Universidad de La Habana
Facultad de Matemática y Computación

Aprendizaje automático orientado a la clasificación de cáncer de pielÑ un enfoque basado en EfficientNetB1

Autor:
Deborah Famadas Rodríguez

Tutores:

Dr. Reinaldo Rodríguez Ramos
Dr. Yudivian Almeida

Trabajo de Diploma
presentado en opción al título de
Licenciado en Ciencia de la Computación

Fecha

github.com/deborahfam/Thesis

A todas esas personitas hermosas que me ayudaron a llegar aquí. Este mérito es suyo.

Agradecimientos

Agradecimientos

Opinión del tutor

Opiniones de los tutores

Resumen

El Machine Learning (Aprendizaje Automático) es una disciplina del campo de la Inteligencia Artificial que, a través de algoritmos, dota a los ordenadores de la capacidad de identificar patrones en datos masivos y elaborar predicciones [1]. Las Redes Neuronales Convolucionales (CNN), una especialización de esta disciplina, conocidas por su eficacia en el procesamiento de imágenes, son ideales para detectar características comunes en imágenes dermatológicas.

Este trabajo de diploma propone un algoritmo de machine learning, específicamente, una red neuronal, que permite extraer, a partir de un conjunto de imágenes de cáncer de piel, características para luego realizar una clasificación.

Abstract

Machine Learning is a discipline in the field of Artificial Intelligence that, through algorithms, gives computers the ability to identify patterns in massive data and make predictions. Convolutional Neural Networks (CNN), a specialization of this discipline, known for their efficiency in image processing, are ideal for detecting common features in dermatological images.

This diploma work proposes a machine learning algorithm, specifically, a neural network, which allows extracting, from a set of skin cancer images, features to then perform a classification.

Índice general

Introducción **1**

1. Estado del Arte **5**
1.1. Teledermatología 5
1.2. Técnicas tempranas de clasificación de imágenes 6
1.2.1. Clasificación y reconocimiento de patrones en imágenes médicas 7
1.3. Introducción y desarrollo de machine learning en el campo médico . . 8
1.3.1. Análisis de imágenes médicas 8
1.4. Datasets de cáncer de piel 9
1.5. Redes neuronales convolucionales en la medicina 9
1.6. Relacionado con EfficientNet 12
1.7. Conclusión del estado del arte 12

2. Propuesta de solución **14**
2.1. Preparación y carga de datos 15
2.1.1. Dataset 15
2.1.2. Transformación de datos 16
2.1.3. Modelación y división del conjunto de datos 17
2.1.4. Generadores de datos y preprocesamiento 17
2.2. Desarrollo del modelo y estrategias de optimización 18
2.2.1. Diseño y entrenamiento del modelo 18
2.2.2. Arquitectura del Modelo y Regularización 20
2.2.3. Ajuste dinámico del learning rate 20
2.3. Experimentos 21
2.3.1. Experimento 1: Evaluación de la eficiencia de la división asimétrica de datos en la clasificación de imágenes de cáncer de piel 21
2.3.2. Experimento 2: Análisis de la estratificación de datos en la clasificación de imágenes de cáncer de piel 22

3. Detalles de implementación y experimentos **23**
3.1. Herramientas y Tecnologías 23
3.1.1. Aumento y división de datos 24
3.2. Capas adicionales y regularización 24
3.2.1. Normalización por lotes 25
3.2.2. Capa densa 25
3.2.3. Dropout 25
3.2.4. Capa de salida 25
3.2.5. Optimización 26
3.3. Distribución de datos 26
3.3.1. Experimento 1: Evaluación de la eficiencia de la división asimétrica de datos en la clasificación de imágenes de cáncer de piel 26
3.3.2. Experimento 2: Análisis de la estratificación de datos en la clasificación de imágenes de cáncer de piel 28

4. Resultados **30**
4.1. Consideraciones Finales 36

Conclusiones **38**

Recomendaciones **39**

Bibliografía **40**

Índice de figuras

2.1. Estadísticas del rendimiento de los modelos de EfficientNet 19

4.1. Curva de aprendizaje a lo largo del proceso de entrenamiento del experimento 1. 32
4.2. Curva de aprendizaje a lo largo del proceso de entrenamiento del experimento 2. 33
4.3. Estadísticas de eficacia del modelo al estimar los resultados en el conjunto de pruebas 34
4.4. Estadísticas de eficacia del modelo al estimar los resultados en el conjunto de pruebas 34
4.5. Gráfico de errores por clase en el conjunto de pruebas 34
4.6. Gráfico de errores por clase en el conjunto de pruebas 35

Ejemplos de código

Introducción

Desde su surgimiento en 2006, el aprendizaje profundo se ha establecido como una importante subdisciplina dentro del aprendizaje automático, especialmente en áreas relacionadas con la percepción visual humana. Esta metodología procesa datos a través de múltiples capas que incluyen estructuras complejas y transformaciones no lineales [2]. Actualmente ha logrado avances significativos en áreas como la visión artificial, el reconocimiento de voz, el procesamiento del lenguaje natural, el reconocimiento de audio y la bioinformática [3]. Desde 2013, el aprendizaje profundo se ha reconocido como uno de los diez avances tecnológicos más significativos, dadas sus amplias aplicaciones potenciales en el análisis de datos [4].

El enfoque del aprendizaje profundo abstrae los datos en distintos niveles, lo que permite su aplicación en tareas complejas como la detección de objetos y la clasificación. Su capacidad para reemplazar la extracción manual de características por algoritmos eficientes de aprendizaje, ya sea no supervisada o semi-supervisada, ha revolucionado múltiples áreas [5]. Esta revolución incluye el campo de la atención médica, donde la gestión y análisis de un volumen abrumador de datos médicos es un desafío crucial.

En el ámbito de la atención médica, especialmente en la dermatología, el aprendizaje profundo ha mostrado un potencial extraordinario. La dermatología, que se enfoca en el estudio y tratamiento de enfermedades de la piel, se enfrenta al desafío del cáncer de piel, el tipo de cáncer más frecuente a nivel mundial. La detección precoz de este cáncer es vital [6], y aquí es donde el aprendizaje profundo, con su habilidad para analizar y clasificar imágenes médicas con precisión, juega un rol transformador. La integración de estas tecnologías en la práctica dermatológica no solo mejora la precisión diagnóstica, sino que también promete revolucionar el tratamiento y manejo de diversas afecciones cutáneas [7]."

Motivación

La motivación detrás del uso de algoritmos de *machine learning* (ML) para el diagnóstico del cáncer de piel es significativa y valiosa. A diferencia de otros tipos de

cáncer, el cáncer de piel se forma en la superficie de la piel y suele ser visible. Esto plantea una oportunidad única para la detección temprana y el tratamiento, lo cual es esencial ya que la mayoría de los casos de cáncer de piel son tratables si se detectan a tiempo [6].

Estos algoritmos pueden identificar patrones complejos con una precisión y consistencia mayor que los métodos de diagnóstico humano, reduciendo así la posibilidad de diagnósticos incorrectos debido a la interpretación subjetiva y variable de los expertos [8]. Además, el ML puede procesar grandes volúmenes de datos rápidamente y su capacidad para aprender y adaptarse con el tiempo significa que la detección y clasificación del cáncer de piel puede mejorar continuamente [8].

Aunque ya existen algoritmos eficientes para la clasificación de melanomas, una forma de cáncer de piel, el desarrollo de un modelo capaz de clasificar varios tipos de cáncer de piel y generalizar entre ellos es un objetivo crucial. Esto ampliaría el alcance de las imágenes dermatológicas procesables, mejorando potencialmente la precisión en la detección y el tratamiento de distintas formas de cáncer de piel."

Antecedentes

El desarrollo de las Redes Neuronales Convolucionales (CNN) ha sido fundamental para la identificación de características en imágenes médicas, una base sobre la cual se construye la motivación actual para aplicar ML en la dermatología. Su uso en medicina ha demostrado ser eficaz para capturar patrones específicos en datos de imágenes con alta precisión ([9]). Por ejemplo, Brinker et al. [10] en 2018, analizan 13 artículos sobre la aplicación de CNN en la clasificación de lesiones cutáneas, resaltando su alto rendimiento y la posibilidad de reutilizar CNNs pre-entrenadas. En 2020, Ameri et al. [11] hacen un avance significativo al implementar una CNN profunda para procesar imágenes dermatoscópicas directamente, lo que mejora la eficacia del proceso de clasificación. Más recientemente, en 2022, Shetty et al. [12] logran una precisión del 95,18 % en la clasificación de lesiones cutáneas utilizando CNN, demostrando su superioridad sobre otros algoritmos de ML.

En otras investigaciones enfocadas al diagnóstico mejorado de lesiones pigmentadas, Tajerian et al. (2023) [13] presentan un enfoque metodológico para el diagnóstico de lesiones cutáneas pigmentadas utilizando CNN, logrando una alta puntuación F1 de 0,93. En 2021 Adegun et al. [14] recoge un conjunto de estudios enfocados en el desarrollo de algoritmos con CNN para la detección de cáncer de piel. Estos incluyen enfoques de segmentación y clasificación, arquitecturas de auto-encoder-decoder, y la implementación de redes pre-entrenadas como AlexNet y VGG16.

Las universidades han desempeñado un papel crucial en la investigación y el desarrollo de tecnologías avanzadas en el campo de la medicina, especialmente en la detección y tratamiento del cáncer. Estos centros académicos no solo proporcionan

una base sólida para la investigación teórica, sino que también fomentan la innovación práctica mediante el uso de tecnologías emergentes como el aprendizaje automático y la inteligencia artificial. En particular, nuestra universidad ha contribuido significativamente a este campo.

Las universidades han desempeñado un papel crucial en la investigación y el desarrollo de tecnologías avanzadas en el campo de la medicina, especialmente en la detección y tratamiento del cáncer. Estos centros académicos no solo proporcionan una base sólida para la investigación teórica, sino que también fomentan la innovación práctica mediante el uso de tecnologías emergentes como el aprendizaje automático y la inteligencia artificial. En particular, nuestra universidad ha contribuido significativamente a este campo.

Un claro ejemplo de esta contribución es la tesis de Darien Viera Barredo titulada *Autómata celular estocástico en redes complejas para el estudio de la invasión, migración y metástasis del cancer* [15] proporciona un marco detallado sobre cómo se aborda el estudio del cáncer desde una perspectiva matemática y computacional avanzada. El modelo propuesto utiliza autómatas celulares estocásticos para simular el crecimiento avascular y vascular del tumor. En el se aborda la complejidad del ciclo vital tumoral, destacando la importancia de su comprensión tanto para la investigación del cáncer como para la salud pública. Tradicionalmente, la modelación matemática y computacional se ha centrado en las etapas tempranas del desarrollo tumoral, donde la mortalidad es baja. Sin embargo, este estudio se enfoca en las fases avanzadas, que son críticas para la vida del paciente.

Complementando esta línea de investigación, la tesis reciente de Claudia Olavarrieta Martínez [16] propone un *ensemble* de redes neuronales para clasificar imágenes dermatoscópicas en cuatro categorías: melanoma, carcinoma basocelular, carcinoma espinocelular y otros utilizando la técnica de transferencia de conocimientos en redes como VGG16, ResNet50 y EfficientNet B0.

Problemática

Es entonces notorio que las investigaciones mencionadas utilizan técnicas de machine learning para llevar a cabo el proceso de clasificación. Esto tiene sentido dado que la problemática central en la detección del cáncer de piel radica en la necesidad de mejorar la precisión y rapidez del diagnóstico. Tradicionalmente, esta tarea recae en dermatólogos y el diagnóstico de melanoma, que depende de la evaluación clínica y los hallazgos clásicos en la biopsia de la lesión. Pero la inspección visual puede no ser suficiente para diferenciar lesiones benignas de tumores malignos, y aunque la biopsia de piel es el estándar de oro, es un procedimiento invasivo con limitaciones. Además, la experiencia, el costo y la disponibilidad son desafíos para el uso generalizado de herramientas no invasivas en el diagnóstico clínico [8].

Objetivos

Este trabajo propone como objetivo fundamental el diseño y validación de un modelo predictivo basado en *deep learning* para el diagnóstico del cáncer de piel mediante la clasificación de imágenes dermatológicas. El modelo diseñado y desarrollado se enfoca en clasificar imágenes, priorizando tanto la precisión de los resultados como la capacidad de generalización del modelo. Esto se concibe así con la idea de desarrollar trabajos posteriores que admitan otros datos de entrada.

Entre los objetivos específicos del proyecto se encuentran:

1. Estudiar el estado del arte sobre las técnicas empleadas en el diagnóstico de imágenes dermatológicas y su efectividad.

2. Crear un modelo de *deep learning* que dado un conjunto de imágenes de cáncer de piel clasifique el tipo al que pertenecen.

3. Decidir mediante un algoritmo de machine learning la mejor distribución de datos para entrenamiento del modelo e implementar mejoras potenciales al modelo e hiperparámetros para mejor precisión del mismo.

4. Implementar técnicas de validación para evaluar la precisión del modelo.

Contribuciones

La metodología propuesta podría contribuir al desarrollo de un sistema de clasificación de imágenes de cáncer de piel más preciso, para luego ser utilizado en la práctica clínica en el diagnóstico de cáncer de piel. De esta forma podrían extenderse los algoritmos de clasificación existentes para el análisis de estos datos.

Estructura de la tesis

El contenido de la tesis se organiza de la siguiente forma. En el capítulo 1 se exponen las principales alternativas presentes en la literatura que se han desarrollado para la clasificación de imágenes. En el capítulo 2 presenta el modelo propuesto para la implementación de un sistema de clasificación de imágenes de cáncer de piel. En los capítulos 3 y 4 se describe los algoritmos y técnicas utilizadas, se describen los experimentos realizados y se exponen los resultados obtenidos y se analiza la efectividad de estos. Finalmente, se presentan las conclusiones de la tesis y las recomendaciones para investigaciones futuras.

Capítulo 1

Estado del Arte

En las últimas décadas, los avances en potencia computacional han permitido un progreso significativo en el análisis automatizado de imágenes. Se ha pasado del análisis básico de imágenes digitales a sofisticados algoritmos capaces de identificar patrones sutiles en las imágenes de lesiones cutáneas. Los progresos en el reconocimiento de melanomas a partir de imágenes dermatológicas, han demostrado que los sistemas automatizados pueden lograr un diagnóstico comparable al de los expertos humanos [17].

1.1. Teledermatología

La teledermatología, una rama emergente de la telemedicina, ha revolucionizado el campo de la atención dermatológica. Impulsada por el auge de las tecnologías digitales a finales del siglo XX, esta modalidad se ha consolidado como una herramienta vital en el diagnóstico y manejo de afecciones cutáneas. Desde la década de 2000, la teledermatología ha permitido consultas dermatológicas remotas, ampliando significativamente el alcance de los servicios de salud [18].

La investigación de Whited et al. en 2002 fue pionera en demostrar la efectividad de esta práctica [19]. El estudio resaltó una reducción notable en los tiempos de respuesta, con una mediana de 5 días para consultas teledermatológicas, en contraste con los 28 días de los métodos tradicionales. Además, se enfatizó su utilidad en casos urgentes o semi-urgentes.

Paralelamente, la teledermatopatología ha mostrado su potencial, ofreciendo una fiabilidad comparable a la evaluación histológica tradicional [18]. Uno de los usos importantes de esta técnica fue detectar cáncer de piel. Un estudio Piccolo et al., [20] publicado en 2002, se centró en la concordancia entre los diagnósticos telepatológicos e histopatológicos convencionales, indicando contribuciones significativas en el campo de la teledermatopatología. Por término medio, el 78 % de los telediagnósticos fueron

correctos (intervalo, 60 %-95 %), mientras que el 85 % de los diagnósticos convencionales fueron correctos (intervalo, 60 %-95 %). Se obtuvo una concordancia diagnóstica perfecta en 7 (35 %) de los 20 casos, y sólo se identificó una diferencia significativa en 1 caso.

Este avance en la teledermatología ha sentado las bases para la siguiente etapa en la medicina digital: el desarrollo de algoritmos computarizados que puedan detectar características en las imágenes, imitando el comportamiento humano, para luego a partir de estos obtener resultados.

1.2. Técnicas tempranas de clasificación de imágenes

Cuando se trata de la clasificación de imágenes, es esencial considerar que nuestro sistema visual humano (SVH) primero recibe las ondas electromagnéticas que pertenecen al espectro visible y luego las interpreta en el cerebro. Sin embargo, en el campo de la visión artificial, cuando introducimos una imagen en un ordenador, lo que se interpreta es una matriz de números generalmente en el rango de [0,255] y con tres dimensiones en caso de que sea una imagen a color (RGB). Como resultado, se puede notar una gran brecha entre el significado semántico de la clase asociada a una imagen y los valores de píxeles que la componen, lo que hace que la tarea de clasificación sea compleja para sistemas artificiales. [9]

El reconocimiento de imágenes es una faceta de la inteligencia artificial que posibilita que los sistemas informáticos analicen e interpreten el contenido visual en imágenes. Esto se consigue detectando patrones y características distintivas que luego se emplean para clasificar y etiquetar objetos. Su propósito es automatizar el análisis visual, optimizando tiempo y recursos en diversas aplicaciones [21]. Esta tecnología se sustenta en algoritmos de aprendizaje automático, que entrenan a las máquinas para reconocer patrones visuales. Mediante el uso de bases de datos de imágenes etiquetadas para el entrenamiento, los algoritmos aprenden a identificar objetos y patrones. Una vez completado el entrenamiento, el modelo puede identificar automáticamente estos elementos en nuevas imágenes.

Desde finales del siglo XX, los ingenieros han dedicado esfuerzos significativos al desarrollo de técnicas y algoritmos para la categorización y reconocimiento de eventos a través de datos. Inicialmente, esto se centró en el procesamiento de texto para la clasificación automática de documentos, seguido por el tratamiento de sonidos e imágenes en diversos formatos.

Un hito importante fue la publicación de un artículo sobre reconocimiento de patrones en 1974 en *IEEE Transactions on Automatic Control* [22], evidenciando que ya en la década de 1970 se estaban implementando estas técnicas en el reconocimiento

de patrones. Los trabajos de este enfatizaron avances teóricos y experimentales significativos que impulsaron el progreso en el reconocimiento automático de patrones y el aprendizaje automático.

1.2.1. Clasificación y reconocimiento de patrones en imágenes médicas

En el ámbito de la medicina, se ha observado que la mayoría de los sistemas artificiales de diagnóstico, en un punto, toman decisiones que están cada vez menos relacionadas con la apariencia física de la imagen tal como la vería un radiólogo. En su lugar, estos sistemas se basan en los detalles del patrón matemático de las características físicas individuales de la imagen, que son extraídas por un sistema de visión artificial o un radiólogo, para tomar su decisión final. Estos patrones matemáticos han sido objeto de estudio durante décadas por científicos que han utilizado diversos métodos analíticos, incluyendo las redes neuronales [9].

Para abordar esta brecha entre la representación de la imagen y su significado, se han desarrollado varios tipos de algoritmos de clasificación. Algunos de estos algoritmos se basan en la detección de bordes, como el algoritmo de Canny [23]. Sin embargo, estos algoritmos son robustos cuando se trata de identificar una clase específica, pero si se desea clasificar una clase diferente, es necesario crear un nuevo modelo desde cero [24].

Esta serie de limitaciones restringieron su capacidad para abordar tareas complejas y desafiantes. La escasez de datos adecuados, la falta de recursos computacionales avanzados, arquitecturas simples, dificultades en el entrenamiento, generalización limitada, problemas de gradiente, falta de interpretabilidad y largos tiempos de entrenamiento fueron obstáculos clave en su desarrollo inicial.

Técnicas utilizadas en los inicios para la clasificación de patrones rompieron con alugnas de las barreras de desarrollo: histograma de gradientes orientados (HOG) [25], Scale-Invariant Feature Transform (SIFT) [26], Binary Robust Independent Elementary Features (BRIEF) [27], Color Histograms [28], entre otras. Sin embargo no nos fue suficiente. Luego, con el avance de la tecnología y la llegada del machine learning el aprendizaje profundo (deep learning) también desarrolló métodos de esta índole. En la bibliografía encontramos métodos de bajo nivel como son segmentación de la imagen por niveles grises, bordes o formas, entre otras y métodos de alto nivel como clasificadores basados en redes neuronales, máquinas de soporte vectorial (SVM), árbol de decisiones, entre otros [29].

Se destaca además que las tres técnicas más usadas en la clasificación automática de imágenes son árboles de decisiones, redes neuronales y máquinas de vectores de soporte siendo las redes neuronales una de las más utilizadas en campo del aprendizaje profundo [29].

1.3. Introducción y desarrollo de machine learning en el campo médico

Las primeras aplicaciones de redes neuronales en imágenes médicas se orientaron hacia el análisis y clasificación de dichas imágenes para apoyar en diagnósticos y tratamientos.

1.3.1. Análisis de imágenes médicas

El análisis de imágenes médicas mediante redes neuronales se ha enfocado en campos de la medicina como resonancia magnética, medicina nuclear y radiología, permitiendo la identificación y clasificación de patologías o condiciones específicas. Además, estas tecnologías han encontrado aplicaciones en áreas como la oftalmología, para el diagnóstico de enfermedades oculares a partir de imágenes de retina, y en la cardiología, para la evaluación de imágenes de ecocardiogramas [9].

Resonancia Magnética: En el contexto de la esclerosis múltiple, se han explorado soluciones de segmentación basadas en redes neuronales convolucionales (CNNs) para segmentaciones rápidas y fiables de lesiones y estructuras de materia gris en imágenes de resonancia magnética multimodal [30].

Medicina Nuclear: En este campo un estudio demuestra cómo el aprendizaje profundo puede restaurar la calidad de imagen diagnóstica y mantener la precisión de la cuantificación de SUV para exploraciones PET con un conteo reducido, lo que podría aumentar la seguridad y reducir el costo de las imágenes PET [31].

Radiología: En radiología, la aplicación de la inteligencia artificial en el análisis de imágenes de cáncer, con un enfoque en radiomía y representaciones derivadas del aprendizaje profundo, y su uso para el soporte de decisiones en la gestión del cáncer [32].

Por lo que el análisis de imágenes médicas a través de redes neuronales representa un avance significativo en diversas áreas de la medicina. Las aplicaciones van desde la identificación de patologías en resonancia magnética, medicina nuclear y radiología, hasta el diagnóstico de enfermedades oculares y la evaluación cardíaca. Estas tecnologías no solo mejoran la precisión en la detección y clasificación de condiciones específicas, sino que también optimizan la eficiencia de los procesos diagnósticos y terapéuticos, destacando el papel crucial de la inteligencia artificial en el futuro de la medicina.

1.4. Datasets de cáncer de piel

En el campo de la dermatología, la generación de imágenes clínicas y dermatoscópicas es una práctica común para supervisar los cambios en las condiciones de la piel. Estas imágenes se han vuelto un recurso crucial para el avance de algoritmos de aprendizaje automático, especialmente en el desarrollo de Redes Neuronales Convolucionales (CNN). Existen varios conjuntos de datos accesibles para la investigación en este ámbito.

Das et al. [8] recoge en su estudio un conjunto de los dataset de imágenes dermatologicas más utilizados en algoritmos de clasificación que se exponen a continuación:

Entre los más destacados, se encuentra el archivo ISIC, que agrupa varios datasets de lesiones de piel clínicas y dermatoscópicas, incluidos los Desafíos ISIC, HAM10000 y BCN20000. Otros conjuntos de datos relevantes incluyen el Atlas Interactivo de Dermoscopia con 1000 ejemplos clínicos, la Biblioteca de Imágenes Dermofit con 1300 fotografías de alta resolución, el conjunto de datos PH2 con 200 imágenes dermatoscópicas, y el MED-NODE con 170 fotos clínicas.

El conjunto de datos de Asan, con 17,125 fotos clínicas, y el Hallym, con 125 fotos de casos de carcinoma basocelular, también son significativos. Además, los conjuntos de datos SD-198 y SD-260 ofrecen una amplia gama de imágenes clínicas de diversas enfermedades de la piel. Dermnet NZ y Derm7pt proporcionan colecciones extensas de fotografías clínicas, dermatoscópicas e histológicas, y The Cancer Genome Atlas presenta una de las mayores colecciones de diapositivas de lesiones cutáneas patológicas.

Entre todos estos conjuntos de datos, el HAM10000 se destaca por su amplia utilización en la investigación del cáncer de piel. Este conjunto de datos es particularmente valioso debido a su extensa colección de imágenes de lesiones de piel, que incluye una variedad de tipos de cáncer de piel. Su uso generalizado en la comunidad científica y su relevancia en estudios recientes lo convierten en el dataset ideal para el desarrollo de esta tesis. La riqueza y diversidad de las imágenes en HAM10000 proporcionan una base sólida para entrenar y evaluar algoritmos de machine learning.

1.5. Redes neuronales convolucionales en la medicina

Una red neuronal convolucional está formada por diferentes capas, entre ellas las principales son las capas convolucionales, las capas de max-pooling, y las capas completamente conectadas.La capa convolucional tiene como objetivo realizar la convolución a la imagen de entrada, para extraer sus características. Realizar una convolución a una imagen, consiste en filtrar dicha imagen utilizando una máscara o ventana. La

máscara se va desplazando por toda la imagen, multiplicándose de forma matricial [9].

En el campo de la medicina, estas facilitan la identificación de características relevantes en las imágenes médicas que pudieran ser indicativas de alguna condición médica particular. Se aprovechan de estructuras específicas de datos, como imágenes, para capturar patrones con mayor precisión, reducir la carga computacional y mejorar la generalización y la interpretabilidad.

La evolución en la clasificación de lesiones cutáneas ha estado marcada por el uso innovador de redes neuronales convolucionales (CNNs). Una revisión sistemática en 2018 por Brinker et al. [10] analizó 13 artículos que implementaban CNNs para esta tarea, destacando su alto rendimiento. Los enfoques más comunes involucraron la reutilización de CNNs ya entrenadas con grandes conjuntos de datos, optimizadas posteriormente para la clasificación específica de lesiones cutáneas. Esta metodología, aunque efectiva, enfrentó retos como la dificultad de comparar distintos métodos debido a la variabilidad en los conjuntos de datos.

Posteriormente, las técnicas de aprendizaje automático, y en particular los modelos de aprendizaje profundo, emergieron como herramientas poderosas para el análisis de imágenes médicas. Un proyecto clave fue *A Deep Learning Approach to Skin Cancer Detection in Dermoscopy Images* [11], donde se utilizó un conjunto de 3400 imágenes dermatoscópicas del HAM10000, incluyendo lesiones melanoma y no melanoma. Se implementó una red neuronal convolucional profunda que procesaba imágenes directamente, identificando características valiosas sin necesidad de segmentación previa, lo cual representó un avance significativo en la simplificación y eficacia del proceso de clasificación.

En 2022, el estudio *Skin lesion classification of dermoscopic images using machine learning and convolutional neural network* [12], publicado en Nature, utilizó un subconjunto del HAM10000 para clasificar lesiones cutáneas mediante aprendizaje automático y CNN. Este enfoque ofreció resultados prometedores en la distinción entre lesiones malignas y benignas, logrando una precisión del 95,18 % con el modelo CNN. La comparación con otros algoritmos de aprendizaje automático resaltó la superioridad de los modelos CNN en términos de precisión.

Tajerian et al. [13] presentó un enfoque metodológico para mejorar el diagnóstico de la lesiones cutáneas pigmentadas utilizando imágenes dermatoscópicas del conjunto de datos HAM10000. Este conjunto de datos se utilizó para analizar lesiones cutáneas pigmentadas. El modelo obtiene los mejores resultados en la detección de lesiones de nevos melanocíticos,con una puntuación F1 de 0,93.

Adegun et al. [14] recoge un conjunto de artículos que contribuyeron al desarrollo de algoritmos además de los anteriormente mencionados:

1. Majtner et al. usaron técnicas de CNN para la extracción de características y preprocesamiento, utilizando el conjunto de datos ISIC con 900 muestras de

entrenamiento y 379 de prueba, clasificadas en benignas y malignas.

2. Vipin et al. implementaron un sistema de dos etapas, segmentación y clasificación, utilizando un conjunto de datos ISIC de 13,000 imágenes, reducido a 7,353 tras eliminar imágenes no utilizables.

3. Nasr-Esfahani et al. desarrollaron su propia CNN para preprocesar, extraer características y clasificar imágenes, con un conjunto de 170 imágenes no dermatoscópicas del UMCG, aumentado a 6,120 imágenes.

4. Attia et al. usaron una CNN completamente conectada con arquitecturas de autoencoder-decoder, alcanzando una precisión del 98 % y una especificidad del 94 %.

5. Mukherjee et al. desarrollaron una arquitectura CNN para la detección de lesiones malignas, logrando precisiones de 90.14 % y 90.58 % en los conjuntos de datos MEDNODE y Dermofit.

6. Sanketh et al. propusieron una CNN para la detección temprana de cáncer de piel, obteniendo un resultado óptimo del 98 %.

7. Rahi et al. propusieron un modelo CNN con varias capas convolucionales y de agrupación máxima, logrando una precisión del 84.76 % y una especificidad del 78.81 %.

8. Gulati et al. emplearon redes preentrenadas como AlexNet y VGG16, obteniendo mejores resultados con VGG16 en modo de aprendizaje transferido.

9. Daghrir et al. combinaron una CNN con técnicas de aprendizaje automático clásicas, alcanzando una precisión individual del 85.5 % con CNN.

10. Acosta et al. incorporaron técnicas de CNN basadas en máscaras y regiones con una estructura ResNet152 preentrenada, logrando una precisión del 90.4 % y una especificidad del 92.5 %.

El avance en la detección de cáncer mediante el uso de la inteligencia artificial (IA) y el aprendizaje automático (ML) ha sido significativo en las últimas décadas, el mismo, ha demostrado un rendimiento excepcional en tareas de reconocimiento de imágenes, que es fundamental en la detección del cáncer de piel. Se llevó a cabo una revisión exhaustiva para evaluar el impacto de las técnicas de aprendizaje profundo en la detección precoz, en la que se analizaron diversos resultados de investigación y se presentaron mediante herramientas, gráficos, tablas y marcos para comprender mejor las técnicas predominantes en este campo [33].

1.6. Relacionado con EfficientNet

Un estudio innovador es el de Ali et al. [34]. Este desarrolló una cadena de procesamiento de imágenes previo al entrenamiento, que incluía la eliminación de cabellos en las imágenes, el aumento de datos y el redimensionamiento de las imágenes para cumplir con los requisitos de cada modelo de CNN. Utilizando transferencia de aprendizaje con pesos pre-entrenados de ImageNet y ajuste fino de las redes, se entrenaron variantes de EfficientNet (B0-B7) en el conjunto de datos HAM10000. El modelo más exitoso, EfficientNet B4, alcanzó una puntuación F1 y una precisión Top-1 del 87 % y 87.91 %, respectivamente, destacando que una complejidad intermedia del modelo puede ser óptima para este tipo de tareas. El modelo en cuestión relacionado con nuestro enfoque (EfficientNetB1) obtuvo una precisión del 86.5 % y una precisión Top-1 del 86.5 %.

Por otro lado Papiththira et al. [35] se centró específicamente en la detección de melanoma utilizando un enfoque basado en transferencia de aprendizaje profundo sin necesidad de preprocesamiento de imágenes. Este método se apoyó en el modelo EfficientNet pre-entrenado, complementado con un módulo de atención de canales para resaltar características específicas del melanoma en la clasificación. Evaluado en los conjuntos de datos UMGC y HAM10000, que incluyen imágenes clínicas y dermoscópicas, el enfoque propuesto superó los métodos del estado del arte con una precisión de clasificación del 84.12 % y 96.32 %, respectivamente.

1.7. Conclusión del estado del arte

Los estudios revisados reflejan un claro progreso en la aplicación de CNNs y otras técnicas de aprendizaje profundo, no solo en términos de eficacia sino también en la reducción de tiempos de espera y mejora en el acceso a diagnósticos. Es notable cómo estos avances han permitido el desarrollo de sistemas de clasificación más robustos y precisos, capaces de distinguir entre lesiones cutáneas malignas y benignas con altos niveles de precisión.

En el contexto de estos avances, esta investigación aporta un valor distintivo en varios aspectos. Primero, la implementación de una arquitectura EfficientNetB1 para la clasificación de cáncer de piel, una técnica relativamente reciente y menos explorada en la literatura comparada con modelos más establecidos como AlexNet o InceptionV3. Esta elección representa un intento de equilibrar la eficiencia y precisión en un campo donde la carga computacional y la exactitud son críticas.

Además, el enfoque de esta investigación hacia el tratamiento de conjuntos de datos desequilibrados aborda una limitación significativa que ha sido un desafío persistente en estudios anteriores. Al proponer y validar métodos que manejan de manera efectiva la desproporción en las categorías de datos, se está contribuyendo a un área de

necesidad crítica, mejorando potencialmente la capacidad del modelo para generalizar y funcionar eficazmente en escenarios clínicos reales.

Capítulo 2

Propuesta de solución

Este capítulo introduce la propuesta desarrollada para enfrentar el reto de clasificar de manera automática el cáncer de piel, empleando técnicas de aprendizaje profundo. La solución se centra en la utilización de una red neuronal profunda pre-entrenada, combinada con un algoritmo propio para el ajuste de precisión.

En el núcleo de nuestra estrategia se encuentra el uso de una red neuronal convolucional pre-entrenada llamada *EfficientNetB1*. Esta red, desarrollada a partir de extensos conjuntos de datos y experiencias previas, ofrece una base sólida y rica en características para nuestro modelo. Al aprovechar este pre-entrenamiento, podemos acelerar significativamente el proceso de aprendizaje del modelo, al tiempo que aumentamos su capacidad para generalizar y reconocer patrones complejos en las imágenes dermatoscópicas.

Para complementar nuestro enfoque metodológico, seleccionamos herramientas tecnológicas avanzadas. Estas herramientas están diseñadas para optimizar el rendimiento del modelo, mejorar la precisión de la clasificación y garantizar una implementación efectiva. Entre ellas, destaca una capa de normalización, una capa densa, una de regularización, una *dropout* y una de salida (capa densa con activación *softmax*). Estas técnicas le permiten al modelo afinar su capacidad de identificar con precisión las diferentes categorías de lesiones cutáneas.

Para la optimización del algoritmo además se llevaron a cabo una serie de experimentos de los que se detallan al final de este capítulo los 2 más relevantes. Estos experimentos se realizaron con el objetivo de encontrar la mejor configuración de datos para el modelo, que permita obtener la mayor precisión posible. Para esto se utilizaron dos técnicas de normalización de datos: division asimétrica con utilización de pesos por clases y estratificación de datos. Además se generaron distintas distribuciones de datos para los diferentes acercamientos.

Con la intensión de evaluar la efectividad y precisión del modelo se utilizo el conjunto de datos mencionado en capítulos anteriores HAM1000. Este, con más de

10000 imágenes, representa una variedad de condiciones de la piel, lo que lo convierte en un recurso valioso para entrenar y evaluar algoritmos de detección de cáncer de piel. Para esto se importan, modelan y dividen los datos para asegurar un aprendizaje efectivo

La sección que sigue detalla la metodología empleada en la preparación y carga de los datos. Esta fase es esencial, ya que la calidad y el tratamiento de los datos tienen un impacto directo en la eficacia del modelo.

2.1. Preparación y carga de datos

En la presente sección, se describe cómo se seleccionó y procesó el conjunto HAM10000. Se detallan técnicas aplicadas para optimizar el rendimiento del algoritmo incluyendo el procesamiento de las imágenes y la conversión de los metadatos a un formato categórico adecuado para su análisis.

Posteriormente, se explica la modelación y división del conjunto de datos, utilizando herramientas como Pandas para estructurar los datos y dividirlos en conjuntos de entrenamiento, validación y prueba. Finalmente, se aborda el desafío del desequilibrio en la representación de las clases dentro del *dataset*. Se detallan las estrategias implementadas para equilibrar el conjunto de datos, garantizando así que el modelo aprenda de manera efectiva a identificar una variedad de lesiones cutáneas sin sesgos hacia las condiciones más comunes.

2.1.1. Dataset

El conjunto de datos HAM10000, acrónimo de *Human Against Machine with 10000 training images* (Humano Contra Máquina con 10000 imágenes de entrenamiento), se presenta como una solución al problema de la falta de diversidad y tamaño reducido en los conjuntos de datos disponibles para el diagnóstico automatizado de lesiones cutáneas pigmentadas. Este conjunto de datos es notable por su extenso alcance y diversidad, abarcando una amplia gama de lesiones cutáneas pigmentadas comunes [36].

HAM10000

Las 10015 imágenes dermatoscópicas del conjunto de datos HAM10000 se recopilaron a lo largo de 20 años desde dos ubicaciones diferentes: el Departamento de Dermatología de la Universidad Médica de Viena, Austria, y la práctica de cáncer de piel de Cliff Rosendahl en Queensland, Australia [36]. En comparación con otros conjuntos de datos, HAM10000 ofrece un conjunto más diverso y completo de imágenes

dermatoscópicas para la investigación del aprendizaje automático. Las imágenes y los metadatos del HAM10000 tienen la siguiente distribución.

Tabla 2.1: Distribución de imágenes por categoría diagnóstica

Categoría Diagnóstica	**Número de Imágenes**	**Porcentaje**
Melanocytic nevi	6705	66.95 %
Melanoma	1113	11.11 %
Benign keratosis-like lesions	1099	10.97 %
Basal cell carcinoma	514	5.13 %
Actinic keratoses	327	3.27 %
Vascular lesions	142	1.42 %
Dermatofibroma	115	1.15 %

Las imágenes almacenadas, originalmente como diapositivas, fueron digitalizadas usando un escáner *Nikon Coolscan 5000 ED*. Posteriormente, se ajustaron manualmente para centrar las lesiones y se aplicaron correcciones al histograma para mejorar el contraste visual y la reproducción del color. Para separar eficientemente las imágenes dermatoscópicas de otros tipos de imágenes (como primeros planos y vistas generales), se utilizó un método automatizado que clasificaba más de 30,000 imágenes. Se empleó una arquitectura *InceptionV3*, entrenada con un conjunto de imágenes etiquetadas manualmente, para categorizar las imágenes. Las imágenes mal clasificadas por este método fueron revisadas y corregidas manualmente. Se realizó una revisión manual final para excluir imágenes con ciertos atributos no deseados, como contenido potencialmente identificable, imágenes fuera de enfoque o con artefactos perturbadores como prendas, y lesiones completamente no pigmentadas. Las imágenes restantes fueron revisadas para asegurar una reproducción de color y luminosidad adecuadas, aplicando correcciones manuales si era necesario [36].

Los datos utilizados como medio de aprendizaje para este proyecto son imágenes y metadatos. El dataset HAM10000 contiene imágenes y un archivo de metadatos que contienen información relacionada con cada imagen en formato *one hot encoding* [37].

2.1.2. Transformación de datos

Los metadatos asociados a la clasificación, etiquetados con el método mencionado (*one hot encoding*), fueron convertidos a un formato *categórico* [38] para su procesamiento. En este proceso a cada elemento se le asignó una etiqueta basada en las categoría de la lesion cutánea, como 'MEL' (Melanoma), 'NV' (Nevus Melanocítico), entre otras. Se elimina del *dataframe* además cualquier columna innecesaria, dejan-

do solo las etiquetas y nombres de imágenes relevantes. De tal forma que los datos quedan distribuidos por clase.

Tabla 2.2: Datos transformados a formato categórico

index	**Imágenes**	**Etiqueta**
0	$ISIC\_0024306.jpg$	NV
1	$ISIC\_0024307.jpg$	NV
2	$ISIC\_0024308.jpg$	NV
3	$ISIC\_0024309.jpg$	NV
4	$ISIC\_0024310.jpg$	MEL

2.1.3. Modelación y división del conjunto de datos

Los datos se modelan a partir de un *dataframe* de Pandas [39]. Estos son divididos en 3 conjuntos: Entrenamiento, Validación y Prueba, utilizando un enfoque simple de division de datos en porcentaje. Estas divisiones son necesarias para que el modelo pueda aprender y ser evaluado correctamente.

Inicialmente, el conjunto de datos, fue dividido en dos subconjuntos: *train*, que se destinó para el entrenamiento, y *dummy*, que fue utilizado como una combinación temporal para los conjuntos de validación y prueba. Luego el segundo conjunto fue separado en *valid*, destinado a la validación y *test*, utilizado para las pruebas.

2.1.4. Generadores de datos y preprocesamiento

Uno de los principales problemas del dataset, como se puede observar en la tabla 2.1, es el desequilibrio en la representación de las clases, un problema habitual en los conjuntos de datos médicos en los que algunas enfermedades son más raras que otras. Para solucionar este problema, el conjunto de datos se equilibra limitando el número máximo de muestras por clase. Esto garantiza que el modelo no esté sesgado hacia las clases más comunes y pueda generalizar mejor entre varios tipos de lesiones cutáneas.

En cada experimento, la cantidad de muestras por clase se ajustó de manera selectiva. Por ejemplo, en el experimento 1 se utilizaron 300 muestras por clase, mientras que en el experimento 2 se incrementó a 500 muestras.

Incluso en el subconjunto de 300 muestras, algunas clases contaban con menos de 300 ejemplos, lo que llevó a la implementación de la técnica de *class weighting* [40]. Con este método se asignó pesos diferenciados a cada clase durante el entrenamiento del modelo, reforzando así la señal de aprendizaje para las clases menos representadas.

Además, a los conjuntos segmentados en entrenamiento, validación y prueba, se les aplicó *data augmentation*. Esta técnica aplicó varias transformaciones a las imágenes,

como rotación, desplazamiento y zoom, durante su carga, generando lotes de imágenes optimizados para el entrenamiento y evaluación del modelo [41].

2.2. Desarrollo del modelo y estrategias de optimización

Para abordar efectivamente el desafío de la detección de cáncer de piel, se seleccionó la arquitectura *EfficientNetB1* [42] como base del modelo. Esta red neuronal convolucional, parte de la familia EfficientNet, se caracteriza por su alta eficiencia y precisión. Utilizando un modelo pre-entrenado, se aprovecharon los pesos derivados de conjuntos de datos extensos, lo que facilitó la adaptación del modelo a nuestro conjunto de datos específico (HAM10000) [43].

2.2.1. Diseño y entrenamiento del modelo

El modelo *EfficientNetB1* se carga pre-entrenado con pesos de *ImageNet* [44], una amplia base de datos de imágenes ampliamente utilizada para entrenamiento y *benchmarking* en visión por computadora. Luego a este se le omitió la capa superior. Al omitir la capa superior del modelo, se permite la incorporación y personalización de capas adicionales que se mencionan más adelante en el capítulo. Además, la salida del modelo base se somete a una serie de transformaciones, incluyendo normalización por lotes, capas densas con regularizaciones, y técnicas de *Dropout* para prevenir el sobre-ajuste, culminando en una capa de salida optimizada para la clasificación.

Arquitectura del modelo EfficientNet

La familia de arquitecturas EfficientNet, desarrollada en [43], surgió con el objetivo de hallar un método adecuado para escalar las CNNs de manera que mejoraran tanto en precisión (i.e., rendimiento del modelo) como en eficiencia (es decir, en términos de parámetros del modelo y FLOPS).

Inicialmente, el enfoque se centraba en aumentar la profundidad o el ancho de la red. Sin embargo, este método de escalado único tenía limitaciones y no se comprendía completamente. La investigación de EfficientNet reveló que un escalado uniforme en todas las dimensiones de la red (profundidad, ancho y resolución) usando un conjunto de coeficientes de escalado fijos podría lograr un mejor equilibrio y rendimiento. Este enfoque innovador, denominado escalado compuesto, aborda el problema del escalado en ConvNets de manera integral. Los resultados empíricos mostraron que escalar cualquier dimensión de la red mejora la precisión, pero el beneficio disminuye en modelos más grandes. Por lo tanto, se propuso un método de escalado que coordina y equilibra estas dimensiones, en lugar de escalar una sola dimensión a la vez.

EfficientNet comenzó con la creación de un modelo base, EfficientNet-B0, utilizando una búsqueda de arquitectura neural multiobjetivo que optimizaba tanto la precisión como los FLOPS. Esta red base se caracterizaba por su bloque principal, el cuello de botella móvil invertido MBConv, y la optimización de compresión y excitación. El proceso de escalado de EfficientNet se realizó en dos etapas principales. Primero, se fijaron ciertos parámetros y luego se escaló la red base para obtener variantes más grandes, desde EfficientNet-B1 hasta B7, utilizando el método de escalado compuesto. Mientras que la arquitectura EfficientNet B0 tiene 5,3 millones de parámetros y acepta imágenes de entrada de $224x224$, EfficientNet B7 cuenta con 66 millones de parámetros y acepta imágenes de $600x600$ [45].

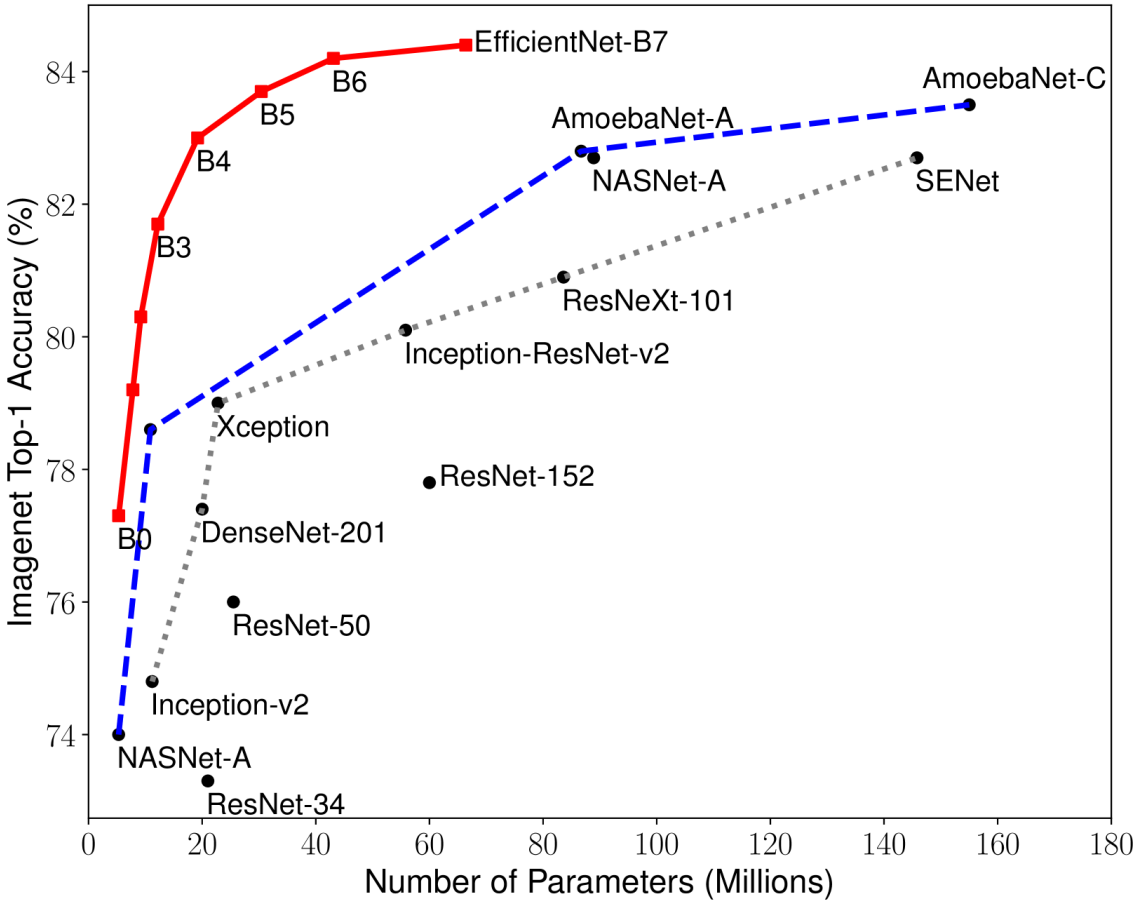

Figura 2.1: Estadísticas del rendimiento de los modelos de EfficientNet

Los experimentos demostraron que el método de escalado compuesto mejoraba la precisión en modelos ya existentes como MobileNets y ResNet. Los modelos Efficient-

Net entrenados en ImageNet mostraron una precisión y eficiencia significativamente mayores en comparación con otros ConvNets, utilizando una cantidad mucho menor de parámetros y FLOPS. Especialmente, EfficientNet-B7 logró una precisión del 84.3% en top-1 en ImageNet, superando modelos anteriores y siendo considerablemente más pequeño y rápido.

En relación con el dataset HAM1000 [46], la implementación del EfficientNetB1 puede ser particularmente beneficiosa para el análisis de datos. EfficientNetB1, entre las variantes de la serie EfficientNet, se encuentra en un punto medio en términos de complejidad y tamaño, ofreciendo un equilibrio entre precisión y eficiencia computacional. Dado que el HAM1000 es un conjunto de datos de imágenes dermatoscópicas que requiere una alta precisión en la identificación y clasificación de lesiones cutáneas, la utilización de EfficientNetB1 podría proporcionar una precisión y eficiencia aceptable en términos de recursos computacionales.

2.2.2. Arquitectura del Modelo y Regularización

Para fortalecer la arquitectura del modelo, se añadieron capas adicionales, incluyendo Dropout y regularizadores L1 y L2, esenciales para combatir el sobre-ajuste. Una capa densa personalizada fue incorporada para facilitar la clasificación precisa de múltiples tipos de tumores. La compilación del modelo se realizó con un enfoque en la clasificación multi-clase, utilizando la pérdida de entropía cruzada categórica (*categorical crossentropy*) y un optimizador *Adamax*.

2.2.3. Ajuste dinámico del learning rate

Un elemento innovador del entrenamiento fue el uso de un callback personalizado para el ajuste dinámico del learning rate [47]. Se implementa un mecanismo de ajuste de la tasa de aprendizaje (LRA) personalizado de Keras. Este, ajusta dinámicamente la tasa de aprendizaje durante el entrenamiento basado en la precisión y la pérdida de validación, con el objetivo de mejorar la eficiencia del entrenamiento y alcanzar una mejor convergencia. Los componentes clave del LRA son:

1. Modelo: La red neuronal sobre la que se aplicará el callback.

2. Paciencia: El número de épocas que se esperará sin mejora en la métrica de rendimiento antes de realizar un ajuste.

3. Umbral (Threshold): Un valor límite que define cuándo se considera que ha habido una mejora significativa en el rendimiento.

4. Factor de Ajuste: La magnitud por la cual se modificará el learning rate en caso de no observarse mejoras.

5. Dwell: Una opción que permite al modelo volver a un estado de pesos anterior si no hay mejoras tras el ajuste del learning rate.

Durante el entrenamiento, el *callback* monitorea constantemente el rendimiento del modelo en términos de precisión y pérdida de validación. Al final de cada época, realiza las siguientes operaciones:

1. Evaluación de Métricas: Se revisan la precisión del entrenamiento y la pérdida de validación para determinar si se ha alcanzado o superado el umbral establecido.

2. Decisión de Ajuste: Basándose en la paciencia y las métricas evaluadas, se decide si se ajustará el learning rate. Si las métricas no han mejorado durante el número de épocas definidas por la paciencia, se procede al ajuste.

3. Aplicación del Ajuste: Si se requiere un ajuste, el learning rate se multiplica por el factor de ajuste. Este cambio tiene como objetivo reaccionar ante el estancamiento del aprendizaje, estimulando al modelo para explorar nuevas áreas del espacio de parámetros.

4. Implementación de Dwell: En caso de no observarse mejora incluso después del ajuste, y si la opción 'dwell' está activada, el modelo puede revertir a un estado de pesos anterior, evitando así el estancamiento en mínimos locales.

5. Reporte de Progreso: El callback proporciona información valiosa sobre el progreso del entrenamiento, incluyendo la tasa de aprendizaje actual y la próxima, y si el enfoque está en la precisión o la pérdida de validación.

2.3. Experimentos

Como se había expresado al inicio del capítulo se llevaron a cabo una serie de experimentos de los cuales analizaremos los siguientes para la distribución y normalización de datos. La metodología general para ambos experimentos coincide, en ambos se utiliza una red convolucional, EfficientNetB1 y capas adicionales.

2.3.1. Experimento 1: Evaluación de la eficiencia de la división asimétrica de datos en la clasificación de imágenes de cáncer de piel

Este experimento se centra en una división de datos altamente asimétrica, con un enfoque predominante en el conjunto de entrenamiento. La técnica de *dummy split* se emplea para mantener proporciones consistentes entre los conjuntos de validación y

prueba. A esta división se le aplica luego un determinado peso para que el algoritmo preste especial atención a las clases menos representadas.

2.3.2. Experimento 2: Análisis de la estratificación de datos en la clasificación de imágenes de cáncer de piel

Este experimento explora la estratificación de datos para mantener una distribución uniforme de etiquetas en cada conjunto de datos. La proporción de los conjuntos de datos es más equilibrada en comparación con el Experimento 1, lo que ofrece *insights* sobre la importancia de la distribución equitativa de datos en el entrenamiento y evaluación de modelos.

Capítulo 3

Detalles de implementación y experimentos

Este capítulo presenta los métodos, técnicas, detalles e hiperparámetros utilizados en la implementación general del modelo y especifica de cada experimento.

3.1. Herramientas y Tecnologías

La implementación del modelo se llevó a cabo utilizando las siguientes herramientas:

- **Lenguajes de Programación:** Python: utilizado por su rica biblioteca de paquetes de aprendizaje automático, incluyendo TensorFlow y Keras para la construcción y entrenamiento del modelo.
- **Frameworks de Aprendizaje Profundo:** TensorFlow y Keras proporcionan las funcionalidades necesarias para diseñar, entrenar y validar modelos de aprendizaje profundo con alta eficiencia y flexibilidad.
- **Técnicas de Preprocesamiento:** Herramientas para la normalización de imágenes, el aumento de datos, y el balanceo de clases serán utilizadas para preparar el dataset para el entrenamiento del modelo.
- **Optimización y Regularización:** Se integrarán técnicas como el ajuste dinámico del learning rate y la regularización L1 y L2 para optimizar el rendimiento del modelo y prevenir el sobre-ajuste.
- **Hardware y Recursos Computacionales:** Se utilizó Google Collab para acelerar el proceso de entrenamiento del modelo, permitiendo la experimentación con diferentes híper-parámetros y arquitecturas de forma eficiente.

Implementación del modelo

Hiperparámetros y configuración

1. Modelo Base: EfficientNetB1 con pesos de ImageNet.
2. Tasa de Aprendizaje Inicial: 0,001.
3. Regularizadores: L2 con $\lambda = 0{,}016$ y L1 con $\lambda = 0{,}006$.
4. Tasa de Dropout: 45 %.
5. Epochs: 40
6. Callback: Incluye el LRA con parámetros específicos de paciencia, umbral, factor de reducción, y control de *dwell*.

3.1.1. Aumento y división de datos

Para la carga y procesamiento de imágenes se utilizó la clase *ImageDataGenerator* de Keras [48]. Esta clase es una parte integral de la biblioteca Keras y proporciona una forma eficiente de manipular imágenes para tareas de aprendizaje automático. Su principal función es facilitar la creación de lotes de imágenes que se utilizan durante el entrenamiento y la evaluación de modelos de Machine Learning, especialmente en el contexto de redes neuronales.

Las transformaciones aplicadas fueron las siguientes

Tabla 3.1: Parámetros de aumento de datos

Parámetro	Descripción	Valor
rotation range	Rango de rotación	20 grados
width shift range	Rango de desplazamiento horizontal	20 %
height shift range	Rango de desplazamiento vertical	20 %
shear range	Rango de corte	20 %
zoom range	Rango de zoom	20 %
horizontal flip	Activación de volteo horizontal	verdadero
fill mode	Modo de relleno para manejar los píxeles faltantes	nearest

3.2. Capas adicionales y regularización

Para ajustar el modelo de EfficientNetB1 a nuestras necesidades, se añaden capas adicionales:

3.2.1. Normalización por lotes

Esta capa se define con un *eje de normalización* establecido en -1, lo que indica que la normalización se aplica a lo largo del último eje en el tensor de entrada. Además, se configura un *momentum* de 0,99 y un valor de *epsilon* de 0,001. El alto valor de momentum ayuda a mantener la estabilidad de las medias y varianzas móviles a lo largo del entrenamiento, mientras que el pequeño valor de epsilon evita divisiones por cero, asegurando así cálculos numéricos estables [49].

3.2.2. Capa densa

Se integra una capa densa con 256 neuronas, que juega un papel clave en la síntesis de las características aprendidas por el modelo. Se utiliza un regularizador *L2* con un *lambda* de 0,016 para los pesos de la capa, lo que ayuda a penalizar y controlar el tamaño de los pesos, reduciendo así el riesgo de sobre-ajuste. Además, tanto el regularizador de actividad como el regularizador de bias se configuran con un regularizador *L1* con un *lambda* de 0,006. Este enfoque impone una penalización en los pesos y los sesgos, promoviendo un modelo más simple y disperso. La función de activación utilizada es *ReLU* , conocida por su eficacia en la introducción de no linealidad en el modelo, lo que permite aprender relaciones complejas entre las características [50].

3.2.3. Dropout

En la arquitectura del modelo, se integra una capa de dropout para aumentar la robustez y prevenir el sobre-ajuste. Esta capa se configura con una tasa de desactivación del 45 % , lo que significa que, durante el entrenamiento, el 45 % de las neuronas se desactivarán aleatoriamente en cada paso.

En una primera iteración del algoritmo tuvo una tasa de desactivación más baja. Luego de varias iteraciones se concluyó que se necesitaba un algoritmo de clasificación que estudiara más a detalle la data fomentando así una mejor generalización y reduciendo el riesgo de sobre-ajuste en el proceso de aprendizaje. Basándonos en esto aumentamos la tasa y obtuvimos mejores resultados.

Para asegurar la reproducibilidad, se establece una semilla(seed) 123. Esta introducción de aleatoriedad ayuda a que el modelo no dependa excesivamente de ninguna característica o neurona específica [51].

3.2.4. Capa de salida

La capa de salida utiliza una activación *softmax* para transformar las salidas del modelo en probabilidades de pertenencia a cada clase. Esto, clasifica las entradas en

categorías distintas, proporcionando probabilidades para cada clase, lo cual es esencial en la clasificación multi-clase.

3.2.5. Optimización

Para el proceso de entrenamiento, se utiliza el optimizador *Adamax* [52]. Este es una variante del conocido optimizador Adam, que combina las ventajas de los métodos adaptativos de tasa de aprendizaje con una implementación más robusta en entornos con gradientes dispersos, lo cual es común en imágenes médicas

La función de pérdida elegida es la *categorical crossentropy* [38], idónea para problemas de clasificación multi-clase. Se configura el modelo para minimizarla y se rastrea la precisión como métrica principal.

3.3. Distribución de datos

3.3.1. Experimento 1: Evaluación de la eficiencia de la división asimétrica de datos en la clasificación de imágenes de cáncer de piel

La distribución de datos fue la siguiente:

1. El 95 % de los datos se destinan al conjunto de entrenamiento.
2. El 2.5 % de los datos restantes se destinan al conjunto de validación.
3. El 2.5 % restante se destina al conjunto de pruebas.

Además se mezclan aleatoriamente los datos y se utiliza una variable fija para garantizar que la división sea reproducible. Aquí se utiliza *dummy split* para mantener la proporción deseada entre validación y prueba. Por lo que este experimento tiene 9514 datos de entrenamiento, 251 de test y 250 de validación.

Generadores de datos y preprocesamiento

Se establece para este un tamaño objetivo de muestras por clase (300 en este caso), y se utiliza un bucle para iterar a través de cada clase única. Se realiza un re-muestreo con reemplazo para clases con un número de muestras menor al objetivo (300), y sin reemplazo para clases con un número igual o mayor al tamaño objetivo.

Como se evidencia anteriormente, al separar la data en clases el conjunto se mantiene desbalanceado. Se hace necesario aplicar un método llamado *Class Weighting* para compensar este desbalanceo. Este método consiste en asignar un peso a cada

Tabla 3.2: Experimento 1: Distribución de imágenes de cáncer de piel en los conjuntos de entrenamiento, test y validación

Diagnostic category	**Training**	**Validation**	**Testing**
NV	300	158	163
MEL	300	25	35
BKL	300	34	30
DF	115	5	3
AKIEC	300	11	7
BCC	300	16	10
VASC	142	1	3

Tabla 3.3: Distribución de muestras por categoría después del sobre-muestreo

Diagnostic category	Sampling
AKIEC	300
BCC	300
BKL	300
DF	115
MEL	300
NV	300
VASC	142

clase inversamente proporcional a su frecuencia. De esta manera, las clases con menor representación tendrán un mayor peso y las clases con mayor representación tendrán un menor peso. Esto permite que el modelo se entrene de manera más equilibrada y que no se sesgue hacia las clases con mayor representación.

Tabla 3.4: Distribución de muestras con peso asignado

Diagnostic category	Sampling	Weighting
AKIEC	300	1.00
BCC	300	1.00
BKL	300	1.00
DF	115	2.60
MEL	300	1.00
NV	300	1.00
VASC	142	2.11

3.3.2. Experimento 2: Análisis de la estratificación de datos en la clasificación de imágenes de cáncer de piel

Descripción: Este experimento explora la estratificación de datos para mantener una distribución uniforme de etiquetas en cada conjunto de datos. La proporción de los conjuntos de datos es más equilibrada en comparación con el Experimento 1, lo que ofrece insights sobre la importancia de la distribución equitativa de datos en el entrenamiento y evaluación de modelos.

La distribución de datos fue la siguiente:

1. El 70 % de los datos se destinan al conjunto de entrenamiento.
2. El 15 % de los datos restantes se destinan al conjunto de validación.
3. El 15 % restante se destina al conjunto de pruebas.

Se utiliza *stratify* en ambas divisiones para mantener la distribución de etiquetas *label* en cada conjunto. Se calcula la proporción del conjunto de prueba sobre la suma del conjunto de test y el de validación.

Tabla 3.5: Experimento 2: Distribución de imágenes de cáncer de piel en los conjuntos de entrenamiento, test y validación

Diagnostic category	**Training**	**Validation**	**Testing**
NV	500	1006	1006
MEL	500	167	167
BKL	500	165	165
DF	500	17	17
AKIEC	500	49	49
BCC	500	77	77
VASC	500	21	22

Luego este quedaría distribuido en 7010 datos de entrenamiento, 1503 de test y 1502 de validación.

Generadores de datos y preprocesamiento

Se establece también un tamaño objetivo de muestras por clase (500). Se utiliza la función *groupby* para agrupar el dataFrame por la etiqueta de clase. Se itera sobre cada grupo, y se realiza un re-muestreo con reemplazo para grupos menores al tamaño deseado y sin reemplazo para los grupos que ya alcanzan o superan el tamaño deseado.

En este, a diferencia del primero, en cada clase se alcanza la misma cantidad de muestras, por lo que no es necesario aplicar *Class Weighting*.

Tabla 3.6: Distribución de muestras por categoría después del sobre-muestreo

Diagnostic category	Sampling
AKIEC	500
BCC	500
BKL	500
DF	500
MEL	500
NV	500
VASC	500

Capítulo 4

Resultados

En este capítulo se presenta un marco experimental para evaluar la efectividad de los experimentos realizados. Para ello se emplean gráficos para visualizar la pérdida y la precisión tanto de entrenamiento como de validación a lo largo de las épocas (iteraciones), marcando la época con menor pérdida de validación y mayor precisión de validación.

Se genera también un gráfico de barras que muestra la distribución de errores por clase en el conjunto de pruebas. Además se genera una matriz de confusión y un informe de clasificación que incluye precisión, recuperación (recall), puntuación F1 y soporte para cada clase. Al final del entrenamiento, se evalúa el modelo en el conjunto de pruebas y se obtiene la precisión del mismo. El modelo con mayor eficiencia luego de varios ajustes tuvo una eficacia cercana al 87%.

La tabla 4.1 al final del capítulo describe las métricas utilizadas para la evaluación del modelo:

Pérdida y presición (loss y accuracy)

En el primer experimento se observa que la precisión de entrenamiento aumenta con cada epoch, la pérdida de entrenamiento disminuye consistentemente y la tasa de aprendizaje permanece constante al principio y luego disminuye para afinar el entrenamiento a medida que el modelo comienza a converger, lo cual indica que el modelo esta entrenando de forma correcta. Sin embargo la división asimétrica de datos en este experimento influye en el aprendizaje del modelo. El uso de división asimétrica puede estar causando que el modelo esté sesgado hacia clases con más muestras, afectando la precisión general y la capacidad de generalizar. Esto se evidencia dado que a pesar de la disminución de la pérdida de validación y el aumento de la precisión de validación, la notable diferencia entre la precisión de entrenamiento y la precisión de validación podría indicar un potencial sobre-ajuste.

Tabla 4.1: Estadísticas básicas del modelo del experimento 1.

E	Loss	Acc	V loss	V acc	LR	M	Batch
1	9.587	40.581	8.95658	56.800	10^{-2}	acc	85.25
2	7.798	67.615	7.67235	66.800	10^{-2}	acc	21.72
3	6.884	79.340	6.96014	69.600	10^{-2}	acc	22.56
4	6.214	87.365	6.35865	71.200	10^{-2}	acc	25.81
5	5.646	91.690	5.94812	75.200	10^{-2}	vloss	23.08
6	5.172	92.999	5.44954	76.800	10^{-2}	vloss	23.23
...	...	...	...	...	...	...	...
34	0.627	99.886	1.24470	76.800	0.00013	val_loss	23.30

En el experimento 2 se utiliza la estratificación de datos para garantizar que cada clase este representada de manera proporcional. Similar al Experimento 1, en el Experimento 2 también se observa una disminución constante en la pérdida de entrenamiento (Loss) con cada epoch. La pérdida de entrenamiento (Loss) muestra una disminución significativa desde la primera hasta la última epoch registrada, pasando de 8.418 a 0.406. Esto indica un aprendizaje efectivo y una mejora continua en la capacidad del modelo para predecir con precisión las clases. La precisión (Acc), que comienza en 48.371 % y alcanza el 98.057 %, corrobora esta mejora constante. A diferencia del Experimento 1, la pérdida de validación (V loss) en este experimento, aunque también muestra una tendencia descendente, tiene una alineación más estrecha entre la precisión de entrenamiento y la precisión de validación. La pérdida de validación disminuye de manera más consistente y la precisión de validación es comparativamente más alta que en el Experimento 1, lo que indica una mejor capacidad de generalización.

Tabla 4.2: Estadísticas básicas del modelo del experimento 2.

E	Loss	Acc	V loss	V acc	LR	M	Batch
1	8.418	48.371	7.41700	66.911	10^{-2}	accuracy	184.55
2	6.362	69.943	5.81767	69.907	10^{-2}	accuracy	107.94
3	5.110	77.686	4.76153	72.969	10^{-2}	accuracy	106.30
4	4.181	81.371	3.86405	78.362	10^{-2}	accuracy	104.60
5	3.417	84.771	3.25588	77.097	10^{-2}	accuracy	101.90
6	2.777	87.314	2.70253	78.495	10^{-2}	accuracy	101.85
...	...	...	...	...	...	...	...
22	0.406	98.057	0.81854	83.955	0.00006	val_loss	101.18

Las tablas 4.1 y 4.2 corresponde a la evaluación del experimento 1 y 2 respectiva-

mente.

Curva de aprendizaje

Los análisis expuestos también se ven evidenciados en los gráficos de curvas de aprendizaje generados. Cada figura a continuación muestra dos gráficos, uno de pérdida y otro de precisión respectivamente, a lo largo de los epoch de entrenamiento y validación de cada experimento.

En el primer experimento la pérdida de entrenamiento (línea roja) y la pérdida de validación (línea verde) disminuyen con el tiempo, lo que indica que el modelo está aprendiendo. La mejor epoch basada en la pérdida de validación es la 31, marcada por un punto azul. La precisión de entrenamiento (línea roja) es casi perfecta, cercana al 100 %, lo que puede ser un indicador de sobre-ajuste. La precisión de validación (línea verde) mejora pero tiene una variabilidad considerable y alcanza su punto más alto en la epoch 29, también marcada con un punto azul. Hay una brecha notable entre la precisión de entrenamiento y validación, lo que puede ser un signo de que el modelo no está generalizando bien.

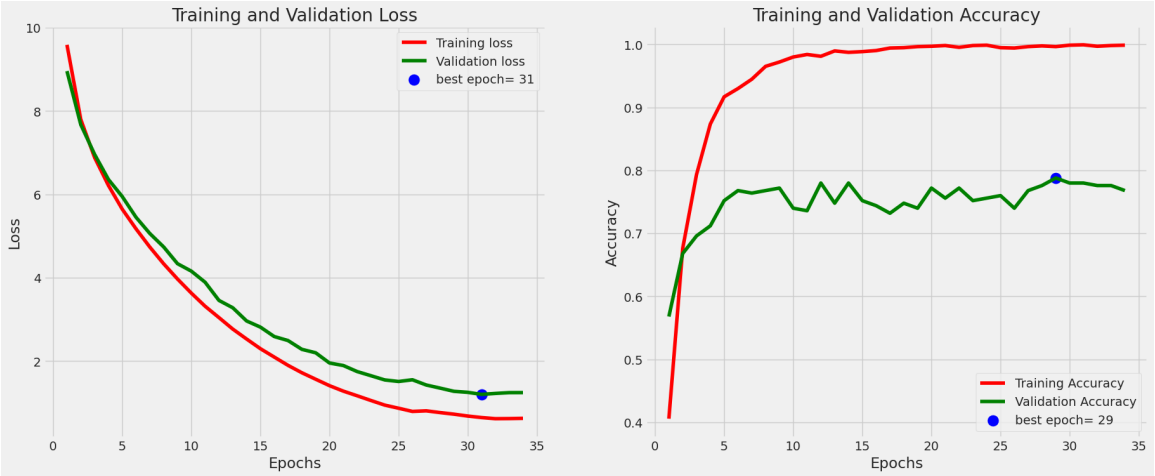

Figura 4.1: Curva de aprendizaje a lo largo del proceso de entrenamiento del experimento 1.

Similar al primer gráfico, en el segundo la pérdida de entrenamiento y validación disminuye, lo que es positivo. La mejor epoch basada en la pérdida de validación es la 19, que ocurre antes que en el primer gráfico, lo que indica una convergencia más rápida. La precisión de entrenamiento también es alta, pero no tan cercana al 100 % como en el primer gráfico, lo que sugiere un menor riesgo de sobre-ajuste. La

precisión de validación muestra menos variabilidad y una alineación más cercana con la precisión de entrenamiento, lo que es un indicador de mejor generalización. La diferencia entre la precisión de entrenamiento y validación es menor en comparación con el primer gráfico, lo que sugiere que el modelo podría estar generalizando mejor.

Figura 4.2: Curva de aprendizaje a lo largo del proceso de entrenamiento del experimento 2.

Las figuras 4.1 y 4.2 corresponde a la curva de aprendizaje del experimento 1 y 2 respectivamente.

Estadísticas de eficacia

La matriz de confusión proporciona información valiosa sobre el rendimiento del modelo en relación de Actual/Predicho, en términos de su capacidad para clasificar correctamente cada una de las siete clases de cáncer de piel. La diagonal principal de la matriz representa los verdaderos positivos (TP), el número de casos en los que el modelo ha predicho correctamente la clase correspondiente. Los valores fuera de la diagonal principal indican errores de clasificación

En ambas matrices la clase con el mayor número de verdaderos positivos es "NV", en la primera con 134 casos predichos correctamente y en la segunda con 862 casos.

En el caso de .ªKIEC", la tasa de TP mejoró significativamente, pasando de 7 de 300 a 47 de 500. Incluso teniendo en cuenta el aumento del tamaño del conjunto de datos, se trata de una clara mejora. Se observan mejoras similares en otras clases, como "BCC", "BKL", "DF", "MELz "NV". La tasa de TP (true positive o verdaderos

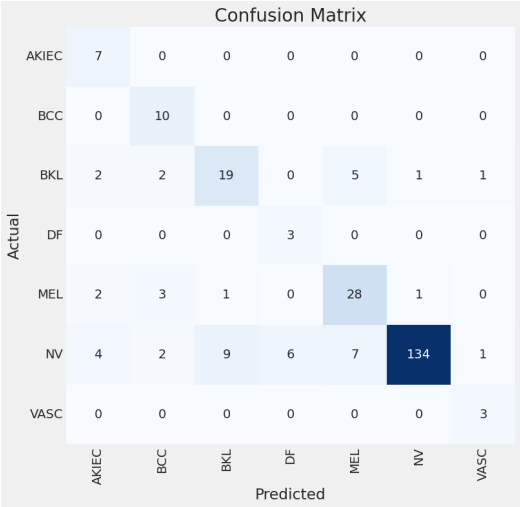

Figura 4.3: Estadísticas de eficacia del modelo al estimar los resultados en el conjunto de pruebas

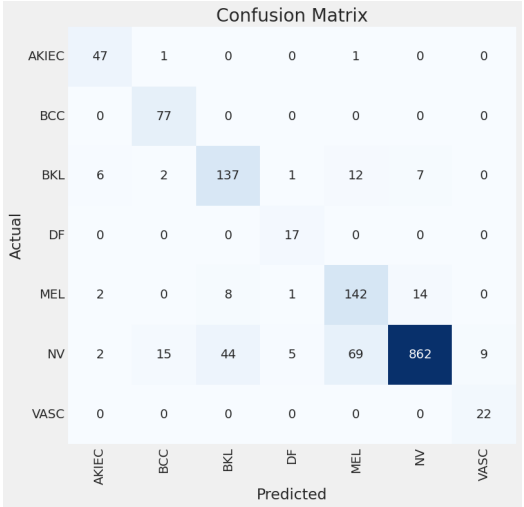

Figura 4.4: Estadísticas de eficacia del modelo al estimar los resultados en el conjunto de pruebas

positivos) ha aumentado no sólo en términos absolutos, sino también proporcionalmente si se tiene en cuenta el mayor tamaño del conjunto de datos. 'VASC' es un caso especial; mientras que la primera matriz no muestra ningún TP y tiene 3 FN (false negative o falsos negativos), la segunda matriz, a pesar de tener un gran número de FP(false positive o falsos positivos) (22), muestra que el modelo ha empezado a reconocer esta clase, cosa que antes no hacía.

En las siguientes figuras se evidencia el margen de error de las clases con mas errores que se obtuvieron en el experimento 1 y 2 respectivamente.

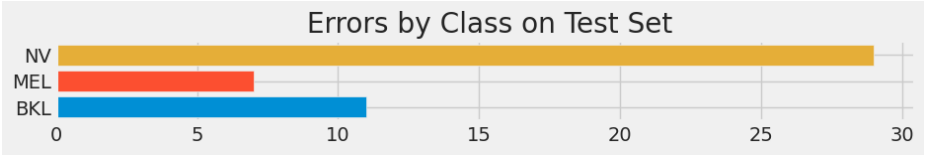

Figura 4.5: Gráfico de errores por clase en el conjunto de pruebas

En el caso de "NV", el número de errores del primer gráfico de barras se correlaciona con un conjunto de pruebas en el que el modelo tiene casi las mismas posibilidades de hacer una predicción correcta que de dar un falso positivo. En el segundo gráfico, a pesar del aumento de errores, el modelo tiene la misma proporción de verdaderos positivos que de falsos positivos, lo que sugiere que el rendimiento del modelo se ha mantenido constante en relación con el tamaño del conjunto de datos. MELz

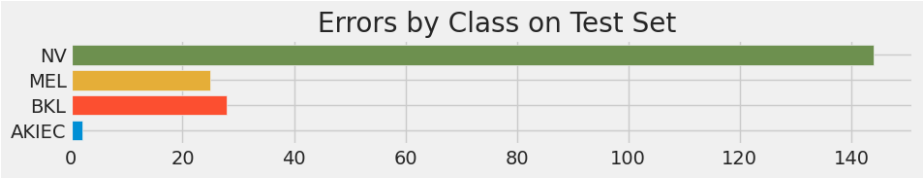

Figura 4.6: Gráfico de errores por clase en el conjunto de pruebas

"BKL"muestran un aumento de los errores en el segundo gráfico de barras, pero también es proporcional al aumento del tamaño del conjunto de datos. La proporción de verdaderos positivos frente a falsos positivos sigue siendo la misma. AKIEC"presenta un número relativamente pequeño de errores en el segundo gráfico de barras, lo que puede deberse al rendimiento relativamente bueno del modelo en esta clase en el conjunto de datos más grande.

Informe de clasificación

La tabla siguientes proporcionan una visión cuantitativa de la precisión, el recall (sensibilidad), el puntaje F1 y el soporte (número de muestras verdaderas) (NM) para cada categoría diagnóstica evaluada. Estos indicadores de rendimiento son esenciales para comprender la capacidad del modelo para identificar correctamente cada condición, así como su confiabilidad general en un conjunto de datos diverso. La métrica de 'Accuracy' refleja la proporción general de predicciones correctas, mientras que los promedios 'Macro' y 'Weighted' proporcionan una perspectiva agregada del rendimiento del modelo, teniendo en cuenta el desequilibrio en el soporte de las clases.

Tabla 4.3: Informe de clasificación combinado para los Experimentos 1 y 2

	Experimento 1				**Experimento 2**			
Categoría	**Acc**	**Recall**	**F1-Score**	**NM**	**Acc**	**Recall**	**F1-Score**	**NM**
AKIEC	0.47	1.00	0.64	7	0.82	0.96	0.89	49
BCC	0.59	1.00	0.74	10	0.81	1.00	0.90	77
BKL	0.66	0.63	0.64	30	0.72	0.83	0.77	165
DF	0.33	1.00	0.50	3	0.71	1.00	0.83	17
MEL	0.70	0.80	0.75	35	0.63	0.85	0.73	167
NV	0.99	0.82	0.90	163	0.98	0.86	0.91	1006
VASC	0.60	1.00	0.75	3	0.71	1.00	0.83	22
Accuracy			0.81	251			0.87	1503
Macro Avg	0.62	0.89	0.70	251	0.77	0.93	0.84	1503
Weighted Avg	0.86	0.81	0.83	251	0.89	0.87	0.87	1503

Se añadieron a la tabla 4.3 los campos *Accuracy*, *MacroAvg* y *WeightedAvg* para un mejor análisis de los resultados.

El campo *accuracy* muestra que la precisión mejora de 0.81 a 0.87, reflejando un aumento en la capacidad general del modelo para hacer predicciones correctas. El Macro Avg (Promedio Macro) aumento de 0.62 a 0.77 en precisión y de 0.70 a 0.84 en la puntuación F1, indicando una mejora en el rendimiento medio del modelo a través de todas las categorías. El Weighted Avg (Promedio Ponderado) muestra un incremento de 0.86 a 0.89 en precisión y de 0.83 a 0.87 en la puntuación F1, mostrando que, teniendo en cuenta el número de muestras (soporte), el rendimiento general del modelo ha mejorado.

Observaciones generales entre los experimentos

En general los experimentos demuestran resultados interesantes en la clasificación. El Experimento 2 muestra mejoras generalizadas en precisión, recall y puntuaciones F1 para la mayoría de las categorías. El soporte aumenta significativamente, lo que indica que se evaluó al modelo con más muestras, proporcionando una base más robusta para la evaluación del rendimiento. A pesar de que el soporte es mayor, lo que generalmente hace más desafiante mantener altas métricas, el Experimento 2 demuestra mejoras en las métricas en general. Este, a diferencia del 1, maneja mejor el desequilibrio de clases, lo que se refleja en una mejor diferenciación entre ciertas categorías diagnósticas, lo que es indicativo de un modelo más preciso y confiable.

4.1. Consideraciones Finales

Los resultados obtenidos son prometedores y sugieren que los modelos de aprendizaje profundo tienen un potencial considerable para mejorar la precisión y la eficiencia del diagnóstico del cáncer de piel. Tomando como referencia los resultados del experimento 2, se obtuvo una eficacia de clasificación de 87 %, lo cual es bajo con respecto a métodos más robustos mencionados en el estado del arte pero prometedor teniendo en cuenta que los algoritmos desarrollados utilizando EfficientNet obtienen resultados entre 84 % y 86.5 % de eficacia [34].

Es notable además la importancia para el dataset específico utilizado (HAM10000) de un balance proporcional en el conjunto de datos, dado que se evidencia en los resultados que la desproporción entre los mismos lleva a errores de sobre-ajuste y a un peor rendimiento del modelo.

Métricas

Tabla 4.4: Descripción de términos clave en el entrenamiento de modelos de aprendizaje automático.

Término	**Descripción**
Epoch	Es una iteración completa sobre todo el conjunto de datos de entrenamiento.
Loss (Pérdida)	Es una medida de cuán bien el modelo está realizando sus predicciones. Los valores decrecientes indican una mejora en el aprendizaje.
Accuracy (Precisión)	Muestra el porcentaje de etiquetas que el modelo predice correctamente para el conjunto de entrenamiento.
V loss (Pérdida de Validación)	Es similar a la pérdida, pero se calcula sobre un conjunto de datos que no se utiliza para el entrenamiento.
V acc (Precisión de Validación)	Muestra el porcentaje de etiquetas que el modelo predice correctamente para el conjunto de datos de validación.
LR (Learning Rate)	La tasa de aprendizaje dicta cuánto se ajustan los pesos del modelo en cada actualización.
Next LR (Próxima Learning Rate)	Indica la próxima tasa de aprendizaje planificada. La adaptación de la tasa de aprendizaje puede ayudar a evitar el estancamiento y mejorar la convergencia.
Monitor	Muestra la métrica que se está utilizando para monitorizar el rendimiento del modelo. Cambia de *accuracy* a *val loss*, lo que probablemente indica que el cambio se hizo para evitar el sobre'ajuste.
Duration (Duración)	Tiempo que tardó cada epoch en completarse. Importante para evaluar la eficiencia del entrenamiento.

Conclusiones

Los resultados obtenidos muestran una alta eficacia (79 % aproximadamente) que puede estar sujeta a futuras mejoras del modelo y del algoritmo. En general, los modelos de detección de cáncer de piel pueden tener una precisión del 85 % al 95 % o superior, dependiendo de la complejidad del problema, la calidad de los datos y la selección del algoritmo utilizado. Sin embargo, el 78 % es un porcentaje elevado para los algoritmos de aprendizaje automático, y con algunas mejoras en el modelo puede incrementarse. También se desprende de este resultado la intención de convertir este proyecto en un producto a gran escala para ser utilizado, inicialmente, por los profesionales sanitarios de atención primaria, como prueba complementaria de alta efectividad a la hora de derivar a un paciente con una lesión cutánea al área de oncología. Con la premisa de convertir los resultados obtenidos en un producto al servicio de la sociedad, el equipo investigador pretende seguir investigando y desarrollando una solución mejor.

Recomendaciones

Datos: procesamiento y estrategias de normalización

1. Enriquecimiento de Datos: Complementar el conjunto de datos con imágenes adicionales de fuentes confiables para mejorar la robustez del modelo.

2. Aumentar la Diversidad de Datos: Utilizar técnicas de aumento de datos para generar variantes adicionales de las imágenes existentes, como rotaciones, zoom, o cambios en el brillo, para mejorar la capacidad del modelo de generalizar a partir de nuevos datos.

3. Balance de Clases mediante Aumento: Para categorías con menos ejemplos, aplicar técnicas de aumento de datos de manera selectiva para equilibrar la distribución de clases y reducir el sesgo del modelo.

4. Estratificación Mejorada: Asegurarse de que la estratificación de datos se aplique de manera que todas las particiones (entrenamiento, validación, prueba) tengan una distribución representativa de cada clase.

Modelo: optimización y ajuste de hiperparámetros

1. Regularización y Arquitectura de Red: Explorar diferentes parámetros de regularización, para dropout y L1/L2 para encontrar el mejor equilibrio entre el rendimiento y la prevención del sobre-ajuste.

2. Presición y sesgos: Implementar modelos más complejos de clasificación para las clases menos representadas.

3. Optimizadores: Utilizar otros optimizadores como RMSprop o SGD para mejorar la velocidad de convergencia y el rendimiento del modelo.

Bibliografía

[1] *'Machine Learning': definición, tipos y aplicaciones prácticas*. Accessed: 2023-11-19. Iberdrola, 2023. URL: https://www.iberdrola.com/innovacion/machine-learning-aprendizaje-automatico (vid. pág. V).

[2] Yann LeCun, Yoshua Bengio y Geoffrey Hinton. «Deep learning». En: *nature* 521.7553 (2015), págs. 436-444 (vid. pág. 1).

[3] Li Deng, Dong Yu y col. «Deep learning: methods and applications». En: *Foundations and trends® in signal processing* 7.3–4 (2014), págs. 197-387 (vid. pág. 1).

[4] Lei Cai, Jingyang Gao y Di Zhao. «A review of the application of deep learning in medical image classification and segmentation». En: *Annals of translational medicine* 8.11 (2020) (vid. pág. 1).

[5] Hyun Ah Song y Soo-Young Lee. «Hierarchical representation using NMF». En: *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part I 20*. Springer. 2013, págs. 466-473 (vid. pág. 1).

[6] *Estadísticas importantes sobre el cáncer de piel tipo melanoma*. Último acceso: 2023-11-27. American Cancer Society, 2023. URL: https://www.cancer.org/es/cancer/tipos/cancer-de-piel-tipo-melanoma/acerca/estadisticas-clave.html (vid. págs. 1, 2).

[7] Fundación Piel Sana - Academia Española de Dermatología y Venereología. *¿Qué es la Dermatología?* https://aedv.fundacionpielsana.es/piel-sana/que-es-la-dermatologia/. Último acceso: 19 de noviembre de 2023 (vid. pág. 1).

[8] Kinnor Das y col. «Machine learning and its application in skin cancer». En: *International Journal of Environmental Research and Public Health* 18.24 (2021), pág. 13409 (vid. págs. 2, 3, 9).

[9] Departamento de Ingeniería Eléctrica. Universidad Nacional de Colombia Alberto Delgado PhD. Profesor Asociado. *Aplicación de las Redes Neuronales en Medicina*. Online; accessed 22-October-2023. 1999. URL: https://repositorio.unal.edu.co/bitstream/handle/unal/32711/19460-64062-1-PB.pdf (vid. págs. 2, 6-8, 10).

[10] Titus Josef Brinker y col. «Skin cancer classification using convolutional neural networks: systematic review». En: *Journal of medical Internet research* 20.10 (2018), e11936 (vid. págs. 2, 10).

[11] A Ameri. «A deep learning approach to skin cancer detection in dermoscopy images». En: *Journal of biomedical physics & engineering* 10.6 (2020), pág. 801 (vid. págs. 2, 10).

[12] Bhuvaneshwari Shetty y col. «Skin lesion classification of dermoscopic images using machine learning and convolutional neural network». En: *Scientific Reports* 12.1 (2022), pág. 18134 (vid. págs. 2, 10).

[13] Amin Tajerian y col. «Design and validation of a new machine-learning-based diagnostic tool for the differentiation of dermatoscopic skin cancer images». En: *Plos one* 18.4 (2023), e0284437 (vid. págs. 2, 10).

[14] Adekanmi Adegun y Serestina Viriri. «Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art». En: *Artificial Intelligence Review* 54 (2021), págs. 811-841 (vid. págs. 2, 10).

[15] Darien Viera Barredo. «Autómata celular estocástico en redes complejas para el estudio de la invasión, migración y metástasis del cancer». Tesis de Licenciatura. Habana, Cuba: Universidad de la Habana, 2019 (vid. pág. 3).

[16] Claudia Olavarrieta Martínez. «Ensemble de redes convolucionales para la clasificación de lesiones de cáncer de piel». Tesis de Licenciatura. Habana, Cuba: Universidad de la Habana, 2022 (vid. pág. 3).

[17] X. Du-Harpur y col. «What is AI? Applications of artificial intelligence to dermatology». En: *Frontiers in Medicine* 7 (2020). DOI: 10.3389/fmed.2020.00100. URL: https://www.frontiersin.org/articles/10.3389/fmed.2020.00100/full (vid. pág. 5).

[18] G Romero y col. «Practice models in teledermatology in Spain: longitudinal study, 2009-2014». En: *Actas Dermo-Sifiliográficas (English Edition)* 109.7 (2018), págs. 624-630 (vid. pág. 5).

[19] John D Whited y col. «Teledermatology's impact on time to intervention among referrals to a dermatology consult service». En: *Telemedicine Journal and e-Health* 8.3 (2002), págs. 313-321 (vid. pág. 5).

[20] D1 Piccolo y col. «Concordance between telepathologic diagnosis and conventional histopathologic diagnosis: a multiobserver store-and-forward study on 20 skin specimens». En: *Archives of dermatology* 138.1 (2002), págs. 53-58 (vid. pág. 5).

[21] Qindel. *Reconocimiento de imágenes: Qué es y cómo funciona*. Último acceso: 2023-12-01. 2023. URL: https://www.qindel.com/que-es-y-como-funciona-el-reconocimiento-de-imagenes/ (vid. pág. 6).

[22] L. Gerhardt. «Pattern recognition and machine learning». En: *IEEE Transactions on Automatic Control* 19.4 (1974), págs. 461-462. DOI: 10.1109/TAC.1974.1100578 (vid. pág. 6).

[23] Datamount. *What Canny Edge Detection Algorithm is All About*. Último acceso: 2023-12-01. 2023. URL: https://medium.com/@datamount/what-canny-edge-detection-algorithm-is-all-about-103d94553d21 (vid. pág. 7).

[24] Weibin Rong y col. «An improved CANNY edge detection algorithm». En: *2014 IEEE international conference on mechatronics and automation*. IEEE. 2014, págs. 577-582 (vid. pág. 7).

[25] *Cómo Extraer Features de una Imagen con HOG en Scikit-Image*. Último acceso: 2023-11-27. DataSmarts Español, 2020. URL: https://datasmarts.net/es/como-extraer-features-de-una-imagen-con-hog-en-scikit-image/ (vid. pág. 7).

[26] Tony Lindeberg. «Scale invariant feature transform». En: (2012) (vid. pág. 7).

[27] Michael Calonder y col. «Brief: Binary robust independent elementary features». En: *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*. Springer. 2010, págs. 778-792 (vid. pág. 7).

[28] *Color Histograms in Image Retrieval*. Disponible en: https://www.pinecone.io/learn/series/image-search/color-histograms/. 2023. URL: https://www.pinecone.io/learn/series/image-search/color-histograms/ (vid. pág. 7).

[29] Isidoro Gil Leiva, Pedro Díaz Ortuño y José Vicente Rodríguez Muñoz. *Técnicas y usos en la clasificación automática de imágenes*. 2019 (vid. pág. 7).

[30] Kang-Woo Lee y col. «Analysis of facial ultrasonography images based on deep learning». En: *Scientific reports* 12.1 (2022), pág. 16480 (vid. pág. 8).

[31] Akshay S Chaudhari y col. «Low-count whole-body PET with deep learning in a multicenter and externally validated study». En: *NPJ digital medicine* 4.1 (2021), pág. 127 (vid. pág. 8).

[32] Kaustav Bera y col. «Predicting cancer outcomes with radiomics and artificial intelligence in radiology». En: *Nature Reviews Clinical Oncology* 19.2 (2022), págs. 132-146 (vid. pág. 8).

[33] Mehwish Dildar y col. «Skin cancer detection: a review using deep learning techniques». En: *International journal of environmental research and public health* 18.10 (2021), pág. 5479 (vid. pág. 11).

[34] Karar Ali y col. «Multiclass skin cancer classification using EfficientNets–a first step towards preventing skin cancer». En: *Neuroscience Informatics* 2.4 (2022), pág. 100034 (vid. págs. 12, 36).

[35] S Papiththira y T Kokul. «Melanoma Skin Cancer Detection Using EfficientNet and Channel Attention Module». En: *2021 IEEE 16th International Conference on Industrial and Information Systems (ICIIS)*. IEEE. 2021, págs. 227-232 (vid. pág. 12).

[36] Philipp Tschandl, Cliff Rosendahl y Harald Kittler. «The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions». En: *Scientific data* 5.1 (2018), págs. 1-9 (vid. págs. 15, 16).

[37] *One-hot encoding*. `https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/`. 2023 (vid. pág. 16).

[38] Vijay Singh. *Keras - Categorical Cross Entropy Loss Function*. Accessed: [25 de noviembre de 2023]. 2021. URL: `https://vitalflux.com/keras-categorical-cross-entropy-loss-function/` (vid. págs. 16, 26).

[39] *Dataframes de Pandas*. `https://pandas.pydata.org/docs/getting_started/intro_tutorials/01_table_oriented.html`. 2023 (vid. pág. 17).

[40] Analytics Vidhya. *How to Improve Class Imbalance using Class Weights in ML?* Online; accessed 16-November-2023. Oct. de 2020. URL: `https://www.analyticsvidhya.com/blog/2020/10/improve-class-imbalance-class-weights/` (vid. pág. 17).

[41] *Aumentación de datos*. `https://machinelearningmastery.com/image-augmentation-deep-learning-keras/`. 2023 (vid. pág. 18).

[42] *EfficientNetB1*. `https://towardsdatascience.com/complete-architectural-details-of-efficientnets-7e50bc115fe0` (vid. pág. 18).

[43] Mingxing Tan y Quoc Le. *EfficientNet: Improving Accuracy and Efficiency through AutoML and Model Scaling*. Accessed: [24 de noviembre de 2023]. 2019. URL: `https://blog.research.google/2019/05/efficientnet-improving-accuracy-and.html` (vid. pág. 18).

[44] Pinecone. *ImageNet - The Dataset that Transformed Image Classification*. 2021. URL: https://www.pinecone.io/learn/series/image-search/imagenet/ (vid. pág. 18).

[45] Mingxing Tan y Quoc V. Le. «EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks». En: *ar5iv* (2019). URL: https://ar5iv.labs.arxiv.org/html/1905.11946 (vid. pág. 19).

[46] *Conjunto de datos HAM10000*. https://www.kaggle.com/kmader/skin-cancer-mnist-ham10000. 2023 (vid. pág. 20).

[47] *Tasa de aprendizaje*. https://www.jeremyjordan.me/nn-learning-rate/. 2023 (vid. pág. 20).

[48] *ImageDataGenerator*. https://keras.io/api/preprocessing/image/. 2023 (vid. pág. 24).

[49] *Regularización*. https://machinelearningmastery.com/how-to-reduce-overfitting-in-deep-neural-networks-with-weight-regularization-in-keras/. 2023 (vid. pág. 25).

[50] *Capas densas*. https://towardsdatascience.com/activated-dense-layer-intuition-and-implementation-93e090cadc34. 2023 (vid. pág. 25).

[51] *Dropout*. https://machinelearningmastery.com/dropout-regularization-deep-learning-models-keras/. 2023 (vid. pág. 25).

[52] *Adamax*. https://keras.io/api/optimizers/adamax/. 2023 (vid. pág. 26).