
        x = self.inp(z)
        x = x.view(x.size(0), 512, self.init_size, self.init_size)
        return self.main(x)
```

```
# --- CRÍTICO (DISCRIMINADOR) ---
class Critic(nn.Module):
    def __init__(self, channels=3, img_size=224):
        super().__init__()
        def sn_conv(in_c, out_c):
            return nn.utils.spectral_norm(
                nn.Conv2d(in_c, out_c, 4, stride=2, padding=1,
bias=False))

        self.main = nn.Sequential(
            sn_conv(channels, 32), nn.LeakyReLU(0.2,
inplace=True),
            sn_conv(32, 64),   nn.InstanceNorm2d(64),
nn.LeakyReLU(0.2, inplace=True),
            sn_conv(64, 128),  nn.InstanceNorm2d(128),
nn.LeakyReLU(0.2, inplace=True),
            sn_conv(128, 256), nn.InstanceNorm2d(256),
nn.LeakyReLU(0.2, inplace=True),
            sn_conv(256, 512), nn.InstanceNorm2d(512),
nn.LeakyReLU(0.2, inplace=True)
        )
        self.out = nn.Sequential(
            nn.Flatten(),
            nn.Linear(512 * (img_size // 32) ** 2, 1) # Salida
lineal (Wasserstein Score)
        )
    def forward(self, x):
        return self.out(self.main(x))
```

B.3. Funciones de pérdida avanzadas (*loss functions*)

Implementación de las funciones de costo utilizadas para mitigar el desbalance en los Escenarios 2, 3, 4 y 7.

```
import tensorflow as tf
import tensorflow.keras.backend as K

# --- FOCAL LOSS (Escenarios 2 y 7) ---
def focal_loss(gamma=2.0, alpha=0.25):
    def focal_loss_fixed(y_true, y_pred):
        y_true = tf.cast(y_true, tf.float32)
        y_pred = tf.cast(y_pred, tf.float32)
        y_pred = tf.clip_by_value(y_pred, K.epsilon(), 1. -
K.epsilon())

        # Cálculo de pt y factor alpha
        pt = tf.where(tf.equal(y_true, 1), y_pred, 1 - y_pred)
        alpha_factor = tf.where(tf.equal(y_true, 1), alpha, 1 -
alpha)

        # Fórmula Focal
        return K.mean(-alpha_factor * K.pow(1 - pt, gamma) *
K.log(pt))
    return focal_loss_fixed

# --- BWCCE LOSS (Escenario 3) ---
# Binary Weighted Categorical Cross-Entropy
```

```
def bwce_loss(class_weights={0: 1.85, 1: 0.69}):
    def bwce_loss_fixed(y_true, y_pred):
        y_true = tf.cast(y_true, tf.float32)
        y_pred = tf.clip_by_value(y_pred, K.epsilon(), 1 -
K.epsilon())
        bce = -(y_true * K.log(y_pred) + (1 - y_true) * K.log(1 -
y_pred))
        weights = y_true * class_weights[1] + (1 - y_true) *
class_weights[0]
        return K.mean(weights * bce)
    return bwce_loss_fixed

# --- LDAM LOSS (Escenario 4) ---
def ldams_loss():
    # (Implementación basada en márgenes de frecuencia de clase)
    # ... [Código LDAM detallado anteriormente] ...
    return ldams_loss_fixed
```

B.4. Estrategias de aumento y balanceo de datos

Código utilizado para preparar los sets de entrenamiento para los Escenarios 5, 6 y 7.

```
# --- ESTRATEGIA ROS (Escenario 5) ---
# Random Over-Sampling: Duplicación simple de la clase minoritaria
from sklearn.utils import resample
X_ros, y_ros = resample(X_minority, y_minority,
                        replace=True, n_samples=n_majority,
random_state=42)
X_train_ros = np.concatenate((X_train, X_ros))

# --- ESTRATEGIA GAN (Escenarios 6 y 7) ---
# Generación de nuevas muestras sintéticas únicas con WGAN-GP
generator.load_state_dict(torch.load('best_generator_stage3.pth'))
noise = torch.randn(num_needed, 128, device=device)
synthetic_images = generator(noise).cpu().detach().numpy()

# Integración al dataset de entrenamiento (Data Augmentation
Sintético)
X_train_gan = np.concatenate((X_train_orig, synthetic_images))
y_train_gan = np.concatenate((y_train_orig, np.zeros(num_needed)))
# Etiqueta 0 o 1
```

B.5. Configuración de los siete escenarios experimentales

Resumen de la implementación y los nombres de archivo de los modelos generados para cada escenario, tal como constan en el repositorio digital.

```
# Rutas base
models_path = '/content/drive/MyDrive/proyecto_completo/models/'

# --- ESCENARIO 1: LÍNEA BASE (Base) ---
# Datos: Desbalanceados | Loss: Binary Crossentropy
```

```
# Archivo generado: ResNet50_neumonia_final.h5
model_s1 = create_ResNet50_model()
model_s1.compile(optimizer=Adam(lr=1e-4),
loss='binary_crossentropy', metrics=['accuracy'])
model_s1.fit(X_train_orig, y_train_orig)

# --- ESCENARIO 2: FOCAL LOSS ---
# Datos: Desbalanceados | Loss: Focal Loss (gamma=2.0)
# Archivo generado: ResNet50_focal_loss_final.h5
model_s2 = create_ResNet50_model()
model_s2.compile(optimizer=Adam(lr=1e-4),
loss=focal_loss(gamma=2.0), metrics=['accuracy'])
model_s2.fit(X_train_orig, y_train_orig)

# --- ESCENARIO 3: BWCCE (Balanced Weighted Cross Entropy) ---
# Datos: Desbalanceados | Loss: Weighted BCE
# Archivo generado: ResNet50_bwcc_final.h5
model_s3 = create_ResNet50_model()
model_s3.compile(optimizer=Adam(lr=1e-4), loss=bwce_loss(),
metrics=['accuracy'])
model_s3.fit(X_train_orig, y_train_orig)

# --- ESCENARIO 4: LDAM LOSS ---
# Datos: Desbalanceados | Loss: Label-Distribution-Aware Margin
# Archivo generado: ResNet50_ldam_final.h5
model_s4 = create_ResNet50_model()
model_s4.compile(optimizer=Adam(lr=1e-4), loss=ldams_loss(),
metrics=['accuracy'])
model_s4.fit(X_train_orig, y_train_orig)

# --- ESCENARIO 5: ROS (Random Over-Sampling) ---
# Datos: Balanceados por Duplicación | Loss: Binary Crossentropy
# Archivo generado: ResNet50_ROS_best.h5
model_s5 = create_ResNet50_model()
model_s5.compile(optimizer=Adam(lr=1e-4),
loss='binary_crossentropy', metrics=['accuracy'])
model_s5.fit(X_train_ros, y_train_ros)

# --- ESCENARIO 6: GAN (Base + GAN Data) ---
# Datos: Balanceados con WGAN-GP | Loss: Binary Crossentropy
# Archivo generado: ResNet50_Base_GAN_FINETUNED_best.h5
# Nota: Incluye una fase de Fine-Tuning tras el entrenamiento
inicial
model_s6 = create_ResNet50_model()
model_s6.compile(optimizer=Adam(lr=1e-5),
loss='binary_crossentropy', metrics=['accuracy'])
model_s6.fit(X_train_gan, y_train_gan)

# --- ESCENARIO 7: HÍBRIDO (Focal Loss + GAN Data) ---
# Datos: Balanceados con WGAN-GP | Loss: Focal Loss
# Archivo generado: ResNet50_Focal_GAN_FINETUNED_best.h5
# Nota: Combina datos sintéticos con la función de pérdida
focalizada
model_s7 = create_ResNet50_model()
model_s7.compile(optimizer=Adam(lr=1e-5),
loss=focal_loss(gamma=2.0), metrics=['accuracy'])
model_s7.fit(X_train_gan, y_train_gan)
```

B.6. Algoritmo de auditoría de robustez (validación *Grad-CAM*)

Script desarrollado para la cuantificación automatizada de la atención del modelo, verificando la coincidencia geométrica entre el mapa de calor y la región anatómica pulmonar (ROI).

```
def ejecutar_auditoria_roi(model, test_images,
layer_name='conv5_block3_out'):
    hits = 0
    total = len(test_images)

    # Definir márgenes de seguridad (ROI: Región Pulmonar Central)
    # Se excluye el 5% del borde para evitar falsos positivos en
etiquetas
    h, w = 224, 224
    margin = int(w * 0.05)
    x_min, x_max = margin, w - margin
    y_min, y_max = margin, h - margin

    for img in test_images:
        # Generar Grad-CAM
        heatmap = make_gradcam_heatmap(img, model, layer_name)
        if np.max(heatmap) == 0: continue

        # Encontrar coordenadas del píxel de máxima activación
(Punto caliente)
        max_idx = np.unravel_index(np.argmax(heatmap),
heatmap.shape)
        cY, cX = max_idx

        # Verificar si la atención cae dentro de la ROI anatómica
        if (x_min < cX < x_max) and (y_min < cY < y_max):
            hits += 1

    accuracy = (hits / total) * 100
    return accuracy
```

Referencias

Ashok, M., Rajeswaran, N., Atheeswaran, A., & VijayaRangan, V. (2024). *Advanced Techniques in Medical Imaging: Computer Vision and Machine Learning Approaches*. Quing Publications. https://books.google.com.pe/books?id=44skEQAAQBAJ

Barrios Roqueme, L. (2022). *Análisis Multidimensional de Indicadores ODS de Educación e Innovación en el Mundo* [Universidad Tecnológica de Bolívar]. https://hdl.handle.net/20.500.12585/11818

Bishop, C. M. (2007). Pattern Recognition and Machine Learning. *Journal of Electronic Imaging*, *16*(4), 049901. https://doi.org/10.1117/1.2819119

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The Balanced Accuracy and Its Posterior Distribution. *2010 20th International Conference on Pattern Recognition*, 3121–3124. https://doi.org/10.1109/ICPR.2010.764

Brownlee, J. (2020). *Imbalanced Classification with Python: Choose Better Metrics, Balance Skewed Classes, and Apply Cost-Sensitive Learning*. Machine Learning Mastery. https://books.google.com.pe/books?hl=es&lr=&id=jaXJDwAAQBAJ&oi=fnd&pg=PP1&ots=CgMC7MUY2S&sig=gdwu2GzxcK6wHtdfa1tEa0Owa0s&redir_esc=y#v=onepage&q&f=false

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, *106*, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

Camargo, N. S. R. (2021). *Clasificación de la insatisfacción de estudiantes utilizando el algoritmo SMOTE en una regresión logística* [Universidad Nacional Mayor de San Marcos]. https://cybertesis.unmsm.edu.pe/bitstream/handle/20.500.12672/19136/Rubina_cn.pdf?sequence=1&isAllowed=y

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. (2019). Class-Balanced Loss Based on Effective Number of Samples. *2019 IEEE/CVF Conference on*

Computer Vision and Pattern Recognition (CVPR), 9260–9269. https://doi.org/10.1109/CVPR.2019.00949

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, *44*(3), 837. https://doi.org/10.2307/2531595

Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, *30*(1), 27–38. https://doi.org/10.1016/j.patrec.2008.08.010

Fotouhi, S., Asadi, S., & Kattan, M. W. (2019). A comprehensive data level analysis for cancer diagnosis on imbalanced data. *Journal of Biomedical Informatics*, *90*, 103089. https://doi.org/10.1016/j.jbi.2018.12.003

Fraiwan, M., Al-Kofahi, N., Ibnian, A., & Hanatleh, O. (2022). Detection of developmental dysplasia of the hip in X-ray images using deep transfer learning. *BMC Medical Informatics and Decision Making*, *22*(1), 216. https://doi.org/10.1186/s12911-022-01957-9

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, *321*, 321–331. https://doi.org/10.1016/j.neucom.2018.09.013

Galán-Cuenca, A., Gallego, A. J., Saval-Calvo, M., & Pertusa, A. (2024). Few-shot learning for COVID-19 Chest X-Ray Classification with Imbalanced Data: An Inter vs. Intra Domain Study. *arXiv preprint arXiv:2401.10129*.

Galdran, A., Carneiro, G., & Ballester, M. A. G. (2021a). Balanced-MixUp for Highly Imbalanced Medical Image Classification. *arXiv preprint arXiv:2109.09850*. https://doi.org/10.1007/978-3-030-87240-3_31

Galdran, A., Carneiro, G., & Ballester, M. A. G. (2021b). *Balanced-MixUp for Highly Imbalanced Medical Image Classification*. https://doi.org/10.1007/978-3-030-87240-3_31

Ganesan, P., Rajaraman, S., Long, R., Ghoraani, B., & Antani, S. (2019). Assessment of Data Augmentation Strategies Toward Performance Improvement of Abnormality Classification in Chest Radiographs. *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 841–844. https://doi.org/10.1109/EMBC.2019.8857516

Gao, Y., Jiang, Y., Peng, Y., Yuan, F., Zhang, X., & Wang, J. (2025). Medical Image Segmentation: A Comprehensive Review of Deep Learning-Based Methods. *Tomography*, *11*(5), 52. https://doi.org/10.3390/tomography11050052

Garro, J. A. C. (2019). *Clasificación de aceptación de campañas para una entidad financiera, usando random forest con datos balanceados y datos no balanceados* [Universidad Ricardo Palma]. https://repositorio.urp.edu.pe/bitstream/handle/20.500.14138/2307/T030_47199993_M%20%20%20CARDENAS%20GARRO%20JOS%C3%89%20ANTONIO.pdf?sequence=1&isAllowed=y

Gaspar, R. J. J. (2023). Utilización de técnicas de aprendizaje profundo para la detección temprana de displasia de cadera en infantes mediante radiografías pélvicas. *Revista de investigación de Sistemas e Informática*, *16*(2), 229–239. https://doi.org/10.15381/risi.v16i2.26710

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks. *Advances in neural information processing systems*, *27*. http://arxiv.org/abs/1406.2661

Haibo He, & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hernández Sampieri, R., Fernández Collado, C., María del Pilar Baptista Lucio, D., & Méndez Valencia Christian Paulina Mendoza Torres, S. (2014). *Metodología de la Investigación* (S. A. D. C. V. McGRAW-HILL /

INTERAMERICANA EDITORES, Cur.; Sexta edición). https://dialnet.unirioja.es/servlet/libro?codigo=775008

Hernández Sampieri, R., & Mendoza Torres, C. P. (2018). *Metodología de la investigación : las rutas cuantitativa, cualitativa y mixta*. https://doi.org/HernándezSampieri,R.,&MendozaTorres,C.P.(2018).Metodologíadelainvestigación:lasrutascuantitativa,cualitativaymixta.https://doi.org/10.22201/fesc.20072236e.2019.10.18.6

HernándezSampieri,R.,FernándezCollado,C.,MaríadelPilarBaptistaLucio,D.,&MéndezValenciaChristianPaulinaMendozaTorres,S.(2014).MetodologíadelaInvestigación(S.A.D.C.V.McGRAW-HILL/INTERAMERICANAEDITORES,Cur.;Sextaedición).https://dialnet.unirioja.es/servlet/libro?codigo=775008

Hoang, V. D., Vo, X. T., & Jo, K. H. (2023). Categorical Weighting Domination for Imbalanced Classification with Skin Cancer in Intelligent Healthcare Systems. *IEEE Access*, *11*, 105170–105181. https://doi.org/10.1109/ACCESS.2023.3319087

Hoens, T. R., & Chawla, N. V. (2013). Imbalanced Datasets: From Sampling to Classifiers. In *Imbalanced Learning* (pp. 43–59). Wiley. https://doi.org/10.1002/9781118646106.ch3

Huarhuachi, J. C. P. (2017). *CLASIFICACIÓN DE FUGA DE CLIENTES EN UNA ENTIDAD FINANCIERA UTILIZANDO EL ALGORITMO SMOTE PARA DATOS DESBALANCEADOS EN UNA REGRESIÓN LOGÍSTICA* [Universidad Nacional Agraria La Molina]. https://repositorio.lamolina.edu.pe/bitstream/handle/20.500.12996/3329/E13-P37-T.pdf?sequence=3&isAllowed=y

Hussain, B. Z., Andleeb, I., Ansari, M. S., Joshi, A. M., & Kanwal, N. (2022). Wasserstein GAN based Chest X-Ray Dataset Augmentation for Deep Learning Models: COVID-19 Detection Use-Case. *2022 44th Annual International Conference of the IEEE Engineering in Medicine y Biology Society (EMBC)*, 2058–2061. https://doi.org/10.1109/EMBC48229.2022.9871519

Iqbal, T., Shaukat, A., Akram, M. U., Muzaffar, A. W., Mustansar, Z., & Byun, Y. C. (2022). A Hybrid VDV Model for Automatic Diagnosis of Pneumothorax Using Class-Imbalanced Chest X-Rays Dataset. *IEEE Access*, *10*, 27670–27683. https://doi.org/10.1109/ACCESS.2022.3157316

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study1. *Intelligent Data Analysis*, *6*(5), 429–449. https://doi.org/10.3233/IDA-2002-6504

Keerthi, R., Kiran, K., Kiran, S. S., & Likitha, P. (2024, october). Advancing Medical Imaging: A Comparative Exploration of Generative Adversarial Networks for Chest X-ray Synthesis. *2024 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI)*. https://doi.org/10.1109/CVMI61877.2024.10782852

Kermany, D., Zhang, K., & Goldbaum, M. (2018). *Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification*. Mendeley Data. https://doi.org/10.17632/rscbjbr9sj.2

Khan, S. H., Hayat, M., Bennamoun, M., Sohel, F. A., & Togneri, R. (2018). Cost-Sensitive Learning of Deep Feature Representations From Imbalanced Data. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(8), 3573–3587. https://doi.org/10.1109/TNNLS.2017.2732482

Kumar, P., Bhatnagar, R., Gaur, K., & Bhatnagar, A. (2021). Classification of Imbalanced Data:Review of Methods and Applications. *IOP Conference Series: Materials Science and Engineering*, *1099*(1), 012077. https://doi.org/10.1088/1757-899x/1099/1/012077

Liang, Z., Huang, J. X., Li, J., & Chan, S. (2020). Enhancing Automated COVID-19 Chest X-ray Diagnosis by Image-to-Image GAN Translation. *Proceedings - 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020*, 1068–1071. https://doi.org/10.1109/BIBM49941.2020.9313466

Liang, Z., Xue, Z., Rajaraman, S., & Antani, S. (2024). COVID-19 PNEUMONIA CHEST X-RAY PATTERN SYNTHESIS BY STABLE DIFFUSION. *2024 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*, 21–24. https://doi.org/10.1109/ssiai59505.2024.10508671

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2017). Focal Loss for Dense Object Detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2999–3007. https://doi.org/10.1109/ICCV.2017.324

López, M., Rubio, Y., & et al., M. S. (2019). Algoritmo de procesamiento de imagenes para señales de transito. *Revista Aristas: Investigación Básica y Aplicada*, *7*, 101–106. https://repositorio.cetys.mx/handle/60000/581

McNemar, Q. (1947). Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, *12*(2), 153–157. https://doi.org/10.1007/BF02295996

Ñaupas, H., Marcelino, P., Valdivia, R., Jesús, D., Palacios, J., Hugo, V., & Delgado, E. R. (2018). *Metodología de la investigación Cuantitativa-Cualitativa y Redacción de la Tesis* (E. de la U, Cur.; 5a.Edición). https://books.google.com.pe/books?id=KzSjDwAAQBAJ&printsec=frontcover&hl=es#v=onepage&q&f=false

Ñaupas-Paitán, H., Mejía-Mejía, E., Novoa-Ramírez, E., & Villagomez-Páucar, A. (2014). *Metodología de la investigación cuantitativa-cualitativa y redacción de la tesis* (4th ed.). Ediciones de la U.

Qin, C., Yao, D., Shi, Y., & Song, Z. (2018). Computer-aided detection in chest radiography based on artificial intelligence: a survey. *BioMedical Engineering OnLine*, *17*(1), 113. https://doi.org/10.1186/s12938-018-0544-y

Qin, X., Bui, F. M., & Nguyen, H. H. (2019). Learning from an Imbalanced and Limited Dataset and an Application to Medical Imaging. *2019 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*, 1–6. https://doi.org/10.1109/PACRIM47961.2019.8985057

Qin, X., Bui, F. M., Nguyen, H. H., & Han, Z. (2022). Learning From Limited and Imbalanced Medical Images With Finer Synthetic Images From GANs. *IEEE Access*, *10*, 91663–91677. https://doi.org/10.1109/ACCESS.2022.3202560

Ramnarine, A. K. (2021). In-Line Image Transformations for Imbalanced, Multiclass Computer Vision Classification of Lung Chest X-Rays. *arXiv preprint arXiv:2104.02238*. https://doi.org/10.1007/978-3-030-89880-9

Reed, J. C. (2019). *Radiologia del Torax. Patrones y Diagnosticos Diferenciales (7a Ed.)* (Amolca, Cur.; 7a Ed.). Amolca.

Sadi, A. A., Chowdhury, L., Jahan, N., Rafi, M. N. S., Chowdhury, R., Khan, F. A., & Mohammed, N. (2022). *LMFLOSS: A Hybrid Loss For Imbalanced Medical Image Classification*. http://arxiv.org/abs/2212.12741

Shen, D., Wu, G., & Suk, H.-I. (2017). Deep Learning in Medical Image Analysis. *Annual Review of Biomedical Engineering*, *19*(1), 221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442

Sindhura, D. N., Pai, R. M., Bhat, S. N., & Pai, M. M. M. (2024). A review of deep learning and Generative Adversarial Networks applications in medical image analysis. *Multimedia Systems*. https://doi.org/10.1007/s00530-024-01349-1

Sinha, S., Ohashi, H., & Nakamura, K. (2021). Class-Wise Difficulty-Balanced Loss for Solving Class-Imbalance. In *Proceedings of the Asian Conference on Computer Vision (ACCV)* (pp. 549–565). https://doi.org/10.1007/978-3-030-69544-6_33

Sun, Y., Kamel, M. S., Wong, A., & Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, *40*, 3358–3378. https://doi.org/10.1016/j.patcog.2007.04.009

Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, *23*(4), 687–719. https://doi.org/10.1142/S0218001409007326

Vargas Romero, J. (s.d.). *Aspectos básicos en radiología de tórax* (3o ed.).

Wang, H., Cui, Z., Chen, Y., Avidan, M., Abdallah, A. Ben, & Kronzer, A. (2018). Predicting Hospital Readmission via Cost-Sensitive Deep Learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *15*(6), 1968–1978. https://doi.org/10.1109/TCBB.2018.2827029

Wang, S., Liu, W., Wu, J., Cao, L., Meng, Q., & Kennedy, P. J. (2016a). Training deep neural networks on imbalanced data sets. *2016 international joint conference on neural networks (IJCNN)*, 4368–4374.

Wang, S., Liu, W., Wu, J., Cao, L., Meng, Q., & Kennedy, P. J. (2016b). Training deep neural networks on imbalanced data sets. *2016 International Joint Conference on Neural Networks (IJCNN)*, 4368–4374. https://doi.org/10.1109/IJCNN.2016.7727770

Yadav*, S. S., & Bhole, G. P. (2020). Learning from Imbalanced Data in Classification. *International Journal of Recent Technology and Engineering (IJRTE)*, *8*(5), 1907–1016. https://doi.org/10.35940/ijrte.E6286.018520

Zeyun Yu, & Bajaj, C. (2004). A fast and adaptive method for image contrast enhancement. *2004 International Conference on Image Processing, 2004. ICIP '04.*, *2*, 1001–1004. https://doi.org/10.1109/ICIP.2004.1419470

Zhou, S. K., Greenspan, H., & Shen, D. (2023). *Deep Learning for Medical Image Analysis*. Academic Press. https://books.google.com.pe/books?id=OYXOEAAAQBAJ

FORMATO DE AUTORIZACIÓN DE PUBLICACIÓN EN EL REPOSITORIO INSTITUCIONAL UNAM

Formulario de autorización para la publicación electrónica en la página web del Repositorio Institucional de la Universidad Nacional de Moquegua, de conformidad con la Ley N° 30035 Ley que regula el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto y Resolución de Comisión Organizadora N° 041-2022-UNAM que aprueba el Reglamento que regula el Repositorio Institucional Digital de la UNAM.

1. DATOS DEL AUTOR

Nombres y Apellidos : Estefany Paola Mamani Gutierrez
DNI/Carnet de Extranjería/Pasaporte N° : 76465200
Correo electrónico : estefanyp.magu@gmail.com
Teléfono/Celular : 927344112

2. GRADO, TITULO O ESPECIALIZACIÓN (Marque con una X)

Bachiller	Título Profesional	Magister	Doctor	Segunda Especialidad
	X			

3. TIPO DE PUBLICACIÓN (Marque con una X)

a) Articulo ( )
b) Tesis de pregrado (X)
c) Tesis de Maestría ( )
d) Tesis de Doctorado ( )
e) Monografía ( )
f) Reporte de caso científico o hallazgos ( )
g) Libro ( )
h) Parte de un libro ( )
i) Revisión ( )
j) Conferencia ( )
k) Presentación ( )
l) Artículo preliminar ( )
m) Articulo antes de ser publicado ( )
n) Artículo de prensa ( )
o) Documentos internos ( )
p) Imagen ( )
q) Documentos técnicos ( )
r) Ponencias presentadas en eventos científicos (Jornada, Congreso, Simposio Conferencias, etc.) ( )
s) Informes de suficiencia profesional ( )
t) Patentes producidas por académicos/as ( )

4. TIPO DE TRABAJO DE INVESTIGACIÓN (Marque con una X)

a) Tesis **(X)**
b) Trabajo de Investigación **( )**
c) Trabajo de suficiencia profesional **( )**
d) Trabajo académico **( )**

5. DATOS DE LA PUBLICACION||||||||||||||||||||||||||||||

a) Título : Técnicas de Balance de Datos para Optimizar el Rendimiento de Modelos de Aprendizaje Profundo en Clasificación de Radiografías Torácicas
b) Campo del conocimiento OCDE (Indicar el código de clasificación OCDE de la obra : https://catalogos.concytec.gob.pe/vocabulario/ocde_ford.html) : 1.02.01
c) Fecha de publicación : Año 2026

6. DATOS DEL ASESOR:

a) Nombres y Apellidos : Juan Carlos Clares Perca
b) DNI/Pasaporte/Cédula : 00450714
c) ORCID : 0000-0001-5146-2879

FIRMA : .....

DOCUMENTO DE IDENTIDAD : 76465200
NOMBRES Y APELLIDOS : ESTEFANY PAOLA MAMANI GUTIERREZ

LICENCIAS

Licencia Estándar

Bajo los siguientes términos, autorizo el depósito de mi trabajo de investigación en el Repositorio Institucional Digital de la UNAM.

Con la autorización de depósito da mi trabajo de investigación, otorgo a la Universidad Nacional de Moquegua, una Licencia No Exclusiva para reproducir. distribuir, comunicar al público, transformar (únicamente mediante su traducción a otros idiomas) y poner a disposición del público mi trabajo de investigación (incluido el resumen), en formato físico o digital, en cualquier medio, conocido o por conocerse, a través de los diversos servicios provistos por la Universidad, creados o por crearse, tales como el Repositorio Institucional Digital de la UNAM, Colección de Tesis, entre otros, en el Perú y en el extranjero, por el tiempo y veces que considere necesarias, y libre de remuneraciones.

En virtud de dicha Licencia, la UNIVERSIDAD NACIONAL DE MOQUEGUA, podrá reproducir mi trabajo de investigación en cualquier tipo de soporte y en más de un ejemplar, sin modificar su contenido, solo con los propósitos de seguridad, respaldo y preservación.

Declaro que el trabajo de investigación es una creación de mi autoría y exclusiva titularidad, o coautoría con titularidad compartida, y me encuentro facultado a conceder la presente licencia y, asimismo, garantizo que dicho trabajo de investigación no infringe derechos de autor de terceras personas.

La UNIVERSIDAD NACIONAL DE MOQUEGUA consignara el nombre del/los autor/es del trabajo de investigación, y no le hará ninguna modificación más de la permitida en la presente licencia.

Autorizo el depósito (marque con una X).

(X) **SI autorizo que se deposite inmediatamente.**

FIRMA : .....



DOCUMENTO DE IDENTIDAD : 76465200
NOMBRES Y APELLIDOS : ESTEFANY PAOLA MAMANI GUTIERREZ
FECHA : 23 DE ENERO DE 2026

ANEXO VIII

DECLARACIÓN JURADA DE ORIGINALIDAD DE TRABAJO DE INVESTIGACIÓN / TESIS / TRABAJO DE SUFICIENCIA PROFESIONAL

Yo, Estefany Paola Mamani Gutierrez,
bachiller de la Escuela Profesional de:
Ingeniería de Sistemas e Informática,
de la Universidad Nacional de Moquegua, identificado(a) con DNI Nº 76465200, autor(a) del Trabajo de Investigación / Tesis / Trabajo de suficiencia profesional denominado:

Técnicas de Balance de Datos para Optimizar el Rendimiento de Modelos de Aprendizaje Profundo en Clasificación de Radiografías Torácicas

Declaro bajo juramento que:

1. El trabajo de investigación es auténtico, siendo resultado de mi trabajo personal, que no se ha copiado, que no se ha utilizado ideas, formulaciones, citas integrales e ilustraciones diversas, sacadas de cualquier tesis, obra, artículo, memoria, etc., (en versión digital o impresa), sin mencionar de forma clara y exacta su origen o autor, tanto en el cuerpo del texto, figuras, cuadros, tablas u otros que tengan derechos de autor. Por tanto, el trabajo de investigación no ha sido plagiado total ni parcialmente.
2. El trabajo de investigación no ha sido autoplagiada, es decir, no ha sido publicada ni presentada anteriormente para obtener algún grado académico previo o título profesional.
3. Los datos presentados en los resultados son reales, no han sido falseados, ni duplicados, ni copiados y por tanto los resultados que se presentan en el trabajo de investigación constituirán aporte a la realidad investigada.

Soy consciente de que, el hecho de no respetar los derechos de autor y hacer plagio, es objeto de sanciones universitarias y/o legales, por lo que asumo cualquier responsabilidad que pudiera derivarse de irregularidades en el trabajo de investigación, así como de los derechos sobre la obra presentada.

Moquegua, 23 de enero de 2026



Huella digital

Nombres y Apellidos: Estefany Paola Mamani Gutierrez
DNI: 76465200

Estefany Paola Mamani Gutierrez

Tesis.pdf

Informe Final de Tesis

TESISTAS EPISI

Universidad Nacional de Moquegua

Detalles del documento

Identificador de la entrega
trn:oid:::1:3417434415

Fecha de entrega
19 nov 2025, 9:40 a.m. GMT-5

Fecha de descarga
19 nov 2025, 9:45 a.m. GMT-5

Nombre del archivo
Tesis.pdf

Tamaño del archivo
7.4 MB

147 páginas

38.178 palabras

223.118 caracteres

21% Similitud general

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

UNAM

Firmado digitalmente por CLARES PERCA Juan Carlos FAU 20449347448 soft
Motivo: Soy el autor del documento
Fecha: 19.11.2025 10:02:15 -05:00

Fuentes principales

18% Fuentes de Internet
12% Publicaciones
12% Trabajos entregados (trabajos del estudiante)

Marcas de integridad

N.° de alertas de integridad para revisión

No se han detectado manipulaciones de texto sospechosas.

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

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

Universidad Nacional de Moquegua

RESOLUCIÓN DE CONSEJO UNIVERSITARIO
N° 098-2026-CU

Moquegua, 13 de febrero de 2026

VISTOS:

El Proveído N° 359-2026-R de fecha 10 de febrero de 2026 emitido por el Rectorado, Informe N° D0063-2026-UNAM-R-SG de fecha 10 de febrero de 2026 emitido por el Secretario General, Informe Técnico N° D0018-2026-SEGE-GT de Grados y Títulos, el Acuerdo del Pleno del Consejo Universitario de Sesión Extraordinaria continuada N° 07-2026-SE-CU-UNAM, de fecha 12 de febrero de 2026, y;

CONSIDERANDO:

Que, conforme al 4to párrafo del artículo 18, de la Constitución Política del Perú, dispone que cada Universidad es autónoma en su régimen normativo, de gobierno, académico, administrativo y económico. Y el artículo 8° de la Ley Nº 30220-Ley Universitaria, establece que la referida autonomía inherente a las Universidades, se ejerce de conformidad con la Constitución y demás normativa aplicable. Concordante con el artículo 1, de la Ley Nº 31520, Ley que restablece la autonomía y la institucionalidad de las universidades peruanas, así como la Ley N° 32141, Ley que modifica la Ley N° 30220-Ley Universitaria, a fin de fortalecer la institucionalidad de las Universidades Públicas licenciadas;

Que, el Artículo 58, de la Ley N° 30220-Ley Universitaria, establece que el Consejo Universitario es el máximo órgano de gestión, dirección y ejecución académica y administrativa de la Universidad. Concordante con el artículo 18° del Estatuto de la Universidad Nacional de Moquegua, aprobado bajo Resolución de Asamblea Universitaria N° 015-2025-AU-UNAM;

Que, el artículo 59 de la Ley Universitaria-Ley N° 30220, establece las atribuciones del Consejo Universitario, y específicamente el numeral 59.9, establece; Conferir los grados académicos y los títulos profesionales aprobados por las Facultades y Escuela de Posgrado, así como otorgar distinciones honoríficas y reconocer y revalidar los estudios, grados y títulos de universidades extranjeras, cuando la universidad está autorizada por la Superintendencia Nacional de Educación Superior Universitaria. Concordante con el numeral 14, del artículo 20° del Estatuto de la Universidad Nacional de Moquegua;

Que, el Sr. Rector es el personero y representante legal de la Universidad Nacional de Moquegua, conforme a lo dispuesto por la Ley Nº 30220-Ley Universitaria, tiene a su cargo la Dirección, Conducción y Gestión del Gobierno Universitario en todos sus ámbitos, concordante con el artículo 24° y numeral 2 del artículo 26° del Estatuto de la Universidad Nacional de Moquegua;

Que, mediante Informe Técnico N° D0018-2026-SEGE-GT de fecha 10 de febrero de 2026, e Informe N° D0063-2026-UNAM-R-SG de fecha 10 de febrero de 2026, el Secretario General de la UNAM, remite al Rectorado los expedientes para el trámite de Título Profesional de Ingeniero de Sistemas e Informática a los Bachilleres: RAMIREZ AGUILAR SERGIO ALAIN, VILCA CONDORI JHONNY ENRIQUE ANDY y MAMANI GUTIERREZ ESTEFANY PAOLA de la Escuela Profesional de Ingeniería de Sistemas e Informática, quienes se encuentran aptos para la expedición del Título Profesional de Ingeniero de Sistemas e Informática, al haber cumplido satisfactoriamente con los requisitos legales y reglamentarios, exigidos en el Reglamento de Grados y Títulos de la Universidad Nacional de Moquegua, conforme lo señala la Dirección de Actividades y Servicios Académicos y la Dirección de la Escuela Profesional de Ingeniería de Sistemas e Informática, razón por la cual, el señor Secretario General incluye dichos expedientes como punto de agenda para ser tratado en Sesión Extraordinaria continuada de Consejo Universitario de fecha 12 de febrero de 2026;

Que, el artículo 44° de la Ley N° 30220, Ley Universitaria señala que las universidades otorgan los grados académicos de Bachiller, Maestro, Doctor y los Títulos Profesionales que correspondan, a nombre de la Nación (...). Asimismo, la Ley N° 31803, modifica el artículo en mención adicionando; La Superintendencia Nacional de Educación Superior Universitaria (SUNEDU) determina los criterios, así como establece y supervisa el procedimiento administrativo para el reconocimiento de grados y títulos otorgados en el extranjero, en el marco de la normativa vigente;

Que, el numeral 45.2 del artículo 45° de la Ley Nº 30220, Ley Universitaria, precisa los requisitos mínimos para la obtención del título profesional, como son: el grado de Bachiller y la aprobación de una tesis o trabajo de suficiencia profesional. Asimismo, establece que el título profesional solo se puede obtener en la universidad en la cual se haya obtenido el grado de bachiller;

Que, mediante Resolución de Comisión Organizadora N° 1153-2021-UNAM, se aprobó el Reglamento de Grados y Títulos de la Universidad Nacional de Moquegua, el mismo que fue modificado a través de las Resoluciones de Comisión Organizadora N° 410; 453 y 890-2022, N° 880-2023-UNAM y N° 081, 484-2024-UNAM el cual establece los requisitos y norma

Domicilio Legal: Prolongación Av., Ancash S/N - Moquegua
RUC: 20449347448 - Teléfono: (053) 670168
Página Web: www.unam.edu.pe

RESOLUCIÓN DE CONSEJO UNIVERSITARIO
N° 098-2026-CU

Moquegua, 13 de febrero de 2026

los procedimientos para la obtención de los grados académicos y títulos profesionales conferidos por esta Casa Superior de Estudios;

Que, bajo el amparo de los considerandos precedentes, el Pleno del Consejo Universitario en Sesión Extraordinaria continuada N° 07-2026-SE-CU-UNAM, de fecha 12 de febrero de 2026, emitió, el siguiente, Acuerdo, POR UNANIMIDAD: CONFERIR, a nombre de la Nación el Título Profesional de Ingeniero de Sistemas e Informática por la modalidad de sustentación de tesis a los Bachilleres: RAMIREZ AGUILAR SERGIO ALAIN, VILCA CONDORI JHONNY ENRIQUE ANDY y MAMANI GUTIERREZ ESTEFANY PAOLA, de la Escuela Profesional de Ingeniería de Sistemas e Informática, de la Facultad de Ingeniería y Arquitectura de la Universidad Nacional de Moquegua;

Que, mediante la Resolución Rectoral Nº 048-2026-R-UNAM de fecha 30 de enero de 2026, se resolvió ENCARGAR, al Dr. Alejandro Manuel Ecos Espino, en su calidad de Vicerrector Académico el Despacho de Rectorado de la Universidad Nacional de Moquegua, para que ejerza las funciones correspondientes en ausencia del Rector, durante los días 02 al 11 de febrero de 2026, por tanto, en aplicación de lo dispuesto por el numeral 5 del artículo 30 del Estatuto de la Universidad Nacional de Moquegua, y conforme a los considerandos del referido acto administrativo, el señor Vicerrector Académico asume el cargo, con todas las prerrogativas de la autoridad;

Que, a través del Memorando N° D0099-2026-UNAM-R de fecha 12 de febrero de 2026, el señor Rector, encarga las funciones de la jefatura de Secretaría General al Abg. Richard Alex Flores Perez, Jefe de la Oficina de Asesoría Jurídica, a partir del 12 de febrero, hasta la designación en el cargo por la autoridad universitaria;

En uso de las facultades y atribuciones conferidas por el Art. 18, de la Constitución Política del Perú, la Ley N° 30220 Ley Universitaria, el Estatuto de la Universidad Nacional de Moquegua, y a lo acordado por el pleno del Consejo Universitario:

SE RESUELVE:

ARTÍCULO 1°: CONFERIR, A NOMBRE DE LA NACIÓN a los Bachilleres

- **RAMIREZ AGUILAR SERGIO ALAIN**
  TESIS : "DESARROLLO DE UNA APLICACIÓN MÓVIL DIDÁCTICA PARA EL APRENDIZAJE DEL LENGUAJE DE SEÑAS PERUANO USANDO DEEP LEARNING", cuyo Asesor Principal fue el MSc. Hugo Euler Tito Chura.
- **VILCA CONDORI JHONNY ENRIQUE ANDY**
  TESIS : "DESARROLLO DE UNA APLICACIÓN MÓVIL DIDÁCTICA PARA EL APRENDIZAJE DEL LENGUAJE DE SEÑAS PERUANO USANDO DEEP LEARNING", cuyo Asesor Principal fue el MSc. Hugo Euler Tito Chura.
- **MAMANI GUTIERREZ ESTEFANY PAOLA**
  TESIS : "TÉCNICAS DE BALANCE DE DATOS PARA OPTIMIZAR EL RENDIMIENTO DE MODELOS DE APRENDIZAJE PROFUNDO EN CLASIFICACIÓN DE RADIOGRAFÍAS TORÁCICAS", cuyo Asesor Principal fue el Mgr. JUAN CARLOS CLARES PERCA, y el Coasesor el Ing. Diego Ismael Mamani Padilla.

El **TÍTULO PROFESIONAL DE INGENIERO DE SISTEMAS E INFORMÁTICA.**

Por la modalidad de SUSTENTACIÓN DE TESIS, de la Escuela Profesional de Ingeniería de Sistemas e Informática, de la Facultad de Ingeniería y Arquitectura de la Universidad Nacional de Moquegua.

ARTÍCULO 2°: ENCARGAR, al Secretario General a través de quien corresponda, para que realice las acciones pertinentes relacionadas con la inscripción correspondiente en el Registro de Grados y Títulos de la Superintendencia Nacional de Educación Superior Universitaria - SUNEDU.

ARTÍCULO 3°: DISPONER, el cumplimiento de la presente Resolución a todas las dependencias de la Universidad Nacional de Moquegua.

ARTÍCULO 4°: DISPONER la publicación de la presente resolución en el Portal Institucional de la Universidad Nacional de Moquegua www.unam.edu.pe

Domicilio Legal: Prolongación Av., Ancash S/N - Moquegua
RUC: 20449347448 - Teléfono: (053) 670168
Página Web: www.unam.edu.pe

RESOLUCIÓN DE CONSEJO UNIVERSITARIO
N° 098-2026-CU

Moquegua, 13 de febrero de 2026

REGÍSTRESE, PUBLÍQUESE Y CÚMPLASE;

ABG. RICHARD ALEX FLORES PEREZ
SECRETARIO GENERAL (e)

DR. HUGO RUBEN MARCA MAQUERA
RECTOR
PRESIDENTE DEL CONSEJO UNIVERSITARIO

DISTRIBUCION: VRA , VRI, FAIN, EPIS, GyT, DASA, OTI Interesados, c.c. /Arch.

Domicilio Legal: Prolongación Av., Ancash S/N - Moquegua
RUC: 20449347448 - Teléfono: (053) 670168
Página Web: www.unam.edu.pe