UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Industrial

Aplicación de técnicas de Inteligencia Artificial (DL y ML) para la clasificación de lesiones cutáneas a partir de imágenes demoscópicas y datos clínicos.

Trabajo Fin de Máster

Máster Universitario en Ingeniería Biomédica

AUTOR/A: Pérez Martínez, Sara

Tutor/a: Pastor López, Oscar

Director/a Experimental: Navarro Aljibe, Salvador Francisco

CURSO ACADÉMICO: 2023/2024

AGRADECIMIENTOS

Me gustaría agradecer sinceramente a todas las personas que han tenido un impacto directo en la ejecución de este Trabajo de Fin de Máster. Primero que nada, mis padres y mi familia por su apoyo inquebrantable a lo largo de cada paso del camino. Gracias por escuchar con paciencia, incluso cuando no entendían en lo más mínimo de qué estaba hablando, y por estar siempre a mi lado con sus palabras de apoyo y fuerza.

Querido Salva, muchísimas gracias por ser tan amable conmigo y por enseñarme a enfocar el trabajo. Tu talento y sabiduría han sido la columna vertebral de todo este trabajo, y te agradezco mucho cada minuto que has pasado intentando enseñarme.

A Óscar y a mis compañeros, muchas gracias por tratarme con tanto amor y paciencia. Vuestro compañerismo y apoyo me han sido tan útiles y han enriquecido este proceso.

Me gustaría agradecer a la Universidad Politécnica de Valencia (UPV) por brindarme los medios y conocimientos adecuados para entrar en este campo. La formación que he recibido en la UPV ha sido la base de mi crecimiento profesional y personal, y estoy muy agradecido por todo lo que he aprendido de ella. A la facultad y a todos los empleados: muchas gracias.

Y para finalizar me gustaría agradecer a la UPV el mejor regalo que he recibido jamás, a mis amigas de grado; Paula, María, Sofía, Carlos y en especial a Inma y a Mireia. Sois todos unos ejemplos a seguir para mí, pero vuestra generosidad, cariño y puro talento me hace agradecer diariamente la existencia de gente tan absolutamente maravillosa. Nos vemos en 40 años tomando café en la *Tarongeria*, no lleguéis tarde.

RESUMEN

La detección temprana de la malignidad de las lesiones cutáneas es crucial debido a la agresividad y alta tasa de mortalidad asociada a diversos tipos de cáncer de piel. Las lesiones cutáneas malignas, aunque representan una pequeña proporción de los casos, son responsables de la mayoría de las muertes por cáncer de piel. Cuando se detectan en etapas iniciales, los tratamientos pueden ser altamente efectivos, incrementando las tasas de supervivencia. La dermoscopia, una técnica de imagen no invasiva ha revolucionado la capacidad de los dermatólogos para evaluar las lesiones cutáneas sospechosas.

Implementar un sistema de cribado inicial utilizando dermoscopia en el contexto de la sanidad pública podría ser de gran relevancia, permitiendo identificar y priorizar rápidamente los casos sospechosos para una evaluación más exhaustiva. Las técnicas de aprendizaje profundo han demostrado ser herramientas eficaces en el análisis y clasificación de imágenes médicas, y su aplicación en la clasificación de lesiones cutáneas. Estas técnicas, basadas en redes neuronales convolucionales (CNN), pueden procesar y analizar grandes cantidades de datos de imágenes dermoscópicas con una precisión que a menudo supera la de los métodos tradicionales. Al aprender automáticamente las características relevantes para la clasificación de lesiones a partir de grandes conjuntos de datos etiquetados, las CNN pueden identificar patrones sutiles y complejos que podrían pasar desapercibidos para el ojo humano.

Además, los modelos de aprendizaje profundo pueden integrarse con datos clínicos adicionales, como características demográficas, para mejorar más la precisión diagnóstica. La implementación de estas tecnologías en sistemas de cribado inicial puede agilizar el proceso de detección, proporcionando evaluaciones rápidas y precisas que permiten a los dermatólogos centrarse en los casos críticos, reducir el número de biopsias innecesarias y monitorizar cambios en las lesiones cutáneas a lo largo del tiempo.

Este trabajo pretende aplicar técnicas de inteligencia artificial para la clasificación de lesiones cutáneas a partir de imágenes dermoscópicas y datos clínicos. El objetivo es desarrollar un sistema que mejore la precisión diagnóstica y facilite la detección temprana de lesiones malignas, optimizando así el proceso de cribado en el contexto de la sanidad pública, y abriendo la posibilidad al diagnóstico remoto mediante teledermatología.

Palabras clave: Dermoscopia, lesiones cutáneas, malignidad, red neuronal convolucional, datos multimodales, aprendizaje profundo, inteligencia artificial, cribado, teledermatología.

RESUM

La detecció primerenca de la malignitat de les lesions cutànies és crucial a causa de l'agressivitat i alta taxa de mortalitat associada a diversos tipus de càncer de pell. Les lesions cutànies malignes, tot i que representen una petita proporció dels casos, són responsables de la majoria de les morts per càncer de pell. Quan es detecten en etapes inicials, els tractaments poden ser altament efectius, incrementant les taxes de supervivència. La dermoscòpia, una tècnica d'imatge no invasiva ha revolucionat la capacitat dels dermatòlegs per avaluar les lesions cutànies sospitoses.

Implementar un sistema de cribratge inicial utilitzant dermoscòpia en el context de la sanitat pública podria ser de gran rellevància, permetent identificar i prioritzar ràpidament els casos sospitosos per a una avaluació més exhaustiva. Les tècniques d' aprenentatge profund han demostrat ser eines eficaces en l' anàlisi i classificació d' imatges mèdiques, i la seva aplicació en la classificació de lesions cutànies. Aquestes tècniques, basades en xarxes neuronals convolucionals (CNN), poden processar i analitzar grans quantitats de dades d'imatges dermoscòpiques amb una precisió que sovint supera la dels mètodes tradicionals. En aprendre automàticament les característiques rellevants per a la classificació de lesions a partir de grans conjunts de dades etiquetades, les CNN poden identificar patrons subtils i complexos que podrien passar desapercebuts per a l' ull humà.

A més, els models d' aprenentatge profund poden integrar-se amb dades clíniques addicionals, com característiques demogràfiques, per millorar més la precisió diagnòstica. La implementació d' aquestes tecnologies en sistemes de cribratge inicial pot agilitar el procés de detecció, proporcionant avaluacions ràpides i precises que permeten als dermatòlegs centrar-se en els casos crítics, reduir el nombre de biòpsies innecessàries i monitoritzar canvis en les lesions cutànies al llarg del temps.

Aquest treball pretén aplicar tècniques d'intel·ligència artificial per a la classificació de lesions cutànies a partir d'imatges dermoscòpiques i dades clíniques. L'objectiu és desenvolupar un sistema que millori la precisió diagnòstica i faciliti la detecció primerenca de lesions malignes, optimitzant així el procés de cribratge en el context de la sanitat pública, i obrint la possibilitat al diagnòstic remot mitjançant teledermatologia.

Paraules clau: Dermoscòpia, lesions cutànies, malignitat, xarxa neuronal convolucional, dades multimodals, aprenentatge profund, intel·ligència artificial, cribratge, teledermatologia.

ABSTRACT

Early detection of the malignancy of skin lesions is crucial due to the aggressiveness and high mortality rate associated with various types of skin cancer. Malignant skin lesions, although they account for a small proportion of cases, are responsible for the majority of skin cancer deaths. When detected in early stages, treatments can be highly effective, increasing survival rates. Dermoscopy, a non-invasive imaging technique, has revolutionized dermatologists' ability to evaluate suspicious skin lesions.

Implementing an initial screening system using dermoscopy in the context of public health could be of great relevance, allowing suspected cases to be quickly identified and prioritized for more thorough evaluation. Deep learning techniques have proven to be effective tools in the analysis and classification of medical images, and their application in the classification of skin lesions. These techniques, based on convolutional neural networks (CNNs), can process and analyze large amounts of dermoscopic imaging data with accuracy that often exceeds that of traditional methods. By automatically learning features relevant to lesion classification from large labeled datasets, CNNs can identify subtle and complex patterns that might go unnoticed by the human eye.

In addition, deep learning models can be integrated with additional clinical data, such as demographic characteristics, to further improve diagnostic accuracy. Implementing these technologies in initial screening systems can streamline the screening process, providing fast and accurate assessments that allow dermatologists to focus on critical cases, reduce the number of unnecessary biopsies, and monitor changes in skin lesions over time.

This work aims to apply artificial intelligence techniques for the classification of skin lesions from dermoscopic images and clinical data. The aim is to develop a system that improves diagnostic accuracy and facilitates the early detection of malignant lesions, thus optimising the screening process in the context of public health, and opening the possibility of remote diagnosis through teledermatology.

Keywords: Dermoscopy, skin lesions, malignancy, convolutional neural network, multimodal data, deep learning, artificial intelligence, screening, teledermatology.

I. MEMORIA

Documentos contenidos en el TFM

- Memoria
- Presupuesto
- Añejo I

ÍNDICE DE LA MEMORIA

ÍNDICE DEL PRESUPUESTO

ÍNDICE DEL AÑEJO I

1. Introducción

1.1. Contexto y motivación

El cáncer de piel es uno de los tipos de cáncer más comunes en Europa. Según la Academia Española de Dermatología y Venerología (*AEDV: Academia Española de Dermatología y Venereología*, s. f.) en España la incidencia del cáncer de piel ha aumentado un 40% en los cuatro últimos años. Anualmente, se diagnostica a más de 78.000 nuevos pacientes y se espera que en 2040 el melanoma, se convierta en el segundo tumor en incidencia global. La detección temprana es fundamental para mejorar las tasas de supervivencia y los resultados del tratamiento, el diagnóstico precoz puede aumentar las posibilidades de supervivencia, alcanzándose una tasa de supervivencia a cinco años del 98% (Esteva et al., 2017).

Entre las técnicas empleadas para detectar estas lesiones malignas, la dermatoscopia se ha consolidado como una herramienta no invasiva esencial. Esta técnica de imagen permite a los dermatólogos observar estructuras subcutáneas que no son visibles a simple vista. A pesar de la experiencia de los dermatólogos, el diagnóstico temprano del melanoma sigue siendo una tarea abrumadora, ya que se presenta en muchas formas, tamaños y colores diferentes, incluso entre muestras de la misma categoría. En el artículo publicado en Nature por Esteva et al. (2017) se afirma que los dermatólogos tienen una tasa de precisión de entre el 65% y el 80% en inspección visual y hasta el 75% y el 84% mediante el uso de imágenes dermatoscópicas.

Las limitaciones de las técnicas actuales conducen a la necesidad de nuevas tecnologías, entre ellas las técnicas de inteligencia artificial y aprendizaje profundo, como las redes neuronales convolucionales (CNN), que han demostrado ser altamente eficaces. Estas técnicas son capaces de procesar grandes volúmenes de datos y aprender de manera autónoma las características más relevantes para la clasificación de las lesiones, alcanzando niveles de precisión diagnóstica que, en algunos casos, superan a los métodos tradicionales. Además, la incorporación de datos clínicos, como la edad, el historial médico del paciente y otras características relevantes, en conjunto con las imágenes dermoscópicas, permite desarrollar modelos más integrales y precisos, aumentando la efectividad del diagnóstico.

Implementar un sistema de cribado inicial que combine dermatoscopia y técnicas de aprendizaje profundo en el ámbito de la sanidad pública podría ofrecer ventajas significativas. Este enfoque no solo ayudaría a identificar y priorizar rápidamente los casos sospechosos para una evaluación más exhaustiva, sino que también optimizaría los recursos sanitarios, reduciendo el tiempo y los costos asociados con el diagnóstico manual, aliviando la carga sobre el sistema de salud y mejorando, en última instancia, la atención al paciente.

1.2. Objetivos del trabajo

El objetivo principal de este trabajo es desarrollar y diseñar un sistema de inteligencia artificial que mejore la clasificación de la malignidad de lesiones cutáneas a partir de imágenes dermoscópicas y datos clínicos. Este sistema busca optimizar el proceso de diagnóstico, facilitando la detección temprana de melanomas y otras lesiones malignas en entornos clínicos.

Para alcanzar este objetivo general, el trabajo se enfoca en los siguientes objetivos específicos:

1. Evaluar diversas arquitecturas de redes neuronales convolucionales (CNN) para la clasificación de imágenes dermoscópicas, con el fin de identificar la mejor estrategia de modelado para diferenciar eficazmente entre lesiones benignas y malignas.

2. Desarrollar modelos de inteligencia artificial que utilicen variables clínicas relevantes para mejorar la precisión en la clasificación de riesgo de malignidad.
3. Diseñar un modelo multimodal que combine la información de imágenes dermoscópicas y datos clínicos, optimizando así la capacidad de diagnóstico del sistema al aprovechar la complementariedad de ambas fuentes de información.
4. Comparar el rendimiento de diferentes enfoques, incluyendo modelos de ensamblado y el modelo multimodal, para justificar la selección del modelo final más eficiente y aplicable en la práctica clínica.

Estos objetivos se alinean con la necesidad de proporcionar una herramienta que no solo mejore la precisión diagnóstica de las lesiones cutáneas malignas, sino que también sea viable para su implementación en diversos contextos clínicos, facilitando el acceso a un diagnóstico más temprano y preciso.

1.3. Objetivos de desarrollo sostenible

El presente TFM puede englobarse en varios Objetivos de Desarrollo Sostenible (ODS) establecidos por las naciones unidas, en especial en lo que respecta a la salud y el bienestar, la educación de calidad, la innovación y la reducción de desigualdades.

ODS 3: Salud y Bienestar

El desarrollo de un modelo basado en técnicas de inteligencia artificial para la clasificación de la malignidad de las lesiones cutáneas directamente contribuye al ODS 3, que busca garantizar una vida sana y promover el bienestar para todos en todas las edades. La mejora en la precisión diagnóstica y la capacidad para detectar tempranamente el cáncer de piel puede salvar vidas y mejorar significativamente la calidad de vida de los pacientes. Al integrar datos clínicos y de imágenes dermoscópicas, este trabajo ayuda a reducir las tasas de mortalidad y morbilidad por cáncer de piel, ofreciendo una herramienta poderosa para el cribado y diagnóstico precoz.

ODS 4: Educación de Calidad

Este proyecto también contribuye al ODS 4 al fomentar la investigación y el desarrollo de habilidades avanzadas en el campo de la inteligencia artificial y la medicina. Al proporcionar un ejemplo práctico de cómo las técnicas de aprendizaje profundo pueden aplicarse en la práctica clínica, este trabajo sirve como un recurso educativo en los campos de la ingeniería biomédica, la informática y la medicina. Además, la colaboración con instituciones académicas y de investigación promueve el intercambio de conocimientos y el desarrollo de capacidades.

ODS 9: Industria, Innovación e Infraestructura

El uso de tecnologías avanzadas como la inteligencia artificial para mejorar los diagnósticos médicos está alineado con el ODS 9, que busca construir infraestructuras resilientes, promover la industrialización inclusiva y sostenible, y fomentar la innovación. Este proyecto no solo introduce innovaciones en el diagnóstico médico, sino que también propone la implementación de sistemas de seguimiento automatizado y telediagnóstico, que son cruciales para modernizar los servicios de salud y hacerlos más accesibles y eficientes.

ODS 10: Reducción de las Desigualdades

La implementación de herramientas de telediagnóstico y el uso de modelos de inteligencia artificial en la atención primaria pueden reducir significativamente las desigualdades en el acceso a servicios de

salud de calidad, alineándose con el ODS 10. Estas tecnologías permiten que las personas en áreas remotas o con acceso limitado a dermatólogos especializados reciban diagnósticos precisos y oportunos. Al mejorar el acceso a la atención médica y garantizar que más personas puedan beneficiarse de diagnósticos tempranos y precisos, este trabajo ayuda a reducir las disparidades en salud.

1.4. Estructura del documento

La estructura del presente documento sigue los principios de la **Ciencia del Diseño** (*Design Science*) (Gregory, 1966), centrando su enfoque en la identificación de un problema relevante y el desarrollo de una solución innovadora. Primero, se presenta la introducción, que establece el contexto, la relevancia del problema y los objetivos de la investigación. A continuación, se incluye una revisión de la literatura y el marco teórico, que fundamenta la necesidad del estudio y proporciona antecedentes sobre las técnicas actuales de diagnóstico dermatológico y las herramientas de inteligencia artificial utilizadas.

El documento prosigue con la definición del problema y los requerimientos del artefacto, seguida de la descripción del diseño y desarrollo del sistema propuesto, detallando los métodos y modelos implementados. Posteriormente, se presenta la evaluación de la solución, donde se analizan los resultados obtenidos y se compara su efectividad con los métodos tradicionales. Finalmente, se discuten las implicaciones de los hallazgos, las limitaciones del estudio y las recomendaciones para futuras investigaciones, concluyendo con un resumen de las contribuciones principales.

Esta estructura permite una comprensión clara del proceso de investigación y del desarrollo del artefacto propuesto, destacando tanto su justificación científica como su potencial aplicabilidad práctica.

2. Marco teórico y estado del arte

2.1. El órgano de la piel y las lesiones cutáneas

La piel es el órgano más grande del cuerpo humano, cubriendo una superficie de aproximadamente 2 m² y pesando unos 3 kg en los adultos (Madheswari & Karthikeyan, 2024). Su función principal es proteger el cuerpo contra amenazas externas, es la primera barrera con el exterior. Está formada por tres capas: la epidermis (la más externa), la dermis y la hipodermis.

A su vez, la epidermis se divide en cuatro capas: la capa córnea, que es protectora y contiene sebo y queratina que se desgastan con el tiempo; la capa granulosa, que puede estar queratinizada; la capa de Malpighi, que tiene terminaciones nerviosas libres responsables de las sensaciones de dolor; y la capa basal, que produce continuamente nuevas células. En esta última se encuentran entre otros los melanocitos, las células de Langerhans y los linfocitos, que son muy importantes para la respuesta inmunológica.

La dermis está compuesta principalmente de colágeno y elastina, y contiene vasos sanguíneos terminales. Además, en la dermis se encuentran glándulas sebáceas, fibroblastos, macrófagos y terminaciones nerviosas responsables del tacto y la hipersensibilidad.

La hipodermis o tejido subcutáneo alberga los folículos pilosos, glándulas sudoríparas, fibras nerviosas y una red capilar que regula la sensación de temperatura. Las diversas capas de la piel se pueden ver esquematizadas en la Ilustración 1.

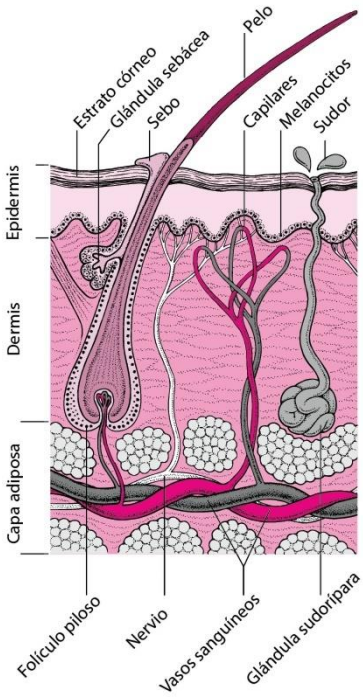


Ilustración 1. Corte transversal de la piel. Fuente: (*Anatomy of the Skin* , s. f.)

Las lesiones cutáneas son cambios o alteraciones en la piel que pueden variar mucho en su apariencia, tamaño, color y textura, debido a su naturaleza altamente heterogénea. Pueden ser causadas por muchos factores, como infecciones, inflamaciones, alergias, enfermedades autoinmunes, exposición al sol, traumatismos y factores genéticos, muchos de los anteriores en combinación. Algunas de las causas más comunes de lesiones cutáneas pueden ser infecciones, como el herpes, inflamaciones, como la dermatitis, alergias, como la urticaria, u otros factores como enfermedades autoinmunes, traumatismos, exposición a radiación o factores genéticos, entre muchos otros.

Se pueden dividir las lesiones cutáneas en dos tipos, dependiendo de su origen. Por un lado, están las lesiones primarias, que aparecen sobre la piel sana, y las lesiones secundarias, que resultan de la evolución de una lesión primarias.

La malignidad se refiere a la capacidad de una lesión de convertirse en cancerosa, es decir, de crecer sin control e invadir tejidos cercanos o diseminarse a otras partes del cuerpo. El melanoma es el tipo de cáncer de piel más peligroso, suele comenzar en los melanocitos y se manifiesta como un lunar asimétrico, con bordes irregulares, varios colores y cambios en tamaño o forma (Goldsmith, Lowell A. Katz, Stephen I. Glichrest, Barbara A. Paller, Amy S. Leffell, David J. Wolff, 2014).

2.1.1. Prevalencia y mortalidad

Aunque el cáncer de piel en general es uno de los más comunes, el melanoma es su forma más agresiva y peligrosa, desarrollándose a partir de los melanocitos con una capacidad notable para crecer y diseminarse rápidamente. El melanoma cutáneo representa una carga significativa tanto a nivel global como europeo y español. Este tipo de cáncer es de particular interés en el contexto de este trabajo debido a su alta mortalidad y a su creciente incidencia, especialmente en poblaciones de piel clara y en edades avanzadas, influenciada principalmente por la exposición a la radiación ultravioleta (UV).

A nivel global, el melanoma se considera una de las formas más letales de cáncer de piel, con una prevalencia que continúa aumentando. La mortalidad asociada al melanoma varía significativamente según la región y la efectividad de los tratamientos disponibles (Goldsmith, Lowell A. Katz, Stephen I. Glichrest, Barbara A. Paller, Amy S. Leffell, David J. Wolff, 2014). En Europa, por ejemplo, se estimaron 55.397 nuevos casos de melanoma cutáneo en hombres y 50.972 en mujeres durante el año 2020, y se registraron 9.457 muertes en hombres y 7.031 en mujeres, destacando la gravedad de esta forma específica de cáncer de piel (*Cancer burden statistics and trends across Europe | ECIS*, s. f.). Además, se trata se trata de una de las enfermedades de mayor auge en nuestro país, con un incremento de la tasa de incidencia del 181,3 % en varones y del 205,3 % en mujeres. La mortalidad se ha visto incrementada en España en las últimas décadas (1,76 % en varones y 1,26 % en mujeres), si bien esta tendencia al alza se ha estabilizado en los últimos años (Sãenz et al., 2005).

En comparación con otros tipos de cáncer de piel, como el carcinoma basocelular o el carcinoma de células escamosas, el melanoma, aunque menos frecuente, es responsable de la mayoría de las muertes relacionadas con el cáncer de piel debido a su comportamiento agresivo.

La elección de centrarse en el melanoma dentro de este trabajo se justifica por su potencial letalidad y la necesidad urgente de mejorar las herramientas de detección y diagnóstico, especialmente aquellas que permitan una intervención temprana. Las técnicas avanzadas basadas en redes neuronales convolucionales y aprendizaje profundo son particularmente prometedoras para la detección temprana de melanoma, debido a su capacidad para identificar patrones complejos en imágenes dermoscópicas y datos clínicos que podrían pasar desapercibidos con métodos tradicionales. Un diagnóstico precoz puede aumentar significativamente las tasas de supervivencia y reducir la mortalidad, justificando así la necesidad de enfocar este estudio en el melanoma mientras se abordan también otros tipos de cáncer de piel.

2.2. Diagnóstico de lesiones cutáneas

El diagnóstico de lesiones cutáneas ha avanzado significativamente gracias a las innovaciones tecnológicas y la mejora en las técnicas diagnósticas. Uno de los métodos más avanzados y ampliamente utilizados es la dermatoscopia. Se trata de una técnica diagnóstica in vivo, no invasiva, desarrollada para estudiar las lesiones cutáneas. Permite mejorar la precisión diagnóstica de las lesiones hiperpigmentadas y el diagnóstico precoz de las lesiones potencialmente malignas, especialmente el melanoma. No incrementa significativamente el tiempo dedicado a la exploración física.

Esta herramienta ha transformado el diagnóstico de lesiones cutáneas, especialmente en la detección temprana de melanoma y otros tipos de cáncer de piel. La dermatoscopia utiliza un dispositivo llamado dermatoscopio (Figura 1), que combina iluminación y magnificación para examinar las lesiones cutáneas (Malvehy et al., 2006). La dermatoscopia se basa en la transiluminación de la lesión cutánea y la amplificación de la imagen mediante lentes, mejorando la visibilidad de estructuras cutáneas profundas (Palacios-Martínez & Díaz-Alonso, 2017).

Figura 1. La figura 2A muestra un dermatoscopio con luz convencional. La figura 2B muestra un dermatoscopio con luz polarizada. Fuente: (*Dermatoscopia - Wikipedia, la enciclopedia libre*, s. f.)

La dermatoscopia ofrece numerosas ventajas, una de ellas siendo la detección temprana de melanoma, ya que permite identificarlos en etapas iniciales, mejorando significativamente el pronóstico. No obstante, la dermatoscopia también puede estar sujeta a errores y sesgos. Al tratarse de una técnica de inspección visual, depende en gran medida de la experiencia del profesional que la realiza. Además,

también puede inducir sesgos la calidad y resolución de las imágenes obtenidas. También cabe mencionar, que, al tratarse de un sistema humano, existe el sesgo de confirmación, ya que los dermatólogos pueden ser influenciados por sus expectativas previas sobre una lesión, afectando así el juicio diagnóstico. En un metaanálisis publicado en *JAMA Dermatology* (Williams et al., 2021) se evaluó la precisión de la dermatoscopia en comparación con el examen clínico visual. Los resultados mostraron que la dermatoscopia aumenta la sensibilidad y especificidad en la detección de melanoma.

2.3. Inteligencia artificial en el diagnóstico dermatológico

El término inteligencia artificial (IA) se refiere a un conjunto de tecnologías en constante evolución que imitan la inteligencia humana y se utilizan para resolver una amplia gama de problemas. Funcionan combinando grandes volúmenes de datos con procesamiento rápido e iterativo y algoritmos avanzados, permitiendo al software aprender automáticamente a partir de las características o patrones presentes en los datos, a menudo invisibles para los humanos. Entre los subcampos de la IA se encuentran el aprendizaje automático (Machine Learning, ML), el procesamiento del lenguaje natural y la robótica, entre otras áreas (Mukhamediev et al., 2022).

En el ámbito del aprendizaje profundo (Deep Learning, DL por sus siglas en inglés), una rama del ML que ha ganado particular relevancia en dermatología son las redes neuronales convolucionales (*Convolutional Neural Networks*, CNN). Estas redes han demostrado ser especialmente efectivas en la clasificación de lesiones cutáneas, diferenciando con alta precisión entre melanomas y otras lesiones. Además, las CNN se están utilizando cada vez más para identificar diversas afecciones cutáneas, mejorando así la capacidad para distinguir la mayoría de los casos que se presentan en la atención primaria de salud. Este creciente reconocimiento de la relevancia de la IA en dermatología se refleja en estudios recientes, como el de Guzmán-Bucio y Vega-Memije (Guzmán-Bucio & Vega-Memije, 2023), que destacan los principales subcampos de la inteligencia artificial con impacto en esta especialidad (ver Figura 3).

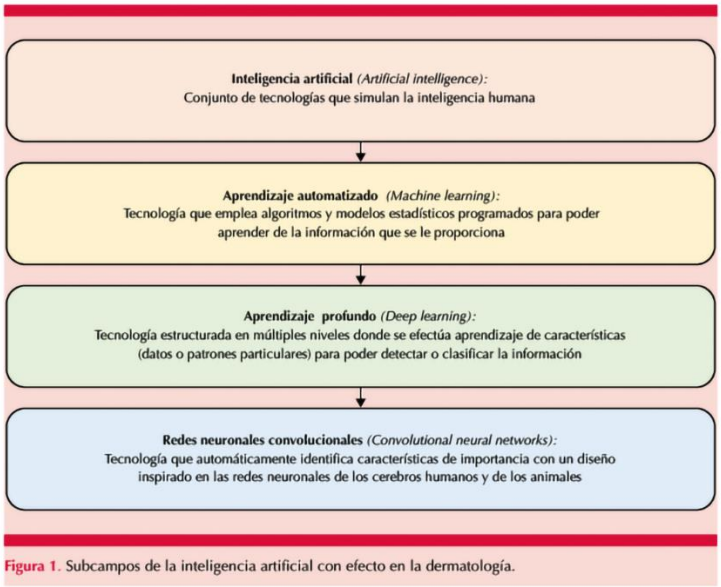


Figura 2. Subcampos de la inteligencia artificial con efecto en la dermatología. Fuente: (Guzmán-Bucio & Vega-Memije, 2023)

Tal y como se ilustra en la Figura 2, las técnicas de IA están ganando relevancia en el campo de la dermatología, gracias a la gran cantidad de imágenes generadas por los dispositivos diagnósticos de imagen, como los dermoscópicos, que permiten encontrar patrones complejos invisibles para el ojo humano, y que pueden ser determinantes para un diagnóstico temprano.

2.3.1. Redes neuronales artificiales

Las redes neuronales artificiales (Artificial Neural Networks, ANN) son modelos computacionales inspirados en la estructura y el funcionamiento del cerebro humano. Están compuestas por unidades básicas llamadas neuronas, organizadas en capas, que trabajan conjuntamente para procesar información y realizar tareas como el reconocimiento de patrones, la clasificación y la predicción (Goodfellow et al., 2016).

Las ANN están formadas por una capa de entrada, múltiples capas ocultas y una capa de salida. Cada capa está compuesta por nodos o neuronas interconectadas con las neuronas de las capas adyacentes. Cada neurona recibe una entrada, aplica una función de activación y transmite la salida a las neuronas de la siguiente capa. Este proceso se ve ilustrado en la Figura 3.

El perceptrón, mostrado en la Figura 3, es la unidad básica de las redes neuronales artificiales. Una red neuronal puede considerarse como una colección de perceptrones organizados en capas. Un solo perceptrón puede resolver problemas linealmente separables, pero para problemas más complejos se utilizan perceptrones multicapa. Las Redes Neuronales Profundas (Deep Neural Networks, DNN) son una extensión de los perceptrones multicapa con muchas capas ocultas, lo que permite a las redes aprender representaciones de características más complejas y jerárquicas.

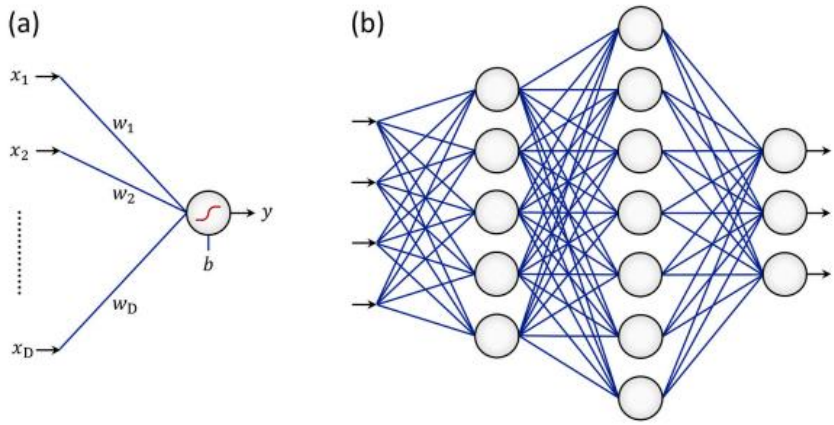


Figura 3. (a) Un perceptrón y (b) un perceptrón multicapa con cuatro entradas, dos capas ocultas y tres salidas. Fuente (Razavi, 2021)

Entrenamiento y optimización

Durante el entrenamiento, los datos se pasan a través de la red desde la capa de entrada hasta la capa de salida, calculando las predicciones. Se utiliza una función de pérdida para medir la discrepancia entre las predicciones de la red y los valores reales. Esta función, también llamada función de coste, cuantifica la discrepancia entre las predicciones de la red comparadas con los resultados esperados y su valor indica el grado de error del modelo. Mediante el algoritmo de retropropagación, los errores se propagan hacia atrás a través de la red para ajustar los pesos de las conexiones entre neuronas, minimizando la función de pérdida.

Para ajustar los pesos y minimizar la función de pérdida, se utiliza un algoritmo de optimización, como el gradiente descendente. Este proceso iterativo permite que la red neuronal aprenda y mejore su precisión al identificar patrones en los datos.

En resumen, el aprendizaje profundo permite a las redes neuronales artificiales analizar y comprender patrones complejos en grandes volúmenes de datos mediante una estructura de múltiples capas, un proceso de entrenamiento que ajusta continuamente los pesos y un algoritmo de optimización que refina los resultados.

Fortalezas y debilidades

Las ANN tienen la capacidad de modelar y aprender relaciones complejas y no lineales en los datos. Su flexibilidad permite utilizarlas en una amplia variedad de tareas, incluyendo clasificación, regresión, reconocimiento de patrones y generación de datos. Además, las ANN son adaptables y pueden mejorar su rendimiento con el tiempo a medida que se entrenan con más información.

Sin embargo, las ANN requieren grandes volúmenes de datos para entrenarse efectivamente y evitar el sobreajuste. El entrenamiento de redes neuronales puede ser intensivo en términos de recursos computacionales, necesitando hardware especializado como GPUs. Además, las redes neuronales a menudo son consideradas como "cajas negras" debido a la dificultad de interpretar cómo se toman las decisiones internas.

2.3.2. Redes neuronales convolucionales

Las redes neuronales convolucionales (*Convolutional Neural Networks*, CNN) son un tipo específico de red neuronal artificial diseñada para procesar datos con una estructura de cuadrícula, como imágenes. Son especialmente eficaces en tareas de reconocimiento y clasificación de imágenes debido a su capacidad para capturar características espaciales y patrones locales en los datos de entrada. El esquema básico de una red neuronal convolucional se muestra resumido en la Figura 4.

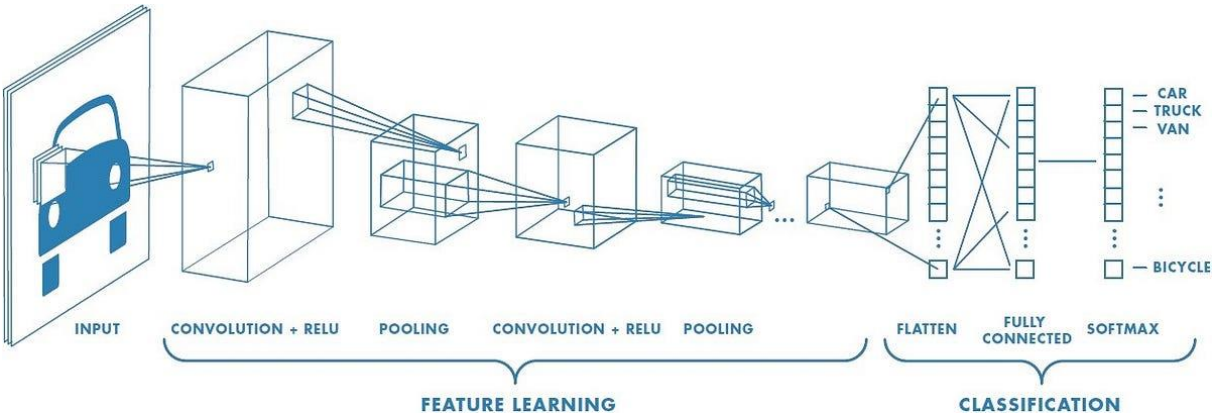


Figura 4. Esquema del funcionamiento de una red neuronal convolucional para clasificación de imágenes. Fuente: (Saha, 2018)

Componentes principales

Capas convolucionales (*Convolutional Layers*)

Las capas convolucionales utilizan filtros (*kernels*), que son pequeñas matrices de pesos aplicadas a la entrada para crear mapas de características. Los filtros se mueven a través de la imagen de entrada, realizando una operación de convolución para detectar características específicas como bordes, texturas y patrones. Esta operación implica multiplicar los valores de los píxeles de la imagen por los valores del filtro y sumar los resultados para producir un valor en el mapa de características.

Capas de *Pooling* (*Pooling Layers*)

Las capas de *pooling*, como *max pooling* y *average pooling*, reducen la dimensionalidad de los mapas de características, reteniendo la información más relevante y reduciendo la complejidad computacional. *Max pooling* selecciona el valor máximo en una ventana de la característica, mientras que *average pooling* calcula el promedio.

Capas de activación (Activation Layers)

Las capas de activación en redes neuronales son componentes esenciales que aplican una función de activación a la salida de cada neurona. Su principal objetivo es introducir no linealidad en el modelo, lo que permite que la red neuronal aprenda y represente relaciones complejas en los datos. La función de activación ReLU (*Rectified Linear Unit*) es común en las CNN e introduce no linealidades en el modelo. La función ReLU establece los valores negativos en cero y mantiene los valores positivos.

Por otro lado, **Swish** es una función de activación utilizada en redes neuronales profundas, introducida por primera vez por Google en 2017. Se define matemáticamente como Swish(x)= x·sigmoid(x). La función *Swish* combina las propiedades lineales y no lineales de manera eficiente, permitiendo que las redes neuronales aprendan representaciones complejas con mayor efectividad. A diferencia de la función ReLU, que puede sufrir de la "muerte de ReLU" (donde los gradientes se vuelven cero y las neuronas dejan de aprender), *Swish* es no-monótona y suaviza las transiciones, lo que puede conducir a una mejor

optimización y rendimiento. Estudios han demostrado que el uso de *Swish* como función de activación puede mejorar la precisión y la velocidad de convergencia en comparación con otras funciones de activación tradicionales.

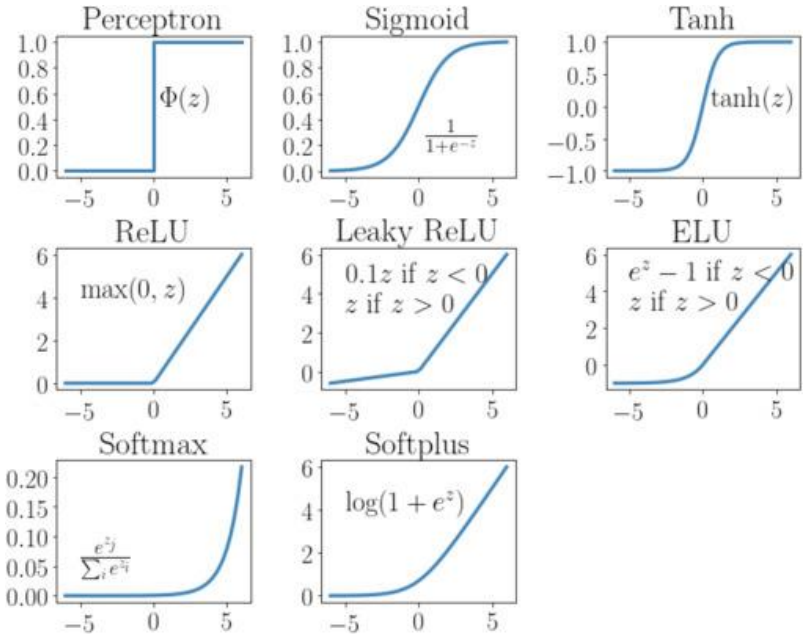


Figura 5. Funciones de activación comunes para redes neuronales artificiales que introducen no linealidades. Fuente: (Johnson et al., 2020)

Dropout

El *dropout* es una técnica de regularización utilizada en redes neuronales para prevenir el sobreajuste. Consiste en desactivar aleatoriamente un porcentaje de neuronas durante el entrenamiento, lo que obliga a la red a no depender excesivamente de ninguna neurona en particular y a aprender características más robustas y generalizables. Este proceso se ve esquematizado en la Figura 6.

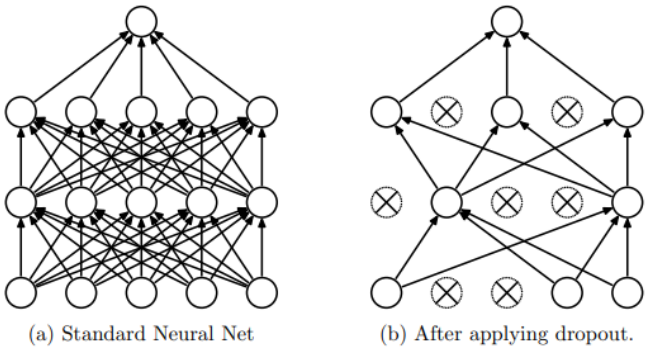


Figura 6. Modelo de red neuronal con *dropout*. (a) Una red neuronal estándar con 2 capas ocultas. (b) Un ejemplo de una red producida al aplicar el *dropout* a la red de la izquierda. Fuente: (Srivastava et al., 2014)

Capas Completamente Conectadas (*Fully Connected Layers*)

Al final de la red, las capas completamente conectadas toman los mapas de características y producen la salida final. Cada neurona en estas capas está conectada a todas las neuronas de la capa anterior, similar a una red neuronal tradicional, característica que les da el nombre.

Fortalezas y debilidades

Una de las principales ventajas de las CNNs es que pueden aprender y extraer automáticamente características relevantes de las imágenes sin necesidad de intervención humana, lo que facilita el procesamiento y análisis de datos visuales complejos. Gracias a las operaciones de convolución, estas redes pueden detectar características de manera consistente, independientemente de su ubicación en la imagen, proporcionando invarianza espacial. Además, las capas de pooling y las convoluciones locales reducen significativamente la cantidad de parámetros en comparación con una red completamente conectada, disminuyendo el riesgo de sobreajuste.

Por otro lado, las CNNs requieren grandes cantidades de datos etiquetados para entrenarse efectivamente, lo cual puede ser un desafío en términos de disponibilidad, etiquetado y calidad de datos. También cabe destacar, que el entrenamiento y la inferencia de CNNs son intensivos en términos de recursos computacionales, necesitando hardware especializado como GPUs. Aunque en menor medida que otras redes profundas, las CNNs también pueden ser consideradas "cajas negras" debido a la dificultad de interpretar cómo se toman las decisiones internas.

2.3.3. Técnicas avanzadas de aprendizaje profundo

Aprendizaje por transferencia y ajuste fino

El aprendizaje por transferencia (*Transfer Learning*) es una técnica en aprendizaje automático donde un modelo entrenado en una tarea se reutiliza como punto de partida para un modelo en una segunda tarea relacionada. El ajuste fino (*fine-tuning*) es el proceso de entrenar un modelo preentrenado en un nuevo conjunto de datos para adaptarlo a una tarea específica. Esta técnica aprovecha el conocimiento adquirido por un modelo en una tarea inicial, generalmente con un gran conjunto de datos, y lo aplica a una tarea diferente pero relacionada, reduciendo significativamente el tiempo de entrenamiento y los recursos computacionales necesarios (Zhuang et al., 2020).

El transfer learning funciona mediante el preentrenamiento de un modelo, como una red neuronal profunda, en un gran conjunto de datos genéricos, para aprender características generales. Las capas del modelo preentrenado, especialmente las capas iniciales que capturan características básicas se transfieren a una nueva tarea. En el ajuste fino, las capas finales del modelo se ajustan o se reemplazan para adaptarse a la nueva tarea específica, permitiendo que el modelo completo sea entrenado nuevamente con un conjunto de datos más pequeño y específico.

En la **Figura 7** se puede ver resumido el proceso. El modelo entrenado mediante transfer learning aprovecha el modelo superior como extractor de características y le añade un nuevo bloque para adaptarse a la nueva tarea, de manera que el modelo preentrenado actúa como extractor de características.

El ajuste fino se centra en las capas superiores del modelo, ya que estas capturan características más especializadas y relevantes para la tarea original. Durante el ajuste fino, solo estas capas finales se reentrenan o se ajustan, mientras que las capas iniciales, que capturan características más generales, suelen permanecer fijas o con un grado mínimo de ajuste. Este enfoque permite que el modelo se adapte rápidamente a las nuevas características de los datos sin requerir un entrenamiento completo desde cero, aprovechando el conocimiento adquirido previamente y mejorando la eficiencia del proceso.

El transfer learning y el ajuste fino presentan varias fortalezas. Al reutilizar un modelo preentrenado, se reduce el tiempo y los recursos computacionales necesarios para entrenar el modelo desde cero. Además, esta técnica es especialmente útil cuando se dispone de un conjunto de datos pequeño para la nueva

tarea, ya que el modelo ya ha aprendido características generales en el preentrenamiento. Los modelos preentrenados pueden ser fácilmente adaptados a una variedad de tareas diferentes, mejorando su versatilidad.

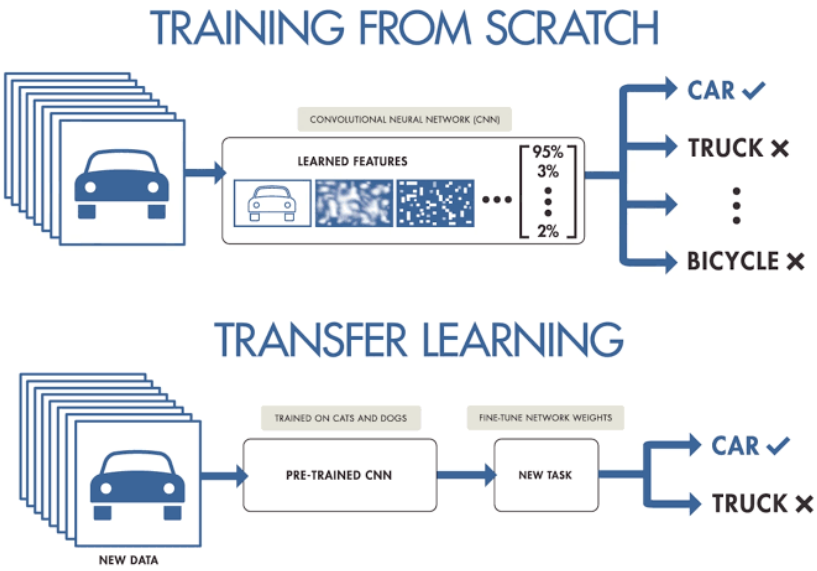


Figura 7. En la parte superior muestra un modelo estrenado desde cero. En la parte inferior se muestra un modelo entrenado mediante transfer learning. Fuente: (Saha, 2018)

Sin embargo, también existen debilidades en estos métodos. El rendimiento del modelo final depende de la calidad y la relevancia del modelo preentrenado y los datos originales en los que se entrenó. Si las características de los datos del conjunto de datos inicial son muy diferentes de las del nuevo conjunto de datos, el rendimiento puede no ser óptimo. Además, si el ajuste fino no se maneja cuidadosamente, el modelo puede sobreajustarse a la nueva tarea, perdiendo la generalización adquirida durante el preentrenamiento.

Bloques de atención

Los bloques de atención se llaman así porque "atienden" a las características más relevantes dentro de los datos procesados por la red neuronal, similar a cómo una persona puede enfocarse en los aspectos más importantes de una escena visual. En esencia, estos bloques recalibran las características aprendidas, asignando más peso a las partes cruciales y menos a las irrelevantes. Fueron propuestas por Hu et al. (2017) (Hu et al., 2017)

Uno de los mecanismos de atención más empleados son los bloques de *Squeeze and Excitation* (SE), que consisten en dos etapas principales: *squeeze* (comprimir) y *excitation* (excitar). En la fase de compresión, se reduce la dimensionalidad de las características para capturar información global. Luego, en la fase de excitación, se recalibran las características aprendidas asignando diferentes pesos a cada una, de modo que las características más importantes se enfatizan mientras que las menos importantes se suprimen.

Funcionamiento

Operación de compresión o *Squeeze*

En la fase de compresión, se aplica una operación de compresión global a la salida de una capa convolucional para generar una descripción de características compacta. Esto se realiza mediante *global average pooling*, que reduce cada canal de características a un solo valor, preservando la información

de las características a nivel global. Esta compactación permite que la red tenga una visión general de las características más importantes de la imagen. Esto se resumen matemáticamente en la **Fórmula 1**.

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_{i,j,c}$$

Fórmula 1. Operación de *Squeeze.*

Donde z_c es el valor comprimido para el canal c_i y $x_{i,j,c}$ es el valor de la característica en la posición (i,j) del canal c

Operación de excitación o *Excitation*

En la fase de excitación, se aplica una red completamente conectada de dos capas a los valores comprimidos para capturar las dependencias entre canales y generar un conjunto de pesos de activación. Estos pesos se utilizan para recalibrar los canales de características originales, enfatizando las características más importantes y suprimiendo las menos relevantes. Este proceso queda resumido en la **Fórmula 2**.

$$s = \sigma(W_2 \delta(W_1 z))$$

Fórmula 2. Operación de Excitation.

Donde δ es una función de activación ReLU, σ es una función sigmoide, W_1 y W_2 son matrices de peso, y z es el vector de características comprimidas.

Recalibración de características

Finalmente, los pesos de activación generados se utilizan para recalibrar las características originales mediante una operación de multiplicación por canal, enfatizando las características más importantes. Esto asegura que la red neuronal presta mayor atención a las características más relevantes, mejorando así su rendimiento en tareas de clasificación y reconocimiento. Podemos ver un esquema visual de este proceso en la **Figura 8**.

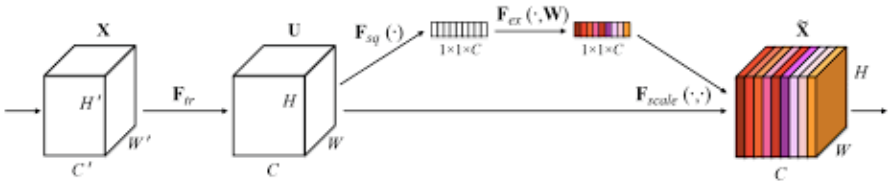


Figura 8. Esquema del proceso *Squeeze and Excitation* Fuente: (Prove, 2017)

Fortalezas y debilidades

Los bloques SE presentan diversas fortalezas. Una de sus principales ventajas es la mejora del rendimiento, ya que pueden ser integrados en cualquier arquitectura CNN existente, incrementando significativamente la precisión en tareas de clasificación y detección de objetos. Además, su eficiencia computacional es notable, ya que añaden una mínima complejidad computacional, involucrando solo operaciones de pooling y redes completamente conectadas. También destacan por su capacidad de

adaptación, permitiendo que la red aprenda a enfatizar dinámicamente las características relevantes, mejorando su capacidad para generalizar en diferentes tareas y conjuntos de datos.

Sin embargo, los bloques SE también tienen algunas debilidades. A pesar de que la sobrecarga de parámetros es mínima, la adición de bloques SE introduce algunos parámetros adicionales que pueden aumentar la complejidad del modelo. Además, la eficacia de los bloques SE puede depender de la calidad y diversidad del conjunto de datos utilizados para el entrenamiento, lo que puede limitar su desempeño en ciertos escenarios.

2.4. Estado del arte

La inteligencia artificial está siendo investigada en distintos campos de la medicina. Entre las principales aplicaciones a este campo podemos encontrar el análisis de imágenes médicas, secuenciación genómica, medicina personalizada, cirugía robotizada y automatización de tareas administrativas. Se espera que el papel de la IA en la atención sanitaria evolucione hacia modelos predictivos personalizados, gestión de enfermedades crónicas, desarrollo de fármacos y salud pública de precisión. Sin embargo, garantizar la confianza, la transparencia, la equidad y la ética será crucial a medida que los humanos y la IA se asocien cada vez más para maximizar los beneficios de la atención sanitaria (Arefin, 2024)

Algunas de estos algoritmos ya están siendo aplicados en el entorno clínico real, es el caso del artículo publicado en Nature por Sandbank et al. (2022) (Sandbank et al., 2022), se detalla el desarrollo, validación y despliegue de un algoritmo de inteligencia artificial (IA) para la detección de patología mamaria en imágenes completas de diapositivas (WSIs) de biopsias de mama. El algoritmo, llamado "*Galen Breast*," fue implementado en diciembre de 2019 en el Instituto de Patología de Maccabi Healthcare Services (MHS) demostrado ser exitosa en la práctica clínica, proporcionando mejoras significativas en la precisión diagnóstica y la eficiencia del flujo de trabajo, sin añadir una carga considerable al tiempo de trabajo de los patólogos.

Los modelos preentrenados, como ResNet, Inception y VGG, se utilizan comúnmente y se ajustan para resolver problemas específicos en dermatología. Estos modelos se entrenan en grandes conjuntos de datos de imágenes y luego se afinan para tareas específicas utilizando *transfer learning*, mejorando la precisión sin necesidad de recopilar enormes cantidades de datos médicos específicos.

- En (Madheswari & Karthikeyan, 2024) se evaluó el uso de arquitecturas de *deep learning*, como VGG-16, Inception V3 e Inception ResNet V2, para la detección temprana del cáncer de piel. Mediante la transferencia de aprendizaje y el ajuste fino, se mejora significativamente la precisión del diagnóstico, demostrando que estos modelos preentrenados pueden ser implementados en entornos clínicos para apoyar a los dermatólogos en la detección temprana y el tratamiento del cáncer de piel.
- En (Mohammed et al., 2024) se analiza la aplicación de redes neuronales convolucionales para la clasificación de imágenes médicas, destacando la integración de datos clínicos para mejorar la precisión del diagnóstico. Los modelos CNN ajustados con datos clínicos adicionales mostraron un mejor rendimiento que aquellos que solo usaban imágenes
- También en el campo de la imagen médica, pero en la enfermedad de Alzheimer, en (Huang et al., 2024) se exploran dos métodos para incorporar mecanismos de atención en las CNN para el diagnóstico de la enfermedad de Alzheimer. La incorporación de bloques de atención en las CNN resultó en una mejora significativa en la precisión del diagnóstico de la enfermedad de Alzheimer en comparación con las CNN estándar sin atención. Además, los modelos con

bloques de atención mostraron una reducción notable en la tasa de falsos positivos y negativos, lo cual es crucial para aplicaciones clínicas.

Sin embargo, existen desafíos y limitaciones. La calidad variable de las imágenes y la falta de datos diversos representan un desafío significativo. Además, muchos modelos tienen dificultades para generalizar fuera de los conjuntos de datos en los que fueron entrenados, especialmente con diferentes demografías o tipos de piel. En comparación con dermatólogos, varios estudios han mostrado que los modelos de IA tienen un rendimiento comparable o superior en la clasificación de condiciones de la piel. Estos modelos no solo pueden asistir en el diagnóstico, sino que también pueden mejorar los flujos de trabajo de triaje y referencia, ayudando a identificar casos que requieren atención urgente.

No obstante, la recopilación y el uso compartido de datos médicos plantean importantes desafíos éticos y de privacidad, lo cual puede limitar la disponibilidad de grandes conjuntos de datos necesarios para entrenar modelos robustos. La inclusión de metadatos del paciente, como la edad, el sexo y el historial médico, puede mejorar significativamente el rendimiento de los modelos de IA en dermatología. Aunque la mayoría de los modelos actuales se entrenan únicamente con imágenes, incorporar información clínica adicional podría llevar a mejores resultados de clasificación.

Existe un potencial significativo de las técnicas de aprendizaje profundo en dermatología, al tiempo que cabe subrayar la necesidad de abordar los desafíos relacionados con la calidad de los datos, la diversidad de los conjuntos de datos y la generalizabilidad de los modelos. Las contribuciones de este TFM al estado del arte se comentan en la sección *3.3 Justificación del proyecto.*

3. Definición del problema y requerimientos

3.1. Problema de investigación

El principal desafío que se abordará en el presente trabajo es el diagnóstico temprano y preciso de malignidad de lesiones cutáneas mediante el uso de imágenes dermatoscópicas y datos clínicos, que es el paso previo a la biopsia en la práctica. A pesar de los avances en técnicas de imagen como la dermatoscopia, la interpretación de las imágenes sigue dependiendo en gran medida de la experiencia y el juicio del dermatólogo, lo que puede conducir a variabilidad en los diagnósticos y, en ocasiones, a errores. Como se ha comentado anteriormente, en un artículo publicado en Nature por Esteva et al. (2017) (Esteva et al., 2017) se afirma que los dermatólogos tienen una tasa de precisión de entre el 65% y el 80% en inspección visual y hasta el 75% y el 84% mediante el uso de imágenes dermatoscópicas. Esta situación se agrava en entornos de atención primaria, donde los recursos especializados son limitados y el tiempo para la evaluación es restringido.

Para ello, se desarrollarán y evaluarán herramientas automatizadas eficaces de inteligencia artificial para apoyar la detección temprana y el cribado de lesiones cutáneas malignas, como el melanoma. Actualmente, las técnicas tradicionales requieren una gran pericia, son costosas y consumen mucho tiempo, lo que limita su aplicabilidad en entornos con alta carga de pacientes. Además, el uso de datos clínicos adicionales, como antecedentes médicos o características demográficas del paciente, rara vez se integra de manera eficiente con las imágenes dermoscópicas en un único modelo de diagnóstico.

Este trabajo propone desarrollar un sistema basado en inteligencia artificial que utilice aprendizaje profundo para mejorar la precisión y eficiencia del diagnóstico de lesiones cutáneas malignas. El objetivo es abordar algunas de las deficiencias claves en los sistemas actuales. El primero de estos es la limitada capacidad de integración de datos multimodales, ya que los enfoques actuales rara vez

combinan de manera eficaz datos clínicos y de imagen, lo que puede llevar a diagnósticos incompletos o imprecisos.

Además, la falta de herramientas de cribado automatizadas en la atención primaria limita la capacidad de los profesionales de la salud para realizar evaluaciones rápidas y precisas, lo que puede resultar en diagnósticos tardíos y un mayor riesgo de mortalidad. Finalmente, es importante destacar la variabilidad en la interpretación humana de imágenes dermoscópicas, que puede introducir subjetividades y posibles errores en la evaluación de las lesiones cutáneas.

La resolución de este problema tiene una relevancia significativa tanto a nivel clínico como social. Un sistema automatizado y preciso para el diagnóstico temprano de lesiones cutáneas malignas podría reducir significativamente la mortalidad asociada al melanoma y otros cánceres de piel, optimizando los recursos disponibles en los sistemas de salud. Al mejorar la eficiencia del cribado en atención primaria, se puede aliviar la carga sobre los dermatólogos, permitiendo que se centren en los casos más críticos y reduciendo así el tiempo de espera para los pacientes.

Este Trabajo de Fin De Máster, por tanto, se centra en desarrollar y validar una solución innovadora que aborde estas deficiencias a través de un enfoque basado en *Design Science*, creando un artefacto (modelo) que no solo sea técnicamente eficaz sino también sea potencialmente práctico para su implementación en un entorno real.

3.2. Requerimientos funcionales y no funcionales

Con el objetivo de centrar el desarrollo de la solución en las especificaciones del problema, a continuación, se listan los requerimientos funcionales y no funcionales del modelo. Los requerimientos funcionales definen que debe hacer el modelo para cumplir con su propósito, mientras que los no funcionales indican cómo debe funcionar, considerando aspectos como la seguridad, eficiencia y explicabilidad.

Requerimientos funcionales

- **Clasificación precisa de lesiones cutáneas**. Es decir, el modelo debe ser capaz de identificar correctamente lesiones cutáneas como benignas o malignas, basándose en imágenes dermoscópicas y datos clínicos relevantes del paciente. Esto implica alcanzar una alta sensibilidad y especificidad, particularmente en la detección temprana de melanomas.
- **Integración de datos multimodales**. Para aumentar la precisión del modelo y asemejar los flujos de trabajo humano, el modelo debe poder combinar de manera eficaz imágenes dermoscópicas con datos clínicos adicionales mejorando así la precisión del diagnóstico.
- **Explicabilidad del modelo**. Dada la novedad de algunos de los métodos empleados en IA, es de especial relevancia poder ganar la confianza de los médicos, es decir, que el modelo debe ser explicable. Esto significa que debe proporcionar información sobre las características y patrones específicos que está utilizando para clasificar cada lesión. Por ejemplo, debe ser capaz de destacar las áreas de las imágenes o las variables clínicas que influyen más en su diagnóstico, facilitando que los médicos entiendan y validen sus resultados.
- **Capacidad de cribado automatizado**. Para poder optimizar los procesos, debe ser capaz de procesar y analizar imágenes y datos clínicos de forma rápida y eficiente, lo que permite su uso en entornos de atención primaria con una alta carga de trabajo.

- **Escalabilidad**. Por último, el modelo debe poder manejar un creciente volumen de datos de imágenes y clínicos, adaptándose a diferentes contextos clínicos y manteniendo un rendimiento óptimo.

Requerimientos no funcionales

- **Fiabilidad y robustez.** El modelo debe ser confiable en su rendimiento, manejando datos incompletos o ruidosos sin errores significativos. Además, debe mantener su precisión a lo largo del tiempo y en diferentes conjuntos de datos.
- **Explicabilidad y transparencia.** Más allá de su precisión, el modelo debe ser interpretativo para los profesionales de la salud, explicando sus decisiones y facilitando la validación clínica de sus resultados. Esto es esencial para su aceptación e integración en entornos médicos.
- **Seguridad y privacidad de los datos**. Aunque el modelo en sí no implica una interfaz de usuario final, cualquier manejo de datos debe cumplir con las normativas de privacidad y protección de datos, como el GDPR en Europa. Esto incluye asegurar que los datos de entrenamiento y evaluación se gestionen de manera segura.
- **Eficiencia computacional**. El modelo debe ser eficiente en términos de tiempo de ejecución y uso de recursos computacionales, permitiendo su implementación en infraestructuras de hardware típicas de entornos clínicos, sin requerir grandes inversiones adicionales en infraestructura.

3.3. Justificación del proyecto

Este proyecto aporta al estado del arte actual en varios aspectos clave, diferenciándose de los estudios mencionados en la revisión del estado del arte en términos de enfoque, integración de datos, y aplicabilidad en la práctica clínica.

En primer lugar, a diferencia de muchos estudios previos que se han centrado exclusivamente en el uso de modelos preentrenados como VGG, ResNet o Inception, este proyecto desarrolla un enfoque multimodal que no solo utiliza imágenes dermoscópicas, sino que también integra datos clínicos relevantes del paciente, como la edad, el historial médico y otros factores demográficos. Aunque algunos trabajos, como el de Mohammed et al. (2024), han explorado la integración de datos clínicos, nuestro enfoque se diferencia al combinar ambos tipos de datos en un único modelo, diseñado específicamente para maximizar la precisión diagnóstica y mejorar la capacidad de generalización del sistema en diferentes contextos clínicos. Esto responde a la necesidad destacada en la literatura de mejorar los resultados de clasificación mediante la inclusión de información clínica adicional, una práctica que aún no ha sido suficientemente adoptada.

Además, este proyecto aborda las limitaciones señaladas en el estado del arte en cuanto a la calidad y la diversidad de los datos. Mientras que muchos estudios previos utilizan conjuntos de datos como ISIC, que presentan problemas de duplicidad y calidad variable de las imágenes, este trabajo emplea un riguroso proceso de preprocesado y limpieza de datos para eliminar duplicados siguiendo las directrices establecidas y asegurar un conjunto de datos más robusto y representativo.

Además, este proyecto incorpora bloques de atención en las arquitecturas de redes neuronales convolucionales, alineándose con las tendencias observadas en otros campos de la medicina, como el diagnóstico de la enfermedad de Alzheimer (Huang et al., 2024). Sin embargo, se diferencia al aplicar esta técnica específicamente al diagnóstico dermatológico, mejorando la capacidad del modelo para centrarse en características relevantes de las imágenes dermoscópicas, lo cual es crucial para reducir

falsos positivos y negativos en aplicaciones clínicas. Esta incorporación de mecanismos de atención no solo optimiza la precisión del diagnóstico, sino que también proporciona explicabilidad al modelo, lo que facilita su aceptación e integración en entornos clínicos.

Otro punto distintivo del enfoque escogido, es comparar el uso del conjunto de datos completo con un alto desequilibrio entre clases (<10% de imágenes malignas) frente un conjunto de datos menor pero más equilibrado (aproximadamente 45% de imágenes malignas). Por último, este proyecto también hace una contribución importante al estado del arte al abordar la explicabilidad de los modelos, incluyendo mecanismos de atención.

En resumen, este proyecto aporta al estado del arte actual mediante el desarrollo de un modelo multimodal que integra eficazmente imágenes dermoscópicas y datos clínicos, abordando desafíos clave relacionados con la generalización, la eficiencia computacional y la aplicabilidad práctica, y ofreciendo una solución más robusta y adaptable para la detección temprana de melanomas en entornos clínicos.

3.3.1. Justificación del uso de redes neuronales convolucionales

El uso de redes neuronales convolucionales (CNN, por sus siglas en inglés) en el contexto del diagnóstico dermatológico está justificado por varias razones clave, especialmente debido a su capacidad para procesar y analizar imágenes médicas de manera eficiente y precisa.

Por un lado, se destaca la eficiencia en la detección de patrones complejos, ya que las CNN están diseñadas específicamente para trabajar con datos de imágenes. Aprovechan su arquitectura jerárquica para extraer automáticamente características relevantes a diferentes niveles de abstracción, desde patrones básicos como bordes y texturas, hasta características más complejas y específicas de las lesiones cutáneas. Esta capacidad de detectar y analizar detalles sutiles es esencial para diferenciar entre lesiones benignas y malignas que a menudo presentan características visuales muy similares.

También es relevante su capacidad de aprender de grandes volúmenes de datos. Las CNN se benefician enormemente de la disponibilidad de grandes conjuntos de datos de imágenes dermoscópicas. A través del entrenamiento en estos conjuntos de datos, las redes neuronales convolucionales pueden identificar patrones que serían difíciles de discernir incluso para dermatólogos experimentados. Esto permite un diagnóstico más preciso, consistente y escalable, particularmente en escenarios donde la variabilidad interobservador puede ser un problema.

Otra característica de las CNN es la versatilidad en la integración de datos multimodales: Además de su eficacia en el análisis de imágenes, las CNN pueden integrarse con otras técnicas de inteligencia artificial para combinar datos de diferentes modalidades, como imágenes dermoscópicas y datos clínicos (por ejemplo, edad del paciente, historial familiar, etc.). Esta integración mejora la precisión diagnóstica al aprovechar múltiples fuentes de información para tomar decisiones más informadas.

Por otro lado, otra de las características que favorecen su uso es la potencial automatización del proceso de diagnóstico. El uso de CNN podría permitir automatizar gran parte del proceso de diagnóstico de lesiones cutáneas, lo que es particularmente útil en entornos de atención primaria donde los recursos especializados son limitados. Al automatizar tareas que de otro modo serían manuales y lentas, las CNN pueden reducir significativamente el tiempo necesario para el cribado inicial de pacientes, permitiendo priorizar rápidamente los casos más críticos para una evaluación adicional.

Capacidad de explicar sus decisiones: Las arquitecturas modernas de CNN, combinadas con técnicas de interpretabilidad, como los mapas de calor o "*Grad-CAM*", permiten identificar qué características de las imágenes influyen más en las decisiones del modelo. Esta capacidad de proporcionar explicaciones visuales aumenta la confianza de los médicos en el sistema, ya que pueden validar las decisiones del algoritmo comparándolas con su propia experiencia clínica.

4. Metodología

4.1. Obtención de datos

Para la elaboración del presente trabajo, se han usado los conjuntos de datos de *International Skin Imaging Collaboration* (ISIC)[1]. El ISIC es una iniciativa global cuyo objetivo principal es mejorar el diagnóstico y la detección de enfermedades de la piel a través de la estandarización y el intercambio de imágenes dermatológicas (*ISIC- The International Skin Imaging Collaboration*, s. f.).

El ISIC organiza anualmente desafíos, conocidos como ISIC *Challenges*, cuyo objetivo principal es impulsar el desarrollo y la evaluación de algoritmos avanzados para el diagnóstico automatizado de enfermedades de la piel. Los ISIC *Challenges* se centran en tareas específicas, como la clasificación de lesiones cutáneas, la segmentación de imágenes y la detección de malignidad, entre otras. Los participantes tienen acceso a extensos *datasets* de imágenes etiquetadas, que incluyen información clínica detallada y diagnósticos verificados. A través de estos concursos, ISIC busca no solo fomentar la innovación en la inteligencia artificial aplicada a la dermatología, sino también establecer estándares de rendimiento y promover la colaboración entre diferentes disciplinas.

Los conjuntos de datos de ISIC, en particular los de 2019 y 2020, son altamente valiosos ya que incluyen miles de imágenes etiquetadas de diversas condiciones de la piel, desde lesiones benignas hasta malignas, junto con información clínica y diagnósticos verificados. Estos datasets son más completos en comparación con otros disponibles, ya que integran imágenes de diferentes fuentes y proporcionan una diversidad significativa en términos de tipos de lesiones y características demográficas, lo que los hace ideales para el entrenamiento y la validación de modelos de inteligencia artificial en dermatología.

El dataset utilizado en este estudio se construyó combinando los datasets de ISIC 2019 y 2020. Este enfoque permite aprovechar la riqueza y diversidad de datos disponibles en ambos años, garantizando una base de datos robusta para el entrenamiento de los modelos. Siguiendo las directrices de (Cassidy et al., 2022), que incluyen la eliminación de duplicados identificados a través de las listas proporcionadas por los propios ISIC Challenges, las cuales indican explícitamente las imágenes duplicadas y los conjuntos de datos de origen. Este proceso permitió obtener un conjunto de datos unificado con 58,032 observaciones y 9 características, de las cuales 3 son variables objetivo, la descripción detallada de estas se incluye en el capítulo “Descripción de datos”. En la Figura 9 se puede ver un ejemplo de imágenes aleatorias obtenidas de este conjunto de datos.

Los datos incluyen imágenes de lesiones cutáneas etiquetadas con información relevante sobre el paciente, como identificadores únicos de imagen y paciente, sexo, edad aproximada, y el sitio anatómico de la lesión, entre otras variables de diagnóstico. Estas características se describen en detalle en la sección *4.2 Descripción de datos*.

[1] https://www.isic-archive.com/

Este dataset presenta ciertos desafíos, como la presencia de datos faltantes en variables críticas que se mencionan en la sección *Completitud del conjunto de datos* que afectan principalmente a las observaciones de lesiones benignas. Además, los datos contienen imágenes con artefactos, como cabello, sombras y reflejos, que complican la clasificación automática. Estos aspectos se abordan mediante técnicas de preprocesado que se detallan en la sección ***¡Error! No se encuentra el origen de la referencia.***: ***¡Error! No se encuentra el origen de la referencia.***.

En comparación con otros conjuntos de datos, el dataset de ISIC fue elegido por su nivel de detalle y amplitud, así como por su reconocimiento y uso extendido en la comunidad de investigación. Su naturaleza estandarizada y su integración de datos provenientes de múltiples fuentes lo hacen particularmente adecuado para la tarea de clasificación de lesiones cutáneas, proporcionando una base sólida para el desarrollo y la validación de los modelos propuestos.

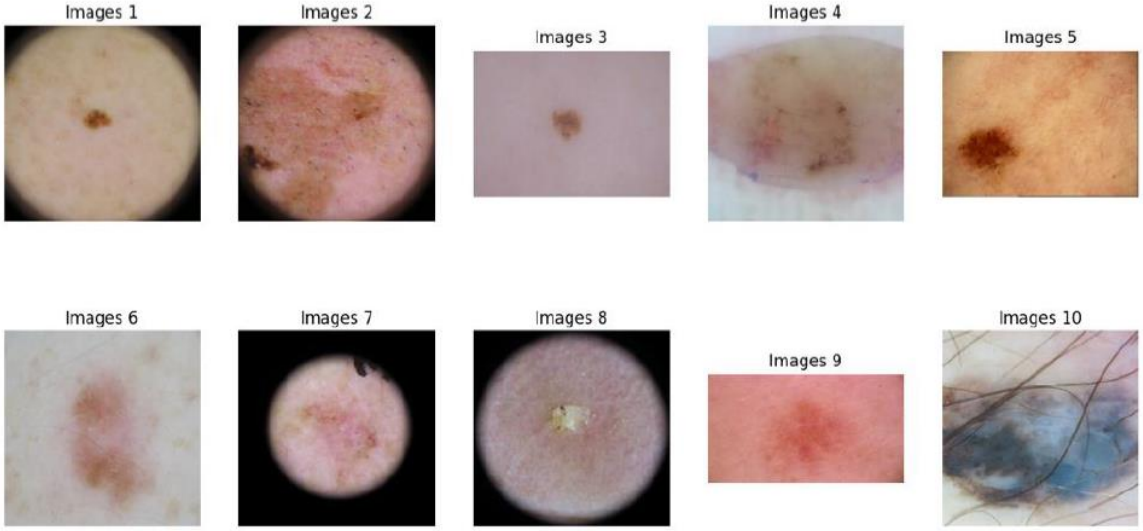


Figura 9. Ejemplos de imágenes extraídas del conjunto de datos combinados de ISIC Challenge 2019 y 2020.

4.2. Descripción de datos

4.2.1. Características generales

El conjunto de datos de trabajo cuenta con 58.032 observaciones y 9 características de las cuales 3 pertenecerían a las variables objetivo. Estas características son:

- *image_name*: identificador único de la imagen.
- *Lesión_id*: Identificador único de lesión.
- *Patient_id*: Identificador único de paciente.
- *Sex*: sexo del paciente donde *male* representa hombre y *female* mujer
- *Age_approx*: Variable numérica discreta que representa la edad aproximada del paciente
- *Anatom_site_general_challenge*: variable categórica que indica el lugar anatómico de la lesión. Los posibles valores son:
  - *Anterior torso*: torso anterior
  - *Lateral torso*: torso lateral
  - *Head/neck*: cabeza y cuello
  - *Upper extremity*: extremidades superiores.
  - *Lower extremity*: extremidades inferiores

- *Oral/genital*: oral y genital
- *Palms/soles*: palmas de las manos y suela de los pies
- *Posterior torso*: torso posterior.
- *Torso*: torso

❖ *Diagnosis*: Una de las variables objetivo, de tipo categórica. La Figura 10 muestra ejemplos de imágenes por tipo de diagnóstico. Los posibles valores que puede tener son:

- *atypical melanocytic proliferation:* proliferación melanocítica atípica
- *cafe-au-lait macule:* mácula café con leche
- *lentigo NOS:* Lentigo
- *lichenoid keratosis:* queratosis liquenoide
- *melanoma:* melanoma
- *nevus:* nevo
- *seborrheic keratosis:* Queratosis seborreica
- *solar lentigo: l*entigo solar
- *unknown:* desconocido.

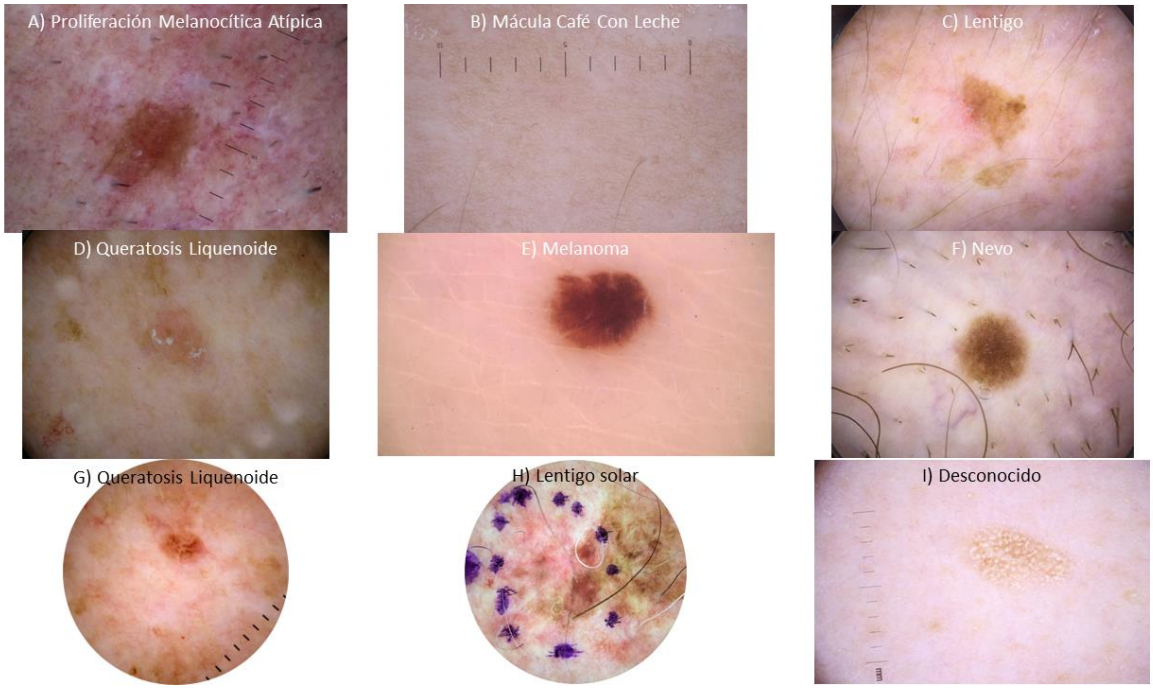


Figura 10. Ejemplo de imágenes por cada diagnóstico de la base de datos. Fuente: propia

❖ *benign_malignant*: Variable objetivo categórica, donde *benign* representa benigno y *malignant* maligno.

❖ *Target*: Variable objetivo binaria {0,1} donde 0 representa benigno y 1 representa maligno

Completitud del conjunto de datos

El análisis de completitud del conjunto de datos empleado revela información crucial sobre la integridad y calidad de los datos disponibles. A continuación, se comentarán los hallazgos principales.

En primer lugar, cabe destacar que el formato de los datos clínicos ha variado en los conjuntos de datos de 2019 y 2020. En ambos casos el identificador único de imagen es *image_name*, este será el nexo único entre los archivos de imágenes y las variables clínicas.

La variable *anatom_site_general_challenge*, que indica el sitio anatómico de la lesión, muestra una completitud del 94,56%, con 3.157 valores faltantes. De estos, 3.021 corresponden a lesiones benignas (5,50% de benignos) y 136 a lesiones malignas (10,77% de malignos). Por otro lado, la variable target, que es la variable objetivo binaria (0 para benigno, 1 para maligno), tiene una completitud del 100%, estando presente en todos los registros.

El conjunto de datos, las variables *image_name* y *target* no presentan valores faltantes, mostrando una completitud del 100%. Las variables *sex* y *age_approx* tienen una completitud muy alta, con un 99,27% y 99,17% respectivamente, aunque aún presentan una pequeña cantidad de valores faltantes. La variable *image_name*, que actúa como el identificador único de cada imagen, muestra una completitud del 100%, estando presente en todos los registros.

4.2.2. Análisis de imágenes dermoscópicas

Se dispone de 58.032 imágenes de lesiones cutáneas identificadas mediante el identificador único de imagen. Son archivos de tipo jpg de diversos tamaños y tomados con diferentes dispositivos. A modo representativo, se han cogido 100 imágenes del conjunto de datos de entrenamiento y se ha analizado el tamaño, como podemos ver en la Figura 11 las imágenes son de diferentes dimensiones, resaltando la importancia del preprocesado previo al entrenamiento del modelo.

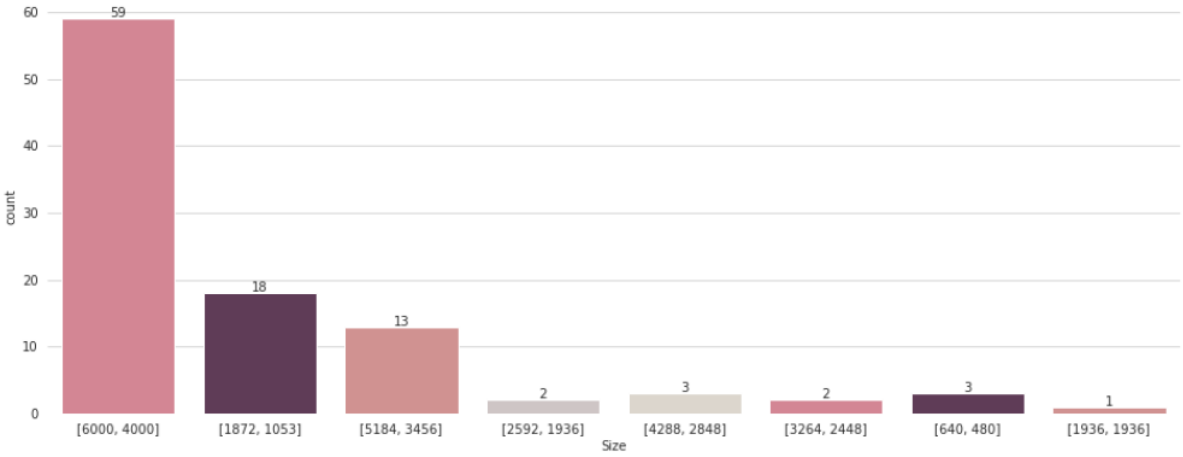


Figura 11. Distribución de tamaños en una muestra de 100 imágenes aleatoriamente seleccionadas.

En la Figura 12, se puede ver una muestra de imágenes benignas escogidas aleatoriamente. Estas presentan variaciones en el brillo y el contraste. Algunas imágenes son más claras y brillantes, mientras que otras son más oscuras y tienen menos contraste. Esto puede deberse a la iluminación utilizada durante la captura de las imágenes o a las características de la piel del paciente. Hay una amplia gama de colores en las imágenes benignas. Predominan los tonos de piel, que varían desde tonos rosados claros hasta marrones más oscuros.

Algunas imágenes muestran áreas con pigmentación más intensa. Las imágenes benignas también muestran variaciones en el encuadre. Algunas están bien centradas en la lesión, mientras que otras incluyen más áreas de piel circundante. Además, la presencia de vello puede afectar la visibilidad de la lesión en algunas imágenes. En general, hay una heterogeneidad significativa en las imágenes benignas en términos de brillo, color y encuadre. Esta variabilidad puede representar un desafío para los algoritmos de análisis de imágenes que dependen de la consistencia en estas características.

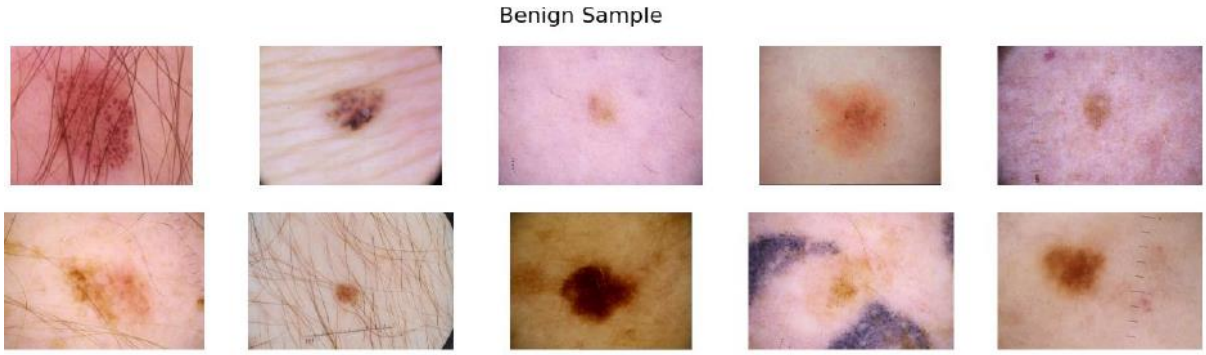


Figura 12. Muestra de imágenes benignas aleatorias del conjunto de datos.

Estas características se repiten en la Figura 13, donde vemos representada una muestra aleatoria de imágenes malignas. Aunque en general, las imágenes malignas presentan una variedad de colores, con una tendencia hacia tonos más oscuros y áreas con pigmentación irregular.

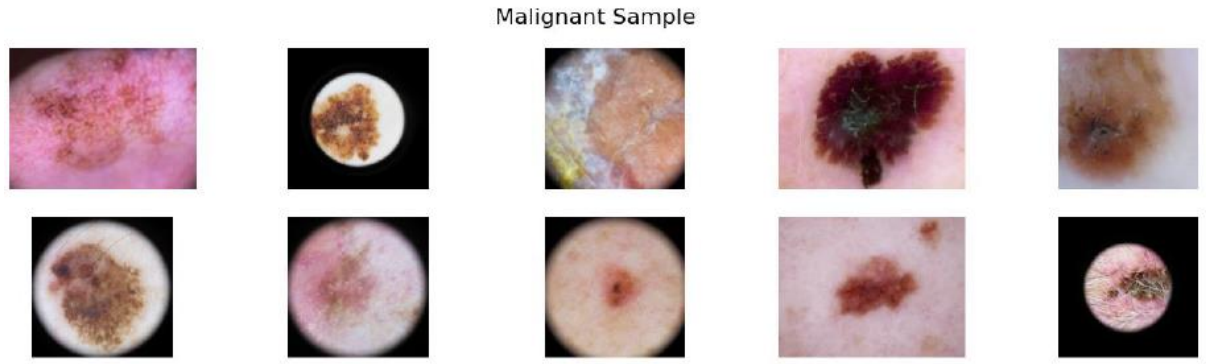


Figura 13. Muestra de imágenes malignas aleatorias del conjunto de datos.

La variación en el brillo y el contraste entre las imágenes benignas y malignas es notable. Esta heterogeneidad puede deberse a diferentes condiciones de iluminación y configuraciones de la cámara durante la captura de las imágenes. Esta falta de uniformidad puede ser un desafío para el análisis automatizado de imágenes. Además, la presencia de vello en algunas imágenes puede interferir con la visibilidad de las lesiones y afectar el análisis.

Además de la heterogeneidad presente por el uso de diferentes dispositivos y diferentes profesionales realizando las imágenes, añadimos una capa de complejidad debida a artefactos, es decir, propiedades recurrentes que aumentan la complejidad del dataset, además de la variabilidad intrínseca del órgano de la piel.

Entre los artefactos más comunes, se encuentran imágenes con formas de lesiones cutáneas irregulares, lesiones grandes que conectan múltiples límites de la imagen y lesiones con bajo contraste con la piel circundante. También se observan lesiones con artefactos de cartas de colores, capilares, tinta de rotulador y reglas de medir. Además, hay lesiones con artefactos de vasos sanguíneos y burbujas de gel, así como imágenes con ruido de viñeta. En algunos casos, las lesiones presentan múltiples artefactos simultáneamente y exhiben múltiples tonos de intensidad de color. Asimismo, se encuentran pequeñas lesiones cutáneas que añaden otra capa de variabilidad al conjunto de datos.

Esta diversidad y la presencia de diversos artefactos resaltan la necesidad de técnicas avanzadas de preprocesamiento y normalización para desarrollar algoritmos de análisis de imágenes efectivos y

precisos. Estas técnicas se explican en detalle en la sección 4.3. ***¡Error! No se encuentra el origen de la referencia.***.

4.2.3. Análisis exploratorio de variables clínicas

ANÁLISIS UNIVARIADO

Variable sex

La variable *sex* representa el sexo del paciente del que proviene la lesión. Este dato está disponible en 57.607 observaciones, lo que representa el 99,27 % del total, dejando un total de 425 datos faltantes, equivalentes al 0,73 %. De estos 425 datos faltantes, 344 pertenecen a la categoría de lesiones benignas, representando un 0,65 % de los casos benignos, mientras que los 81 datos faltantes restantes corresponden a la categoría de malignos y representan un 6,74 % de los registros malignos del conjunto de datos. Si observamos el histograma de la variable, en la **Figura 14**, podemos ver que el sexo de los pacientes se encuentra equilibrado, con una predominancia de hombres que representan el 52,35 % de la variable sex.

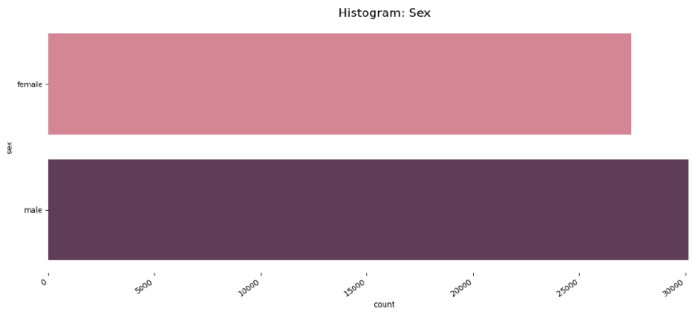


Figura 14. Histograma de la variable sex. Fuente: propia

Variable age_approx

La variable *age_approx* representa la edad aproximada del paciente del que proviene la lesión. Este dato está disponible en 57.607 observaciones, lo que representa el 99,27 % del total, con un total de 425 datos faltantes, equivalentes al 0,73 %. De estos 425 datos faltantes, 396 pertenecen a la categoría de lesiones benignas, representando un 0,69 % de los casos benignos. Los 85 datos faltantes restantes corresponden a la categoría de malignos y representan un 6,54 % de los registros malignos del conjunto de datos.

Esta abarca desde aproximadamente 0 hasta 90 años. La distribución, mostrada en la **Figura 28**, muestra una amplia gama de edades, indicando la presencia de grupos de edad diversos en el conjunto de datos. El pico más alto se encuentra alrededor de los 35 años, señalando una mayor concentración de pacientes en sus treinta y tantos. La distribución parece ser relativamente simétrica alrededor de este pico de 35 años, lo que sugiere una representación equilibrada de edades por encima y por debajo de este punto. La densidad disminuye constantemente desde el pico hacia grupos de edad más jóvenes y mayores, con una notable disminución en el número de observaciones a medida que la edad aumenta más allá de los 50 años.

La presencia de puntos de datos en los extremos (cerca de 0 y 90 años) podría indicar valores atípicos potenciales o grupos de edad menos frecuentes en el conjunto de datos. Esta variabilidad en la distribución de edades resalta la importancia de considerar la edad como un factor relevante en el análisis y modelado de los datos.

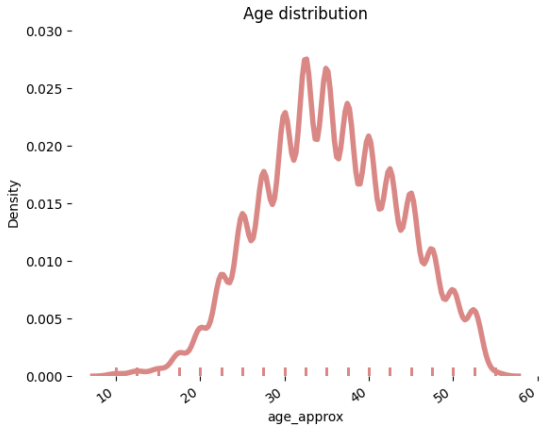


Figura 15. Distribución de la variable age_approx.

Variable anatom_site_general_challenge

La variable `*anatom_site_general_challenge*` representa el sitio anatómico donde se encuentra la lesión cutánea. Este dato está disponible en 54.875 observaciones, lo que representa el 94,55 % del total, con un total de 3.157 datos faltantes, equivalentes al 5,45 %. De estos 3.157 datos faltantes, 3.021 pertenecen a la categoría de lesiones benignas, representando un 5,50 % de los casos benignos. Los 136 datos faltantes restantes corresponden a la categoría de malignos y representan un 10,77 % de los registros malignos del conjunto de datos.

El histograma de la **Figura 29** muestra el recuento de observaciones para diferentes sitios anatómicos donde se localizan las lesiones cutáneas. Los sitios anatómicos incluyen torso anterior, extremidad superior, torso posterior, extremidad inferior, torso lateral, cabeza/cuello, palmas/plantas, oral/genital y torso. El sitio anatómico más común es el torso, con el mayor recuento de observaciones, superando las 16.000. Le sigue la extremidad inferior, con un recuento cercano a las 14.000.

Las extremidades superiores y cabeza/cuello también tienen recuentos sustanciales, cada uno superando las 6.000 observaciones. Las palmas/plantas y la zona oral/genital tienen los recuentos más bajos, lo que indica que las lesiones en estas áreas son menos comunes en el conjunto de datos. El torso lateral también tiene un recuento relativamente bajo en comparación con otros sitios anatómicos. Esta variabilidad en la distribución de los sitios anatómicos resalta la importancia de considerar la ubicación de las lesiones como un factor relevante en el análisis y modelado de los datos.

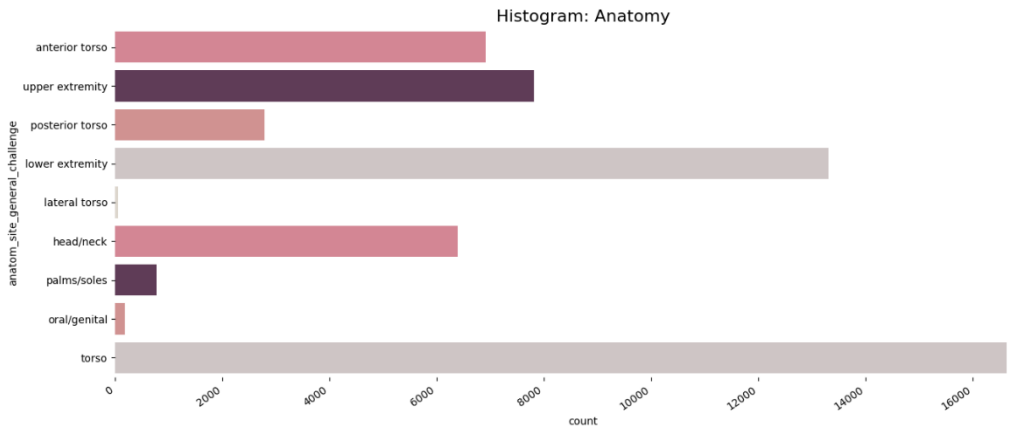


Figura 16. Histograma de valores apra la variable anatom_site_general_challenge. Fuente: propia

ANÁLISIS MULTIVARIADO

Para profundizar en la relación entre diferentes variables del conjunto de datos y obtener una comprensión más detallada de cómo interactúan entre sí, se ha realizado un análisis bivariado. El objetivo de este análisis es identificar y evaluar las asociaciones y patrones que pueden existir entre pares de variables. Este análisis es fundamental porque permite detectar interacciones significativas que pueden mejorar la precisión y efectividad de los modelos predictivos, así como proporcionar información valiosa para el desarrollo de estrategias de diagnóstico.

Edad según el diagnóstico (*age_approx* frente a *anatom_site_general_challenge*)

Para entender mejor la relación entre la edad del paciente y los diferentes diagnósticos de lesiones cutáneas, se realizó un análisis detallado utilizando técnicas de visualización y pruebas estadísticas.

En la Figura 17B, se puede observar un *box plot* que muestra la distribución de la edad según el diagnóstico, incluyendo la mediana, los cuartiles y los posibles valores atípicos. La mediana de edad varía según los diferentes diagnósticos, con algunas categorías presentando rangos más amplios. En la Figura 17A se combina un *box plot* con un gráfico de densidad de *kernel* para mostrar la forma de la distribución de edades para cada categoría de diagnóstico. La densidad de las distribuciones de edad proporciona información adicional sobre la concentración de edades dentro de cada diagnóstico.

La categoría “desconocido” presenta un amplio rango de edades, pero se centra en una mediana de edad más alta. Por otro lado, nevus muestra un rango de edad relativamente estrecho en comparación con otras categorías. Si nos fijamos en melanoma, se observa una distribución más amplia con una mediana de edad más alta, sugiriendo que los pacientes de mayor edad son más comúnmente diagnosticados con melanoma. Otros diagnósticos como 'lentigo NOS', 'queratosis liquenoide', 'lentigo solar', 'mancha café-au-lait' y 'proliferación melanocítica atípica' muestran distribuciones de edad variables, destacando la diversidad en la población de pacientes.

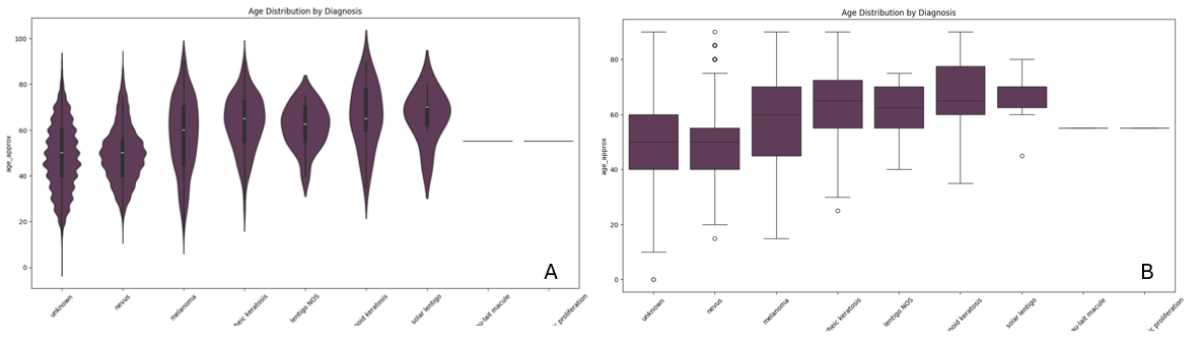


Figura 17. A) Gráfico de violín que representa la distribución de edad por diagnóstico. B) Box plot representando la distribución de edad por diagnóstico. Fuente: propia.

Los resultados del test ANOVA indican un estadístico de 87,824 y un valor p de 1,0437e-109, significativamente inferior al nivel alfa común de 0,05. Este resultado sugiere que existen diferencias estadísticamente significativas en la edad media entre las diferentes categorías de diagnóstico. Por lo tanto, se confirma que las edades medias para al menos algunas de las categorías de diagnóstico son significativamente diferentes entre sí. Esto se evidencia en las visualizaciones, donde categorías como 'nevus' y 'melanoma' tienen diferentes edades medianas y distribuciones de edad.

Edad frente a malignidad

Para entender mejor la relación entre la edad del paciente y la malignidad de las lesiones cutáneas, se realizó un análisis comparativo.

Si observamos la Figura 18, podemos ver la distribución de la edad para lesiones benignas y malignas, incluyendo la mediana, los cuartiles y los posibles valores atípicos. Tanto las lesiones benignas como las malignas abarcan un amplio rango de edades, desde aproximadamente 0 hasta más de 80 años. Las distribuciones cubren un espectro amplio de edades, indicando que las lesiones pueden ocurrir en todos los grupos etarios. La mediana de edad para las lesiones malignas parece ser ligeramente mayor que la de las lesiones benignas. Las lesiones benignas tienen una mediana de edad alrededor de los 50 años, mientras que las lesiones malignas tienen una mediana de edad alrededor de los 55 años.

El rango intercuartil (IQR) para ambas lesiones, benignas y malignas, es similar, lo que indica una dispersión comparable de edades dentro del 50 % central de los datos. Sin embargo, el IQR para las lesiones malignas es ligeramente más amplio, sugiriendo más variabilidad en las edades para los casos malignos. Hay algunos valores atípicos en los grupos de edad más jóvenes tanto para las lesiones benignas como malignas. Estos valores atípicos podrían representar casos raros de pacientes muy jóvenes con lesiones cutáneas.

Los resultados del test T indican un valor p de 0,0, significativamente menor que el nivel alfa común de 0,05. Esto indica que la diferencia en las edades medias entre las lesiones benignas y malignas es estadísticamente significativa. El estadístico t de -43,153 es muy grande en magnitud, lo que respalda aún más la presencia de una diferencia sustancial entre los dos grupos. El estadístico t negativo sugiere que la edad media para las lesiones malignas es mayor que para las lesiones benignas. Esto se alinea con las observaciones del *box plot*, donde la mediana de edad para los casos malignos era ligeramente mayor que para los casos benignos.

Este análisis resalta la importancia de considerar la edad del paciente en el diagnóstico y manejo de las lesiones cutáneas, ya que las diferencias en las distribuciones de edad entre las lesiones benignas y malignas pueden proporcionar pistas adicionales para una evaluación más precisa.

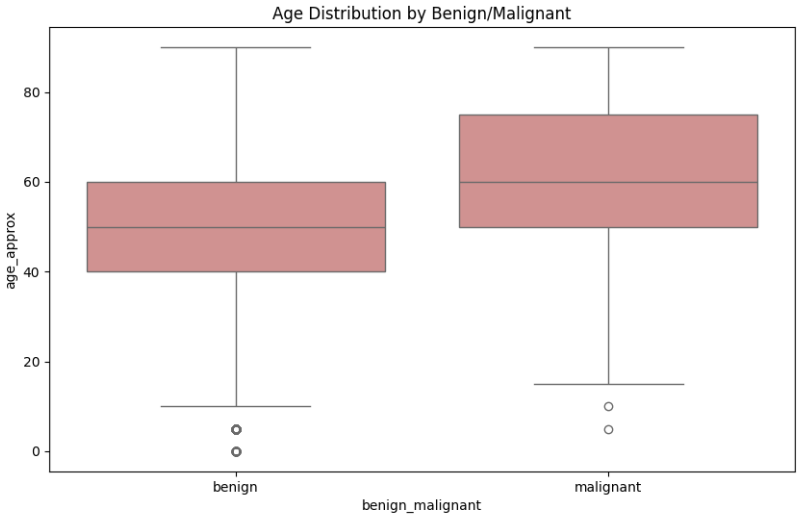


Figura 18. Box plot que representa edad frente a malignidad de la lesión. Fuente: propia.

Diagnósticos frente a sitio anatómico

Para comprender mejor la relación entre el sitio anatómico de las lesiones cutáneas y sus diagnósticos, se realizó un análisis detallado que se muestra en la **Figura 19**. Esta figura ilustra el conteo de diferentes diagnósticos en varios sitios anatómicos, incluyendo categorías como 'desconocido', 'nevus', 'melanoma', 'queratosis seborreica', 'lentigo NOS', 'queratosis liquenoide', 'lentigo solar', 'mancha café-au-lait' y 'proliferación melanocítica atípica'.

La categoría 'desconocido' presenta los conteos más altos en múltiples sitios anatómicos, especialmente en el torso y las extremidades. 'Nevus' también muestra una presencia significativa en varios sitios, notablemente en la extremidad superior y cabeza/cuello. El torso anterior muestra una distribución diversa de diagnósticos con conteos notables para 'nevus' y 'queratosis seborreica'. En contraste, las palmas/suelas tienen pocas observaciones en los diagnósticos, sugiriendo que las lesiones en estas áreas son raras, al igual que las zonas oral/genital.

Los resultados del Test Chi-Cuadrado indican un estadístico de 605,124 y un valor p de 4,7588e-102, significativamente menor que el nivel alfa de 0,05. Esto indica una asociación estadísticamente significativa entre el sitio anatómico y el diagnóstico. El alto valor del estadístico chi-cuadrado respalda aún más la presencia de una relación significativa. Esta asociación sugiere que ciertos diagnósticos son más probables en sitios anatómicos específicos.

Estos resultados destacan la importancia de considerar el sitio anatómico como una característica clave en los modelos predictivos para diagnosticar lesiones cutáneas. Incorporar esta variable puede mejorar la precisión y fiabilidad de los modelos, ayudando a identificar patrones específicos que pueden ser cruciales para una evaluación y tratamiento adecuados.

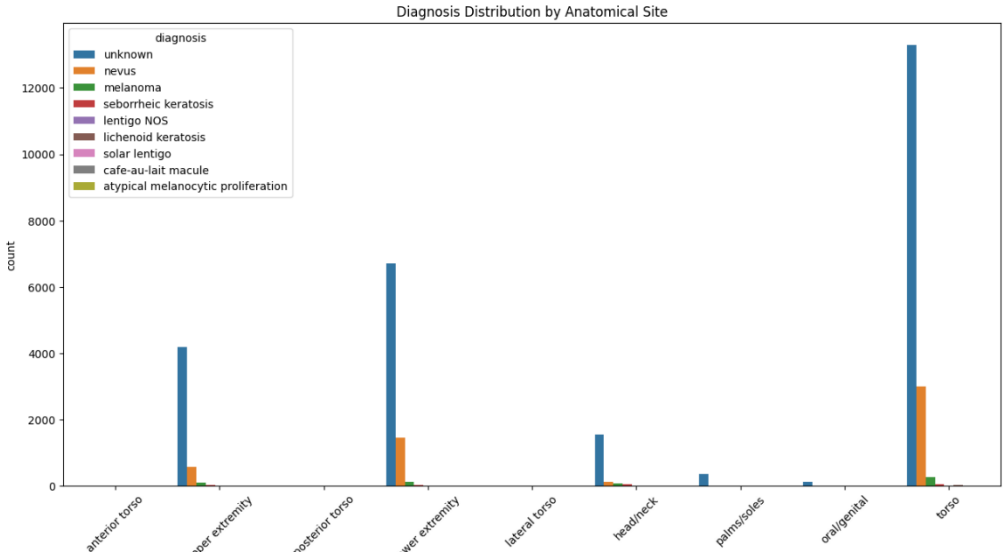


Figura 19. Recuento de diagnóstico por lugar anatómico de la lesión. Fuente: propia

Sitio anatómico frente a malignidad

Para analizar la relación entre el sitio anatómico de las lesiones cutáneas y su estado benigno o maligno, se realizó un estudio detallado, cuyos resultados se muestran en la Figura 20. Esta figura ilustra el conteo de lesiones benignas y malignas en varios sitios anatómicos.

El torso tiene el conteo más alto de lesiones tanto benignas como malignas, con un conteo significativamente mayor de lesiones benignas. La extremidad inferior también presenta un alto conteo

de lesiones benignas, pero relativamente menos lesiones malignas. En la mayoría de los sitios anatómicos, las lesiones benignas son más comunes que las malignas. Sin embargo, la proporción de lesiones malignas es relativamente mayor en áreas como la cabeza/cuello y la extremidad superior en comparación con otros sitios.

Los resultados del Test Chi-Cuadrado indican un estadístico de 2827,113 y un valor p de 1,2e-10, significativamente menor que el nivel alfa de 0,05. Esto indica una asociación estadísticamente significativa entre el sitio anatómico y el estado benigno/maligno de las lesiones. El alto valor del estadístico chi-cuadrado respalda aún más la presencia de una relación significativa. Esta asociación sugiere que ciertos sitios anatómicos son más propensos a tener lesiones benignas o malignas.

Estos resultados reafirman la importancia de considerar el sitio anatómico como una característica clave en los modelos predictivos para clasificar las lesiones como benignas o malignas. Incluir el sitio anatómico en el modelo predictivo puede mejorar la precisión y fiabilidad de los modelos, resaltando la asociación significativa con el estado benigno/maligno. Esta perspectiva puede contribuir significativamente a la mejora de los diagnósticos y tratamientos de las lesiones cutáneas, proporcionando una base sólida para decisiones clínicas más informadas.

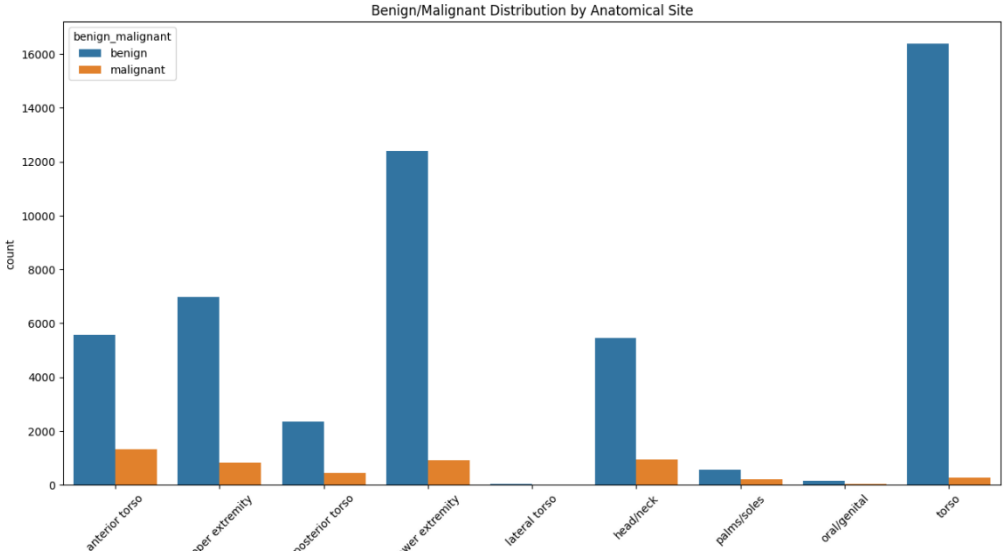


Figura 20. Gráfico de barras que muestra la distribución de lesiones malignas por sitio anatómico. Fuente: propia.

4.3. Distribución y particionado del conjunto de datos

El conjunto de datos combinado de los retos de 2019 y 2020 contienen un total de 58.032 imágenes, siendo 5.103 imágenes pertenecientes a la clase objetivo “maligna”, representando un 8,79% del total de los datos, como se puede ver en la Figura 21.

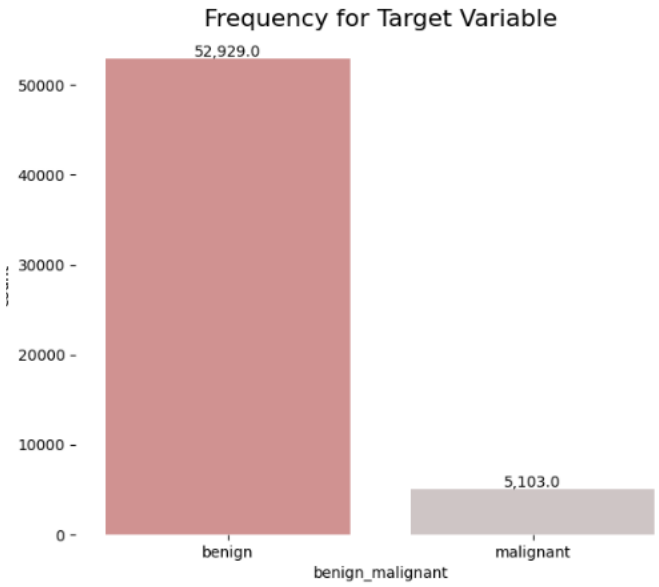


Figura 21. Gráfico de barras que muestra la frecuencia de casos malignos y benignos. Fuente: propia.

Las representaciones mostradas en la Figura 22 muestran cómo se mantiene más o menos equilibrada la distribución de género entre lesiones malignas y benignas.

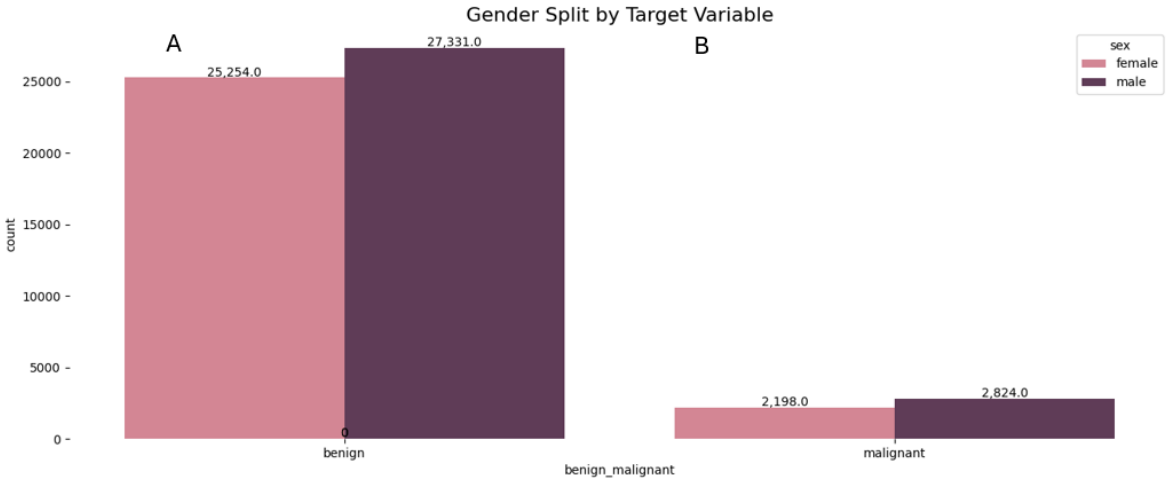


Figura 22. Distribución de sexo en el subconjunto de datos benigno (A) y maligno (B). Fuente: propia.

El análisis bivariado de la densidad de casos benignos y malignos en función de la edad revela patrones significativos en la distribución de las lesiones cutáneas (Figura 23). El gráfico de densidad de casos benignos, sombreado en azul, muestra que las lesiones benignas son más comunes en personas más jóvenes y de mediana edad, alcanzando un pico alrededor de los 50 años y disminuyendo más allá de esta edad. También se observan algunos casos benignos en el grupo de edad muy joven (menos de 10 años).

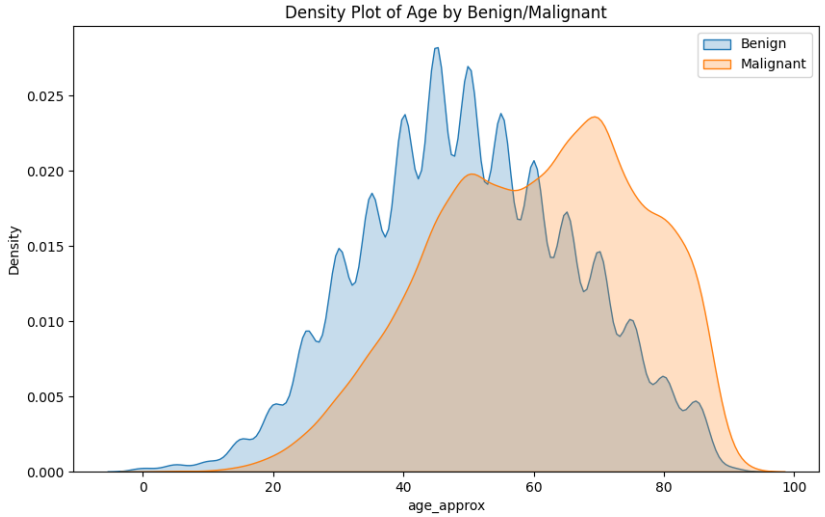


Figura 23. Gráfico de densidad de edad en los subconjuntos benigno (azul) y maligno (naranja. Fuente: propia

Por otro lado, el gráfico de densidad de casos malignos, sombreado en naranja, indica que las lesiones malignas son más prevalentes en personas mayores en comparación con las benignas. La distribución de los casos malignos alcanza su punto máximo alrededor de los 60 años y muestra una presencia notable en personas de 40 a 80 años. Hay menos casos malignos en los grupos de edad muy jóvenes y muy mayores.

La superposición entre las dos distribuciones sugiere que existen rangos de edad donde ocurren con frecuencia lesiones tanto benignas como malignas, especialmente entre los 40 y 60 años. Sin embargo, los casos malignos tienden a ser más prevalentes en personas mayores, mientras que los casos benignos se distribuyen de manera más uniforme en un rango de edad más amplio.

Los distintos picos y la extensión de las distribuciones sugieren que la edad puede ser una característica útil para distinguir entre lesiones benignas y malignas, especialmente en grupos de mayor edad. Esta información es valiosa para el desarrollo de modelos predictivos y para la toma de decisiones clínicas, ya que resalta la importancia de considerar la edad del paciente al evaluar la probabilidad de que una lesión cutánea sea maligna.

A modo de resumen, el análisis exploratorio de datos (EDA) sobre las variables clínicas, reveló una importante heterogeneidad en las imágenes y confirmó la necesidad de manejar cuidadosamente los datos faltantes y los artefactos en las imágenes. Además, se observó un desbalanceo de clases significativo.

El EDA también destacó la importancia de incluir variables como la edad y el sitio anatómico en los modelos predictivos para mejorar su precisión y fiabilidad. Además, será crucial para el modelo manejar los valores faltantes y atípicos adecuadamente. Debido al notable desbalanceo de datos

4.3.1. Creación de conjunto de datos equilibrado

El análisis exploratorio de datos ha revelado una distribución desbalanceada entre las lesiones benignas y malignas en el conjunto de datos completo. Este desbalance puede afectar negativamente el rendimiento de los modelos de DL, ya que los algoritmos tienden a ser sesgados hacia la clase mayoritaria, en este caso, las lesiones benignas. Para abordar esta problemática, se ha creado un nuevo conjunto de datos “equilibrado” donde la proporción de imágenes benignas y malignas es similar.

La creación de un conjunto de datos balanceado implica tomar todas las imágenes de lesiones malignas y una selección aleatoria de 6000 entradas de lesiones benignas. Este enfoque tiene varias justificaciones:

Mejora en la generalización del modelo. Un conjunto de datos balanceado permite que el modelo de DL aprenda de manera equitativa sobre ambas clases, evitando el sesgo hacia la clase mayoritaria. Esto mejora la capacidad del modelo para generalizar y desempeñarse bien en datos no vistos, asegurando que las predicciones sean precisas tanto para lesiones benignas como malignas.

Reducción del sesgo del modelo. Los modelos de DL entrenados en conjuntos de datos desbalanceados tienden a predecir la clase mayoritaria con mayor frecuencia, ignorando la clase minoritaria. Al balancear el conjunto de datos, se garantiza que el modelo tenga suficientes ejemplos de ambas clases para aprender, reduciendo así el sesgo y mejorando la equidad de las predicciones.

Aumento de la sensibilidad para la clase minoritaria. En el contexto médico, es crucial que los modelos sean sensibles a las lesiones malignas, ya que un diagnóstico incorrecto puede tener consecuencias graves. Un conjunto de datos balanceado ayuda a aumentar la sensibilidad del modelo hacia la detección de lesiones malignas, mejorando así su capacidad para identificar correctamente casos críticos.

Facilitación del entrenamiento y optimización. Un conjunto de datos balanceado simplifica el proceso de entrenamiento y optimización del modelo. Los algoritmos de DL pueden converger más rápidamente y con mayor estabilidad, lo que resulta en modelos más robustos y eficientes. La Tabla 1 resumen las características del conjunto de datos equilibrado.

Tabla 1. Resumen de la distribución de datos de entrenamiento y test (Holdout) del conjunto de datos balanceado.

	Malignas	Benignas	Total
Train	4.082 (45,96 %)	4.800 (54,04 %)	8.882 (80,00 %)
Test	1.021 (45,97 %)	1.200 (54,03 %)	2.221 (20,00 %)
Total	5.103 (45,96 %)	6.000 (54,04 %)	11.103

A modo de comparación, se ejecutó un entrenamiento bajo las mismas condiciones con la diferencia del uso del conjunto de datos y la función de pérdidas. Por un lado, se empleó el conjunto de datos original con función de pérdida Focal Loss, esta se explica en detalle en las secciones de

Diseño e

Implementación, pero a grandes rasgos, aborda el problema del desbalance de clases, enfocándose más en los ejemplos de la clase minoritaria al reducir la contribución de los ejemplos fáciles a la pérdida total. En contraste, la *BCE with Logits Loss (Binary Cross-Entropy*) trata todos los ejemplos por igual sin dar importancia especial a los ejemplos de la clase minoritaria.

El experimento de comparación entre el conjunto de datos completo y el conjunto de datos equilibrado se justifica para abordar los desafíos inherentes a la distribución desbalanceada de las clases en el conjunto original, donde las lesiones benignas dominan. Este desbalance puede sesgar los modelos de aprendizaje profundo hacia la clase mayoritaria, reduciendo su capacidad de identificar correctamente las lesiones malignas, que son de mayor importancia clínica. Utilizar Focal Loss con el conjunto de datos completo permite mitigar parcialmente este problema, al ajustar la pérdida para enfocarse más en los ejemplos difíciles de la clase minoritaria. Sin embargo, al crear un conjunto de datos equilibrado y

aplicar la función de pérdida BCE with Logits Loss, se asegura que ambos tipos de lesiones estén igualmente representados durante el entrenamiento.

4.3.2. Preprocesado de imágenes

Como se ha indicado anteriormente, uno de los artefactos más comunes y que pueden dificultar la clasificación de lesiones cutáneas es la presencia de vello en las imágenes. Para eliminar el pelo, se ha utilizado el algoritmo *DullRazor* (Toossi et al., 2013) que aplica transformaciones sobre la imagen para crear una máscara binaria y reducir la interferencia del vello.

La eliminación del vello tiene como objetivo mejorar la calidad de las imágenes al reducir el ruido visual, lo que podría facilitar una mejor extracción de características y, por tanto, mejorar la precisión de los modelos de clasificación automática. Se puede ver resumido el proceso de obtención de máscara binaria y consecuente aplicación en la Figura 24.

En la Figura 24 podemos ver el proceso de obtención de la máscara binaria con su consiguiente eliminación de la imagen. Primero, la imagen se convierte a escala de grises, destacando las características esenciales sin la interferencia de la información de color. Se aplica una operación morfológica de *blackhat* para resaltar el vello, obteniendo una imagen donde los pelos aparecen más prominentes sobre un fondo oscuro. A continuación, se genera una máscara binaria mediante el umbralado de la imagen de *blackhat*, identificando claramente las regiones ocupadas por el vello. Finalmente, esta máscara se utiliza para eliminar el vello de la imagen original, reemplazando los píxeles correspondientes con valores interpolados de los alrededores, resultando en una imagen limpia y sin la interferencia del vello.

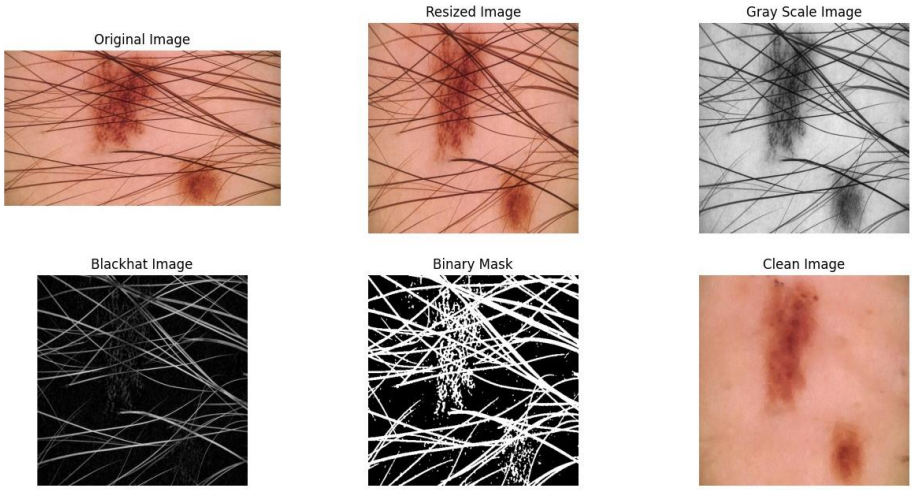


Figura 24. Proceso de aplicación del algoritmo DullRazor para eliminar el vello de las lesiones.

Para preparar las imágenes antes de entrenar los modelos de clasificación de lesiones cutáneas, se ha implementado un conjunto de transformaciones que varían según la arquitectura probada. Estas transformaciones incluyen técnicas de aumento de datos (*data augmentation*) y normalización, ajustadas específicamente para mejorar el rendimiento y la robustez de los modelos.

Tras la aplicación de DullRazor en el conjunto de datos global, la aplicación de aumento de datos de las imágenes en el conjunto de entrenamiento incluye una serie de transformaciones definidas para aumentar la variabilidad de los datos y mejorar la capacidad del modelo para generalizar. En primer

lugar, se aplica una transformación de `*RandomResizedCrop*`, que recorta aleatoriamente una porción de la imagen y la redimensiona al tamaño deseado, que varía con la arquitectura de entrada del modelo a probar, con una escala de entre 80% y 100% de la imagen original. Luego, se aplican `*RandomHorizontalFlip*` y `*RandomVerticalFlip*`, que voltean la imagen de manera aleatoria en direcciones horizontal y vertical, respectivamente. Estas transformaciones ayudan a que el modelo sea más robusto a variaciones en la orientación y escala de las lesiones cutáneas. Posteriormente, las imágenes se convierten a tensores y se normalizan usando valores de media y desviación estándar calculados a partir del conjunto de datos, lo que ayuda a estandarizar los valores de píxeles y mejorar la estabilidad del entrenamiento.

Para el conjunto de pruebas finales, las transformaciones se simplifican para garantizar que las imágenes se mantengan consistentes y comparables. Se aplica una transformación de `*Resize*` para ajustar todas las imágenes a al tamaño de entrada de la red en cuestión, seguida de la conversión a tensores y una normalización para estandarizar los valores de píxeles.

El uso de técnicas de *data augmentation*, como las aplicadas en este proyecto, es crucial para mejorar el rendimiento del modelo. *Data augmentation* aumenta artificialmente el tamaño del conjunto de datos mediante la creación de variaciones de las imágenes existentes. Esto no solo ayuda a prevenir el sobreajuste, ya que el modelo se expone a una mayor diversidad de ejemplos durante el entrenamiento, sino que también mejora la capacidad del modelo para generalizar a nuevas imágenes que no ha visto antes. Al introducir variaciones en la escala, orientación y otras propiedades de las imágenes, se entrena al modelo para que sea más robusto y capaz de manejar las variaciones naturales que ocurren en las imágenes del mundo real.

4.3.3. Preprocesado de variables clínicas

El análisis de las variables clínicas mediante técnicas de aprendizaje automático (ML) requiere un preprocesado cuidadoso de los datos para asegurar la precisión y robustez de los modelos. Dado que en el análisis exploratorio de datos se identificó la presencia de datos faltantes y la variabilidad en las características de los datos, es esencial abordar estos aspectos para mejorar la calidad del conjunto de datos.

La imputación de datos faltantes, la normalización de las variables y la transformación de datos categóricos son pasos críticos que ayudan a minimizar el impacto de las inconsistencias y a mejorar la capacidad predictiva de los modelos ML. Este preprocesado permite que los algoritmos de aprendizaje automático manejen mejor las peculiaridades de los datos clínicos, asegurando una evaluación más precisa y confiable de las variables en el contexto de la clasificación de lesiones cutáneas.

Las variables clínicas seleccionadas para el desarrollo del modelo de clasificación han sido la edad aproximada del paciente, el sexo y la ubicación anatómica de la lesión. Para la imputación de valores faltantes en la variable 'sexo', se ha utilizado la moda, que es la categoría más frecuente. La elección de la moda es adecuada para la imputación de variables categóricas, ya que preserva la distribución original de los datos, asegurando que la proporción de cada categoría se mantenga lo más fiel posible a la realidad.

En cuanto a la variable 'edad', se ha empleado la mediana para imputar los datos faltantes. La mediana es menos sensible a los *outliers* en comparación con la media, proporcionando un valor central que no se ve afectado por valores extremadamente altos o bajos. Esto es crucial para mantener la representatividad de la variable sin distorsionar su distribución central. Además, se ha realizado un

tratamiento de *outliers* mediante *Winsorización*, limitando los valores extremos a los percentiles 5% y 95%. Esta técnica permite reducir el impacto de los valores anómalos sin eliminar datos, ajustando los valores extremos a límites más razonables, lo cual preserva la mayor parte de la información mientras se controla la variabilidad.

Adicionalmente, se ha llevado a cabo una transformación logarítmica de la variable 'edad' para reducir su variabilidad y asegurar que la distribución sea más normal. Esta transformación es beneficiosa para manejar distribuciones sesgadas y reducir el impacto de valores extremos, lo que puede mejorar el rendimiento de los modelos de aprendizaje automático. La normalización posterior de los datos, ajustándolos a una media de 0 y una desviación estándar de 1, es particularmente importante para los modelos de aprendizaje profundo, ya que mejora la velocidad de convergencia y la precisión del modelo.

Finalmente, las variables 'sexo' y 'ubicación anatómica' se han transformado a variables binarizadas, lo que facilita su uso en modelos de aprendizaje automático. Esto ha resultado en un conjunto final de características de entrada para el modelo, que incluye: *'age_approx_scaled'*, *'anatom_site_unknown'*, *'anatom_site_general_challenge_head/neck'*, *'anatom_site_general_challenge_lateral torso'*, *'anatom_site_general_challenge_lower extremity'*, *'anatom_site_general_challenge_oral/genital'*, *'anatom_site_general_challenge_palms/soles'*, *'anatom_site_general_challenge_posterior torso'*, *'anatom_site_general_challenge_torso'*, *'anatom_site_general_challenge_upper extremity'* y *'sex'*. Estos pasos de preprocesamiento aseguran que los datos sean adecuados para el análisis y que los modelos puedan aprender de manera efectiva a partir de ellos.

4.4. Materiales

4.4.1. Python

En este estudio se utilizaron varias herramientas y bibliotecas de Python, un lenguaje de programación de código abierto, interpretado, orientado a objetos y de alto nivel con semántica dinámica. Python es muy apreciado para el desarrollo rápido de aplicaciones gracias a sus estructuras de datos de alto nivel y su tipificación y vinculación dinámicas. Python admite módulos y librerías, fomentando la segmentación del programa y la reutilización del código. El intérprete de Python y su extensa biblioteca estándar están disponibles sin coste alguno para las principales plataformas y pueden distribuirse libremente. Las bibliotecas de Python, que son colecciones de módulos relacionados, simplifican y facilitan la programación, desempeñando un papel crucial en campos como el aprendizaje automático, la ciencia de datos y la visualización de datos (*What is Python? Executive Summary*, s. f.).

PyTorch

Una de las bibliotecas utilizadas en este proyecto es PyTorch, una biblioteca de aprendizaje automático y aprendizaje profundo de código abierto desarrollada por Facebook, Inc. PyTorch proporciona una alternativa más rápida a NumPy gracias a su buen uso de las GPU y una plataforma flexible y rápida para el aprendizaje profundo. Las ventajas de usar PyTorch frente a otras bibliotecas como TensorFlow o Keras incluyen un entorno más flexible, aunque con una automatización ligeramente reducida, lo que permite crear arquitecturas más complejas y tener más flexibilidad en el entrenamiento (Chirodea et al., 2021).

Otras bibliotecas

Scikit-learn es otra biblioteca clave utilizada en este proyecto. Es una biblioteca de aprendizaje automático de código abierto que soporta el aprendizaje supervisado y no supervisado, y proporciona

diversas herramientas para el ajuste de modelos, el preprocesamiento de datos, la selección de modelos y la evaluación de modelos. Se hace hincapié en la facilidad de uso, el rendimiento, la documentación y la coherencia de la API. Tiene dependencias mínimas y se distribuye bajo la licencia BSD simplificada, lo que fomenta su uso tanto en entornos académicos como comerciales. En este estudio, se emplearon en especial los módulos para la evaluación del modelo, como sklearn.metrics, que implementa varias funciones de pérdida, puntuación y utilidad para medir el rendimiento de la clasificación (Pedregosa et al., 2011).

NumPy y **Pandas** son otras dos bibliotecas fundamentales utilizadas en este proyecto. NumPy es esencial para la computación científica en Python, proporcionando un objeto de matriz multidimensional, varios objetos derivados y un surtido de rutinas para realizar operaciones rápidas con matrices. Pandas es un paquete de código abierto utilizado sobre todo para tareas de ciencia de datos, análisis de datos y aprendizaje automático. Está construido sobre el paquete NumPy, que proporciona soporte para matrices multidimensionales. Pandas se usó en este estudio para integrar los datos tabulares gracias a una de sus estructuras de datos, el DataFrame, que es una estructura de datos bidimensional etiquetada con columnas de tipos potencialmente diferentes.

Matplotlib es una biblioteca fundamental para la creación de gráficos en Python. Matplotlib es muy flexible y personalizable, lo que la hace ideal para la creación de gráficos tanto simples como complejos. Se utiliza frecuentemente en combinación con otras bibliotecas como Seaborn para mejorar la presentación visual de los datos.

Seaborn es una biblioteca para la visualización de datos basada en Matplotlib. Proporciona una interfaz de alto nivel para crear gráficos estadísticos atractivos y complejos de manera sencilla. Seaborn es especialmente útil para explorar y entender mejor los datos, ya que incluye herramientas para el análisis de distribuciones, relaciones y categorizaciones.

PIL (*Python Imaging Library*), ahora mantenida como *Pillow*, es una biblioteca que proporciona capacidades para abrir, manipular y guardar muchos formatos de archivos de imagen. PIL es especialmente útil en el procesamiento de imágenes, permitiendo realizar tareas como redimensionamiento, recorte, rotación y conversión de formatos, entre otras.

4.4.2. Kaggle

Kaggle es una plataforma en línea que facilita la realización de proyectos de ciencia de datos y aprendizaje automático. Este se basa en las libretas Jupyter, que son entornos interactivos de computación en los que se pueden crear y compartir documentos que contienen código en vivo, ecuaciones, visualizaciones y texto explicativo. Estas libretas son ideales para el análisis de datos y el desarrollo de modelos de aprendizaje automático.

Una de las ventajas de Kaggle es el acceso a GPUs, que aceleran significativamente el entrenamiento de modelos de aprendizaje profundo. Además, Kaggle permite la subida fácil de bases de datos grandes, lo que facilita el acceso y uso compartido de datos en proyectos colaborativos. Esta combinación de recursos y herramientas hace de Kaggle una plataforma poderosa para el desarrollo y la implementación de soluciones de ciencia de datos y aprendizaje automático.

GPU

Para la ejecución de los diferentes modelos, se han usado las GPU disponibles en *Kaggle*, específicamente del tipo T4X2. La GPU T4x2 es un recurso de procesamiento gráfico de alto rendimiento diseñado para acelerar tareas intensivas en cómputo, como el entrenamiento de modelos de

DL y el procesamiento de grandes volúmenes de datos. Esta GPU pertenece a la familia NVIDIA Turing, conocida por su eficiencia energética y su capacidad para manejar cargas de trabajo complejas.

En el contexto de *Kaggle*, la GPU T4x2 permite a los usuarios ejecutar sus *notebooks o* libretas con una potencia de procesamiento significativamente mayor que la de las CPUs tradicionales. Esto es especialmente útil para proyectos que involucran redes neuronales profundas, donde el tiempo de entrenamiento puede reducirse drásticamente. La T4x2 proporciona un entorno ideal para desarrollar y probar modelos complejos sin la necesidad de invertir en hardware costoso.

Además, *Kaggle* facilita el acceso a estas GPUs mediante un entorno basado en libretas Jupyter, permitiendo a los usuarios cargar grandes bases de datos de manera sencilla y aprovechar la capacidad de procesamiento de la T4x2 para tareas de análisis y modelado de datos.

5. Diseño

5.1. Descripción general de los modelos

El diseño de los modelos se centra en desarrollar y evaluar varios modelos de aprendizaje profundo para la clasificación de lesiones cutáneas mediante el uso de imágenes dermatoscópicas y variables clínicas. Para ello se ha evaluado el poder clasificatorio de cada modalidad de datos (imágenes y datos clínicos) por separado, para posteriormente combinar ambas fuentes de información en un modelo multimodal. Finalmente se han utilizado técnicas de ensamblado de modelos para mejorar el rendimiento del sistema global.

Se distinguirán los modelos por el tipo de datos empleados para la clasificación, siendo estos:

- **Modelos basados en imágenes dermoscópicas.** Emplean como dato de entrada las imágenes de las lesiones distinguidas por el identificador único de imagen. En esta linea se han comparado el uso de distintas distribuciones de imágenes y diferentes arquitecturas avanzadas.
    - **Distintas distribuciones de imagen.** Se ha trabajado inicialmente con un conjunto de datos de imágenes donde la clase de lesiones malignas estaba subrepresentada. Tras comprobar que el uso de un conjunto de datos equilibrado (donde las lesiones malignas representan aproximadamente el 46% del total) mejoraba significativamente la sensibilidad del modelo, se decidió utilizar este conjunto balanceado en los experimentos posteriores. Este ha sido una toma de decisión crítica que ha afectado al resto de modelos ya que ha definido la distribución de datos empleados.
    - **Distintas arquitecturas.** Se han probado varias arquitecturas de redes neuronales convolucionales (CNN) para clasificar las imágenes. Estas incluyen:
        - **VGG16**: Utilizada como modelo de base por su eficacia comprobada en tareas de clasificación de imágenes.
        - **SE-ResNet**: Una variante de ResNet que incorpora mecanismos de atención (*Squeeze-and-Excitation blocks*), lo que permite al modelo recalibrar dinámicamente las características más relevantes de las imágenes.
        - **EfficientNet (variantes B1 a B4)**: Seleccionada por su capacidad de escalar el tamaño del modelo de forma eficiente, probando varias variaciones para maximizar el rendimiento en la etapa de ensamblado.

- **Modelos basados en el uso de variables clínicas.** Estos modelos emplean como entrada las variables clínicas, y se han evaluado diferentes técnicas de ML y DL aplicado a variables clínicas.
    - **Métodos de ML**. En paralelo, se ha evaluado el potencial clasificatorio de las variables clínicas mediante métodos tradicionales de aprendizaje automático (ML), que son el estándar en la clasificación de datos tabulares.
    - **Métodos de DL**. Además, se ha utilizado una red neuronal artificial (ANN) similar a la que se empleará en el modelo multimodal. Esta ANN tiene la misma arquitectura de extracción de características y clasificador que se usará posteriormente para integrar ambas modalidades de datos, con el fin de observar su comportamiento específico con los datos clínicos.
- **Integración de imágenes y datos clínicos en un modelo multimodal**. En esta fase, se han combinado las imágenes dermoscópicas con las variables clínicas en una única red neuronal multimodal. Este modelo utiliza una arquitectura conjunta que integra las características aprendidas tanto por la CNN (a partir de las imágenes) como por la ANN (a partir de los datos clínicos), permitiendo una fusión efectiva de la información visual y clínica para mejorar la precisión diagnóstica.
- **Ensamblado de modelos.** Finalmente, se han aplicado técnicas de ensamblado para optimizar el rendimiento general del sistema. Estas técnicas incluyen:
    - Voto por mayoría: Donde la clase final se decide por la mayoría de votos entre los modelos.
    - Voto ponderado: En este caso, se asignan diferentes pesos a las salidas de cada modelo en función de su precisión.
    - Metamodelo utilizando stacking: Aquí, las probabilidades y clases predichas por los modelos individuales se utilizan como entradas para un metamodelo adicional, que aprende a combinar de manera óptima las decisiones de los modelos previos.

5.2. Arquitecturas de los modelos

5.2.1. Modelos basados en imágenes

Modelos basados en imágenes sin mecanismos de atención

Con el fin de crear un modelo de DL que empleara solo las imágenes de lesiones y con las limitaciones explicadas anteriormente, se han usado las técnicas de transferencia de aprendizaje y ajuste fino para desarrollar dicho modelo. Para adaptar la arquitectura de VGG16 al problema específico de la clasificación de lesiones cutáneas, se ha utilizado transfer learning (aprendizaje por transferencia) aprovechando los pesos preentrenados en el conjunto de datos ImageNet. Este enfoque permite que el modelo comience con un conocimiento preexistente de características visuales generales, lo que reduce significativamente el tiempo de entrenamiento y mejora la eficiencia al trabajar con un conjunto de datos más pequeño y específico. Posteriormente, se realiza un ajuste fino de las últimas capas de la red para especializarla en la tarea de clasificación binaria de lesiones cutáneas, optimizando así su capacidad para identificar patrones relevantes en las imágenes dermoscópicas y mejorar la precisión diagnóstica en el contexto clínico.

El desarrollo se enfoca en la implementación de una red neuronal convolucional basada en la arquitectura VGG16 (Simonyan & Zisserman, 2014) para la clasificación de imágenes de lesiones cutáneas entre benignas y malignas. La elección de VGG16 se justifica por su demostrado rendimiento

en tareas de clasificación y su arquitectura bien estructurada, que permite el uso efectivo de técnicas de aprendizaje por transferencia.

VGG16 es una arquitectura de CNN desarrollada por Karen Simonyan y Andrew Zisserman del *Visual Geometry Group* de la Universidad de Oxford. Esta arquitectura ganó el primer y segundo lugar en detección y clasificación de objetos en el desafío ImageNet de 2014. Y su arquitectura se puede ver resumida en la Figura 25.

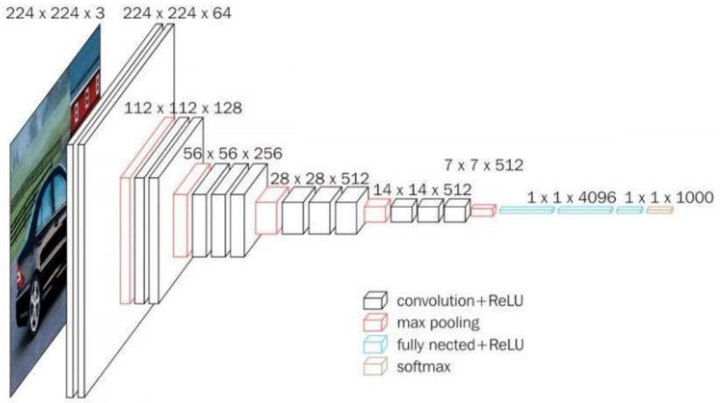


Figura 25. Esquematización de la arquitectura de VGG16. Fuente: (*Everything you need to know about VGG16*, 2021)

La "16" en VGG16 se refiere a las 16 capas con pesos entrenables. La arquitectura incluye 13 capas convolucionales que utilizan filtros pequeños de 3x3, permitiendo capturar características espaciales con detalle. Además, cuenta con 5 capas de *max pooling* que utilizan filtros de 2x2 con un *stride* de 2 para reducir la dimensionalidad y el costo computacional. La arquitectura incluye 3 capas completamente conectadas, donde las dos primeras tienen 4.096 unidades cada una y la última capa originalmente diseñada para clasificación ILSVRC tiene 1.000 unidades, seguida de una capa *softmax* para la clasificación en 1000 categorías. VGG16 toma un tensor de entrada de tamaño 224x224 con 3 canales RGB y sigue una estructura consistente de capas convolucionales y de pooling.

Para adaptar VGG16 a la clasificación de imágenes de lesiones cutáneas entre benignas y malignas, se realizaron varias modificaciones. La capa final de VGG16 diseñada para 1.000 categorías de ImageNet fue eliminada para adaptar la red a el problema de clasificación binaria. Se añadieron capas completamente conectadas personalizadas para ajustar el modelo a nuestra tarea específica, ayudando a la red a aprender combinaciones no lineales de las características extraídas por las capas convolucionales. Además, se añadió una capa de *dropout* antes de la capa final para prevenir el sobreajuste, desactivando aleatoriamente un porcentaje de las unidades de la capa durante el entrenamiento, lo que promueve una mejor generalización del modelo. Finalmente, se añadió una capa completamente conectada con una unidad y una activación sigmoide para realizar la clasificación binaria. Esto se puede ver resumido en la Figura 26

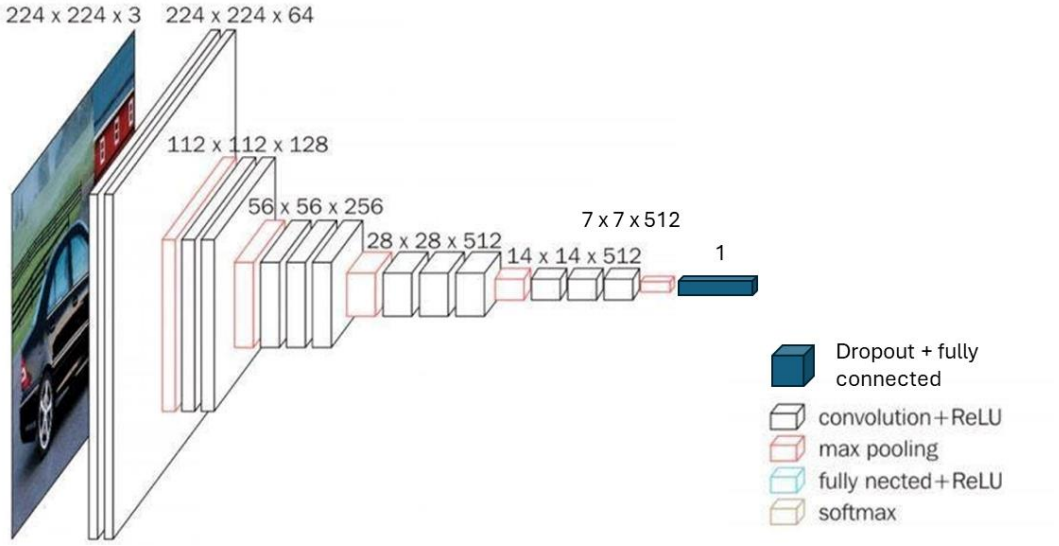


Figura 26. Esquema de la arquitectura propuesta basada en VGG16 modificada. Elaboración propia

VGG16 ha demostrado un alto rendimiento en múltiples tareas de clasificación de imágenes, con una precisión del 92,7 % en el desafío ImageNet. La profundidad de VGG16 permite capturar características complejas y detalladas en las imágenes, lo cual es crucial para diferenciar entre lesiones benignas y malignas.

Implementar VGG16 con capas adicionales completamente conectadas y una capa de *dropout* permite crear un modelo robusto y eficiente para la clasificación de imágenes de lesiones cutáneas entre benignas y malignas. La elección de VGG16 se justifica por su rendimiento comprobado, su arquitectura profunda y la facilidad de uso en transfer learning, proporcionando una base sólida para el análisis y la clasificación precisa de imágenes médicas.

Modelos basados en imágenes con mecanismos de atención

SE-Resnet

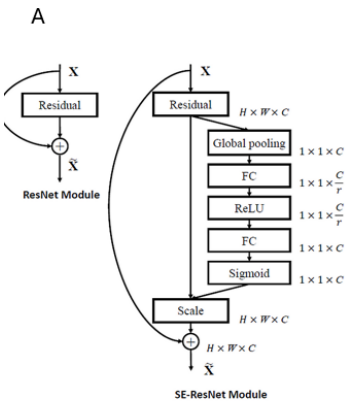
SE-ResNet es una arquitectura que combina las capacidades de la red ResNet (He et al., 2015) con los bloques *Squeeze-and-Excitation* (SE)(Hu et al., 2017), mejorando la representación de características al recalibrar dinámicamente las respuestas de las características canal por canal. Esta combinación ha demostrado ser altamente efectiva en diversas tareas de clasificación de imágenes, incluyendo la clasificación de lesiones cutáneas.

La elección de SE-ResNet para la clasificación de imágenes de lesiones cutáneas se basa en varias razones. En primer lugar, los bloques SE permiten que la red preste más atención a las características importantes y suprima las irrelevantes, mejorando la capacidad de la red para diferenciar entre diferentes tipos de lesiones cutáneas. Esta recalibración dinámica es crucial para tareas donde los detalles finos son importantes, como en la clasificación de imágenes dermatológicas. Además, SE-ResNet ha mostrado una mejora significativa en la precisión de la clasificación en comparación con las arquitecturas de ResNet estándar.

Los bloques SE permiten que la red modele las interdependencias entre canales, lo cual es crucial para capturar características sutiles y complejas en imágenes de lesiones cutáneas. La recalibración de características mejora la discriminación entre diferentes tipos de lesiones, reduciendo errores de clasificación y mejorando la precisión diagnóstica. Además, la adición de bloques SE a la arquitectura ResNet introduce un costo computacional mínimo, manteniendo la eficiencia del modelo mientras se

mejora su rendimiento. En la Figura 27 se muestra cómo se incorporan los bloques SE a la arquitectura base.

El uso de SE-ResNet para la clasificación de lesiones cutáneas está justificado por su capacidad para mejorar la recalibración de características, su desempeño superior en tareas de clasificación de imágenes y su adopción en estudios recientes que han demostrado mejoras significativas en métricas de precisión, sensibilidad y especificidad. Los bloques SE permiten que la red enfoque mejor las características relevantes, cruciales para la discriminación precisa de diferentes tipos de lesiones cutáneas.



Output size	ResNet-50	SE-ResNet-50	SE-ResNeXt-50 (32 × 4d) B
112 × 112	conv, 7 × 7, 64, stride 2		
56 × 56	max pool, 3 × 3, stride 2		
	$\begin{bmatrix} \text{conv}, 1\times1, 64 \\ \text{conv}, 3\times3, 64 \\ \text{conv}, 1\times1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} \text{conv}, 1\times1, 64 \\ \text{conv}, 3\times3, 64 \\ \text{conv}, 1\times1, 256 \\ fc, [16, 256] \end{bmatrix} \times 3$	$\begin{bmatrix} \text{conv}, 1\times1, 128 \\ \text{conv}, 3\times3, 128 \quad C=32 \\ \text{conv}, 1\times1, 256 \\ fc, [16, 256] \end{bmatrix} \times 3$
28 × 28	$\begin{bmatrix} \text{conv}, 1\times1, 128 \\ \text{conv}, 3\times3, 128 \\ \text{conv}, 1\times1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} \text{conv}, 1\times1, 128 \\ \text{conv}, 3\times3, 128 \\ \text{conv}, 1\times1, 512 \\ fc, [32, 512] \end{bmatrix} \times 4$	$\begin{bmatrix} \text{conv}, 1\times1, 256 \\ \text{conv}, 3\times3, 256 \quad C=32 \\ \text{conv}, 1\times1, 512 \\ fc, [32, 512] \end{bmatrix} \times 4$
14 × 14	$\begin{bmatrix} \text{conv}, 1\times1, 256 \\ \text{conv}, 3\times3, 256 \\ \text{conv}, 1\times1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} \text{conv}, 1\times1, 256 \\ \text{conv}, 3\times3, 256 \\ \text{conv}, 1\times1, 1024 \\ fc, [64, 1024] \end{bmatrix} \times 6$	$\begin{bmatrix} \text{conv}, 1\times1, 512 \\ \text{conv}, 3\times3, 512 \quad C=32 \\ \text{conv}, 1\times1, 1024 \\ fc, [64, 1024] \end{bmatrix} \times 6$
7×7	$\begin{bmatrix} \text{conv}, 1\times1, 512 \\ \text{conv}, 3\times3, 512 \\ \text{conv}, 1\times1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} \text{conv}, 1\times1, 512 \\ \text{conv}, 3\times3, 512 \\ \text{conv}, 1\times1, 2048 \\ fc, [128, 2048] \end{bmatrix} \times 3$	$\begin{bmatrix} \text{conv}, 1\times1, 1024 \\ \text{conv}, 3\times3, 1024 \quad C=32 \\ \text{conv}, 1\times1, 2048 \\ fc, [128, 2048] \end{bmatrix} \times 3$
1 × 1	global average pool, 1000-d fc, softmax		

Figura 27. Esquema de las características y arquitectura de SE-ResNet. A) Esquema del funcionamiento de los bloques SE dentro de la arquitectura ResNet. B) Comparativa de parámetros de ResNet, SE-ResNet y SE-ResNeXt. Fuente: (Tsang, 2019)

En la Figura 28 se muestra esquematizada la arquitectura modificada para la clasificación de lesiones cutáneas, a la arquitectura base de SE-Resnet34 se le ha eliminado el bloque clasificatorio y se le ha añadido una capa de *dropout* y una *fully connected* para la clasificación binaria de malignidad. Se ha empleado la red preentrenada.

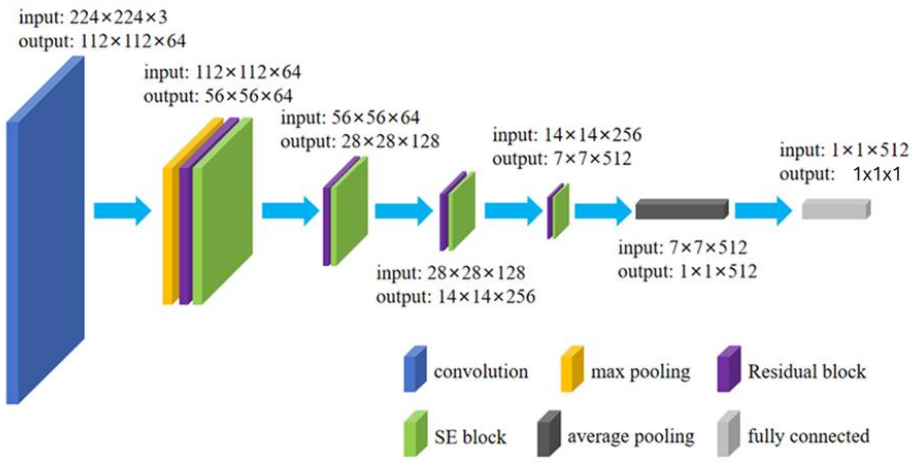


Figura 28. Esquema de la arquitectura de SE-Resnet34 modificada para la clasificación de lesiones cutáneas. Adaptada de (Xu et al., 2023)

Efficientnet

EfficientNet es una familia de redes neuronales convolucionales que optimiza la eficiencia computacional y el rendimiento. Introducida por Mingxing Tan y Quoc V. Le de Google Research en

su artículo "*EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*" (Tan & Le, 2019), EfficientNet emplea una técnica de escalado compuesta para ajustar simultáneamente el ancho, la profundidad y la resolución de la red. Esta metodología ha demostrado ser especialmente eficaz para tareas de clasificación de imágenes, incluidas las lesiones cutáneas.

La elección de EfficientNet para la clasificación de lesiones cutáneas se basa en varias razones. En primer lugar, EfficientNet utiliza un método de escalado compuesto que ajusta de manera equilibrada las dimensiones de ancho, profundidad y resolución de la red, mejorando significativamente el rendimiento sin un aumento proporcional de los recursos computacionales. Las variantes de EfficientNet, como EfficientNet-B0 a EfficientNet-B7, han demostrado un rendimiento superior en comparación con arquitecturas previas como ResNet, Inception y DenseNet. Además, EfficientNet incorpora bloques de atención (SE) que recalibran dinámicamente las características de cada canal, mejorando la capacidad de la red para enfocarse en características relevantes y aumentar la precisión en la clasificación de lesiones cutáneas.

Utilizar múltiples versiones de EfficientNet en el ensamblado proporciona diversidad en los modelos base. Cada versión tiene diferentes características y fortalezas, lo que ayuda a mejorar la robustez y la capacidad de generalización del modelo de ensamblado. Al combinar predicciones de modelos con diferentes arquitecturas, se pueden mitigar las debilidades individuales de cada modelo, logrando un rendimiento global superior.

Comparar los modelos SE-ResNet y EfficientNet es crucial para entender cuál de estas arquitecturas proporciona un mejor rendimiento en la clasificación de lesiones cutáneas, dado que ambas han demostrado ser efectivas, pero con enfoques diferentes. SE-ResNet mejora la representación de características mediante bloques de atención que recalibran dinámicamente las respuestas canal por canal, mientras que EfficientNet optimiza la eficiencia computacional utilizando una técnica de escalado compuesto que ajusta simultáneamente el ancho, la profundidad y la resolución de la red, además de incorporar bloques de atención como los bloques SE. Evaluar estos modelos permite identificar cuál ofrece una mejor discriminación de lesiones cutáneas en términos de precisión, sensibilidad y especificidad, y cómo se comportan en escenarios clínicos reales.

Además, comparar las diferentes versiones de EfficientNet aporta valor porque cada versión presenta una arquitectura con distinta complejidad y capacidad de modelado, lo que ayuda a determinar el balance óptimo entre rendimiento y eficiencia computacional. Utilizar múltiples versiones en un ensamblado puede mejorar la robustez y la capacidad de generalización del modelo, combinando fortalezas individuales y mitigando debilidades específicas, para lograr un rendimiento superior en la tarea de clasificación

5.2.2. Modelos basados en variables clínicas

Como se ha mencionado anteriormente, para abordar la clasificación de lesiones cutáneas utilizando variables clínicas, se han explorado tanto modelos de Machine Learning (ML) tradicionales como modelos de Deep Learning (DL). Los modelos de ML, como los árboles de decisión o los modelos de regresión logística, son altamente adecuados para analizar datos tabulares como las variables clínicas debido a su capacidad para manejar relaciones lineales y no lineales de manera eficiente con un menor costo computacional. Sin embargo, los modelos de DL, como las redes neuronales artificiales (ANN), ofrecen la ventaja de capturar patrones más complejos y relaciones ocultas en los datos al utilizar múltiples capas de procesamiento. Esto permite al modelo aprender combinaciones no lineales de variables clínicas, proporcionando un enfoque complementario que puede mejorar la precisión cuando

se combinan ambos tipos de modelos en una arquitectura multimodal para el diagnóstico de lesiones cutáneas.

Modelos de ML

A continuación, se detallan los modelos utilizados en el desarrollo del proyecto y las razones específicas para su selección, basadas en sus características y ventajas particulares.

La **regresión logística** es un modelo estadístico utilizado para la clasificación binaria, que calcula la probabilidad de que una instancia pertenezca a una clase particular mediante una función logística. Este modelo es apreciado por su simplicidad y facilidad de interpretación, permitiendo entender la influencia de cada característica en la predicción. Además, es computacionalmente eficiente y rápido de entrenar, incluso con conjuntos de datos grandes, lo que lo convierte en una herramienta valiosa para una primera aproximación en problemas de clasificación.

Random Forest, por otro lado, es un algoritmo de aprendizaje conjunto que construye múltiples árboles de decisión y combina sus resultados para mejorar la precisión y evitar el sobreajuste. Este modelo se destaca por su robustez frente al ruido y su capacidad para manejar datos con alta dimensionalidad. Además, puede capturar relaciones complejas entre las características y la variable objetivo, lo que mejora su precisión en problemas de clasificación complicados.

Gradient Boosting es otro algoritmo de conjunto que crea modelos de predicción de manera secuencial, donde cada nuevo modelo intenta corregir los errores de los modelos anteriores. Este enfoque proporciona un alto rendimiento gracias a su capacidad para mejorar iterativamente, y su flexibilidad permite ajustar finamente los hiperparámetros para optimizar el rendimiento del modelo.

Support Vector Machine (SVM) es un algoritmo de clasificación que encuentra el hiperplano que mejor separa las clases en el espacio de características. Este modelo es eficiente en espacios de alta dimensión y eficaz incluso cuando el número de dimensiones es mayor que el número de muestras. Con el uso de núcleos, SVM puede manejar relaciones no lineales entre las características y la variable objetivo, lo que lo hace adecuado para una variedad de problemas de clasificación.

Finalmente, **k-Nearest Neighbors (k-NN)** es un algoritmo basado en la similitud que clasifica una instancia en función de las clases de sus k vecinos más cercanos en el espacio de características. Este modelo es simple de implementar y entender, y su naturaleza no paramétrica lo hace útil en situaciones donde las distribuciones de los datos son desconocidas o complejas.

La selección de estos modelos se justifica por la diversidad de enfoques que representan, desde modelos lineales y basados en árboles hasta métodos basados en la distancia y el aprendizaje conjunto. Esto permite un balance entre interpretabilidad y rendimiento, así como la capacidad de manejar diferentes tipos de datos y relaciones, asegurando una comparación integral para seleccionar el modelo más adecuado para el problema específico.

Modelo de DL

Con el fin de aproximar el comportamiento del uso de las variables clínicas en la red neuronal artificial, se ha construido un modelo de clasificación de DL que emplea como entrada las variables clínicas. Para ello toma como entrada una capa completamente conectada o "*Linear*" de tamaño de entrada 11, el número de variables clínicas disponibles, y procede a realizar una normalización de lote, con activación *Swish* y un *dropout* de 0,3. Posteriormente pasa por tras capa linear de tamaño de entrada 512 y tamaño de salida 128, también con su respectiva normalización de lotes y activación *Swish*. Estas dos capas

crean la fase de extracción de características. Después, pasan por la etapa clasificatoria, donde incluimos un *dropout* de 0,5 y la capa clasificatoria. El resumen de la arquitectura se puede ver en la **Figura 29**.

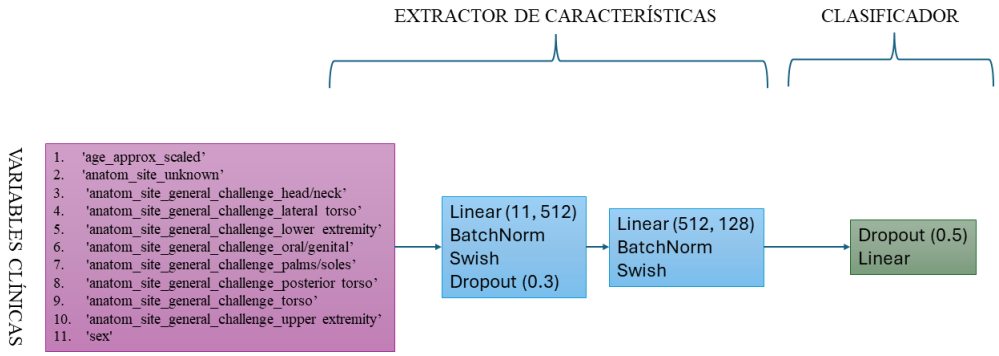


Figura 29. Esquema del modelo de red neuronal artificial para variables clínicas tabulares (TNN)

Este funcionamiento basado en la extracción de características y su clasificación, es similar al que se empleará para la red neuronal multimodal, en el que ambas modalidades de datos (imágenes y variables clínicas) pasarán por sus propios extractores de características paralelamente, juntando todas estas en un único vector de características, que finalmente pasará por la etapa clasificatorio, de características similares a la descrita para el modelo TNN, pero con diferente tamaño de entrada al variar el número de características extraídas.

5.2.3. Modelo integrador o Multimodal

Para el diseño de la red neuronal multimodal, se debe considerar la arquitectura como dos etapas; la fase de extracción de características, y la etapa clasificatoria. A su vez, la primera está dividida en dos partes que trabajan paralelamente. Por un lado, se usa la fase extractora de características de una CNN, en nuestro caso usaremos la base de EfficientNet entrenada en experimentos anteriores, eliminando la etapa clasificatoria. Paralelamente, los datos clínicos circularán por un extractor de características de datos tabulares.

Llegado a este punto, se dispondrá de dos vectores de características, uno proveniente de las imágenes y otro proveniente de las variables clínicas, que serán concatenados para tener un único vector de características. Este será alimentado a la etapa clasificatoria, obteniendo así una clasificación que tenga en cuenta las imágenes y los datos clínicos del paciente. Se muestra un esquema visual en la Figura 30.

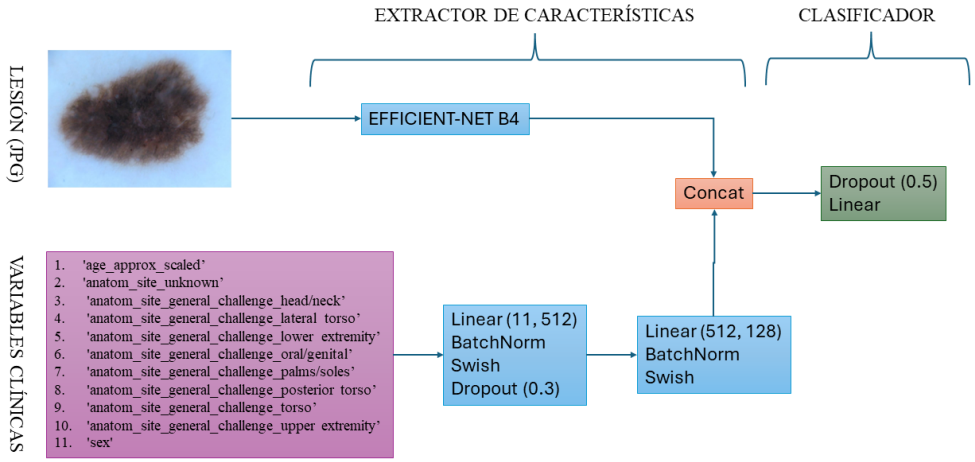


Figura 30. Esquema del funcionamiento de la red neuronal multimodal.

5.2.4. Modelos de ensamblado

El ensamblado de modelos (*model ensembling*) es una técnica de aprendizaje automático que combina las predicciones de múltiples modelos para mejorar la precisión y robustez general. En la clasificación de lesiones cutáneas, el ensamblado de modelos puede ser particularmente útil debido a la variabilidad y complejidad de las imágenes dermatológicas.

El ensamblado ofrece varias ventajas. Ayuda a reducir la variabilidad y el sesgo, mejorando la precisión general. Al combinar múltiples modelos, se mejora la capacidad del sistema para manejar casos atípicos y ruidosos. Además, el ensamblado a menudo resulta en una mejora significativa en métricas de rendimiento como la precisión y el recall.

El uso de ensamblado de modelos en la clasificación de lesiones cutáneas se justifica por varias razones. El uso de técnicas de ensamblado ha demostrado consistentemente mejoras en la precisión de la clasificación al combinar las predicciones de múltiples modelos, aprovechando sus diferentes fortalezas. Los modelos ensamblados son más robustos y generalizan mejor en conjuntos de datos no vistos, lo cual es crucial en la clasificación de imágenes médicas donde la variabilidad es alta y los datos pueden ser ruidosos.

Las imágenes dermatológicas pueden variar significativamente en términos de color, textura y forma, y el ensamblado de modelos ayuda a capturar esta complejidad al combinar las fortalezas de diferentes arquitecturas de modelos. Además, el ensamblado ayuda a reducir el sesgo y la varianza al promediar las predicciones de múltiples modelos, resultando en una mejor estabilidad y rendimiento del modelo final.

Métodos de ensamblado empleados

Para el ensamblado de modelos en la clasificación de lesiones cutáneas, se han utilizado tres técnicas específicas: voto mayoritario, voto ponderado y *stacking* (*majority voting, weighted majority voting* y *stacking)*. A continuación, se justifica el uso de cada una de estas técnicas en el contexto del problema dermatológico.

Majority Voting

Majority voting o voto mayoritario es una técnica de ensamblado simple pero efectiva que combina las predicciones de múltiples modelos y selecciona la clase que recibe la mayoría de los votos. Esta técnica es especialmente útil en problemas de clasificación de lesiones cutáneas debido a su capacidad para reducir el sesgo individual de los modelos y mejorar la robustez general del sistema. Al combinar varios modelos, se disminuye la probabilidad de errores cometidos por modelos individuales, lo que es crucial en aplicaciones médicas donde la precisión es esencial. *Majority voting* es fácil de implementar y comprender, proporcionando una mejora sin añadir complejidad adicional.

Weighted Majority Voting

Weighted majority voting o voto ponderado es una extensión de la técnica de *majority voting*, donde se asignan pesos a las predicciones de los modelos en función de su rendimiento individual. En el contexto de la clasificación de lesiones cutáneas, esta técnica permite que los modelos más precisos y confiables tengan una mayor influencia en la decisión final. Esto es especialmente beneficioso cuando algunos modelos tienen un mejor desempeño en ciertas características de las imágenes dermatológicas. Al ponderar las contribuciones de los modelos, se puede mejorar aún más la precisión y la robustez del

sistema, asegurando que las predicciones más confiables tengan un mayor impacto en la clasificación final.

Stacking

Stacking es una técnica de ensamblado más avanzada que utiliza un modelo meta-aprendizaje para combinar las predicciones de varios modelos base. En el caso de la clasificación de lesiones cutáneas, stacking es particularmente útil porque permite aprovechar las fortalezas individuales de diferentes modelos y capturar relaciones más complejas entre las características de las imágenes y los datos clínicos. El modelo meta-aprendizaje aprende a optimizar la combinación de las predicciones de los modelos base, mejorando la capacidad del sistema para manejar la variabilidad y complejidad inherente en las imágenes dermatológicas. Stacking tiende a ofrecer mejoras significativas en la precisión y capacidad de generalización del modelo ensamblado, haciéndolo ideal para aplicaciones donde la exactitud del diagnóstico es crítica. Para la realización de esta técnica se ha empelado los modelos de *Logistic Regression, Random Forest, Gradient Boosting y SVM* explicados en apartados anteriores.

6. Implementación

6.1. Desarrollo de entrenamiento y optimización

Para el desarrollo de este proyecto, se ha utilizado la plataforma Kaggle, que proporciona un entorno interactivo basado en libretas Jupyter. Kaggle facilita el acceso a potentes GPU, específicamente del tipo T4x2, que aceleran significativamente el entrenamiento de los modelos de *deep learning*. Estas GPU permiten ejecutar los modelos con una capacidad de procesamiento superior, reduciendo drásticamente el tiempo de entrenamiento y permitiendo el manejo eficiente de grandes volúmenes de datos.

Además del uso de Kaggle y sus recursos de GPU, se ha llevado a cabo un proceso de optimización de hiperparámetros para mejorar el rendimiento de los modelos. La optimización de hiperparámetros es una etapa crucial que implica ajustar parámetros como la tasa de aprendizaje, el tamaño de lote y las funciones de activación, entre otros. Este proceso se realiza para encontrar la combinación óptima que maximice la precisión y eficiencia del modelo en la clasificación de lesiones cutáneas.

6.1.1. Parámetros de entrenamiento

Función de pérdidas

Binary Cross Entropy With Logits Loss

La función de pérdida *BCEWithLogitsLoss* es una combinación de la Entropía Cruzada Binaria (*Binary Cross-Entropy, BCE*) y la función sigmoide. Esta función se utiliza comúnmente en problemas de clasificación binaria donde las salidas del modelo son *logits*, que son los valores sin procesar que un modelo de clasificación binaria produce antes de aplicar la función sigmoide para convertirlos en probabilidades, en lugar de probabilidades. Al combinar estos dos pasos en una sola función, se mejora la estabilidad numérica y la eficiencia del cálculo.

BCEWithLogitsLoss funciona de la siguiente manera:

Primero, la entropía cruzada binaria (BCE) mide la discrepancia entre dos distribuciones de probabilidad, generalmente entre las salidas predichas por el modelo y las etiquetas verdaderas. La fórmula de BCE aparece en la Fórmula 3.

$$BCE(p, y) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)]$$

Fórmula 3. Calcula de entropía cruzada binaria o BCE.

donde y_i son las etiquetas verdaderas y p_i son las probabilidades predichas.

Por otro lado, la función sigmoide convierte los logits (valores en el rango de -∞ a +∞ en probabilidades (valores entre 0 y 1). La fórmula de la función sigmoide es:

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

Fórmula 4. Función sigmoide.

Para combinar la BCE la función sigmoide se aplica la función sigmoide a las salidas del modelo y luego se calcula la BCE, reduciendo problemas de estabilidad numérica asociados con la transformación de *logits* en probabilidades antes de calcular la BCE. La fórmula combinada se muestra en Fórmula 5.

$$BCEWithLogitsLoss(x, y) = BCE(\sigma(x), y)$$

Fórmula 5. Ecuación de la función de pérdidas Binary Cross Entropy With Logits Loss.

Alguna de las ventajas de usar *BCEWithLogitsLoss*:

- Estabilidad numérica: La combinación de la sigmoide y la BCE en una sola función evita problemas de desbordamiento y pérdida de precisión que pueden ocurrir si se aplican por separado.
- Eficiencia computacional: Al realizar ambas operaciones en un solo paso, se mejora la eficiencia computacional, reduciendo el tiempo de cálculo y el uso de memoria.
- Aplicabilidad en clasificación binaria: Es particularmente útil en tareas de clasificación binaria donde las salidas del modelo son *logits*, como en la clasificación de lesiones cutáneas.

En estudios recientes, el uso de *BCEWithLogitsLoss* ha demostrado mejorar la precisión y la eficiencia de los modelos de clasificación de lesiones cutáneas. La estabilidad numérica adicional permite un entrenamiento más robusto y preciso. Además, *BCEWithLogitsLoss* es altamente adaptable y se puede integrar fácilmente con diversos tipos de arquitecturas de modelos, incluidas las redes neuronales convolucionales (CNN) utilizadas en la clasificación de imágenes médicas.

Focal Loss

La función de pérdida Focal Loss es una extensión de la función de Entropía Cruzada Binaria (BCE) diseñada para abordar problemas de desequilibrio en los datos de clasificación. Se enfoca en reducir el impacto de las muestras bien clasificadas y aumentar el peso de las muestras difíciles de clasificar, ayudando al modelo a centrarse en los casos más desafiantes.

El funcionamiento se basa en la modulación del factor de focalización (γ): Este parámetro ajusta la pérdida de acuerdo con la dificultad de clasificación de cada muestra. Un valor alto de γ aumenta el peso de las muestras mal clasificadas, reduciendo el impacto de las muestras bien clasificadas. Esto permite que el modelo se enfoque más en las muestras difíciles, mejorando su capacidad para aprender de ellas.

Por otro lado, el factor de ponderación (α) actúa como un factor de ponderación para las clases, ajustando el peso de las clases minoritarias para equilibrar su impacto en la función de pérdida. Esto es

especialmente útil en conjuntos de datos desbalanceados, donde las clases minoritarias tienden a ser subrepresentadas.

La fórmula de Focal Loss incluye tanto la BCE como la modulación del factor de focalización, lo que permite una combinación efectiva de ambos enfoques. La fórmula combinada aparece en la Fórmula 6

$$FL(p_t) = -\alpha_t(1 - p_t)^{\gamma} \log(p_t)$$

Fórmula 6. Ecuación de la función de pérdidas Focal Loss.

La principal ventaja de usar Focal Loss es el manejo de desequilibrio en los datos, ya que Focal Loss es particularmente efectiva en conjuntos de datos desbalanceados, donde las clases minoritarias pueden ser subrepresentadas. Ajustando los pesos de las clases y enfocándose en las muestras difíciles, se mejora la capacidad del modelo para aprender de todas las clases de manera equilibrada.

También es ventajosa la reducción de la influencia de muestras bien clasificadas. Al reducir el impacto de las muestras que el modelo ya clasifica correctamente, permite que el modelo se concentre más en mejorar las predicciones de las muestras difíciles, lo que puede llevar a un mejor rendimiento global. Por último, Focal Loss se puede ajustar fácilmente mediante los parámetros α y γ, lo que permite adaptar la función de pérdida a diferentes características del conjunto de datos y objetivos del modelo.

Optimizador

El optimizador Adam (*Adaptive Moment Estimation*) (Kingma & Ba, 2014) es un algoritmo de optimización de gradiente descendente estocástico que combina las ventajas de dos otros métodos populares: *AdaGrad* y *RMSProp*. Adam calcula adaptativamente las tasas de aprendizaje para cada parámetro, utilizando estimaciones del primer y segundo momento del gradiente. Las fortalezas de Adam incluyen su eficiencia computacional, su bajo requerimiento de memoria y su capacidad para manejar problemas con grandes cantidades de datos y parámetros. Además, Adam es robusto frente a gradientes ruidosos y funciona bien con problemas que involucran datos dispersos o características raras.

Las debilidades de Adam pueden incluir su tendencia a no converger tan bien como otros optimizadores en algunos problemas específicos y su sensibilidad a la configuración de hiperparámetros como las tasas de aprendizaje y los factores de decaimiento.

Tasa de aprendizaje

La tasa de aprendizaje (*learning rate*) es uno de los hiperparámetros más cruciales en el entrenamiento de modelos de DL, ya que influye directamente en la velocidad y estabilidad de la convergencia del modelo. Un *learning rate* demasiado alto puede hacer que el modelo no converja o incluso diverja, mientras que uno demasiado bajo puede resultar en una convergencia muy lenta, impidiendo que el modelo alcance el óptimo rendimiento.

Explorar un rango amplio de tasas de aprendizaje, desde 1e-6 hasta 1, permite identificar un valor óptimo que equilibra entre una convergencia rápida y una estabilidad adecuada. Este enfoque, conocido como "*Learning Rate Finder*", ha sido propuesto por Leslie N. Smith en su artículo "Cyclical Learning Rates for Training Neural Networks"(Smith, 2015). La idea principal de este método es entrenar el modelo durante unas pocas iteraciones mientras se incrementa exponencialmente el *learning rate* en cada paso. Esto permite visualizar cómo cada tasa de aprendizaje afecta la pérdida, ayudando a identificar el rango de valores más efectivo.

El *Learning Rate Finder* tiene varias ventajas, entre ellas, la rápida exploración de un rango amplio de learning rates sin requerir un entrenamiento completo y la visualización clara de la curva de pérdida resultante, que facilita la selección de un valor óptimo.

Una vez identificado un *learning rate* inicial óptimo, utilizamos la técnica de "*ReduceLROnPlateau*" durante el entrenamiento completo del modelo. Esta técnica reduce automáticamente la tasa de aprendizaje cuando la métrica de interés (como las pérdidas) se estabiliza, ayudando a afinar los últimos pasos de la optimización. Los beneficios de "*Reduce LR On Plateau*" incluyen su adaptabilidad, ya que ajusta dinámicamente el *learning rate* durante el entrenamiento, y su capacidad para mejorar la estabilidad al reducir la tasa de aprendizaje en partes planas, previniendo que el modelo oscile alrededor de mínimos locales y mejorando la estabilidad en las últimas fases del entrenamiento.

En cuanto al experimento, se inicia cargando el modelo VGG16 modificado y los *datasets* de entrenamiento y validación, definiendo un rango amplio de tasas de aprendizaje para la búsqueda. Durante la ejecución del *Learning Rate Finder*, el modelo se entrena durante un número limitado de iteraciones, incrementando exponencialmente el *learning rate* en cada paso y registrando la pérdida para identificar el comportamiento del modelo con diferentes tasas de aprendizaje. Posteriormente, se analiza la curva de pérdida obtenida del *Learning Rate Finde*r para seleccionar el punto donde la pérdida comienza a disminuir rápidamente antes de que aumente debido a un gradiente excesivamente grande.

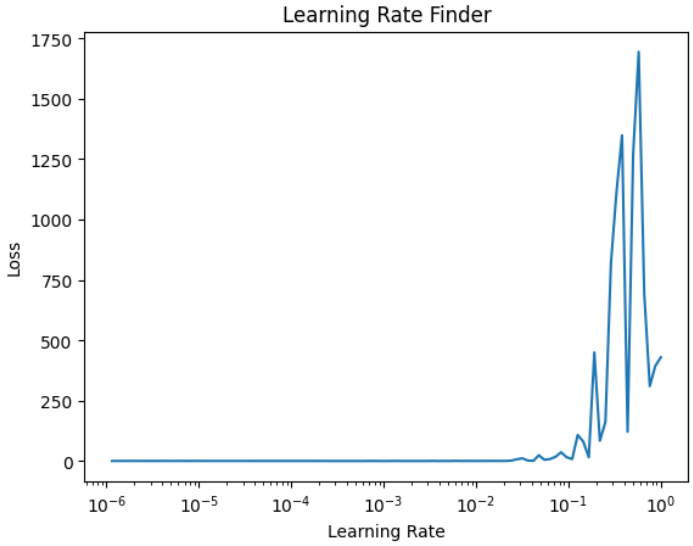


Figura 31. Resultados obtenidos de la búsqueda de tasa de aprendizaje o learning rate Finder.

En la Figura 31, se grafican los resultados del experimento, donde se muestra la pérdida (Loss) en función de la tasa de aprendizaje (*Learning Rate*) en una escala logarítmica. En el eje X, la escala logarítmica muestra la tasa de aprendizaje variando desde 10^{-6} hasta 1. En el eje Y, se muestra la pérdida correspondiente a cada tasa de aprendizaje probada. La curva de pérdida se interpreta de la siguiente manera:

En la región plana al comienzo, en el rango de tasas de aprendizaje muy bajas (10^{-6} a 10^{-4}), la pérdida se mantiene casi constante y baja. Esto indica que la tasa de aprendizaje es demasiado baja para que el modelo aprenda efectivamente. A medida que la tasa de aprendizaje aumenta hacia 10^{-2}, se observa una ligera disminución en la pérdida, lo que sugiere que el modelo comienza a aprender de manera efectiva.

Más allá de 10^{-1}, la pérdida comienza a aumentar drásticamente y se vuelve muy inestable. Esto indica que la tasa de aprendizaje es demasiado alta, causando que el modelo no converja y que los gradientes se vuelvan demasiado grandes.

Para determinar la tasa de aprendizaje óptima, se debe buscar la región justo antes de que la pérdida comience a aumentar rápidamente. En este caso, la tasa de aprendizaje seleccionada se encuentra en el rango de 10^{-2}, donde la pérdida empieza a disminuir y muestra un buen comportamiento.

Tamaño de lote

Seleccionar el tamaño de lote (*batch size*) en modelos de DL requiere un balance entre varios factores teóricos y prácticos que influyen en el entrenamiento de las redes neuronales.

Un aspecto clave es el equilibrio entre eficiencia computacional y convergencia. Los tamaños de lote más grandes permiten una mayor paralelización y mejor uso de la memoria GPU, acelerando el entrenamiento debido a que el hardware moderno, como las GPUs, es más eficiente con grandes lotes de datos. Sin embargo, los tamaños de lote más pequeños pueden proporcionar un entrenamiento más estable y una mejor generalización al introducir más ruido en la estimación del gradiente, ayudando a evitar mínimos locales y a explorar mejor el espacio de soluciones.

Otro factor es el impacto en el gradiente. Un tamaño de lote más grande ofrece una mejor estimación del gradiente de la función de pérdida sobre el conjunto de entrenamiento, llevando a actualizaciones de parámetros más precisas. Por otro lado, un tamaño de lote más pequeño introduce más variabilidad en las estimaciones del gradiente, lo que puede ayudar a salir de mínimos locales y mejorar la capacidad de generalización del modelo.

Finalmente, las consideraciones de memoria son cruciales. La capacidad de la memoria de la GPU puede limitar el tamaño del *batch*, especialmente con modelos grandes que requieren más memoria para almacenar parámetros y gradientes durante el *backpropagation*.

La elección del *batch size* en estos experimentos se basa en un equilibrio entre eficiencia computacional y la estabilidad de convergencia. Siguiendo la investigación de (Smith, 2018), utilizamos un tamaño de lote de 64, ya que proporciona una buena estimación del gradiente al tiempo que permite un uso eficiente de la memoria de la GPU, maximizando la paralelización.

Número de épocas

Para el entrenamiento de los modelos de clasificación de lesiones cutáneas, hemos elegido un número de 30 épocas. Este número de épocas permite que los modelos tengan suficiente tiempo para aprender las características de los datos sin incurrir en un entrenamiento excesivo. No obstante, para evitar el sobreajuste, hemos implementado la técnica de detención anticipada *(early stopping*). Esta técnica detiene el entrenamiento si la métrica de validación, en este caso el área bajo la curva de la curva ROC (ROC *Score*), no mejora durante tres épocas consecutivas.

La elección de 30 épocas se justifica por la necesidad de proporcionar a los modelos un tiempo adecuado para converger y captar las complejas relaciones en los datos. Sin embargo, continuar el entrenamiento más allá del punto en el que la métrica de validación deja de mejorar puede llevar al sobreajuste, donde el modelo se adapta demasiado a los datos de entrenamiento y pierde capacidad de generalización a nuevos datos.

Por el contrario, aplicar este método da pie a un posible subentrenamiento. En algunos casos, la detención anticipada podría detener el entrenamiento demasiado pronto si la métrica de validación

muestra una mejora temporal y luego vuelve a mejorar después de más épocas. Además, existe una dependencia de la métrica de validación, es decir, la efectividad de la detención anticipada depende en gran medida de la métrica de validación seleccionada. Si esta métrica no es adecuada, podría llevar a decisiones subóptimas sobre cuándo detener el entrenamiento.

7. Evaluación del desempeño de los modelos

7.1.1. Estrategias de validación

Validación cruzada

El método de validación utilizado en este estudio implica una división estratificada de los datos en conjuntos de entrenamiento, validación y prueba. En este enfoque, se reserva el 20% de los datos para la validación final (o prueba), es decir, el *Holdout*, y el 10% para la validación interna dentro del conjunto de entrenamiento. Este método ha sido ampliamente justificado y aplicado en diversos estudios en el campo de la dermatología y la clasificación de imágenes médicas.

La estratificación asegura que la distribución de clases en los conjuntos de datos de entrenamiento, validación y prueba sea representativa del conjunto de datos original. Esto es particularmente importante en problemas de clasificación de lesiones cutáneas, donde algunas clases pueden estar subrepresentadas. Al mantener la proporción de clases constante, se evita el sesgo que podría afectar negativamente el rendimiento del modelo.

El principal beneficio de este sistema frente a otros métodos de validación, como la validación cruzada simple, es la reducción del sesgo de selección y la mejora en la capacidad de generalización del modelo. La validación cruzada simple puede no garantizar una representación adecuada de todas las clases en cada subconjunto de datos, especialmente en conjuntos de datos desequilibrados.

División de datos de entrenamiento y prueba

Para evaluar de manera efectiva el rendimiento del modelo, realizamos una división estratificada del conjunto de datos en dos subconjuntos principales: entrenamiento (*train*) y prueba (*test*). Esta división asegura que la proporción de las clases se mantenga constante entre ambos conjuntos, garantizando una representación adecuada de las lesiones benignas y malignas en cada subconjunto.

El gráfico de distribución de las clases en el conjunto de entrenamiento (Figura 32) muestra un equilibrio razonable entre las clases "benigna" y "maligna", con un conteo ligeramente mayor de lesiones benignas. En el conjunto de prueba, observamos una proporción similar, lo que indica que la estratificación ha sido efectiva en mantener la consistencia de las clases entre los conjuntos de datos.

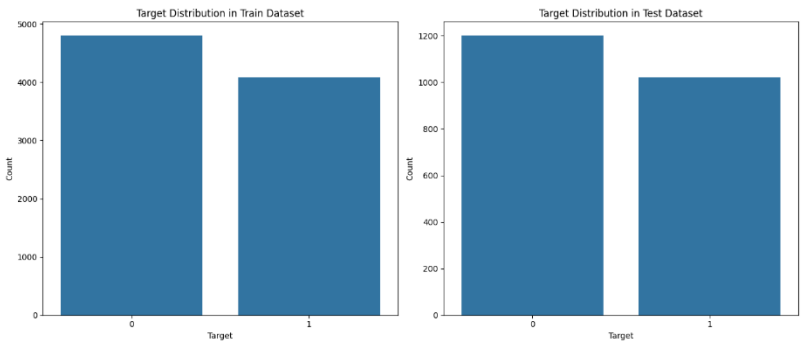


Figura 32. Distribución de la variable objetivo 'target' en el conjunto de datos de entrenamiento 'train' y test, donde '0' representa lesiones benignas y '1' lesiones malignas.

Métricas de validación

Matriz de confusión

La matriz de confusión es un método muy popular para resolver problemas de clasificación. Se emplea para validar problemas de clasificación binaria y multiclase. Después de la clasificación, se utilizó la matriz de confusión para evaluar el rendimiento de los métodos utilizados. La estructura de la matriz de confusión se muestra en la **Figura 33** para la clasificación binaria.

La matriz de confusión muestra gráficamente el número de clasificaciones realizadas divididas entre acertadas y falladas. Por un lado, tenemos las clasificaciones acertadas divididas en verdaderos negativos y verdaderos positivos. Los errores a su vez quedan divididos en falsos positivos (FP), que eran instancias negativas que el modelo ha clasificado como positivos, y falsos negativos (FN) que son errores de la clase positiva catalogados erróneamente como negativos.

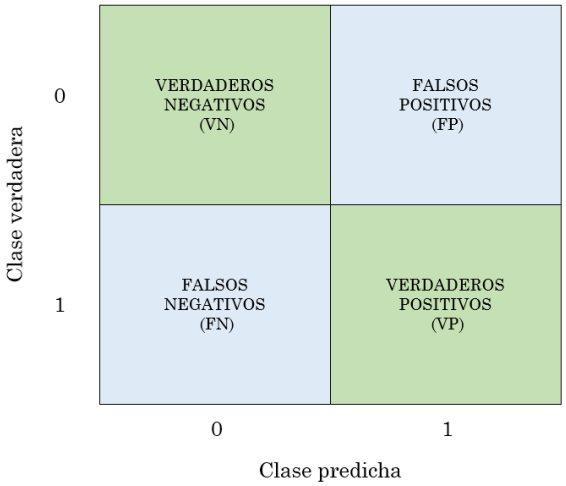


Figura 33. Matriz de confusión.

Exactitud

La métrica de la exactitud está basada en los valores de la matriz de confusión, su fórmula se puede ver en la Fórmula 7. Representa el número de muestras clasificadas correctamente sobre el total.

$$Exactitud = \frac{VP + VN}{VP + VN + FP + FN}$$

Fórmula 7. Ecuación para calcular la métrica de exactitud.

Sensibilidad o Recall

La sensibilidad indica la capacidad del clasificador para discriminar los casos positivos de los negativos. Es representada como la fracción de verdaderos positivos, es decir los valores positivos correctamente predichos frente al total de valores positivos en el conjunto de datos. Su cálculo se muestra en la Fórmula 8. En el contexto de este trabajo, la sensibilidad refleja la capacidad del modelo para detectar correctamente las lesiones malignas (clase 1) entre todas las lesiones malignas presentes en el conjunto de datos.

$$Sensibilidad = \frac{VP}{VP + FN}$$

Fórmula 8. Ecuación para calcular la métrica de sensibilidad o Recall.

Especificidad o tasa de verdaderos negativos

Esta expresa la capacidad del clasificador para distinguir los casos negativos. Se representa como la fracción de datos negativos predichos correctamente (verdaderos negativos) frente a todos los datos negativos del conjunto. Su cálculo se puede ver en la Fórmula 9. En el contexto de este trabajo, la especificidad refleja la capacidad del modelo para identificar correctamente las lesiones benignas (clase 0) entre todas las lesiones benignas presentes en el conjunto de datos.

$$Especificidad = \frac{VN}{VN + FP}$$

Fórmula 9. Cálculo de la especificidad.

Precisión

La calidad de una predicción positiva del modelo se mide por su precisión. El número de verdaderos positivos dividido por el número total de predicciones positivas se conoce como precisión. La **Fórmula 10** muestra su cálculo. n el contexto de este trabajo, la precisión indica cuántas de las lesiones clasificadas como malignas (clase 1) por el modelo realmente son malignas.

$$Precisión = \frac{VP}{VP + FP}$$

Fórmula 10. Calculo de la precisión.

F1 SCORE

Los dos componentes fundamentales de la puntuación F1 son la sensibilidad y la precisión. El objetivo de la puntuación F1 es unir las métricas de sensibilidad y precisión en una sola. La **Fórmula 11** muestra la puntuación F1 como la media armónica de sensibilidad y precisión.

$$F1 = 2 \times \frac{Precisión \times Sensibilidad}{Precisión + Sensibilidad}$$

Fórmula 11. Ecuación de la puntuación F1 o *F1 Score*

Curva ROC Y Área bajo la curva (AUC o ROC Score)

Para evaluar el rendimiento de un clasificador, se utilizan generalmente gráficos ROC (*Receiver Operating Characteristic*). El eje X representa la especificidad y el eje Y representa la sensibilidad. Uno de los métodos más confiables para evaluar la capacidad de clasificación de modelos es el análisis ROC, ya que mide el rendimiento de clasificación del modelo trazando la sensibilidad frente a la especificidad. El área bajo la curva (AUC, por sus siglas en inglés) del gráfico ROC es una medida cuantitativa del rendimiento general del modelo: un AUC de 1.0 indica una clasificación perfecta, mientras que un AUC de 0.5 sugiere que el modelo no tiene capacidad discriminativa (equivalente a una clasificación aleatoria).

El presente apartado de evaluación contiene los resultados de los modelos en el conjunto de datos de Test o *Holdout*, es decir, el conjunto de datos nunca visto por la red durante el entrenamiento, y que se emplea para aproximar el rendimiento del modelo en datos con los que no ha sido entrenado. El Añejo I, contiene el análisis de la evolución de las métricas principales durante la etapa de entrenamiento.

7.1.2. Evaluación del rendimiento según distribución de datos

Con el objetivo de comparar el rendimiento de los modelos con diferentes distribuciones de datos, se ha comparado el conjunto de datos completo (8,79% de lesiones malignas) frente a un conjunto de datos más equilibrado (45,96% de lesiones malignas). Para la comparación, se entrenó el modelo de imágenes VGG16 (Modelos basados en imágenes sin mecanismos de atención). Se entrenó con los parámetros descritos en *6.1.1 Parámetros de entrenamiento*, con la diferencia de que en el caso del conjunto de datos equilibrado se empleó la función de pérdidas *BCEwithLogitsLoss*, mientras que en el caso del conjunto de datos no equilibrado se empleó *Focal Loss* tomando como parámetros alfa=2 y gamma=5. El parámetro alfa ayuda a equilibrar la contribución de ambas clases a la función de pérdida, mientras que gamma modula la pérdida de acuerdo con la dificultad de clasificación de cada muestra. Cuando gamma=5, se da más peso a las muestras difíciles de clasificar y se reduce la pérdida para las muestras que el modelo ya clasifica correctamente. Esto significa que, durante el entrenamiento, el modelo se enfocará más en las muestras mal clasificadas (aquellas con alta incertidumbre) y menos en las muestras bien clasificadas (aquellas con baja incertidumbre).

Los parámetros escogidos para Focal Loss (alfa=2 y gamma=5) han sido escogidos por su recurrencia en la literatura, en (Lin et al., 2020) establecen que un valor alto de gamma (por ejemplo, gamma=5) pone más énfasis en las muestras difíciles de clasificar, lo que es beneficioso en escenarios donde hay una gran variabilidad en la dificultad de las muestras. Además, un alfa específico ayuda a manejar el sesgo de clase al darle mayor peso a la clase minoritaria. Este enfoque es especialmente útil cuando la clase minoritaria es de mayor interés clínico, como en la detección de lesiones malignas en imágenes dermatológicas.

Tabla 2. Resumen de las principales métricas obtenidas en el conjunto de datos de test entrenando VGG16 con el dataset completo y Focal Loss y el dataset equilibrado con BCE.

Modelo	Exactitud (Accuracy)	Precisión	Sensibilidad (Recall)	Especificidad	F1 Score	ROC AUC
VGG16 todo dataset	0,9372	0,6901	0,5191	0,9775	0,5925	0,7936
VGG16 equilibrado	0,8041	0,7738	0,8110	0,7983	0,7920	0,9000

Los resultados en cuanto a matriz de confusión y principales métricas se pueden ver en la Tabla 2 y la Figura 34. Aunque la exactitud (*accuracy*) es mayor en el conjunto de datos no equilibrado, esto se alinea con la proporción original de muestras malignas, que es aproximadamente del 8%. Sin embargo, al observar métricas más críticas como la precisión y la sensibilidad, se evidencia que ambos son significativamente menores en comparación con el conjunto de datos equilibrado. Esto se refleja claramente en la métrica F1 Score y el área bajo la curva (ROC *Score*), que demuestran una capacidad de discriminación más limitada en el conjunto no equilibrado, resaltando la importancia de utilizar un dataset equilibrado para mejorar la calidad del modelo en términos de sensibilidad y capacidad de detección de lesiones malignas.

De este punto en adelante se usó el conjunto de datos equilibrado principalmente ya que los resultados respecto a la detección de la clase maligna eran más prometedores, además de la eficiencia que conlleva entrenar con un conjunto de datos menor, desde un punto de vista informático.

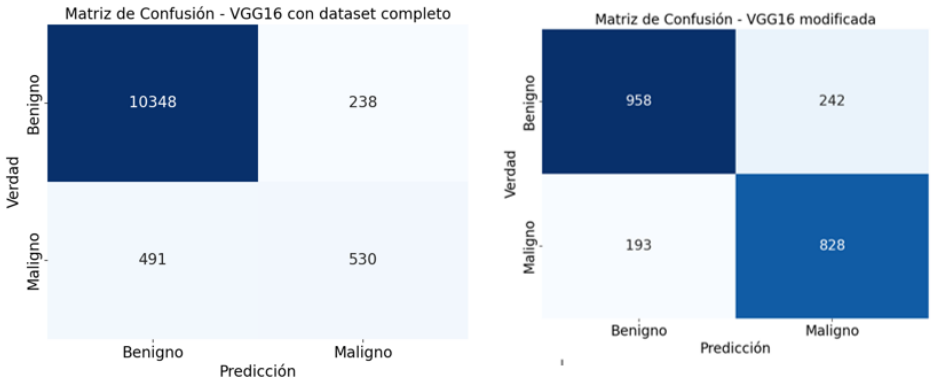


Figura 34. Matriz de confusión del modelo VGG 16 con el conjunto de datos completo (<10% casos malignos) evaluado sobre el conjunto de datos de test.

7.1.3. Evaluación de los modelos de imágenes

Modelo de imagen sin mecanismos de atención: VGG16.

En la Figura 35 se muestra el área bajo la curva de la curva ROC y la matriz de confusión obtenido en el conjunto de datos de test o *Holdout (*nunca vistos durante entrenamiento). Podemos ver que la red ha obtenido un área bajo la curva (ROC Score) de 0,90, habiendo clasificado erróneamente 193 lesiones como benignas que realmente eran malignas y 242 lesiones catalogadas como malignas cuando realmente eran benignas.

En la Tabla 3 se aprecian resumidas las principales métricas obtenidas sobre el conjunto de datos de test, siendo la exactitud final de 0,80 y la sensibilidad final de 0,81.

Tabla 3. Resumen de las métricas en el conjunto de datos de test del modelo VGG16.

Modelo	Exactitud (Accuracy)	Precisión	Sensibilidad (Recall)	Especificidad	F1 Score	ROC AUC
VGG16 modificado	0,8041	0,7738	0,8110	0,7983	0,7920	0,9000

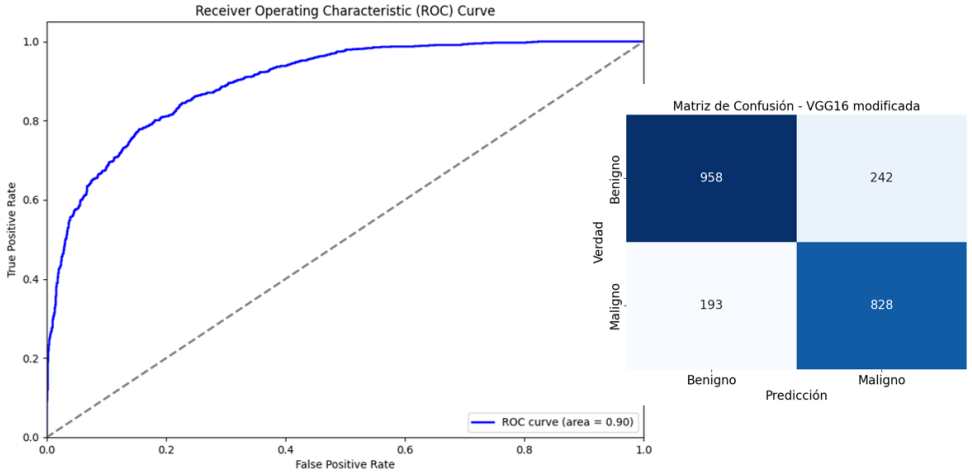


Figura 35. A la izquierda, la curva ROC del modelo VGG16 sobre el conjunto de datos de test. A la derecha la matriz de confusión sobre el mismo.

Modelo de imagen con mecanismos de atención: SE-ResNet

El modelo SE-ResNet es un modelo que toma como entrada las imágenes de lesiones cutáneas y emplea mecanismos de atención, específicamente *Squeeze and Excitation*. Probando el modelo en el conjunto de datos de test, se calculó un 0,93 de área bajo la curva en la curva ROC, como se aprecia en la Figura 36, mostrando una mayor habilidad para diferenciar entre clases que en la arquitectura anterior, que no empleaba mecanismos de atención. En la matriz de confusión es de notar que la red clasificó incorrectamente 187 casos malignos como benignos, obteniendo una sensibilidad del 82% aproximadamente. Esta y las otras métricas de validación, se pueden ver resumidas en la Tabla 4.

Tabla 4. Resultados en el conjunto de datos de test del modelo SE-ResNet.

Modelo	Exactitud (Accuracy)	Precisión	Sensibilidad (Recall)	Especificidad	F1 Score	ROC AUC
SE-ResNet	0,8483	0,8476	0,8168	0,875	0,8319	0,9304

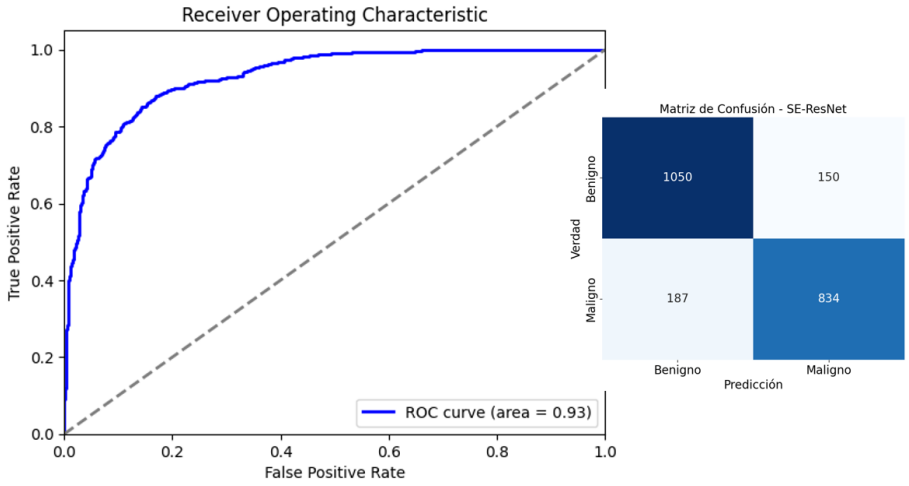


Figura 36. A la izquierda, la curva ROC del modelo SE-ResNet sobre el conjunto de datos de test. A la derecha la matriz de confusión sobre el mismo.

Modelo de imagen con mecanismos de atención: EfficientNet

EfficientNet B1

La Figura 37 muestra los resultados del modelo de imágenes con mecanismos de atención EfficientNet B1 respecto al conjunto de datos nunca visto durante el entrenamiento, o conjunto de test, en la que podemos ver la curva ROC, que muestra un área bajo la curva (AUC o ROC *Score*) de 0.94. Este valor es adecuado, ya que un ROC *Score* cercano a la unidad que indica que el modelo tiene una gran capacidad para distinguir entre las clases positivas y negativas. Llevando la atención a la matriz de confusión, se observa que, de los 1.200 casos benignos, 1.038 fueron clasificados correctamente como benignos, mientras que 162 fueron identificados incorrectamente como malignos. Del mismo modo, de los 1.021 casos malignos, 879 fueron clasificados correctamente como malignos y 142 fueron erróneamente identificados como benignos. Estos resultados muestran que el modelo hace un trabajo satisfactorio de distinguir entre las dos clases, aunque todavía hay algunos errores que deben ser considerados.

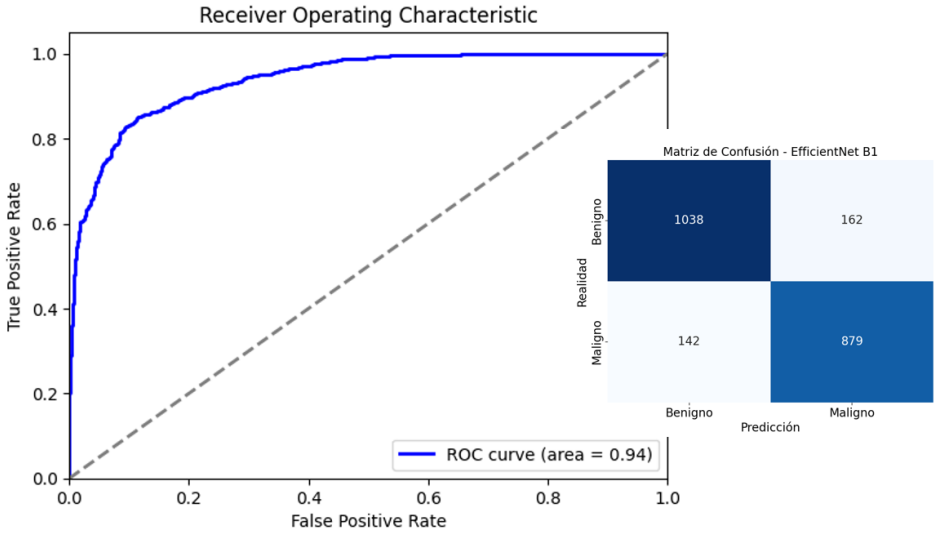


Figura 37. A la izquierda, la curva ROC del modelo EfficientNet B1 sobre el conjunto de datos de test. A la derecha la matriz de confusión sobre el mismo.

Por último, en la Tabla 5 se destacan varias métricas de rendimiento: la precisión es de 0,8444, el recall de 0,8609, la puntuación F1 de 0,8526 y la precisión general del modelo es del 86,31%.

Tabla 5. Resultados en el conjunto de datos de test del modelo EfficientNet B1.

Modelo	Exactitud (Accuracy)	Precisión	Sensibilidad (Recall)	Especificidad	F1 Score	ROC AUC
EfficientNet B1	0,8631	0,8444	0,8609	0,8650	0,8526	0,9397

EfficientNet B2

Durante la validación final, o conjunto de datos de test nunca vistos durante el entrenamiento del modelo de imagen con mecanismos de atención EfficientNet B2, se ha obtenido una exactitud del 86% con un *recall* o sensibilidad del 85,6%. La precisión también ha sido elevada con un 85,19%, obteniendo pues una puntuación F1 de 85,19% y un área bajo la curva de 0,93. Estos resultados quedan resumidos en la Tabla 6.

Tabla 6. Resultados en el conjunto de datos de test del modelo EfficientNet B2.

Modelo	Exactitud (Accuracy)	Precisión	Sensibilidad (Recall)	Especificidad	F1 Score	ROC AUC
EfficientNet B2	0,8631	0,8477	0,8560	0,8692	0,8519	0,9387

En la Figura 38 se muestra la curva ROC y la matriz de confusión obtenidas en el conjunto de test. El área bajo la curva es elevada, mostrando buen poder de diferenciación. De los 1.021 casos malignos en este conjunto, la red ha detectado exitosamente 874, dejando pasar 147. Aunque los resultados son positivos, existe un amplio margen de mejora en cuanto a sensibilidad de la red.

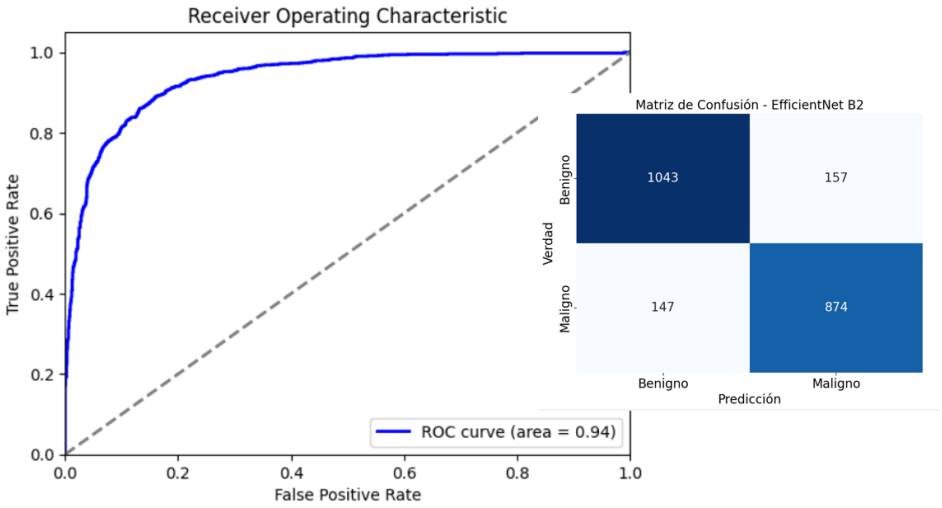


Figura 38. A la izquierda, la curva ROC del modelo EfficientNet sobre el conjunto de datos de test. A la derecha la matriz de confusión sobre el mismo.

EfficientNet B3

Los resultados en cuanto a curva ROC y matriz de confusión sobre el conjunto de datos nunca vistos durante entrenamiento o conjunto de datos de test del modelo EfficientNet B3 (modelo de imagen con mecanismos de atención), se encuentran representados en la Figura 39. Similar a los resultados de validación el área bajo la curva de esta red ha sido de 0,94 con 136 errores catalogados erróneamente como benignos y 163 clasificados erróneamente como malignos.

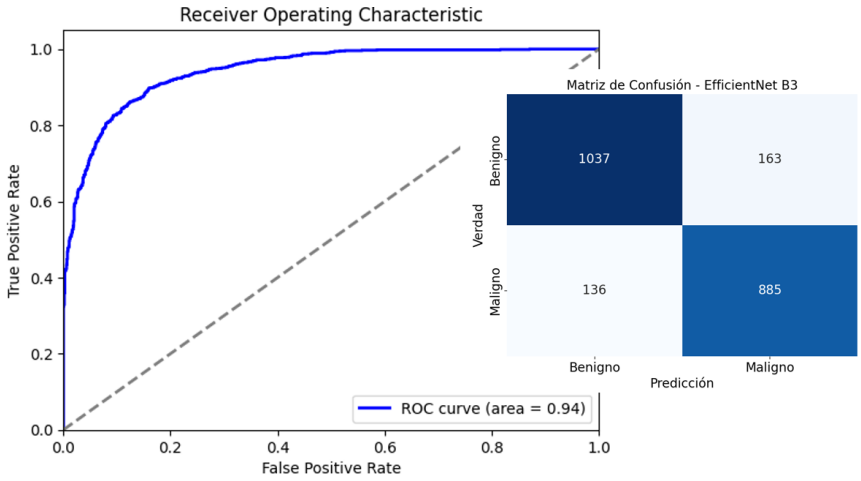


Figura 39. A la izquierda, la curva ROC del modelo EfficientNet B3 sobre el conjunto de datos de test. A la derecha la matriz de confusión sobre el mismo.

Finalmente, en la Tabla 7 se resumen las principales métricas del modelo, teniendo una exactitud de 0,8654, un Recall de 0,8668 y una precisión de 0,8445.

Tabla 7. Resultados en el conjunto de datos de test del modelo EfficientNet B3

Modelo	Exactitud (Accuracy)	Precisión	Sensibilidad (Recall)	Especificidad	F1 Score	ROC AUC
EfficientNet B3	0,8654	0,8445	0,8668	0,8642	0,8555	0,9438

EfficientNet B4

En la Figura 40, se muestran la curva ROC y la matriz de confusión del modelo EfficientNet B4 (modelo de imagen con mecanismos de atención). Podemos ver un área bajo la curva de 0,94, similar a otras versiones de EfficientNet, en este caso con 149 falsos negativos y 139 falsos positivos. El resumen de las métricas calculadas se puede ver en la Tabla 8.

Tabla 8. Resultados en el conjunto de datos de test del modelo EfficientNet B4

Modelo	Exactitud (Accuracy)	Precisión	Sensibilidad (Recall)	Especificidad	F1 Score	ROC AUC
EfficientNet B4	0,8703	0,8625	0,8541	0,8842	0,8583	0,9442

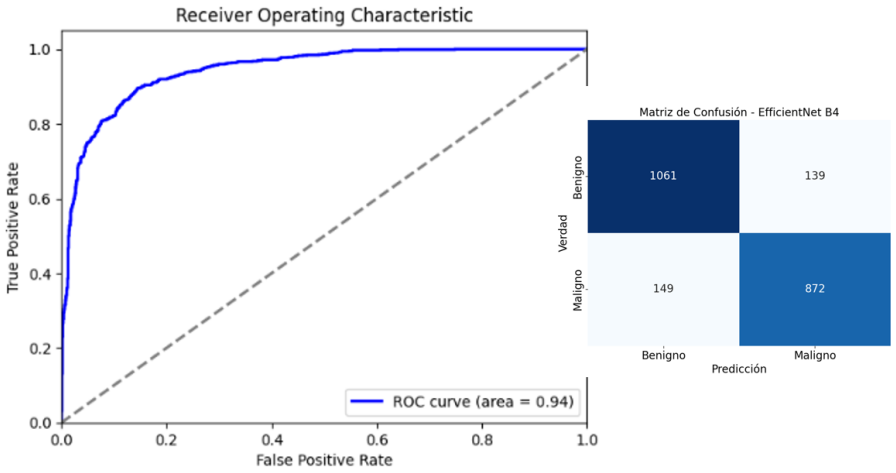


Figura 40. A la izquierda, la curva ROC del modelo EfficientNet B4 sobre el conjunto de datos de test. A la derecha la matriz de confusión sobre el mismo.

Resumen de los resultados de los modelos de imagen

Al comparar los diferentes modelos de imágenes en el conjunto de datos de test nunca visto durante el entrenamiento, se observan diferencias clave en su rendimiento y capacidades de discriminación. El modelo VGG16, sin mecanismos de atención, logra un área bajo la curva (AUC) de 0,90, pero muestra limitaciones en la precisión y sensibilidad, con una exactitud final del 80% y una sensibilidad del 81%. Además, presenta inestabilidad en la sensibilidad, lo que sugiere dificultades para clasificar correctamente lesiones malignas, probablemente debido a su arquitectura más simple y a la falta de mecanismos de atención.

En contraste, el modelo SE-ResNet, que utiliza bloques *Squeeze and Excitation* para mejorar la atención a características relevantes, presenta una mejora notable con un AUC de 0,93 y una sensibilidad ligeramente superior (82%), lo que indica una mejor capacidad para identificar lesiones malignas. Este modelo logra reducir los errores en comparación con VGG16, destacando la importancia de los mecanismos de atención en la mejora de la clasificación.

Las distintas versiones de EfficientNet (B1 a B4) muestran un rendimiento consistentemente alto en el conjunto de test, con ROC Score de hasta 0,94. Estas redes, optimizadas para eficiencia computacional y precisión mediante un escalado compuesto, superan a las otras arquitecturas en la mayoría de las métricas. Aunque algunos modelos presentan signos de sobreajuste, EfficientNet B4 logra un buen equilibrio entre precisión, sensibilidad y eficiencia, mostrando que puede discriminar mejor entre clases con menos errores de clasificación.

Comparar estos modelos ha permitido identificar cómo diferentes arquitecturas afectan la capacidad del modelo para clasificar lesiones cutáneas. Las diferencias en las métricas entre los modelos demuestran la importancia de incluir mecanismos de atención, como en SE-ResNet y EfficientNet, que mejoran la capacidad de los modelos para centrarse en características relevantes, aumentando la precisión y reduciendo los errores en la clasificación. Este análisis comparativo subraya la necesidad de explorar diversas arquitecturas para optimizar el rendimiento del modelo en tareas críticas de diagnóstico, como la clasificación de lesiones cutáneas.

7.1.4. Evaluación de los modelos de variables clínicas

Evaluación de los modelos de ML

La explicación de la selección de los modelos de ML empleados para el desarrollo se justifica en la sección Modelos de ML, y los resultados se pueden ver resumidos en la **Tabla 9**.

Tabla 9. Resumen de los resultados de las métricas en el conjunto de test para los modelos de ML con las variables clínicas.

Modelo	Exactitud (Accuracy)	Precisión	Sensibilidad (Recall)	Especificidad	F1 Score	ROC AUC
Regresión Logística	0,6994	0,6657	0,6938	0,7042	0,6795	0,7622
Random Forest	0,7037	0,6905	0,6445	0,7542	0,6667	0,7736
Gradient Boosting	0,6979	0,6639	0,6944	0,7542	0,7008	0,7715
SVM	0,7145	0,6997	0,6641	0,7542	0,6814	0,7458
KNN	0,7145	0,6997	0,6641	0,7542	0,6814	0,7150
TNN	0,7015	0,6711	0,6876	0,7133	0,6792	0,7702

Regresión logística

Los resultados mostrados en las imágenes representan la evaluación del modelo de regresión logística utilizando el conjunto de datos de test. El gráfico ROC a la izquierda de la **Figura 41** muestra la relación entre la tasa de verdaderos positivos (*True Positive Rate*) y la tasa de falsos positivos (*False Positive Rate*) para el modelo de regresión logística. El área bajo la curva (AUC) es de 0,76, lo cual indica la capacidad del modelo para distinguir entre las clases benigno y maligno. Un AUC de 0,76 sugiere que el modelo puede discriminar entre las clases, pero este resultado es notablemente inferior que en los modelos que emplean únicamente las imágenes de las lesiones.

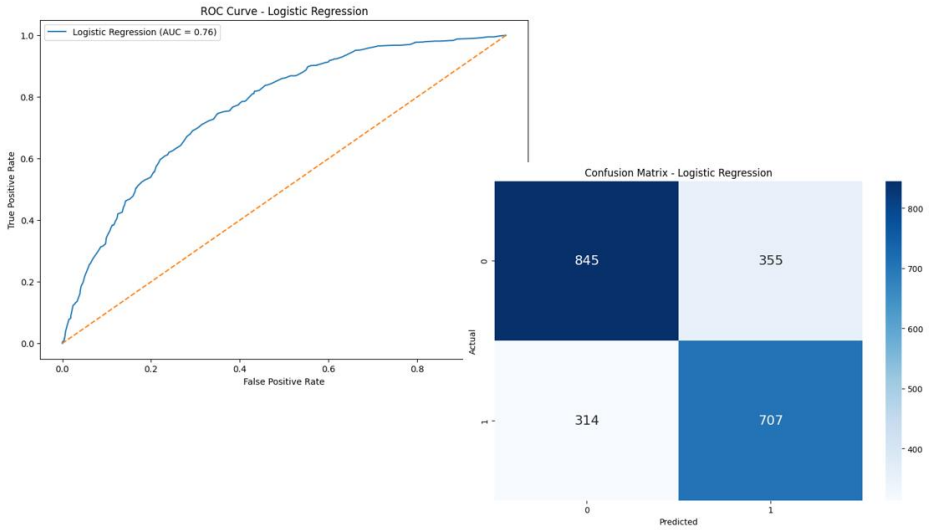


Figura 41. A la izquierda, la curva ROC del modelo de regresión logística sobre el conjunto de datos de test. A la derecha la matriz de confusión sobre el mismo.

La matriz de confusión, a la derecha, proporciona un resumen visual del desempeño del modelo en términos de predicciones correctas e incorrectas. Los resultados indican que hubo 845 casos clasificados correctamente como benignos y 707 casos correctamente como malignos. Sin embargo, también se observaron errores: 355 casos de lesiones benignas fueron clasificadas erróneamente como malignas, y 312 casos de lesiones malignas fueron clasificadas erróneamente como benignas.

En la **Tabla 9** se recogen los resultados de este y otros modelos de ML, mostrando aproximadamente un 70% de exactitud, un 69% de sensibilidad o Recall y una precisión del 67%.

Random Forest

En la evaluación final del modelo empleando *Random Forest* se ha obtenido, como se muestra en la **Figura 42**, un área bajo la curva ROC de aproximadamente 0,77, similar al modelo anterior de regresión logística. Habiendo clasificado erróneamente 363 casos malignos como benignos, se ha calculado una sensibilidad o Recall de 0,64 y una precisión de aproximadamente el 69%. El resto de métricas de interés están resumidas en la **Tabla 9**.

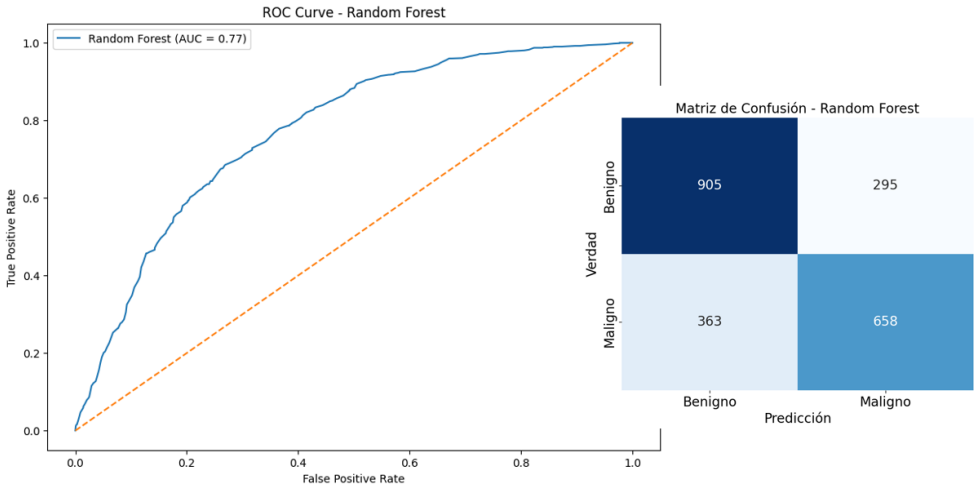


Figura 42. A la izquierda, la curva ROC del modelo SE-ResNet sobre el conjunto de datos de test. A la derecha la matriz de confusión sobre el mismo.

Gradient Boosting

En cuanto al modelo de *Gradient Boosting*, tal y como se observa en la **Figura 43**, el modelo ha sido capaz de discernir con éxito 841 casos benignos y 709 casos malignos, obteniendo una exactitud aproximada del 70%. Además, la red ha categorizado erróneamente 312 casos malignos como benignos y 359 casis benignos como malignos, lo cual equivale a una precisión de 0,66, un *Recall* de 0,69 y una especificidad de 0,70. Estos resultados se complementan con un área bajo la curva de ROC de 0,77.

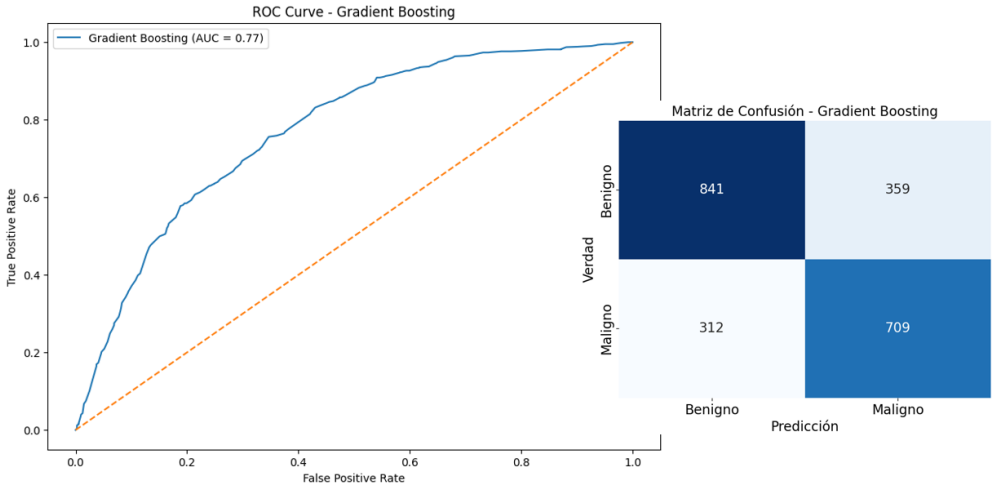


Figura 43. A la izquierda, la curva ROC del modelo *Gradient Boosting* sobre el conjunto de datos de test. A la derecha la matriz de confusión sobre el mismo.

Support Vector Machine (SVM)

En la **Figura 44** se grafican los resultados de la curva ROC y la matriz de confusión del modelo *Support Vector Machine*. Los resultados muestran 1.587 casos clasificados correctamente, siendo 909 de ellos pertenecientes a la clase benigna y 678 a la clase maligna. El modelo ha obtenido en el conjunto de datos de test unas métricas que, como se resumen en la **Tabla 9**, reúnen una exactitud de 0,71, una precisión de 0,69, una sensibilidad de 0,66 y una especificidad de 0,75. Finalmente, el área bajo la curva obtenida es de 0,75 aproximadamente.

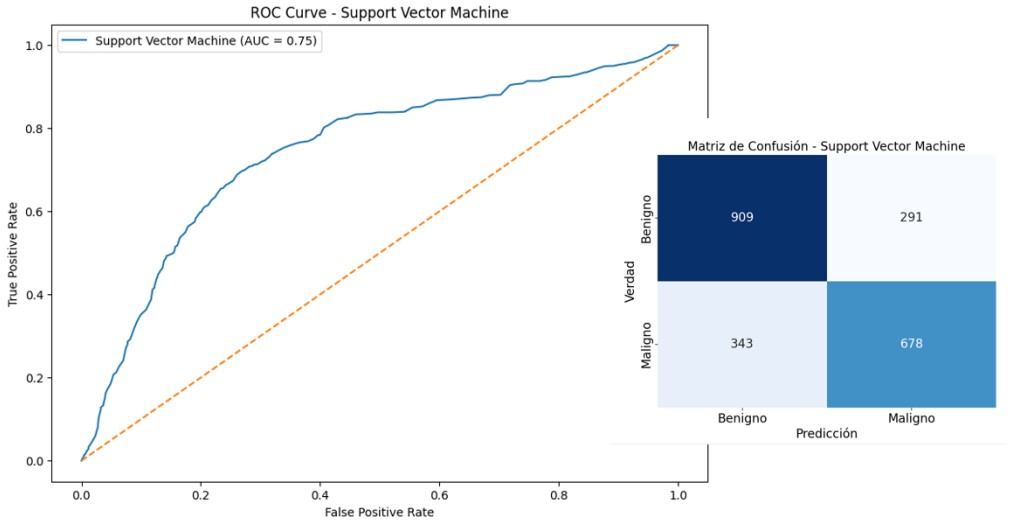


Figura 44. A la izquierda, la curva ROC del modelo SVM sobre el conjunto de datos de test. A la derecha la matriz de confusión sobre el mismo.

K-Nearest Neighbors (KNN)

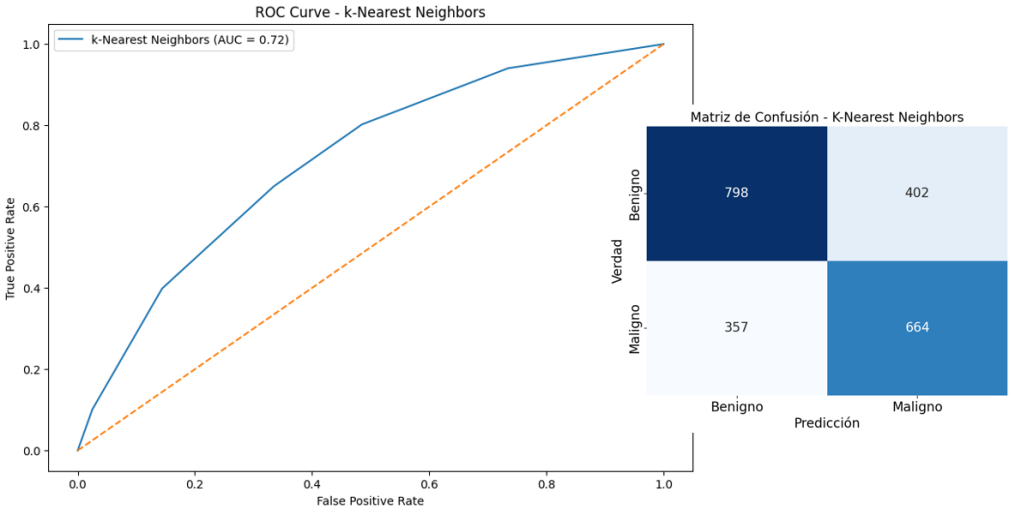


Figura 45. A la izquierda, la curva ROC del modelo KNN sobre el conjunto de datos de test. A la derecha la matriz de confusión sobre el mismo.

Respecto al modelo KNN y como podemos ver resuido en la **Figura 45** y la **Tabla 9**, el modelo ha obtenido un rendimiento en cuanto a exactitud de 66% aproximadamente habiendo acertado 798 lesiones benignas y 664 lesiones malignas. También se ha calculado una precisión de 0,62, una sensibilidad de 0,65 y una espcificidad de 0,66 con un área bajo la curva ROC de 0,72.

Evaluación del modelo de DL para datos clínicos (TNN)

Quedan resumidos los resultados durante la fase de evaluación final del modelo de DL para datos clínicos, en la **Figura 46** y la Tabla 9, obteniendo una exactitud de 0,70 una precisión de 0,67, una sensibilidad de 0,68 y una especificidad de 0,71. Finalmente, el área bajo la curva ROC de este modelo es aproximadamente de 0,77. El significado y las causas para estos resultados se dicuten en el apartado 8.1 Análisis crítico de los resultados

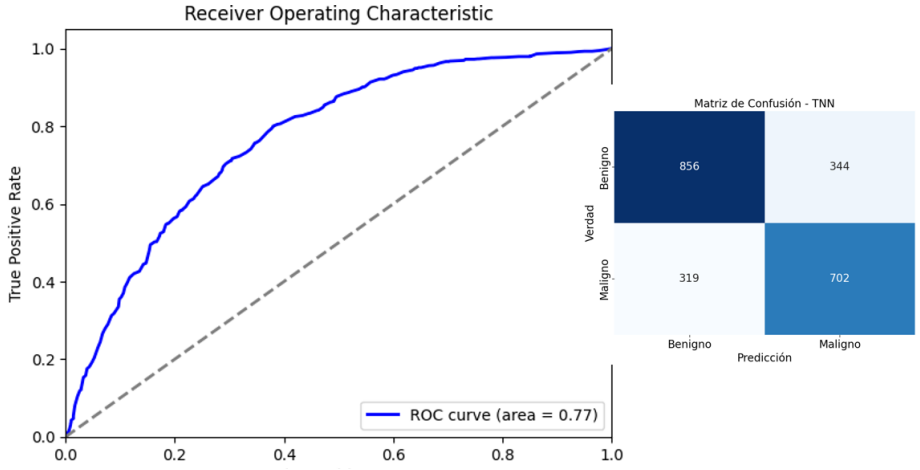


Figura 46. A la izquierda, la curva ROC del modelo TNN sobre el conjunto de datos de test. A la derecha la matriz de confusión sobre el mismo.

7.1.5. Evaluación del modelo integrador o multimodal

El modelo integrador o multimodal, emplea como datos de entrada las imágenes de las lesiones y las variables clínicas. Como se muestra en la Figura 47 y la Tabla 10 las métricas han sido ligeramente mejores que en modelos anteriores, obteniendo un área bajo la curva de aproximadamente 0,95 en el conjunto de datos de test, errando únicamente en 272 lesiones, habiendo clasificado 123 como benignas siendo malignas y 149 como malignas siendo benignas. Esto significa una exactitud del 88%, una precisión de 86%, una sensibilidad de 88% con una especificidad de 88%. Obtenemos así la mayor área bajo la curva hasta el momento, con un valor de 0,95 aproximadamente.

Tabla 10. Resultados en el conjunto de datos de test del modelo de red neuronal multimodal.

Modelo	Exactitud (Accuracy)	Precisión	Sensibilidad (Recall)	Especificidad	F1 Score	ROC AUC
Multimodal	0,8775	0,8577	0,8795	0,8758	0,8685	0,9494

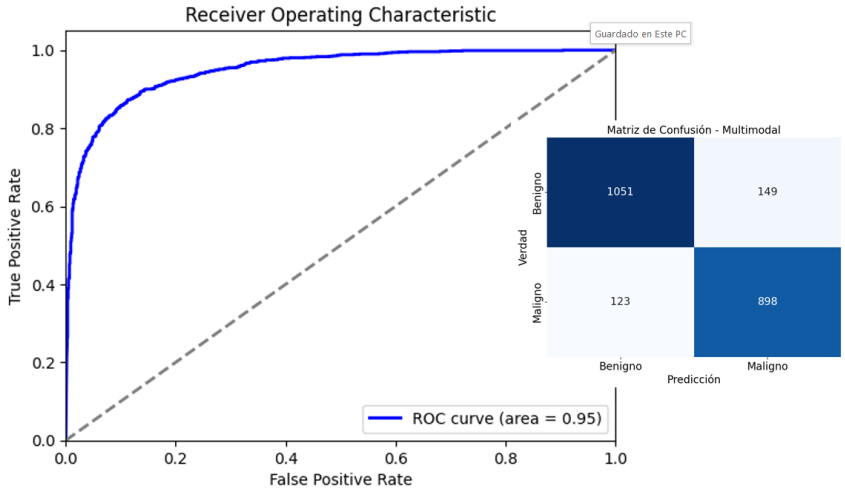


Figura 47. A la izquierda, la curva ROC de la red neuronal multimodal sobre el conjunto de datos de test. A la derecha la matriz de confusión sobre el mismo.

7.1.6. Evaluación de los modelos ensamblados

Se han resumido las métricas de validación del ensamblado de modelos en la Tabla 11. En general, los valores se encuentran en la linea del resto de modelos, con métricas por encima del 0,8.

Tabla 11. Resultados del ensamblado de modelos.

Modelo	Exactitud (Accuracy)	Precisión	Sensibilidad (Recall)	Especificidad	F1 Score	ROC AUC
Majority Voting (MV)	0,8460	0,8398	0,8217	0,8667	0,8307	0,8442
Weighted MV	0,8667	0,8636	0,8433	0,8867	0,8533	0,8650
Stacking 1: Logistic Regression	0,8879	0,8807	0,8746	0,8992	0,8776	0,9517
Stacking 2: Random Forest	0,8874	0,8806	0,8737	0,8992	0,8771	0,9484
Stacking 3: Gradient Boosting	0,8843	0,8694	0,8805	0,8875	0,8749	0,9542
Staking 4: SVM	0,8892	0,8825	0,8756	0,9008	0,8791	0,9146

Voto mayoritario o *Majority Voting*

En la Figura 48 se muestra la curva y la matriz de confusión obtenidas mediante el método de *majority voting*. En este caso, el área bajo la curva (AUC) es de 0,84, lo que indica un buen rendimiento, aunque notablemente menor que en modelos de imagen.

La matriz de confusión complementa esta evaluación proporcionando una visión detallada de las predicciones del modelo. En la matriz, se observa que el modelo ha clasificado correctamente 1.040 casos de lesiones benignas y 839 casos de lesiones malignas. Sin embargo, se han producido 160 falsos positivos (lesiones benignas clasificadas como malignas) y 182 falsos negativos (lesiones malignas clasificadas como benignas). Estos resultados reflejan una buena capacidad del modelo para distinguir entre lesiones benignas y malignas, aunque con un margen de mejora en la reducción de falsos negativos para asegurar una detección más precisa de lesiones malignas.

En cuanto a las métricas de evaluación, se presentan en la Tabla 11 que la exactitud del modelo es del 84%, lo que significa que un alto porcentaje de las predicciones del modelo son correctas. La precisión y la sensibilidad son del 84% y 82%, respectivamente, lo que indica un equilibrio adecuado entre la capacidad del modelo para identificar correctamente las lesiones malignas y minimizar los falsos positivos. La especificidad del 87% muestra que el modelo también es eficaz en la identificación de lesiones benignas. El F1 Score de 0,83, que es una medida combinada de precisión y sensibilidad, sugiere que el modelo mantiene un buen rendimiento general.

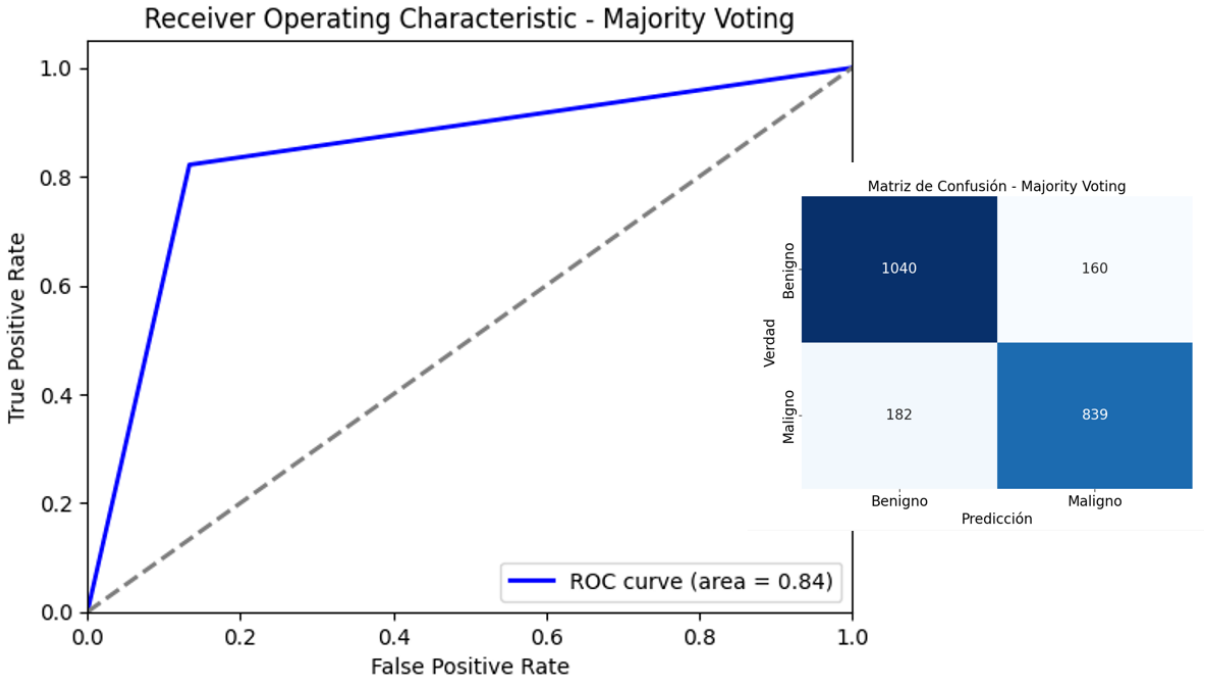


Figura 48. A la izquierda, la curva ROC del modelo ensamblado mediante voto de la mayoría sobre el conjunto de datos de test. A la derecha la matriz de confusión sobre el mismo.

Voto ponderado o Weighted Majority Voting

En la Figura 49 se muestra la curva ROC junto con la matriz de confusión obtenida mediante el método de *Weighted Majority Voting*. Para este modelo, se ha obtenido un área bajo la curva (AUC) de 0.86, lo que refleja su capacidad para discriminar entre lesiones cutáneas malignas y benignas.

La matriz de confusión detalla el desempeño del modelo en términos de predicciones correctas e incorrectas. Se puede observar que el modelo ha identificado correctamente 1.064 lesiones benignas y 861 lesiones malignas. Sin embargo, ha cometido 136 errores al clasificar lesiones benignas como malignas (falsos positivos) y 160 errores al clasificar lesiones malignas como benignas (falsos negativos). Esto demuestra una mejora en comparación con el método de *Majority Voting*, especialmente en la reducción de falsos negativos, lo que es crucial para la detección temprana de lesiones malignas.

La exactitud del modelo, que se sitúa en un 85%, indica que una gran mayoría de las predicciones realizadas son correctas. La precisión, con un valor del 86%, muestra la proporción de verdaderos positivos entre las predicciones positivas realizadas. La sensibilidad, con un 84%, refleja la capacidad del modelo para identificar correctamente las lesiones malignas. La especificidad, que alcanza un 89%, indica la eficacia del modelo en identificar las lesiones benignas. Finalmente, el F1 Score de 0,85 representa un equilibrio entre precisión y sensibilidad, destacando el buen rendimiento general del modelo.

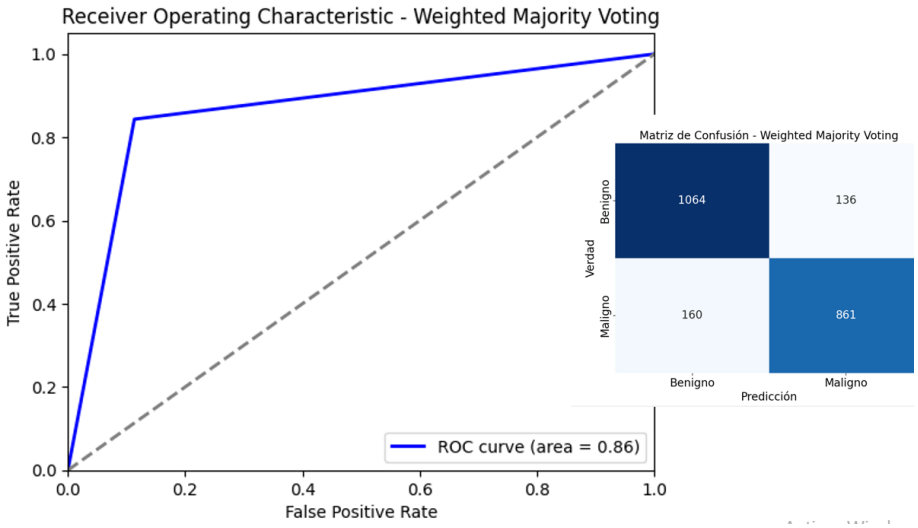


Figura 49. A la izquierda, la curva ROC del modelo ensamblado mediante voto ponderado sobre el conjunto de datos de test. A la derecha la matriz de confusión sobre el mismo.

Stacking

Elección de características

Para la elaboración de metamodelos empleando las salidas de los modelos anteriores, se disponían de cada modelo un vector con probabilidades de malignidad y una de clase obtenida. Para seleccionar el enfoque más adecuado para las características de entrada de estos modelos se han analizado las métricas claves empleando sólo las clases predichas, sólo las probabilidades y finalmente ambas.

En general, el uso de probabilidades (variable numérica continua) mejora los resultados respecto a las clases predichas (variable binaria). El uso de solo clases da unos resultados competentes, pero no alcanzan los valores más altos observados tal y como se muestra en la Tabla 12.

El mejor enfoque es el de uso de probabilidades y clases ya que presentan los valores más elevados de exactitud, precisión y F1 score. El ROC AUC de 0,9523 destaca como el mejor valor, reflejando un rendimiento elevado del modelo en cuanto a la discriminación entre clases.

En los apartados siguientes se evaluaron los resultados de los modelos usando como características de entrada las clases predichas y las probabilidades de los modelos simples. Elegir el enfoque que alimenta los modelos de stacking con ambas características, tanto clases como probabilidades, está justificado por varias razones. Utilizar tanto las clases como las probabilidades proporciona una mayor cantidad de información al modelo de stacking. Las clases ofrecen una decisión categórica, mientras que las probabilidades proporcionan información sobre el grado de certeza de las predicciones individuales.

Los resultados muestran que este enfoque de combinar predicciones y probabilidades como entrada de los modelos de stacking, obtiene las mejores métricas en términos de exactitud, precisión, F1 Score y ROC AUC, lo cual indica un mejor rendimiento general del modelo. Además, incorporar ambas características permite al modelo de stacking ser más robusto y flexible al manejar diferentes tipos de entradas, mejorando su capacidad para generalizar sobre datos nuevos. Al combinar la información probabilística y categórica, se reduce la probabilidad de cometer errores en la clasificación, especialmente en casos difíciles o ambiguos. Por ello, a partir de este punto se comentarán los resultados de este enfoque.

Tabla 12. Resultados de los modelos de Stacking sobre las clases predichas de los modelos (verde), las probabilidades (rosa) y ambas (azul).

	Características	Exactitud (Accuracy)	Precisión	Sensibilidad (Recall)	Especificidad	F1 Score	ROC AUC
Clase	Stacking 1: Logistic Regression	0,8897	0,8872	0,8707	0,9058	0,8789	0,9409
	Stacking 2: Random Forest	0,8847	0,8798	0,8678	0,8992	0,8738	0,922
	Stacking 3: Gradient Boosting	0,8847	0,8768	0,8717	0,8958	0,8743	0,9413
	Staking 4: SVM	0,8811	0,8766	0,8629	0,8967	0,8697	0,9132
Probabilidades	Stacking 1: Logistic Regression	0,8879	0,8807	0,8746	0,8992	0,8776	0,9517
	Stacking 2: Random Forest	0,8874	0,8806	0,8737	0,8992	0,8771	0,9484
	Stacking 3: Gradient Boosting	0,8843	0,8694	0,8805	0,8875	0,8749	0,9542
	Staking 4: SVM	0,8892	0,8825	0,8756	0,9008	0,8791	0,9146
Clase + Probabilidades	Stacking 1: Logistic Regression	0,8906	0,8859	0,8746	0,9042	0,8802	0,9523
	Stacking 2: Random Forest	0,8883	0,8808	0,8756	0,8992	0,8782	0,9509
	Stacking 3: Gradient Boosting	0,8838	0,8686	0,8805	0,8867	0,8745	0,9543
	Staking 4: SVM	0,8847	0,8798	0,8678	0,8992	0,8738	0,9276

Regresión logística o Logistic Regression

En la Figura 50 se presenta la curva ROC y la matriz de confusión obtenidas mediante el modelo de regresión logística. La curva ROC es una herramienta crítica para evaluar el rendimiento de un modelo de clasificación, mostrando la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos en diversos umbrales de decisión. En este caso, el área bajo la curva (AUC) es de 0,95, lo que indica un rendimiento bueno del modelo en la discriminación entre lesiones cutáneas benignas y malignas.

Según la matriz de confusión mostrada en la Figura 50, el modelo ha identificado correctamente 1.085 lesiones benignas y 893 lesiones malignas. Sin embargo, ha cometido 115 errores al clasificar lesiones benignas como malignas (falsos positivos) y 128 errores al clasificar lesiones malignas como benignas

(falsos negativos). Estos resultados reflejan una notable capacidad del modelo para distinguir entre ambos tipos de lesiones, con un menor número de falsos negativos en comparación con otros métodos.

Las métricas de evaluación clave aparecen resumidas en la Tabla 13. La exactitud del modelo es del 89%, lo que indica que una gran mayoría de las predicciones son correctas. La precisión del 89% muestra la proporción de verdaderos positivos entre las predicciones positivas realizadas. La sensibilidad del 87% refleja la capacidad del modelo para identificar correctamente las lesiones malignas. La especificidad del 91% indica la eficacia del modelo en la identificación de lesiones benignas. Finalmente, el F1 Score de 0,88 sugiere un buen equilibrio entre precisión y sensibilidad, destacando un satisfactorio rendimiento general del modelo.

Tabla 13. Resultados del ensamblado mediante Stacking en el conjunto de datos de test.

Modelo	Exactitud (Accuracy)	Precisión	Sensibilidad (Recall)	Especificidad	F1 Score	ROC AUC
Stacking 1: Logistic Regression	0,8906	0,8859	0,8746	0,9042	0,8802	0,9523
Stacking 2: Random Forest	0,8883	0,8808	0,8756	0,8992	0,8782	0,9509
Stacking 3: Gradient Boosting	0,8838	0,8686	0,8805	0,8867	0,8745	0,9543
Staking 4: SVM	0,8847	0,8798	0,8678	0,8992	0,8738	0,9276

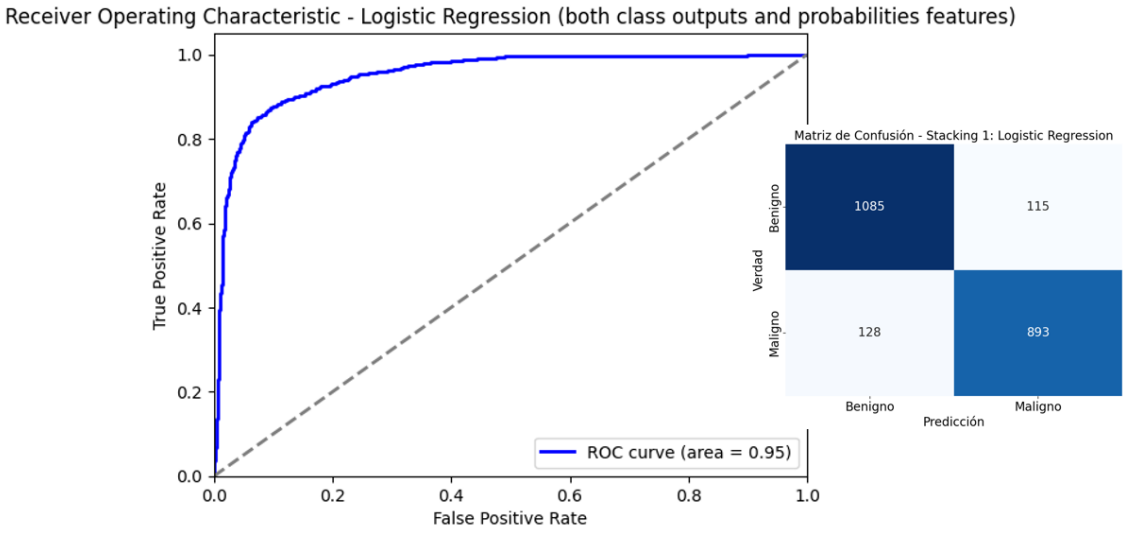


Figura 50. A la izquierda, la curva ROC del modelo de Stacking mediante regresión logística sobre el conjunto de datos de test. A la derecha la matriz de confusión sobre el mismo.

Random Forest

La curva ROC y la matriz de confusión que se obtuvo utilizando el modelo de *Random Forest* se muestran en las figuras. El modelo muestra una capacidad para distinguir entre lesiones cutáneas benignas y malignas que se refleja en el área bajo la curva (AUC) de este caso de 0,95, en la Figura 51.

El 88% de precisión del modelo significa que la mayoría de las predicciones son correctas. La proporción de verdaderos positivos entre las predicciones positivas realizadas se muestra con una precisión del 88%. La capacidad del modelo para identificar correctamente las lesiones malignas se refleja en la sensibilidad del 88%. La especificidad del 90% demuestra que el modelo es efectivo para

identificar lesiones benignas. Finalmente, con un F1 Score de 0,88, se puede ver un buen equilibrio entre precisión y sensibilidad, lo que destaca el rendimiento excepcional general del modelo.

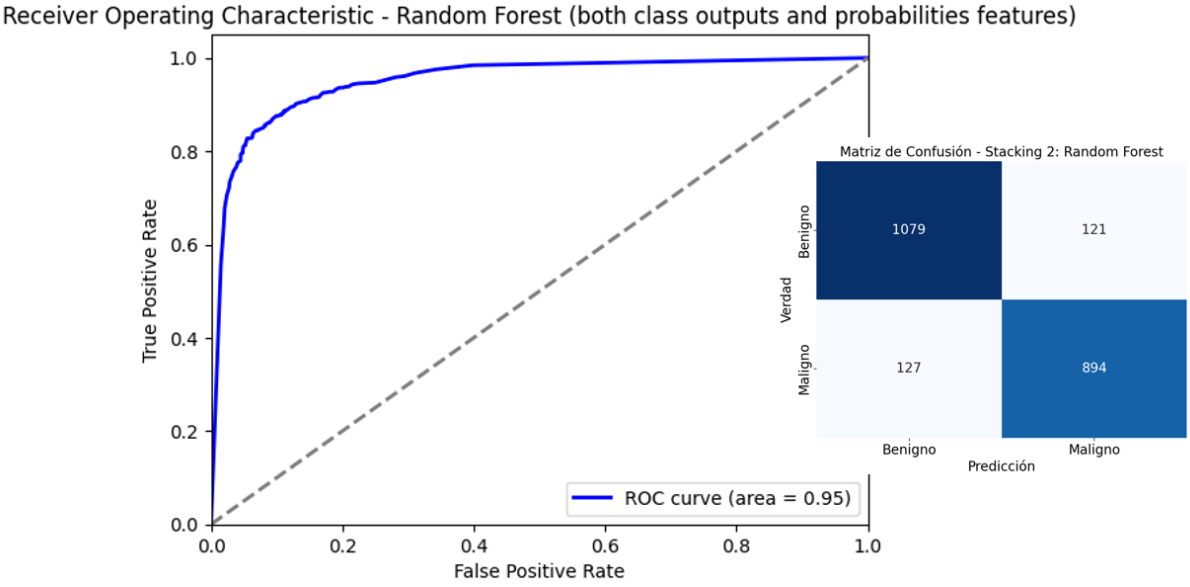


Figura 51. A la izquierda, la curva ROC del modelo de Stacking mediante *Random Forest* sobre el conjunto de datos de test. A la derecha la matriz de confusión sobre el mismo.

Gradient Boosting

El modelo de stacking mediante *gradient boosting* ha obtenido un área bajo la curva de la curva de ROC de 0,95, similar al resto de modelos de las mismas características, clasificando correctamente el 88% de las lesiones. El modelo ha clasificado incorrectamente 122 casos malignos como benignos y 136 casos benignos como malignos. Estos resultados se muestran gráficamente en la Figura 52.

En cuanto a las métricas obtenidas, el modelo posee una exactitud del 88%, con una precisión de 87% y un Recall del 88%, la especificidad es similar siendo esta de 88,67%.

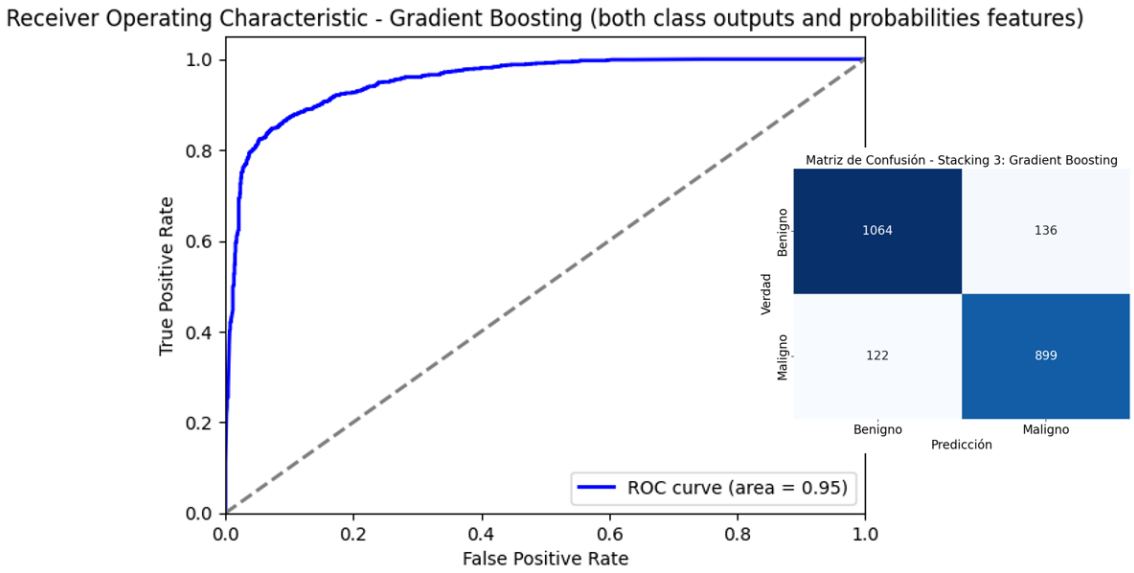


Figura 52. A la izquierda, la curva ROC del modelo de Stacking mediante Gradient Boosting. A la derecha la matriz de confusión sobre el mismo.

Los resultados son relativamente similares al resto de modelos de stacking, pero con un resultado ligeramente inferior respecto a especificidad y precisión y ligeramente superior en cuanto a sensibilidad y área bajo la curva.

SVM

En la Figura 53 se presenta la curva ROC la matriz de confusión obtenidas mediante el modelo de SVM. En este caso, el área bajo la curva (AUC) es de 0,93, lo que refleja una buena capacidad del modelo para discriminar entre lesiones cutáneas benignas y malignas.

Acorde a la matriz de confusión representada en la Figura 53, el modelo ha identificado correctamente 1.079 lesiones benignas y 886 lesiones malignas. Sin embargo, ha cometido 121 errores al clasificar lesiones benignas como malignas (falsos positivos) y 135 errores al clasificar lesiones malignas como benignas (falsos negativos). Estos resultados demuestran una notable precisión en la clasificación de lesiones cutáneas, aunque con un ligero incremento en falsos negativos en comparación con el modelo de *Random Forest.*

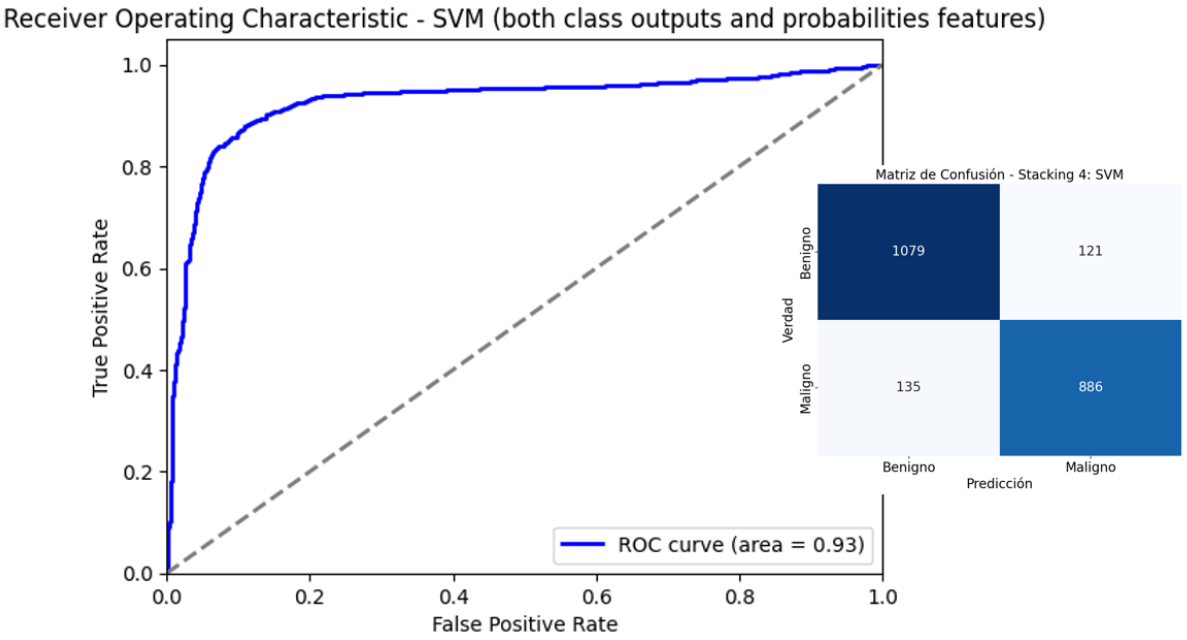


Figura 53. A la izquierda, la curva ROC del modelo de Stacking mediante SVM. A la derecha la matriz de confusión sobre el mismo.

La exactitud del modelo es del 88%, lo que indica que una alta proporción de las predicciones son correctas. La precisión del 88% muestra la proporción de verdaderos positivos entre las predicciones positivas realizadas. La sensibilidad del 87% refleja la capacidad del modelo para identificar correctamente las lesiones malignas. La especificidad del 90% indica la eficacia del modelo en la identificación de lesiones benignas. Finalmente, el F1 Score de 0,88 sugiere un buen equilibrio entre precisión y sensibilidad, destacando el rendimiento general notable del modelo.

7.2. Selección del modelo óptimo

7.2.1. Comparativa de rendimiento de los modelos

En la Tabla 14 se pueden ver resumidas las métricas de validación del conjunto de datos de test (*Holdout*, nunca vistos durante entrenamiento) para todos los modelos entrenados con el conjunto de datos equilibrado. En rosa, se encuentra el experimento de red neuronal convolucional sin atención. En

naranja, las redes neuronales que emplean mecanismos de atención, entre ellas SE-ResNet y Las versiones 1 a 4 de EfficientNet. Podemos ver un patrón de resultados, siendo en general la exactitud de los modelos que emplean las imágenes, superior al 80%, sin embargo, existe una diferencia notable en las métricas de precisión y especificidad entre los modelos que emplean mecanismos de atención, siendo estos últimos, en general, mejores en distinguir entre clases. Esto también se ve reflejado en el área bajo la curva de la curva ROC, siendo la diferencia mínima de un 3% entre el modelo convolucional y el convolucional con atención.

Por otro lado, podemos ver que los datos clínicos por si solos tienen unos resultados muy inferiores al uso único de imágenes, siendo el desempeño máximo en cuanto a exactitud de aproximadamente un 71% con sensibilidades entorno al 69% en el mejor de los casos. Esto representa una limitación intrínseca de los datos clínicos por si solos para detectar malignidad, que se discutirá más adelante.

Sin embargo, al combinar los datos clínicos con las imágenes, el desempeño del modelo mejora en comparación al caso de uso de las imágenes por si solas. La mejora es discreta, mejorando un 2% la sensibilidad respecto al mejor modelo de imágenes, superando en un 0,52% el área bajo la curva del mejor modelo de imágenes.

Finalmente, en verde podemos ver los modelos de ensamblados. Cabe destacar, que los modelos que empleaban las estrategias de votos, tanto voto simple como voto ponderado, en general empeoraban los resultados de las redes neuronales de imágenes, debido a que emplea todos los modelos para calcular la clase, y como se ha mencionado, los modelos de datos clínicos solo tenían un comportamiento menos acertado que los homólogos de imágenes.

El uso de estrategias de stacking con características de clase y probabilidades ha demostrado mejorar el comportamiento, siendo el área bajo la curva un 1% superior que el mejor de los modelos simples. Obtenido los mejores resultados en cuantos a AUC ROC y sensibilidad con un 95,43% y un 88,05% respectivamente.

Tabla 14. Resumen de los resultados obtenidos por los modelos empleando el conjunto de datos equilibrado.

	Modelo	Exactitud (Accuracy)	Precisión	Sensibilidad (Recall)	Especificidad	F1 Score	ROC AUC
CNN	VGG16 modificado	0,8041	0,7738	0,8110	0,7983	0,7920	0,9000
CNN + Atención	SE-ResNet	0,8483	0,8476	0,8168	0,8750	0,8319	0,9304
	EfficientNet B1	0,8631	0,8444	0,8609	0,8650	0,8526	0,9380
	EfficientNet B2	0,8631	0,8477	0,8560	0,8692	0,8519	0,9387
	EfficientNet B3	0,8654	0,8445	0,8668	0,8642	0,8555	0,9438
	EfficientNet B4	0,8703	0,8625	0,8541	0,8842	0,8583	0,9442
ML datos clínicos	Regresión Logística	0,6994	0,6657	0,6938	0,7042	0,6795	0,7622
	Random Forest	0,7037	0,6905	0,6445	0,7542	0,6667	0,7736
	Gradient Boosting	0,6979	0,6639	0,6944	0,7542	0,7008	0,7715
	SVM	0,7145	0,6997	0,6641	0,7542	0,6814	0,7458
	KNN	0,7145	0,6997	0,6641	0,7542	0,6814	0,7150
	TNN	0,7015	0,6711	0,6876	0,7133	0,6792	0,7702
	Multimodal	0,8775	0,8577	0,8795	0,8758	0,8685	0,9494
Ensamblado	Majority Voting (MV)	0,8460	0,8398	0,8217	0,8667	0,8307	0,8442
	Weighted MV	0,8667	0,8636	0,8433	0,8867	0,8533	0,8650
	Stacking 1: Logistic Regression	0,8906	0,8859	0,8746	0,9042	0,8802	0,9523
	Stacking 2: Random Forest	0,8883	0,8808	0,8756	0,8992	0,8782	0,9509
	Stacking 3: Gradient Boosting	0,8838	0,8686	0,8805	0,8867	0,8745	0,9543
	Staking 4: SVM	0,8847	0,8798	0,8678	0,8992	0,8738	0,9276

7.2.2. Justificación de la selección del modelo final

Tras evaluar exhaustivamente los diferentes modelos desarrollados, se observa que, aunque los modelos basados en técnicas de *stacking* (como el *Gradient Boosting*) presentan una ligera ventaja en términos de métricas de rendimiento, como el área bajo la curva (AUC) y la sensibilidad, esta mejora marginal viene acompañada de una carga computacional significativamente mayor. Esta carga computacional se debe al hecho de que los modelos de *stacking* requieren la ejecución de múltiples algoritmos de clasificación (incluidos modelos de aprendizaje profundo y automático) para cada instancia, lo que implica un tiempo de procesamiento mucho más largo y un consumo elevado de recursos.

En un entorno de práctica clínica, donde la rapidez y eficiencia del diagnóstico son críticas, esta complejidad adicional podría traducirse en limitaciones significativas. La necesidad de clasificar cada muestra utilizando todos los modelos desarrollados no solo incrementa el tiempo de respuesta, sino que también exige infraestructuras tecnológicas más avanzadas y costosas, lo cual puede ser poco práctico para su implementación en entornos de atención primaria o en hospitales con recursos limitados.

Por otro lado, el modelo multimodal que integra imágenes dermoscópicas y datos clínicos ofrece una solución equilibrada y eficiente. Este modelo aprovecha la capacidad de las redes neuronales convolucionales (CNN) para identificar patrones visuales en las imágenes dermoscópicas y la capacidad de los modelos de aprendizaje automático para procesar variables clínicas relevantes, logrando una integración coherente de ambas fuentes de información. Aunque los resultados de este modelo en términos de AUC y sensibilidad son ligeramente inferiores a los del mejor modelo de stacking, el modelo multimodal proporciona una mejora significativa en comparación con el uso exclusivo de imágenes o datos clínicos, lo que demuestra su valor en un contexto clínico real.

Además, el modelo multimodal mantiene una carga computacional moderada, adecuada para su uso en tiempo real, sin requerir la ejecución simultánea de múltiples algoritmos. Esto lo hace más viable para su integración en flujos de trabajo clínicos, donde el equilibrio entre precisión diagnóstica, velocidad de procesamiento y costo es fundamental. En términos de aplicabilidad, el modelo multimodal facilita un diagnóstico más completo, considerando tanto la información visual como los datos clínicos, lo cual puede aumentar la confianza de los profesionales de la salud en los resultados del modelo.

Finalmente, la explicabilidad del modelo multimodal, al combinar datos de imagen y variables clínicas, permite una interpretación más clara de los factores que contribuyen a las decisiones diagnósticas, facilitando su aceptación entre los médicos y otros profesionales de la salud. Este enfoque integrador respalda un diagnóstico más informado, ayudando a reducir tanto los falsos positivos como los falsos negativos de manera eficiente y adaptándose mejor a las necesidades prácticas del entorno clínico.

En conclusión, la elección del modelo multimodal se justifica no solo por su rendimiento sólido y equilibrado, sino también por su menor carga computacional, su facilidad de implementación y su potencial para mejorar la práctica clínica mediante una integración efectiva de múltiples fuentes de datos.

7.2.3. Análisis de precisión, sensibilidad y especificidad en la clasificación de malignidad

En el contexto de herramientas de cribado y diagnóstico médico, es crucial mantener un equilibrio adecuado entre precisión, sensibilidad y especificidad. Estas métricas son fundamentales para evaluar la efectividad de las pruebas diagnósticas y determinar su utilidad en la práctica clínica. Estas métricas ofrecen una visión complementaria de la eficacia de una prueba diagnóstica, y una sola de estas no es suficiente para tener una visión global del rendimiento y la utilidad en el entorno clínico.

En el caso de la sensibilidad, esta es crucial en situaciones donde es fundamental detectar todos los casos posibles de una enfermedad, como en el cribado de enfermedades graves o contagiosas. Una alta sensibilidad asegura que pocos casos verdaderos se pasen por alto, reduciendo el riesgo de falsos negativos. Sin embargo, una sensibilidad alta a menudo puede ir acompañada de una menor especificidad. En el cribado, una alta especificidad es esencial para minimizar los falsos positivos, que pueden llevar a tratamientos innecesarios, ansiedad y costos adicionales. Es particularmente importante en enfermedades donde un diagnóstico falso positivo podría resultar en intervenciones perjudiciales o innecesarias.

El equilibrio entre sensibilidad y especificidad es crítico porque mejorar una de estas métricas a menudo puede resultar en la disminución de la otra. En herramientas de cribado, este equilibrio depende del contexto clínico y de las consecuencias de los falsos positivos y falsos negativos. Por ejemplo, en el cribado de cáncer, es preferible tener una alta sensibilidad para asegurarse de detectar la mayoría de los casos, incluso si eso significa aceptar una menor especificidad y algunos falsos positivos. En contraste, para enfermedades donde un falso positivo podría resultar en un tratamiento invasivo o costoso, una alta especificidad puede ser más deseable.

Estas implicaciones es importante tenerlas en cuanta, en el contexto de una herramienta de cribado para el cáncer de piel, ya que, puestos a optimizar sensibilidad o especificidad, las consecuencias de una menor sensibilidad son mucho más graves ya que podría implicar que el paciente no reciba tratamiento temprano. Al contrario, si la prueba fuera positiva en malignidad iría acompañada de estudios más específicos del tejido con una especificidad mucho mayor que ayudaría a solventar el error de la

herramienta de cribado. Aunque esta opción no es óptima, el coste económico es preferible al coste que puede tener sobre la vida de un paciente la detección tardía del melanoma.

De esto se deriva, que la elección de métricas de rendimiento depende del objetivo del cribado. En el trabajo de Santos-Neto et al. (de Alencar Neto & Santos-Neto, 2024) se aborda la limitación de utilizar sensibilidad y especificidad como métricas aisladas para evaluar pruebas diagnósticas en la práctica clínica. Aunque estas métricas son fundamentales en la educación médica, su aplicación exclusiva puede ser engañosa debido a su naturaleza retrospectiva. Los autores argumentan que los cocientes de verosimilitud (LR+, LR-) son más adecuados para la práctica clínica prospectiva porque permiten a los médicos actualizar las probabilidades diagnósticas basadas en nueva evidencia. Los LRs se derivan de la sensibilidad y especificidad, pero proporcionan una perspectiva diferente al indicar cuánto aumenta o disminuye la probabilidad de una enfermedad con un resultado positivo o negativo de la prueba.

Dado que le umbral de decisión es un parámetro modificable en este proyecto se ha explorado la reducción del umbral de decisión. Hasta el momento, los resultados mostrados han sido con un umbral del 50%, es decir, si la probabilidad de malignidad era igual o superior a 0,5, se clasificaba como maligno, mientras que si era menor se clasificaba como benigno. En este caso se ha empleado un umbral de 0,25, es decir, que si la probabilidad de malignidad es igual o superior a 0,25 se ha clasificado como maligno.

La matriz de confusión de la Figura 54 muestra que, de los casos positivos verdaderos, el modelo seleccionado (multimodal) correctamente clasificó 949 como malignos (verdaderos positivos), mientras que clasificó incorrectamente 72 como benignos (falsos negativos). Por otro lado, de los casos negativos verdaderos, 963 fueron correctamente identificados como benignos (verdaderos negativos) y 237 fueron erróneamente clasificados como malignos (falsos positivos). Esto se corresponde con una exactitud del 86%, una sensibilidad del 93% y una puntuación F1 de 86%.

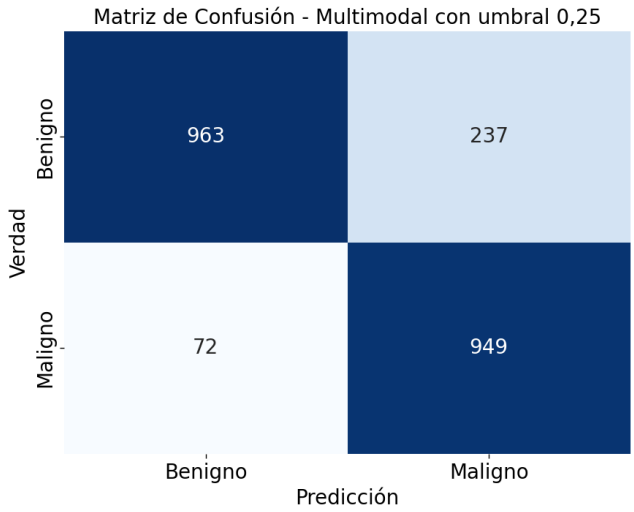


Figura 54. Matriz de confusión del modelo de stacking mediante el modelo Multimodal con un umbral de 0,25.

El umbral del 25% se ha elegido para equilibrar tanto la sensibilidad como la especificidad del modelo. Modificar el umbral puede alterar este equilibrio, aumentando la sensibilidad a expensas de la especificidad o viceversa. Este *trade-off* es una consideración fundamental en el diseño y evaluación de modelos diagnósticos, como se destaca en estudios sobre la evaluación de pruebas diagnósticas (Rezk et al., 2022). El umbral del 25% se ha elegido específicamente para maximizar la sensibilidad del modelo sin comprometer excesivamente su especificidad, teniendo en cuenta las características del

conjunto de datos y el objetivo clínico del cribado. Este valor se ha determinado como un punto de equilibrio que permite detectar un mayor número de casos malignos (reduciendo los falsos negativos) mientras se mantiene un número aceptable de falsos positivos.

La elección de un umbral del 25% está apoyada en la necesidad de asegurar que cualquier caso con al menos una baja probabilidad de malignidad sea tratado como maligno, lo que es crucial en contextos donde el costo de pasar por alto una lesión maligna es mucho mayor que el costo de realizar pruebas adicionales en lesiones benignas. Este umbral permite un enfoque más conservador, lo cual es preferible en el cribado de cáncer de piel, ya que prioriza la detección temprana. Además, estudios previos sugieren que reducir el umbral a niveles inferiores al 50% mejora la sensibilidad del modelo en contextos clínicos donde la seguridad del paciente es prioritaria (Rezk et al., 2022), lo que justifica la selección de este valor específico en nuestro caso. No se han utilizado técnicas de optimización de umbral más sofisticadas debido a la naturaleza clínica del problema y la necesidad de una implementación práctica en entornos reales.

Finalmente, y en el contexto de la aplicación clínica, se han empleado los datos no utilizados previamente, "descartados" durante el proceso de equilibrado inicial del conjunto de datos, todos de la categoría benigna, para evaluar el mejor modelo sencillo con un umbral de decisión de 0,25. Los resultados se muestran en la Figura 55. Como se puede ver, la red ha clasificado como malignos más de 5.000 casos con una exactitud y una especificidad, que por las características del dataset son la misma métrica, de un 89,22%.

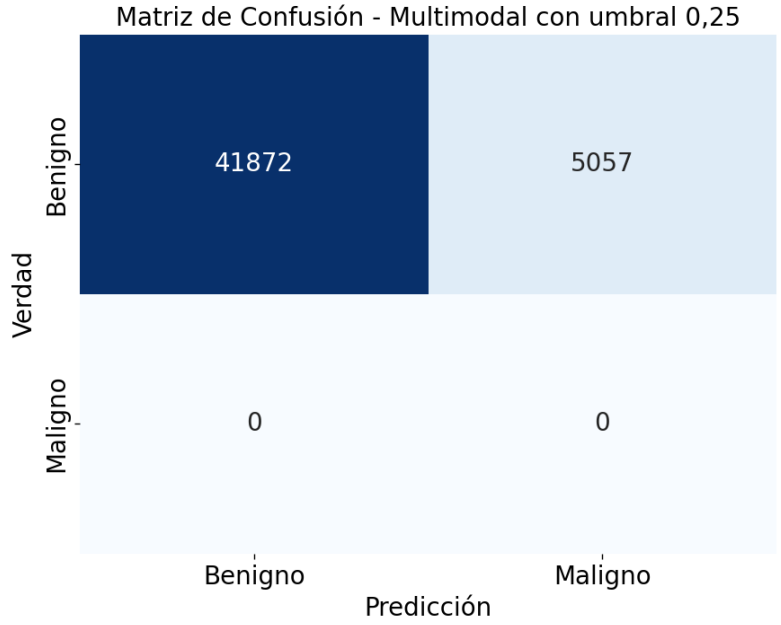


Figura 55. Matriz de confusión de la prueba de datos benignos con el modelo multimodal y un umbral de 0,25.

8. Discusión

8.1. Análisis crítico de los resultados

En el marco del presente Trabajo de Fin de Máster, se han llevado a cabo numerosos experimentos destinados a la creación y evaluación de modelos de *deep learning* (DL) y *machine learning* (ML) para la clasificación de lesiones cutáneas en benignas o malignas. Estos experimentos han generado una vasta cantidad de resultados que requieren un análisis meticuloso y una interpretación cuidadosa para

determinar la mejor vía de acción. La correcta evaluación de estos resultados es esencial para optimizar el rendimiento de los modelos y mejorar su aplicabilidad en entornos clínicos reales.

Uno de los problemas actuales en la inteligencia artificial y el DL es la necesidad de un gran número de datos etiquetados para entrenar modelos efectivos. Sin embargo, no solo es crucial la cantidad de datos, sino también su calidad y relevancia para el caso específico en cuestión. Nuestro estudio pone de manifiesto que la sensibilidad del modelo, es decir, su capacidad para identificar correctamente las lesiones malignas se ve significativamente afectada por la distribución y el equilibrio de los datos.

Los resultados de nuestros experimentos indican que los modelos funcionan mejor con conjuntos de datos que, aunque más pequeños, están equilibrados en cuanto al número de imágenes de lesiones benignas y malignas. En particular, se ha observado que un dataset con proporciones aproximadamente iguales de datos malignos y benignos presenta una mejor sensibilidad en comparación con un dataset más grande donde menos del 10% de las imágenes corresponden a lesiones malignas, a pesar de que este último contiene cuatro veces más imágenes. Esto se ve apoyado en la literatura en estudios como Mooijman et al. (2023) (Mooijman et al., 2023) donde se analiza cómo los datasets desequilibrados afectan negativamente el rendimiento de los algoritmos de aprendizaje automático.

El mejor rendimiento del dataset equilibrado en términos de sensibilidad puede atribuirse a varios factores. Por un lado, un conjunto de datos equilibrado proporciona una representación más completa y precisa de ambos tipos de lesiones, lo que permite al modelo aprender características distintivas de manera más efectiva. Además, conjuntos de datos altamente desbalanceados pueden conducir a modelos que están sesgados hacia la clase mayoritaria (en este caso, lesiones benignas), lo que disminuye su capacidad para detectar correctamente las lesiones malignas.

Por otro lado, un dataset equilibrado facilita un entrenamiento más eficaz del modelo, ya que cada clase está adecuadamente representada durante el proceso de aprendizaje, lo que mejora la generalización del modelo a nuevos datos. Además la reducción notable del número de datos para entrenamiento ha permitido un entrenamiento más rápido en términos temporales.

La primera línea de experimentos se centró en la clasificación de lesiones cutáneas a partir de imágenes. Para ello, se implementaron dos enfoques distintos. En primer lugar, se empleó un modelo de CNN sencillo entrenado mediante transferencia de aprendizaje utilizando la arquitectura VGG16, sin mecanismos de atención. Este modelo se ha empleado como una base para comparar con otros más avanzados. En segundo lugar, se han empleado modelos que incorporan bloques de atención, específicamente la técnica conocida como "*squeeze and excitation*". Dentro de este enfoque, probamos dos arquitecturas: SE-ResNet y EfficientNet, evaluando las variantes de EfficientNet desde B1 hasta B4. Utilizar SE-ResNet permite observar cómo la recalibración de características a nivel de canal afecta la capacidad del modelo para centrarse en las regiones más relevantes de las imágenes. Por otro lado, EfficientNet no solo incorpora bloques SE, sino que también optimiza su rendimiento mediante un escalado compuesto de ancho, profundidad y resolución, ofreciendo una comparación directa entre un enfoque de atención basado en canales y una arquitectura que combina atención con eficiencia computacional.

Comparar estas dos arquitecturas con mecanismos SE aporta una visión más amplia sobre cómo distintos enfoques de atención y eficiencia afectan el rendimiento del modelo. Además, al evaluar diferentes variantes de EfficientNet (B1 a B4), podemos analizar cómo se comportan estas redes con distintas configuraciones de complejidad, ayudando a determinar qué versión ofrece el mejor balance entre rendimiento y recursos computacionales. Esta comparación es fundamental para entender cuál de estas

arquitecturas proporciona un mejor rendimiento y cómo pueden combinarse eficazmente en un modelo de ensamblado para optimizar la clasificación de lesiones cutáneas.

Las redes neuronales, dependiendo de su arquitectura y complejidad, tienden a enfocarse en distintos aspectos de las imágenes. Los modelos más sencillos sin atención, como VGG16, pueden centrarse en características generales y menos específicas, lo que puede limitar su capacidad para distinguir entre lesiones benignas y malignas con alta precisión. Por el contrario, los modelos que emplean bloques de atención, como SE-ResNet y EfficientNet, tienen la capacidad de recalibrar características específicas y ajustar el enfoque a los detalles más relevantes de las imágenes. Esto les permite identificar patrones sutiles y diferenciadores que pueden ser cruciales para una clasificación precisa.

Los resultados obtenidos mostraron una mejora significativa en el desempeño de los modelos con atención en comparación con el modelo sencillo de VGG16. Esta mejora puede atribuirse a varias razones. En primer lugar, los bloques de atención permiten que la red preste más atención a las características importantes mientras suprime las menos relevantes, optimizando así el proceso de aprendizaje. Además, estas arquitecturas más avanzadas están diseñadas para capturar interdependencias entre canales de características, lo que mejora la capacidad del modelo para discernir entre las diferentes clases de lesiones. El uso de SE-ResNet y EfficientNet no solo incrementa la precisión del modelo, sino que también mejora su sensibilidad, un aspecto crítico en la detección de lesiones malignas donde los falsos negativos pueden tener consecuencias graves.

Por último, la experimentación con variantes de EfficientNet, desde B1 hasta B4, nos permitió identificar las configuraciones más efectivas para el ensamblado de modelos. Cada variante de EfficientNet aporta una combinación única de eficiencia y rendimiento, ajustándose de manera diferente a los datos y proporcionando perspectivas diversas que, cuando se combinan, pueden ofrecer una solución más robusta y precisa para la clasificación de lesiones cutáneas.

Los mapas de activaciones de clase, representados en forma de mapa de calor como los que se muestran en la Figura 56, proporcionan una visualización clara de las áreas en las que los modelos se enfocan para tomar sus decisiones de clasificación. Observando las imágenes aleatorias presentadas, podemos comparar el comportamiento de la VGG16 con el de EfficientNet-B1, es decir, modelo de imagen sin y con mecanismo de atención, respectivamente.

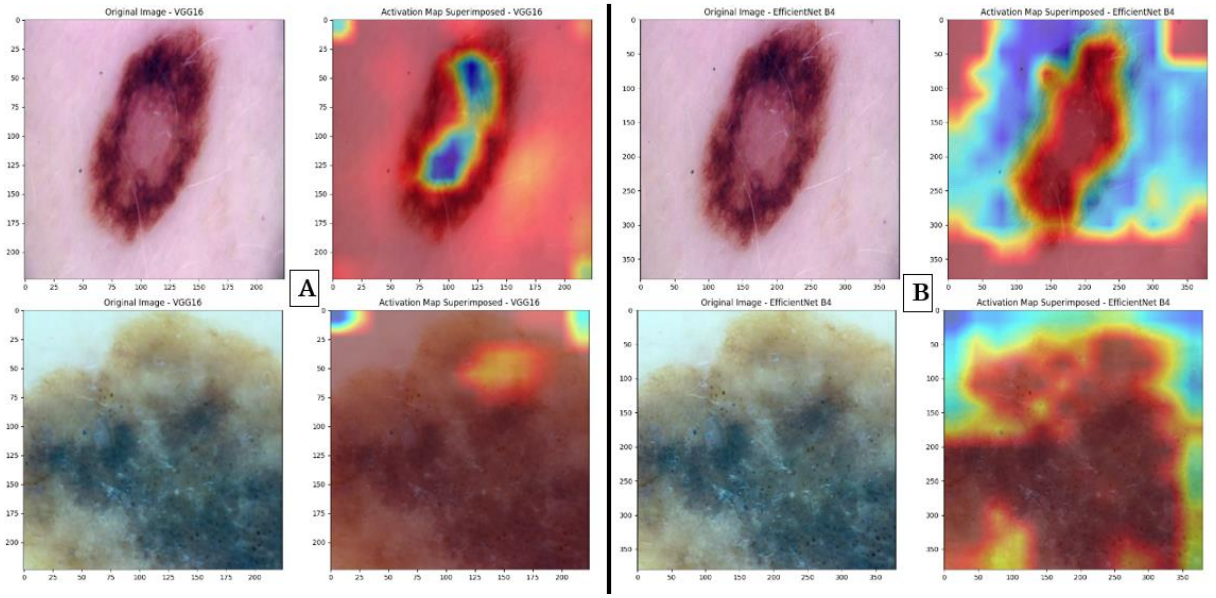


Figura 56. Mapas de activación de clase para el modelo sin mecanismos de atención VGG16 (A) y con mecanismos de atención, EfficientNetB4 (B)

La Figura 56 muestra los mapas de activación con escala de color de tipo JET, esto significa que los colores fríos, es decir zonas azules o verdes, representan zonas de baja activación de neuronas, es decir, el modelo no encuentra estas áreas relevantes para la clase maligna y son regiones donde las neuronas de la capa seleccionada no están activas o están menos activas. Por otro lado, los colores cálidos como amarillos naranjas o rojos representan zonas de alta activación. Es decir, son las regiones donde las neuronas de la capa seleccionada están más activas y, por tanto, están influyendo más en la decisión del modelo.

Si comparamos para una misma imagen, por ejemplo el caso de la imagen superior, las activaciones para los diferentes modelos, podemos ver que el modelo sin mecanismos de atención, ha considerado relevantes las zonas de tejido circundante y bordes de la lesión, ignorando el centro de esta. En contraste, el modelo con mecanismos de atención se está fijando en el interior de la lesión y los bordes, ignorando el tejido circundante a esta.

El caso de la lesión inferior de la Figura 56 es un caso interesante, ya que parece que el centro de la lesión está fuera de encuadre y esta ocupa la mayoría de la imagen. En el lado A del modelo sin atención podemos ver que ha considerado prácticamente toda la imagen de igual importancia (excepto algunas esquinas), por otro lado, el modelo con atención ha se ha centrado en las zonas más pigmentadas de la lesión, ignorando las zonas que no pertenecen a esta.

La comparación visual entre los mapas de activación de VGG16 (modelo de imagen sin mecanismos de atención) y EfficientNet-B1(modelo de imagen con mecanismos de atención) resalta la ventaja de utilizar modelos con bloques de atención. Mientras que VGG16 distribuye su atención de manera más amplia y menos específica, EfficientNet-B1 se concentra en las regiones más informativas de las imágenes, lo que conduce a una mejora significativa en los resultados de clasificación. Esta capacidad de los modelos avanzados para enfocarse en detalles críticos es fundamental para tareas como la detección de lesiones malignas, donde cada detalle puede ser crucial para un diagnóstico preciso.

Los mapas de activación también contribuyen significativamente a la explicabilidad del modelo. Al proporcionar una visualización clara de las áreas específicas en las que el modelo se enfoca para tomar

sus decisiones, estos mapas ayudan a convertir los modelos de "caja negra" en modelos de "caja gris". Esta mayor transparencia en el proceso de toma de decisiones ofrece una ventaja crucial en entornos clínicos, donde la confianza en las herramientas de diagnóstico es fundamental. Al poder visualizar y entender mejor cómo el modelo llega a sus conclusiones, los clínicos pueden sentirse más seguros y respaldados en su uso de estas tecnologías avanzadas para el diagnóstico de lesiones cutáneas.

En la segunda línea de experimentos, nos centramos exclusivamente en el uso de datos clínicos para la clasificación de la malignidad de las lesiones cutáneas. Los datos utilizados incluyen variables como el sexo, la edad y la ubicación de la lesión. Sin embargo, esta aproximación presenta limitaciones intrínsecas. La deducción de la malignidad basada únicamente en estos factores es inherentemente restrictiva, ya que no proporcionan suficiente información para un diagnóstico preciso. Por lo tanto, el rendimiento de estos modelos está seriamente limitado.

A pesar de estas limitaciones, los modelos basados en datos clínicos tienen la ventaja de ser rápidos y eficientes. Además, su precisión y utilidad podrían mejorar significativamente con la inclusión de datos clínicos adicionales. Por ejemplo, factores como antecedentes familiares de cáncer de piel, presencia de dolor, cambios recientes en la lesión, entre otros, aportarían información crucial que podría ayudar a los modelos a realizar predicciones más precisas.

Es importante destacar que, en esta instancia, los datos clínicos deben considerarse como complementarios. Aunque por sí solos no son suficientes para aplicaciones en entornos clínicos debido a su limitación informativa, su combinación con análisis de imágenes y otros datos más detallados puede mejorar notablemente el rendimiento general de los modelos de clasificación de lesiones cutáneas. De este modo, los datos clínicos aportan un contexto adicional que, cuando se integra adecuadamente con otros métodos de análisis, contribuye a un diagnóstico más robusto y fiable.

En la tercera línea de experimentos, desarrollamos un modelo "multimodal" que integra tanto las imágenes de las lesiones cutáneas como los datos clínicos de los pacientes. Este enfoque ha demostrado ser el más eficaz, proporcionando los mejores resultados en términos de precisión y sensibilidad. Los motivos detrás de este éxito radican en la capacidad del modelo para combinar la información visual detallada con el contexto clínico del paciente, emulando así el proceso de decisión humana. Mientras que la imagen ofrece una representación visual directa de la lesión, los datos clínicos proporcionan información adicional crucial, como la edad, el sexo, y la localización de la lesión, que en conjunto permiten una evaluación más completa y precisa.

Este modelo multimodal tiene la ventaja adicional de ser más eficiente en términos de cálculo en comparación con el ensamblado de modelos. Al no requerir que la información pase a través de múltiples modelos para ensamblar las probabilidades y predicciones, el proceso es más rápido y directo. Esta eficiencia mejora la aplicabilidad del modelo en entornos clínicos donde el tiempo es un factor crítico.

El enfoque multimodal se asemeja más al proceso de decisión de un clínico, quien no solo observa la lesión visualmente, sino que también considera el historial y el contexto del paciente para tomar una decisión informada. Esta capacidad de integración de diferentes tipos de datos no solo mejora la precisión del diagnóstico, sino que también aumenta la confianza en las predicciones del modelo.

Finalmente, en la última línea de experimentos, exploramos el ensamblado de modelos mediante técnicas como el "*majority voting*" y el "*weighted majority voting"*. Sin embargo, estos enfoques no aportaron beneficios significativos en comparación con el modelo multimodal, que hasta ese punto había demostrado ser el más efectivo. Esta falta de mejora podría atribuirse al uso de modelos basados en datos clínicos que, por sus métricas inferiores, no lograron complementar eficazmente los otros modelos.

En futuras investigaciones, sería interesante explorar diferentes combinaciones de modelos para determinar si pueden mejorar los resultados.

Las técnicas de "*stacking*" lograron modestas mejoras en los resultados, pero incrementaron considerablemente la complejidad del modelo. Dependiendo de varios modelos complejos, estas técnicas requieren tiempo y recursos computacionales significativos, lo cual puede ser una limitación en contextos clínicos donde estos recursos no siempre están disponibles. No obstante, esta limitación es actualmente menor, ya que aumentar la sensibilidad es crucial para la aplicación de estas herramientas como modelos de cribado masivo.

Finalmente, es importante destacar que la capacidad de personalizar el umbral de decisión del modelo puede ser una herramienta valiosa para los clínicos. Ajustar este umbral permite a los profesionales de la salud equilibrar la sensibilidad y la especificidad del modelo según las necesidades específicas del contexto clínico, optimizando así la utilidad práctica del modelo en diferentes escenarios de cribado y diagnóstico.

8.2. Revisión de las limitaciones

En este proyecto se presentan varias limitaciones que es importante considerar. Una de las principales limitaciones es la disponibilidad limitada de datos clínicos. En un contexto de atención primaria, es posible que se necesiten incorporar otras características clínicas relevantes, como antecedentes familiares, peso y etnia, para mejorar la precisión y utilidad del modelo. Sin embargo, la falta de estos datos en el conjunto utilizado puede restringir la capacidad del modelo para generalizar a una población más amplia y diversa.

Otra limitación significativa es la poca variedad racial en el dataset. La mayoría de los datos provienen de individuos de origen racial homogéneo, lo que puede sesgar los resultados y reducir la efectividad del modelo en poblaciones racialmente diversas. Esto se ha discutido en diversos estudios que subrayan la importancia de incluir datos de múltiples grupos étnicos para desarrollar modelos robustos y aplicables globalmente (Rezk et al., 2022).

El modelo utilizado en este proyecto también enfrenta desafíos inherentes a los modelos de aprendizaje profundo, comúnmente referidos como "modelos de caja negra". Esto significa que, aunque el modelo puede ofrecer predicciones precisas, la interpretación de cómo se toman esas decisiones puede ser difícil. Incluso con la incorporación de mapas de activación que proporcionan cierta transparencia, el modelo sigue siendo en gran medida opaco para los usuarios finales, lo que puede limitar la confianza y la aceptación en contextos clínicos.

Además, el uso de unidades de procesamiento gráfico (GPU) es esencial para entrenar modelos de aprendizaje profundo con eficiencia. No obstante, esto introduce restricciones de memoria y recursos computacionales que pueden no estar disponibles en todos los entornos. Las limitaciones de hardware pueden afectar la capacidad para entrenar modelos más complejos o para manejar conjuntos de datos muy grandes, lo que a su vez puede limitar el rendimiento y la escalabilidad del sistema propuesto.

Otra consideración es la necesidad de datos etiquetados de alta calidad para el entrenamiento de los modelos. La anotación de datos médicos es un proceso costoso y propenso a errores, y cualquier inconsistencia en las etiquetas puede afectar negativamente la precisión del modelo. Asimismo, la variabilidad en la calidad de las imágenes dermoscópicas debido a diferentes dispositivos y condiciones de captura puede introducir ruido en los datos, complicando aún más el proceso de entrenamiento y validación del modelo.

Finalmente, la implementación práctica de este modelo en un entorno clínico también puede enfrentar desafíos relacionados con la integración de sistemas y la aceptación por parte de los profesionales de la salud. La integración de un modelo de aprendizaje profundo en los flujos de trabajo existentes requiere un diseño cuidadoso y la colaboración entre ingenieros, médicos y administradores para garantizar que la herramienta sea efectiva y fácil de usar.

Estas limitaciones subrayan la necesidad de una investigación continua y mejoras en el diseño y la implementación de modelos de aprendizaje profundo para la clasificación de lesiones cutáneas, con el fin de superar estos obstáculos y maximizar el impacto clínico positivo.

9. Conclusiones

En este trabajo se ha desarrollado un sistema basado en inteligencia artificial para la clasificación de lesiones cutáneas, combinando imágenes dermoscópicas y datos clínicos con el fin de mejorar la precisión diagnóstica y facilitar la detección temprana de malignidad en entornos clínicos.

Para alcanzar este objetivo, se implementaron diferentes modelos de aprendizaje profundo y automático. En primer lugar, se desarrollaron varias arquitecturas de redes neuronales convolucionales (CNN) comparando entre el uso de mecanismos de atención, utilizando técnicas de aprendizaje por transferencia y ajuste fino, con el propósito de analizar imágenes dermoscópicas y distinguir entre lesiones benignas y maligna.

Además, se exploró el uso de modelos de aprendizaje automático tradicionales y redes neuronales artificiales para la clasificación de datos clínicos tabulares. Posteriormente, se integraron estas dos modalidades de datos en un modelo multimodal, que combinó las características aprendidas de las imágenes y los datos clínicos.

Finalmente, en linea con los *Objetivos del trabajo* se comparó el uso de diferentes enfoques como métodos de ensamblado con el fin de alcanzar el modelo más adecuado para cumplir con los requerimientos definidos en Requerimientos funcionales y no funcionales. De esto podemos concluir lo siguiente:

- La investigación destaca que un dataset equilibrado, con una proporción similar de imágenes de lesiones malignas y benignas, mejora la sensibilidad de los modelos, aumentando su capacidad para identificar correctamente las lesiones malignas. Este hallazgo subraya la necesidad de prestar atención a la distribución de los datos para optimizar el rendimiento del modelo, especialmente en contextos clínicos donde la detección temprana de lesiones malignas es crucial.
- Los resultados del trabajo demuestran que el uso de modelos de redes neuronales convolucionales (CNN) avanzados con mecanismos de atención, como SE-ResNet y EfficientNet, mejora significativamente la precisión y sensibilidad en la clasificación de lesiones cutáneas en comparación con modelos más simples, sin mecanismos de atención como VGG16. Esto resalta la importancia de los bloques de atención que optimicen el enfoque en las características relevantes de las imágenes.
- Aunque los modelos que utilizan únicamente datos clínicos presentan limitaciones debido a la falta de información y la limitación intrínseca de estos (edad, sexo y lugar anatómico de la lesión), demostraron ser eficientes en términos de rapidez y requerimientos computacionales. Los resultados de este TFM muestran que estos modelos deben considerarse complementarios

en lugar de independientes, ya que, cuando se integran con datos de imágenes, pueden proporcionar un enfoque diagnóstico más robusto y fiable.

- El modelo multimodal, que integra tanto imágenes dermoscópicas como datos clínicos, ha mostrado ser la estrategia más efectiva, proporcionando los mejores resultados en términos de precisión y sensibilidad. Este enfoque se asemeja al proceso de decisión clínica humana, combinando la riqueza de la información visual con el contexto clínico del paciente.

Este trabajo aporta una herramienta de inteligencia artificial que mejora la capacidad de detección de lesiones cutáneas malignas, combinando múltiples enfoques y fuentes de datos para ofrecer un diagnóstico más preciso y temprano. La integración de modelos avanzados y técnicas multimodales proporciona una solución viable para su implementación en diversos contextos clínicos, facilitando el acceso a un diagnóstico más fiable y, en última instancia, mejorando los resultados en salud para los pacientes.

9.1. Implicaciones para la investigación futura

El modelo desarrollado en este trabajo tiene implicaciones clínicas que podrían aplicarse a la práctica de la dermatología y la atención primaria. Primero, la integración de imágenes dermoscópicas con datos clínicos proporciona un enfoque más holístico y preciso para la clasificación de lesiones cutáneas, mejorando la precisión diagnóstica. Esta mejora en la precisión puede reducir el número de biopsias innecesarias y permitir una detección más temprana de las lesiones malignas, lo cual es crucial para mejorar las tasas de supervivencia de los pacientes.

Además, el modelo tiene un gran potencial para el seguimiento automatizado de lesiones cutáneas. Al poder monitorizar cambios en las lesiones a lo largo del tiempo, los dermatólogos pueden detectar el crecimiento o la transformación de lesiones sospechosas de manera más eficiente y oportuna. Este seguimiento automatizado puede aliviar la carga de trabajo de los profesionales de la salud y proporcionar una vigilancia continua y precisa.

El telediagnóstico es otra área donde este modelo puede tener un impacto significativo. La capacidad de realizar diagnósticos remotos utilizando imágenes dermoscópicas y datos clínicos permite a los pacientes recibir evaluaciones y diagnósticos sin la necesidad de visitas presenciales. Esto es especialmente beneficioso para personas en áreas rurales o con acceso limitado a dermatólogos especializados, mejorando el acceso a la atención médica y la equidad en salud.

El modelo podría servir como una herramienta de soporte para la toma de decisiones clínicas. Al proporcionar a los dermatólogos una segunda opinión basada en datos precisos y modelos avanzados de inteligencia artificial, se puede aumentar la confianza en los diagnósticos y tratamientos recomendados. Esto puede resultar en mejores resultados para los pacientes y un mayor uso de prácticas basadas en evidencia en la dermatología.

Las futuras líneas de investigación para este proyecto incluyen varias áreas prometedoras. Primero, se propone trabajar en la segmentación de imágenes dermoscópicas para mejorar la precisión del modelo.

Además, se planea incluir otro tipo de variables clínicas para aumentar la dimensionalidad del modelo. Esto podría incluir factores como el historial médico, antecedentes familiares, características demográficas, y datos genómicos. La inclusión de datos genómicos puede ofrecer una perspectiva completamente nueva sobre la predisposición individual a ciertos tipos de cáncer de piel y mejorar la personalización del diagnóstico y tratamiento.

La colaboración con instituciones de investigación y clínicas para incluir estos datos genómicos es una de las prioridades. Estos datos pueden proporcionar información valiosa sobre las mutaciones genéticas y otros factores de riesgo, permitiendo un enfoque más integrado y preciso para la detección y tratamiento de las lesiones cutáneas malignas.

10. Bibliografía

AEDV: Academia Española de Dermatología y Venereología. (s. f.). Recuperado 13 de septiembre de 2024, de https://aedv.es/

Anatomy of the Skin . (s. f.). Stanford Medicine Children's Health. Recuperado 10 de septiembre de 2024, de https://www.stanfordchildrens.org/es/topic/default?id=anatomy-of-the-skin-85-P04436

Arefin, S. (2024). AI Revolutionizing Healthcare: Innovations, Challenges, and Ethical Considerations. *MZ Journal of Artificial Intelligence*, *1*(2), 1−17-1−17. https://mzjournal.com/index.php/MZJAI/article/view/193

Cancer burden statistics and trends across Europe | ECIS. (s. f.). Recuperado 28 de junio de 2024, de https://ecis.jrc.ec.europa.eu/

Cassidy, B., Kendrick, C., Brodzicki, A., Jaworek-Korjakowska, J., & Yap, M. H. (2022). Analysis of the ISIC image datasets: Usage, benchmarks and recommendations. *Medical Image Analysis*, *75*, 102305. https://doi.org/10.1016/J.MEDIA.2021.102305

Chirodea, M. C., Novac, O. C., Novac, C. M., Bizon, N., Oproescu, M., & Gordan, C. E. (2021). Comparison of Tensorflow and PyTorch in Convolutional Neural Network - Based Applications. *Proceedings of the 13th International Conference on Electronics, Computers and Artificial Intelligence, ECAI 2021*. https://doi.org/10.1109/ECAI52376.2021.9515098

de Alencar Neto, J. N., & Santos-Neto, L. (2024). The Post Hoc Pitfall: Rethinking Sensitivity and Specificity in Clinical Practice. *Journal of General Internal Medicine*, *39*(8), 1506-1510. https://doi.org/10.1007/S11606-024-08692-Z/FIGURES/3

Dermatoscopia - Wikipedia, la enciclopedia libre. (s. f.). Recuperado 28 de junio de 2024, de https://es.wikipedia.org/wiki/Dermatoscopia

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature 2017 542:7639*, *542*(7639), 115-118. https://doi.org/10.1038/nature21056

European Agency for Safety and Health at Work. (2008). *Occupational skin diseases and dermal exposure in the European Union (EU-25) : policy and practice overview*. Office for Official Publications of the European Communities.

Everything you need to know about VGG16. (2021, septiembre 23). https://medium.com/@mygreatlearning/everything-you-need-to-know-about-vgg16-7315defb5918.

Goldsmith, Lowell A. Katz, Stephen I. Glichrest, Barbara A. Paller, Amy S. Leffell, David J. Wolff, Klaus. (2014). *Fitzpatrick. Dermatología en medicina general*. 3100.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning* (1.ª ed.). The MIT Press.

Gregory, S. A. (1966). Design Science. *The Design Method*, 323-330. https://doi.org/10.1007/978-1-4899-6331-4_35

Guzmán-Bucio, S., & Vega-Memije, M. E. (2023). Inteligencia artificial en Dermatología. *Dermatología Revista Mexicana*, *67*(6), 905-910.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-December*, 770-778. https://doi.org/10.48550/arxiv.1512.03385

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2017). Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(8), 2011-2023. https://doi.org/10.1109/TPAMI.2019.2913372

Huang, Y., Shan, K., Yan, Y., & Li, W. (2024). Computer-aided diagnosis of Alzheimer's disease based on structural magnetic resonance imaging. *Chinese Medical Journal*. https://doi.org/10.1097/CM9.0000000000003180

ISIC- The International Skin Imaging Collaboration. (s. f.). Recuperado 29 de junio de 2024, de https://www.isic-archive.com/

Johnson, N. S., Vulimiri, P. S., To, A. C., Zhang, X., Brice, C. A., Kappes, B. B., & Stebner, A. P. (2020). Invited review: Machine learning for materials developments in metals additive manufacturing. *Additive Manufacturing*, *36*, 101641. https://doi.org/10.1016/J.ADDMA.2020.101641

Kingma, D. P., & Ba, J. L. (2014). Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. https://arxiv.org/abs/1412.6980v9

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2020). Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 318-327. https://doi.org/10.1109/TPAMI.2018.2858826

Madheswari, K., & Karthikeyan, N. (2024). Deep Learning Enhanced Skin Cancer Analysis: Advancements in Melanoma Detection with Medical Support. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(3), 3695-3708. https://ijisae.org/index.php/IJISAE/article/view/6046

Malvehy, J., Puig, S., Braun, R. P., Marghoob, A. A., & Kopf, A. W. (2006). Handbook of Dermoscopy. *Handbook of Dermoscopy*, 1-97. https://doi.org/10.1201/B14613/HANDBOOK-DERMOSCOPY-JOSEP-MALVEHY-RALPH-BRAUN-SUSANA-PUIG-ASHFAQ-MARGHOOB-ALFRED-KOPF

Mohammed, F. A., Tune, K. K., Assefa, B. G., Jett, M., & Muhie, S. (2024). Medical Image Classifications Using Convolutional Neural Networks: A Survey of Current Methods and Statistical Modeling of the Literature. *Machine Learning and Knowledge Extraction*, *6*(1), 699-735. https://doi.org/10.3390/MAKE6010033/S1

Mooijman, P., Catal, C., Tekinerdogan, B., Lommen, A., & Blokland, M. (2023). The effects of data balancing approaches: A case study. *Applied Soft Computing*, *132*, 109853. https://doi.org/10.1016/J.ASOC.2022.109853

Mukhamediev, R. I., Popova, Y., Kuchin, Y., Zaitseva, E., Kalimoldayev, A., Symagulov, A., Levashenko, V., Abdoldina, F., Gopejenko, V., Yakunin, K., Muhamedijeva, E., & Yelis, M. (2022). Review of Artificial Intelligence and Machine Learning Technologies: Classification, Restrictions, Opportunities and Challenges. *Mathematics*, *10*(15). https://doi.org/10.3390/MATH10152552

Palacios-Martínez, D., & Díaz-Alonso, R. A. (2017). Dermatoscopia para principiantes (i): características generales. *Medicina de Familia. SEMERGEN*, *43*(3), 216-221. https://doi.org/10.1016/J.SEMERG.2015.11.009

Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot andÉdouardand, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830. http://scikit-learn.sourceforge.net.

Prove, P.-L. (2017, octubre 17). *Squeeze-and-Excitation Networks.* Towards Data Science.

Razavi, S. (2021). Deep learning, explained: Fundamentals, explainability, and bridgeability to process-based modelling. *Environmental Modelling & Software*, *144*, 105159. https://doi.org/10.1016/J.ENVSOFT.2021.105159

Rezk, E., Eltorki, M., & El-Dakhakhni, W. (2022). Improving Skin Color Diversity in Cancer Detection: Deep Learning Approach. *JMIR Dermatol 2022;5(3):e39143 https://derma.jmir.org/2022/3/e39143*, *5*(3), e39143. https://doi.org/10.2196/39143

Sãenz, S., Conejo-Mir, J., & Cayuela, A. (2005). Epidemiología del melanoma en España. *Actas Dermo-Sifiliográficas*, *96*(7), 411-418. https://doi.org/10.1016/S0001-7310(05)73105-7

Saha, S. (2018, diciembre 15). *A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way* . Towards Data Science. https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

Sandbank, J., Bataillon, G., Nudelman, A., Krasnitsky, I., Mikulinsky, R., Bien, L., Thibault, L., Albrecht Shach, A., Sebag, G., Clark, D. P., Laifenfeld, D., Schnitt, S. J., Linhart, C., Vecsler, M., & Vincent-Salomon, A. (2022). Validation and real-world clinical application of an artificial intelligence algorithm for breast cancer detection in biopsies. *npj Breast Cancer 2022 8:1*, *8*(1), 1-11. https://doi.org/10.1038/s41523-022-00496-w

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. https://arxiv.org/abs/1409.1556v6

Smith, L. N. (2015). Cyclical Learning Rates for Training Neural Networks. *Proceedings - 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017*, 464-472. https://doi.org/10.1109/WACV.2017.58

Smith, L. N. (2018). *A disciplined approach to neural network hyper-parameters: Part 1 -- learning rate, batch size, momentum, and weight decay*. https://arxiv.org/abs/1803.09820v2

Srivastava, N., Hinton, G., Krizhevsky, A., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*, 1929-1958.

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *36th International Conference on Machine Learning, ICML 2019*, *2019-June*, 10691-10700. https://arxiv.org/abs/1905.11946v5

Toossi, M. T. B., Pourreza, H. R., Zare, H., Sigari, M. H., Layegh, P., & Azimi, A. (2013). An effective hair removal algorithm for dermoscopy images. *Skin Research and Technology*, *19*(3), 230-235. https://doi.org/10.1111/SRT.12015

What is Python? Executive Summary. (s. f.). Recuperado 17 de junio de 2022, de https://www.python.org/doc/essays/blurb/

Williams, N. M., Rojas, K. D., Reynolds, J. M., Kwon, D., Shum-Tien, J., & Jaimes, N. (2021). Assessment of Diagnostic Accuracy of Dermoscopic Structures and Patterns Used in Melanoma Detection: A Systematic Review and Meta-analysis. *JAMA Dermatology*, *157*(9), 1078-1088. https://doi.org/10.1001/JAMADERMATOL.2021.2845

Xu, J., Leng, L., & Kim, B. G. (2023). Gesture Recognition and Hand Tracking for Anti-Counterfeit Palmvein Recognition. *Applied Sciences (Switzerland)*, *13*(21). https://doi.org/10.3390/APP132111795

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Member, S., Xiong, H., & He, Q. (2020). *A Comprehensive Survey on Transfer Learning*.

II. PRESUPUESTO

Objetivo

El presente presupuesto tiene como objetivo determinar el valor económico del Trabajo Final de Máster realizado. Además, permite recalcular el coste del trabajo en caso de cambio de algún factor.

Presupuesto desglosado

Coste mano de obra

Según la revisión de 2023 de las recomendaciones en la elaboración de presupuestos en actividades de I+D+I del servicio de gestión de la I+D+I de la UPV, el coste por mano de obra participante en el proyecto se elaborará de acuerdo con la fórmula 12.

$$Coste(€) = Ch \times Dh$$

Fórmula 12

Siendo Ch = Coste horario (€/h) y Dh = Dedicación (h)

En la Tabla 15 y Tabla 16 extraídas de este mismo documento se indican las tarifas 2020 recomendadas para personal de plantilla UPV y para personal externo.

Tabla 15. Tarifas 2022 recomendadas para personal de plantilla UPV

CATEGORIA PLANTILLA UPV	CATEGORÍA EN LA ACTIVIDAD	Horas/año facturables[2]	Coste directo por hora[3]
Catedrático/a de Universidad	Responsable	1.650	58,54
Titular de Universidad	Experto	1.650	45,63
Prof. Contratado Doctor		1.650	42,76
Ayudante Doctor	Técnico	1.650	28,33
Ayudante		1.650	18,3
Catedrático/a de Escuela Universitaria		1.650	42,17
Titular de Escuela Universitaria		1.650	34,46
Profesor Colaborador		1.650	42,26

Tabla 16. Tarifas 2022 recomendadas para personal eventual

CATEGORIA	Retribución anual bruta		Coste anual con S.S. (32,1%) e indemnización (3,04%).		Coste horario (incluido S.S.)	
	Mín	Máx	Mín	Máx	Mín	Máx
Doctor contratado	17.341,52	40.956,86	23.435,33	55.349,10	13,32	31,45
Titulado superior	17.341,52	32.540,48	23.435,33	43.975,20	13,32	24,99
Titulado medio	14.994,84	26.032,44	19.866,72	35.180,24	11,51	19,99
Especialista técnico	C16: 20.396,60	C19: 21.846,58	C16: 27.563,97	C19: 29.523,47	C16: 15,66	C19: 16,77

A continuación, se desglosan las actividades de cada tipo de personal para el cálculo de las horas. Se ha dividido en tres tipos de personal, siendo el primero el Ingeniero Biomédico o Titulado Superior (Tabla 17), el segundo el doctor contratado (Tabla 18), seguido del catedrático o tutor (Tabla 19).

Tabla 17. Desglose de actividades del ingeniero biomédico.

Id	Concepto	Unidad	Nº de unidades
1	Recopilación de información	Horas	45
2	Planificación del proyecto	Horas	25
3	Recogida de datos	Horas	60
4	Diseño y implementación y optimización de modelos	Horas	200
5	Análisis de resultados	Horas	100
6	Redacción del proyecto	Horas	60
7	Revisión final	Horas	20
8	Presentación	Horas	16
TOTAL		Horas	526

En la Tabla 17 se presenta un desglose detallado de las horas dedicadas a diferentes actividades en un proyecto realizado por un ingeniero biomédico. En la fase de recopilación de información, se invirtieron 45 horas, reflejando el tiempo dedicado a investigar y reunir datos relevantes para el proyecto. La planificación del proyecto requirió 25 horas, asegurando una organización y estructuración adecuadas de las tareas a realizar.

La recogida de datos ocupó 60 horas, involucrando la obtención y preparación de los datos necesarios para el análisis. En el diseño e implementación y optimización de modelos, se dedicaron 200 horas, evidenciando el esfuerzo en desarrollar y perfeccionar los modelos de inteligencia artificial utilizados en el estudio. El análisis de resultados tomó 100 horas, durante las cuales se interpretaron los datos obtenidos y se extrajeron conclusiones significativas. La redacción del proyecto consumió 60 horas, tiempo necesario para documentar detalladamente el trabajo realizado y los hallazgos obtenidos. En la revisión final, se emplearon 20 horas para asegurar la calidad y coherencia del documento final. Finalmente, la presentación del proyecto requirió 16 horas, preparando y ensayando la exposición de los resultados. En total, el proyecto demandó 526 horas, distribuidas en diversas actividades clave para asegurar un desarrollo riguroso y completo del estudio.

Tabla 18. Actividades del doctor contratado.

Id	Concepto	Unidad	Nº de unidades
1	Supervisión y apoyo técnico	Horas	50
TOTAL		Horas	50

La Tabla 18 detalla las horas dedicadas por el doctor contratado muestra que se invirtieron un total de 50 horas en supervisión y apoyo técnico. Este tiempo refleja el esfuerzo necesario para proporcionar asistencia técnica y supervisión especializada durante las diversas etapas del proyecto. La experiencia y conocimientos técnicos del doctor han sido esenciales para guiar y apoyar el desarrollo del proyecto, asegurando que se sigan las mejores prácticas y se solucionen problemas técnicos a medida que surgen.

En cuanto a la tabla correspondiente al catedrático, se observa que se dedicaron 40 horas a supervisión y orientación. Este tiempo ha sido crucial para proporcionar una visión estratégica y orientación académica al proyecto. La supervisión del catedrático ha sido fundamental para garantizar que el proyecto siga un enfoque riguroso y que los objetivos académicos y científicos se cumplan de manera efectiva.

Tabla 19. Actividades del catedrático.

Id	Concepto	Unidad	Nº de unidades
1	Supervisión y orientación	Horas	40
TOTAL		Horas	40

La Tabla 20 detalla el coste total de los recursos humanos involucrados en el proyecto. El ingeniero biomédico, con un coste por hora de 24,99 euros y dedicando 526 horas, representa la mayor parte del presupuesto, sumando un total de 13.144,74 euros. El catedrático de universidad, con un coste por hora de 58,54 euros y dedicando 40 horas, contribuye con 2.341,60 euros al total. Por último, el doctor contratado, con un coste por hora de 31,45 euros y 50 horas de trabajo, añade 1.572,50 euros. En total, el proyecto implica 616 horas de trabajo, con un coste global de 17.058,84 euros. Esta distribución refleja el esfuerzo y la inversión necesarios para asegurar el éxito del proyecto.

Tabla 20. Coste de mano de obra.

Id	Concepto	Coste/hora (€/	Horas	Total (€)
1	Ingeniero biomédico junior	24,99	526	13.144,74 €
2	Catedrático de universidad	58,54	40	2.341,60 €
3	Doctor contratado	31,45	50	1.572,50 €
Total			616	17.058,84 €

Coste de materiales

En el proyecto se han identificado distintos materiales que se clasifican en dos categorías: costes directos y costes con amortización. La Tabla 21 contiene el desglose de los costes con amortización, que corresponden a aquellos materiales cuyo uso y beneficio se extienden más allá del año en curso y, por lo tanto, se distribuyen en el tiempo. Esta categoría incluye materiales que se registran como activos y se amortizan en función de su vida útil estimada, que en este caso se ha determinado en un año. En esta tabla, encontramos el ordenador de sobremesa, adquirido por un valor de 1,200 euros, un equipo esencial para realizar tareas de procesamiento y análisis de datos que se espera utilice en varios proyectos durante su vida útil. Otro material que se amortiza es el disco duro de 4TB, con un coste de 120 euros, utilizado para el almacenamiento de grandes volúmenes de datos que, dada su naturaleza, seguirá proporcionando capacidad de almacenamiento a lo largo del tiempo. Asimismo, se amortiza el costo de la licencia de software Microsoft Windows Enterprise, valorada en 300 euros, que es fundamental para el funcionamiento del equipo de computación, este software tiene una licencia que, aunque se utiliza desde el inicio del proyecto, mantendrá su utilidad operativa durante su periodo de vigencia, extendiéndose más allá del año actual.

Tabla 21. Costes de materiales con amortización

Nº	Material	Descripción	Costo (€)	Vida útil (años)	Amortización anual (€)
2.1	Ordenador de sobremesa	Equipo informático para la ejecución del proyecto	1,200.00 €	5	240.00 €
2.2	Disco Duro de 4TB	Almacenamiento adicional de datos	120.00 €	3	40.00 €
2.3	Microsoft Windows Enterprise	Licencia de software operativo	300.00 €	3	100.00 €
Total			**1,620.00 €**		**380.00 €**

En la Tabla 22 se presentan los costes de materiales directos, es decir, aquellos que corresponden a gastos que se imputan de manera inmediata al presupuesto del proyecto, ya que se consumen o utilizan completamente durante el periodo en curso y no tienen una vida útil prolongada más allá de este tiempo. Entre los materiales clasificados como costes directos se encuentra la suscripción a Kaggle Pro, que es una plataforma de ciencia de datos utilizada para acceder a herramientas y competencias necesarias durante la ejecución del proyecto, esta suscripción tiene un coste de 50 euros y es considerado un gasto directo porque su uso se limita al periodo actual del proyecto sin generar valor a largo plazo. También se incluyen los servicios de Google Cloud GPUs, con un coste de 292 euros, utilizados para la computación en la nube que permite acelerar el entrenamiento de modelos de aprendizaje profundo y su gasto se realiza íntegramente durante el año del proyecto, siendo esencial para operaciones puntuales y específicas sin extender su beneficio a futuros periodos.

Tabla 22. Costes directos de materiales

Nº	Concepto	Costo (€)
2.4	Kaggle Pro	50.00 €
2.5	Google Cloud GPUs	292.00 €
	Total	**342.00 €**

Coste total

La Tabla 23 presenta un desglose detallado del presupuesto total para la ejecución del proyecto. El coste principal corresponde a la mano de obra, con un total de 17.058,84 euros, seguido por el coste de materiales, que suma 722,00 euros. Estos dos componentes representan los gastos directos más significativos del proyecto.

Adicionalmente, se incluyen los gastos generales, calculados al 13% del coste de mano de obra y hardware/software, resultando en 2.311,51 euros. La suma de estos costos da un total de 20.092,35 euros.

Finalmente, se añade el 21% de IVA, que asciende a 4.219,39 euros, llevando el presupuesto de ejecución por contrata a un total de 24.311,74 euros. Este presupuesto refleja todos los costos necesarios para la realización completa del proyecto, asegurando que se cubren tanto los gastos directos como los indirectos y el beneficio empresarial.

Tabla 23. Cálculo del presupuesto del proyecto

COSTE	IMPORTE (euros)
Coste Mano de obra	17,058.84 €
Costes directos de materiales	342.00 €
Costes de materiales amortizados	380.00 €
13% gastos generales	2,311.51 €
Suma	20,092.35 €
21% IVA	4,219.39 €
Presupuesto ejecución pro contrata	24,311.74 €

AÑEJO 1: MÉTRICAS DURANTE ENTRENAMIENTO

1. Objetivo

El presente añejo recoge las principales métricas evaluadas durante las fases de entrenamiento y validación durante el entrenamiento, con el fin de ilustrar el proceso de entrenamiento de los diferentes modelos de aprendizaje profundo empleados en el TFM.

2. Modelos de imagen

2.1. Modelo de imagen sin mecanismos de atención: VGG16.

En la Figura 57, se observa la evolución de las principales métricas de validación durante la fase de entrenamiento y validación durante entrenamiento. Un fenómeno que se repite en todas las métricas es la divergencia entre las curvas de entrenamiento y validación, a partir de la época 4, se produce un crecimiento estable de las métricas de entrenamiento, mientras que las de validación tienen un crecimiento mucho menor o en algunos casos un estancamiento, como es el ejemplo del área bajo la curva de la curva ROC (AUC). También es notable, la inestabilidad en el conjunto de validación de la métrica de *Recall* o sensibilidad, que varía entre 0,6 y 0,9 durante la evolución de épocas. Esta inestabilidad puede estar debida a la dimensión del conjunto de validación, que es un 10% del conjunto de entrenamiento, por lo que pequeñas variaciones en predicciones, tienen un fuerte impacto sobre la métrica.

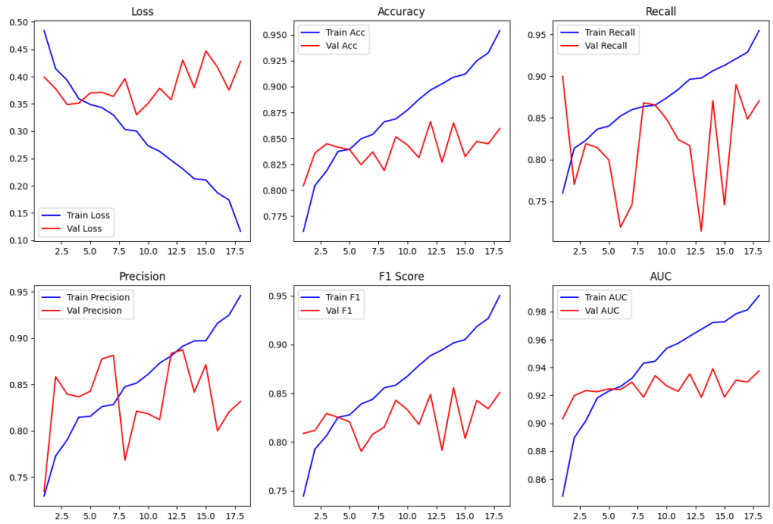


Figura 57. Evolución de las principales métricas de validación durante el entrenamiento en el conjunto de datos de entrenamiento (azul) y validación (rojo) del modelo VGG16.

2.2. Modelo de imagen con mecanismos de atención: SE-ResNet

En la Figura 58, aparece la evolución de las principales métricas de seguimiento durante el entrenamiento y validación. En estas podemos ver una evolución constante y de mejora en las métricas de entrenamiento y un crecimiento menor e incluso estancamiento de las métricas de validación

específicamente a partir de la octava época. Destaca visualmente la métrica de *recall* o sensibilidad, en la que se muestra una inestabilidad y una evolución menos marcada.

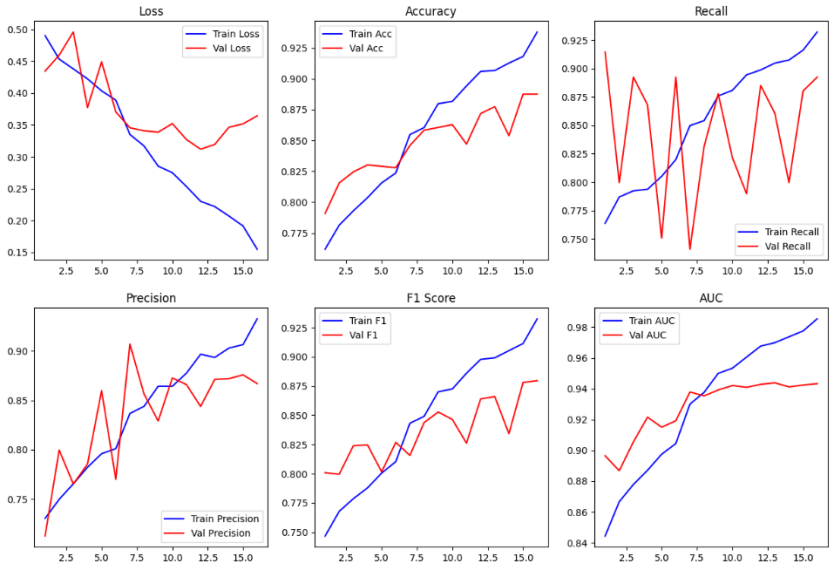


Figura 58. Evolución de las principales métricas de validación durante el entrenamiento en el conjunto de datos de entrenamiento (azul) y validación (rojo) del modelo SE-ResNet.

2.3. Modelo de imagen con mecanismos de atención: EfficientNet

2.3.1. EfficientNet B1

En la Figura 59 se muestran varias gráficas que ilustran las métricas durante el entrenamiento y la validación. En la primera fila, podemos ver la evolución de la pérdida, exactitud y recall tanto para el conjunto de entrenamiento como para el de validación. Se puede ver que la pérdida disminuye de manera constante en ambos conjuntos, mientras que la precisión y el *recall* aumentan a medida que avanza el entrenamiento.

Sin embargo, en la validación, estos incrementos no son tan pronunciados, lo cual podría indicar un leve sobreajuste del modelo. En la segunda fila, se muestra la evolución de la precisión, la puntuación F1 y el AUC, mostrando una tendencia similar, donde las métricas de entrenamiento mejoran más significativamente que las de validación, lo que sugiere que el modelo podría beneficiarse de técnicas adicionales para mejorar su capacidad de generalización.

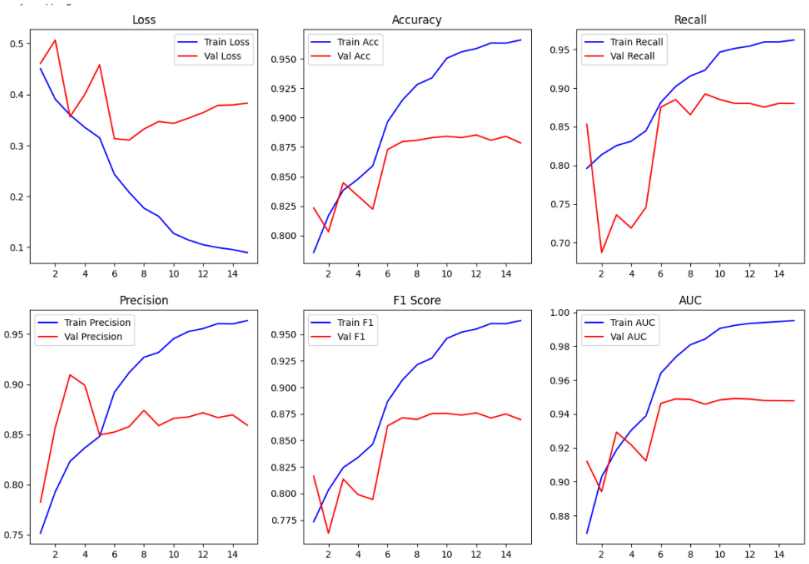


Figura 59. Evolución de las principales métricas de validación durante el entrenamiento en el conjunto de datos de entrenamiento (azul) y validación (rojo) del modelo EfficientNet B1.

2.3.2. EfficientNet B2

En la Figura 60 podemos ver las métricas durante el entrenamiento de la red EfficientNet B2. La red se ha entrenado durante un total de 17 épocas, en esta última se ha ejecutado la parada temprana o *early stopping* y se ha detenido el entrenamiento. Como se puede observar la red partía de una base sólida de resultados con los coeficientes preentrenados ya que los valores de exactitud (*accuracy*) variaban entre el 70% y el 98% en todas las épocas. Se puede observar en todas las métricas un sobreajuste, que se aprecia en la diferencia entre las curvas de entrenamiento y validación. Este fenómeno se hace especialmente evidente desde épocas muy tempranas, por ejemplo, en la gráfica de área bajo la curva ROC, es especialmente evidente a partir de la cuarta época.

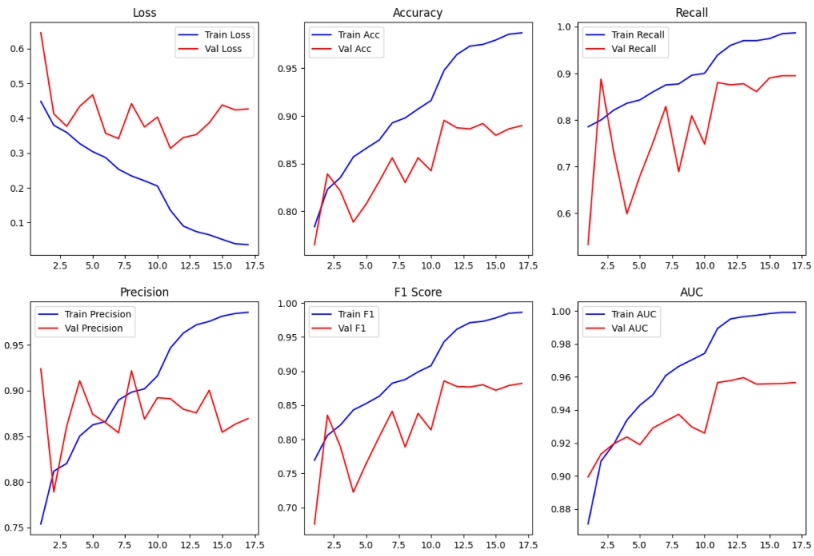


Figura 60. Evolución de las principales métricas de validación durante el entrenamiento en el conjunto de datos de entrenamiento (azul) y validación (rojo) del modelo EfficientNet B2.

2.3.1. EfficientNet B3

En la Figura 61 podemos ver la evolución de las métricas durante el entrenamiento de la CNN EfficientNet B3. Como en redes con estructura similar, se ha producido un sobreajuste, en este caso es especialmente notable en la métrica de pérdidas, aproximadamente a partir de la décima época, donde las curvas de entrenamiento y validación difieren notablemente. Es especialmente llamativa la inestabilidad en comparación a las otras métricas de la sensibilidad o *recall*, especialmente en el conjunto de validación. Esto puede estar debido a las dimensiones del conjunto de validación, ya que al tratarse de un 10% de los datos de entrenamiento, cambios en pocas predicciones cambia notablemente los resultados de las métricas. El entrenamiento a finalizado en la época 24 con un área bajo la curva en el entrenamiento de aproximadamente 1 con una exactitud de 0,975, y en el set de validación con un área de 0,96 aproximadamente con una exactitud notablemente menor con un valor aproximado o 0,875.

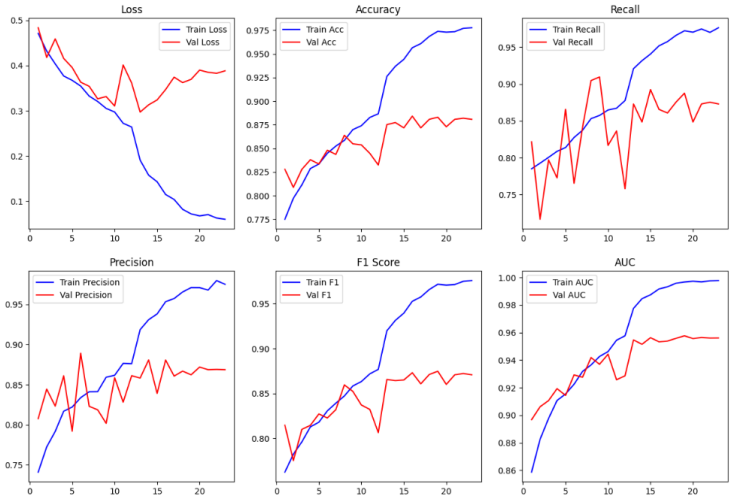


Figura 61. Evolución de las principales métricas de validación durante el entrenamiento en el conjunto de datos de entrenamiento (azul) y validación (rojo) del modelo EfficientNet B3.

2.3.1. EfficientNet B4

En la Figura 62, se presenta la evolución de las métricas de evaluación durante las fases de entrenamiento y validación durante el entrenamiento de la red EfficientNet B4. Como en las versiones anteriores de la familia EfficientNet, se produce un sobreajuste que es notable entorno a la época 7, ya que es donde divergen las curvas de entrenamiento y validación. Se presentan también valores elevados de todas las métricas durante el entrenamiento, siendo menores durante la validación.

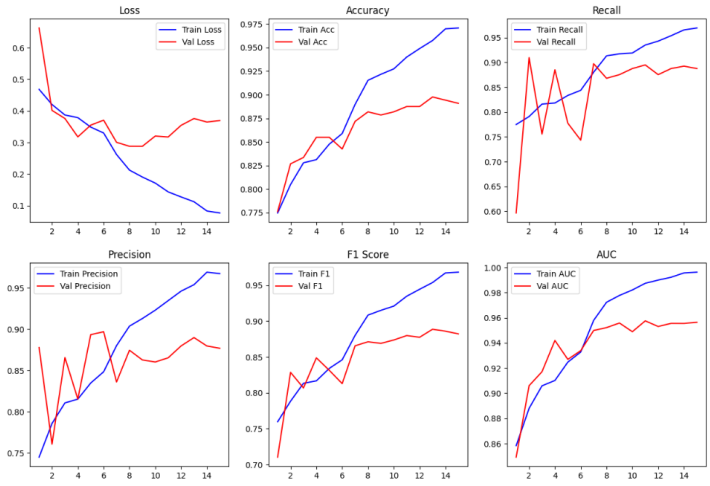


Figura 62. Evolución de las principales métricas de validación durante el entrenamiento en el conjunto de datos de entrenamiento (azul) y validación (rojo) del modelo EfficientNet B4.

3. Modelo de variables clínicas

La **Figura 63** muestra la evolución de las métricas evaluadas durante el entrenamiento y la validación de la red neuronal para variables clínicas. En este caso, se ha ejecutado solo durante 7 épocas, en la cual se ha iniciado la parada temprano debido a alcanzar el número de épocas sin mejora. En este caso, se produce un fenómeno por el cual las métricas, excepto las pérdidas, tienen en general mejores resultados que durante el entrenamiento. Esto puede deberse a varias razones. Por un lado, puede ser debido a el ruido en los datos de entrenamiento, es decir, el modelo podría estar ajustándose al ruido presente en los datos de entrenamiento, mientras que los datos de validación, al ser menos ruidosos, permiten al modelo generalizar mejor. Por otro lado, la distribución de datos de validación podría ser más representativa de la verdadera distribución de datos, permitiendo una mejor evaluación de la capacidad discriminativa del modelo.

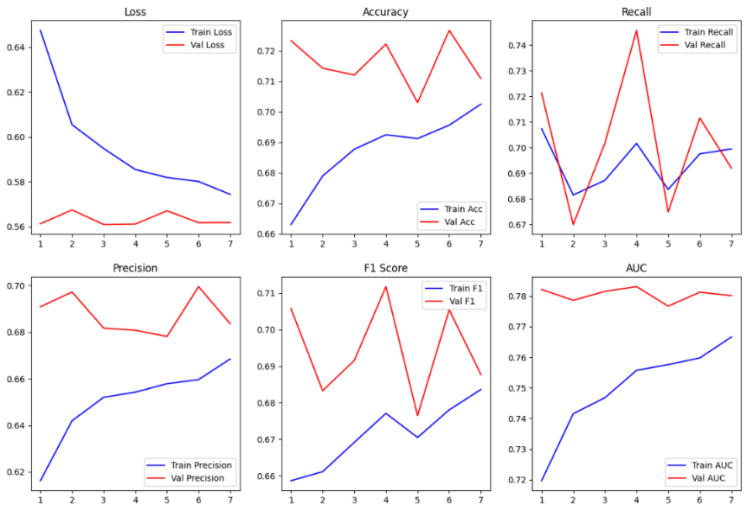


Figura 63. Evolución de las principales métricas de validación durante el entrenamiento en el conjunto de datos de entrenamiento (azul) y validación (rojo) del modelo TNN.

4. Modelo integrador o multimodal

En la Figura 64 se muestra la evolución de las métricas durante la fase de entrenamiento. Se ha entrenado durante un total de 18 épocas, en esta última se ha ejecutado la parada temprana al no haber mejoría en las métricas, siguiendo el criterio de mejora en el área bajo la curva (AUC). Podemos ver que efectivamente la curva de entrenamiento es similar en todas las métricas, sin embargo, en este caso se produce un crecimiento de la métrica de pérdida durante la validación, lo que puede significar una posible optimización de los hiperparámetros. Se puede apreciar que ya en épocas tempranas se ha producido el sobreaprendizaje produciendo una elevada diferencia entre las curvas de entrenamiento y validación.

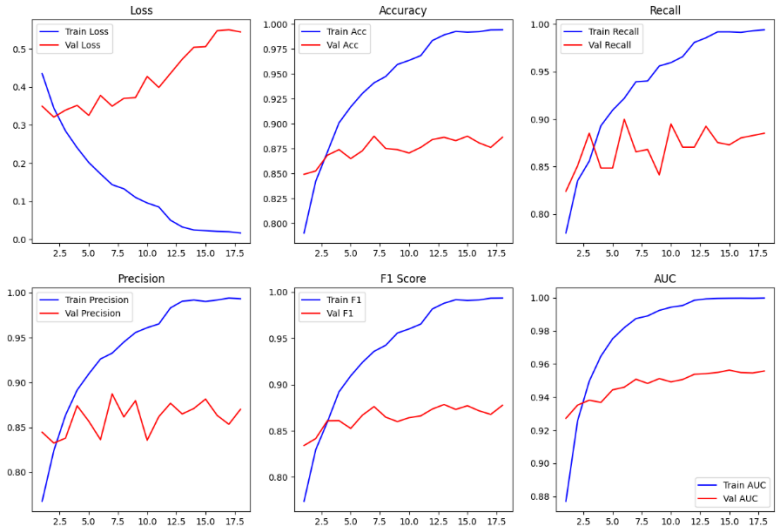


Figura 64. Evolución de las principales métricas de validación durante el entrenamiento en el conjunto de datos de entrenamiento (azul) y validación (rojo) de la red neuronal integradora o multimodal