TRABAJO FIN DE GRADO

INGENIERÍA INFORMÁTICA

Titulo del Proyecto

Subtitulo del Proyecto

Autor

Cristhian Moya Mota (alumno)

Directores

Diego Jesús García Gil

Julián Luengo Martín

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

—

Granada, 2 de marzo de 2024

Título del proyecto

Subtítulo del proyecto.

Autor
Cristhian Moya Mota

Directores
Nombre Apellido1 Apellido2 (tutor1)
Nombre Apellido1 Apellido2 (tutor2)

Título del Proyecto: Subtítulo del proyecto

Cristhian Moya Mota

Palabras clave: palabra_clave1, palabra_clave2, palabra_clave3, ...

Resumen

Poner aquí el resumen.

Project Title: Project Subtitle

First name, Family name (student)

Keywords: Keyword1, Keyword2, Keyword3, ...

Abstract

Write here the abstract in English.

Yo, **Nombre Apellido1 Apellido2**, alumno de la titulación TITULACIÓN de la **Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación de la Universidad de Granada**, con DNI XXXXXXXXX, autorizo la ubicación de la siguiente copia de mi Trabajo Fin de Grado en la biblioteca del centro para que pueda ser consultada por las personas que lo deseen.

Fdo: Nombre Apellido1 Apellido2

Granada a X de mes de 201 .

D. **Nombre Apellido1 Apellido2 (tutor1)**, Profesor del Área de XXXX del Departamento YYYY de la Universidad de Granada.

D. **Nombre Apellido1 Apellido2 (tutor2)**, Profesor del Área de XXXX del Departamento YYYY de la Universidad de Granada.

Informan:

Que el presente trabajo, titulado ***Título del proyecto, Subtítulo del proyecto***, ha sido realizado bajo su supervisión por **Nombre Apellido1 Apellido2 (alumno)**, y autorizamos la defensa de dicho trabajo ante el tribunal que corresponda.

Y para que conste, expiden y firman el presente informe en Granada a X de mes de 201 .

Los directores:

Nombre Apellido1 Apellido2 (tutor1) **Nombre Apellido1 Apellido2 (tutor2)**

Agradecimientos

Poner aquí agradecimientos...

Índice general

Índice de figuras

Índice de tablas

1 Introducción

1.1. Motivación

El cáncer es una de las causas de muerte principales en el mundo. Su gran agresividad, así como su dificultad de diagnóstico, debido a su gran variedad de ubicaciones y manifestaciones, provoca que un alto porcentaje de casos no sean diagnosticados a tiempo correctamente. Tan solo en 2023, aproximadamente se registraron 20 millones de nuevos casos de cáncer a nivel mundial, y produciéndose algo menos de 10 millones de defunciones.

Estos registros provocan una gran inquietud en la población y entre los expertos de la materia; debido al aumento que se produce cada año, se espera que para el año 2050, el número de nuevos casos sea un 70 % mayor. Por desgracia, no existen formas de prevención claras para este tipo de enfermedad, ni un tratamiento efectivo que permita al paciente recuperarse fácilmente.

La única opción probada es la realización de pruebas rutinarias a colectivos de riesgo, para así acelerar la detección de posibles tumores, y aumentar la esperanza de supervivencia. Esto se ve reflejado en las cifras de los dos tipos de cáncer más frecuentes: el cáncer de mama, y el cáncer colorrectal. Si se detectan en fases iniciales, la correcta recuperación del colon podría aumentarse hasta el 90 %, mientras que en el cáncer de mama, podría reducirse su mortalidad entre un 25-31 %. Gracias a la existencia de pruebas rutinarias programadas por servicio de salud, se puede reducir la mortalidad.

El gran problema de estos tipos de cánceres son la escasa visibilidad y síntomas de los mismos; cuando muestran señales, es probable que la tasa de supervivencia sea mucho menor, sobre todo en el colon. Pero existe otro tipo de cáncer que sí se manifiesta de forma más visible y que puede alarmar al paciente de forma más temprana: el cáncer de piel.

Este tipo de tumores puede manifestarse en las diferentes capas de la dermis, y su origen se atribuye a la exposición prolongada a la luz solar sin hacer uso de protección. Debido a los daños que sufre la capa de ozono, y otros factores ambientales, la cantidad de rayos ultravioleta que llegan hasta la superficie ascendió desde que se tienen registros. Si bien la capa de ozono parece recuperarse, debemos ser cautos,

y tener cuidado de nuestra piel; los rayos ultravioleta pueden dañar células de la misma, y provocar alteraciones en su material genéctico. Son las que dan lugar al crecimiento incontrolado de células, son las que forman los tumores cancerígenos en la piel.

Se estima que en el mundo, los tumores de la piel representan un tercio de los casos de cáncer diagnosticados. Esta distribución sigue valores parecidos en España, y al igual que las cifras de otros tipos de tumores, los casos diagnosticados aumentan año tras año. Las muertes debidas a esta enfermedad son principalmente, por ser indentificadas en fases tardías de su evolución. Debido a que la piel es el órgano más grande del cuerpo humano, y que está en contacto con todos los capilares sanguineos y el sistema linfático, las células cancerosas se pueden extender por ellos hacia otros lugares del cuerpo.

Aunque este cáncer puede ser identificado de forma más sencilla por su portador, la escasa información acerca del tema, y la confusión con otras lesiones benignas de la piel como verrugas o lunares, provoca una disminución en las posibilidades de supervivencia. Por ello, el objetivo de este trabajo es aportar una nueva forma de diagnóstico que permita a los usuarios obtener una orientación acerca de qué posible lesión están experimentando en la piel, y sirvan como complemento del experto. O bien, ayudar a los expertos a tomar la decisión, acortando los tiempos de diagnóstico para aumentar las posibilidades de supervivencia. Esta tarea será realizada gracias al uso de uno de las herramientas en auge en la actualidad: la inteligencia articificial, y concretamente, el uso de DeepLearning para visión por computador.

Mediante una nueva arquitectura, el propósito es conseguir un buen modelo, capaz de segmentar las manchas de interés en la piel que estén recogidas en una fotografía. Dicha fotografía será capturada con el telefóno movil del usuario, retirando así la necesidad de disponer de dispositivos especializados. Y posteriormente, clasificar dichas manchas para ofrecer al usuario final una respuesta sólida acerca del posible tipo de lesión de piel que sufre.

1.2. El cáncer de piel

Prosiguiendo con el cáncer de piel, su diagnóstico si dificulta, sobre todo, por su amplia variedad de formas, tamaños, texturas y manifestaciones. Aunque su visibilidad pueda parecer evidente, (ya que es observable a nivel macroscópico) puede ser confundido fácilmente con lesiones benignas. Normalmente, suele dividirse entre dos tipos diferentes:

- **Melanomas de la piel**. Son la variante más peligrosa. Su origen se encuentra en los melanocitos, las células encargadas de dar el color bronceado a la piel. Éstas pueden comenzar a crecer sin control originando tumores, los cuales crecen y se diseminan rápidamente hacia otras regiones del organismo, provocando la metástasis, una extensión a nivel total del organismo. Es el más grave de los diagnósticos. Puede identificarse como una mancha oscura en la piel, formando tumores de color café oscuro. Sin embargo, debido a la gran variedad de reacciones, pueden darse de color rosado si dejan de producir melanina. Este aspecto dificulta su diagnóstico, por lo que el papel de las herramientas de visión por computador pueden ayudar a su identificación.
- **Cánceres no melanomas**. Este tipo de cánceres no se ubican en los melanocitos, y pueden ser tratados mediante otras ténicas menos agresivas debido a su rara probabilidad de expansión. Los más comunes, son los tumores de células basales y los de células escamosas:
  - Células basales. Componen la capa inferior de la piel, y son las células encargadas de sustituir aquellas que componen la capa más externa de la piel. Se encuentran , por tanto, en constante reproducción para cubrir aquellas que mueren en la superificie. Si experimentan alguna mutación, producen tumores de color similar al de piel del paciente, con la posibilidad de aparecer en colores como negro brillante en las pieles más oscuras.
  - Células escamosas. Son las células externas de la piel, con forma plana. Se regeneran constantemente gracias a las células basales, que producen estas células las cuales se aplanan a medida que ascienden hacia la capa externa. Es frecuente, de nuevo, en zonas expuestas al sol, sobre todo la cara. Normalmente, se encuentran bien localizados, y puede procederse a su extirpación. En casos en los que se haya extendido, se hace uso de radioterapia.

Aunque en base a su descripción parezcan distinguibles, son fácilmente confudidos por su variedad con otros tumores benignos de la piel, como:

- **Lunares(nevus)**: hiperpigmentación benigna en la piel.
- **Verrugas**: tumores benignos de piel, frecuente debido a virus como el del papiloma humano.

- **Lesiones vasculares**: varices, derrames, y otro tipo de problemas circulatorios.
- **Lipomas**: tumores de tacto blando, debido a su contenido en lípidos (grasa).
- **Queratosis seborreica**: son manchas cerosas, comúnmente desarrolladas en la espalda. De aspecto oscuro y gran relieve, no suponen ninguna amenaza más allá de posible incomodidad al roce o estética.

El uso de aprendizaje profundo para este fin resulta interesante como forma de mejora del diagnóstico ante casos malignos y benignos de gran similitud, los cuales pueden confunir y dificultar la labor incluso a expertos dermatólogos.

2 Estado del arte

En este apartado, evaluaremos los modelos usados para la clasificación de tumores y lesiones cutáneas en la literatura, para así comprender las ventajas que ofrece cada modelo, y los puntos en contra, y obtener una referencia de la tendencia general.

2.1. Aprendizaje profundo en dispositivos móviles

2.1.1. Vertientes actuales

Una de las tendencias actuales en auge consiste en la integración de modelos de aprendizaje en aplicaciones de uso diario para mejorar la predictibilidad de ciertos fenómenos, o bien adecuarse a los hábitos y la vida diaria del usuario. Esto requiere una buena capacidad de cómputo en los dispositivos móviles, cuya principal limitación son las restricciones de espacio y consumo. Por este motivo, existen numerosas vertientes de investigación para conseguir arquitecturas convolucionales preentrenadas, a semejanza de arquitecturas como ResNet o VGGNet, pero con un menor número de parámetros para reducir la cantidad de operaciones aritméticas necesarias para el aprendizaje y la inferencia de los resultados.

Podemos encontrar corrientes de investigación de Google que apuestan por este tipo de arquitecturas, como MobileNet. Ésta se trata de un modelo centrado en la dimensionalidad de la red, así como en el incremento de la velocidad de cálculo, haciendo uso de los conceptos empleado en los módulos Xception.

2.1.2. Comparativa de uso entre Deep Learning y Machine learning

2.1.3. Recursos disponibles

La obtención de datos es un proceso fundamental en la resolución de problemas de Machine Learning. Este tipo de problemas requieren un gran número de imágenes que aporten variedad, y permitan construir un modelo general que se capaz de adaptarse a cambios de iluminación, diferentes puntos de vista y composiciones.

Es clave, por tanto, disponer de diferentes tipos de lesiones, tanto benignas como malignas, así como diferentes tonos de piel. La inexistencia de un tipo de piel en el conjunto de entrenamiento, o la inexistencia de un tipo de lesión podrían provocar resultados sesgados indeseados durante la predicción de la imagen tomada.

Podemos encontrar en la red varios datasets de acceso público que permiten su utilización de forma abierta con fines académicos. Dada a la gran cantidad de publicaciones disponibles, resulta complejo averiguar si los datos a los cuales hace referencia se encuentran disponibles públicamente, si son de acceso restringido, o bien, ya no se encuentran disponibles debido a cambios en su política o la falta de mantenimiento.

Debido a que el estudio de la evolución y el diagnóstico del cáncer de piel de forma temprana es un tema en auge, existen gran cantidad de publicaciones especializadas únicamente en el análisis de los conjuntos de datos públicamente accesibles, como es el caso de la lista propuesta por M. Goyal et Al. [16], o el reciente estudio realizado por Sana Nazari y Rafael García (2023)[30]. Basados en su modelo de estudio y las referencias recomendadas por sus artículos, se ha elaborado el siguiente plan de búsqueda para saber qué tipo de datos utilizar y cuáles descartar:

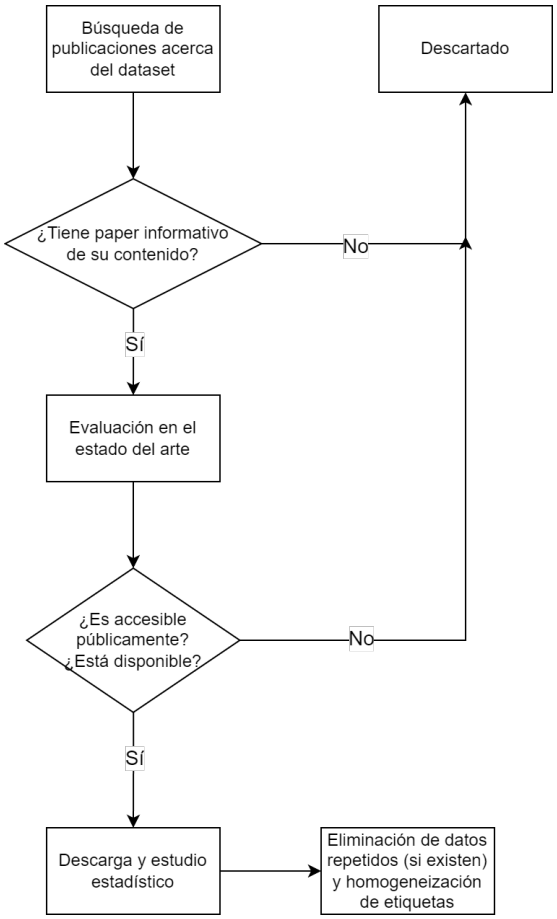

Figura 2.1: Proceso de búsqueda seguido

Siguiendo dicho procedimiento, se han encontrado 6 datasets diferentes, cuyo

origen son instituciones públicas que han cedido datos con fines académicos y de investigación.

Un factor determinante para la elección de estos conjuntos es la diversidad. Es indispensable, para este proyecto, encontrar datos lo suficientemente variados como para distinguir lesiones cancerígenas y no cancerígenas, y disponer de diferentes tonos de piel para entrenar. Aunque los tonos de piel más oscuras sufren lesiones de tipo cancerígeno en menor proporción gracias a su protección natural, también pueden sufrir este tipo de patologías, y es clave ser capaces de detectarlas en cualquier posible paciente.

ISIC Data

El repositorio ISIC (International Skin Imaging Collaboration [1]), contiene imágenes demoscópicas de lesiones principalmente cancerosas. Existe una gran cantidad de publicaciones acerca de este conjunto de datos, debido a su utilización anual durante los años 2016-2020 para la realización de un reto virtual de Machine Learning descrito en [1]. El objetivo, consiste en identificar los melanomas frente a lesiones no cancerosas (ISIC Challenge-2020[2]) o bien, identificar diferentes subtipos de lesiones cancerosas frente a lesiones benignas (ISIC Challenge 2019, [3]).

En el estado del arte, destacan las soluciones que hacen uso de métodos de DeepLearning, como el planteado por Ian Pan [4], finalista de la competición para el año 2020. El ganador de la competición del año 2020, cuyo análisis podemos encontrar en [5], realiza por su parte un enfoque híbrido entre el uso de DeepLearning para la clasificación de los datos de tipo imagen, y clasificación mediante Machine Learning para los metadatos, y así obtener un resultado más preciso.

Debido a que cada uno de estos subconjuntos de datos anuales podían ser pequeños, normalmente se recurría a la reutilización de los conjuntos anteriores para enriquecer el conjunto de entrenamiento. Este método es bastante recomendable, ya que, a mayor conjunto de entrenamiento, más cercanos se encontrarán los parámetros de interés del problema general a solucionar. Sin embargo, hay que realizar dicha fusión con especial cuidado, ya que existen volúmenes de datos considerables que se repitieron en las competiciones de cada año para aumentar el tamaño del dataset, y si se realizase una simple concatenación de los datos, estaríamos desperdiciando esfuerzo computacional en clasificar imágenes redundantes (comportamiento para nada deseable al trabajar sobre entornos móviles de menor potencia.).

Un análisis extenso de los datos asociados a cada Challenge podemos encontrarlo en [6]. En él, se recogen otros modelos del estado del arte utilizables para este fin, así como una forma de tratar los datos duplicados. Los datos pueden ser separados

en los siguientes subconjuntos:

Challenge Dataset Year	Train	Test	Total	Tipo de problema
ISIC 2016	900	379	1279	Clas. binaria
ISIC 2017	2000	600	2600	Clas. Multiclase
ISIC 2018	10015	1512	11527	Clas. Multiclase
ISIC 2019	25331	8238	33569	Clas. Multiclase
ISIC 2020	33126	10982	44108	Clas. Binaria

- ISIC 2016 [7]: Es el dataset de menor tamaño de todos los propuestos. Hace distinción únicamente de los casos malignos y benignos. Contiene imágenes dermoscópicas anotadas con información acerca de la localización de la mancha, y la edad del paciente. Contiene información adicional para la segmentación de la mancha pigmentada de interés (máscaras).

- ISIC 2017 [8] Es un conjunto de mayor tamaño al anterior, y hace alusión a 4 clases diferentes: melanomas, nevus, y seborrheic keratosis. Contiene también información acerca de la edad del paciente, y otros metadatos de interés. La escasa cantidad de datos provoca que normalmente en la literatura este dataset se utilice también como clasificación binaria entre nevus y keratosis, u otros enfoques similares.

- ISIC 2018 [9]. Este dataset contiene un número de imágenes considerables, siendo un total de 10015 imágenes para entrenamiento, y 1512 para test. En este caso, se realiza subclasificación de tipos, a través de las clases melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma y lesiones vasculares. Es de especial interés destacar que este dataset proviene, a su vez, de HAM10000 (Human against machine, [11]) y MSK Dataset [12]. El challenge original comprendía, de nuevo, la clasificación de los diferentes tipos realizando previamente una discriminación de la mancha en cuestión mediante segmentación. Existen gran cantidad de publicaciones que tratan este conjunto de datos, como [13], donde se emplea este dataset para demostrar mejores resultados al emplear transformaciones polares de la imagen y aumentar la invarianza.

- ISIC 2019 [2]. Se trata del mayor conjunto de datos para clasificación multiclase propuesto por ISIC [1]. Se trata del mismo dataset que el año 2018 [9], con la adición de BCN_20000 Dataset [14], cuyos datos provienen del Hospital Clínic de Barcelona [14]. Las clases a clasificar se amplían hasta 9, encontrando subtipos de melanomas en el conjunto.

- ISIC 2020 [3]. El último dataset propuesto públicamente, contiene únicamente datos binarios acordes a melanomas y no malignos.

Todos estos datos pueden ser acoplados entre sí para dar un dataset global de ISIC [6], donde obtendríamos las siguientes clases:

Clase	2017	2018	2019	2020
Melanoma	374	1113	4522	584
Atypical melanocytic proliferation	-	-	-	1
Cafe-au-lait macule	-	-	-	1
Lentigo NOS	-	-	-	44
Lichenoid keratosis	-	-	-	37
Nevus	-	-	-	5193
Seborrheic keratosis	254	-	-	135
Solar lentigo	-	-	-	7
Melanocytic nevus	-	6705	12.875	-
Basal cell carcinoma	-	514	3323	-
Actinic keratosis	-	327	867	-
Benign keratosis	-	1099	2624	-
Dermatofibroma	-	115	239	-
Vascular lesion	-	142	253	-
Squamous cell carcinoma	-	-	628	-
Other / Unknown	1372	-	-	27.124
Total	2000	10.015	25.331	33.126

Sin embargo, sería necesario tener en cuenta la eliminación de imágenes repetidas, debido a que durante cada edición de ISIC, un número considerable de imágenes han sido incluidos en varios años. Este procedimiento engloba:

1. Eliminar las imágenes idénticas por hash. Todas las imágenes de ISIC están numeradas de forma única para facilitar la identificación de cada una de ellas. Si unimos todos los datatesets, y tomamos las repeticiones, podemos remover:

	2016	2017	2018	2019	2020
Train	291	1283	0	0	0
Test	95	594	0	0	0

Tabla 2.1: Número de imágenes duplicadas recogidas por [6]

2. Eliminación del ISIC 2018. Como éste se encuentra contenido en la composición para el año 2019, puede prescindirse totalmente de él a favor de la versión de 2019.

3. Eliminación de imágenes "downsampled" del conjunto. En los años 2019 y 2020, se añadieron imágenes de challenges anteriores con una reducción en

resolución. Para ahorar en espacio y tiempo de cómputo, pueden eliminarse las imágenes reducidas para quedarnos con una única copia de mayor calidad de la lesión, y luego realizarles manualmente un reescalado en caso de que sea necesario.

Atendiendo de nuevo a los resultados propuestos por [6], obtenemos el siguiente conjunto:

Year	Task No.	Images Removed	Images Remaining
2016	3	826	74
2017	3	801	1199
2018	3	10,015	0
2019	1	2235	23,096
2020	-	433	32,693
Total	-	14,310	57,0621

Tabla 2.2: Tabla de imágenes únicas extraída de [6]. En este caso, el autor descarta el uso del dataset de 2016 por su baja aportación

Obtendríamos un total de 57000 imágenes, los cuales podrían clasificarse, con sus respectivas clases extraídas de los metadatos. Componen, en resumen, un conjunto de datos robusto que puede formar parte del dataset de entrenamiento de este trabajo.

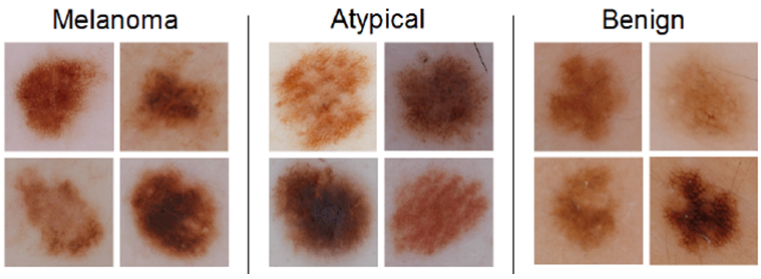

Figura 2.2: Ejemplo de imágenes de ISIC 2017 [23]

ASAN Dataset

ASAN (Seung Seog Han 2018)[15][18][19] es un conjunto de datos de origen surcoreano compuesto por lesiones malignas y benignas de la piel. Nos permite obtener un mayor grado de variedad de las imágenes, ya que el repositorio ISIC se centra sobre todo en lesiones de piel de población europea.

Tipo de lesión	Número de ejemplares
Actinic keratoses and intraepithelial carcinoma (AKIEC)	651
Basal Cell Carcinoma (BCC)	1082
Dermatofibroma (DF)	1247
Hereditary angioedema (HAO)	2715
Intraepithelial Carcinoma (IC)	918
Lentigo (LEN)	1193
Melanoma (ML)	599
Nevus (NV)	2706
Pyogenic Granuloma (PG)	375
Squamous Cell Carcinoma (SCC)	1231
Seborrhoeic Keratosis (SK)	1423
Wart	2985
Total	**17125**

Tabla 2.3: Distribución de clases de ASAN dataset

Tal y como se describe en [17] (M Goyal 2019), este dataset tiene 12 tipos de enfermedades, sumando un total de 17125 imágenes clínicas. Estas imágenes están compuestas en su mayoría por imágenes en miniatura, pero existe un repositorio con imágenes de mayor tamaño, donde sería necesario realizar tareas de segmentación- Sin embargo, dichas imágenes son de acceso restringido, y se requieren permisos especiales del hospital para acceder a ellos. Por ese motivo, tendremos únicamente en cuenta loas 17125 miniaturas.

Adicionalmente, podemos encontrar también imágenes proporcionadas por Hallym, un dataset complementario de 125 imágenes pertenecientes a lesiones de tipo melanoma cancerosas.

Las clases más destacadas de este dataset en su conjunto son la presencia de lesiones benignas de la piel, detalle que no encontramos en ISIC, y que permiten así contrastar información de la piel con lesiones benignas con la piel cancerosa. Podemos encontrar 4 clases benignas: lentigos (manchas solares fruto del envejecimiento y la exposición prolongada al sol), nevus (lunares comunes), verrugas y granulomas benignos.

Figura 2.3: Ejemplo de lunares beningnos en ASAN (Nevus)

Los resultados han sido confirmados por expertos dermatólogos y los resultados verificados en su mayoría mediante biopsia, por lo que las etiquetas asociadas a cada lesión están completamente verificadas.

El formato de las imágenes es una disposición matricial de las miniaturas, donde cada fichero que contiene las subimágenes representa en su conjunto una clase. Por desgracia, no se aporta otro tipo de información adicional más allá de la etiqueta por motivos de privacidad.

Dermnetz

Podemos encontrar extraer este dataset de un atlas online de enfermedades cutáneas recogidas de pacientes alrededor de todo el mundo. Contiene tanto lesiones benignas como malignas, existiendo además manchas y lesiones vinculadas a enfermedades infecciosas y hongos.

Existen gran cantidad de herramientas para realizar esta extracción de datos, como la que podemos encontrar en [21]. En total, se pueden obtener hasta 23000 imágenes, existiendo un total de 23 clases no balanceadas. Podemos encontrar lesiones de tipo alérgico, así como acné, dermatitis severa o celulitis.

Carece de metadatos asociados, ya que dicha información es de carácter reservado por su mantenedor. El dataset no se encuentra listo para usar de forma inmediata, ya que desde 2019, las imágenes deben ser extraídas de la propia web, pues dispone un índice donde se pueden acceder a las enfermedades de interés. El fichero contenedor del dataset fue retirado en 2019 del libre acceso, junto a sus metadatos. Es necesario

solicitar su acceso y aportar una cantidad económica.

PH2

El conjunto de datos PH2 [22] es un conjunto de 200 imágenes obtenidas gracias al hospital Pedro Hispano de Portugal. Está compuesto por imágenes de alta resolución que contienen 3 posibles casos de lesiones:

- Lunar común (Common Nevus), 80 ejemplares
- Lunar atípico (Atypical Nevus), 80 ejemplares
- Melanomas, 40 ejemplares.

Además de las 200 imágenes, podemos encontrar metadatos asociados a cada una de ellas, como el color, su extensión, textura, forma del borde, localización, entre otros. Su acceso es libre para fines académicos desde su página oficial [22], que contiene las imágenes en formato jpg, y varios ficheros .csv con la información de la imagen y su clasificación.

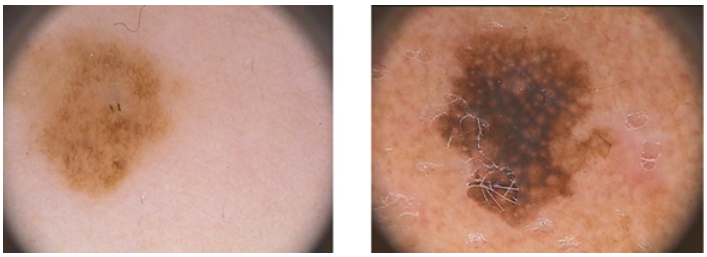

Figura 2.4: Nevus maligno y benigno en PH2

PAD-UFES 20

PAD-UFES-20 [25] se trata de un conjunto de datos recopilado de diferentes poblaciones, que contiene diagnósticos para 1.641 lesiones cutáneas únicas recopiladas, comprendiendo un total de 2.298 imágenes.

Entre sus clases, podemos encontrar tres enfermedades y tres cánceres de piel. Todos estos datos han sido recogidos y verificados mediante biopsia en un 100 % de los casos cancerosos, por lo que su diagnóstico está totalmente verificado.

Podemos encontrar, además del diagnóstico, metadatos acerca de:

- ID de paciente
- ID de lesión,

- ID de imagen
- Si la lesión benigna fue o no probada por biopsia.
- Información del paciente: fumador o no, localización de la lesión, edad, exposición a químicos, historial cancerígeno, etc.

Los datos han sido recogidos mediante teléfonos móviles en formato PNG, siendo las imágenes validadas por el Hospital Pathological Anatomy Unit of the University Hospital Cassiano Antȳnio Moraes (HUCAM) de la Federal University of Espírito Santo (Brasil). En su publicación original [25], podemos encontrar un resumen de su contenido de forma más específica:

Diagnostico	**Ejemplares**	**% biopsied**
Actinic Keratosis (ACK)	730	24.4 %
Basal Cell Carcinoma of skin (BCC)	845	100 %
Malignant Melanoma (MEL)	52	100 %
Melanocytic Nevus of Skin (NEV)	244	24.6 %
Squamous Cell Carcinoma (SCC)	192	100 %
Total	**2298**	**58.4 %**

Tabla 2.4: Tabla de casos diagnosticados en PAD-UFES20

Donde podemos apreciar que todos los casos de enfermedades cancerígenas han sido probados mediante biopsia, y el cáncer de célula basal se trata del tipo de enfermedad más frecuente.

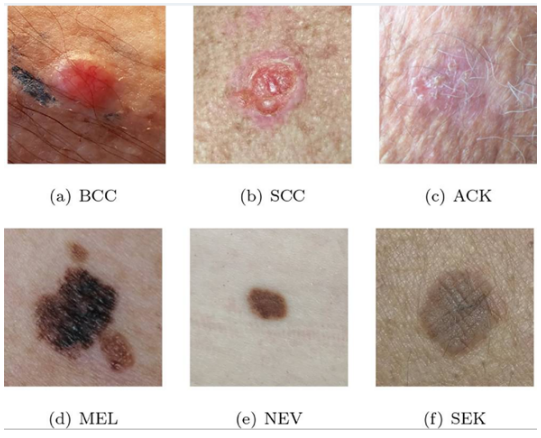

Figura 2.5: Batch de ejemplo de PAD-UFES 20 [25]

Severance

Se trata de un conjunto de imágenes de lesiones cutáneas [24] recopiladas de pacientes de Corea del Sur. Recibe dicho nombre a que los datos recopilados cuentan con la colaboración del Hospital Severance, en el mismo país.

En su variante A, que es la única disponible públicamente, podemos encontrar el diagnóstico y otra información asociada sobre 10426 imágenes, cuya valoración se encuentra entre las 38 posibles clases que contiene este conjunto de datos.

Seleccionando las 6 clases más comunes contenidas en este dataset, encontraremos que comprenden aproximadamente el 75 % del conjunto. Está compuesto por actinickeratosis (22.5 %), angiofibromas (14.4 %), angiokeratomas(13.8 %), cáncer de tipo basal cell (8.1 %), Becker nevus (7.5 %), bluenevus (6.2 %), y la enfermedad de Bowen (carcinomas)(6.1 %).

El interés en este dataset se debe a que algunas de estas clases mayoritarias, como los nevus azules y de Becker, son condiciones benignas que suelen ser retirados únicamente con fines estéticos, permitiendo complementar con el resto de los diagnósticos negativos. Este tipo de lunares son los más complejos de diagnosticar, debido a sus colores similares a un melanoma, y suelen requerir una biopsia, por lo que su diagnóstico suele alargarse.

Las imágenes se encuentran en formato matriz, por lo que es necesario proceder a su separación previo a su utilización con fines de Deep Learning:

Figura 2.6: Imágenes de ejemplo provenientes del dataset Severance

Otros datasets

Existen otros datasets ampliamente referenciados que son de acceso público. Sin embargo, en los últimos años, éstos han sido retirados y han quedado inaccesibles. Es el caso de DermQuest, un atlas virtual que contenía lesiones cutáneas y otras patologías. Ese dataset fue contenido posteriormente por Derm101, pero ambas versiones

fueron retiradas para su descarga. Alternativamente, podemos encontrar algunas de sus imágenes en los datasets SD-198 y SD-260 [26][29], pero únicamente permanece en activo el primero de ellos, bajo solicitud. En total, SD-198 contiene más de 6500 imágenes, mientras que SD-260 alcanzaba las 20000 imágenes.

En el estado del arte actual, podemos encontrar otros datasets ampliamente utilizados, como el caso de DermIS [27], un atlas online de patologías de la piel. También existen publicaciones reciente sobre nuevos conjuntos de datos utilizados de uso restringido, que permiten observar que la tendencia de investigación de este campo sigue en alza; es el caso del estudio propuesto por Papadakis et Al (2021) [28], que recoge datos sobre pacientes con melanoma de grado 3 para estudiar su evolución durante un período de 3 años, para estimar su crecimiento y potencial grosor del tumor.

Debido a las restricciones de acceso, ninguno de estos datasets será empleado como parte del entrenamiento del modelo diseñado para este estudio.

Conjunto resultado

Una vez examinados todos los conjuntos mencionados anteriormente, podemos llevar a cabo la unión de todos los datos en un único subconjunto. Esto nos permitirá conseguir un dataset completo y variado con diferentes tipos de piel y diferentes lesiones que nos permitirán identificar multitud de tipos de patologías, siendo posible ajustar el grado de granularidad en función de la agrupación o no de posibles subclases.

Inicialmente, el conjunto de datos construido contendrá todos los subtipos de lesiones cutáneas vistos, pero dispondrán de una segunda etiqueta que indicará si se trata de un caso canceroso o no, atendiendo a su subclase que lo etiqueta. Si agrupamos por lesiones benignas, cancerosas, y potencialmente cancerosas, obtenemos:

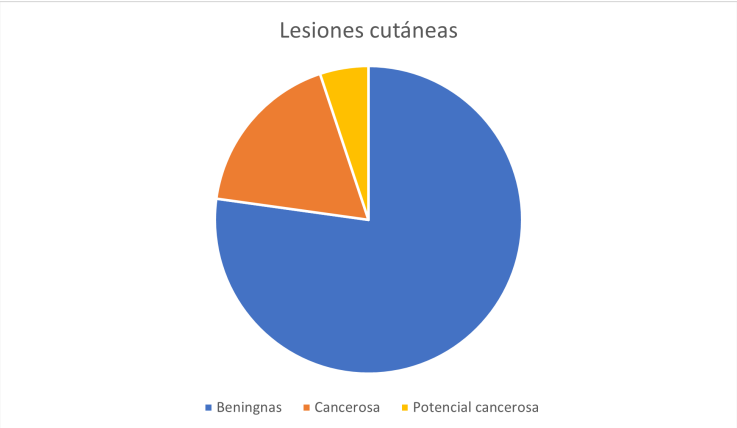

Figura 2.7: Distribución de clases

Se puede observar cómo la mayoría de imágenes disponibles engloban problemas de piel no cancerosos, mientras que el segundo tipo más común de lesión si es la cancerosa. Si atendemos a clasificar las subclases de cada tipo de patología, encontramos 52 posibles etiquetas.

2.2. Procesado de imágenes cutáneas

2.2.1. Técnicas de reducción de ruido

2.2.2. Normalización

2.2.3. Extracción de características