Universidad de La Habana
Facultad de Matemática y Computación

Propuesta de modelo de clasificación con alta sensibilidad al Melanoma

Autor:

Daniel Abad Fundora

Tutora:

Dra. Marta Lourdes Baguer Díaz Romañach

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

Febrero del 2025

Repositorio de GitHub

A mi madre

Agradecimientos

Sería injusto de mi parte escribir una sección de agradecimientos sin mencionar, en primer lugar, a quien, de hecho, me enseñó a escribir, así como a leer, a patinar, a montar bicicleta, aritmética básica, casi todo lo que sé de física, y tantas cosas más que me sería imposible enumerarlas. A quien debo lo que sé de inglés, de natación, de judo, de ajedrez (aunque, siendo sincero, nunca fui realmente bueno en estas cosas). A quien siempre me ha impulsado a seguir adelante, a quien siempre me ha motivado a superarme, a quien me enseñó que el conocimiento nunca sobra, a quien me enseñó a creer en mí y en que puedo lograr lo que me proponga, a quien me ha apoyado en cada una de las decisiones de mi vida, a quien jamás me ha fallado, a quien más admiro, a mi mejor maestra, a mi mejor amiga, **a mi madre**.

A pesar de la distancia, mi familia ha sido una alentadora constante en mi vida, quisiera hacer una mención especial a mis tres abuelos, quienes siempre han estado ahí cuando lo he necesitado.

Empecé esta carrera en el 2021, en aquel momento, tenía dudas acerca de si había escogido la carrera apropiada, dudas que se desvanecieron a los pocos meses de llegar a esta facultad. Agradezco a los que me han acompañado en esta travesía, a mis inseparables compañeros: Anabel Benítez, Alex Sierra y Raudel Gómez, quienes han estado a mi lado en todo momento, en cada asignatura, con quienes he discutido incontables veces y a quienes siempre he admirado, aunque jamás les haya dicho. A Javier, por hacernos de tutor durante nuestro primer año y a Omar, que lamentablemente la vida separó nuestros caminos. A Enzo por los proyectos conjuntos y por su apoyo constante. A Amanda, Ana Melissa, Nahomi, Sherlyn y Juan Carlos por su amistad.

Siempre he soñado con hacer de este mundo un lugar mejor, hacer grandes aportes a esta humanidad, cuyo futuro es incierto, y preocupante. Ya sea erradicando epidemias, evitando disputas, apaciguando necesidades... sueños exagerados, y ridículos, tal vez, pero en esta facultad aprendí que los sueños no son para cumplirse, sino para caminar, nos hacen avanzar aunque no los logremos. Y aprovecho esta oportunidad para que le conste por escrito a todos los mencionados y a los que, injustamente, dejé de mencionar, que todo avance en mi camino, todo aporte que haga este nuevo Científico de la Computación a este mundo, siempre será, también, mérito vuestro.

Opinión de la tutora

El estudio de modelos de clasificación que se presentan en esta tesis constituye un resultado planificado en el proyecto PN223LH010-036 "Wavelets, frames, técnicas espectrales, ecuaciones en derivadas parciales y aprendizaje automático científico en el análisis de imágenes" que ejecuta desde enero 2024 hasta diciembre de 2026 asociado al Programa Nacional Ciencias Básicas y Naturales. Daniel encontró como antecedentes de su tesis el modelo propuesto por Claudia Olavarrieta en 2022 que clasifica en la versión 2.0 de DermaUH, un ensemble de tres redes neuronales en el que se experimentó con diferentes técnicas de fusión. Sin embargo, entre las principales preocupaciones se señalaron dificultades en la predicción del melanoma y de la clase otros. Consideramos entonces la clasificación basada en vectores de características. Jordan Pla presenta en 2023 en su tesis modelos para clasificar en cuatro y ocho clases, estudia un gran conjunto de características y experimenta ampliamente utilizando diferentes métodos para la clasificación. Los resultados superaron por poco los obtenidos por el emsemble propuesto por Claudia. Por supuesto, hay que considerar aspectos concernientes a la preparación de los data sets en ambas tesis. Le encargamos a Daniel entonces la tarea de mejorar los resultados de la clasificación, hablamos en los encuentros de modelos híbridos encontrados en la literatura y en las aplicaciones en el mercado que o bien alertan de malignidad o clasifican en melanoma y no melanoma. Considerando que el melanoma es el cáncer más agresivo los resultados hasta el momento de comenzar su estudio debían ser definitivamente mejorados. Solo dimos algunas ideas. Daniel con su carácter inquieto, de búsqueda incesante realizó el estudio e implementación de varias combinaciones de métodos y fue conformando con todo lo experimentado los resultados que hoy se presentan. Toda la experimentación, la búsqueda constante de mejores variantes, la prueba en imágenes cubanas, descartada al final, le han permitido presentar modelos con una alta sensibilidad al melanoma. Con mucha más confianza, esa que se adquiere del trabajo serio, de la discusión con los especialistas y del estudio de la literatura disponible vemos hoy a Daniel explicar sus propuestas. No fueron pocos los obstáculos, por una parte, las imágenes cubanas, la duda de las etiquetas, preparar su conjunto de imágenes lo cual le llevó buena parte del tiempo y la mejora día a día de su documento escrito. Daniel trabajó con absoluta independencia y creatividad, participó

en las discusiones con especialistas, fue ponente en el recién organizado Taller del 21 de enero y sobre todo discutió mano a mano con el Dr. Rigoberto García, Jefe del Grupo Nacional de Dermatología mostrando su capacidad de análisis. Es muy grato para un profesor, transcurrido un periodo de trabajo ver cómo los estudiantes se han adueñado del problema, han crecido profesionalmente y han mejorado su capacidad de trabajo en equipo para resolver problemas reales y eso me ha sucedido con Daniel. Ahora culmina con este ejercicio académico una etapa de formación, pero sigue otra de desarrollo profesional en la que con absoluta certeza continuará creciendo y se desenvolverá satisfactoriamente. Fue un placer haber compartido este tiempo y me alegro de poder seguir contando con él en el equipo de DermaUH.

Dra. C. Marta Lourdes Baguer Díaz Romañach

Resumen

La incidencia de cáncer de piel ha aumentado considerablemente en los últimos años, y el melanoma, por su alta tasa de mortalidad, representa uno de los mayores desafíos en dermatología. La detección temprana de esta lesión es crucial, ya que mejora significativamente el pronóstico del paciente. Por lo tanto, lograr una elevada sensibilidad en la identificación del melanoma reduciría la cantidad de falsos negativos, lo que podría salvar numerosas vidas.

En este estudio, se aborda la clasificación de imágenes de lesiones cutáneas con un enfoque que prioriza una alta sensibilidad en la detección del melanoma, minimizando la posibilidad de que estas lesiones pasen desapercibidas. Para ello, se implementó un esquema de clasificación en dos fases: una primera etapa de clasificación binaria para diferenciar entre lesiones melanocíticas y no melanocíticas, seguida de una clasificación en cuatro categorías en caso de ser clasificado como "no melanoma".

Se emplearon modelos basados en la arquitectura *Transformer*, reconocidos por su capacidad de extraer representaciones globales de las imágenes, lo que permite capturar patrones complejos relevantes para la detección del melanoma. Los resultados obtenidos muestran que este enfoque logró una sensibilidad elevada para la clase melanoma, lo que refuerza su potencial como herramienta de apoyo en el diagnóstico temprano. No obstante, esta mejora se alcanzó a expensas de una disminución en otras métricas de clasificación.

También, se presentan y comparan los resultados obtenidos utilizando distintos modelos de extracción automática de características, contrastándolos con aquellos basados en criterios médicos predefinidos. Esta comparación permite evaluar la eficacia de ambos enfoques y su aplicabilidad en la mejora de los sistemas de diagnóstico asistido por computadora en dermatología.

Los hallazgos de este estudio evidencian el potencial de los modelos basados en aprendizaje profundo para la detección automatizada del melanoma, ofreciendo un enfoque prometedor para complementar el trabajo de los especialistas. La integración de estas herramientas en entornos clínicos podría contribuir significativamente a una detección más temprana, ayudando a reducir la mortalidad asociada a esta enfermedad.

Abstract

The incidence of skin cancer has increased significantly in recent years, and melanoma, due to its high mortality rate, represents one of the greatest challenges in dermatology. Early detection of this lesion is crucial, as it significantly improves patient prognosis. Therefore, achieving high sensitivity for the Melanoma class would lead to a lower number of false negatives, ultimately saving numerous lives.

This study addresses the classification of skin lesion images with an approach that prioritizes high sensitivity in melanoma detection, minimizing the likelihood of missing these lesions. To achieve this, a two-phase classification scheme was implemented: an initial binary classification stage to differentiate between melanocytic and non-melanocytic lesions, followed by a four-class classification in cases identified as non-melanoma.

Transformer-based models were employed, recognized for their ability to extract global image representations, enabling the capture of complex patterns relevant to melanoma detection. The results obtained show that this approach achieved high sensitivity for the melanoma class, reinforcing its potential as a support tool for early diagnosis. However, this improvement came at the cost of a decrease in other classification metrics.

Additionally, the study presents and compares the results obtained using different automatic feature extraction models, contrasting them with those based on predefined medical criteria. This comparison allows for an evaluation of both approaches' effectiveness and their applicability in enhancing computer-aided diagnosis systems in dermatology.

The findings of this study demonstrate the potential of deep learning models for automated melanoma detection, offering a promising approach to complement specialists' work. The integration of these tools into clinical settings could significantly contribute to earlier and more accurate detection, helping to reduce the mortality associated with this disease.

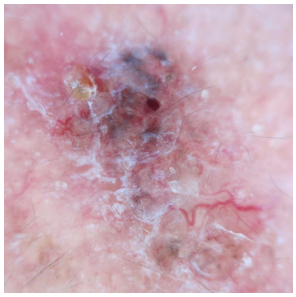
Índice general

Índice de figuras

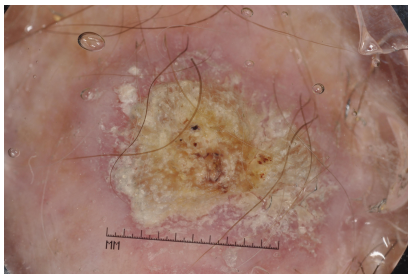
Introducción

El cáncer de piel se ha convertido en un problema de salud cada vez más preocupante a nivel mundial. Entre sus principales causas se encuentra la exposición excesiva a la radiación ultravioleta (UV), proveniente tanto del sol como de fuentes artificiales (por ejemplo, camas de bronceado). El debilitamiento de la capa de ozono, que actúa como filtro natural de los rayos UV, permite que una mayor cantidad de esta radiación alcance la superficie terrestre, incrementando el riesgo de padecer enfermedades cutáneas. Entre los tipos de cáncer de piel, destacan:

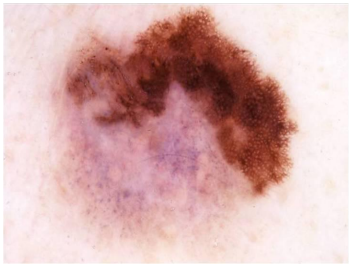
- **Carcinoma Basocelular** [31]: representa aproximadamente el 60 % de los diagnósticos de cáncer de piel. Aunque raramente hace metástasis, puede ocasionar daños locales significativos si no se detecta y trata oportunamente.
- **Carcinoma Espinocelular** [31]: constituye alrededor del 10-20 % de los cánceres de piel y es el segundo más frecuente tras el Carcinoma Basocelular. Frecuentemente se origina de queratinocitos epidérmicos que sufren mutaciones por la exposición repetida a radiación UV. Cuando se detecta tempranamente, su tasa de curación puede llegar al 95 %.
- **Melanoma** [39]: es menos común que los anteriores, pero particularmente peligroso debido a su alta capacidad de diseminación. La detección precoz es fundamental para lograr un mejor pronóstico; de ahí la relevancia de las campañas de prevención y de la evaluación médica periódica de lesiones sospechosas.



Carcinoma Basocelular



Carcinoma Espinocelular



Melanoma

Figura 1: Imágenes representativas de los principales tipos de cáncer de piel

A inicios del siglo XX, el diagnóstico de lesiones cutáneas se basaba casi exclusivamente en la inspección macroscópica y en la experiencia clínica, lo que conllevaba un alto riesgo de error en estadios tempranos de la enfermedad. La llegada de la dermatoscopia en la década de 1980 supuso un hito, al permitir la observación ampliada de estructuras subdérmicas y mejorar de forma considerable la precisión diagnóstica frente a la inspección visual tradicional. Con el avance de la Ciencia de la Computación y la digitalización de las imágenes médicas, comenzaron a emplearse algoritmos de aprendizaje automático —como Máquinas de Soporte Vectorial (SVM, por sus siglas del inglés *Support Vector Machines*)[40, pag. 202] y métodos de K-Vecinos Más Cercanos(KNN por las siglas en inglés de K-Nearest Neighbors)[41] que ofrecían un apoyo adicional al especialista. Sin embargo, la efectividad de estas aproximaciones dependía en gran medida de la extracción manual de características y de la disponibilidad de grandes conjuntos de datos (*datasets*) representativos.

La introducción de AlexNet [22] en 2012 y su éxito en el desafío ImageNet marcó el inicio de una nueva etapa. Las redes neuronales convolucionales (Convolutional Neural Networks, CNNs) [38, pag. 7] mostraron un rendimiento nunca antes visto en el procesamiento de grandes volúmenes de datos, lo que impulsó su adopción en diferentes ámbitos médicos. Posteriormente, arquitecturas como ResNet [18] y EfficientNet [29] consolidaron la pertinencia de las CNNs en la clasificación de imágenes dermatoscópicas.

Más recientemente, el surgimiento de los *Transformers* [46] —originalmente diseñados para el procesamiento del lenguaje natural— ha demostrado un gran potencial también en el campo de la Visión Computacional. En particular, los Vision Transformers (ViT) [10] han conseguido resultados competitivos e incluso superiores a los métodos basados en convoluciones, gracias a su capacidad de capturar relaciones globales en la imagen. Paralelamente, se han desarrollado nuevos métodos de extracción de características, como DINOv2 [32], que permiten generar *embeddings* potentes para

tareas de clasificación, incluso cuando la disponibilidad de datos anotados es limitada.

En la Facultad de Matemática y Computación de la Universidad de La Habana, se han desarrollado dos trabajos significativos relacionados con la clasificación de imágenes dermatoscópicas como apoyo al diagnóstico médico:

1. **Ensemble de redes convolucionales para la clasificación de lesiones de cáncer de piel**, de Claudia Olavarrieta [28]: En este trabajo fueron tomadas 4623 imágenes de tipo dermatoscópico del ISIC Archive, con las que fue entrenado y evaluado un*Ensemble* [47] de tres redes neuronales: una VGG16 [24],ResNet[18] y EfficientNet[29] para la clasificación de lesiones cutáneas.

2. **Características en lesiones cutáneas: El uso de la Inteligencia Artificial en la clasificación de imágenes dermatoscópicas, de Jordan Plá[12]**: Este trabajo aborda la clasificación de lesiones cutáneas desde una perspectiva cercana a los procedimientos que utilizan los dermatólogos con este propósito. Se realiza una extracción de características basada en criterios de diagnóstico y se utilizan para entrenar algoritmos de clasificación.

Debido a la elevada tasa de mortalidad del melanoma y la necesidad de temprana detección y atención de esta lesión, surge la necesidad de aumentar la sensibilidad a la clase Melanoma. Este trabajo propone un enfoque de clasificación en 2 fases haciendo uso de modelos basados en extracción automática de características utilizando la arquitectura *Transformer*.

Este estudio parte de las siguientes **hipótesis**:

- **H1:** El uso de algoritmos basados en la arquitectura *Transformer* para la extracción automática de características produce *embeddings* más robustos que las características manuales, mejorando los resultados de clasificación, sobre todo en la distinción melanoma vs. no melanoma.

- **H2:** Al realizar una clasificación en dos etapas (binaria y luego multicategoría), se consigue una mayor sensibilidad en la detección del melanoma sin sacrificar significativamente el rendimiento global en la clasificación de otras lesiones.

Objetivo general

Estudiar e implementar nuevos modelos de clasificación de lesiones de cáncer de piel como apoyo al diagnóstico, buscando lograr una mayor sensibilidad a la clase Melanoma, haciendo uso de modelos basados en la arquitectura *Transformer*.

Objetivos específicos

- Revisar la bibliografía sobre técnicas de clasificación de imágenes basadas en *Transformers*.
- Basado en dicho estudio, seleccionar arquitecturas y modelos preentrenados que permitan una clasificación satisfactoria de las imágenes
- Implementar y entrenar algoritmos de clasificación con los vectores de características extraídos usando modelos basados en *Transformers* y comparar los resultados con aquellos obtenidos mediante un enfoque de extracción de rasgos utilizando criterios médicos en [12] entrenando y evaluando en el conjunto utilizado en esta tesis.
- Entrenar los modelos con imágenes médicas del *dataset* ISIC, el cual constituye un conjunto altamente variado, utilizar técnicas de aumento de datos y realizar un balanceo de clases para garantizar el satisfactorio complimiento de este objetivo.
- Evaluar el desempeño de los modelos implementados, tanto en la clasificación binaria (melanoma vs. no melanoma) como en la clasificación de múltiples categorías, así como del enfoque de dos fases en un subconjunto de ISIC separado del conjunto de entrenamiento para ser utilizado con este propósito.

Estructura de la tesis

Este documento se organiza en seis capítulos. El **Capítulo 1** profundiza en la revisión bibliográfica, haciendo énfasis en la aplicación de Transformers para la clasificación de lesiones dermatoscópicas. En el **Capítulo 2** se aborda el tema de las imágenes digitales y la técnica de la dermatoscopía. El **Capítulo 3** introduce los conceptos fundamentales de la Inteligencia Artificial, el Aprendizaje de Máquinas y el Aprendizaje Profundo. A continuación, en el **Capítulo 4**, se describen las Redes Neuronales y algunas técnicas avanzadas utilizadas en su implementación, se presenta la arquitectura Transformer, su adaptación a la Visión Computacional y el algoritmo DINOv2. En el **Capítulo 5** Se describe el conjunto de datos empleado, así como las técnicas de aumento para su preparación, se detallan los modelos seleccionados y el proceso de entrenamiento. Finalmente, en el **Capítulo 6** se presentan los resultados obtenidos, discutiendo su relevancia y limitaciones para la validación final de los modelos.

Capítulo 1

Estado del Arte

Gracias a su capacidad para capturar relaciones complejas entre los píxeles de las imágenes, los *Vision Transformers* (ViTs) han alcanzado el estado del arte en diversas tareas de Visión Computacional, igualando, e incluso superando [34], los resultados de su predecesor, las Redes Neuronales Convolucionales (CNNs). Entre estas tareas se encuentra la dermatoscopia. Satoshi Takahashi et al. [43] ofrecen una revisión sistemática que compara *Vision Transformers* y Redes Neuronales Convolucionales en el análisis de imágenes médicas, revisando 36 estudios previos de comparación entre ambos enfoques en diferentes áreas de la medicina. La revisión concluye que, en la mayoría de los casos, los ViTs igualan o superan a las CNNs, aunque algunos estudios muestran mejores resultados para las CNNs. Las principales ventajas de los *Vision Transformers* incluyen su eficiencia computacional y su capacidad para capturar relaciones globales en la imagen, mientras que sus desventajas suelen ser su capacidad limitada para capturar características locales y la gran cantidad de datos requeridos para su entrenamiento [10].

Además del modelo original de ViT, se han desarrollado diversas variantes que han mostrado resultados competitivos en el análisis de imágenes dermatoscópicas. Entre estas variantes destacan el *Swin Transformer*, presentado por Ze Liu et al. [26], esta arquitectura introduce ventanas deslizantes para capturar características locales, superando así la limitación original de los ViTs, y posee una estructura jerárquica que mejora la eficiencia al procesar imágenes de alta resolución. Otra variante relevante es el *Data-Efficient Image Transformer* (DeiT), propuesto por Hugo Touvron et al. [45], diseñado para entrenarse de manera más eficiente utilizando menos datos, incorporando técnicas como el *distillation token* para mejorar el aprendizaje supervisado. Otros estudios, como el realizado por Somaiya Khan et al. [21], han propuesto modelos híbridos entre CNNs y ViTs, con el objetivo de combinar lo mejor de ambos enfoques.

En el 2021, Mathilde Caron et al. [3] presentaron el algoritmo de DINO. En

este trabajo, los autores proponen un enfoque de destilación para el entrenamiento de *Vision Transformers* sin la necesidad de etiquetas. La técnica se basa en el aprendizaje auto-supervisado, lo que permite generar representaciones visuales de alta calidad sin la necesidad de anotaciones etiquetadas, mejorando así el rendimiento de los modelos en tareas de visión por computadora.

Posteriormente, en el 2023, Maxime Oquab et al. [32], presentaron DINOv2, una versión mejorada de DINO que ofrece representaciones visuales de mayor calidad, además de ser más eficiente y escalable, logrando avances significativos en la capacidad de los modelos para generalizar en una variedad de tareas de visión por computadora. Jayanth Mohan et al. [30] abordan la clasificación de enfermedades cutáneas mediante arquitecturas avanzadas de aprendizaje profundo basadas en *Transformers*. Los autores emplean un conjunto de datos que abarca 31 clases de enfermedades de la piel y comparan diversas arquitecturas de *Transformers*, incluyendo *Vision Transformers, Swin Transformers* y DINOv2, así como modelos convolucionales tradicionales. Mediante el uso de aprendizaje por transferencia con pesos preentrenados en ImageNet1k, el modelo DINOv2 alcanzó una precisión de prueba del 96.48 % y una puntuación F1 de 0.9727, superando en aproximadamente un 10 % los resultados de referencia previos. Además, se evaluó la robustez del modelo DINOv2 utilizando los conjuntos de datos HAM10000 y Dermnet, obteniendo mejoras marginales en precisión y puntuación F1 en estos conjuntos de datos.

Con respecto a la clasificación binaria (Melanoma/No melanoma) utilizando vectores de características, Jayanth Mohan et al. [23] proponen un método para identificar si una muestra dada es o no un melanoma. En este artículo, aplanan la matriz de la imagen en un arreglo de intensidades de píxeles, que utilizan posteriormente para entrenar una SVM, logrando una exactitud de alrededor de 90 %, un enfoque similar para clasificación multicategórica utilizando vectores de características fue presentado en [1], donde utilizan un enfoque de 4 etapas: pre-procesamianto, segmentación, extracción de características y clasificación, para esta última fase, son utilizados Árboles de Decisión [40, pág. 250], SVMs, KNN y métodos de Ensemble, por cada uno de los clasificadores, se prueban distintos tipos de Kernel, utilizándose una Validación Cruzada(*Cross-Validation*) para el entrenamiento y evaluación de los modelos obteniéndose los mejores resultados con la SVM con kernel cuadrático.

Capítulo 2

Visión, Imágenes y Dermatoscopia

En este capítulo se introducen los conceptos de visión y de visión computacional. Luego se explica qué es una imagen, prestando fundamental atención a las imágenes digitales. Finalmente se presenta el concepto de dermatoscopia, tema sobre el que se desarrolla esta tesis.

2.1. Visión y Visión Computacional

La visión [15] es un regalo extraordinario que la naturaleza nos ha otorgado. Desde el primer destello de luz que nuestros ojos perciben al nacer hasta los paisajes que admiramos con asombro, este sentido ha sido un componente esencial en la manera en que interactuamos con el mundo. La luz, que sostiene la vida en nuestro planeta, no solo nos permite existir, sino también interpretar y comprender nuestro entorno. Nuestra capacidad para percibirla y traducirla en información visual es un triunfo evolutivo.

Los humanos capturamos la luz a través de un lente que la concentra en nuestras retinas, donde se convierte en señales eléctricas. Estas señales viajan a través del sistema nervioso hasta el cerebro, que reconstruye esta información para darnos una representación coherente de nuestro entorno. Este proceso, conocido como visión, es tan crucial que los científicos han especulado que el desarrollo de sistemas nerviosos centralizados está estrechamente relacionado con la aparición de este sentido. Sin los vastos volúmenes de información que los ojos capturan, no habría razón para desarrollar el complejo aparato cerebral que poseemos.

La visión también ha definido nuestra interacción con el mundo de maneras más prácticas. Considere algo tan simple como patear una pelota. En ese instante, el cerebro realiza una serie de cálculos intrincados en fracciones de segundo: identifica la pelota, rastrea su movimiento, predice su trayectoria, calcula la velocidad a la que llegará, ajusta la fuerza y el ángulo del impacto, y envía las señales necesarias para que

el pie se coloque en la posición correcta. Este proceso fluido y automático ejemplifica la capacidad del cerebro para transformar una imagen visual en una acción.

Este nivel de procesamiento ocurre sin una educación formal. No aprendemos cálculos explícitos para determinar la fuerza necesaria para un tiro; lo hacemos por ensayo y error desde niños. Sin embargo, replicar incluso la tarea más sencilla de este proceso en un sistema artificial es desafiante. Por ejemplo, identificar una pelota. Podríamos intentar definir qué es una pelota y buscarla exhaustivamente en una imagen. Sin embargo, ¿cómo definimos una pelota? Su tamaño varía desde pequeñas pelotas de tenis hasta enormes esferas usadas en deportes extremos. Factores como estos complican la definición.

Esta capacidad de los humanos para reconocer objetos no depende exclusivamente de reglas estrictas. En su lugar, generalizamos conceptos relacionados y usamos pistas contextuales. Reconocemos un balón de fútbol incluso si está cubierto de plumas, pero no consideramos un volante de bádminton como una pelota, a pesar de su uso en juegos. Este tipo de comprensión implícita, acumulada a través de experiencias y memorias, es algo que los sistemas artificiales tradicionales no poseen.

Aquí radica la importancia de la visión computacional. Este campo busca construir sistemas que puedan interpretar información visual y tomar decisiones de manera similar a los humanos. Sin embargo, la visión computacional no intenta simplemente replicar la visión humana. Su enfoque incluye problemas que serían demasiado tediosos, costosos o propensos a errores si se realizaran manualmente. Por ejemplo, un modelo capaz de rastrear una pelota puede agilizar decisiones en eventos deportivos, haciendo el juego más justo y accesible.

Además, con los avances en modelos de texto a imagen y de imagen a texto, es posible describir eventos deportivos en tiempo real para personas con discapacidades visuales, brindándoles una experiencia inclusiva. De esta manera, incluso las aplicaciones aparentemente simples pueden tener un impacto significativo en la sociedad.

Vivimos en un momento emocionante, donde la inteligencia artificial está expandiendo los límites de lo posible. Podemos entrenar modelos para detectar patrones que los humanos no perciben y generar nuevas representaciones visuales desde descripciones textuales. La visión computacional está en todas partes: desde nuestros teléfonos inteligentes hasta aplicaciones industriales[42].

2.2. ¿Qué es una imagen?

Una imagen [14] es una representación visual de un objeto, una escena, una persona o incluso un concepto. Puede ser una fotografía, una pintura, un dibujo, un esquema, una exploración digital y mucho más. Otra forma de representar una imagen es mediante una función. Inicialmente consideremos una imagen como una función bidimensional, F(X, Y), donde X e Y representan coordenadas espaciales. Las coorde-

nadas espaciales no son más que un sistema que utilizamos para describir posiciones en un espacio físico, siendo el más común el sistema cartesiano 2D. Al evaluar F en un par de coordenadas (x, y) se le hace corresponder la intensidad, nivel de gris o color de la imagen en ese punto, lo que permite percibir luces y sombras.

Las imágenes digitales son discretas en su representación, aunque los procesos que las generan suelen ser continuos, ya que solo es posible almacenar una cantidad finita de valores. Usualmente cuando tenemos el par de coordenadas (x_i, y_i) , nos referimos a ella como píxel(*picture element*) [5, pág. 49]. La cantidad de píxeles usados para cubrir el espacio visual, es decir, la cantidad de píxeles por columnas y por filas es conocida como resolución [5, pág 3].

Un tipo distinto de imagen es la volumétrica o tridimensional (3D). En este caso, la función es F(X, Y, Z), donde las coordenadas (x, y, z) describen un vóxel (volume element). Estas imágenes pueden obtenerse directamente en 3D mediante escaneos médicos, resonancias magnéticas o ciertos tipos de microscopios.

Otra característica importante de las imágenes es el color, representado mediante canales. Los canales de una imagen son componentes individuales de color, como el rojo, verde y azul en el sistema RGB [5, pág. 10]. En este caso, F(X, Y) tiene valores independientes para cada componente de color. La intensidad de un canal indica cuán predominante es ese color en un punto específico. Por ejemplo, en un canal de rojo, una alta intensidad representa un rojo vibrante, mientras que una baja intensidad indica poca o ninguna presencia de ese color.

Existen también imágenes especiales conocidas como imágenes etiquetadas. En estas, las coordenadas no describen niveles de intensidad, sino que asignan etiquetas a cada píxel. Por ejemplo, en una imagen binaria, el primer plano puede etiquetarse como 1 y el fondo como 0. Estas imágenes son útiles en tareas de segmentación, como separar objetos del fondo.

En campos como la biomedicina, se habla incluso de imágenes en 4D o 5D. Estas imágenes incorporan dimensiones adicionales, como el tiempo o diferentes modalidades de captura (por ejemplo, combinando una foto con una radiografía). La idea es que cada nueva fuente de información agrega una dimensión, permitiendo un análisis más completo de los datos visuales.

En computadoras, las imágenes suelen representarse como matrices, lo que facilita su manipulación. Cada entrada en la matriz corresponde a un píxel y su valor refleja la intensidad o el color en ese punto. Alternativamente, una imagen también puede representarse como un grafo, donde los nodos son coordenadas y los bordes conectan nodos vecinos. Esta flexibilidad permite aplicar algoritmos tanto de procesamiento de imágenes como de teoría de grafos.

En resumen, una imagen es mucho más que una representación visual; es un conjunto de datos ricos en información espacial. Las diferencias clave entre tipos de imágenes radican en su resolución espacial (2D o 3D), sistemas de color (RGB u otros)

y componentes adicionales como el tiempo.

2.3. Dermatoscopia

La dermatoscopia [33] es una técnica diagnóstica no invasiva y de fácil ejecución, utilizada para examinar las lesiones cutáneas de forma detallada. Esta herramienta permite observar estructuras cutáneas que no son visibles a simple vista, facilitando así el diagnóstico y la evaluación de las lesiones, especialmente aquellas hiperpigmentadas [6]. Es particularmente útil en el contexto de atención primaria, ya que mejora el diagnóstico diferencial entre melanoma y otras lesiones hiperpigmentadas, lo que resulta en una intervención más temprana y precisa.

Desde su desarrollo, se han identificado numerosas estructuras dermatoscópicas que sirven como parámetros para caracterizar las lesiones cutáneas. Estas estructuras, también conocidas como criterios dermatoscópicos, corresponden a hallazgos profundos en la piel que tienen una estrecha relación con ciertas características histopatológicas. La dermatoscopia proporciona una perspectiva «horizontal» de la lesión, ya que permite observar patrones y distribuciones de estructuras en un plano superficial, mientras que la anatomía patológica brinda una perspectiva «vertical» al analizar cortes transversales del tejido para evaluar las capas involucradas en profundidad. Ambas técnicas se complementan entre sí, combinando sus enfoques para lograr una mayor precisión diagnóstica.

La dermatoscopia ha sido objeto de varios estudios en los últimos años, con el objetivo de identificar las estructuras que presentan una mayor sensibilidad y especificidad, y que además sean fácilmente reproducibles por diferentes observadores. Entre los parámetros dermatoscópicos más relevantes para el diagnóstico de lesiones hiperpigmentadas se encuentran: la pigmentación y el color, el retículo pigmentado, los puntos, los glóbulos, las proyecciones radiales, los pseudópodos, las lagunas rojo-azuladas, las estructuras vasculares, las estructuras en forma de hoja de arce, las estructuras en forma de rueda de carro, los nidos grandes ovalados azulados, los glóbulos múltiples, el parche central blanco, las fisuras, las criptas, los quistes tipo millium y los tapones córneos. Además de estos, existen otros hallazgos, como el pseudorretículo pigmentado y las estructuras de regresión, que también pueden ser de gran utilidad en el diagnóstico.

Capítulo 3

Inteligencia Artificial

En este capítulo se explica brevemente en que consiste la Inteligencia Artificial, el aprendizaje de máquinas, así como subdisciplinas del mismo como lo son el aprendizaje supervisado y no supervisado.

3.1. ¿Qué es la Inteligencia Artificial?

La Inteligencia Artificial (IA) [37, pág 1] es una rama de la informática que se centra en la creación de sistemas capaces de realizar tareas que normalmente requieren la intervención de la inteligencia humana. Estas tareas incluyen habilidades como el aprendizaje, la comprensión del lenguaje, la toma de decisiones, la percepción del entorno, el razonamiento y la resolución de problemas.

El objetivo de la IA es desarrollar máquinas que puedan imitar, de alguna manera, las capacidades cognitivas humanas, como la percepción visual, el procesamiento del lenguaje o la capacidad de aprender de la experiencia. Aunque la IA no se refiere a una sola tecnología, se basa en la idea de que es posible crear algoritmos y modelos que permitan a las máquinas realizar tareas complejas de forma autónoma.

Desde sus inicios en la década de 1950, la IA ha avanzado a pasos agigantados. En sus primeros años, los investigadores se centraron en crear programas que pudieran simular ciertas capacidades humanas, como el juego de ajedrez o la resolución de problemas matemáticos. Con el tiempo, la IA ha evolucionado y se ha integrado en muchas áreas de la vida cotidiana. Hoy en día, se utiliza para desarrollar asistentes virtuales como Siri o Alexa, mejorar sistemas de recomendación en plataformas como Netflix y Amazon, ayudar en diagnósticos médicos, e incluso hacer posible la conducción autónoma de vehículos.

Un aspecto fundamental de la IA es su capacidad para aprender de los datos. En lugar de programar explícitamente cada tarea, los sistemas de IA pueden mejorar su rendimiento con el tiempo a medida que se exponen a más información. Esto les

permite adaptarse a nuevas situaciones y mejorar su capacidad para realizar tareas de forma más precisa.

En general, la IA busca ampliar las capacidades humanas y proporcionar soluciones innovadoras a problemas complejos, transformando la forma en que interactuamos con la tecnología y mejorando diversas industrias y servicios.

3.2. Aprendizaje de Máquinas

El Aprendizaje de Máquinas [37, pág 650] (Machine Learning, ML) es una subdisciplina de la IA que se enfoca en el diseño de algoritmos y modelos que permiten a las máquinas aprender a partir de datos. En lugar de ser programados explícitamente para cada tarea, los sistemas de ML identifican patrones y relaciones en los datos para realizar predicciones o decisiones.

El ML se ha convertido en un pilar fundamental de la IA moderna debido a su capacidad para abordar una amplia gama de problemas prácticos. Desde la personalización de experiencias de usuario hasta la optimización de procesos industriales, las aplicaciones del ML están presentes en todos los ámbitos de la sociedad. Entre las características del aprendizaje de maquinas son destacables:

1. Generalización: Habilidad de un modelo para realizar predicciones precisas en datos no vistos previamente.
2. Adaptabilidad: Capacidad para ajustar modelos a medida que se obtienen nuevos datos.
3. Automatización: Reducción de la intervención manual en el diseño de soluciones específicas.

Existen tres categorías principales de aprendizaje de máquinas según el tipo de datos y el enfoque utilizado:

- **Aprendizaje supervisado [37, pág 653] :** Utiliza datos etiquetados.
- **Aprendizaje no supervisado [37, pág 775]:** Busca patrones en datos no etiquetados.
- **Aprendizaje por refuerzo [37, pág 789]:** Entrena a un agente a través de recompensas y castigos.

El ML ha permitido avances significativos en diversas áreas, como el reconocimiento de imágenes, el procesamiento del lenguaje natural [37, pág 823]: y la detección de fraudes.

Aprendizaje Supervisado:

En el aprendizaje supervisado, los algoritmos aprenden a partir de un conjunto de datos donde cada entrada está asociada a una salida conocida. Este enfoque es ampliamente utilizado en problemas como:

- **Clasificación:** Asignar etiquetas a ejemplos, como identificar correos electrónicos como spam o no spam.
- **Regresión:** Predecir valores continuos, como el precio de una vivienda en función de sus características.

Los modelos supervisados se entrenan minimizando una función de pérdida, que mide la discrepancia entre las predicciones del modelo y las salidas reales. Este proceso de entrenamiento incluye:

1. **Preparación de datos:** Limpieza, selección de características y normalización para asegurar que el modelo reciba entradas relevantes y consistentes.
2. **División de datos:** Separar los datos en conjuntos de entrenamiento, validación y prueba.
3. **Entrenamiento:** Ajuste de parámetros del modelo para minimizar la función de pérdida utilizando técnicas de optimización como el descenso por gradiente.
4. **Evaluación:** Medición del rendimiento del modelo en datos no vistos utilizando métricas como precisión, sensibilidad o error cuadrático medio.

   El aprendizaje supervisado es fundamental para tareas que requieren predicciones precisas basadas en grandes volúmenes de datos etiquetados.

Aprendizaje No Supervisado

El **aprendizaje no supervisado** es un tipo de enfoque de aprendizaje automático en el que el modelo intenta identificar patrones o estructuras en los datos sin utilizar etiquetas predefinidas. A diferencia del **aprendizaje supervisado**, donde el modelo se entrena con ejemplos que tienen entradas y sus respectivas salidas (etiquetas), en el aprendizaje no supervisado el modelo solo tiene acceso a las entradas (datos sin etiquetar) y debe descubrir por sí mismo alguna estructura subyacente o patrón.

El objetivo principal del aprendizaje no supervisado es explorar los datos y encontrar alguna relación o regularidad. Existen varias técnicas populares en este enfoque:

- **Clustering (Agrupamiento):** El modelo organiza los datos en grupos (o clusters) basados en características similares. Un ejemplo clásico es el algoritmo **K-means**, que busca dividir los datos en un número especificado de *clusters*, donde los elementos dentro de cada grupo son más similares entre sí que con los de otros grupos.
- **Reducción de dimensionalidad:** En muchos casos, los datos pueden tener un número muy alto de características, lo que puede dificultar su análisis. La reducción de dimensionalidad busca encontrar una representación de los datos en un espacio de menor dimensión, preservando la mayor cantidad posible de la información. Algunos ejemplos son **PCA (Análisis de Componentes Principales)** y **t-SNE (t-distributed Stochastic Neighbor Embedding)** [27].
- **Modelado de densidad:** Este enfoque intenta identificar la distribución subyacente de los datos, lo que puede ayudar a entender su estructura. **Estimación de densidad** es un ejemplo, donde se busca aprender una función que pueda generar una estimación de la probabilidad de que un punto dado pertenezca a un conjunto de datos.

Algunas aplicaciones comunes del aprendizaje no supervisado incluyen:

- **Segmentación de clientes:** En marketing, el aprendizaje no supervisado puede ayudar a agrupar clientes en diferentes segmentos según su comportamiento, sin tener una etiqueta de "segmento"previamente definida.
- **Análisis de anomalías:** Detectar puntos de datos que son diferentes de la mayoría, lo cual es útil en aplicaciones como la detección de fraudes o fallos en sistemas.
- **Análisis de imágenes:** Técnicas como el clustering pueden ser utilizadas para agrupar imágenes similares, o para reducir la dimensionalidad de características visuales.

En resumen, el aprendizaje no supervisado se enfoca en descubrir estructuras ocultas en los datos sin necesidad de que se les haya proporcionado una supervisión explícita, como las etiquetas.

3.3. Aprendizaje Profundo

El aprendizaje profundo (Deep Learning) es un subcampo del ML que utiliza redes neuronales profundas para modelar relaciones complejas en los datos. Estas redes

están compuestas por múltiples capas de neuronas artificiales, que extraen representaciones jerárquicas de los datos.

El aprendizaje profundo ha sido un catalizador en la evolución de la IA, permitiendo avances en:

- **Visión por computadora:** Reconocimiento de imágenes, detección de objetos y segmentación semántica.
- **Procesamiento del lenguaje natural:** Traducción automática, generación de texto y análisis de sentimientos.
- **Juegos:** Entrenamiento de agentes que superan el desempeño humano en juegos como Go y StarCraft.

Las redes neuronales profundas requieren grandes cantidades de datos y potencia computacional para entrenarse, pero su capacidad para aprender representaciones complejas las hace invaluables en tareas avanzadas de IA.

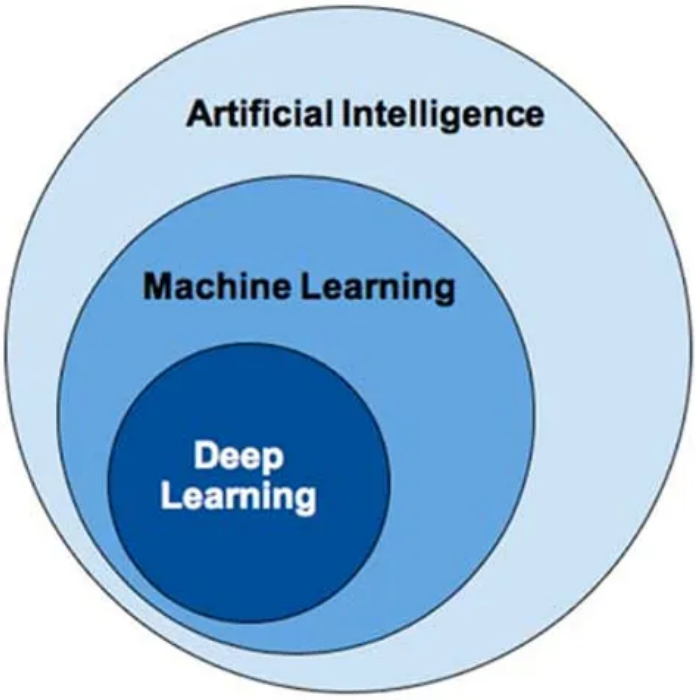


Figura 3.1: Diagrama de la relación entre Inteligencia Artificial, Aprendizaje Automático y Aprendizaje Profundo, extraída de [2]

Capítulo 4

Redes Neuronales Artificiales

En este capítulo se profundiza acerca de las redes neuronales, en la sección 4.1, se introduce el perceptrón, la red neuronal más básica, así como su adaptación para múltiples salidas y su extensión para representar funciones más complejas utilizando múltiples capas. En 4.2 se presenta la arquitectura *Transformer*, introducida en el 2017, revolucionando el campo del Procesamiento de Lenguaje Natural, posteriormente, en 4.3, se introduce su adaptación para tareas de Visión por Computadoras, el *Vision Transformer*. En 4.4 se introduce el aprendizaje por Transferencia, una técnica de gran utilidad cuando se cuenta con recursos de cómputo o datos limitados. En 4.5 se aborda el aprendizaje autosupervisado y en 4.6 se presenta el novedoso algoritmo de DINOv2, un modelo que actúa como extractor automático de características.

4.1. El Perceptrón

El perceptrón es uno de los modelos más simples y fundamentales en el campo de las redes neuronales artificiales. Consiste en una única unidad de salida conectada a un conjunto de unidades de entrada mediante pesos que son ajustados durante el proceso de entrenamiento. Este modelo básico es capaz de realizar tareas de clasificación binaria aplicando una función de activación sobre una combinación lineal de las entradas.

Estructura del Perceptrón

La estructura de un perceptrón se compone de los siguientes elementos:

- **Unidades de entrada**: Representadas por $v_1, v_2, \ldots, v_D$, son los valores de entrada que describen una observación o dato.

- **Pesos**: Cada conexión entre una unidad de entrada y la unidad de salida tiene un peso asociado $w_1, w_2, \ldots, w_D$ que determina la influencia de esa entrada en la salida. Estos pesos no son introducidos por el usuario, en cambio, son calculados mediante el algoritmo de Retropropagación [38, pag. 6].
- **Sesgo** (bias): Representado como w_0, es un término adicional que permite ajustar la función de activación.
- **Unidad de salida**: Calcula el valor de salida y aplicando una función de activación $f(\cdot)$ sobre la suma ponderada de las entradas.

El cálculo de la salida se expresa matemáticamente como:

$$y(v;\theta) = f\left(\sum_{i=1}^{D} v_i w_i + w_0\right) = f(w \cdot v + w_0), \tag{4.1}$$

donde $\theta = \{w, w_0\}$ representa el conjunto de parámetros del modelo. La función de activación comúnmente utilizada en el perceptrón es la sigmoide logística, definida como $\sigma(z) = \frac{1}{1+\exp(-z)}$, ideal para tareas de clasificación binaria.

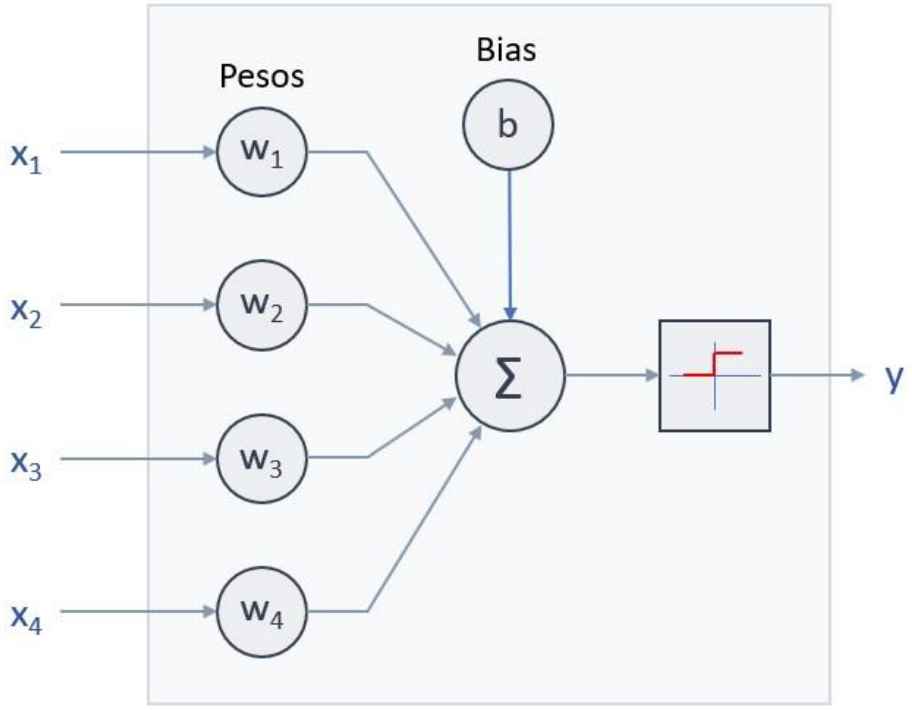


Figura 4.1: Formulación matemática del Perceptrón. Extraído de [19]

Extensiones del Perceptrón para Tareas con múltiples salidas

El perceptrón puede extenderse para manejar tareas con múltiples salidas, como clasificación multiclase. Esto se logra añadiendo varias unidades de salida, cada una con su propio conjunto de pesos:

$$y_k(v;\theta) = f\left(\sum_{i=1}^{D} v_i w_{ki} + w_{k0}\right) = f(w_k \cdot v + w_{k0}), \tag{4.2}$$

donde w_{ki} denota el peso que conecta la entrada v_i con la salida y_k.

Para la clasificación multiclase, se utiliza típicamente la función *softmax*, que convierte las salidas en probabilidades normalizadas:

$$s(z_k) = \frac{\exp(z_k)}{\sum_{l=1}^{K} \exp(z_l)}. \tag{4.3}$$

4.1.1. Perceptrón Multicapa

El **perceptrón multicapa** (MLP, por sus siglas en inglés: Multi-Layer Perceptron) es una extensión del perceptrón simple, el cual es un tipo de red neuronal artificial utilizada para tareas de clasificación y regresión. A diferencia del perceptrón simple, que consta únicamente de una capa de entrada y una capa de salida, el perceptrón multicapa está compuesto por múltiples capas: una capa de entrada, al menos una capa oculta y una capa de salida. La introducción de las capas ocultas permite que el MLP pueda modelar relaciones no lineales complejas, a diferencia del perceptrón simple, que solo puede resolver problemas lineales.

Estructura del Perceptrón Multicapa

El MLP está compuesto por tres tipos principales de capas:

- **Capa de entrada:** Es la capa inicial de la red, encargada de recibir los datos de entrada. Cada nodo en esta capa representa una característica de los datos de entrada.
- **Capas ocultas:** Estas capas se encuentran entre la capa de entrada y la capa de salida. Los nodos en las capas ocultas realizan cálculos utilizando pesos y funciones de activación no lineales. El número de capas ocultas y la cantidad de nodos por capa son parámetros ajustables que afectan el rendimiento del modelo.
- **Capa de salida:** Esta es la capa final de la red, cuya salida representa el resultado de la red neuronal. En tareas de clasificación, la salida suele ser la probabilidad de pertenecer a cada clase o la clase predicha.

Funcionamiento del Perceptrón Multicapa

El MLP realiza su cálculo en varias etapas a través de las diferentes capas de la red. Cada capa lleva a cabo una transformación de los datos utilizando una combinación de pesos, sesgos y una función de activación. El proceso general se describe a continuación:

1. **Propagación hacia adelante (Forward Propagation):** Para cada capa l, las salidas $y_k^{(l)}$ se calculan como una función no lineal $f^{(l)}$ de la suma ponderada de las entradas de esa capa:

$$y_k^{(l)} = f^{(l)}\left(\sum_m w_{km}^{(l)} y_m^{(l-1)}\right) \tag{4.4}$$

donde:

- $w_{km}^{(l)}$ son los pesos de la capa l,
- $y_m^{(l-1)}$ son las salidas de la capa anterior (o las entradas en el caso de la capa inicial),
- $f^{(l)}$ es la función de activación aplicada en la capa l.

2. **Funciones de activación:** Se emplean funciones no lineales, como la sigmoide (logística) o la tangente hiperbólica (tanh), para garantizar que la red sea capaz de modelar relaciones no lineales. Si no se utilizaran funciones de activación no lineales, la red de múltiples capas se reduciría a una red de una sola capa, equivalente a una transformación lineal.

Formulación Matemática

El MLP puede representarse de manera matemática mediante una composición de funciones de activación y sumas ponderadas a través de las capas. Para una red de dos capas, la salida y_k se expresa como:

$$y_k(v;\Theta) = f^{(2)}\left(\sum_j w_{kj}^{(2)} f^{(1)}\left(\sum_i w_{ji}^{(1)} v_i\right)\right) \tag{4.5}$$

donde:

- v_i son los valores de entrada,
- $w_{ji}^{(1)}$ son los pesos de la capa oculta,
- $f^{(1)}$ es la función de activación en la capa oculta,

- $w_{kj}^{(2)}$ son los pesos de la capa de salida,
- $f^{(2)}$ es la función de activación en la capa de salida.

Para un MLP con L capas ocultas, la función de salida se describe como una composición de funciones de activación a través de las capas:

$$y_k = f^{(L)}\left(\sum_l w_{kl}^{(L)} f^{(L-1)}\left(\sum_m w_{lm}^{(L-2)}\left(\dots f^{(1)}\left(\sum_i w_{ji}^{(1)} x_i\right)\right)\right)\right) \tag{4.6}$$

Limitaciones del Perceptrón Simple y Ventajas del MLP

El perceptrón simple (sin capas ocultas) tiene la limitación de que solo puede aprender funciones lineales, lo que significa que solo puede clasificar correctamente datos que sean separables de manera lineal. En contraste, el MLP, al incluir capas ocultas con funciones de activación no lineales, puede aprender representaciones más complejas y no lineales de los datos, lo que le permite resolver problemas mucho más complejos, como la clasificación de datos no linealmente separables.

Por lo tanto, la capacidad del MLP para incorporar capas ocultas y funciones de activación no lineales permite superar la limitación del perceptrón simple y expandir su capacidad para resolver una mayor variedad de tareas.

4.2. Transformers

En los últimos años, los modelos de aprendizaje profundo han experimentado una revolución gracias a la introducción de los *Transformers*, una arquitectura propuesta por Vaswani et al. en su trabajo seminal *Attention is All You Need* en 2017. Este modelo representó un avance significativo en el campo del procesamiento del lenguaje natural (NLP), al abordar de manera eficiente la complejidad inherente al manejo de secuencias largas y dependencias contextuales.

Antes de los *Transformers*, las arquitecturas predominantes para tareas secuenciales eran las redes neuronales recurrentes (RNNs) y sus variantes como las Long Short-Term Memory(LSTM) [13]. Si bien estos modelos lograban buenos resultados, su naturaleza secuencial limitaba la capacidad de paralelismo durante el entrenamiento y dificultaba la captura de dependencias a largo plazo. Los transformers solucionaron estos problemas al introducir el mecanismo de atención, que permite procesar todas las palabras de una secuencia simultáneamente, asignando pesos de relevancia entre ellas sin necesidad de recorrerlas de manera secuencial.

El impacto de los transformers ha trascendido el área de NLP, influyendo también en campos como la visión por computadora, donde arquitecturas como los Vision

Transformers (ViT) han demostrado ser competitivas frente a las redes convolucionales tradicionales. En el contexto de esta tesis, los Vision Transformers son particularmente relevantes, ya que permiten analizar y clasificar de manera precisa imágenes dermatoscópicas, capturando patrones complejos presentes en las lesiones cutáneas.

4.2.1. Arquitectura

La arquitectura de los transformers consta de varios componentes fundamentales que trabajan en conjunto para procesar secuencias de datos de manera eficiente. A continuación, se describen los componentes principales y el flujo de datos a través del modelo.

- **Capa de embedding:** Esta capa transforma las entradas discretas, como palabras o píxeles, en vectores continuos de dimensiones fijas. Este paso inicial es crucial para que el modelo pueda manejar y procesar los datos de entrada en un espacio vectorial.
- **Positional Encoder:** Dado que los transformers no procesan los datos de manera secuencial, el *positional encoding* agrega información sobre la posición relativa de los elementos en la secuencia. Esto se logra mediante funciones trigonométricas que generan patrones que el modelo puede aprender a interpretar.
- **Bloques de Atención Multi-Cabezas:** El mecanismo de atención [25] multi-cabezas permite que el modelo enfoque su atención en diferentes partes de la secuencia de manera simultánea. Cada cabeza de atención opera de forma independiente, aprendiendo relaciones específicas entre los elementos de la secuencia. Posteriormente, las salidas de todas las cabezas se combinan para formar una representación enriquecida.
- **Redes Feed-Forward:** Cada bloque de atención es seguido por una red completamente conectada que procesa las representaciones generadas, aplicando transformaciones no lineales para enriquecer la representación de los datos.
- **Capa de Normalizacion y Residual Connections:** La normalización (*Layer Normalization*) estabiliza el entrenamiento al asegurar que las distribuciones de activaciones permanezcan consistentes a lo largo de las capas. Las conexiones residuales permiten que las representaciones originales se sumen a las salidas transformadas, mejorando la propagación del gradiente y acelerando la convergencia.

El flujo de datos en un transformer comienza con la entrada a los embeddings, seguido por la adición del *positional encoding*. Esta representación enriquecida pasa por múltiples bloques de atención multi-cabezas y redes feed-forward, cada uno

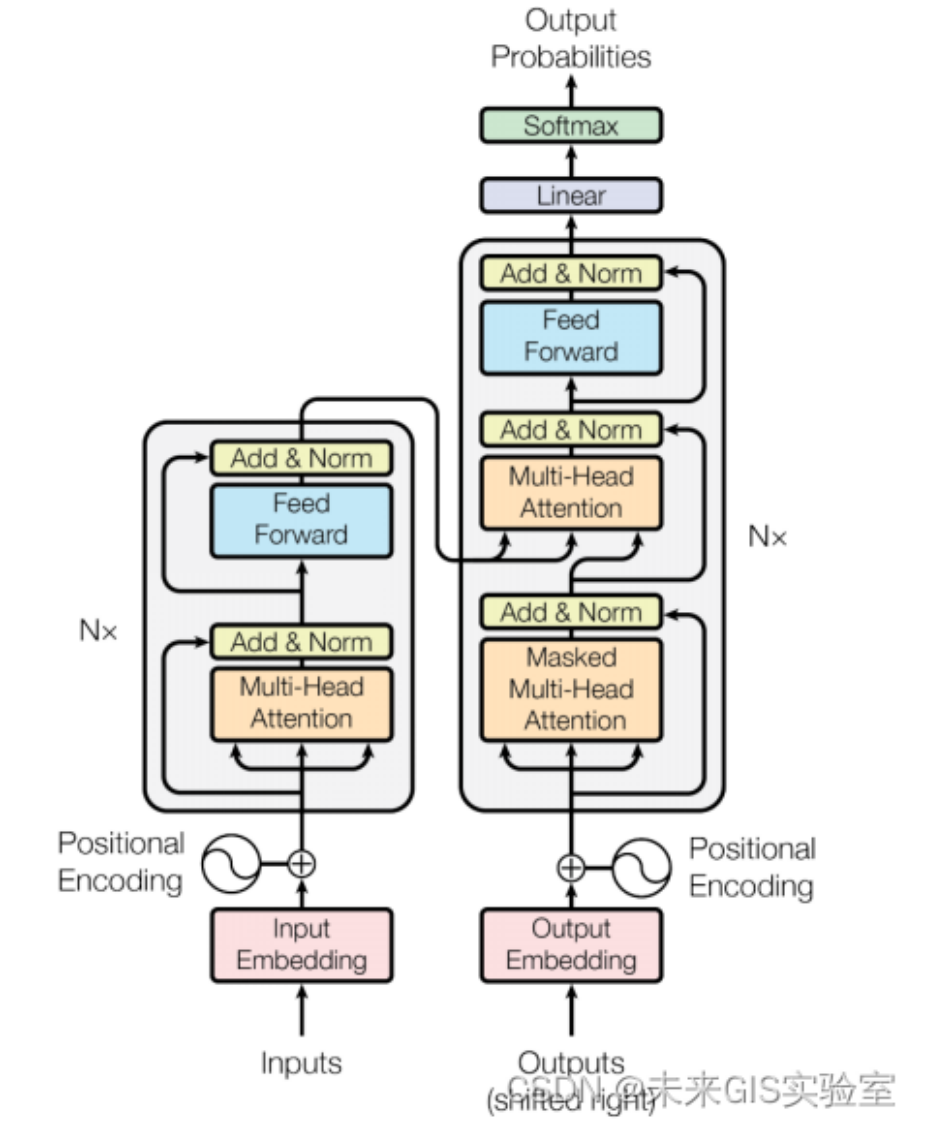


Figura 4.2: Arquitectura Transformer, presentada en [46]

equipado con normalización y conexiones residuales. Finalmente, las representaciones resultantes pueden ser utilizadas para diversas tareas, como clasificación, traducción o generación de secuencias.

4.3. Vision Transformer

El Vision Transformer (ViT) es un modelo de aprendizaje profundo que ha revolucionado el campo de la clasificación de imágenes, marcando un cambio significativo en la forma en que las arquitecturas de reconocimiento visual son diseñadas y aplicadas. A diferencia de las redes neuronales convolucionales (CNN), que han sido históricamente la norma para tareas de visión por computador, el ViT utiliza principios de transformación que fueron originalmente desarrollados para el procesamiento de lenguaje natural. Este modelo fue presentado por primera vez en el artículo [10].

En el proceso de diseño del ViT, las imágenes son divididas en parches fijos, de manera similar a cómo se segmentan las palabras en un texto. Cada parche es tratado como un "token", que es representado por un vector en un espacio de alta dimensión después de ser linealmente incrustado y complementado con información de posición (Alexey Dosovitskiy, 2020). Esta representación permite que el modelo use mecanismos de autoatención para evaluar las relaciones entre todos los parches de la imagen, capturando contextos y patrones a largo alcance en lugar de limitarse a las profundas estructuras locales que caracterizan a las CNN. Así, el ViT puede identificar características globales y realizar clasificaciones basadas en estas interacciones complejas.

Una de las innovaciones clave que introduce el Vision Transformer es su capacidad para ser entrenado en grandes conjuntos de datos, lo cual es crucial para lograr un rendimiento óptimo. Con el preentrenamiento realizado en extensos registros como ImageNet, el modelo puede ser afinado para tareas específicas, logrando resultados de clasificación que a menudo superan a sus predecesores, las CNN, incluso utilizando menores recursos computacionales. Sin embargo, también presenta desafíos, como su tendencia a necesitar más datos para generalizar adecuadamente, lo que puede dificultar su eficacia en aplicaciones con conjuntos de datos más pequeños. A pesar de estas limitaciones, el ViT se establece como una alternativa prometedora y altamente eficiente en el reconocimiento y la clasificación de imágenes en muchos campos, incluidos los ámbitos médicos y de dermatología.

En resumen, la presentación del Vision Transformer en octubre de 2020 significó un hito en el ámbito del aprendizaje automático, introduciendo una arquitectura que desafía las normas establecidas y ofrece nuevos ciclos de innovación en la clasificación de imágenes. Esta evolución continúa influyendo en la manera en que se aborda el procesamiento de imágenes y se fomenta la investigación en el campo del aprendizaje profundo.

4.3.1. Estructura del Vision Transformer

El principio central del Vision Transformer radica en la utilización de imágenes como secuencias de parches en lugar de tratarla como un todo indivisible. Este enfo-

que, similar al tratamiento de palabras en el procesamiento del lenguaje, implica que una imagen se divide en parches de tamaño fijo, que luego se aplanan y convierten en vectores. Por ejemplo, una imagen de 224x224 píxeles puede ser segmentada en parches de 16x16 píxeles, generando así 196 tokens que serán procesados.

Una vez que los parches son convertidos a vectores, cada representación se alimenta a un modelo Transformer. A diferencia de las CNN, que dependen en gran medida de las convoluciones locales, el ViT utiliza el mecanismo de atención que evalúa la importancia relativa de cada parche en el contexto de los demás, permitiéndole aprender características globales sin perder la estructura espacial de la imagen.

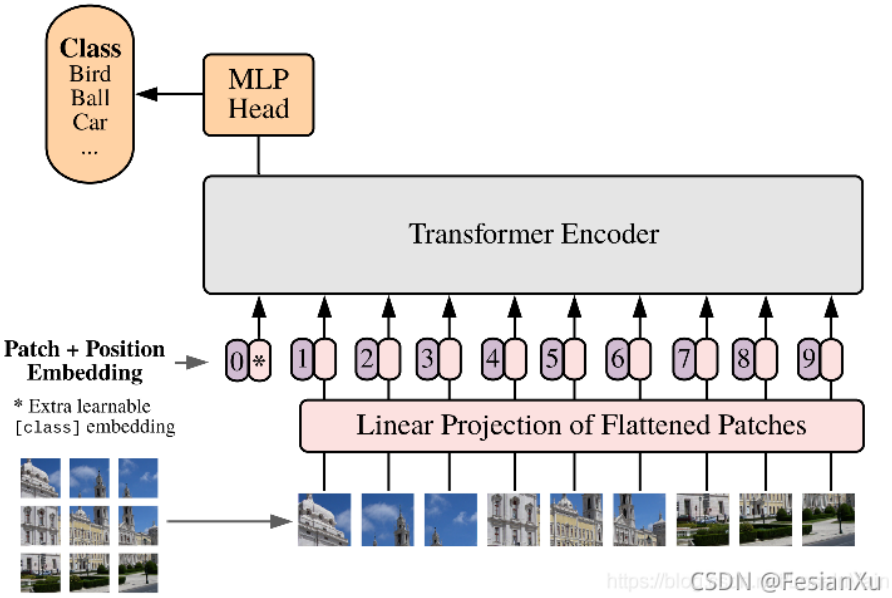


Figura 4.3: Arquitectura del Vision Transformer, presentada en [3]

4.4. Aprendizaje por Transferencia

El *Transfer Learning*, o aprendizaje por transferencia, es una técnica dentro del campo del aprendizaje automático que permite mejorar el rendimiento de un modelo en una tarea específica al aprovechar el conocimiento adquirido en una tarea relacionada previa. Esta metodología es especialmente valiosa en contextos donde el tamaño del conjunto de datos para la nueva tarea es limitado, como ocurre a menudo en la clasificación de imágenes médicas. En lugar de entrenar un modelo desde cero, como tradicionalmente se hace, el Transfer Learning utiliza modelos previamente entrenados que han sido expuestos a grandes volúmenes de datos [8].

Este enfoque es particularmente útil en el ámbito médico, donde las bases de datos pueden ser más pequeñas y difíciles de etiquetar. Por ejemplo, en la clasificación de lesiones cutáneas, modelos preentrenados en grandes conjuntos de datos de imágenes generales pueden ajustarse a la categorización de condiciones específicas de la piel, como melanoma, carcinoma basal y carcinoma escamoso, mediante un proceso de ajuste fino o *finetuning*. Este ajuste permite que el modelo adapte su conocimiento a características relevantes para la nueva tarea, mejorando así su capacidad de generalización y reduciendo el riesgo de sobreajuste.

Además, el Transfer Learning es conocido por disminuir los costos computacionales y el tiempo de entrenamiento, lo que se traduce en una implementación más rápida y económica en aplicaciones prácticas, especialmente en entornos clínicos con recursos limitados. Sin embargo, es crucial tener en cuenta que la calidad de los datos es fundamental; un mal uso del Transfer Learning, como la aplicación de un modelo inadecuado a un conjunto de datos con características muy diferentes, puede resultar en un rendimiento degradado. Por lo tanto, entender adecuadamente las características de las tareas y dominios involucrados en el proceso de transferencia es esencial para el éxito de esta estrategia en la clasificación de imágenes dermatoscópicas.

4.5. Aprendizaje autosupervisado

El aprendizaje autosupervisado es una técnica de aprendizaje automático que utiliza datos no etiquetados para generar señales de supervisión, permitiendo que los modelos aprendan representaciones útiles sin la necesidad de conjuntos de datos etiquetados manualmente [16].

En lugar de depender de etiquetas externas, los modelos autosupervisados crean tareas auxiliares a partir de los propios datos, como predecir una parte de la entrada a partir de otra. Este enfoque ha demostrado ser efectivo en áreas como el procesamiento del lenguaje natural y la visión por computadora [44].

Por ejemplo, en el procesamiento del lenguaje natural, modelos como BERT utilizan el aprendizaje autosupervisado para predecir palabras ocultas en una oración, lo que les permite aprender representaciones lingüísticas profundas sin necesidad de grandes cantidades de datos etiquetados [44].

Este enfoque reduce la dependencia de datos etiquetados manualmente y permite aprovechar grandes volúmenes de datos no etiquetados, facilitando la creación de modelos más robustos y generalizables.

4.6. DINOv2

DINOv2 [32] es una potente técnica de aprendizaje autosupervisado lanzada por Meta AI en abril de 2023, que marca un avance significativo en el campo de la visión por computador. Este modelo fue presentado oficialmente el 17 de abril de 2023, y se ha destacado por su capacidad para entrenar modelos de visión sin requerir grandes cantidades de datos etiquetados, lo cual es especialmente valioso en contextos donde la anotación de datos resulta costosa y laboriosa.

La principal innovación de DINOv2 radica en su capacidad para obtener representaciones visuales robustas y generalizables a partir de grandes conjuntos de datos. Este modelo se entrena en un corpus diverso que consta de 142 millones de imágenes y tiene como objetivo aprender de forma eficiente las características de estas imágenes mediante un procedimiento de pre-entrenamiento y ajuste fino. A diferencia de otros modelos que requieren suposiciones estructuradas y diseños específicos, DINOv2 se beneficia de su diseño flexible y se puede aplicar a tareas como la segmentación, detección de objetos y reconocimiento de imágenes de manera efectiva.

Una de las características más destacadas de DINOv2 es su habilidad para funcionar sin ajuste fino (*fine-tuning*) en las tareas específicas de segmentación y clasificación, lo que implica que puede implementarse listo para ejecutar en diversos contextos y aplicaciones, desde el análisis médico hasta la clasificación general de imágenes. Esta propiedad hace que DINOv2 sea un modelo muy atractivo para investigaciones en curso y aplicaciones comerciales, ya que reduce el tiempo y los recursos necesarios para poner en marcha soluciones de inteligencia artificial, maximizando la eficacia del proceso de entrenamiento.

Además de su rendimiento sobresaliente en tareas de visión, DINOv2 enfrenta retos relacionados con la adaptación a diferentes dominios y la complejidad de las imágenes, así como la necesidad de interpretar la robustez del modelo en situaciones del mundo real. Sin embargo, gracias a su enfoque auto-supervisado, ha mostrado ser eficiente incluso en contextos donde los datos etiquetados son limitados o difíciles de obtener.

DINOv2 representa un avance significativo en la capacidad de los modelos de visión por computador para aprender y generalizar de manera efectiva en tareas complejas, sin la necesidad de extensos datos anotados. Con su diseño innovador y su versatilidad, este modelo está preparado para jugar un papel crucial en el futuro del aprendizaje automático y la inteligencia artificial aplicada a la visión.

Capítulo 5

Propuesta

El melanoma es el tipo de cáncer de piel más peligroso debido a su alta tasa de mortalidad y capacidad de propagación. Por lo tanto, es fundamental minimizar el riesgo de que un melanoma sea erróneamente clasificado como otro tipo de lesión. Este objetivo requiere que el modelo propuesto alcance una alta sensibilidad para la clase melanoma. Con este propósito, se plantea una estrategia de clasificación dividida en dos fases:

1. **Clasificación binaria (Melanoma/No Melanoma):** En esta primera fase, el modelo determina si la lesión pertenece o no a la categoría de melanoma. Si el resultado es positivo (melanoma), se reporta esta clasificación de inmediato. En caso contrario, la muestra avanza a la siguiente fase.

2. **Clasificación multicategórica (4 clases):** En esta segunda etapa, el modelo clasifica la muestra entre cuatro categorías: melanoma, carcinoma basocelular, carcinoma espinocelular y otras lesiones. Este resultado se proporciona como salida final únicamente si en la primera fase se determinó que no era melanoma.

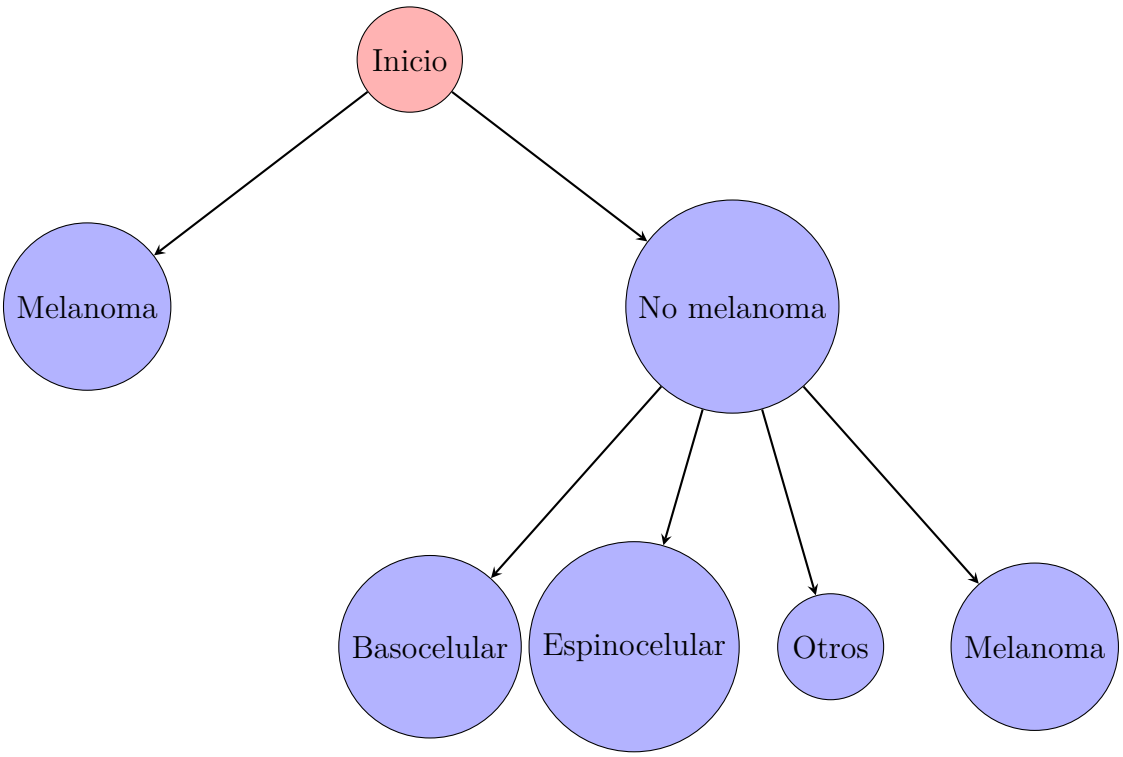


Figura 5.1: Arquitectura Propuesta

5.1. Conjunto de datos utilizados

El entrenamiento de los modelos y su evaluación fue realizado en el conjunto de datos *International Skin Imaging Collaboration (ISIC)*, el cual es un recurso ampliamente utilizado en la investigación dermatológica. Este conjunto contiene un total de **52803 imágenes** correspondientes a **27 clases diferentes** de lesiones cutáneas. Estas imágenes abarcan una amplia variedad de condiciones dermatológicas, lo que permite entrenar modelos capaces de distinguir entre múltiples tipos de enfermedades de la piel, como melanomas, queratosis seborreica, nevus benignos, entre otros. A continuación se detalla la composición del *dataset*:

Tabla 5.1: Cantidad de Imágenes por Clase en el Dataset

Lesión	**Número de Imágenes**
Acrocordón	301
Queratosis actínica	1350
Angiofibroma o pápula fibrosa	4
Angioqueratoma	13
Angioma	71
Tumor de Spitz atípico	5
Carcinoma basocelular	4767
Acantoma de células claras	6
Dermatofibroma	398
Léntigo simple	43
Queratosis liquenoide	311
Melanoma	7226
Metástasis de melanoma	39
Neurofibroma	26
Nevus	32608
Nevus spilus	1
Otros	33
Queratosis benigna pigmentada	1339
Granuloma piógeno	3
Cicatriz	29
Adenoma sebáceo	2
Hiperplasia sebácea	7
Queratosis seborreica	1905
Léntigo solar	558
Carcinoma escamocelular	1301
Lesión vascular	338
Verruga	119

En primer lugar fue separado un subconjunto de imágenes para ser utilizado como conjunto de evaluación, este quedó constituido por:

- 300 imágenes de Melanomas
- 300 imágenes de Carcinomas basocelulares
- 150 imágenes de Carcinomas Espinocelulares
- 400 imágenes pertenecientes a la clase "otros"

Las imágenes de las tres primeras categorías fueron seleccionadas de manera aleatoria del conjunto ISIC descrito anteriormente utilizando un script de Python, esto no se realizó con la clase "otros" debido a la gran cantidad de imágenes de Nevus que poseía, por lo que una evaluación en un subconjunto aleatorio de este conjunto podría conllevar a resultados no veraces, dando resultados superiores en un modelo que haya aprendido a identificar imágenes de Nevus para clasificarlas como *otros*, por lo que fueron tomados imágenes de 100 Nevus aleatorios y 300 imágenes al azar del resto de las clases.

5.1.1. Aumento de datos

Dado que el número de imágenes es relativamente limitado para entrenar modelos profundos y complejos, y con el objetivo de **mejorar la generalización** del modelo y prevenir el **sobreajuste** (overfitting)[11, pág. 12], se implementaron diversas técnicas de . Estas técnicas consisten en aplicar transformaciones geométricas y modificaciones sobre las imágenes originales, lo que permite generar nuevas imágenes a partir de las originales sin perder la relevancia del diagnóstico. De esta forma, se incrementa la diversidad del conjunto de datos sin necesidad de adquirir imágenes adicionales. A continuación, se detallan las técnicas empleadas: Dado que el número de imágenes es relativamente limitado para entrenar modelos profundos y complejos, y con el objetivo de **mejorar la generalización** del modelo y prevenir el **sobreajuste** (overfitting)[11, pág. 12], se implementaron diversas técnicas de aumento de datos (*data augmentation*). Estas técnicas consisten en aplicar transformaciones geométricas y modificaciones sobre las imágenes originales, lo que permite generar nuevas imágenes a partir de las originales sin perder la relevancia del diagnóstico. De esta forma, se incrementa la diversidad del conjunto de datos sin necesidad de adquirir imágenes adicionales. A continuación, se detallan las técnicas empleadas:

1. **Rotaciones con Diferencia de 45 Grados**: Una de las técnicas utilizadas fue la *rotación determinista* de las imágenes en múltiplos de **45 grados**. Cada imagen del conjunto de datos fue rotada en incrementos de 45 grados, es decir,

las imágenes se sometieron a rotaciones de 45°, 90°, 135°, 180°, 225°, 270°, 315° y 360°. Este enfoque no solo amplía la cantidad de imágenes, sino que también ayuda a que el modelo aprenda a reconocer las lesiones cutáneas desde diferentes ángulos, mejorando la robustez del modelo frente a posibles variaciones en la orientación de las lesiones cuando se encuentren en entornos clínicos reales.

2. **Volteos**: Otra técnica utilizada fue la *reflexión o volteo* de las imágenes en torno a los ejes **X** y **Y**. Al reflejar la imagen sobre el eje horizontal (X) y el eje vertical (Y), se simulan diferentes posiciones que podrían ser observadas en un escenario clínico. Por ejemplo, una lesión en un lado del cuerpo puede ser reflejada en la imagen para que el modelo sea capaz de identificar lesiones tanto en el lado izquierdo como en el derecho de una persona, sin importar la orientación original de la imagen. Esta técnica es especialmente útil en imágenes médicas, donde las lesiones pueden aparecer en cualquier área del cuerpo.

3. **Traslaciones Aleatorias**: Se aplicaron *traslaciones aleatorias* a cada imagen, desplazando la imagen en direcciones horizontales y verticales dentro de un rango de **±15 píxeles**. Esto permite que el modelo sea capaz de aprender patrones relevantes sin depender de la posición exacta de las lesiones en la imagen. De esta forma, se simulan variaciones de localización dentro del campo de visión, lo que aumenta la capacidad del modelo para generalizar y reconocer lesiones que pueden estar ligeramente descentradas o fuera de lugar debido a variaciones en el proceso de captura de la imagen.

4. **Rotaciones Aleatorias**: Además de las rotaciones deterministas, se implementaron *rotaciones aleatorias* para cada imagen. Se realizaron un total de **8 rotaciones aleatorias** con ángulos variables entre 0° y 359°. Este tipo de variabilidad en los ángulos de rotación permite que el modelo se entrene en una mayor diversidad de vistas posibles, mejorando su capacidad para identificar lesiones independientemente de la orientación precisa de la imagen. De este modo, el modelo se vuelve más flexible y capaz de manejar variaciones de ángulos que podrían ocurrir debido a distintas posiciones del paciente al momento de capturar la imagen.

El resultado de la aplicación de estas técnicas de aumento de datos sobre el conjunto original de **50,804 imágenes** fue un notable **incremento en el tamaño del conjunto de datos**, alcanzando un total de **1,473,867 imágenes**. Este proceso requiere un alto poder de cómputo, tomando cada una de las técnicas utilizadas mas de 10 días. Para acelerar el proceso, se utilizó programación paralela, utilizando los 16 núcleos de un microprocesasdor Intel Core i7 de 13ra generación, con lo cual las técnicas realizadas tomaban poco más de 1 día. Este aumento significativo en la cantidad

de datos no solo refuerza la capacidad del modelo para aprender patrones más robustos, sino que también contribuye a mejorar su desempeño en tareas de clasificación y generalización.

La implementación de estas estrategias de aumento de datos tiene como principal objetivo **evitar el sobreajuste** al aumentar la diversidad del conjunto de datos y permitir que el modelo aprenda a reconocer lesiones de la piel bajo diferentes condiciones de rotación, traslación y orientación. Esto facilita un mejor rendimiento en el análisis de imágenes de piel, permitiendo que el modelo sea más eficaz al enfrentar nuevos datos no vistos durante el entrenamiento. Debido a limitaciones de nuestros recursos computacionales, no se pudo utilizar este gran volumen de datos para el entrenamiento de los modelos de este trabajo, tema del cual se trata en la siguiente sección. Todo el conjunto generado fue guardado para ser utilizado en futuros trabajos.

5.2. División de clases y trabajo con el conjunto de clases

Se comenzó realizando dos particiones distintas del *Dataset*, una para la clasificación binaria y otra para la clasificación en 4 clases. A continuación se describen ambos conjuntos.

En el caso de la clasificación binaria, de las 52,804 imágenes del conjunto de datos, solo 7,226 corresponden a la clase melanoma, por lo que se tomaron solamente 11713 imágenes de no melanoma, de las cuales 761 fueron tomadas para evaluación y las 10,952 restantes para el entrenamiento. De los 7226 melanomas, 300 pertenecen al conjunto de evaluación para la clasificación en 4 clases y 675 fueron separados para la evaluación, quedando 6251 disponibles para el entrenamiento, por lo que se tomaron 3,756 imágenes aumentadas, dando un total de 10007 imágenes usadas para el entrenamiento.

En el caso de la clasificación en cuatro clases, el conjunto de datos original estaba compuesto por 52,804 imágenes, distribuidas de la siguiente manera:

- **Carcinoma de Células Basales:** 4,557 imágenes, de las cuales se reservaron 300 para evaluación y las restantes para entrenamiento.
- **Melanoma:** 7,226 imágenes, de las cuales 300 se destinaron a evaluación.
- **Carcinoma de Células Escamosas:** 1,301 imágenes, con 150 reservadas para evaluación.
- **Otros:** Las imágenes restantes correspondieron a esta clase, con 400 imágenes destinadas a evaluación.

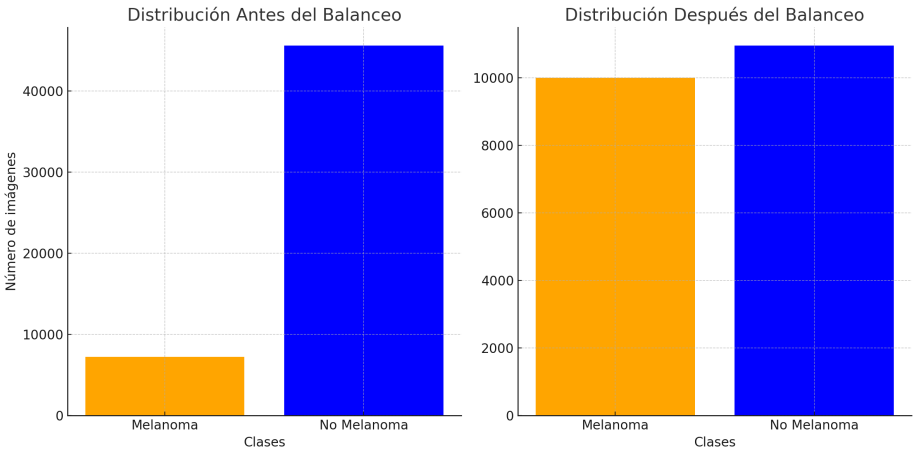


Figura 5.2: Balanceo de clases del Dataset Binario

Para abordar el desequilibrio en las clases, se tomaron imágenes aumentadas para incrementar la cantidad de imágenes en las clases minoritarias. El conjunto final de entrenamiento quedó distribuido de la siguiente manera:

- **Carcinoma de Células Basales:** 11,433 imágenes.
- **Melanoma:** 11,486 imágenes.
- **Carcinoma de Células Escamosas:** 11,010 imágenes.
- **Otros:** 11,459 imágenes.

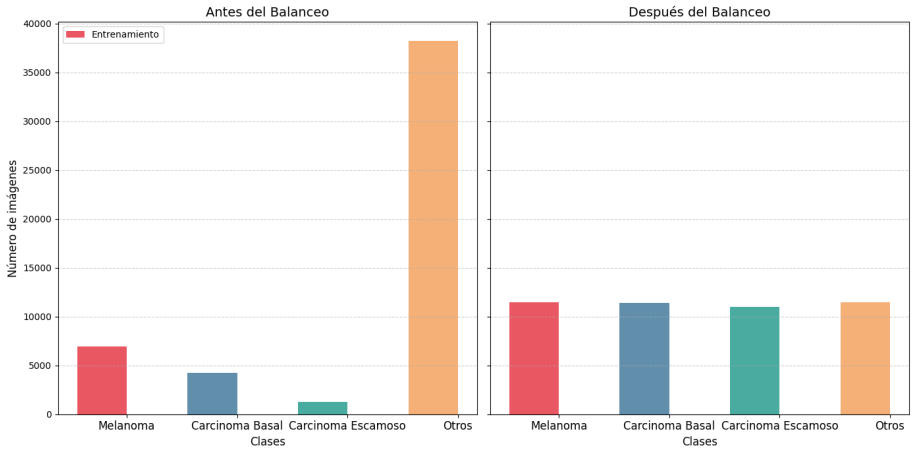


Figura 5.3: Balanceo de clases del Dataset de 4 clases

5.3. Modelos utilizados

Fueron entrenados los siguientes modelos:

- Se le ha realizado Fine Tuning a un Vision Transformer preentrenado.
- Se ha utilizado DINO-V2 para realizar extracción automática de características, estos vectores de características fueron posteriormente utilizados para entrenar modelos de SVM, XGBoost [4] y LightGBM como clasificadores.

Todos los modelos fueron entrenados tanto para clasificación binaria como para clasificación en 4 clases, utilizando como conjuntos de entrenamiento y prueba los descritos en la sección anterior. Además, se realizaron experimentos con los conjuntos de entrenamiento y prueba utilizados en la tesis [12] con el objetivo de comparar la robustez de las características extraídas de forma automática con los modelos basados en *Transformers* con las extraídas según criterios médicos en esta tesis.

5.4. Métricas analizadas

Para evaluar la eficacia de los métodos propuestos, se analizaron un conjunto de métricas que permiten obtener una visión más detallada del comportamiento de los algoritmos en los diferentes conjuntos de evaluación. Estas métricas son fundamentales para evaluar la calidad de las predicciones de un modelo, considerando tanto los aciertos como los errores cometidos durante la clasificación.

En primer lugar, es necesario definir los términos fundamentales utilizados en la evaluación de los algoritmos. Los valores que se obtienen al comparar las predicciones del modelo con las etiquetas verdaderas se organizan en una matriz de confusión, cuyos elementos clave son:

- **Verdaderos positivos (TP)**: El número de instancias correctamente clasificadas como positivas.
- **Falsos negativos (FN)**: El número de instancias que son positivas pero clasificadas incorrectamente como negativas.
- **Verdaderos negativos (TN)**: El número de instancias correctamente clasificadas como negativas.
- **Falsos positivos (FP)**: El número de instancias que son negativas pero clasificadas incorrectamente como positivas.

A partir de estos valores, se definen las métricas que se utilizan para evaluar el rendimiento del modelo:

- **Exactitud (Accuracy)**: Esta métrica representa la proporción de predicciones correctas, tanto positivas como negativas, en relación con el total de predicciones. Se calcula como:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

  La exactitud es útil cuando las clases están balanceadas, pero puede no ser confiable en situaciones de desbalanceo de clases.

- **Precisión (Precision)**: La precisión mide la proporción de predicciones positivas que son realmente correctas. Es útil para evaluar modelos en los que los falsos positivos tienen un alto costo. Se calcula como:

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Sensibilidad (Sensitivity)**: También conocida como recobrado (Recall) mide la proporción de instancias positivas que el modelo ha identificado correctamente. Es particularmente importante cuando el costo de los falsos negativos es alto. Se calcula como:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **Especificidad (Specificity)**: La especificidad mide la proporción de instancias negativas que el modelo ha identificado correctamente. Es relevante cuando se desea minimizar los falsos positivos, y se calcula como:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

- **Medida F1 (F1-Score)**: La medida F1 es la media armónica entre la precisión y el recall, y proporciona una métrica única que balancea tanto los falsos positivos como los falsos negativos. Es útil cuando se necesita un balance entre precisión y recall, especialmente en problemas con clases desbalanceadas. Se calcula como:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Cada una de estas métricas ofrece una perspectiva diferente del rendimiento del modelo, y la elección de las métricas a analizar depende de las características específicas del problema y de los objetivos del análisis.

5.4.1. Matriz de confusión

La matriz de confusión es una herramienta fundamental para evaluar el rendimiento de un modelo de clasificación, proporcionando una representación detallada de las predicciones realizadas por el modelo en comparación con las etiquetas reales. Esta matriz permite identificar no solo los aciertos, sino también los tipos de errores cometidos, lo que ofrece una visión más completa del comportamiento del algoritmo.

En su forma general, la matriz de confusión es una tabla cuadrada donde las filas representan las instancias de las clases reales (verdaderas), y las columnas representan las instancias de las clases predichas por el modelo. Los valores en la matriz indican cuántas veces se ha producido una predicción específica para cada combinación de clase real y clase predicha.

Para una clasificación binaria, la matriz de confusión tiene una estructura 2x2, pero en el caso de una clasificación multiclase (por ejemplo, con 4 clases), la matriz será de mayor dimensión. Cada celda de la matriz de confusión indica la cantidad de veces que una clase particular ha sido predicha como cada una de las clases, siendo los valores de la diagonal la cantidad de aciertos del modelo.

Por ejemplo, en una clasificación multiclase de 4 clases, la matriz de confusión tendrá la siguiente forma:

$$\begin{bmatrix} \mathrm{TP}_1 & \mathrm{FP}_{1,2} & \mathrm{FP}_{1,3} & \mathrm{FP}_{1,4} \\ \mathrm{FP}_{2,1} & \mathrm{TP}_2 & \mathrm{FP}_{2,3} & \mathrm{FP}_{2,4} \\ \mathrm{FP}_{3,1} & \mathrm{FP}_{3,2} & \mathrm{TP}_3 & \mathrm{FP}_{3,4} \\ \mathrm{FP}_{4,1} & \mathrm{FP}_{4,2} & \mathrm{FP}_{4,3} & \mathrm{TP}_4 \end{bmatrix}$$

Donde:

- **TP** (*True Positives*): Son los valores correctos, es decir, las instancias correctamente clasificadas en la clase correspondiente.
- **FP** (*False Positives*): Son las instancias que han sido clasificadas incorrectamente en una clase diferente a la real.

Cada fila de la matriz muestra cuántas veces se ha clasificado correctamente o incorrectamente una instancia de cada clase. Las celdas fuera de la diagonal principal representan los errores de clasificación, donde el modelo ha asignado incorrectamente una instancia de una clase a otra.

La matriz de confusión es extremadamente útil, ya que permite observar con claridad los patrones de los errores, como la tendencia del modelo a confundir ciertas clases. Además, a partir de ella se pueden derivar métricas clave, como precisión, sensibilidad, exactitud y medida F1, que proporcionan una visión más precisa del rendimiento del modelo en las diferentes clases.

La interpretación de la matriz de confusión es crucial para identificar áreas de mejora en los modelos de clasificación, especialmente en escenarios de clasificación multiclase, donde un modelo puede tener un desempeño dispar entre diferentes clases. En estos casos, la matriz ofrece información clave para ajustar el modelo y optimizar su rendimiento.

Capítulo 6

Detalles de Implementación y Experimentos

En este capítulo se muestran los parámetros utilizados y los resultados obtenidos por los modelos empleados. Se hacen comparaciones entre estos. Además, se realizan entrenamientos y evaluaciones en el conjunto de datos utilizado en la tesis de Plá. [12].

6.1. Detalles técnicos

Para toda la experimentación realizada, fue utilizada una computadora equipada con un procesador Intel Core i7 de 13^{ra} generación, 16 GB de memoria RAM y una tarjeta gráfica NVIDIA GeForce GTX 4070 con 8 GB de memoria dedicada. Se utilizó el lenguaje de programación Python 3.11.7. Se empleó PyTorch como entorno de desarrollo para el trabajo con los modelos de aprendizaje automático.

6.2. Clasificación binaria

Para abordar la tarea de clasificación binaria melanoma/no melanoma, se llevaron a cabo los siguientes experimentos:

- **Vision Transformer**:Se realizó un proceso de *fine-tuning* utilizando el modelo **Vision Transformer (ViT) Large Patch16 224 preentrenado en ImageNet-21k**, una arquitectura de transformador desarrollada por Google para tareas de reconocimiento de imágenes. Este modelo cuenta con 24 capas de *Transformer* y con aproximadamente 304 millones de parámetros. Está diseñado para trabajar con imágenes de 224x224 píxeles, las cuales se dividen en parches de 16x16 píxeles.

- Se utilizó el modelo DINOv2 en 4 configuraciones:
  - **DINOv2 Small:** Este modelo cuenta con 22 millones de parámetros, está compuesto por 12 bloques *Transformer*, 6 cabezas de atención y genera embeddings de 384 dimensiones.
  - **DINOv2 Base:** Con 86 millones de parámetros, este modelo incluye 12 bloques *Transformer*, 12 cabezas de atención y genera embeddings de 768 dimensiones.
  - **DINOv2 Large:** Este modelo cuenta con 304 millones de parámetros, está compuesto por 24 bloques *Transformer*, 16 cabezas de atención y genera embeddings de 1024 dimensiones.
  - **DINOv2 Giant:** Este modelo cuenta con 1,000 millones de parámetros, está compuesto por 40 bloques *Transformer*, 16 cabezas de atención y genera embeddings de 1536 dimensiones.

  Todas las configuraciones se emplearon como extractores automáticos de características, generando vectores de características a partir del conjunto de datos binario. Posteriormente, estos vectores fueron utilizados para entrenar tres modelos clásicos de aprendizaje automático: una máquina de soporte vectorial (SVM), un modelo basado en *Gradient Boosting* utilizando XGBoost, y otro utilizando LightGBM. En esta sección se muestran los resultados alcanzados y se detallan las métricas propuestas con

6.2.1. Fine Tuning a Vision Transformer

En un primer experimento, el Vision Transformer fue entrenado con el conjunto de entrenamiento concebido para la clasificación binaria descrito en el capítulo anterior y evaluado con el conjunto de evaluación correspondiente. Los resultados obtenidos se pueden apreciar en la siguiente matriz de confusión:

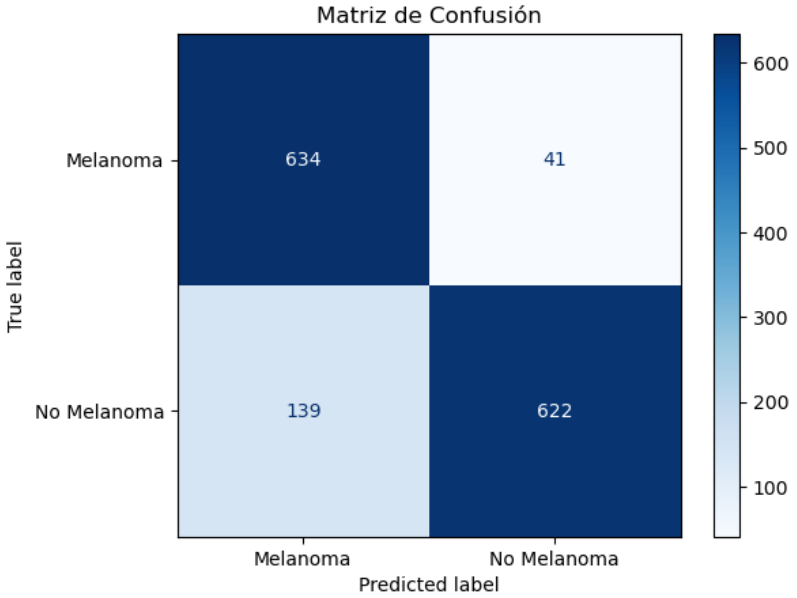


Figura 6.1: Matriz de confusión obtenida en el conjunto de evaluación.

La matriz de confusión refleja el rendimiento del modelo al clasificar las imágenes entre melanomas y no melanomas. A partir de esta, se calcularon las métricas de evaluación clave:

- **Exactitud (Accuracy):** 87,44 %
- **Precisión (Precision):** 82,03 %
- **Sensibilidad (Recall):** 93,93 %
- **Especificidad (Specificity):** 81,73 %
- **Medida F1 (F1-Score):** 87,52 %

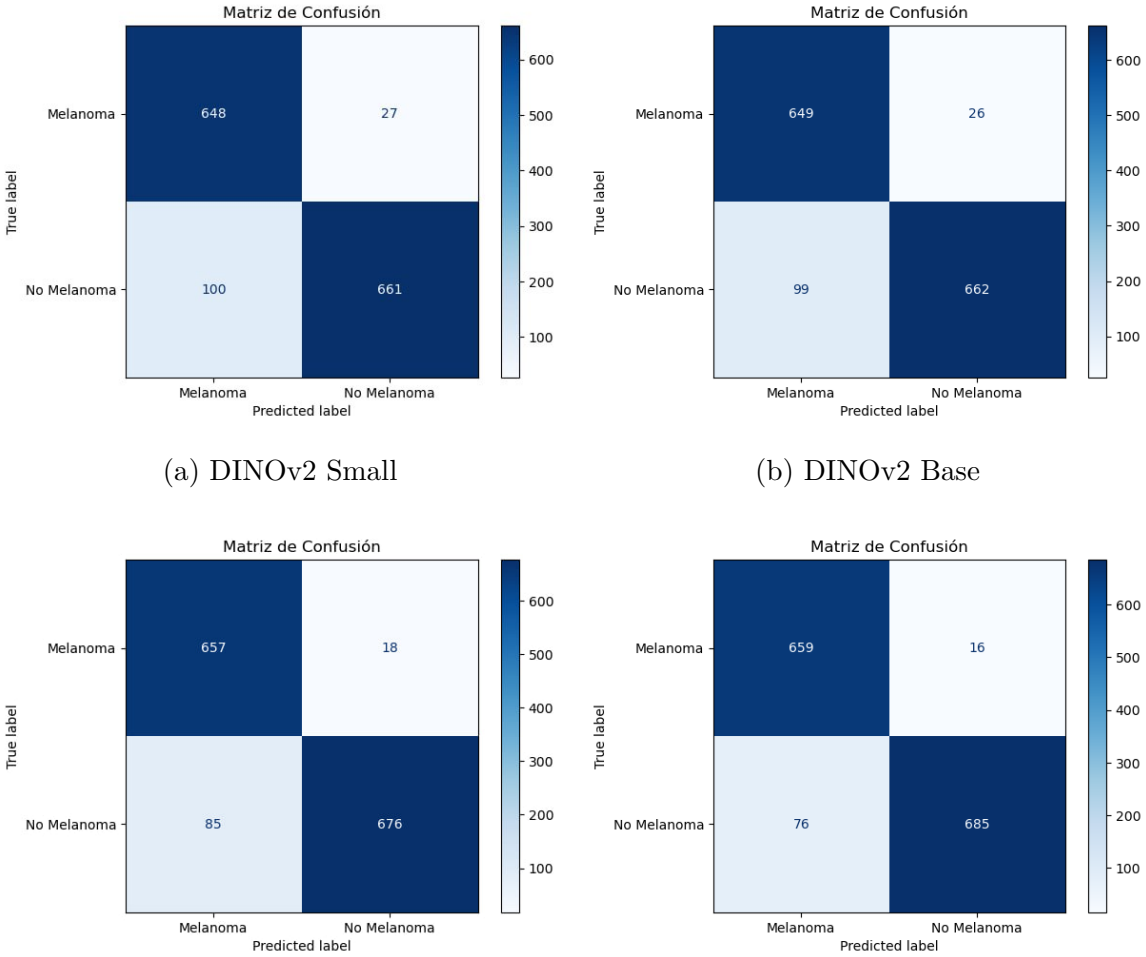
Estas métricas indican un desempeño sólido en términos de clasificación de melanomas, especialmente en sensibilidad, lo que demuestra que el modelo tiene un alto porcentaje de detección de casos positivos. Sin embargo, la precisión y la especificidad sugieren la presencia de falsos positivos, lo que podría derivar en clasificaciones incorrectas de imágenes no melanomas como melanomas.

Posteriormente fueron utilizadas las distintas versiones de *DINOv2* para extraer características de las imágenes, estos vectores de características fueron posteriormente utilizados para entrenar una *Support Vector Machine*. Para la selección de hiperparámetros, fueron tomados los vectores generados por la versión *small* del modelo y se utilizó la técnica de *GridSearch* para encontrar la combinación de hiperparámetros donde la CrossValidation diera los mejores resultados sobre el conjunto destinado a entrenamiento, la combinación hallada fue la siguiente:

Tabla 6.1: Parámetros del modelo SVM

Hiperparámetro	**Valor**
Kernel	rbf (Base Radial)
Gamma	0.001
C	10

Finalmente el modelo fue entrenado con todo el conjunto de entrenamiento. Debido a los recursos de cómputo disponibles y la cantidad de entrenamientos requeridos para emplear *CrossValidation*, no fue posible realizar esta búsqueda exhaustiva de hiperparámetros en los modelos de mayor tamaño, por los que fueron utilizados los encontrados para el modelo *small*. Las siguientes matrices de confusión muestran los resultados obtenidos:



(a) DINOv2 Small

(b) DINOv2 Base

(c) DINOv2 Large

(d) DINOv2 Giant

Figura 6.2: Matrices de confusión obtenidas para la clasificación binaria (melanoma/no melanoma) utilizando una SVM entrenada con los embeddings generados por diferentes versiones de DINOv2.

De las cuales se pueden calcular las siguientes métricas:

Tabla 6.2: Métricas de rendimiento de los modelos DINOv2 para clasificación binaria utilizando una SVM

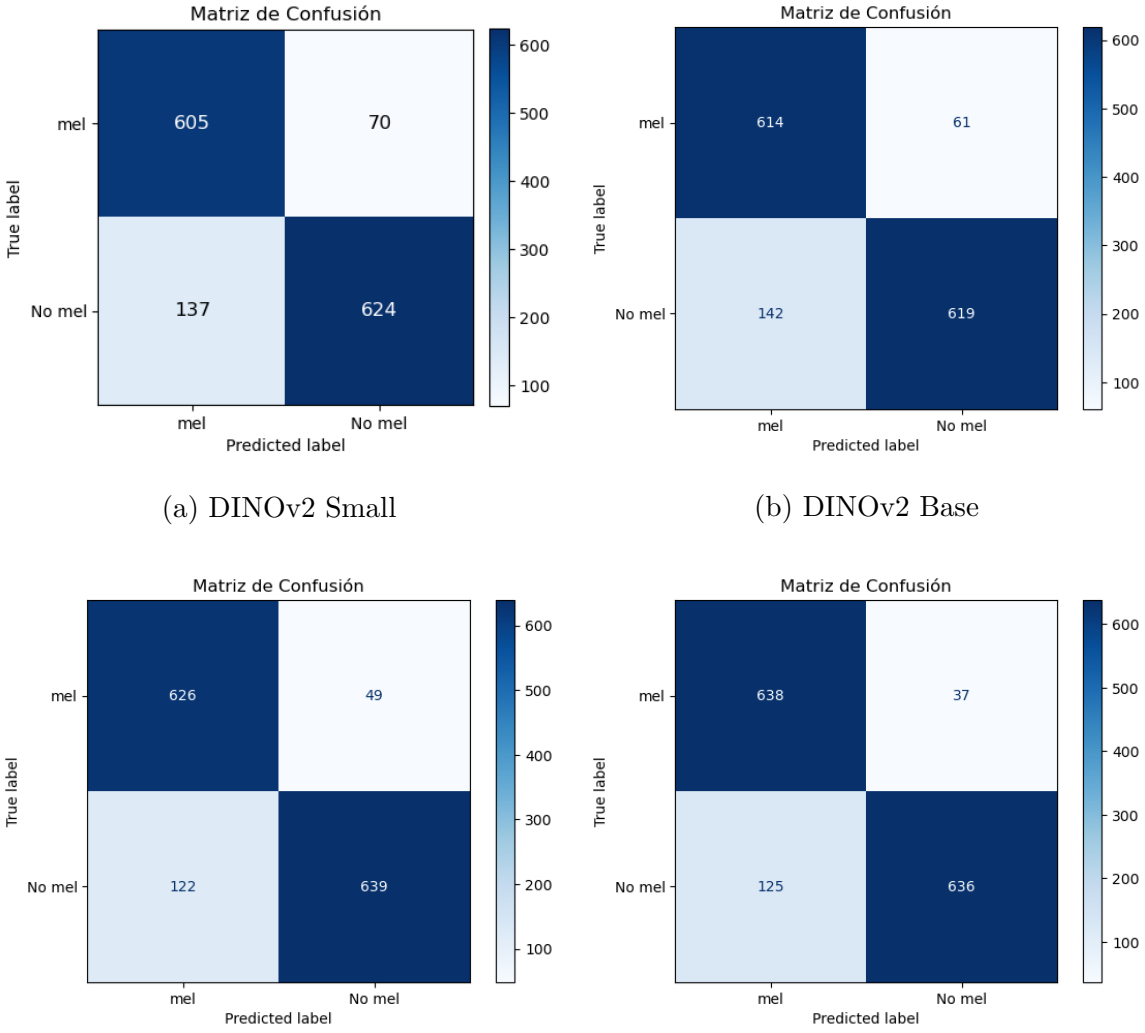
Modelo de DINOv2	**Accuracy**	**Precision**	**Recall**	**Specificity**	**F1-Score**
DINOv2 Small	91.1%	86.6%	96%	86.8%	91.1%
DINOv2 Base	91.3%	86.7%	96.1%	87%	91.2%
DINOv2 Large	92.8%	88.5%	97.3%	88.8%	92.7%
DINOv2 Giant	93.6%	89.6%	97.6%	90%	93.4%

Posteriormente fue utilizado un modelo de XGBoost como clasificador, entrenándolo con los vectores generados por las diferentes versiones de *DINOv2*. Para la selección de hiperparámetros se utilizó Optuna para realizar una búsqueda aleatoria para elegir la combinación con mejores resultados en una *Cross-Validation* de 5 pliegues. A continuación se muestran la combinación obtenida:

Tabla 6.3: Hiperparámetros utilizados para el modelo XGBClassifier

Hyperparameter	**Value**
n_estimators	495
max_depth	9
learning_rate	0.0751
subsample	0.8104
colsample_bytree	0.7897
gamma	0.2933
reg_alpha	8.9689
reg_lambda	6.4685
min_child_weight	8

A continuación se muestran las matrices de confusión con los resultados de cada uno de los experimentos:



(a) DINOv2 Small

(b) DINOv2 Base

(c) DINOv2 Large

(d) DINOv2 Giant

Figura 6.3: Matrices de confusión obtenidas para la clasificación binaria (melanoma/no melanoma) utilizando un modelo XGBoost entrenada con los embeddings generados por diferentes versiones de DINOv2.

De estos resultados se pueden calcular las siguientes métricas:

Tabla 6.4: Métricas de rendimiento del XGBoost para clasificación binaria utilizando los vectores extraidos con diferentes modelos de DINOv2

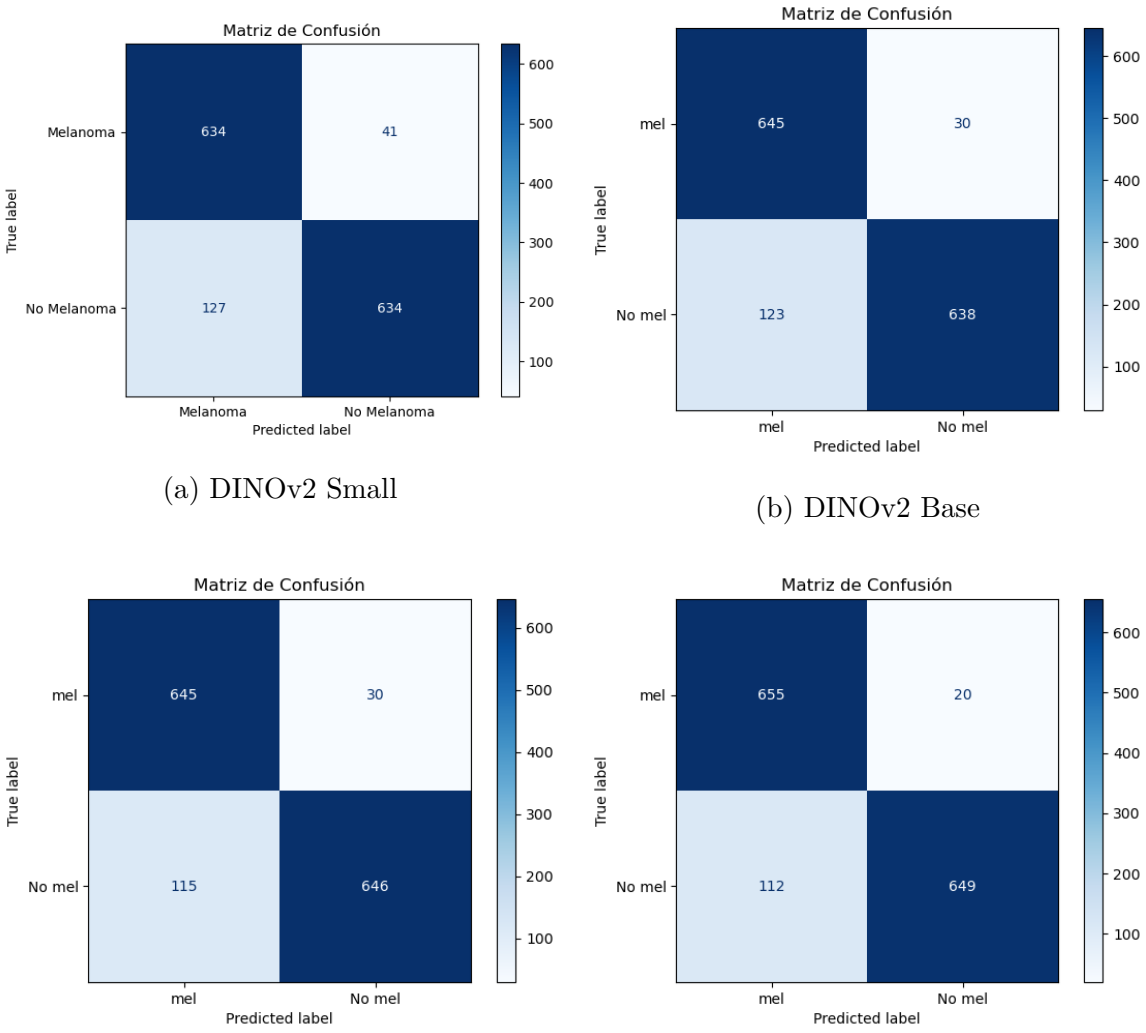
Modelo de DINOv2	**Accuracy**	**Precision**	**Recall**	**Specificity**	**F1-Score**
DINOv2 Small	85.6%	81.5%	89.6%	82%	85,4%
DINOv2 Base	85.8%	81.2%	90.9%	81.3%	85.8%
DINOv2 Large	88.3%	81.5%	92.7%	83.9%	86.7%
DINOv2 Giant	88.7%	83.6%	94.5%	83.5%	88,7%

En un próximo experimento se utilizó *LightGBM* como clasificador, dándole también como entrada los *embedings* generados por los diferentes modelos de *DINOv2*. Se utilizó una Grid-Search con una Cross-Validation de 5 pliegues para la selección de hiperparámetros, obteniéndose los mejores resultado con los siguintes hiperparámetros:

Tabla 6.5: Hiperparámetros del modelo LightGBM

Hyperparameter	**Value**
boosting_type	`gbdt`
num_leaves	`100`
learning_rate	`0.2`
n_estimators	`500`
max_depth	`10`
min_child_samples	`30`

A continuación se muestran las matrices de confusión con los resultados obtenidos:



(a) DINOv2 Small

(b) DINOv2 Base

(c) DINOv2 Large

(d) DINOv2 Giant

Figura 6.4: Matrices de confusión obtenidas para la clasificación binaria utilizando el modelo LightGBM entrenado con los *embeddings* generados por diferentes versiones de DINOv2

Al analizar estos resultados se pueden obtener las siguientes métricas:

Tabla 6.6: Métricas de rendimiento de los modelos DINOv2 para clasificación binaria utilizando una SVM

Modelo de DINOv2	**Accuracy**	**Precision**	**Recall**	**Specificity**	**F1-Score**
DINOv2 Small	88.3%	83.3%	93.9%	83.3%	88.2%
DINOv2 Base	89.3%	83.9%	95.5%	83.8%	89,3%
DINOv2 Large	89.9%	84.8%	95.5%	84.8%	89,8%
DINOv2 Giant	90.8%	85.3%	97%	85.2%	90.8%

Al analizar los resultados obtenidos se observa que los modelos de LightGBM entrenado muestran un rendimiento superior a los modelos XGBoost utilizando los vectores generados por los mismos modelos de *DINOv2* pero inferior al obtenido al emplear *SVMs*. Al igual que en los experimentos anteriores, los resultados mejoran conforme aumenta el tamaño del modelo DINOv2.

6.3. Clasificación en 4 clases

En esta sección se detallan los resultados obtenidos en la clasificación en 4 clases. Primeramente se analizan los resultados obtenidos al hacer Fine-Tuning al Vision Transformer, luego se analizan los resultados obtenidos al emplear modelos que utilizan los embedings extraídos por *DINOv2*.

Utilización de un Vision Transformer preentrenado

Para la clasificación en 4 clases se comenzó haciendo Fine-Tuning al Vision Transformer con el conjunto de entrenamiento de 4 clases descrito en el capitulo anterior, los resultados se detallan en la siguiente matriz de confusión:

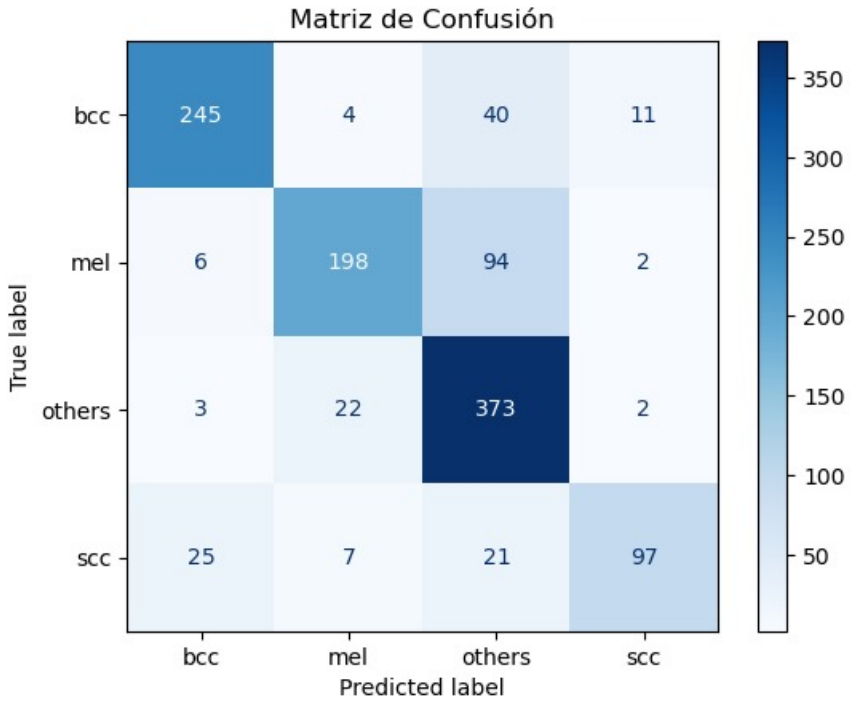


Figura 6.5: Matriz de confusión obtenida en el conjunto de evaluación del *dataset* de 4 clases usando el ViT

Al analizar estas predicciones se obtienen las siguientes métricas:

Tabla 6.7: Resultados de las métricas por clase y promedio

Clase	**Precisión**	**Sensibilidad**	**Medida F1**	**Soporte**
Carcinoma Basocelular	0.88	0.82	0.85	300
Melanoma	0.86	0.66	0.75	300
Otros	0.71	0.93	0.80	400
Carcinoma Espinocelular	0.87	0.65	0.74	150
Promedio Macro	0.83	0.76	0.78	1150
Promedio Ponderado	0.81	0.79	0.79	1150
Exactitud		0.79		

Si bien las métricas obtenidas en general fueron deficientes, se puede observar que en la clase otros (clase con peores resultados en [12] y [28]) se obtiene una sensibilidad del 93 %.

Posteriormente fue realizado otro experimento en el cual el Vision Transformer fue entrenado con el mismo conjunto de datos y evaluado con el mismo conjunto de evaluación que el modelo entrenado en la tesis [12] Los resultados se recogen en la siguiente matriz de confusión:

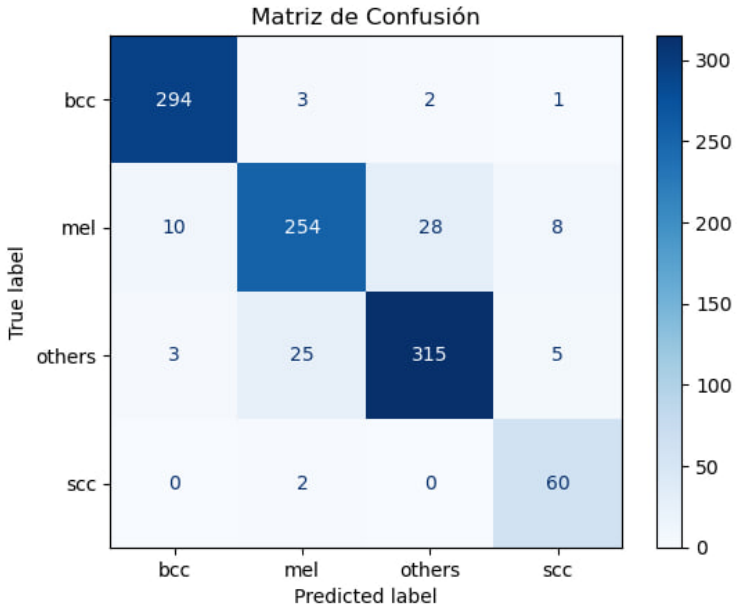


Figura 6.6: Matriz de confusión obtenida entrenando y evaluando el ViT en el *dataset* utilizado en [12]

Al analizar los resultados plasmados en esta matriz son obtenidas las siguientes métricas:

Tabla 6.8: Resultados de las métricas por clase y exactitud

Clase	**Precisión**	**Sensibilidad**	**F1-Score**
Carcinoma Basocelular	0.9591	0.9867	0.9727
Melanoma	0.8958	0.8367	0.8653
Otros	0.9124	0.9000	0.9061
Carcinoma Espinocelular	0.7742	0.9677	0.8613
Exactitud		0.9191	

Estos resultados muestran una mejoría con respecto a los resultados obtenidos en [12], en la cual el mejor resultado obtenido fue del 87,5 % en este conjunto de datos.

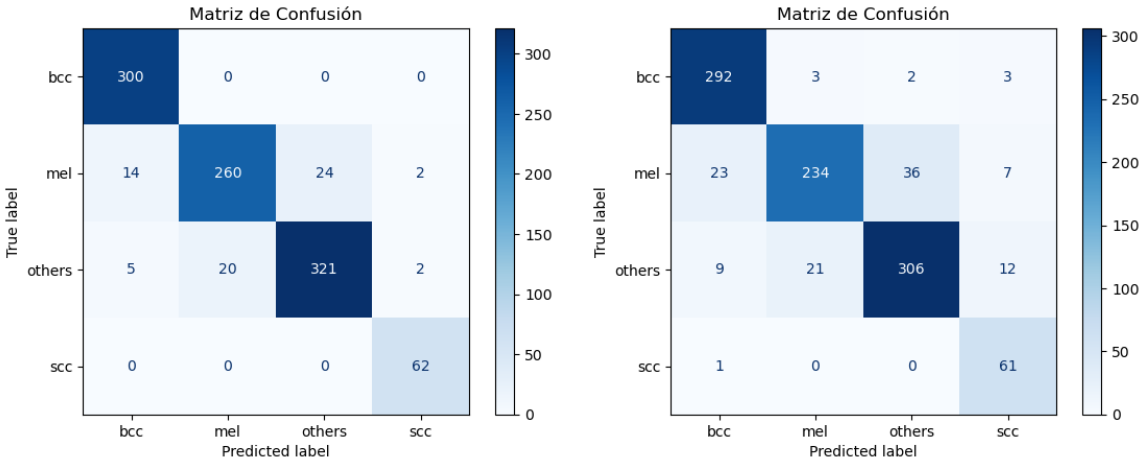
6.3.1. Utilización de modelos basados en los vectores de características extraídos por DINOv2

En esta sección se muestran los resultados de la clasificación en 4 clases utilizando los vectores de características extraídos por DINOv2.

Comparación con los resultados obtenidos en la clasificación utilizando extracción de características según criterios médicos

Con el objetivo de comparar la robustez de las características extraídas automáticamente por DINOv2 con las extraídas según criterios médicos en [12], se utilizó tanto la versión *Base* como *Giant* de *DINOv2* para extraer características del conjunto de datos que se utilizó en el para el entrenamiento del modelo de 4 clases propuesto en la tesis de Plá [12] y se evaluó con el conjunto de evaluación correspondiente.

En primer lugar se utilizó la versión *Giant* de DINOv2 para extraer los vectores de características y fueron entrenados un modelo de *SVM*, *XGBoost* y *LightGBM* con estos vectores. A continuación mostramos los resultados obtenidos en el conjunto de evaluación utilizado en la tesis de Plá [12]:



Matriz de confusión - SVM

Matriz de confusión - XGBoost

Figura 6.7: Matrices de confusión para la clasificación con SVM y XGBoost utilizando los vectores extraídos de DINOv2 Giant

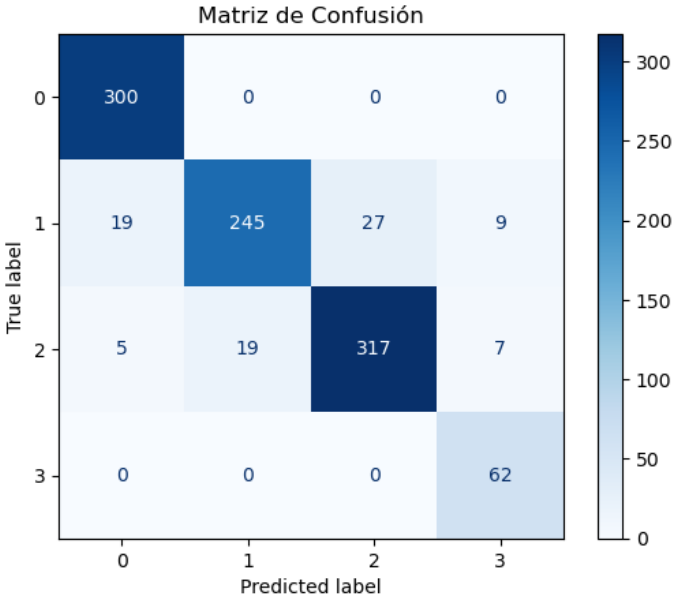


Figura 6.8: Matriz de confusión - LightGBM

Figura 6.9: Matriz de confusión para la clasificación con LightGBM utilizando los vectores extraídos de DINOv2 Giant

Al comparar estos resultados con los obtenidos en la tesis de Plá [12], se puede observar que los resultados logrados utilizando extracción automática de características con *DINOv2* por lo general son superiores a los obtenidos al hacer uso de los vectores de características extraídos según criterios médicos, teniendo en cuenta que el modelo con mejores resultados en [12] fue el modelo de *LightGBM*, el cual obtuvo una exactitud del 87.5 %, mientras que al utilizar los *embeddings* de *DINOv2*, esta fue del 91.5 %. (Nótese que en este caso, el modelo con mejores resultados en el conjunto de evaluación fue el modelo de SVM, con una exactitud del 93.4 % sin embargo, en [12] no fue utilizado *SVM* para la clasificación).

Clasificación en 4 clases en el conjunto de datos propuesto

A continuación se muestran los resultados al utilizar DINOv2 en su versión Giant como extractor de características, para luego entrenar un *Support Vector Clasifier (SVC)* [35] para clasificación. En la siguiente matriz de confusión se muestran los resultados obtenidos en el conjunto de evaluación:

Analizando los resultados de esta matriz se obtienen las siguientes métricas

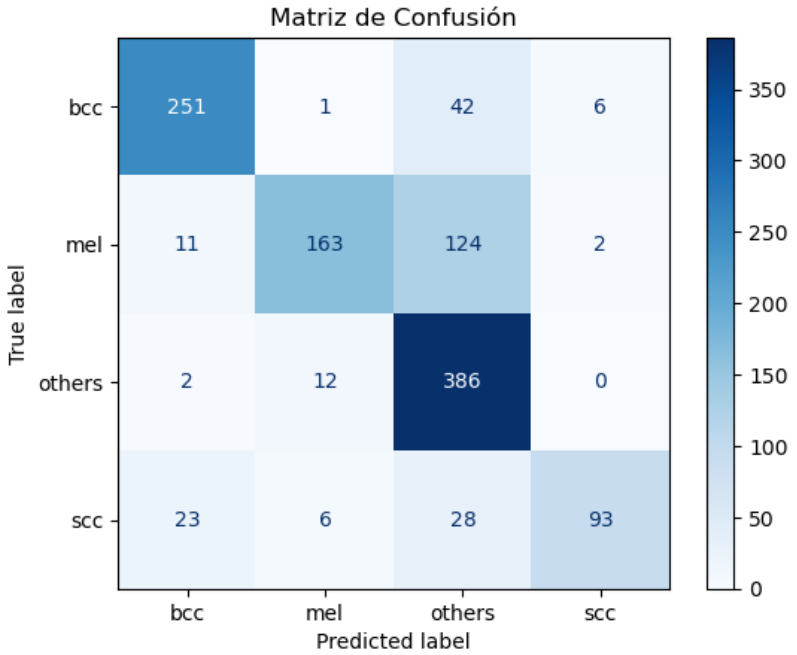


Figura 6.10: Matriz de confusión de los resultados de evaluar la SVM en los vectores de características extraídos por *DINOv2* versión Giant en el *dataset* de 4 clases

Tabla 6.9: Informe de Clasificación y Precisión

Clase	**Precisión**	**Sensibilidad**	**Medida F1**
Carcinoma basocelular	0.87	0.84	0.86
Melanoma	0.90	0.54	0.68
Otros	0.67	0.96	0.79
Carcinoma espinocelular	0.92	0.62	0.74
Exactitud	0.78		

Al analizar estos resultados se puede observar una menor sensibilidad a la clase Melanoma que la lograda por el Vision Transfomer. Por otro lado, se muestra una mayor sensibilidad a las otras clases, lográndose una exactitud ligeramente inferior.

6.4. Resultados de la clasificación en dos fases

En esta sección se muestran los resultados de la experimentación utilizando la clasificación de dos fases con el objetivo de mejorar la sensibilidad al melanoma.

6.4.1. SVM para clasificación binaria y ViT para 4 categorías

En esta sección se muestran los resultados obtenidos al utilizar la SVM entrenada con los vectores de características extraídos con DINOv2 Giant presentada en la sección 6.1.2 para clasificación binaria, para, en caso de esta clasificar como melanoma, clasificar como tal, en caso contrario, pasar a la clasificación en 4 clases y reportar la predicción en esta categoría. para esta fase fue utilizado el Vision Transformer cuyos resultados individuales fueron presentados al comienzo de la sección 4.3. A continuación se muestra la matriz de confusión con los resultados obtenidos en el conjunto de evaluación:

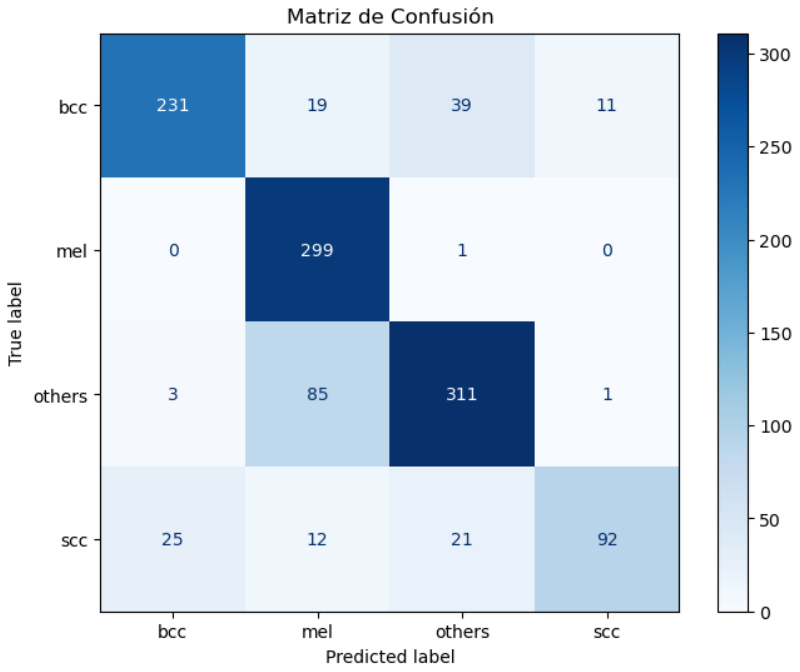


Figura 6.11: Matriz de confusión obtenida en el conjunto de evaluación luego de utilizar el modelo de dos fases

Al analizar los resultados mostrados en esta matriz de obtienen las siguientes métricas:

Tabla 6.10: Resultados de las métricas por clase y exactitud

Clase	**Precisión**	**Sensibilidad**	**Medida F1**
Carcinoma Basocelular	0.8976	0.7897	0.8407
Melanoma	0.7188	0.9967	0.8354
Otros	0.8015	0.7821	0.7917
Carcinoma espinocelular	0.8835	0.6452	0.7450
Exactitud		0.7948	

Como se puede observar, se logró una elevada sensibilidad a la clase Melanoma, por otra parte, esto conllevó a una disminución significativa de las otras métricas, principalmente la precisión del melanoma y la sensibilidad de la clase "Otros". Sin embargo, estos resultados son satisfactorios, ya que es preferible sacrificar otras métricas antes de permitir falsos positivos a una lesión tan mortal como el melanoma.

6.4.2. SVC tanto para clasificación binaria como para 4 categorías

En esta sección se muestran los resultados obtenidos al utilizar dos SVCs entrenados con los vectores de características de DINOv2 Giant, para la clasificación se utilizó la *SVM* presentada en la sección 6.1.2 y para la clasificación multicategórica se utilizó el presentado al final de la sección 6.3.1. En la siguiente matriz de confusión se muestran los resultados obtenidos:

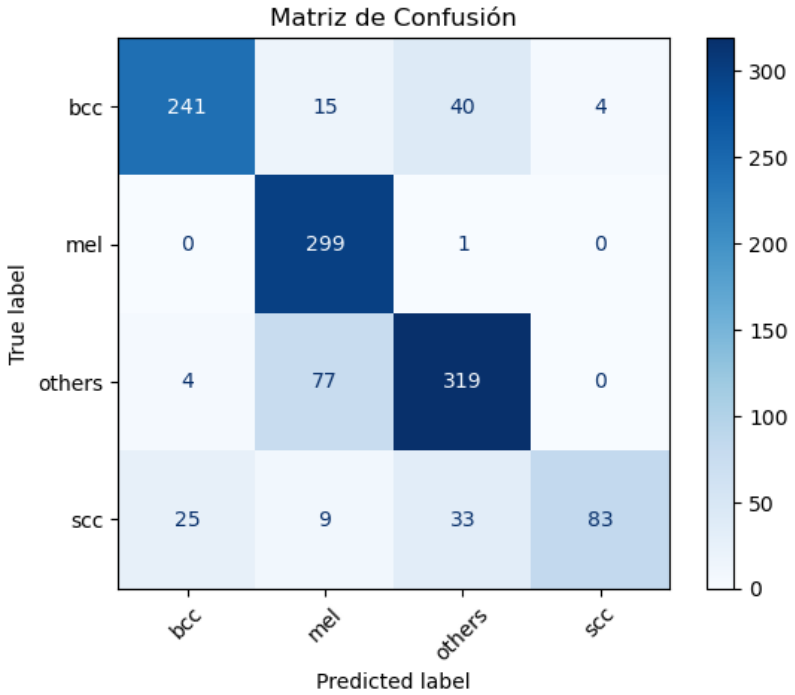


Figura 6.12: Matriz de confusión obtenida en el conjunto de evaluación luego de utilizar el modelo de dos fases.

A partir de estos resultados, se extraen las siguientes métricas:

Tabla 6.11: Resultados de las métricas por clase y exactitud

Clase	**Precisión**	**Sensibilidad**	**Medida F1**
Carcinoma Basocelular	0.893	0.803	0.846
Melanoma	0.748	0.997	0.854
Otros	0.812	0.797	0.805
Carcinoma Espinocelular	0.954	0.553	0.700
Exactitud		0.819	

Al analizarlas, observamos que ocurre una disminución en la sensibilidad al carcinoma espinocelular con respecto al anterior, aumentando su precisión, mientras aumenta ligeramente la sensibilidad al carcinoma basocelular. Todo esto manteniendo la elevada sensibilidad a la clase Melanoma.

Conclusiones

En este trabajo se ha desarrollado un modelo de clasificación de lesiones cutáneas que prioriza la sensibilidad en la detección del melanoma, implementando un enfoque de dos fases utilizando modelos basados en la arquitectura Transformer.

La comparación con [12] muestra que los vectores de características extraídos mediante DINOv2 proporcionan un mejor rendimiento en la clasificación en comparación con aquellos obtenidos a partir de criterios médicos. Asimismo, al contrastar los resultados de este estudio con los obtenidos mediante Fine-Tuning del Vision Transformer, se logró un desempeño superior al reportado en [12] para este conjunto de datos. Estos hallazgos corroboran la primera de las hipótesis planteadas en la introducción de este trabajo.

El conjunto de datos utilizado ha demostrado ser variado, realista y desafiante, ya que contiene un total de 27 clases desbalanceadas. A pesar de aplicar una estrategia extensiva de aumento de datos, alcanzando más de un millón de imágenes, este gran volumen de información no pudo ser empleado en su totalidad para el entrenamiento debido a limitaciones computacionales.

En la fase de clasificación binaria, se obtuvieron mejores resultados al utilizar los vectores de características extraídos por DINOv2 en comparación con el uso del Vision Transformer preentrenado. Al emplear distintas versiones de DINOv2, se observó que el desempeño mejoraba al aumentar la complejidad del modelo, siendo la versión Giant la que generó los vectores de características más robustos. Para la clasificación basada en estos embeddings, el modelo que obtuvo mejores resultados fue la SVM, logrando una sensibilidad del 97 % para la clase Melanoma.

El modelo propuesto alcanzó una alta sensibilidad en la detección del melanoma, lo que refuerza su potencial como herramienta de apoyo en el diagnóstico temprano de esta patología. Sin embargo, una de sus principales deficiencias es la baja sensibilidad en la detección del carcinoma espinocelular, que, si bien generalmente presenta una baja tasa de mortalidad, puede volverse peligroso si no se trata adecuadamente.

Recomendaciones

- Desarollar un modelo híbrido que combine el uso de las características extraídas automáticamente con *DINOv2* con las características extraídas según criterios médicos en la tesis de Plá [12].
- Entrenar la versión "Huge" del modelo de *Vision Transformer* utilizado, que no pudo ser entrenado para en la realización de este trabajo por insuficiencia de recursos computacionales.
- Incorporar uno de los modelos de *Vision Transformer* entrenados al *Ensemble* de Redes Convolucionales de la tesis de Olavarrieta [28] buscando obtener lo mejor de ambos mundos, combinando la capacidad de capturar características locales de las CNNs con la capacidad de capturar características globales de los ViTs, y utilizar un enfoque de dos fases similar al propuesto para garantizar una elevada sensibilidad a la clase Melanoma.
- Recopilar imágenes de lesiones cutáneas cubanas para el entrenamiento de los modelos.

Referencias bibliográficas

[1] Nazia Hameed et al. «An Intelligent Computer-Aided Scheme for Classifying Multiple Skin Lesions». En: *Computers* (2019).

[2] Sumo Analytics. *¿Cuál es la diferencia entre el aprendizaje profundo y el aprendizaje automático?* 2025. URL: https://www.sumoanalytics.ai/post/cual-es-la-diferencia-entre-el-aprendizaje-profundo-y-el-aprendizaje-automatico.

[3] Mathilde Caron et al. «Emerging Properties in Self-Supervised Vision Transformers». En: *arXiv preprint arXiv:2104.14294* (2021). URL: https://arxiv.org/abs/2104.14294.

[4] Tianqi Chen y Carlos Guestrin. «XGBoost: A Scalable Tree Boosting System». En: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. ACM, ago. de 2016, págs. 785-794. DOI: 10.1145/2939672.2939785. URL: http://dx.doi.org/10.1145/2939672.2939785.

[5] Toby Breckon Chris Solomon. *Fundamentals of Digital Image Processing. A Practical Approach with Examples in Matlab*. Wiley-Blackwell, 2011.

[6] Shinjita Das. «Hiperpigmentación». En: *Manuales MSD* (2024). URL: https://www.merckmanuals.com/es-us/hogar/trastornos-de-la-piel/alteraciones-de-la-pigmentaci%C3%B3n/hiperpigmentaci%C3%B3n.

[7] Jia Deng et al. «Imagenet: A large-scale hierarchical image database». En: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, págs. 248-255.

[8] Niklas Donges. *What Is Transfer Learning? A Guide for Deep Learning | Built In*. Ago. de 2024. URL: https://builtin.com/data-science/transfer-learning.

[9] Alexey Dosovitskiy. «An image is worth 16x16 words: Transformers for image recognition at scale». En: *arXiv preprint arXiv:2010.11929* (2020).

[10] Alexey Dosovitskiy et al. «An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale». En: *International Conference on Learning Representations*. 2021. URL: https://arxiv.org/abs/2010.11929.

[11] Richard O. Duda, Peter E. Hart y David G. Stork. *Pattern Classification*. 2nd. New York, NY, USA: Wiley, 2001. ISBN: 978-0-471-05669-0.

[12] Jordan Plá González. *Características en lesiones cutáneas: el uso de la Inteligencia Artificial en la clasificación de imágenes dermatoscópicas*. 2024.

[13] Sepp Hochreiter y Jürgen Schmidhuber. «Long short-term memory». En: *Neural computation* 9.8 (1997), págs. 1735-1780. DOI: 10.1162/neco.1997.9.8.1735.

[14] Hugging Face. *Image*. Accedido: 6 de febrero de 2025. 2023. URL: https://huggingface.co/learn/computer-vision-course/unit1/image_and_imaging/image.

[15] Hugging Face. *The Motivation Behind Creating Artificial Systems Capable of Simulating Human Vision and Cognition*. Accedido: 4 de enero de 2025. 2023. URL: https://huggingface.co/learn/computer-vision-course/unit1/chapter1/motivation.

[16] IBM. *¿ Qué es el aprendizaje autosupervisado?* 2023. URL: https://www.ibm.com/es-es/topics/self-supervised-learning.

[17] IBM. *Support Vector Machine (SVM)*. [Online; accessed 21-January-2025]. 2025. URL: https://www.ibm.com/think/topics/support-vector-machine.

[18] Deep Residual Learning for Image Recognition. *Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun*. 2015. URL: https://arxiv.org/abs/1512.03385v1.

[19] Interactivechaos. *Formulación matemática del perceptrón*. s.f. URL: https://interactivechaos.com/es/manual/tutorial-de-machine-learning/formulacion-matematica-del-perceptron.

[20] Guolin Ke et al. «LightGBM: A Highly Efficient Gradient Boosting Decision Tree». En: *Advances in Neural Information Processing Systems*. Ed. por I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

[21] Somaiya Khan, Athar Shahzad Fazal y Amna Khan. «An Automated Skin Lesions Classification Using Hybrid CNN and Transformer Based Deep Learning Model». En: *Proceedings of the 2023 8th International Conference on Biomedical Imaging, Signal Processing (ICBSP)*. 2023. DOI: 10.1145/3634875.3634879. URL: https://dl.acm.org/doi/10.1145/3634875.3634879.

[22] Alex Krizhevsky, Ilya Sutskever y Geoffrey E. Hinton. «ImageNet Classification with Deep Convolutional Neural Networks». En: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[23] N Vikranth Kumar et al. «Classification of Skin diseases using Image processing and SVM». En: *2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)*. 2019, págs. 1-5. DOI: 10.1109/ViTECoN.2019.8899449.

[24] Very Deep Convolutional Networks for Large-Scale Image Recognition. *Karen Simonyan, Andrew Zisserman*. 2015. URL: https://arxiv.org/abs/1409.1556.

[25] Aprende Machine Learning. *¿Cómo funcionan los Transformers? Español NLP, GPT, BERT*. Accedido el 22 de enero de 2025. n.d. URL: https://www.aprendemachinelearning.com/como-funcionan-los-transformers-espanol-nlp-gpt-bert/.

[26] Ze Liu et al. «Swin Transformer: Hierarchical Vision Transformer using Shifted Windows». En: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), págs. 10012-10022. URL: https://arxiv.org/abs/2103.14030.

[27] Laurens van der Maaten y Geoffrey Hinton. «Visualizing Data using t-SNE». En: *Journal of Machine Learning Research* 9.Nov (2008), págs. 2579-2605. URL: https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf.

[28] Claudia Olavarrieta Martínez. «Ensemble de redes convolucionales para la clasificación de lesiones de cáncer de piel». En: *Facultad de Matemática y Computacion, Universidad de La Habana* (2023).

[29] Quoc V. Le Mingxing Tan. *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. 2020.

[30] Jayanth Mohan et al. «Enhancing Skin Disease Classification Leveraging Transformer-based Deep Learning Architectures and Explainable AI». En: *arXiv preprint arXiv:2407.14757* (2024). Submitted to Computers in Biology and Medicine. URL: https://arxiv.org/abs/2407.14757.

[31] Eduardo Nagore. «Cáncer de piel no melanoma». En: *Revista Médica Clínica Las Condes* 22.6 (2011), págs. 731-736. DOI: 10.1016/S0716-8640(11)70486-2. URL: https://www.elsevier.es/es-revista-revista-medica-clinica-las-condes-202-articulo-cancer-piel-no-melanoma-S0716864011704862.

[32] Maxime Oquab et al. «DINOv2: Learning Robust Visual Features without Supervision». En: *arXiv preprint arXiv:2304.07193* (2023). URL: https://arxiv.org/abs/2304.07193.

[33] Daniel Palacios-Martínez y Ricardo A. Díaz-Alonso. «Dermatoscopia para principiantes (i): características generales». En: *Medicina de Familia. SEMERGEN* 43.3 (2017), págs. 216-221. DOI: 10.1016/j.semerg.2015.11.009.

[34] Ilias Papastratis. *Comparison of Convolutional Neural Networks and Vision Transformers (ViTs)*. Medium. Accessed: 2025-01-05. 2023. URL: https://medium.com/@iliaspapastratis/comparison-of-convolutional-neural-networks-and-vision-transformers-vits-a8fc5486c5be.

[35] F. Pedregosa et al. *Scikit-learn: Machine Learning in Python*. 2011, págs. 2825-2830. URL: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html.

[36] Roche Plus. *La Inteligencia Artificial como herramienta para el diagnóstico del melanoma*. Accessed: 2025-01-21. n.d. URL: https://www.rocheplus.es/innovacion/inteligencia-artificial/melanoma.html.

[37] Stuart Russell y Peter Norvig. *Artificial Intelligence: A Modern Approach*. 3rd. Upper Saddle River, NJ: Pearson, 2016.

[38] Hayit Greenspan S. Kevin Zhou y Dinggang Shen. *Deep Learning for Medical Image Analysis, Second Edition*. MICCAI, 2024.

[39] A. R. Schwartz. «Melanoma maligno y diagnóstico diferencial de lesiones pigmentadas en piel». En: *Revista Médica Clínica Las Condes* (2011).

[40] Shai Shalev-Shwartz y Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[41] Shai Shalev-Shwartz y Ben-Davin. *Understanding Machine Learning, From Theory To Algorithms*. Cambridge University Press, 2014.

[42] Richard Szeliski. *Computer Vision: Algorithms and Applications (2nd ed.)* Springer, 2022.

[43] Satoshi Takahashi et al. «Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review». En: *Journal of Medical Systems* 48.1 (2024), pág. 84. DOI: 10.1007/s10916-024-02105-8. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11393140/.

[44] Telefónica Tech. *¿Qué es el self-supervised learning?* 2023. URL: https://telefonicatech.com/blog/que-es-el-self-supervised-learning.

[45] Hugo Touvron et al. «Training data-efficient image transformers & distillation through attention». En: *arXiv preprint arXiv:2012.12877* (2021). URL: https://arxiv.org/abs/2012.12877.

[46] Ashish Vaswani et al. «Attention is All You Need». En: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017. URL: https://arxiv.org/abs/1706.03762.

[47] Zhi-Hua Zhou. *Ensemble Methods: Foundations and Algorithms*. Chapman y Hall/CRC, 2012.