

# Un enfoque jerárquico y contrastivo para la clasificación de lesiones cutáneas

Miguel Gómez Prieto<sup>1</sup>, GPT-5

<sup>1</sup> Estudiante CDIA, UPM

## Resumen:

Este trabajo aborda la clasificación automática de lesiones cutáneas utilizando el conjunto de datos HAM10000, con especial énfasis en la detección de lesiones malignas, que se encuentran infrarepresentadas. Se propone un enfoque jerárquico compuesto por una primera etapa binaria (benigno vs. maligno) y una segunda etapa multiclasificación sobre las lesiones malignas, contrastando su rendimiento frente a un modelo multiclasificación directo. En la etapa multiclasificación, se analizan distintos enfoques basados en aprendizaje contrastivo junto con alternativas no contrastivas y diversas estrategias de entrenamiento, con el objetivo de evaluar su eficacia en la generación de representaciones discriminativas. Para aportar interpretabilidad, se incorpora Grad-CAM, que permite identificar las regiones más relevantes en las predicciones. Un aspecto central del trabajo es el tratamiento del desbalanceo de clases, mediante técnicas de reponderación y aumento de datos, con el fin de mitigar la disparidad entre categorías y mejorar la robustez del modelo. Los resultados muestran que la combinación de estrategias contrastivas y equilibrio de clases contribuye a una clasificación más precisa y confiable, ofreciendo una solución escalable e interpretable con potencial relevancia clínica.

**Index Terms:** Clasificación de lesiones cutáneas, Aprendizaje contrastivo, Balanceo de clases, Grad-CAM, Diagnóstico asistido por ordenador (CAD)

## 1 Introducción

El cáncer de piel es uno de los tipos de cáncer más frecuentes a nivel mundial, siendo el melanoma su variante más agresiva y potencialmente mortal [1]. La detección temprana y la clasificación precisa de las lesiones cutáneas son fundamentales para mejorar el pronóstico y la eficacia del tratamiento. Sin embargo, el diagnóstico manual basado en imágenes dermatoscópicas presenta desafíos importantes debido a la similitud visual entre lesiones benignas y malignas, la variabilidad entre especialistas y la necesidad de experiencia clínica especializada. La creciente incidencia del melanoma y la limitada disponibilidad de especialistas en muchas regiones hacen que la implementación de sistemas automáticos de apoyo al diagnóstico pueda ser útil.

En los últimos años, las técnicas de aprendizaje profundo han demostrado ser prometedoras para automatizar la clasificación de lesiones cutáneas, alcanzando rendimientos comparables a los de dermatólogos expertos. Entre estas técnicas, el aprendizaje contrastivo ha emergido como una estrategia eficaz para generar representaciones robustas y semánticamente significativas, especialmente en contextos con datos etiquetados limitados y distribuciones desbalanceadas.

Este trabajo presenta un estudio comparativo sobre la clasificación automática de lesiones cutáneas utilizando el conjunto de datos HAM10000 [4]. El uso de este dataset estandarizado permite evaluar de manera objetiva diferentes metodologías, fomentando la reproducibilidad y la comparación justa entre enfoques. En este contexto, el aprendizaje profundo no solo ofrece mejoras en precisión, sino también la posibilidad de desarrollar herramientas escalables e interpretables que contribuyan a una atención médica más equitativa y con potencial relevancia clínica como apoyo al diagnóstico dermatológico.

El código fuente utilizado en este trabajo, junto con la reproducción de resultados, está disponible públicamente<sup>1</sup>.

### 1.1 Contribuciones

Las principales contribuciones de este trabajo son las siguientes:

- Se realiza un estudio comparativo entre un pipeline jerárquico y un modelo multiclasificación directo, evaluando su impacto en la clasificación de lesiones cutáneas.
- En la primera etapa del pipeline jerárquico, se aborda la clasificación binaria entre lesiones benignas y malignas, con especial énfasis en mejorar la detección de las lesiones malignas, que se encuentran infrarepresentadas en el dataset.
- Se aplican técnicas de reponderación y aumento de datos para mitigar el desequilibrio de clases y mejorar la sensibilidad hacia las categorías minoritarias.
- En la segunda etapa, se comparan enfoques basados en aprendizaje contrastivo frente a alternativas no contrastivas para la clasificación multiclasificación de las lesiones malignas, analizando su capacidad para generar representaciones discriminativas.
- Se realiza un análisis visual de los *embeddings* para explorar relaciones entre lesiones malignas, formación de clústeres e identificación de casos ambiguos.
- Se incorpora Grad-CAM (Gradient-weighted Class Activation Mapping) como herramienta de interpretación, permitiendo visualizar las regiones más relevantes en las predicciones y facilitar la validación clínica.

## 2 Antecedentes

El diagnóstico de lesiones cutáneas ha cobrado relevancia en los últimos años debido al aumento en la incidencia del cáncer de piel y a la necesidad de herramientas clínicas que complementen la evaluación dermatológica. Las imágenes dermatoscópicas permiten observar estructuras subcutáneas con mayor detalle, pero su interpretación requiere experiencia especializada y puede estar sujeta a variabilidad entre profesionales. En este contexto, los sistemas de aprendizaje profundo han demostrado ser eficaces para asistir en la clasificación de lesiones, mejorando la precisión diagnóstica y reduciendo la carga clínica.

<sup>1</sup> <https://github.com/MiguelGP-13/PCD-Proyect>

## 2.1 Revisión de literatura

Diversos estudios han explorado el uso de redes neuronales convolucionales (CNN) para la clasificación de lesiones cutáneas. En particular, se han utilizado arquitecturas como ResNet, Inception y EfficientNet para distinguir entre lesiones benignas y malignas, así como para realizar clasificación multiclas sobre tipos específicos de lesiones pigmentadas. Un análisis sistemático reciente destaca que los modelos basados en aprendizaje profundo superan a los métodos tradicionales en tareas de clasificación y detección, especialmente cuando se entran con conjuntos de datos como HAM10000 y ISIC [2].

Por otro lado, el aprendizaje contrastivo supervisado ha emergido como una técnica prometedora para mejorar la calidad de los *embeddings* en contextos médicos. Aunque su aplicación en dermatología aún es limitada, estudios recientes han demostrado su utilidad en tareas de segmentación y clasificación en imágenes médicas, permitiendo representar similitudes semánticas entre muestras con mayor fidelidad [5]. Esta técnica resulta especialmente útil cuando se dispone de datos etiquetados escasos o desequilibrados, como ocurre en muchos escenarios clínicos.

En este trabajo, se propone integrar ambas aproximaciones: una primera etapa de clasificación binaria mediante CNN, seguida de una segunda etapa multiclas basada en aprendizaje contrastivo supervisado, con el objetivo de mejorar la precisión y la interpretabilidad del sistema.

## 3 Metodología

Este trabajo propone un estudio comparativo sobre la clasificación de lesiones cutáneas utilizando el conjunto de datos HAM10000. Se analizan dos enfoques principales: un **pipeline jerárquico**, compuesto por dos etapas consecutivas, comparándolo con un **modelo multiclas único** que clasifica todas las lesiones en una sola etapa. El objetivo es evaluar qué estrategia resulta más eficaz para la detección de lesiones malignas, que se encuentran infrarepresentadas en el *dataset*. A continuación se detallan los enfoques utilizados en cada caso.

### 3.1 Clasificación binaria: lesión benigna vs maligna

La primera etapa del pipeline jerárquico tiene como objetivo distinguir entre lesiones benignas y malignas. Para esta tarea se comparó el entrenamiento desde cero frente al transfer learning. Dado el desbalanceo del *dataset*, se aplicaron técnicas de reponderación de clases y aumento de datos para mitigar la disparidad entre categorías y mejorar la sensibilidad hacia las lesiones malignas.

**3.1.1 Entrenamiento desde cero** Inicialmente se entrenó una red convolucional desde cero utilizando el conjunto de datos HAM10000. A pesar de aplicar técnicas de regularización y ajuste de hiperparámetros, el modelo no logró alcanzar métricas de rendimiento satisfactorias. Esto se atribuye a la complejidad visual de las lesiones cutáneas, lo que motivó la exploración de enfoques más robustos.

**3.1.2 Transfer learning** Se implementaron modelos preentrenados sobre ImageNet para aprovechar representaciones previamente aprendidas. Se evaluaron tres arquitecturas: VGG16, ResNet50 e InceptionV3, todas adaptadas para aceptar imágenes de

tamaño  $224 \times 224$  píxeles. Se congelaron las capas convolucionales iniciales y se entrenaron las capas superiores para la tarea binaria. Este enfoque mejoró significativamente la precisión y la estabilidad del entrenamiento.

## 3.2 Grad-CAM

Para mejorar la interpretabilidad del modelo en la etapa de clasificación binaria, se aplicó la técnica Grad-CAM sobre el modelo con mejor rendimiento: VGG16 con transferencia de aprendizaje. Grad-CAM permite visualizar las regiones de la imagen que más influyen en la decisión del modelo, proporcionando una explicación visual de las predicciones.

Esta técnica genera mapas de calor superpuestos sobre las imágenes originales, destacando las áreas que el modelo considera más relevantes para clasificar una lesión como benigna o maligna. De este modo, se facilita la validación clínica del sistema y se promueve la confianza en su uso como herramienta de apoyo diagnóstico.

### 3.3 Clasificación multiclas: contrastivo vs. no contrastivo

Las muestras clasificadas como malignas en la primera etapa del pipeline se procesan en una segunda etapa multiclas. Aquí se comparan tres enfoques principales:

1. **Contrastivo por pares:** basado en la construcción explícita de parejas de muestras, donde cada par se etiqueta como positivo si ambas imágenes pertenecen a la misma clase o como negativo si corresponden a clases distintas. La función de pérdida busca minimizar la distancia entre los *embeddings* de los pares positivos y maximizar la distancia entre los negativos mediante un margen fijo.
2. **Contrastivo supervisado:** basado en aprendizaje contrastivo supervisado (SupConLoss [3]), que aprovecha el contexto completo del *batch* para generar *embeddings* discriminadores y capturar similitudes semánticas entre lesiones.
3. **No contrastivo:** basado en pérdida de entropía cruzada estándar, que sirve como referencia para evaluar el beneficio del enfoque contrastivo.

**3.3.1 Función de pérdida SupConLoss** En el enfoque contrastivo supervisado se emplea la pérdida SupConLoss, que utiliza las etiquetas para distinguir entre pares de muestras de la misma clase (positivos) y de clases distintas (negativos).

Dado un *batch* de  $B$  muestras con *embeddings*  $\mathbf{z}_i \in \mathbb{R}^D$  y etiquetas  $y_i$ , los *embeddings* se normalizan mediante  $\ell_2$  para evitar que la magnitud de los vectores afecte a la comparación:

$$\tilde{\mathbf{z}}_i = \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|_2}.$$

La similitud entre dos muestras se calcula como:

$$S_{ij} = \frac{\tilde{\mathbf{z}}_i \cdot \tilde{\mathbf{z}}_j}{\tau}, \quad S \in \mathbb{R}^{B \times B},$$

donde  $\tau$  es un parámetro de temperatura que controla la concentración de la distribución de similitudes.

Para identificar qué pares deben considerarse positivos, se construye una máscara binaria  $M_{ij}$  que marca con 1 las parejas de la misma clase (excepto la comparación consigo misma) y con 0 el resto:

$$M_{ij} = \begin{cases} 1 & \text{si } y_i = y_j \text{ y } i \neq j, \\ 0 & \text{en otro caso.} \end{cases}$$

La pérdida para  $i$  se obtiene promediando sobre sus positivos  $P(i)$ :

$$\ell_i = -\frac{1}{|P(i)|} \sum_{j \in P(i)} \log \left( \frac{\exp(S_{ij})}{\sum_{k \neq i} \exp(S_{ik})} \right),$$

La pérdida total del *batch* es:

$$\mathcal{L}_{\text{SupCon}} = \frac{1}{B} \sum_{i=1}^B \ell_i.$$

### 3.4 Modelo multiclasé único

Como referencia, se entrena el modelo multiclasé directo que clasifica todas las lesiones en una sola etapa, sin separación binaria previa. Este enfoque permite comparar el rendimiento global frente al pipeline jerárquico y analizar las ventajas y limitaciones de cada estrategia en la detección de lesiones malignas.

## 4 Experiments

### 4.1 Dataset

Se utilizó el conjunto HAM10000 [4], compuesto por 10015 imágenes dermatoscópicas correspondientes a 7 tipos de lesiones pigmentadas. La distribución de imágenes por clase es la siguiente:

Código	Clase	Número de imágenes
nv	Nevus (benigno)	6705
bkl	Lesión queratósica (benigno)	1099
vasc	Lesiones vasculares (benigno)	142
df	Dermatofibroma (benigno)	115
mel	Melanoma (maligno)	1113
bcc	Carcinoma basocelular (maligno)	514
akiec	Queratosis actínica / carcinoma epidermoide (maligno)	327

Cuadro 1

Distribución de imágenes por clase en el *dataset* HAM10000.

A partir de esta distribución se construyeron tres subconjuntos para los experimentos:

- Subconjunto base:** Benignas (7919), *mel* (1113), *bcc* (514), *akiec* (327).
- Subconjunto binario:** Benignas (*nv*, *bkl*, *vasc*, *df*) → 7919 imágenes. Malignas (*mel*, *bcc*, *akiec*) → 2096 imágenes.
- Subconjunto multiclasé maligno:** *mel* (1113), *bcc* (514), *akiec* (327).

### 4.2 Preprocesamiento

Las imágenes se redimensionaron a  $224 \times 224$  píxeles, se normalizaron en el rango [0,1] y se aplicaron técnicas de aumento de datos: rotaciones, flips horizontales y variaciones de brillo.

Además, dependiendo del modelo, para los preentrenados, se les aplicaba la función de preproceso correspondiente.

Para entrenar, los datos se dividieron en 3 conjuntos, manteniendo la distribución de las clases: entrenamiento (85 % imágenes), evaluación (5 % imágenes) y test (10 % imágenes)

### 4.3 Pipeline

**4.3.1 Clasificación binaria** Al entrenar los modelos, observamos que tendían a estancarse prediciendo todas las muestras como benignas. Esto generaba un *accuracy* aparentemente alto, pero engañoso, debido al fuerte desequilibrio de clases.

Para mitigar este problema implementamos técnicas de *oversampling*, consiguiendo una mejora del *accuracy* en la clase maligna únicamente en el conjunto de entrenamiento. Sin embargo, este efecto no se trasladó al conjunto de validación, ya que las imágenes generadas para el oversampling eran muy similares entre sí y el modelo terminaba sobreaprendiendo (*overfitting*).

También experimentamos con distintos valores de pesos para las clases para contrarrestar el desbalanceo original, finalmente asignando:

$$\text{Benignas} = 0,6212, \quad \text{Malignas} = 2,5623$$

Esta estrategia ayudó a que el modelo prestara más atención a la clase minoritaria, mejorando mucho las métricas, alcanzando un F1-score de 0.87 en test.

Por último, se probó un *ensemble* con los 3 mejores modelos obtenidos para esta tarea. Pero, como esperábamos, el resultado no mejoró y en cambio, el tiempo computacional fue muy superior.

**4.3.2 Clasificación entre malignas** Se empezó buscando un enfoque contrastivo, para poder añadir nuevas clases de manchas malignas sin tener que reentrenar el modelo. Pero al entrenar, nunca llegó a converger, quedándose en el 50 % de *accuracy* al predecir si dos imágenes son de la misma clase o no, es decir, completamente aleatorio.

Por lo que, entrenamos un modelo convolucional que alcanzó un *accuracy* del 81 %, como modelo a batir o por lo menos igualar, pero con la ventaja de la posibilidad de hacer *zero-shot* con el contrastivo.

Por último, se probó la pérdida SupConLoss, que mezcla ambos enfoques. Se consiguió igualar el *accuracy* del modelo convolucional. Pero curiosamente no se consiguió con una red siamesa, que se quedaba en el 60 %, sino con un KNN de los *embeddings*. El KNN se entrenó con las muestras del conjunto de entrenamiento y alcanzó el 82 % en *accuracy* en el conjunto de prueba. Esto probablemente se deba a que para la red siamesa, se intentó obtener el centroide de la clase y comparar el embedding de la imagen con los 3 centroides, determinando cuál era el más parecido. Como los embeddings eran muy dispersos, el centroide no capturaba suficientemente la variabilidad de las clases, disminuyendo el porcentaje de acierto.

Al requerir SupConLoss grandes batches, el principal problema que se enfrentó en esta etapa fue la memoria RAM de la gráfica. Obligó a reescalar las imágenes a 128x128 píxeles, y no se pudo probar modelos con transfer learning.

#### 4.4 Modelo completo

Al entrenar un modelo desde cero, el *accuracy* no superaba el 50 %. Probablemente debido a la escasa profundidad de la arquitectura utilizada, que era la misma que se había probado previamente para la clasificación binaria entre muestras malignas y benignas. Por ello, se optó por realizar fine tuning sobre el modelo que mejores resultados obtuvo en la clasificación binaria, el VGG16. Este modelo mostró una mejora, alcanzando un 68 % de *accuracy*. Aumentar aún más la profundidad no resultaba conveniente, ya que el modelo comenzaba a presentar overfitting al continuar el entrenamiento.

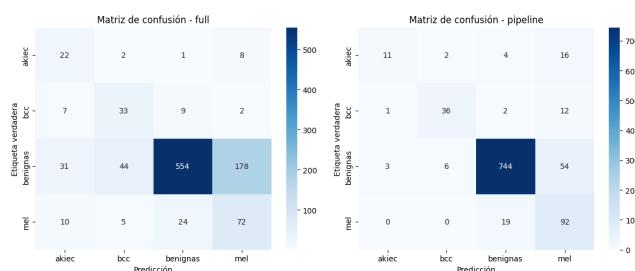
### 5 Resultados

#### 5.1 Comparación resultado final

Aunque la diferencia en rendimiento no es muy grande, los resultados muestran que, contrariamente a lo esperado, donde un modelo convolucional multiclasa debería aprovechar mejor toda la información disponible, el enfoque basado en pipeline ofrece un mejor rendimiento. No solo permite la incorporación de nuevas clases malignas de manera más flexible, sino que también consigue un resultado superior en comparación con el modelo convolucional directo.

Modelo	Accuracy	F1-score	Recall
Pipeline	<b>0.74</b>	<b>0.76</b>	<b>0.74</b>
Modelo completo	0.68	0.72	0.68

**Cuadro 2**  
Comparación de rendimiento en la clasificación final.



**Figura 1.** Comparación de matrices de confusión.

Se observa en las matrices de confusión que el pipeline comete menos errores, especialmente en la separación entre lesiones malignas y benignas, que es la más relevante. El principal fallo del pipeline que tiende a predecir melanoma más de lo que debería. Esto se debe a que es la clase con más muestras dentro de las malignas.

#### 5.2 Pipeline

**5.2.1 Clasificación binaria** Los resultados muestran que los modelos basados en transferencia de aprendizaje superan claramente al entrenamiento desde cero. El mejor desempeño se obtuvo con VGG16. Pero, ResNet50 destaca categorizando bien el *Recall en malignas* del 85 % frente al 75 % de VGG16 en las lesiones malignas. Esto implica que comete pocos falsos negativos, es decir, que hay

pocas manchas malignas que clasifique como benignas, que es el objetivo más importante.

Modelo	Accuracy	F1-score	Recall en malignas
Desde cero	0.79	0.77	0.79
VGG16 (transfer)	<b>0.87</b>	<b>0.87</b>	0.75
ResNet50 (transfer)	0.82	0.83	<b>0.85</b>
InceptionV3 (transfer)	0.77	0.79	0.61

**Cuadro 3**  
Comparación de rendimiento en la clasificación binaria.

**5.2.2 Clasificación de clases malignas** En la segunda etapa del pipeline se evaluaron tres variantes para la clasificación multiclasa de las lesiones malignas: contrastivo por pares, contrastivo supervisado y no contrastivo.

Enfoque	Accuracy	F1-score	Recall
Contrastivo supervisado	<b>0.82</b>	0.80	<b>0.82</b>
No contrastivo	0.81	0.80	0.81

**Cuadro 4**  
Comparación de rendimiento en la clasificación multiclasa de lesiones malignas.

Ambos enfoques obtuvieron resultados muy similares. En cambio, el contrastivo por pares no convergió, por lo que se quedaría en el 30 % de *accuracy*.

#### 5.3 Visualización de embeddings



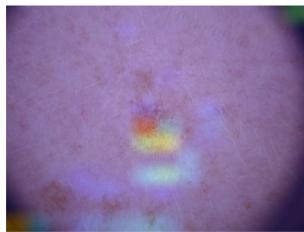
**Figura 2.** Comparación de embeddings.

Para evaluar la calidad de las representaciones aprendidas, se realizó un análisis visual de los *embeddings* generados en la etapa multiclasa. Se puede ver que la clase *melanoma*, la más dominante en el conjunto de datos, forma un clúster más definido y separado respecto a las demás categorías malignas. Sin embargo, las clases *carcinoma basocelular* y *queratosis actínica* presentan una mayor superposición y sobretodo dispersión, lo que dificulta su diferenciación y genera casos ambiguos.

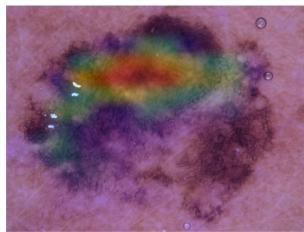
Este comportamiento sugiere que, aunque el aprendizaje contrastivo supervisado contribuye a mejorar la discriminación entre clases, la capacidad de separación sigue siendo limitada para las categorías minoritarias. Probablemente esto se deba a que el modelo que genera los *embeddings* no era lo suficientemente profundo, o que el *batch* debería ser más grande, para ser capaz de capturar patrones más sutiles en las representaciones latentes.

En términos clínicos, la identificación de clústeres bien definidos resulta útil para reforzar la confianza en las predicciones, mientras que la detección de regiones de solapamiento permite señalar casos que requieren una revisión más cuidadosa por parte del especialista. De este modo, el análisis visual de *embeddings* no solo aporta información técnica sobre el rendimiento del modelo, sino que también contribuye a su validación clínica al destacar posibles fuentes de error o incertidumbre.

#### 5.4 Grad-CAM



**Figura 3.** Visualización con Grad-CAM sobre una mancha benigna



**Figura 4.** Visualización con Grad-CAM sobre una mancha maligna.

Las visualizaciones con Grad-CAM muestran que el modelo concentra su atención en regiones pigmentadas y estructuralmente complejas, coincidiendo con criterios clínicos utilizados por dermatólogos. Esto refuerza la interpretabilidad del sistema y su potencial utilidad como herramienta de apoyo diagnóstico al profesional, ayudándole a repasar ciertas zonas, para intentar asegurar que comete los mínimos errores posibles. Las zonas en rojo indican mayor relevancia para la predicción.

## 6 Discusión

Desde la perspectiva clínica, la mejora en la detección de lesiones malignas es de gran importancia, siendo la correcta diferenciación entre benignas y malignas el aspecto más crítico en la práctica dermatológica.

Los resultados obtenidos en este estudio muestran que el pipeline jerárquico ha ofrecido un rendimiento superior al modelo multiclase directo. Aunque la diferencia no es muy grande, la separación inicial entre lesiones benignas y malignas parece favorecer la sensibilidad hacia las categorías minoritarias, lo que resulta especialmente relevante en contextos clínicos. Además, el pipeline puede aportar flexibilidad para incorporar nuevas clases malignas sin necesidad de reentrenar el sistema completo.

La incorporación de Grad-CAM contribuye a que el modelo pueda ser utilizado como apoyo al médico, al permitir validar las predicciones mediante la visualización de las regiones más relevantes en la decisión del sistema.

En el plano técnico, las estrategias de reponderación y aumento de datos han sido determinantes para mitigar el desequilibrio del dataset HAM10000, mejorando la sensibilidad hacia las clases minoritarias. Asimismo, el aprendizaje contrastivo supervisado combinado con un clasificador KNN sobre los *embeddings* ha mostrado resultados competitivos frente a enfoques de aprendizaje profundo con convolucionales, aunque el contraste por pares no logró converger en nuestros experimentos. Resulta especialmente llamativo que el KNN aplicado sobre los *embeddings* haya alcanzado un rendimiento superior al obtenido mediante la comparación directa de

distancias en modelos contrastivos, e incluso comparable al de un modelo convolucional entrenado específicamente para la tarea. Este hallazgo sugiere que la calidad de las representaciones aprendidas puede ser más determinante que la complejidad del clasificador utilizado.

## Trabajo futuro

Como línea de investigación futura, sería interesante explorar enfoques de *few-shot learning* orientados a la incorporación de nuevas clases malignas dentro del pipeline sin necesidad de un reentrenamiento completo. Este tipo de técnicas permitiría ampliar la cobertura del sistema hacia categorías poco representadas, manteniendo la flexibilidad del enfoque jerárquico. En particular, aplicar *few-shot* al aprendizaje contrastivo podría facilitar la integración de nuevas clases de lesiones malignas y reforzar la utilidad clínica del modelo en escenarios más diversos. Asimismo, validar el pipeline en otros conjuntos de datos clínicos será necesario para confirmar su robustez y transferibilidad.

## 7 Conclusiones

En este trabajo se ha realizado un estudio comparativo entre un pipeline jerárquico y un modelo multiclase directo para la clasificación automática de lesiones cutáneas. En nuestro caso, el pipeline ha mostrado un mejor desempeño, especialmente en la separación entre lesiones benignas y malignas, lo que resulta clave para reducir errores clínicamente relevantes. Además, la segunda etapa multiclase basada en aprendizaje contrastivo supervisado y KNN sobre *embeddings* ha alcanzado métricas similares a las de modelos convolucionales, aportando flexibilidad para la incorporación de nuevas clases malignas.

La incorporación de Grad-CAM ha reforzado la utilidad clínica del sistema, al proporcionar interpretabilidad y permitir que los especialistas validen las predicciones sobre la base de criterios visuales relevantes. De este modo, el modelo no solo mejora la precisión en la clasificación, sino que también se presenta como una herramienta práctica de apoyo al médico.

En conjunto, los resultados sugieren que la combinación de técnicas de transferencia de aprendizaje, estrategias de balanceo y enfoques contrastivos constituye una vía prometedora para el desarrollo de sistemas automáticos de apoyo al diagnóstico dermatológico. No obstante, será necesario validar estos hallazgos en otros datasets y explorar nuevas estrategias, como el *few-shot learning*, para seguir mejorando la capacidad del sistema en escenarios clínicos reales.

## Agradecimientos

Se agradece a Kaggle y a los creadores del dataset HAM10000 por poner a disposición los datos y por ofrecer un entorno con gráficas que facilitó el entrenamiento y análisis de los modelos.

Se reconoce también el respaldo de la Universidad Politécnica de Madrid por el entorno académico y los recursos brindados.

Se expresa gratitud a la comunidad de software libre, en especial a los desarrolladores de librerías como *Tensorflow*, *scikit-learn* y otras herramientas que hicieron posible la implementación de este trabajo. Asimismo, se agradece a plataformas como GitHub y Overleaf, así como a José Areia por la plantilla de LaTeX *NobArticle*, que facilitaron la documentación, organización y maquetación del proyecto.

Finalmente, se reconoce la enseñanza del Profesor Jesús Bobadilla, en el curso de PCD de la ETSISI-UPM, que proporcionó el marco académico adecuado para la realización de esta investigación.

## Referencias

1. Apalla Z, Lallas A, Sotiriou E, Lazaridou E y Ioannides D. Epidemiological trends in skin cancer. *Dermatology Practical Conceptual* 2017; 7:1-6. doi: 10.5826/dpc.0702a01
2. Debele TG, Kebede YA, Yohannes D, Alemu B, Abate S, Abebe T, Tsegaye S, Abebe F y Abebe T. Skin Lesion Classification and Detection Using Machine Learning Techniques: A Systematic Review. *Diagnostics* 2023; 13. Received: 30 August 2023; Revised: 22 September 2023; Accepted: 24 September 2023; Published: 7 October 2023:3147. doi: 10.3390/diagnostics13193147. Available from: <https://www.mdpi.com/2075-4418/13/19/3147>
3. Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, Maschinot A, Liu C y Krishnan D. Supervised Contrastive Learning. *arXiv preprint arXiv:2004.11362* 2020 Apr :1-15. doi: 10.48550/arXiv.2004.11362. Available from: <https://arxiv.org/abs/2004.11362>
4. Tschandl P, Rosendahl C y Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data* 2018; 5:180161. doi: 10.1038/sdata.2018.161. Available from: <https://doi.org/10.1038/sdata.2018.161>
5. Wang WC, Ahn E, Feng D y Kim J. A Review of Predictive and Contrastive Self-supervised Learning for Medical Images. *Machine Intelligence Research* 2023; 20:483-513. doi: 10.1007/s11633-022-1406-4. Available from: <https://link.springer.com/article/10.1007/s11633-022-1406-4>