

Clasificación del tipo de carcinoma en imágenes dermatológicas.

Claudia Olavarrieta Martínez

CLAUDIA.OLAVARRIETA@ESTUDIANTES.MATCOM.UH.CU

C-511

Marcos Adrián Valdivié Rodríguez

MARCOS.VALDIVIE@ESTUDIANTES.MATCOM.UH.CU

C-512

Damián O'Hallorans Toledo

DAMIAN.OHALLORANS@ESTUDIANTES.MATCOM.UH.CU

C-512

1. Introducción

La telemedicina se está convirtiendo en un campo de gran interés para la sociedad actual. Una de las ramas principales es la teledermatología, debido a que se puede realizar una clasificación preliminar de un tipo de lesión debido a sus patrones visuales con una imagen tomada por el paciente.

Con el objetivo de facilitar el diagnóstico a los dermatólogos existen enfoques dirigidos a realizar un clasificación automatizada a partir de una imagen de la lesión. Determinar el tipo de cáncer de piel es de los objetivos primordiales, donde los trabajos principales realizados están enfocados en la clasificación del melanoma.

Uno de los principales exponentes en el apoyo al uso de la visión de computadoras para la dermatología es la International Skin Imaging Collaboration (ISIC). La ISIC es una asociación académica e industrial con el objetivo de facilitar imágenes digitales de la piel para ayudar a reducir la mortalidad por melanoma [2].

En este trabajo el objetivo es la clasificación de dos tipos de cáncer no melanomatoso: el carcinoma basocelular y el carcinoma espinocelular. Para esto se utilizarán diferentes modelos con imágenes de entrenamiento obtenidas del ISIC Archive. Se realizará una comparación de los resultados obtenidos para determinar el modelo con mayor precisión y mejor funcionamiento.

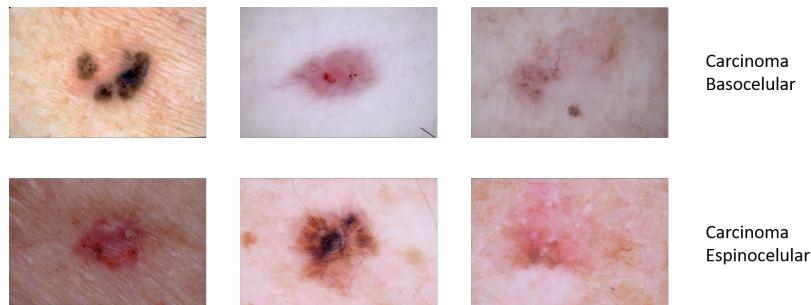


Figura 1: Ejemplo de imágenes del *dataset* de dos tipos de carcinomas a clasificar

2. Conjunto de Datos

Las imágenes a trabajar son extraídas de ISIC Archive, el cual es una plataforma de código abierto disponible públicamente con imágenes de lesiones cutáneas. Las imágenes provienen de centros especializados en melanoma en todo el mundo, aunque también aceptan contribuciones desde otras fuentes bajo determinadas condiciones de uso. Dentro de los objetivos actuales del archivo se encuentran complementar las imágenes de dermatoscopía con imágenes clínicas de primer plano y una representación más amplia de los diferentes tipos de piel [2].

La mayoría de las imágenes están anotadas y asociadas a los siguientes metadatos:

- Atributos de Diagnóstico:

- Benigno o Maligno: benigno, maligno, intermedio, intermedio benigno e intermedio maligno.

- Diagnóstico: nevus, melanoma, carcinoma de células basales, carcinoma de células escamosas, lentigo solar, lesión vascular y otros.

- Atributos Clínicos:

- Sexo
- Edad aproximada: agrupada por décadas.
- Sitio Anatómico General: extremidad inferior, torso anterior, torso posterior, extremidad superior, cabeza/cuello, torso lateral, palmas/suelas y oral/genital.
- Tamaño clínico: Mayor diámetro dado en mm. Agrupado en intervalos de tamaño 10.
- Tipo de diagnóstico: Histopatologías, Microscopía confocal con dermatoscopia concéntrica, las imágenes en serie no muestran ningún cambio, consenso de expertos.
- Antecedentes personales y familiares de melanoma.
- Elementos relacionados con el melanoma: clase, índice mitótico, grosor, ulceración, tipo, entre otros.
- Tipo de nevus: nevus NOS, combinado, azul, spitz, persistente/recurrente, célula fusiforme pigmentada de caña, halo, penetrante profundo y célula fusiforme plexiforme.

Con el objetivo de utilizar metadatos en el entrenamiento se utilizaron solo el sexo, la edad y el sitio anatómico. La mayoría de estos padecimientos se dan en personas mayores y en las zonas de más exposición al sol, por lo que pudieran influenciar en la mejora del diagnóstico.

Para la clasificación del carcinoma se utilizaron las imágenes de diagnóstico de carcinoma basocelular, la cual contenía un total de 1547 imágenes, y el carcinoma espinocelular con un total de 657. Puede apreciarse el desbalance de las cantidades de datos lo cual se debe a que el carcinoma basocelular es un padecimiento mucho más frecuente.

Debido a que el conjunto de datos no resulta suficiente para el entrenamiento debieron utilizarse diferentes técnicas para intentar evitar realizar overfitting.

3. Aumentar conjunto de datos

El aumento del conjunto de datos (*Data Augmentation*) resulta particularmente eficaz en tareas de clasificación de imágenes, ya que es costoso obtener ejemplos etiquetados, y también, porque las clases de imágenes no deberían cambiar bajo pequeñas perturbaciones locales [4].

El aumento de datos consiste en generar más datos de entrenamiento mediante una serie de transformaciones aleatorias que dan lugar a imágenes de aspecto creíble. El objetivo es que en el momento del entrenamiento el modelo nunca observa exactamente la misma imagen dos veces. Esto ayuda a exponer el modelo a más aspectos de los datos y a generalizar mejor. [1].

4. Aprendizaje por transferencia

El aprendizaje por transferencia o *Transfer Learning*, en inglés, es una técnica de Aprendizaje Profundo que permite aprovechar una red neuronal previamente entrenada. Esta se utiliza aunque no haya sido creada y entrenada con los mismos propósitos. Por lo tanto, en lugar de entrenar la red desde cero se simplifica el problema sustancialmente [3]. Esta técnica resulta de gran utilidad en el caso de tener pocas imágenes para realizar el entrenamiento o de no poseer poder de cómputo suficiente, como resulta el caso de este trabajo.

Aunque las imágenes que utilicen las redes entrenadas previamente no estén relacionadas con la clasificación que se desea realizar, se debe tener en cuenta, que la generalidad y reutilización de las representaciones extraídas en las capas convolucionales depende de su profundidad. Es decir, las primeras capas extraen características más generales como líneas, bordes, texturas, entre otras; mientras las capas profundas aprenden a identificar conceptos más abstractos [1].

Las dos formas implementadas para reutilizar una red preentrenada fueron la extracción de características (*feature extraction*) y ajuste de precisión (*fine tuning*).

4.1 Extracción de características

La extracción de características consiste en utilizar las representaciones aprendidas por una red previamente entrenada para obtener rasgos interesantes de las nuevas muestras. Estas capas permanecen fijas o congeladas (en inglés, se conocen como *frozen layers*), es decir no se recalculan sus pesos. A continuación, estas características se pasan a un nuevo clasificador, que se entrena desde cero (Figura 2 b).

4.2 Ajuste de precisión

El ajuste de precisión o ajuste fino (*fine tuning*, en inglés) consiste en "descongelar" o volver a calcular los pesos de capas superiores y entrenar entonces estas capas y el modelo clasificador añadido. De esta forma se ajustan ligeramente las representaciones más abstractas del modelo preentrenado que se esté reutilizando [1]. (Figura 2 c).

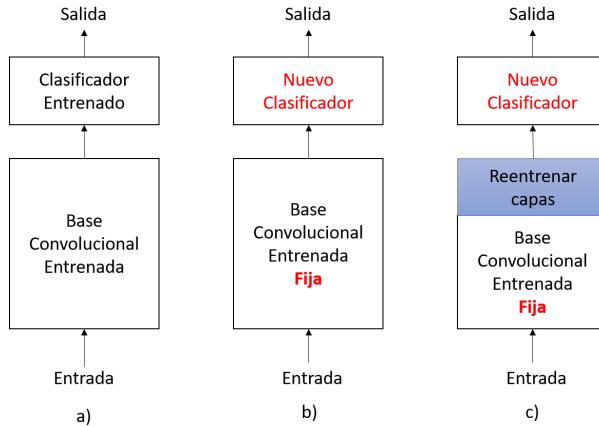


Figura 2: Ejemplo de transfer learning. a) Red previamente entrenada b) Extracción de características, se entrena solamente el nuevo clasificador. c) Ajuste de precisión, se reentrenan las últimas capas y el nuevo clasificador

4.3 VGG

La red empleada para realizar *transfer learning* es la VGG16, esta destaca por su simplicidad y su amplio uso en tareas de reconocimiento y clasificación de imágenes. En cuanto a su arquitectura, está compuesta por bloques de múltiples (normalmente 1, 2 o 3) capas de convolución de tamaño de filtro 3×3 , seguida de una capa de max-pooling. La repetición de estas configuraciones le otorga la profundidad a la red, en este caso 16 (Figura 3).

La entrada de la primera capa de esta arquitectura es una imagen RGB de tamaño, $224 \times 224 \times 3$. Al final de la red se encuentran tres capas *fully connected*, las dos primeras con 4096 neuronas, y la tercera 1000, equivalente a las clases para las que fue preentrenada. Finalmente se encuentra la capa softmax, encargada de clasificar en las diferentes clases según las probabilidades establecidas [3].

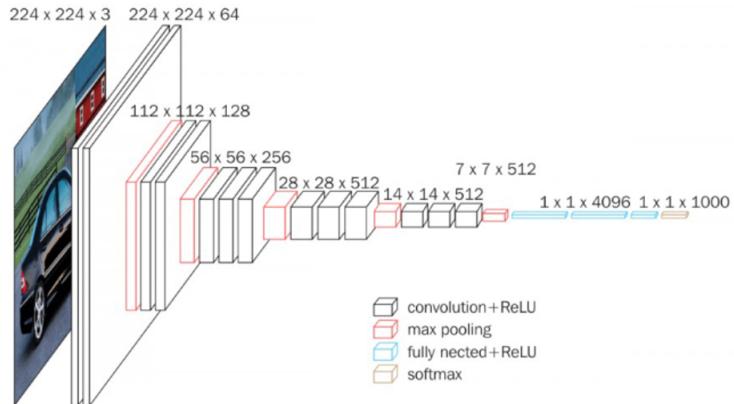


Figura 3: Arquitectura de VGG16. Extraído de [3]

5. Mapas de activación de clases

Se empleó también una técnica de visualización que resulta útil para entender qué partes de una imagen han llevado a la red a decidir un tipo de clasificación. Esto resulta útil para depurar el proceso de decisión, especialmente en el caso de un error; además, permite localizar objetos específicos en una imagen.

Esta categoría general de técnicas se denomina visualización de mapas de activación de clases y consiste en producir mapas de calor de la activación de clases sobre las imágenes de entrada. Un mapa de activación de clase es una cuadrícula 2D de puntuaciones asociadas a una clase de salida específica, calculada para cada lugar de cualquier imagen de entrada, indicando la importancia de cada lugar con respecto a la clase considerada [1].

6. Experimentación

A continuación se presentan los distintos tipos de modelos realizados con el objetivo de obtener la mayor precisión. Se entrenaron los modelos configurando los hiperparámetros con diferentes valores. De forma general se probó con 40, 30 y 25 epochs y un batch size de 8. La función de pérdida utilizada fue *binary crossentropy* y el optimizador RMSProp con una tasa de aprendizaje de 0.001. Considerando que no se cuenta con poder de cómputo suficiente no resultó posible realizar más experimentaciones debido al largo tiempo que tomaba realizar los entrenamientos. Para la implementación de los modelos se utilizó el lenguaje python y la API de Deep Learning escrita sobre Tensorflow: Keras.

6.1 End to End

El primer modelo se entrenó de principio a fin solo con las imágenes disponibles, por lo que debido a la poca cantidad existente no se obtuvieron buenos resultados. Cuenta con un total de parámetros (entrenables) 19 263 809 y las siguientes capas:

Capa (tipo)	Tamaño salida	Parámetros
conv2d (Conv2D)	(None, 222, 222, 32)	896
max_pooling2d (MaxPooling2D)	(None, 111, 111, 32)	0
conv2d_1 (Conv2D)	(None, 109, 109, 64)	18496
max_pooling2d_1 (MaxPooling2D)	(None, 54, 54, 64)	0
conv2d_2 (Conv2D)	(None, 52, 52, 128)	73856
max_pooling2d_2 (MaxPooling2D)	(None, 26, 26, 128)	0
conv2d_3 (Conv2D)	(None, 24, 24, 256)	295168
max_pooling2d_3 (MaxPooling2D)	(None, 12, 12, 256)	0
flatten (Flatten)	(None, 36864)	0
dense (Dense)	(None, 512)	18874880
dense_1 (Dense)	(None, 1)	513

Debido a que la cantidad de datos a analizar es muy pequeña y están desequilibradas las clases, se utilizó Stratified K Fold para lograr un mejor entrenamiento. Se realizó el entrenamiento con 40 épocas y un tamaño de batch de 8. Se graficó el promedio de precisión y pérdida entre los resultados de cada fold, para obtener una comprensión más general de los modelos pudiéndose obtener los gráficos de la figura 8. Se aprecia que el validation accuracy no aumenta, sino que se mantiene en un rango de 0.68 a 0.7.

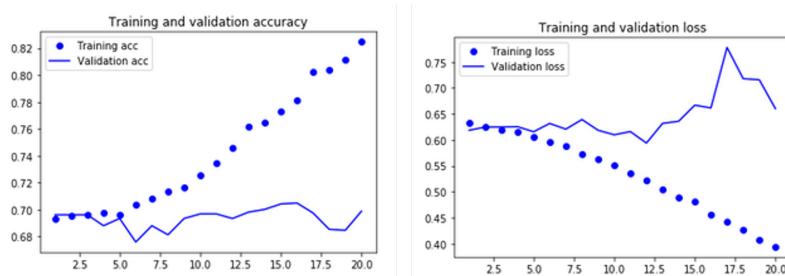


Figura 4: Gráficos para modelo end to end

6.2 End to end. Data Augmentation

Empleando la técnica para aumentar los datos y reutilizando el modelo de la sección 6.1 se obtuvieron resultados distintos.

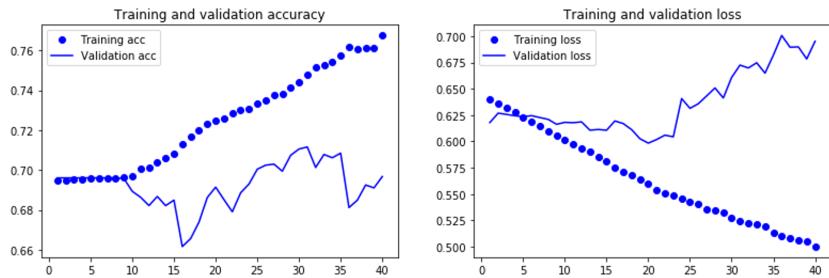


Figura 5: Gráficos para modelo end to end con data augmentation

Se puede ver que el proceso de entrenamiento aumenta la precisión de forma lenta, además, puede verse que el validation accuracy comienza a descender al llegar a diez epochs. La evaluación en el conjunto de prueba dio como resultado un valor de pérdida de **0.70** y una precisión de **0.68**.

6.3 Transfer Learning

Se utiliza la red preentrenada VGG16. En el caso de la extracción de características es eliminando el clasificador existente (capas 14,15,16 Figura 6) y la capa softmax y se añade un nuevo clasificador. Para el ajuste de precisión, se reentrena el último bloque de convolución (capas 11,12,13 Figura 6) y se añade el nuevo clasificador. El clasificador añadido consiste en una capa Flatten, una capa Densa, una capa Dropout y otra capa Densa con tamaño de salida 1 y función de activación softmax.

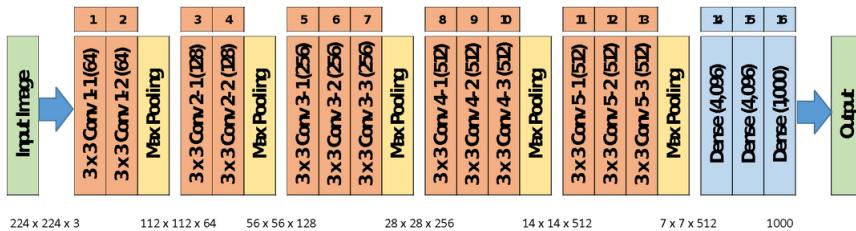


Figura 6: Arquitectura de VGG16. Extraído de [5]

En el modelo de extracción de características para la configuración de 25 epochs, batch size de 8 se obtuvieron los gráficos de la Figura 7. La precisión del entrenamiento aumenta linealmente con cada epoch. Los valores que se obtuvieron para la validación se mantuvieron casi constantes.

Para este modelo, evaluando en el conjunto de prueba se alcanzó una precisión de **0.61** y una pérdida de **0.66**. Teniendo en cuenta que tratamos con enfermedades estos resultados resultan negativos.

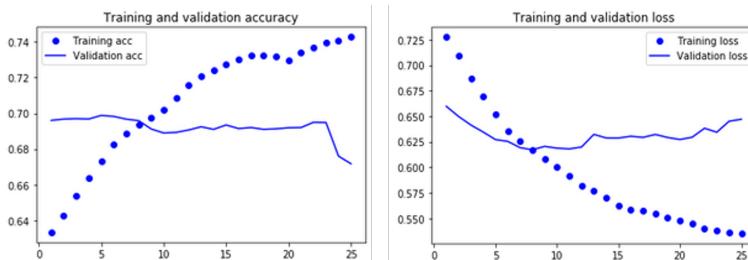


Figura 7: Gráficos de precisión y pérdida obtenidos con 25 epochs y tamaño de batch 8 para el modelo que utiliza la red preentrenada VGG16 y congela todas las capas.

En el caso de volver a entrenar el último bloque de capas convolucionales se obtuvieron los resultados de la Figura 8. En el caso de la precisión del conjunto de validación se puede apreciar que apenas aumenta sus valores, y la pérdida en el conjunto de validación aumenta de manera drástica, alcanzando valores de 1. Realizando la evaluación en el conjunto de prueba se obtuvo una precisión de **0.70** y una pérdida de **1.06**.

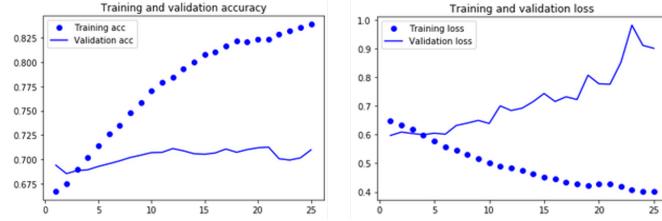


Figura 8: Gráficos de precisión y pérdida obtenidos con 25 epochs y tamaño de batch 8 para el modelo que utiliza la red preentrenada VGG16 y reentrena el último bloque de capas convolucionales.

6.4 Metadatos

Se realizó un modelo con el objetivo de utilizar los metadatos de las imágenes. Se emplearon el sexo, la edad y la posición de la herida, como eran datos categóricos se utilizó *one hot encoding* para convertirlos a vectores de enteros. Se creó una red que concatenara los metadatos y las imágenes para después clasificar los resultados. La arquitectura de la red puede verse en la figura 9. El entrenamiento de esta arrojó los resultados que pueden apreciarse en la figura 10, se puede notar que las funciones de validación no tienen resultados buenos. Este modelo en el conjunto de prueba obtuvo una pérdida de **0.94** y una precisión de **0.65**

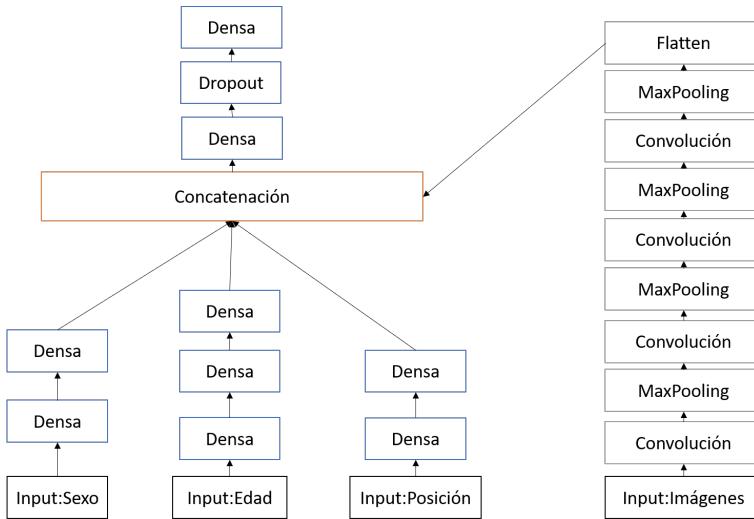


Figura 9: Arquitectura de la red para emplear los metadatos.

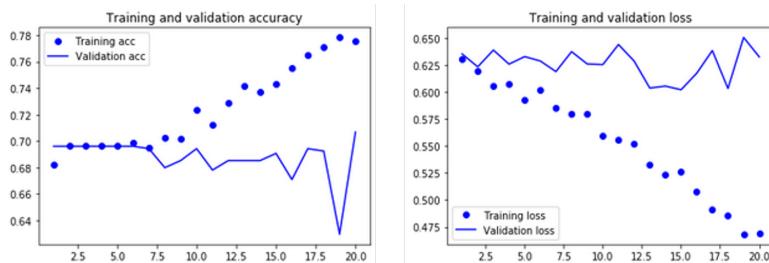


Figura 10: Entrenamiento de la red que emplea los metadatos.

7. Predicciones

Como otra forma de evaluar los modelos se obtuvieron imágenes de ambos padecimientos que no habían sido vistas antes por ningún modelo. Se utilizaron los modelos que mejor resultados mostraron y se comparó la predicción de todos según la clase que debían obtener. La clasificación toma valores entre 0 y 1, donde la clase basocelular se corresponde con los valores de cero y el espinocelular con 1. Se escogieron tres imágenes de cada lesión calculando el promedio de las predicciones de cada modelo, exceptuando el que incluye los metadatos. (Figura 11).

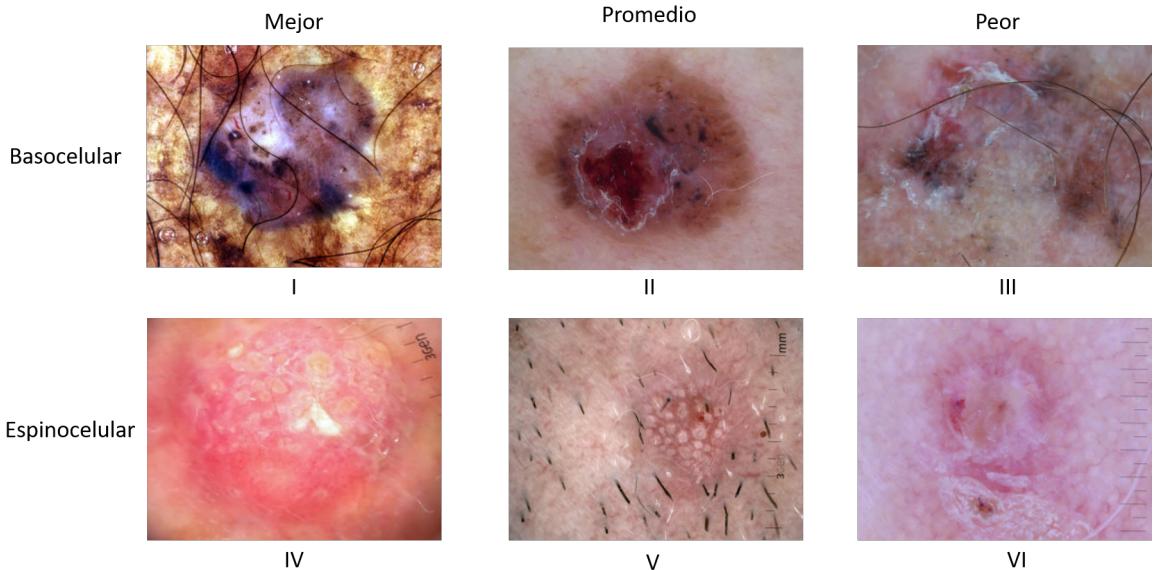


Figura 11: Imágenes de lesiones no vistas por los modelos usados en las predicciones.

Imagen	End-to-End	Data Augmentation	VGG16 Feature Extraction	VGG16 Fine Tuning
Basocelular (I)	2.8e-16	8.7e-10	1.85e-7	1.5e-19
Basocelular (II)	0.02	0.51	0.62	0.21
Basocelular (III)	0.72	0.37	0.6	0.5
Espinocelular (IV)	1.0	0.88	0.49	0.87
Espinocelular (V)	0.69	0.16	0.64	0.99
Espinocelular (VI)	0.01	0.10	0.51	0.03

Para las pruebas con los mapas de activación se muestran en la Figura 12 los resultados obtenidos con el modelo que VGG16 Fine Tuning el cual se aplicó en las imágenes donde mostraba el mejor y el peor resultado. Se puede ver que en las imágenes con mejores resultados (Basocelular I y Espinocelular II) la activación, o zonas de calor como también se les denomina, son más abundantes en el centro del área lesionada. Mientras que en las imágenes con peor comportamiento, no se marcan las partes que resultan relevantes a la lesión y toma como zona importante mayor área que la comprometida por la herida.

Estos mapas de activación de clases permiten, en los casos donde se realicen predicciones correctas, denotar el área que influye en la decisión, que es por tanto el área de la herida o la zona que presenta algún patrón característico.

8. Conclusiones

Pudo observarse en los gráficos que los valores de precisión en el conjunto de validación oscilaron entre 0.65 y 0.70 en la mayoría de los modelos, además de que la pérdida fue considerablemente alta en todos los modelos estudiados alcanzando valores desde 0.6 a 0.7.

Por otra parte, el modelo implementado para emplear los datos anotados de las imágenes no tuvo resultados positivos, de hecho resultó peor que alguna de las otras redes entrenadas. Como trabajo futuro se podría recomendar la búsqueda de una arquitectura que resulte más efectiva, ya que debido a la naturaleza de estas lesiones y los datos disponibles es probable que su inclusión influencie de forma positiva el diagnóstico automático.

Se pudo apreciar que ninguno de los modelos obtuvo un buen comportamiento, todos realizaron overfitting y esto se debe principalmente a que la cantidad de imágenes disponibles no resultaron suficientes para alcanzar valores representativos.

No resulta posible definir como superior a algún modelo debido a que las métricas obtenidas con los entrenamientos resultaron ser negativas para poder darle alguna aplicación práctica.

Como trabajo futuro pudiera recomendarse la búsqueda de nuevas imágenes de los dos tipos de lesiones presentados. Además, con equipos de mayor poder de cómputo, se pudieran realizar entrenamientos y ajuste de parámetros de forma tal que se consiga un modelo que pueda resultar más confiable en la clasificación de algo tan sensible como el cáncer de piel.

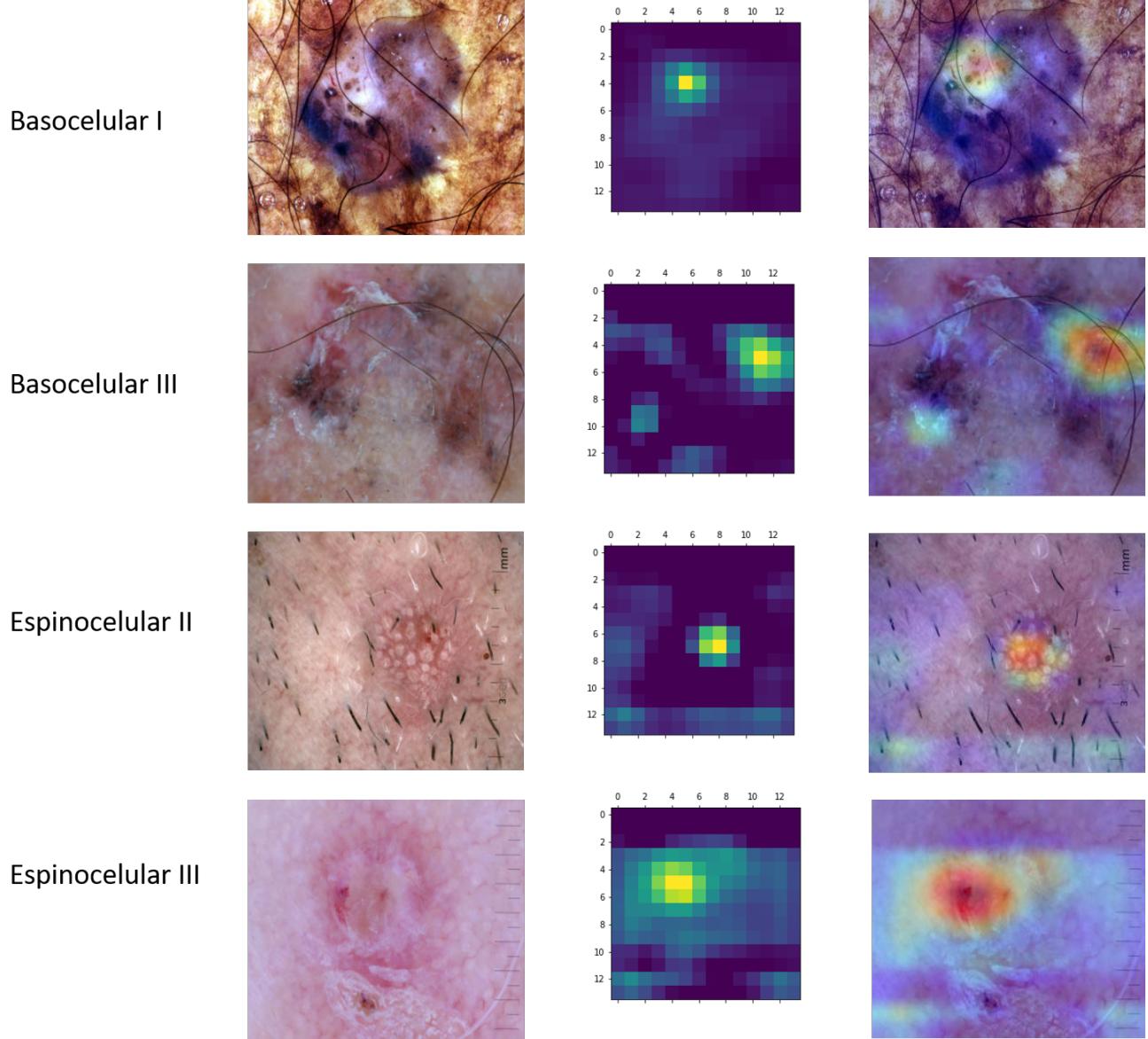


Figura 12: Mapas de activación de clases utilizando el modelo VGG16 Fine Tuning.

Bibliografía

- [1] Francois Chollet. *Deep learning with Python*. Simon y Schuster, 2021.
- [2] ISIC. *ISIC Archive*. 2022. URL: <https://www.isic-archive.com/#!/topWithHeader/tightContentTop/about/isicArchiveContent>.
- [3] Cristina Pérez Lorenzo y col. “Detección precoz de cáncer de piel en imágenes basado en redes convolucionales”. B.S. thesis. 2019.
- [4] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.
- [5] Great Learning Team. *Introduction to VGG16*. 2021. URL: <https://www.mygreatlearning.com/blog/introduction-to-vgg16/>.