Universidad Nacional
de Educación a Distancia

Escuela Técnica Superior de Ingeniería Informática

Estudio de la aplicación de técnicas de aprendizaje profundo sobre resonancia magnética nuclear de mama

Autor: Javier Bel Díaz

Director: Jorge Pérez Martín
Co-director: Francisco Javier Díez Vegas

Trabajo de Fin de Máster

Máster Universitario
en Ingeniería y Ciencia de Datos

Convocatoria de Septiembre

Agradecimientos

Desde el momento en que se me propuso investigar en relación al cáncer de mama, no lo dudé. Me pareció un tema de enorme motivación por la importancia que conlleva cualquier avance o progreso logrado en este campo, pero también una gran responsabilidad, sabiendo bien que habría de poner un gran esfuerzo y trabajo sobre la mesa para que el camino recorrido mereciera la pena y sirviera en última instancia mi granito de arena para mejorar la vida de los pacientes que sufren esta enfermedad.

Desde luego que no ha sido fácil. Todo estudio tiene altibajos, momentos donde uno se siente reconfortado o recompensado por las horas de trabajo ofrecidas y otros instantes en que uno, por el contrario, siente más frustración por toparse con caminos sin salida o no alcanzar los resultados que podía esperar. No obstante, puedo afirmar que conseguí mantenerme firme en aquellos momentos más complicados para lograr finalmente presentar el estudio de este documento y en ello tienen mucho que ver personas que me han acompañado durante este recorrido y a las cuales deseo agradecer.

En primer lugar, gracias a mi familia por estar una vez más siempre disponible para escuchar mis inquietudes, desasosiegos y progresos. Gracias a vuestros consejos y ánimos he podido llegar a la meta pretendida, por lo que el resultado finalmente alcanzado se debe en gran parte a vosotros. GRACIAS.

Deseo también agradecer enormemente a mis tutores Jorge y Francisco Javier. Gracias a vuestros innumerables consejos e incalculable experiencia puestos a mi servicio, he podido pulir poco a poco y con paciencia el análisis que aquí recojo. Larga es la lista de aspectos aprendidos que me llevo de este proyecto gracias a vosotros. No me olvido tampoco de Mikel. Gracias por ayudarme en aquellos momentos delicados en que parecía que no avistaba ningún puerto concreto con los análisis realizados. GRACIAS.

En último lugar, quiero agradecer también a la institución de la UNED, la cual hace posible que estudiantes como yo tengamos la oportunidad de enrolarnos e involucrarnos en proyectos tan apasionantes como éste. GRACIAS.

Qué mejor que aplicar la ciencia para mejorar el bienestar y la vida de las personas. Si esta ilusión también te representa, te invito a que sigas leyendo este documento.

Resumen

El cáncer de mama es un tipo de cáncer que se forma en las células de las mamas. En un país de referencia como Estados Unidos, el cáncer de mama es el tipo más comúnmente diagnosticado en mujeres después del cáncer de piel.

El considerable apoyo para la concienciación y la financiación de investigaciones sobre el cáncer de mama ha ayudado a crear avances en el diagnóstico y tratamiento de esta enfermedad. Las tasas de supervivencia al cáncer de mama han aumentado y el número de muertes asociadas a esta enfermedad está disminuyendo constantemente, en gran medida debido a factores como la detección temprana, un nuevo enfoque de tratamiento personalizado y una mejor comprensión de la enfermedad.

En esta línea, son varias las técnicas existentes en la actualidad para abordar la detección y el diagnóstico del cáncer de mama, como son la mamografía y la ecografía. Además, la resonancia magnética nuclear es aplicada en este campo, especialmente en situaciones en las que el resto de técnicas no ofrecen buenos resultados.

Por otro lado, ha transcurrido tiempo desde la irrupción de la inteligencia artificial y el aprendizaje automático y su aplicación al campo médico. Así, a día de hoy, está bien contrastado el éxito que tienen ciertas arquitecturas de aprendizaje profundo en el reconocimiento de imágenes médicas, ayudando en última instancia al médico en la tarea de detección y diagnóstico de tumores.

De esta forma, la conjunción dada por la resonancia magnética nuclear aplicada al cáncer de mama y la aplicación del aprendizaje profundo al reconocimiento de dichas imágenes es la que caracterizará el estudio que se ha llevado a cabo y que se presentará en este documento. Por tanto, este estudio tratará de aplicar técnicas de aprendizaje profundo al reconocimiento de imágenes procedentes de resonancias magnéticas nucleares de mama con el objetivo de comprobar si es posible mejorar los resultados expuestos hasta hoy en el estado del arte existente.

Palabras clave: Cáncer de mama, resonancia magnética nuclear, aprendizaje profundo.

Abstract

Breast cancer is a type of cancer that forms in the cells of the breasts. In a reference country like the United States, breast cancer is the most commonly diagnosed type in women after skin cancer.

Considerable support for breast cancer awareness and research funding has helped create advances in the diagnosis and treatment of this disease. Breast cancer survival rates have increased and the number of deaths associated with the disease is steadily declining, largely due to factors such as early detection, a new personalized treatment approach and a better understanding of the disease.

In this line, there are various techniques currently available to address the detection and diagnosis of breast cancer, such as mammography and echography. In addition, nuclear magnetic resonance is applied in this field, especially in situations where other techniques do not offer good results.

On the other hand, time has passed since the irruption of artificial intelligence and machine learning and its application to the medical field. Thus, nowadays, the success of certain deep learning architectures in the recognition of medical images has been well proven, ultimately helping the doctor in the task of detecting and diagnosing tumors.

In this way, the conjunction given by nuclear magnetic resonance applied to breast cancer and the application of deep learning to the recognition of such images is what will characterize the study that has been performed and that will be presented in this document. Therefore, this study will try to apply deep learning techniques to the recognition of breast magnetic resonance images with the aim of verifying whether it is possible to improve the results presented up to date in the existing state of the art.

Keywords: Breast cancer, magnetic resonance imaging, deep learning.

Notación

OMS	Organización Mundial de la Salud
ACS	American Cancer Society / Sociedad Americana contra el Cáncer
RMN	Resonancia Magnética Nuclear
MRI	Magnetic Resonance Imaging / Resonancia Magnética Nuclear
BRCA1	Breast Cancer Gene 1 / Gen de Cáncer de Mama 1
BRCA2	Breast Cancer Gene 2 / Gen de Cáncer de Mama 2
IA	Inteligencia Artificial
CNN	Convolutional Neural Network / Red Neuronal Convolucional
TFM	Trabajo Fin de Máster
GPU	Graphics Processing Unit / Unidad Gráfica de Procesamiento
DCIS	Ductal Carcinoma in Situ / Carcinoma Ductal in Situ
T1	Tiempo de relajación longitudinal (cesión de energía de tejido)
T2	Tiempo de relajación transversal (cesión de energía de tejido)
Gd-DTPA	Gadopentetato Dimeglumina (contraste)
DWI	Diffusion Weighted Imaging / Imagen Ponderadas por Difusión
CPU	Central Processing Unit / Unidad Central de Procesamiento
ANN	Artificial Neural Network / Red Neuronal Artificial
ReLU	Rectified Linear Unit / Unidad Lineal Rectificada
EMA	Exponential Moving Average / Media Móvil Exponencial
CLAHE	Contrast Limited Adaptive Histogram Equalization / Ecualización Adaptativa del Histograma Limitada por Contraste
ADN	Ácido Desoxirribonucleico
DCE-MRI	Dynamic Contrast-Enhanced Magnetic Resonance Image / Imagen de Resonancia Magnética con Realce por Contraste Dinámico
CADe	Computer-Aided Detection / Detección Asistida por Ordenador

CADx	Computer-Aided Diagnosis / Diagnóstico Asistido por Ordenador
MIP	Maximum Intensity Projection / Proyección de Intensidad Máxima
HER2	Human Epidermal Growth Factor Receptor 2 / Receptor 2 del Factor de Crecimiento Epidérmico Humano
FGT	Fibroglandular Tissue / Tejido Fibroglandular
BPE	Background Parenchymal Enhancement / Realce Parenquimatoso de Fondo
ROC	Receiver Operating Characteristic / Característica Operativa del Receptor
AUC	Area Under the Curve / Área Bajo la Curva
EUS	Ensemble of Under-Sampled / Conjunto de Submuestreo
ROI	Region of Interest / Región de Interés
LAC	Localized Active Contour / Contorno Activo Localizado
CGN	Convolutional Gating Network / Red Convolucional de Compuertas
DTL	Deep Transfer Learning / Aprendizaje Profundo por Transferencia
AUROC	Area Under the Receiver Operating Characteristic Curve / Área Bajo la Curva ROC
DSC	Dice Similarity Coefficient / Coeficiente de Similitud de Dice
RIDER	Reference Image Database to Evaluate Therapy Response / Base de Datos de Imágenes de Referencia para Evaluar la Respuesta a la Terapia
PACS	Picture Archiving and Communication System / Sistema de Archivo y Comunicación de Imágenes
TCIA	The Cancer Imaging Archive / El Archivo de Imágenes de Cáncer
DICOM	Digital Imaging and Communications in Medicine / Imagen y Comunicación Digitales en Medicina
LCIS	Lobular Carcinoma in Situ / Carcinoma Lobular in Situ
HTML	HyperText Markup Language / Lenguaje de Marcas de Hipertexto
TPU	Tensor Processing Unit / Unidad de Procesamiento Tensorial
GB	Gigabyte
PNG	Portable Network Graphics / Gráficos de Red Portátiles

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción, motivación, objetivos y estructura del documento

1.1. Introducción y motivación

Según la Organización Mundial de la Salud (OMS), el cáncer de mama es el tumor maligno más frecuente en la población femenina (el 99 % de los casos de cáncer de mama se dan en mujeres), tanto en los países de renta baja como en los de renta alta. Su prevalencia está aumentando en las regiones de renta baja debido al aumento de la esperanza de vida y otros factores socio-económicos.

Los métodos más comunes para la detección del cáncer de mama son la exploración física, la mamografía y la ecografía. Entre éstos, la mamografía es la técnica de referencia. Los programas de cribado y diagnóstico del cáncer de mama mediante mamografías periódicas han demostrado que pueden reducir significativamente la mortalidad de esta enfermedad.

No obstante, la mamografía presenta algunos inconvenientes importantes:

- Su sensibilidad (capacidad para dar como casos positivos aquellos casos realmente enfermos, es decir, la capacidad para detectar la enfermedad) es menor con senos muy densos o en mujeres jóvenes (menores de 45 años).

- Implica la exposición a radiación ionizante, la cual aumenta el riesgo de desarrollar cáncer de mama, lo que limita la frecuencia de su aplicación.

- Para obtener un resultado adecuado hay que comprimir los senos, lo que causa dolor en algunas mujeres y disminuye la adherencia al cribado.

- Su especificidad (capacidad para dar como casos negativos aquellos casos realmente sanos, es decir, la capacidad para descartar la enfermedad en pacientes sanas) no es muy alta.

Las guías recientes de la Sociedad Americana contra el Cáncer o *American Cancer Society* (ACS) recomiendan el uso de resonancia magnética nuclear (RMN) junto a mamografía para perfiles de alto riesgo (por ejemplo, mujeres con mutaciones BRCA1 o BRCA2 o con antecedentes familiares). La RMN puede servir para tener una evaluación adicional cuando hay anomalías detectadas por la mamografía o servir de técnica principal en aquellos casos donde la mamografía no tiene suficiente precisión (alta densidad mamaria, cicatrices, anomalías pequeñas, etc.). Así, la resonancia magnética ofrece mayor sensibilidad que la mamografía (es decir, detecta más casos de cáncer), pero menor especificidad (tiene más falsos positivos). Adicionalmente, la RMN permite evaluar el tamaño y la ubicación precisa de las lesiones.

Por otro lado, la tecnología es una gran aliada de la ciencia médica, incorporando nuevas maneras de realizar diagnósticos en los que el procesamiento de grandes cantidades de datos asiste al especialista. Este es el caso de la inteligencia artificial (IA). Las herramientas ofrecidas por la IA se están aplicando a diferentes campos de la medicina, encontrando nuevas maneras de ayudar a la detección de enfermedades diversas.

En el ámbito de los algoritmos de procesamiento de imágenes, la combinación de la IA con una disponibilidad de mayor cantidad de datos y mayor capacidad de cómputo nos ha permitido resolver retos que no eran viables mediante las técnicas clásicas deterministas de visión por computador centradas en métodos como los filtrados, gradientes, contornos adaptables, crecimiento de regiones o la creación de atlas, entre otros.

El motivo por el que este cambio ha supuesto una disrupción positiva reside en gran medida en las redes neuronales convolucionales (*Convolutional Neural Networks*, CNNs). Las CNNs están especialmente adaptadas para el manejo de datos espaciales, sobre todo de imágenes, tanto bidimensionales como tridimensionales y han tenido también un gran impacto en otros dominios, como es el caso del reconocimiento de voz. Las arquitecturas de redes neuronales pueden incluir un elevado número de capas ocultas entre la entrada y la salida, que hacen que a esta tecnología se la conozca con el nombre de aprendizaje profundo o *deep learning*. Así, por ejemplo, para el campo de la radiología, el aprendizaje profundo se ha aplicado principalmente para la detección de lesiones y para la clasificación de las imágenes en patologías y sus subgrupos, así como para la segmentación automática de órganos. Los primeros avances, cuando no había disponibilidad de muchos datos, fueron posibles en parte gracias a la transferencia de conocimiento o *transfer learning*, por medio de redes ya preentrenadas para el reconocimiento y análisis de imágenes de la vida cotidiana que se han adaptado a la evaluación de imágenes radiológicas. En la actualidad, donde resulta notable el aumento de datos disponibles etiquetados, se están ya entrenando redes de manera exclusiva con datos radiológicos, lo que permite una mejora significativa en rendimiento.

Con todo esto, a modo de resumen, las razones que motivan la realización de este Trabajo Fin de Máster (TFM) son:

- El cáncer de mama constituye una de las enfermedades con mayor incidencia en el mundo, por lo que se desea contribuir a desarrollos científicos que puedan ser aplicados

para mejorar la detección y diagnóstico de esta enfermedad y, por tanto, aumentar las posibilidades de supervivencia de estas pacientes.

- Las técnicas de aprendizaje profundo ofrecen enormes posibilidades para el reconocimiento de imágenes médicas, por lo que se desea hacer uso de las mismas para alcanzar la primera motivación comentada.

- En cuanto a estas motivaciones, es importante comentar también que este estudio se enmarca dentro de un proyecto orientado principalmente a las técnicas de mamografía, ecografía y termografía. Puesto que la aplicación del aprendizaje profundo a la resonancia magnética nuclear tiene menos recorrido que la correspondiente a estas otras técnicas de obtención de imágenes, se decide abordar la resonancia magnética al constituir un campo más inexplorado con mayor potencial de mejora. Por tanto, conviene aclarar que este estudio es un trabajo exploratorio, preliminar y prospectivo, cuyas conclusiones servirán para el establecimiento de líneas futuras del proyecto.

Por tanto, mediante el desarrollo de este Trabajo Fin de Máster, se estudiará y probará la aplicación de diversas técnicas de aprendizaje profundo sobre resonancias magnéticas nucleares de mama, evaluando cada modelo de acuerdo a un conjunto de métricas que permita comparar la calidad de los mismos con los ya existentes.

1.2. Alcance y objetivos del estudio

Tras la introducción y los aspectos que motivan el presente estudio, se presenta ahora el alcance del mismo y los objetivos concretos que desearán completarse.

Este Trabajo Fin de Máster busca principalmente aplicar modelos propios de técnicas de aprendizaje profundo a las imágenes obtenidas de resonancias magnéticas nucleares de mama. La actuación o desempeño alcanzados por los modelos desarrollados se pretenden comparar con los correspondientes a otros modelos procedentes del estado del arte para dilucidar si ha sido posible mejorar dichos modelos ya existentes.

Como todo estudio, el presentado en este documento se caracteriza por una serie de limitaciones que perfilan el alcance del mismo. Estas limitaciones se tratarán a lo largo de esta memoria y se resumirán en la discusión de la misma.

Los objetivos planteados en este estudio constituyen preguntas que desean responderse. De esta forma, el objetivo principal de este trabajo viene dado por la siguiente pregunta a la cual se le pretenderá dar respuesta:

¿Es posible inferir con un modelo de aprendizaje profundo una característica elegida (se expondrá más adelante) a partir de las imágenes presentes en el conjunto de datos?

A partir del objetivo principal de arriba, se suceden diferentes sub-objetivos u objetivos específicos. Éstos vienen dados de nuevo por cuestiones que buscan responderse por medio de los análisis efectuados en este trabajo. Estos objetivos específicos se muestran a continuación:

- ¿Qué modelo diagnostica mejor: uno preentrenado o no?
- ¿Se consigue una mejora notable en la actuación del modelo al usar técnicas de pre-procesamiento de datos?
- ¿Hay sobreajuste por parte de los modelos? En caso afirmativo, ¿es posible corregirlo?

Para responder a estas preguntas de investigación y cumplir los objetivos, se ha propuesto la siguiente metodología para este TFM:

1. Se comenzará abordando el estudio del marco teórico asociado a este proyecto.
2. Posteriormente, se analizará el estado del arte para ver qué tendencias existen en el campo que nos ocupa, además de los resultados alcanzados por los modelos propuestos hasta hoy.
3. Tras analizar el estado del arte, se abordará la tarea de obtención de los datos con los que posteriormente serán entrenados los modelos desarrollados. Para ello, se buscarán y analizarán bases de datos en abierto que ofrezcan RMN de mama. Una vez alcanzados estos datos, serán convenientemente descritos y analizados.
4. Cuando se disponga de un conjunto sólido de datos, se procederá a la creación de diferentes modelos de redes neuronales. La actuación de estos modelos será comparada por medio de varias métricas de evaluación. A partir de esta comparación, se escogerá el modelo con mejor desempeño y, con este modelo, se llevará a cabo una serie de experimentos que tendrán como finalidad el comprobar la influencia que tiene sobre el modelo la aplicación de ciertas técnicas, como son el aumento de datos y los filtros de imágenes.
5. En base a todo lo anterior, se expondrán la discusión y las conclusiones alcanzadas tras todo el estudio realizado y se propondrán líneas de continuación/mejora para este proyecto.

Con todo esto, el desarrollo de este TFM se dividirá en las siguientes etapas o fases:

- Etapa de documentación: Se estudiará el estado del arte de la aplicación de técnicas de aprendizaje profundo con RMN para la adquisición de ideas y tendencias que vienen siendo aplicadas y que se tomarán como referencia en la siguiente etapa. Se partirá del estudio de Sheth and Giger [2020] y se realizará, además, una búsqueda bibliográfica

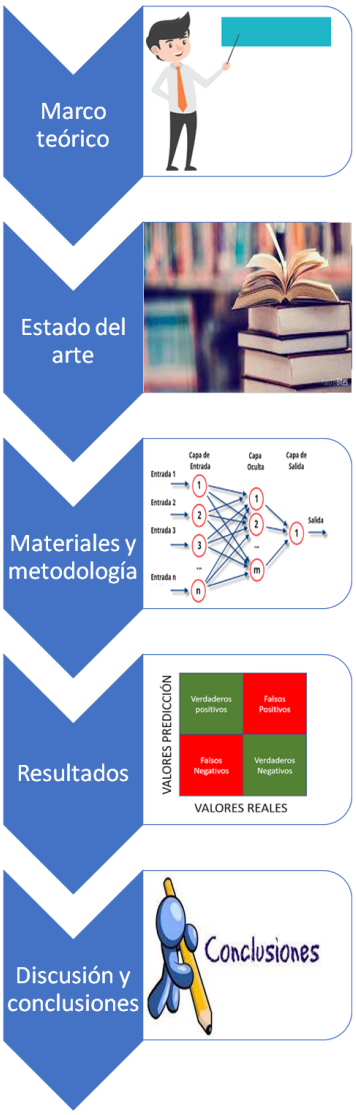

Figura 1.1: Estructura del presente documento.

con artículos recientes que analicen la RMN con técnicas de aprendizaje profundo. Se prevé una duración de unas 100 horas sobre el total de 360 horas supuesto para la dedicación al TFM (12 créditos por 30 horas de cada crédito ECTS).

- Etapa de desarrollo: En base a lo encontrado en la literatura, se escogerá una tarea concreta que sea de interés y se construirán modelos que permitan resolverla. Una vez se tuvieran modelos con buena actuación, se llevaría a cabo una búsqueda de hiperparámetros y se comprobaría su bondad mediante un conjunto de datos separados (conjunto de test). La duración estimada es de 200 horas.

- Redacción de la memoria, preparación de la presentación y defensa del trabajo realizado ante el tribunal del TFM. Se emplearán unas 60 horas en estas tareas.

Para completar la consecución de los objetivos introducidos anteriormente, se hace uso

de diversas herramientas. Así, por ejemplo, para la generación de los modelos y el desarrollo de los experimentos se emplea Google Colab (https://colab.research.google.com/), el cual permite la ejecución de celdas de código Python desde el navegador. Google Colab es un servicio gratuito de notebook alojado de Jupyter que no requiere configuración y que provee acceso gratuito a recursos computacionales, como las GPU de NVIDIA K80, T4, P4 y P100.

Se empleará, a su vez, Google Drive (https://drive.google.com/) como repositorio del proyecto, es decir, se usará para el almacenamiento de los datos con los que serán entrenados y evaluados los modelos desarrollados. Este servicio también es gratuito, por lo que el coste de este TFM será el relativo a los recursos humanos de los tutores y el estudiante.

Además, se empleará la plataforma de Overleaf, la cual constituye un entorno colaborativo y será usada para la redacción de la memoria haciendo uso de Latex.

1.3. Estructura del documento

Considerando todo lo expuesto previamente, se ha decidido estructurar el presente documento de la siguiente manera:

1. Explicación del marco teórico.
2. Presentación del estado del arte.
3. Descripción de los materiales y la metodología aplicados.
4. Exposición de los resultados alcanzados.
5. Presentación de la discusión, las conclusiones obtenidas y propuesta de líneas de continuación/mejora.

A modo visual, esta estructura de documento aparece representada en la Figura 1.1.

Capítulo 2

Marco teórico

En este capítulo se va a exponer el marco teórico sobre el que se sustenta todo el estudio llevado a cabo y presentado en este documento. Así, los bloques fundamentales que componen este marco teórico se listan a continuación:

- Cáncer de mama
- Resonancia magnética nuclear (RMN)
- Aprendizaje profundo
- Red neuronal convolucional (CNN)
- Mecanismos de regularización
- Grado de Nottingham

2.1. Cáncer de mama

El cáncer de mama[1] es una enfermedad en la que las células del tejido mamario crecen sin control. El cáncer de mama se puede presentar tanto en hombres como en mujeres. No obstante, es mucho más común en las mujeres. Existen diferentes tipos de cáncer de mama, dependiendo de qué células de la mama tornan en células tumorales.

El cáncer de mama puede comenzar en diferentes partes de la mama. Una mama se compone de tres partes principales: lóbulos, conductos y tejido conectivo. Los lóbulos son las glándulas que producen la leche, los conductos son tubos que llevan la leche al pezón y el tejido conectivo (que consta de tejido fibroso y graso) rodea y mantiene todo unido. Estas diferentes partes de la mama aparecen representadas en la Figura 2.1[1]. La mayoría de los cánceres de mama comienzan en los conductos o en los lóbulos.

[1] https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm

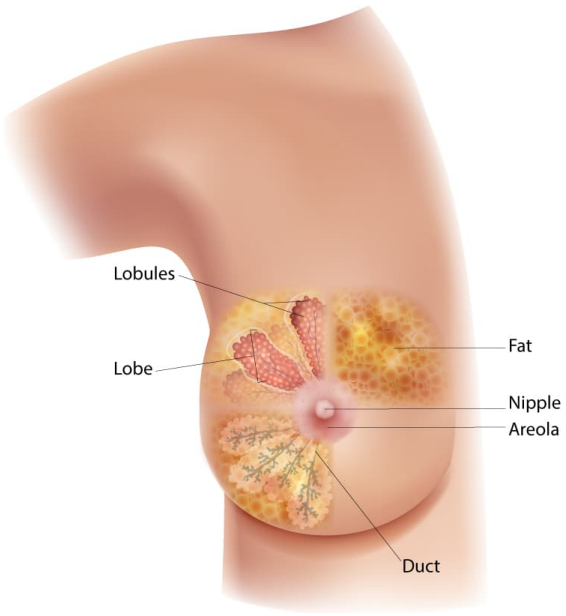

Figura 2.1: Partes principales de la mama.

El cáncer de mama se puede diseminar fuera del seno a través de los vasos sanguíneos y los vasos linfáticos. Cuando el cáncer de mama se propaga a otras partes del cuerpo, se dice que se ha producido metástasis.

Los tipos más comunes de cáncer de mama son:

- Carcinoma ductal invasivo. Las células cancerosas comienzan en los conductos y luego crecen fuera de los conductos hacia otras partes del tejido mamario. Las células cancerosas invasivas también pueden diseminarse o hacer metástasis a otras partes del cuerpo.

- Carcinoma lobulillar invasivo. Las células cancerosas comienzan en los lóbulos y luego se diseminan desde los lóbulos hasta los tejidos del seno que están cerca. Estas células cancerosas invasivas también pueden diseminarse a otras partes del cuerpo.

Hay otros tipos de cáncer de mama menos comunes, como la enfermedad de Paget, el cáncer de mama medular, el mucinoso y el inflamatorio. Además, el carcinoma ductal in situ (DCIS) es una enfermedad de la mama que puede provocar cáncer de mama invasivo. En este caso, las células cancerosas sólo se encuentran en el revestimiento de los conductos y no se han propagado a otros tejidos de la mama.

2.2. Resonancia magnética nuclear (RMN)

De entre las técnicas de detección del cáncer de mama existentes en la actualidad, el presente estudio se centra concretamente en la resonancia magnética nuclear (RMN).

Heywang et al. [1986] y Kaiser and Zeitler [1989] introdujeron esta técnica de forma independiente en la década de 1980. La resonancia magnética es un método de estudio para generar imágenes detalladas de órganos y tejidos por medio de un campo magnético y ondas de radio que cambian rápidamente. La resonancia magnética es una interacción entre un campo magnético externo, ondas de radiofrecuencia y núcleos atómicos. Cuando un cuerpo es sometido a un campo magnético y posteriormente se le estimula mediante ondas electromagnéticas, se logra la resonancia de los núcleos de sus átomos.

La base de la obtención de imágenes [Costa and Soria, 2021] es la medición de la energía liberada por los núcleos atómicos y el tiempo en que vuelven a su estado de relajación una vez que dejan de estar estimulados. La tasa de cesión de energía viene determinada por las propiedades intrínsecas de relajación de cada tejido, caracterizadas por los tiempos de relajación longitudinal (T1) y relajación transversal (T2). El T1 representa la recuperación de la magnetización longitudinal en la dirección del campo magnético principal. El T2 refleja la pérdida de la magnetización en el plano transversal, perpendicular al eje del campo. Las sustancias que tienen un T1 largo (por ejemplo, los líquidos) aparecerán oscuras en las imágenes potenciadas en T1, mientras que aquéllas con un T1 corto (tejidos grasos) mostrarán alta intensidad de señal. Por el contrario, en las imágenes potenciadas en T2, una sustancia con un T2 largo (líquido) aparecerá brillante. Las imágenes utilizadas habitualmente están potenciadas en T1 o en T2.

El contraste más usado es un fármaco que acorta el T1, denominado gadopentetato dimeglumina (o Gd-DTPA), puesto que contiene gadolinio, el cual es un agente paramagnético (es decir, que tiene átomos con electrones no apareados en sus capas externas).

Con todo esto, diagnósticos precisos son posibles de alcanzar de tejidos, órganos y estructuras vasculares.

Basándose en lo comentado anteriormente y tal y como es explicado por Mann et al. [2019], la resonancia magnética con material de contraste evalúa la permeabilidad de los vasos sanguíneos mediante el uso de un agente de contraste intravenoso que acorta el tiempo T1 local, lo que lleva a una señal más alta en las imágenes ponderadas en T1 [Knopp et al., 1999]. El principio subyacente es que la neoangiogénesis[2] conduce a la formación de vasos con fugas que permiten una extravasación más rápida de los agentes de contraste [Carmeliet and Jain, 2000], lo que conduce a un realce local rápido. A pesar de las mejoras en la técnica de resonancia magnética de mama, este principio sigue siendo la base de todos los protocolos clínicos de resonancia magnética. Sin embargo, la mayoría de los protocolos hoy en día son multiparamétricos [Marino et al., 2018, Rahbar and Partridge, 2016]. De esta forma, la resonancia magnética de mama ha evolucionado de una técnica principalmente potenciada con contraste a una técnica multiparamétrica en la que se realizan rutinariamente imágenes

[2]La neoangiogénesis es la formación de nuevos vasos sanguíneos a partir de otros preexistentes. Los tumores malignos al crecer segregan unas sustancias capaces de hacer ramificarse a los vasos sanguíneos de alrededor. Esto lo hacen porque requieren los nutrientes aportados por la sangre para seguir desarrollándose.

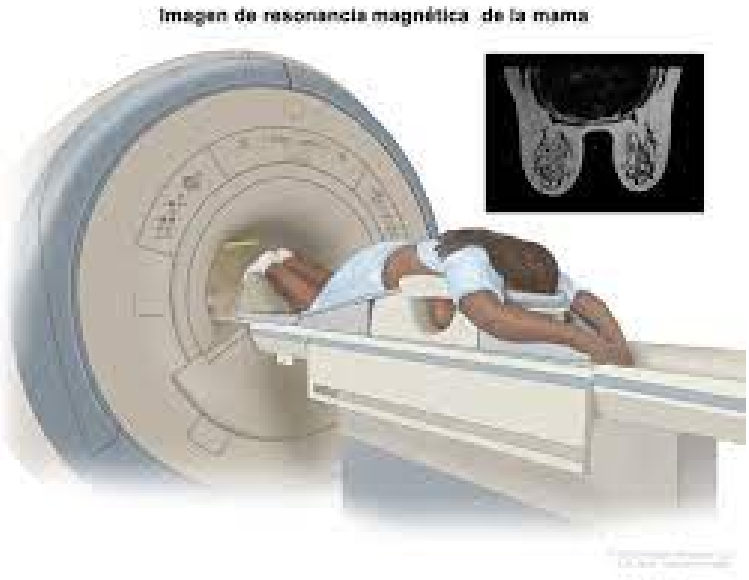

Figura 2.2: Resonancia magnética de mama.

ponderadas en T2 y potenciadas en difusión (DWI). Aun así, la base para cualquier protocolo de resonancia magnética es una secuencia dinámica potenciada en T1 con contraste.

La utilización de una bobina dedicada es obligatoria para obtener imágenes de calidad diagnóstica. Las mujeres se acuestan en posición decúbito prono[3] con las mamas libres introducidas en los huecos de la bobina. Este diseño permite que el tejido mamario se extienda, lo que facilita la detección de anomalías y evita los artefactos de movimiento inducidos por la respiración [Konyer et al., 2002, Yeh et al., 2014]. Una imagen representativa de cómo se lleva a cabo la obtención de la resonancia magnética de mama aparece en la Figura 2.2 [4].

Tal y como afirman Bevers et al. [2009], la resonancia magnética tiene una mayor sensibilidad para detectar el cáncer de mama que la mamografía, aunque la especificidad es menor, lo cual resulta en una tasa más alta de resultados falsos positivos, como se concluye en [Lord et al., 2007]. Además, las microcalcificaciones no son detectables con la resonancia magnética [Mann et al., 2008, Schnall and Orel, 2006]. Por lo tanto, se necesita una cuidadosa selección de pacientes para los exámenes de detección adicionales con resonancia magnética.

Aunque la evidencia actual no respalda el uso de la resonancia magnética de mama para evaluar a las mujeres con un riesgo medio de desarrollar cáncer de mama (ya que ofrece demasiados falsos positivos y no resulta efectivo en términos de coste), varios estudios han demostrado los beneficios de la detección mediante esta técnica en mujeres con una predisposición genética al cáncer de mama. Esto fue concluido, por ejemplo, por Kuhl et al. [2005], Warner [2008] y Lehman and Smith [2009]. La Sociedad Americana contra el Cáncer

[3]Posición anatómica del cuerpo humano tendido boca abajo con los miembros superiores extendidos y el cuello en posición neutra.

[4]https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/imagen-por-resonancia -magnetica-nuclear

(ACS) publicó directrices que recomiendan el uso de RMN de mama como complemento de la mamografía para la detección en ciertas poblaciones de mujeres con alto riesgo de desarrollar cáncer de mama [Saslow et al., 2007].

Se recomienda una resonancia magnética anual como complemento de la mamografía de detección y la examinación clínica de mama para mujeres de 25 años o más con una predisposición genética o antecedentes familiares importantes de cáncer de mama. También se recomienda la consideración de una resonancia magnética anual para las mujeres que tienen un riesgo de por vida de desarrollar cáncer de mama superior al 20 %, según lo definido por modelos que se basan en gran medida en los antecedentes familiares, tal y como se describe en las pautas de la ACS [Saslow et al., 2007]. Este consenso acerca de la necesidad de la RMN como examinación adicional se basa en múltiples estudios que demostraron que esta técnica identificó la enfermedad en una etapa más temprana que la mamografía y que, la RMN y la mamografía combinadas, se asocian con mejores tasas de supervivencia [Warner et al., 2008, Evans et al., 2016].

2.3. Aprendizaje profundo

Hoy en día, la inteligencia artificial, el aprendizaje automático y el aprendizaje profundo son tres términos populares que a veces se usan indistintamente para describir sistemas o software que se comportan de manera inteligente. En la Figura 2.3 [Sarker, 2021a] se ilustra la posición del aprendizaje profundo, comparándolo con el aprendizaje automático y la inteligencia artificial. De acuerdo con dicha figura, el aprendizaje profundo es parte del aprendizaje automático, así como parte también del área amplia de inteligencia artificial. En general, la inteligencia artificial incorpora el comportamiento humano y la inteligencia a las máquinas o sistemas [Sarker et al., 2021], mientras que el aprendizaje automático es el método para aprender de los datos o la experiencia [Sarker, 2021b], que automatiza la construcción de modelos analíticos. El aprendizaje profundo también representa métodos de aprendizaje a partir de datos donde el cálculo se realiza a través de redes neuronales y procesamiento de múltiples capas.

Teniendo más claras las diferencias entre aprendizaje profundo y los otros dos conceptos de inteligencia artificial y aprendizaje automático, el aprendizaje profundo se caracteriza por los siguientes aspectos:

- Generalmente depende de una gran cantidad de datos para construir un modelo enmarcado en un dominio particular de problema. La razón es que, cuando el volumen de datos es pequeño, los algoritmos de aprendizaje profundo a menudo funcionan mal [LeCun et al., 2015].

- Los algoritmos de aprendizaje profundo requieren grandes operaciones computacionales cuando se entrena un modelo con grandes conjuntos de datos. Cuanto más grandes

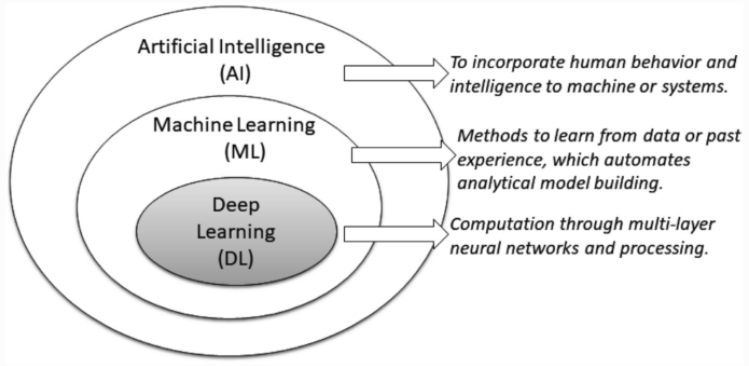

Figura 2.3: Marco donde se encuadra el aprendizaje profundo.

sean los cálculos, mayor será la ventaja de una GPU sobre una CPU. Por lo tanto, el aprendizaje profundo se basa más frecuentemente en máquinas de alto rendimiento con GPU que en métodos estándar de aprendizaje automático [Coelho et al., 2017, Xin et al., 2018].

- La ingeniería de características es el proceso de extraer características (características, propiedades y atributos) de datos sin procesar utilizando el conocimiento del dominio. Una distinción fundamental entre aprendizaje profundo y otras técnicas de aprendizaje automático es el intento de extraer características de alto nivel directamente de los datos [Deng and Yu, 2014, Sarker, 2021b]. Por lo tanto, el aprendizaje profundo reduce el tiempo y el esfuerzo necesarios para construir un extractor de características para cada problema.

- En general, entrenar un algoritmo de aprendizaje profundo lleva mucho tiempo de computación debido a una gran cantidad de parámetros en el algoritmo. Por lo tanto, el proceso de formación del modelo lleva más tiempo. Durante las pruebas, los algoritmos de aprendizaje profundo tardan muy poco tiempo en ejecutarse [Xin et al., 2018], en comparación con ciertos métodos de aprendizaje automático.

- En cuanto a la interpretabilidad, es difícil explicar cómo se obtiene un resultado de aprendizaje profundo, pareciéndose a una "caja negra". Por otro lado, una parte de los algoritmos de aprendizaje automático (como la regresión lineal, el árbol de decisión o los k-vecinos más próximos) tienden más a estar basados en reglas [Sarker, 2021b] que proporcionan normas lógicas explícitas para la toma de decisiones y que son fácilmente interpretables para las personas.

El modelado de aprendizaje profundo es extremadamente útil cuando se tiene una gran cantidad de datos debido a su capacidad para procesar un alto número de características para construir un modelo efectivo basado en datos. En términos de desarrollo y entrenamiento de modelos de aprendizaje profundo, se basa en operaciones con matrices y tensores ejecutadas

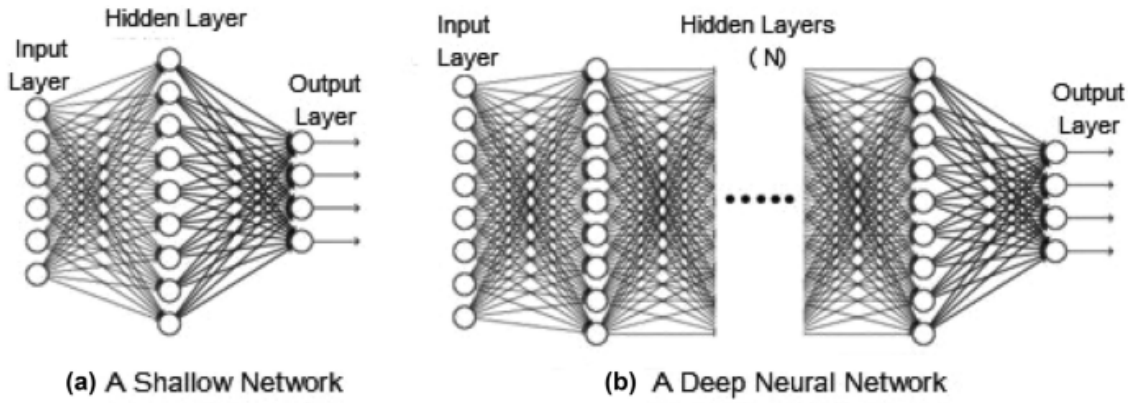

Figura 2.4: Comparación de red neuronal poco profunda (izquierda) con red neuronal más propia del aprendizaje profundo al tener más capas ocultas (derecha).

en paralelo, así como gradientes de cálculo y optimización. Varias bibliotecas y recursos de aprendizaje profundo [Géron, 2019] como PyTorch (con una API de alto nivel llamada Lightning) y TensorFlow (que también incorpora Keras como una API de alto nivel) ofrecen estas utilidades principales que incluyen muchos modelos preentrenados, así como muchas otras funciones necesarias para la implementación y construcción de modelos de aprendizaje profundo.

Entrando seguidamente más de lleno en la estructura básica propia del aprendizaje profundo, es interesante mencionar que la inspiración para las redes neuronales artificiales (ANNs), o simplemente redes neuronales, surgió de la admiración por la forma en que el cerebro humano calcula procesos complejos [Montesinos et al., 2022, McCulloch and Pitts, 1943]. Las ANNs son máquinas implementadas mediante programas informáticos para realizar tareas específicas imitando el funcionamiento del cerebro humano y que están formadas por cientos o incluso miles de millones de neuronas artificiales o unidades de procesamiento. La red neuronal artificial se implementa mediante el desarrollo de un algoritmo de aprendizaje computacional que no necesita programar todas las reglas, ya que es capaz de construir sus propias reglas de comportamiento a través de lo que solemos denominar "experiencia". La implementación práctica de las redes neuronales es posible debido a que son sistemas de computación masivamente paralelos formados por una gran cantidad de unidades básicas de procesamiento (neuronas) que están interconectadas y aprenden de su entorno, y los pesos sinápticos capturan y almacenan las fortalezas de las neuronas interconectadas. El trabajo del algoritmo de aprendizaje profundo consiste en modificar los pesos sinápticos de la red de forma secuencial y supervisada para alcanzar un objetivo específico [Haykin, 2009]. Existe evidencia de que las neuronas que trabajan juntas pueden aprender relaciones complejas de entrada y salida lineales y no lineales mediante el uso de procedimientos de entrenamiento secuencial.

En general, el trabajo de un modelo de red neuronal artificial lo realizan, tal y como se ha dicho, elementos simples llamados neuronas. Las señales se transmiten entre las neuronas

a través de enlaces de conexión. Cada enlace de conexión tiene un peso asociado que, en una red neuronal típica, multiplica la señal transmitida. Cada neurona aplica una función de activación (generalmente no lineal) a las entradas de la red (suma de las señales de entrada ponderadas) para determinar su signo correspondiente.

El aprendizaje profundo realmente se define como una generalización de la ANN donde se usa más de una capa oculta, lo que implica que se usan más neuronas para implementar el modelo. Por esta razón, una red neuronal artificial con múltiples capas ocultas se denomina red neural profunda (DNN) y la práctica de entrenar este tipo de redes se denomina aprendizaje profundo. Por tanto, las técnicas de redes neuronales profundas suelen considerar varias capas de etapas de procesamiento de información en estructuras jerárquicas para aprender. Una red neuronal profunda típica contiene varias capas ocultas, además de las capas de entrada y salida. La Figura 2.4 [Sarker, 2021a] muestra una estructura general de una red neuronal profunda (donde si el número de capas ocultas se denota con N, en una red neuronal profunda se tiene $N \geq 2$) en comparación con una red poco profunda que contiene una única capa oculta.

2.4. Red neuronal convolucional (CNN)

Una red neuronal convolucional, o CNN, es una red neuronal de aprendizaje profundo diseñada para procesar matrices estructuradas de datos, tales como imágenes. Las redes neuronales convolucionales se utilizan ampliamente en la visión por computadora y se han convertido en el estado del arte para muchas aplicaciones visuales, como la clasificación de imágenes.

Las redes neuronales convolucionales son muy buenas para detectar patrones en una imagen de entrada, como líneas, gradientes, círculos o incluso ojos y caras. Es esta propiedad la que hace que las redes neuronales convolucionales sean tan poderosas para la visión artificial. A diferencia de los algoritmos de visión por computadora anteriores, las redes neuronales convolucionales pueden operar directamente en una imagen sin procesamiento previo.

La red neuronal convolucional o CNN, propuesta originalmente por Lecun et al. [1998], es un modelo de red neuronal con tres ideas arquitectónicas principales: campos receptivos locales, pesos compartidos y mapas de características. La red se diseñó inicialmente para el reconocimiento de patrones de imágenes bidimensionales y se ha convertido en una de las principales arquitecturas en aprendizaje profundo. La CNN tiene muchas fortalezas. Primero, la extracción y clasificación de características están integradas en una estructura y son totalmente adaptables. En segundo lugar, la red extrae características de imágenes 2D a escalas diádicas crecientes. Tercero, es relativamente robusta frente al ruido de la imagen y las distorsiones geométricas locales.

Una CNN consta de tres tipos principales de capas: (i) capas de convolución, (ii) capas de submuestreo o agrupación y (iii) una capa de salida. Las capas de red están dispuestas en

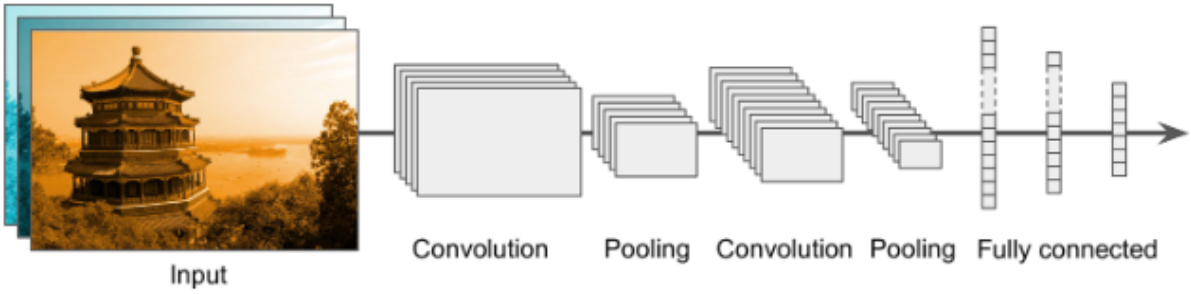

Figura 2.5: Ejemplo de arquitectura empleada por una CNN.

una estructura de avance: cada capa de convolución va seguida de una capa de submuestreo y la última capa de convolución va seguida de la capa de salida. Las capas de convolución y submuestreo se consideran capas 2D, mientras que la capa de salida se considera una capa 1D. En CNN, cada capa 2D tiene varios planos. Un plano consta de neuronas que están dispuestas en una matriz bidimensional. La salida de un plano se denomina mapa de características.

El algoritmo de entrenamiento por lotes basado en la retropropagación o RPROP o el algoritmo de descenso del gradiente estocástico son algoritmos eficientes para entrenar una CNN. Así, por ejemplo, en el algoritmo RPROP, el paso de aprendizaje se incrementa o se reduce, dependiendo únicamente del signo del gradiente del error. Se ha demostrado que este algoritmo converge más rápido en comparación con el algoritmo de descenso de gradiente estándar. RPROP funciona bien incluso cuando el gradiente tiene magnitudes muy pequeñas. Además, es computacionalmente eficiente porque sólo se requiere la derivada de primer orden de la función de error.

Un ejemplo de arquitectura empleada por una CNN se representa en la Figura 2.5 [Géron, 2019].

Como ya se ha comentado arriba, la arquitectura de una red neuronal convolucional es una red neuronal de avance de múltiples capas, como la capa de entrada, capas convolucionales, capas de agrupación y capas completamente conectadas. La capa convolucional aplica filtros a la imagen de entrada para extraer características, la capa de agrupación reduce la muestra de la imagen para reducir el cálculo y la capa completamente conectada hace la predicción final. Es este diseño secuencial el que permite que las redes neuronales convolucionales aprendan características jerárquicas. La red aprende los filtros óptimos a través de los algoritmos de retropropagación y descenso de gradiente ya introducidos.

La capa convolucional constituye el bloque de construcción central de una CNN y es donde ocurre la mayoría de los cálculos. Requiere algunos componentes, los cuales vienen dados por los datos de entrada, un filtro y un mapa de características. Como dato de entrada se puede tener, por ejemplo, una imagen. También se tendría un detector de características, conocido como kernel o filtro, que se moverá a través de los campos receptivos de la imagen

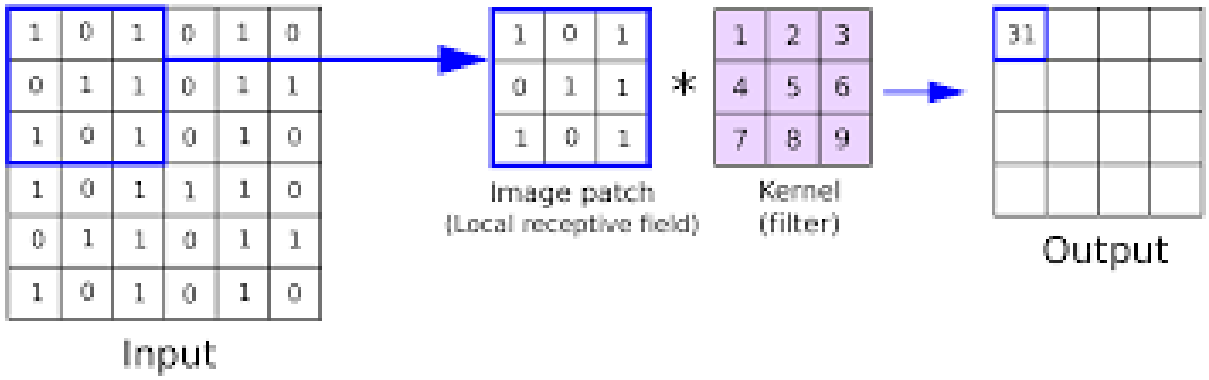

Figura 2.6: Operación de convolución.

verificando si la característica está presente. Este proceso se conoce como convolución.

El detector de características es una matriz bidimensional de pesos, la cual representa parte de la imagen. Si bien pueden variar en tamaño, las dimensiones del filtro suelen venir dadas por una matriz de 3×3; esto también determina el tamaño del campo receptivo. Luego, el filtro se aplica a un área de la imagen y se calcula un producto escalar entre los píxeles de entrada y el filtro. Este producto escalar se introduce posteriormente en una matriz de salida. Seguidamente, el filtro se desplaza según un paso o *stride*, repitiendo el proceso hasta que el filtro ha barrido toda la imagen. El resultado final de la serie de productos escalares de la entrada y el filtro es conocido como mapa de características. Esta operación de convolución es representada en la Figura 2.6 [5] y la Figura 2.7 [6].

Después de cada operación de convolución, una CNN aplica una función de activación al mapa de características. Esta función de activación tiene el efecto de introducir no linealidad en el modelo. Una función de activación común es la función ReLU, también conocida como unidad lineal rectificada, que es lo mismo que tomar la componente positiva de la entrada.

Otra capa de convolución puede seguir a la capa de convolución inicial. Cuando esto sucede, la estructura de la CNN puede volverse jerárquica, ya que las capas posteriores pueden ver los píxeles dentro de los campos receptivos de las capas anteriores. De esta manera, la arquitectura de la CNN con varias de estas capas permite que la red se concentre en características pequeñas de bajo nivel en la primera capa oculta, luego las ensambla en características más grandes de nivel superior en la siguiente capa oculta y así sucesivamente. Esta estructura jerárquica es común en las imágenes del mundo real, lo cual es una de las razones por las que las CNNs funcionan tan bien para el reconocimiento de imágenes.

Con todo esto, los parámetros principales a especificar en relación a una capa convolucional son:

[5] https://anhreynolds.com/blogs/cnn.html
[6] https://keepcoding.io/blog/tipos-capas-red-neuronal-convolucional/

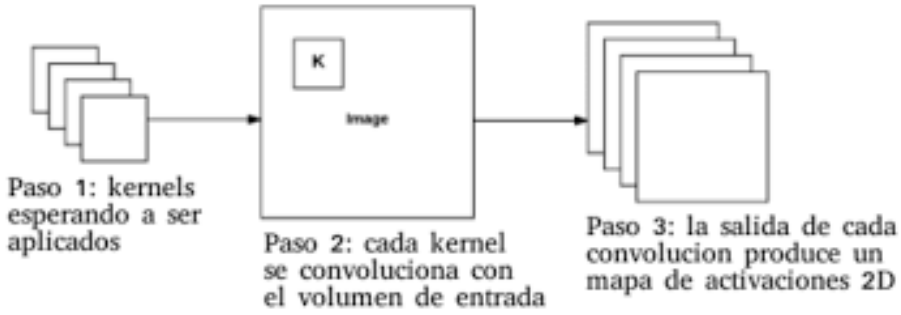

Figura 2.7: Esquema del funcionamiento de una capa convolucional.

- Número de filtros. Se establece el número de filtros o kernels de convolución a aplicar en esa capa sobre la imagen. Una capa llena de neuronas que usa el mismo filtro genera un mapa de características, el cual resalta las áreas en una imagen que activan más el filtro.
- Tamaño o dimensiones del filtro.
- Función de activación.
- Padding. Para que una capa tenga el mismo alto y ancho que la capa anterior, es común agregar ceros alrededor de las entradas. Esto es lo que se conoce como *zero padding*.

La siguiente capa característica de una CNN es la de agrupamiento o *pooling*. El objetivo de esta capa es el de submuestrear (es decir, reducir) la imagen de entrada para reducir la carga computacional, el uso de memoria y la cantidad de parámetros (lo que limita el riesgo de sobreajuste). De forma similar a la capa convolucional, cada neurona en una capa de pooling está conectada a las salidas de un número limitado de neuronas en la capa anterior ubicadas dentro de un pequeño campo receptivo. Además, la operación de agrupamiento barre con un filtro a lo largo de toda la entrada, pero la diferencia es que este filtro no tiene ningún peso. En cambio, el filtro aplica una función de agregación a los valores dentro del campo receptivo, poblando la matriz de salida. Hay dos tipos principales de agrupación: el máximo (a medida que el filtro se mueve a través de la entrada, selecciona el píxel con el valor máximo para enviarlo a la matriz de salida) y el promedio (a medida que el filtro se mueve a través de la entrada, calcula el valor promedio dentro del campo receptivo para enviarlo a la matriz de salida).

Los parámetros a definir para esta capa son su tamaño, el denominado *stride* y el tipo de *padding*. Una representación gráfica de la capa de máximo agrupamiento aparece en la Figura 2.8 [Géron, 2019].

Otra capa importante es la que representa una capa de neuronas densamente conectadas o, lo que es lo mismo, una capa densa. En esta capa, cada nodo de la capa de salida se conecta

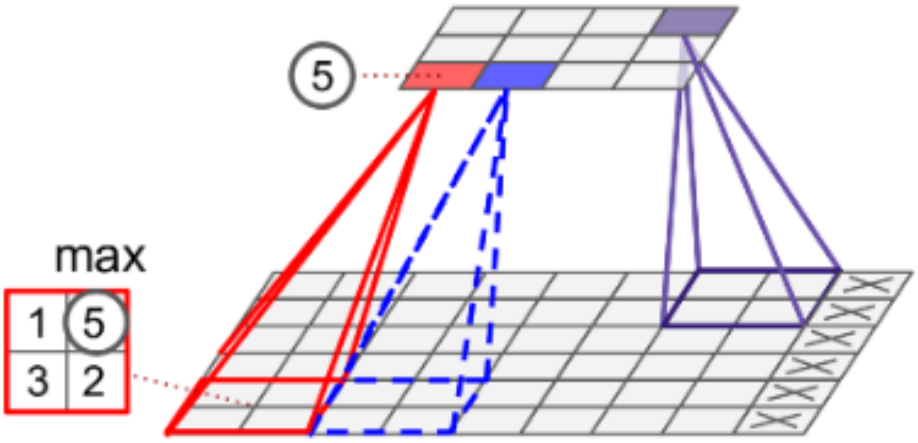

Figura 2.8: Capa de máximo agrupamiento (Max Pooling).

directamente a un nodo de la capa anterior. Realiza la tarea de clasificación en base a las características extraídas a través de las capas anteriores y sus diferentes filtros. Mientras que las capas convolucionales y de agrupación tienden a usar funciones ReLU, las capas densas generalmente aprovechan una función de activación de tipo *softmax* para clasificar las entradas de manera adecuada, produciendo una probabilidad de 0 a 1.

Una capa adicional usada en un modelo de red neuronal convolucional es la capa de aplanamiento o *flattening*. En ella, los mapas de características resultantes se aplanan en un vector unidimensional después de las capas de convolución y agrupamiento para que puedan pasar a una capa densamente conectada para la tarea de clasificación. Esto es debido a que las capas convolucionales y de pooling son capas 2D, mientras que las capas densas son de tipo 1D, por lo que el vector de características ha de ser adaptado al pasar de una parte a la otra del modelo.

2.5. Mecanismos de regularización

El sobreajuste u *overfitting* es un desafío común en el entrenamiento de modelos de aprendizaje automático [Goodfellow et al., 2016]. En general, el sobreajuste ocurre cuando el modelo tiene un buen desempeño en los datos de entrenamiento, pero una mala actuación en los datos de validación o prueba, es decir, el modelo tiene un error de entrenamiento bajo y un error de validación o prueba altos. La regularización es un conjunto de técnicas utilizadas para reducir este sobreajuste [Goodfellow et al., 2016]. En el caso del aprendizaje profundo, algunas de estas técnicas son aplicadas por medio de capas adicionales añadidas en el modelo, como son el dropout y la normalización por lotes, y otras actúan sobre los datos de entrada creando artificialmente nuevos datos de entrenamiento [Pérez and Wang, 2017],

como son el aumento de datos y el uso de filtros de imágenes (estas últimas técnicas son especialmente particulares del campo de reconocimiento de imágenes, el cual es abordado en este proyecto).

Al abordar el sobreajuste para el aprendizaje profundo, la desactivación o *dropout* [Srivastava et al., 2014] propone cambiar aleatoriamente la arquitectura de la red para minimizar los riesgos de que los valores de pesos aprendidos estén altamente personalizados para los datos de entrenamiento subyacentes y, por lo tanto, no se puedan generalizar bien para los datos de validación o prueba. No requiere cambios fundamentales en la arquitectura de la red más allá de la agregación de las capas de dropout. Agregar estas capas a una red aumenta el tiempo de convergencia.

La idea central de la técnica de dropout [Srivastava et al., 2014], como técnica para reducir el sobreajuste, es entrenar submodelos derivados del modelo principal eliminando aleatoriamente unidades o neuronas para cada lote de entrenamiento. Una imagen gráfica de la aplicación de esta técnica se muestra en la Figura 2.9 [Géron, 2019].

Al eliminar repetidamente unidades aleatorias, la desactivación fuerza a las neuronas a ser más robustas, aprendiendo características por sí mismas sin depender de otras unidades. El número de unidades a desactivar se controla mediante un hiperparámetro, el cual se denomina tasa de desactivación. Esta tasa de desactivación puede ser vista como la probabilidad p que tiene cada neurona (incluidas las neuronas de entrada, pero siempre excluyendo las neuronas de salida) en cada paso de entrenamiento de ser desactivada temporalmente, lo que significa que se ignorará por completo durante ese paso de entrenamiento, pero puede estar activa durante el siguiente paso. Por lo general, los valores de esta tasa de desactivación se establecen entre el 10 % y el 50 %: más cerca del 20 % al 30 % en las redes neuronales recurrentes y más cerca del 40 % al 50 % en las redes neuronales convolucionales. Después del entrenamiento, las neuronas ya no se desactivan.

Tal y como explican Garbin et al. [2020], otro mecanismo de regularización es el ofrecido por la técnica de normalización por lotes. Antes de que se introdujera ésta [Ioffe and Szegedy, 2015], el tiempo requerido para alcanzar la convergencia en el entrenamiento de una red dependía significativamente de la inicialización cuidadosa de los hiperparámetros (por ejemplo, valores iniciales de los pesos) y del uso de tasas de aprendizaje pequeñas, lo cual extendía el tiempo de entrenamiento. El proceso de aprendizaje se complica aún más por la dependencia que tiene una capa de sus capas precedentes, ya que los pequeños cambios en una capa pueden amplificarse cuando fluyen a través de las otras capas de la red. La normalización por lotes reduce significativamente el tiempo de entrenamiento al normalizar la entrada de cada capa en la red, no solo la capa de entrada. Este enfoque permite el uso de tasas de aprendizaje más altas (no hay tanto riesgo de acabar en un mínimo local de la función de pérdida que se esté tratando de optimizar), lo que a su vez reduce el número de pasos de entrenamiento que la red necesita para converger al mínimo global (Ioffe and Szegedy [2015] reportaron 14 veces menos pasos en algunos casos).

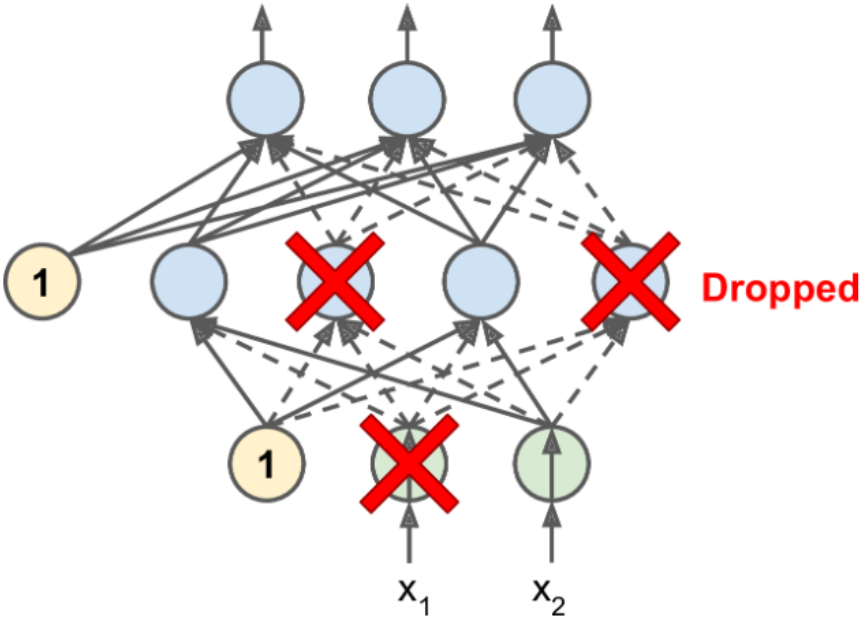

Figura 2.9: Capa de dropout.

Durante el entrenamiento de una red neuronal, la distribución de los valores de entrada de cada capa se ve afectada por todas las capas anteriores. Esta variabilidad reduce la velocidad de entrenamiento (tasas de aprendizaje más bajas). La normalización por lotes se creó para resolver esta variabilidad y acelerar el aprendizaje. Normalizar los valores de cada muestra antes de usarla en la capa de entrada de la red es una técnica bien conocida. La normalización por lotes va un paso más allá y normaliza cada capa de la red, no solo la capa de entrada. La normalización se calcula para cada mini-lote.

La Figura 2.10 muestra los pasos[7] aplicados durante el proceso de normalización por lotes, los cuales se enumeran a continuación:

1. Activaciones (*activations*). Las activaciones de la capa anterior se pasan como entrada a la normalización por lotes. Hay un vector de activación para cada característica en los datos.

2. Cálculo de la media y la varianza (*mean and standard deviation computation*). Para cada vector de activación por separado, se calcula la media y la varianza de todos los valores en el mini-lote.

3. Normalización (*normalization*). Se calculan los valores normalizados para cada vector de características procedente de la activación utilizando la media y la varianza correspondientes. Estos valores normalizados tienen media cero y varianza unitaria.

[7]https://towardsdatascience.com/batch-norm-explained-visually-how-it-works-and-why-neural-networks- need-it-b18919692739

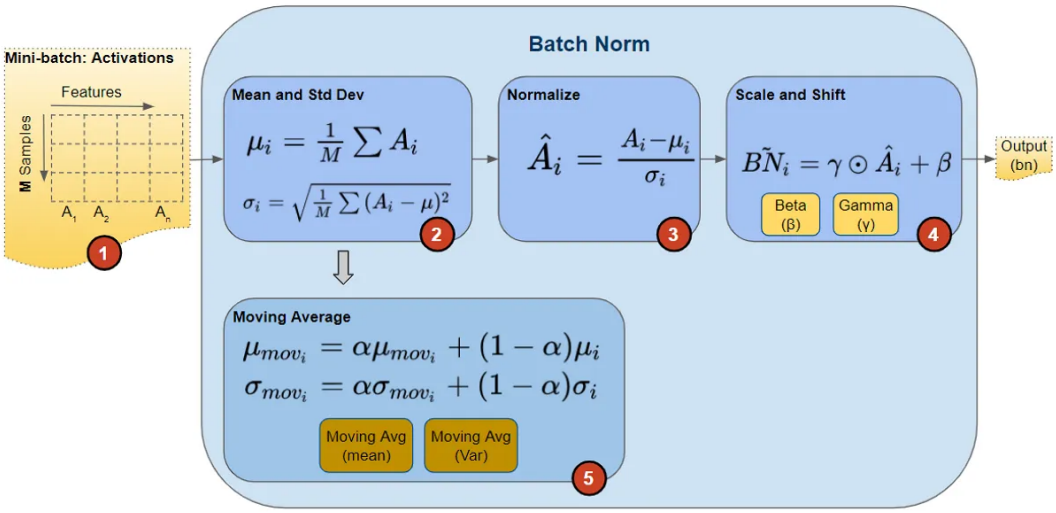

Figura 2.10: Capa de normalización por lotes.

4. Escalado y traslado (*scaling and shift*). Éste constituye el paso innovador del proceso de normalización por lotes. A diferencia de la capa de entrada, que requiere que todos los valores normalizados tengan una media de cero y una varianza unitaria, la normalización por lotes permite trasladar sus valores (a una media diferente) y escalar (a una varianza diferente). Lo hace multiplicando los valores normalizados por un factor, *gamma*, y añadiéndole un factor, *beta*. Se trata de una multiplicación por elementos, no de una matriz. Lo que hace que esta innovación sea ingeniosa es que estos factores no son hiperparámetros (es decir, constantes proporcionadas por el diseñador del modelo), sino parámetros entrenables que aprende la red. En otras palabras, cada capa de normalización por lotes puede encontrar de manera óptima los mejores factores por sí misma y, por lo tanto, puede trasladar y escalar los valores normalizados para obtener las mejores predicciones.

5. Promedio móvil (*moving average*). La normalización por lotes también mantiene un conteo continuo de la media móvil exponencial, o *exponential moving average* (EMA) en inglés, de la media y la varianza. Durante el entrenamiento, simplemente calcula esta EMA, pero no hace nada con ella. Al final del entrenamiento, guarda este valor como parte del estado de la capa para usarlo durante la fase de inferencia.

La normalización por lotes es más una técnica de ayuda al entrenamiento que una estrategia de regularización en sí misma. Esto último se logra realmente aplicando algo adicional conocido como *momentum*. La idea de este momentum es que, cuando se introduzca un nuevo mini-lote de entrada (N muestras procesadas en paralelo), no se usen una media y una desviación para la normalización muy distintas a las de la iteración anterior, sino que se tenga en cuenta el histórico. Por tanto, en ese caso se elegiría una constante que ponderase la importancia de los valores del mini-lote actual frente a los valores del anterior. Gracias a todo esto, se conseguiría reducir el sobreajuste.

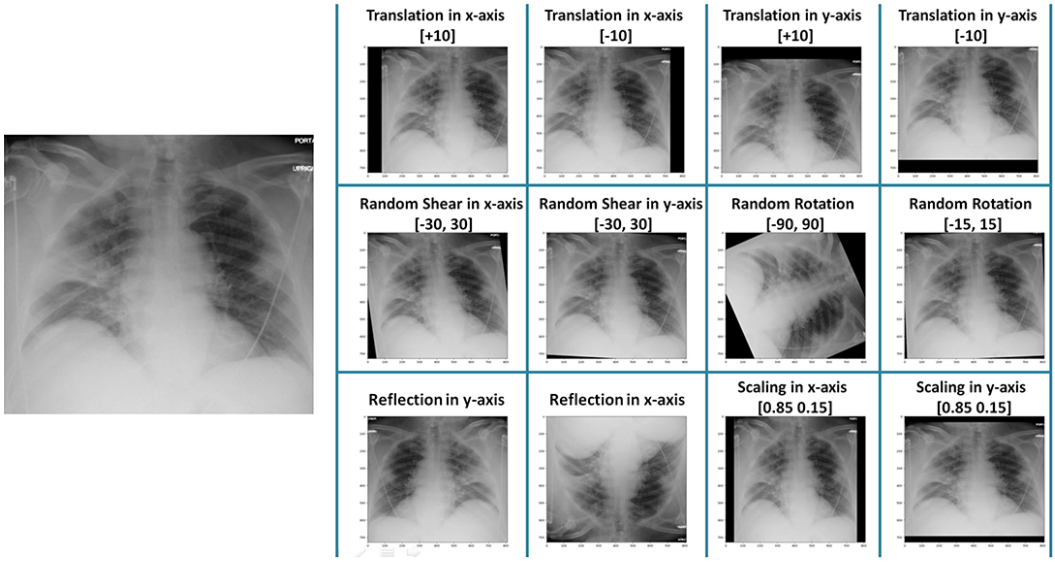

Figura 2.11: Aplicación de la técnica del aumento de datos a una imagen médica.

Por último, es interesante mencionar que la forma en que funciona la normalización por lotes, al ajustar el valor de las unidades para cada lote, y el hecho de que los lotes se crean aleatoriamente durante el entrenamiento, genera más ruido durante el proceso de entrenamiento. El ruido actúa como un regularizador. Este efecto de regularización es similar al introducido por el dropout. Como resultado, el dropout puede eliminarse por completo de la red o debería reducirse significativamente la tasa de desactivación si se usa junto con la normalización por lotes.

Otra técnica a introducir en la lucha contra el sobreajuste es el aumento de datos. Esta técnica adopta el enfoque de generar más datos de entrenamiento a partir de los datos disponibles. En el caso de imágenes, esto lo consigue aplicando una serie de transformaciones aleatorias a la imagen que producen nuevas imágenes de aspecto creíble. El objetivo es que en el momento del entrenamiento el modelo nunca verá exactamente la misma imagen en las diferentes épocas. Esto ayuda a exponer el modelo a más aspectos de los datos y a generalizar mejor.

Una misma imagen de entrada será procesada por la red neuronal tantas veces como épocas se ejecuten de entrenamiento, provocando que la red acabe memorizando la imagen si se entrena demasiado. Por tanto, lo que se hace es aplicar transformaciones de forma aleatoria cada vez que se vuelva a introducir la imagen a la red. La consideración importante que hay que tener en cuenta en relación a estas transformaciones es que las imágenes producidas tras las mismas han de ser realistas. Por tanto, debe tenerse cuidado con la elección de las técnicas específicas de aumento de datos utilizadas, considerando el contexto del conjunto de datos de entrenamiento y el conocimiento del dominio del problema, para no generar imágenes que nunca podrían encontrarse en la realidad, ya que de esta manera se estaría empeorando el entrenamiento.

Ejemplos de transformaciones son:

- Voltear la imagen en horizontal/vertical.
- Rotar la imagen X grados.
- Recortar, añadir relleno, redimensionar, etc.
- Aplicar deformaciones de perspectiva.
- Ajustar brillo, contraste, saturación, etc.
- Introducir ruido, defectos, etc.

Un ejemplo de la aplicación de este tipo de transformaciones a una imagen médica se representa en la Figura 2.11 [8].

De esta forma, se contará con más información para el entrenamiento sin necesidad de obtener imágenes adicionales y también sin alargar los tiempos. Lo más ventajoso es que si la red se dedica a clasificar imágenes, esta técnica conseguirá que el modelo sea capaz de obtener buenos resultados para imágenes tomadas desde distintos ángulos o bajo distintas condiciones de luz. Por tanto, se consigue que la red no sufra por sobreajuste y que generalice mejor.

Es importante resaltar que se realiza la transformación de manera online durante el procesamiento, permitiendo hacer el proceso automático mientras se realiza el entrenamiento sin necesidad de modificar los datos almacenados en disco. Así, el modelo ve una imagen generada aleatoriamente una sola vez. Evidentemente, estas transformaciones se podrían realizar previamente e incluirlas en el conjunto de datos de entrenamiento. De esta manera, el preprocesado resultaría más rápido, puesto que no se realizarían las transformaciones en tiempo de ejecución, pero en cambio el espacio de almacenamiento y el tiempo de carga de los datos en memoria serían más elevados.

Por último, otra técnica que puede resultar muy importante para combatir el sobreajuste en modelos de aprendizaje profundo que abordan una tarea de reconocimiento de imágenes es la aplicación de filtros de imagen. Se va a asumir una imagen en escala de grises. Algunos ejemplos de éstos son filtros de ecualización del histograma, detección de bordes, eliminación de ruido (por ejemplo, CLAHE, Equalized, Gaussiano, Sobel o Canny). Los filtros permiten muchas veces ampliar contrastes y facilitar al modelo el aprendizaje de las características más relevantes.

Teniendo esto en cuenta, se introducen a continuación algunos de los filtros de imágenes más comúnmente utilizados.

Así, el primer ejemplo es la *ecualización de histogramas*, la cual es típicamente usada para ajustar el contraste de una imagen en escala de grises. La ecualización de histogramas

[8] https://www.v7labs.com/blog/data-augmentation-guide

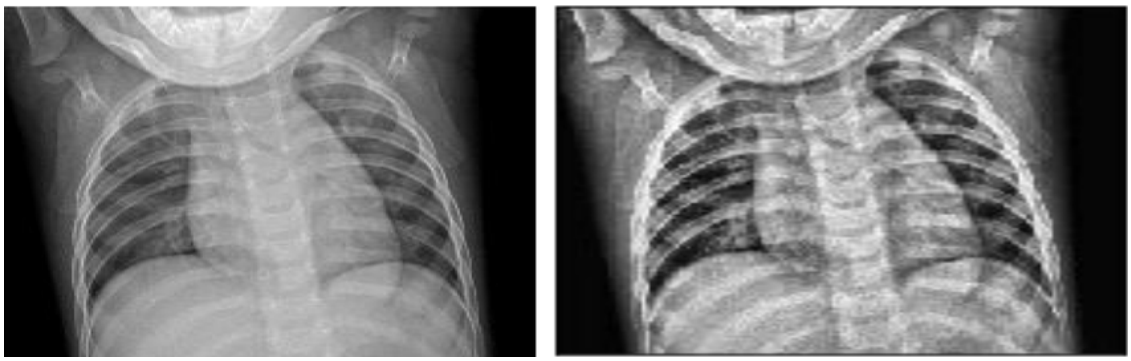

Figura 2.12: Aplicación del filtro CLAHE a una imagen médica.

consiste en transformar los valores de intensidad para que el histograma de la imagen de salida coincida aproximadamente con un histograma especificado. Un caso concreto es cuando se desea que la imagen de salida tenga valores de píxeles distribuidos uniformemente en todo el intervalo. Por lo tanto, la ecualización del histograma de una imagen en escala de grises consiste en la aplicación de un algoritmo que redistribuye los valores de intensidad de ésta, logrando una distribución más uniforme. Esto, a su vez, dará como resultado un mayor contraste en la misma. No obstante, también es cierto que, si bien este filtro incrementa el contraste de la imagen, también aumenta el nivel de ruido que pueda tener ésta, además de hacer parecer la imagen como difuminada, especialmente en las regiones más claras.

Debido a los defectos comentados arriba, puede ser más interesante la aplicación de otro algoritmo: el *Contrast Limited Adaptive Histogram Equalization* (o *CLAHE*). Este algoritmo centrará sus esfuerzos en realizar una redistribución de valores de forma local, dividiendo la imagen en celdas sobre cuyos píxeles se hará el proceso de ecualización y cuyas dimensiones han de definirse buscando un mejor resultado, definiéndose también un umbral límite para dicha transformación. Un ejemplo de la aplicación de este tipo de filtro a una imagen médica se presenta en la Figura 2.12 [Ayalew et al., 2023].

Otra técnica de gran valor es la detección de bordes. Los bordes de una imagen se pueden definir como transiciones entre dos regiones de niveles de gris significativamente distintos. Suministran una valiosa información sobre las fronteras de los objetos y puede ser utilizada para segmentar la imagen, reconocer objetos, etc. La mayoría de las técnicas para detectar bordes emplean operadores locales basados en distintas aproximaciones discretas de la primera y segunda derivada de los niveles de grises de la imagen. Hasta hoy, se han desarrollado numerosos algoritmos o filtros de detección de bordes (Roberts, Prewitt, Sobel, Canny, etc.). Aquí se van a introducir los filtros de Sobel y Canny.

En lo que respecta al algoritmo de Sobel, éste detecta los bordes horizontales y verticales separadamente sobre una imagen en escala de grises. Utiliza como filtro una matriz de coeficientes de 3×3. El operador de Sobel tiene un filtro prácticamente idéntico al de Prewitt con la única diferencia de que en este filtro se le da un mayor peso a la fila o columna central del filtro. La matriz de coeficientes para este operador es definida para el caso horizontal y

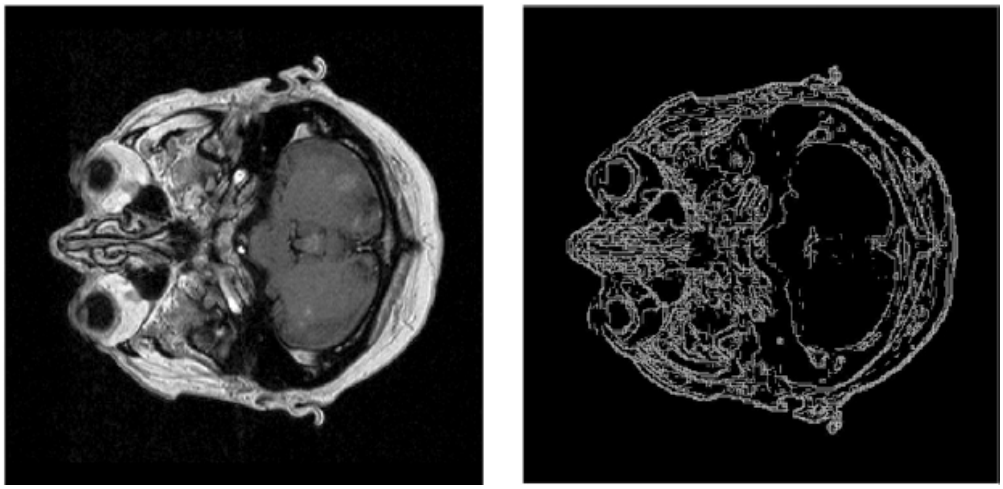

Figura 2.13: Aplicación del filtro de Sobel a una imagen médica para la detección de bordes.

vertical respectivamente como:

$$S_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$$

$$S_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{2.1}$$

Los resultados del filtro de Sobel producen estimaciones del gradiente local para todos los píxeles de la imagen en sus dos direcciones. En la Figura 2.13 [Karras and Mertzios, 2009] se muestra una imagen original de resonancia magnética a la izquierda y, a la derecha, la detección de sus bordes por medio de la aplicación del filtro de Sobel a la imagen original.

Por otro lado, el algoritmo de Canny para la detección de bordes incluye las siguientes etapas:

- Filtra el ruido en la imagen mediante un filtro gaussiano. El filtro gaussiano es igual que un promediado, pero ponderado. Esta ponderación se hace siguiendo la campana de Gauss. Con esto se consigue dar más importancia a los píxeles que están más cerca del centro que a los que están más alejados.

- Encuentra el gradiente. El gradiente define dos valores: la dirección según la cual el cambio de intensidad es máximo y la magnitud de esa dirección.

- Después de obtener la magnitud y la dirección del degradado, se realiza un escaneo completo de la imagen para eliminar los píxeles no deseados que pueden no constituir el borde. Esto elimina los píxeles que no se consideran parte de un borde. Por lo tanto, solo quedarán líneas finas (bordes candidatos).

- Se decide cuáles son los bordes que son realmente bordes y cuáles no. Para esto, se aplica un umbral por histéresis, el cual requiere de dos valores de umbral (uno mínimo y otro máximo). Los bordes con un gradiente de intensidad superior al umbral máximo seguramente serán bordes y aquéllos por debajo del umbral mínimo seguramente no serán bordes, por lo que se descartan. Aquéllos que se encuentran entre estos dos umbrales se clasifican como bordes o no bordes en función de su conectividad. Si están conectados a píxeles de "borde seguro", se consideran parte de los bordes. De lo contrario, también se descartan.

2.6. Grado de Nottingham

Para concluir este capítulo dedicado al marco teórico, resulta interesante e ilustrativo describir con detalle qué mide y en qué se basa la característica que se predecirá en este trabajo exploratorio. Como ya se comentará más adelante, la característica escogida será la conocida como *grado de Nottingham*.

Tras la realización de una biopsia de mama y el correspondiente diagnóstico de cáncer de mama, es necesario saber qué tratamientos son los mejores para la paciente y cuál será su pronóstico. Para hacerlo, es necesario estadificar el cáncer. La estadificación del cáncer se refiere al tamaño o extensión de un tumor sólido y si se ha diseminado o no a otros órganos y tejidos. Tiene en cuenta múltiples factores para establecer la gravedad del cáncer y qué tratamientos son los más adecuados. Como ejemplo de medida de la estadificación del cáncer de mama, aparece el grado de Nottingham.

El sistema de clasificación de Nottingham[9] es una actualización de los criterios de clasificación anteriores, el sistema Bloom-Richardson, el cual se estableció por primera vez en 1957. Nottingham evalúa la estructura y distribución de las células cancerosas para determinar cómo de agresiva será la malignidad.

Los tumores de grado bajo, los cuales se parecen más a las células normales, tienden a crecer lentamente, mientras que los tumores de grado alto tienen un aspecto anormal y se diseminan rápidamente.

Hay tres factores que un patólogo tendrá en cuenta al evaluar las células tumorales: formación de túbulos, tasa mitótica y grado nuclear. A cada uno se le da una puntuación de 1 (más normal) a 3 (menos normal). Posteriormente, se suman estos valores, cuyo total indicará el nivel de malignidad del tumor.

A continuación, se describen los tres factores considerados en el cálculo del grado de Nottingham [Wasserman, 2022].

[9] https://healthyatri.com/tumor-grade-and-pathology-AJQ

Formación de túbulos

Un túbulo es un grupo de células conectadas entre sí para formar una estructura redonda similar a un anillo. Los túbulos se ven similares, pero no son exactamente iguales a las glándulas que normalmente se encuentran en el seno. En este caso, se revisa la cantidad de tejido tumoral que tienen los conductos mamarios o, dicho de otra forma, el porcentaje de células cancerosas que forman túbulos. Las posibles puntuaciones son las siguientes:

- 1: Más del 75 % de las células son normales.
- 2: Entre el 10 % y el 75 % de las células son normales.
- 3: Menos del 10 % de las células son normales.

Tasa mitótica

Las células se dividen para crear nuevas células. El proceso de creación de una nueva célula se llama mitosis y una célula que se está dividiendo se llama figura mitótica. Así, se cuenta la cantidad de figuras mitóticas o células en división en un área específica (llamada campo de alta potencia) y se usa ese número para otorgar una puntuación entre 1 y 3. La puntuación es la siguiente:

- 1: Se observan menos de 10 células mitóticas (tumores con muy pocas figuras mitóticas).
- 2: Se observan entre 10 y 19 células mitóticas.
- 3: Se ven al menos 20 células mitóticas (tumores con muchas figuras mitóticas).

Grado o pleomorfismo nuclear

El núcleo es una parte de la célula que contiene la mayor parte del material genético (ADN). Pleomorfismo (o pleomórfico) es una palabra que usan los patólogos cuando el núcleo de una célula tumoral se ve muy diferente del núcleo de otra célula tumoral. Se da una puntuación de 1 a 3 para el pleomorfismo nuclear, es decir, es la evaluación del tamaño y la forma del núcleo en las células tumorales. Las posibles puntuaciones incluyen:

- 1: Los núcleos son pequeños y uniformes.
- 2: Hay variaciones intermedias en tamaño y forma.
- 3: Hay variaciones marcadas en tamaño y forma.

Las tres puntuaciones anteriores se combinan para determinar finalmente el nivel de malignidad del tumor. Los grados más altos confieren una mayor agresividad y una mayor propensión a propagarse.

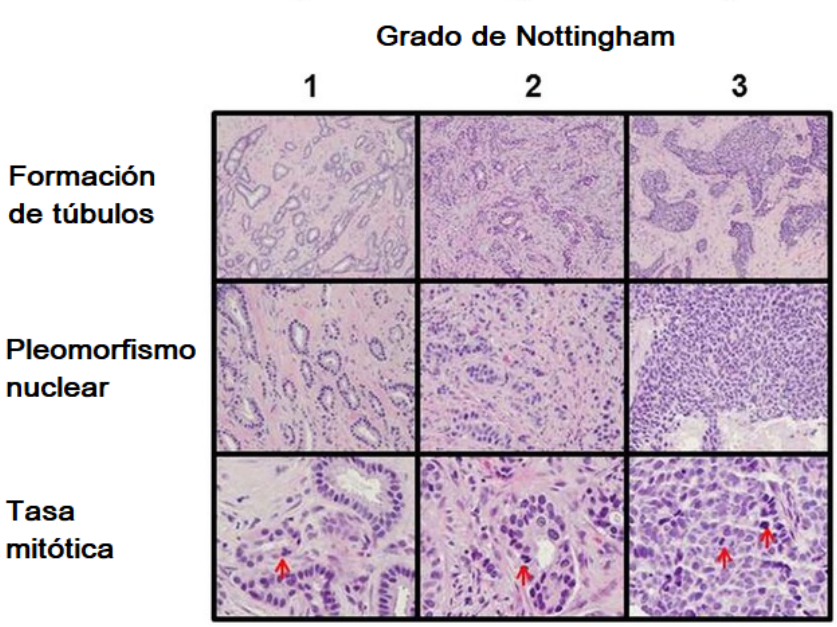

Figura 2.14: Factores considerados para determinar el grado de malignidad del tumor de mama.

Según la puntuación final alcanzada, se puede diferenciar entre los siguientes grados de Nottingham:

- De 3 a 5: Tumor de grado 1. El tumor es bien diferenciado (parece normal, crece lentamente, no es agresivo).
- De 6 a 7: Tumor de grado 2. El tumor es moderadamente diferenciado (semi-normal, crecimiento moderadamente rápido).
- De 8 a 9: Tumor de grado 3. El tumor es poco diferenciado (anormal, crece rápidamente, es agresivo).

La Figura 2.14 [Ping et al., 2016] ofrece una imagen gráfica acerca del criterio seguido para determinar el grado de malignidad del tumor de mama.

Como ya se comentó, el grado de Nottingham se usa para estadificar el cáncer de mama, es decir, para saber cómo de avanzado está el tumor. La estadificación ayuda al médico a decidir qué tratamiento puede erradicar por completo la neoplasia maligna con la menor cantidad de daño. Por ejemplo, el cáncer en estadio temprano puede requerir cirugía o radiación, mientras que el cáncer en estadio avanzado puede necesitar ser tratado con quimioterapia.

Capítulo 3

Estado del Arte

En este capítulo se va a presentar el estado del arte asociado al problema abordado en este proyecto. La tarea de este estudio está principalmente relacionada con la aplicación de técnicas de aprendizaje profundo al reconocimiento de imágenes de resonancia magnética de mama con el objetivo de llevar a cabo actividades de clasificación enmarcadas dentro del campo del diagnóstico médico. Por tanto, las conclusiones obtenidas del estado del arte y que son aquí mostradas estarán principalmente enfocadas hacia dicho objetivo.

Además, es importante mencionar que los datos empleados en este estudio, los cuales se presentan en el siguiente capítulo, ejercen también una gran influencia en la bibliografía revisada, por lo que aquí se mostrarán a su vez conclusiones procedentes de artículos que han utilizado dichos datos.

Así, es interesante comenzar revisando cuáles son los objetivos de la IA en lo que se refiere a la resonancia magnética para el cáncer de mama [Sheth and Giger, 2020]. Son cuatro los bloques principalmente abordados: detección, diagnóstico, respuesta al tratamiento y riesgo de recurrencia y evaluación de riesgo y prevención.

La detección se refiere generalmente a la localización de objetos de interés en imágenes. Sistemas ayudados por ordenador han sido desarrollados para ayudar a radiólogos en la tarea de interpretar MRIs de mama, especialmente imágenes de resonancia magnética con realce por contraste dinámico (DCE-MRIs). Estos sistemas tienen el potencial de reducir el tiempo de lectura y reducir los errores de diagnóstico, ya que pueden resaltar las lesiones sospechosas, incluidas aquéllas que los radiólogos pueden malinterpretar o pasar por alto durante la detección [Gubern-Mérida et al., 2016]. La mayoría de los algoritmos de análisis de resonancia magnética de mama se basan en el protocolo de MRI de mama completa, incluida la información temporal de las exploraciones de fase tardía y la información morfológica de las exploraciones de fase temprana. Los investigadores han informado sobre los sistemas CADe [Giger, 2004, Giger et al., 2008], los cuales indican ubicaciones de posibles lesiones sospechosas en DCE-MRI según la morfología y la cinética [Chang et al., 2014, Gubern-Mérida et al., 2015]. Mención especial recibe el sistema CADe desarrollado por Dalmis et al. [2018] utilizando el aprendizaje profundo y que se basa en la información espacial de la fase

inicial en lugar de la información temporal de la fase tardía. Su sistema CADe funcionó significativamente mejor que su sistema CADe anterior, lo que demuestra su posible uso en protocolos de resonancia magnética abreviados. Los futuros sistemas CADe que pueden extraer información de exploraciones de fase temprana serán cada vez más valiosos para los radiólogos.

El diagnóstico generalmente se refiere al estudio de una lesión mamaria una vez que se detecta mediante imágenes u otros medios, como un examen físico. Por lo tanto, no es una tarea de localización, sino más bien una tarea de clasificación y, en este paso clínico, a menudo se utilizan múltiples modalidades mamarias, lo que requiere la integración de los hallazgos. El diagnóstico se refiere a la segmentación, caracterización y estadificación de los tumores, conocidos colectivamente como diagnóstico asistido por computadora (CADx). Un sistema CADx implica la caracterización automática de una región o tumor, indicada inicialmente por un radiólogo o una computadora, después de lo cual la computadora caracteriza la lesión de interés y estima la probabilidad de malignidad, dejando el plan de manejo terapéutico o de estudio a los médicos [Giger et al., 2008]. Con un sistema CADx, el objetivo es generar características de la lesión (es decir, características radiómicas), así como una firma tumoral relacionada con la probabilidad de que la lesión en cuestión sea un tumor canceroso. Durante décadas, Giger y otros compañeros han desarrollado y traducido clínicamente un sistema de este tipo, que incluye la segmentación automática de lesiones, la extracción de características y la fusión de características en una firma tumoral y han demostrado una mejora en el desempeño de los radiólogos en la tarea de clasificar entre lesiones malignas y benignas [Gilhuijs et al., 1998, Shimauchi et al., 2011].

Dichos sistemas CADx varían en la extracción de tumores, puesto que algunos utilizan métodos semiautomáticos y otros incorporan métodos automatizados de segmentación de lesiones. Meinel et al. [2007] crearon un sistema CAD de resonancia magnética de mama que mejoró el rendimiento de todos los lectores de interpretación en la clasificación de lesiones. El sistema CAD se basó en una red neuronal de retropropagación que se entrenó en 80 lesiones de resonancia magnética de mama que un radiólogo experto segmentó manualmente. Dalmis et al. [2016] desarrollaron un sistema CADx para DCE-MRI de alta resolución espaciotemporal utilizando características radiómicas. Se ha demostrado que la caracterización de las regiones más realzadas dentro de la lesión es beneficiosa para evaluar la probabilidad de malignidad. Chen et al. [2006] construyeron un método para identificar los vóxeles más potenciadores dentro de un tumor mediante la técnica de clustering de *fuzzy c-means* no supervisado, el cual fue validado posteriormente por Chang et al. [2012].

El aprendizaje profundo a través de redes neuronales convolucionales se ha aplicado a la tarea de caracterización en el diagnóstico de tumores de mama en resonancia magnética [Antropova et al., 2017, 2018]. En la tarea de distinguir entre lesiones mamarias malignas y benignas, se ha utilizado el aprendizaje de transferencia con una CNN previamente entrenada, ya sea a través de la extracción de características o del ajuste fino. Cuando se utiliza

el aprendizaje profundo en la clasificación, los datos de la imagen del tumor dentro de una región de fondo, a diferencia de la imagen completa, se ingresan en la CNN para evitar que las regiones de imagen distales no relacionadas con el tumor interfieran con el método de aprendizaje profundo.

Debido a los conjuntos de datos limitados, los investigadores han buscado el beneficio de usar varias formas de datos de imágenes procesadas como entrada a la CNN. Por ejemplo, Antropova et al. [2018] encontraron con DCE-MRI que el uso de imágenes MIP (proyección de intensidad máxima) en lugar de imágenes de sustracción o sin sustracción de fase temprana produjo un mejor rendimiento del modelo de aprendizaje profundo.

Algunos estudios también han reportado correlaciones significativas entre la cinética de realce de la resonancia magnética y los subtipos moleculares del cáncer de mama [Li et al., 2016]. En un estudio de investigación multiinstitucional del Instituto Nacional del Cáncer, se investigaron las relaciones entre las características radiómicas de la resonancia magnética extraída por computadora y varios marcadores clínicos, moleculares y genómicos [Zhu et al., 2015, Wu et al., 2017]. Se observaron asociaciones estadísticamente significativas entre las características radiómicas de la textura de realce y los subtipos moleculares (como luminal A, luminal B, enriquecido con HER2 o de tipo basal).

En cuanto a la respuesta al tratamiento y el riesgo de recurrencia, el papel de las imágenes para evaluar la respuesta a la terapia se está expandiendo a recomendaciones terapéuticas y pronósticas traducibles y clínicamente relevantes. Los métodos actuales para evaluar la respuesta del tumor a la terapia neoadyuvante consisten en un examen físico y estudios de imágenes de mama convencionales con mamografía, ultrasonido y/o resonancia magnética. La resonancia magnética de mama es la modalidad más sensible para la detección del cáncer de mama y es la más precisa para evaluar la respuesta del tumor a la terapia neoadyuvante [Yuan et al., 2010, Wu et al., 2012, Mariscotti et al., 2014]. Usando una CNN para predecir la respuesta patológica completa de la base de datos ISPY, Ravichandran et al. [2018] pudieron producir mapas de calor de probabilidad que demostraron las regiones tumorales más fuertemente asociadas con la respuesta terapéutica.

Otra aplicación de la radiogenómica es la correlación de las características de resonancia magnética del cáncer de mama con ensayos genómicos clínicamente disponibles, que proporcionan puntajes de pronóstico que representan el riesgo de recurrencia del cáncer y, por lo tanto, pueden ser útiles para guiar las decisiones de tratamiento [Veer et al., 2002, Prat et al., 2012]. Los datos cualitativos extraídos de la resonancia magnética podrían ser un biomarcador de imagen potencial para la recurrencia del riesgo de cáncer de mama. Ashraf et al. [2014] investigaron la relación entre las características de la estructura, la función y la heterogeneidad del tumor extraídas por computadora de la resonancia magnética y la expresión génica utilizando OncotypeDx. Cuatro características (relacionadas con el patrón de realce del tumor y el tamaño del tumor) se correlacionaron con la puntuación de recurrencia, lo que demuestra que los tumores con mayor neoangiogénesis se asociaron con un mayor riesgo

de recurrencia. Li et al. [2016] indagaron más en este concepto mediante el uso de fenotipos de resonancia magnética de mama extraídos por computadora para predecir el riesgo de recurrencia del cáncer de mama mediante ensayos multigénicos clínicamente disponibles, lo que demostró que los tumores con un alto riesgo de recurrencia eran más grandes, con un realce más heterogéneo.

Con todo esto, los enfoques personalizados para tumores individuales pueden conducir a un inicio más temprano de los planes terapéuticos, una comprensión más profunda de los pronósticos de los tumores y mejores resultados para las pacientes. Éste fue el objetivo, por ejemplo, de Zhu et al. [2015], los cuales trataron de mapear características radiómicas de MRI extraídas por computadora a marcadores genómicos.

Por último, en lo que se refiere a la evaluación de riesgo y prevención, existen herramientas que permiten estimar el riesgo de cáncer de mama a lo largo de la vida de una mujer. Esta evaluación precisa permite la aplicación de regímenes de detección estratificados por riesgo y terapias preventivas para reducir dicho peligro. Se ha demostrado que la densidad mamaria y los patrones parenquimatosos de la densidad mamaria desempeñan un papel en la estimación del riesgo de cáncer de mama [Boyd et al., 2011]. La densidad mamaria describe la cantidad relativa de tejido fibroglandular al tejido adiposo en la mama. Así, la probabilidad de padecer cáncer de mama aumenta constantemente con el incremento de la densidad mamaria mamográfica [Saftlas et al., 1991]. El riesgo de cáncer de mama en mujeres con mamas mamográficamente densas es de tres a cinco veces mayor que en mujeres con mamas predominantemente no densas (grasas) mamográficamente [Boyd et al., 1995].

Al igual que la densidad mamaria en la mamografía, la cantidad de tejido fibroglandular (FGT) que se observa en las MRIs y el nivel de realce parenquimatoso de fondo (BPE) después de la administración de contraste son características del tejido mamario normal. La proporción de FGT determina la densidad del seno, tal y como se dijo anteriormente. Por otro lado, el nivel de BPE se refiere al volumen y la intensidad del realce del tejido mamario normal después de la administración de material de contraste intravenoso. Dalmis et al. [2017] utilizaron un método de aprendizaje profundo denominado U-Net para cuantificar con precisión la densidad mamaria. Este método de aprendizaje profundo superó a los métodos existentes y pudo hacer frente a las variaciones en los protocolos de resonancia magnética y las formas y los volúmenes de los senos. King et al. [2011] encontraron que el nivel de BPE es un predictor altamente significativo del riesgo de cáncer de mama, aumentando significativamente para las mujeres con BPE moderado o marcado. Así, Dontchos et al. [2015] concluyeron que las mujeres con un BPE leve, moderado o marcado tenían nueve veces más probabilidades de desarrollar cáncer de mama que las mujeres con un nivel mínimo de BPE, sugiriendo que el BPE puede indicar tejido mamario fisiológicamente activo, el cual podría conducir a la transformación maligna.

Además, recientemente se están desarrollando modelos generales de evaluación de riesgos basados en imágenes. Portnoi et al. [2019] desarrollaron un modelo de aprendizaje profundo

basado en imágenes que utiliza una sola imagen de proyección de intensidad máxima 2D realzada por contraste para predecir el riesgo de cáncer de mama en una población de pacientes de alto riesgo.

De los cuatro bloques de aplicación introducidos hasta ahora, se va a continuar estudiando en detalle especialmente el correspondiente a diagnóstico, puesto que, tal y como se ha comentado con anterioridad, es el bloque donde se enmarca la tarea a resolver en el presente proyecto.

De esta forma, un estudio interesante realizado en esta línea es el de Herent et al. [2019], el cual tiene el propósito de evaluar el potencial de un modelo de aprendizaje profundo para discriminar entre lesiones mamarias benignas y malignas mediante resonancia magnética y caracterizar diferentes subtipos histológicos de lesiones mamarias.

Para la tarea de extracción de características de las imágenes se usó una red neuronal de 50 capas, ResNet-50 [He et al., 2016], preentrenada en el conjunto de datos conocido como ImageNet y de la que se eliminaron las dos últimas capas. El vector de características era alimentado a una única capa densamente conectada con cinco neuronas para cada tarea de clasificación (malignidad, tejido normal, otras lesiones benignas, carcinoma ductal invasivo y otras lesiones malignas). El principal inconveniente de este enfoque es que no diferencia regiones de poco interés, como el tórax o el fondo.

Un desafío encontrado fue la heterogeneidad en la apariencia y el tamaño de las lesiones mamarias. De esta forma, se facilitó el aprendizaje al descomponer la clasificación en dos pasos: (i) detección de anomalías presentes en las imágenes de resonancia magnética y (ii) clasificación de estas lesiones. Estos dos pasos fueron realizados simultáneamente por dos ramas del mismo modelo. Para la primera, se creó y usó etiquetas adicionales para la localización. Estas etiquetas consistían en cuadros delimitadores que rodeaban las lesiones. Estas anotaciones no requirieron una caracterización precisa, sino que fueron realizadas rápidamente por un residente de radiología de 5^{o} año, el cual tenía experiencia limitada en resonancia magnética de mama.

Para cada imagen se generó una máscara binaria del mismo tamaño, indicando la presencia o ausencia de lesiones. El tamaño de esta máscara se redujo para que coincidiera con las dimensiones de salida de la ResNet.

El módulo de localización era una única convolución 1×1, aplicada a la salida de la ResNet. Esto transformaba la representación de $2048 \times 8 \times 11$ en una sola imagen con las dimensiones 8×11, a la que se aplicaba una función sigmoidea para generar una predicción entre 0 y 1. Este módulo fue entrenado para reproducir la máscara binaria generada a partir de las anotaciones. Esta predicción local se utilizó luego para guiar el módulo principal responsable de determinar la presencia y caracterización de lesiones en la imagen. Además, se usó la predicción local para calcular un promedio ponderado del mapa de características final.

La predicción final era realizada por una capa densamente conectada con cinco neuronas,

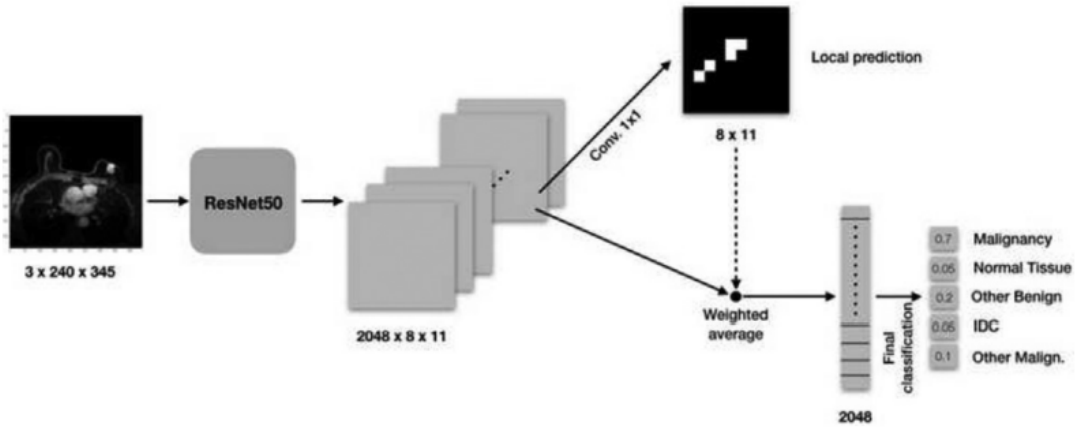

Figura 3.1: Arquitectura del modelo de red neuronal empleado por Herent et al. [2019].

una para cada predicción tal y como se mencionó anteriormente: clasificación de malignidad de la lesión, tejido normal, otras lesiones benignas, carcinoma ductal invasivo y otras lesiones malignas. La arquitectura comentada aparece representada en la Figura 3.1.

Además, el mecanismo usado permite la interpretación de las predicciones del modelo. Así, se redimensionaba el mapa correspondiente a las dimensiones de la imagen original y luego se podía superponer este mapa sobre la imagen para ver las áreas consideradas por el modelo para tomar su decisión.

En cuanto a la implementación, el modelo fue entrenado simultáneamente en las tres tareas evaluadas (detección de lesiones, diagnóstico de malignidad y clasificación de lesiones). Esta técnica multitarea limitó el sobreajuste. Sin embargo, las tareas no se aprendieron al mismo ritmo. Así, se guardaron tres copias de los pesos, elegidos en función del rendimiento del modelo en un conjunto de validación. Cuando el modelo alcanzaba su mejor área bajo la curva ROC (AUC) para la detección de lesiones, se guardaba la primera copia de sus pesos, los cuales se utilizaron solo para esta tarea. Se usó el descenso de gradiente estocástico con impulso de Nesterov [Nesterov, 1983] para entrenar los modelos. Los resultados fueron muy variables debido a la pequeña cantidad de datos y, por lo tanto, se realizó una validación cruzada triple, repetida en tres divisiones diferentes de los datos. Por ello, se repitieron nueve experimentos durante los cuales se seleccionó aleatoriamente 223 imágenes (de las 335 imágenes del conjunto de entrenamiento) para entrenar la red neuronal y se estimó su rendimiento calculando un AUC sobre las 112 imágenes restantes. Luego, las puntuaciones medias de esos nueve experimentos diferentes fueron calculadas para evaluar el modelo antes de ejecutarlo en el conjunto de test.

Con todo esto, el modelo alcanzó un AUC ponderado (cada una de las cinco tareas de clasificación tenía un peso) de 0.816 en el conjunto de datos de test.

Otro sistema interesante a mencionar en la tarea de diagnóstico es el sistema CADx para MRI propuesto por Lu et al. [2017], el cual está centrado en el diagnóstico de masas mamarias malignas, basado en la selección de características y el aprendizaje conjunto. En

primer lugar, se extrajeron las características morfológicas, de textura y de Gabor para caracterizar las masas de cáncer de mama. Luego, se empleó el algoritmo conocido como *Relief* [Kira and Rendell, 1992] para encontrar el subconjunto de características óptimo para el entrenamiento del clasificador. Estas funciones se incorporaron posteriormente a un nuevo marco de aprendizaje conjunto basado en la combinación de lo que se conoce como *Ensemble of Under-Sampled* (EUS) y la técnica subespacial. Los resultados experimentales indicaron que la propuesta supera a los otros métodos del estado del arte en sensibilidad diagnóstica, pero la tasa de clasificación de falsos positivos aumenta ligeramente.

Las principales contribuciones de este estudio se pueden resumir de la siguiente manera:

- La dimensionalidad de las características utilizadas es mayor que en la mayoría de los métodos más avanzados. Se extrajeron varias características, incluidas de tipo morfológico, de Gabor y de textura para caracterizar de manera integral las masas mamarias.

- Se selecciona un subconjunto de características óptimo del conjunto original utilizando Relief, en función de su tipo, lo que ayuda a reducir las características redundantes e irrelevantes y tiene en cuenta el significado físico de las características.

- Se propone un marco de aprendizaje conjunto novedoso basado en la combinación de EUS, subespacio y Adaboost, el cual ayuda a aliviar el problema del desequilibrio de datos (cuando el número de datos de cada clase está descompensado) y mejora la precisión de clasificación general del sistema CADx.

La estructura general de este sistema CADx desarrollado se muestra en la Figura 3.2. Acorde con ésta, debido a que la segmentación automática de regiones de interés (ROIs) puede causar algunos errores, se decidió segmentar manualmente las ROIs de las imágenes de resonancia magnética de mama. Tras esto, una vez se extraen las características para la caracterización de las imágenes de cáncer de mama, se aplica el método mencionado de Relief para efectuar una selección de características. Posteriormente, se aplica la técnica de subespacio y se aplica un aprendizaje conjunto para el procesamiento de los datos desequilibrados o descompensados. Finalmente, el clasificador escogido se entrena y se emplea para acometer predicciones.

Teniendo esto en cuenta, en primer lugar, el objetivo de la extracción de características es adquirir variables numéricas que puedan reflejar las propiedades intrínsecas de la imagen [Ping Tian, 2013]. Así, las masas mamarias malignas suelen tener, por ejemplo, un límite espiculado, rugoso y borroso, mientras que las masas benignas suelen tener un límite redondo, liso y bien definido [Mudigonda et al., 2000]. En este estudio, tal y como ya se ha mencionado, se extrajeron características morfológicas, de Gabor y de textura.

En segundo lugar, la selección de características juega un papel importante en el entrenamiento de clasificadores. Suponiendo que el espacio de características original contiene D

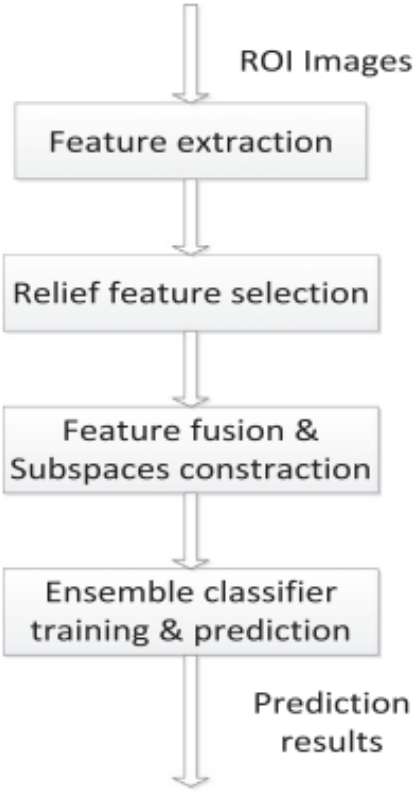

Figura 3.2: Arquitectura general del sistema CADx propuesto por Lu et al. [2017].

características, el objetivo del proceso de selección de características es encontrar un subconjunto óptimo que contenga solo d ($d \leq D$) características. La selección de características puede reducir la complejidad computacional y mejorar la precisión de la clasificación. Los métodos de selección de características se pueden clasificar principalmente en dos tipos: contenedor (*wrapper*) y filtro (*filter*) [Kohavi and John, 1997]. Generalmente, los métodos de tipo filtro son más rápidos que los métodos de tipo contenedor y no son sensibles al clasificador. Entre los métodos de tipo filtro, se encuentra el ya mencionado método Relief, el cual se basa en la ponderación de características y se considera ampliamente como un método eficaz de selección. El método Relief es fácil de implementar y tiene una menor complejidad computacional y, por ello, fue escogido para la selección de características.

En tercer lugar, el desequilibrio de datos es un problema común en las tareas de clasificación médica. Teniendo en cuenta que la mayoría de los algoritmos de clasificación binaria se diseñaron bajo el supuesto de que no hay una diferencia significativa entre el tamaño de la clase minoritaria (positiva) y la mayoritaria (negativa), el rendimiento de los clasificadores podría verse seriamente disminuido como resultado del desequilibrio de datos. Para resolver este problema, se propuso en este estudio un método de conjunto que combina submuestreo, subespacio y boosting basado en la directriz de la descomposición del error en varianza y sesgo. Los métodos de aprendizaje conjunto pueden hacer combinaciones de múltiples clasificadores y así reducir el error de generalización de diferentes maneras. Así, los métodos

basados en bagging se centran en reducir la varianza de los clasificadores base mediante submuestreo, mientras que los métodos basados en boosting se enfocan en reducir el sesgo mediante estrategias de ponderación.

Para los métodos de aprendizaje conjunto, diferentes métodos de agregación de resultados pueden generar diferentes salidas de clasificación. La votación por mayoría, la votación ponderada y la agregación del valor funcional son tres métodos ampliamente utilizados para realizar la agregación. Sin embargo, los resultados experimentales han demostrado que no existe una diferencia significativa en el rendimiento de clasificación entre estos métodos [Kang and Cho, 2006]. Además, la estrategia ponderada y la agregación del valor funcional pueden provocar un sobreajuste [Zhou, 2016]. Por lo tanto, se eligió el método de votación por mayoría.

Además, como clasificador base para la tarea de clasificación se escogió el conocido como C4.5 [Quinlan, 2014]. Este clasificador es un árbol de decisión que elige características y nodos apropiados en función de la relación de ganancia de información. Es un clasificador "débil" y puede mostrar un gran rendimiento en métodos de aprendizaje conjunto[1].

Finalmente, el sistema CADx propuesto alcanzó una sensibilidad de 0.90, una especificidad de 0.96 y un AUC de 0.96. También quedó patente que el sistema propuesto puede aumentar la especificidad con solo una pequeña disminución en la sensibilidad.

Continuando con los sistemas de diagnóstico existentes en el estado del arte, aparece otro estudio realizado por Rasti et al. [2017], el cual propone un nuevo sistema CAD para DCE-MRI de mama. El sistema tiene dos grandes etapas: i) segmentación del candidato tumoral en base a la intensidad e información morfológica de las masas en la imagen y ii) clasificación de tumores basada en una nueva agrupación de aprendizaje profundo de redes neuronales convolucionales. Las actuaciones del algoritmo de segmentación propuesto y el modelo de agrupación de diagnóstico se evaluaron en un conjunto de datos reales constituido por 112 pacientes (mujeres con riesgo alto o intermedio).

Para mejorar la precisión de la clasificación, se han desarrollado métodos de agrupación o conjunto donde varias salidas de CNNs se fusionan utilizando una red adicional o una capa de tipo softmax.

Los primeros pasos a acometer con los datos eran la aplicación de ciertas técnicas de segmentación y la selección de la región de interés de una imagen de mama de entrada. De esta manera, los pasos generales del algoritmo empleado para llevar a cabo la selección automática de la región de interés o ROI aparecen en la Figura 3.3.

El primer paso fue la reducción de fondo a través de la primera sustracción poscontraste, seguida del realce por contraste y el recorte de las regiones mamarias. El objetivo de la fase de recorte de las regiones mamarias era ubicar aproximadamente los tejidos mamarios

[1]Los clasificadores "débiles" no son demasiado precisos, ya que son simples y funcionan solo ligeramente mejor que una clasificación aleatoria. Sin embargo, existen algoritmos o métodos contrastados, como el *bagging* y el *boosting*, que generan un clasificador más preciso combinando muchos clasificadores "débiles".

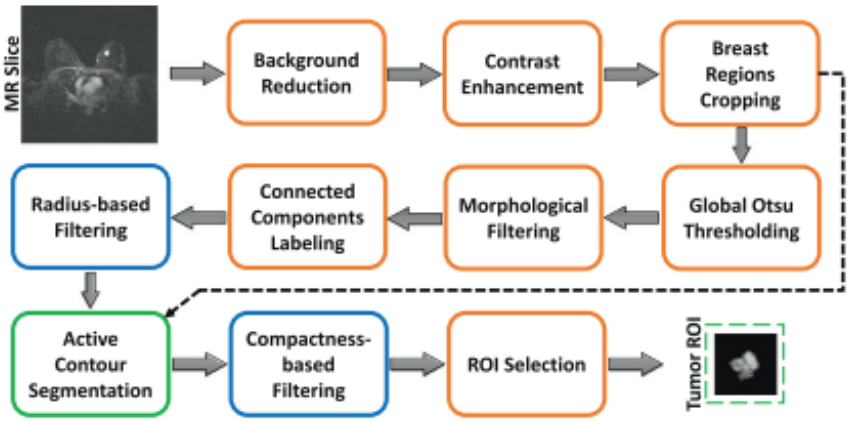

Figura 3.3: Pasos principales para la selección de la región de interés (ROI) en DCE-MRI de mama según Rasti et al. [2017].

y reducir los efectos perturbadores de otras estructuras anatómicas como el tórax. El tamaño y la posición del recorte rectangular se determinaron empíricamente en función de la configuración de adquisición de imágenes.

A continuación, se efectuó la umbralización global de Otsu [Otsu, 1979] y el filtrado morfológico de paso alto para eliminar las estructuras que no eran lesiones [Dougherty and Lotufo, 2003]. Después del etiquetado de los componentes conectados, las pseudolesiones se aislaron para su posterior procesamiento. Luego, se aplicó el filtrado basado en un radio para eliminar regiones con radio fuera de un determinado rango. Posteriormente, se aplicó la segmentación del contorno activo localizado (LAC) a cada lesión restante de interés.

El paso de segmentación de LAC fue útil porque recuperó píxeles tumorales que fueron eliminados por operaciones morfológicas. Para reducir los falsos positivos, se aplicó a continuación un filtrado basado en la compacidad. Se eliminaba una región si su compacidad era inferior a 0.2.

En general, se extrajeron 562 ROIs (244 malignas y 318 benignas) con el algoritmo de selección de ROI en todo el conjunto de cortes.

Además, fue propuesta una agrupación conjunta o conjunto mixto de redes neuronales convolucionales para la clasificación de una imagen ROI como benigna o maligna. En la Figura 3.4 se muestra una descripción general del modelo de conjunto mixto. Hay un número L de expertos y una red de compuerta (CGN) que comparten la misma entrada. Cada experto podría estar especializado en una región del espacio de entrada de alta dimensión. La CGN está capacitada para producir ponderaciones adaptativas de entrada que se utilizan para fusionar las salidas de los expertos. En el modelo propuesto, las CNNs se utilizan como expertos y como red de compuerta. Además, todos los componentes (expertos y red de acceso) se entrenan simultáneamente a través de un proceso de optimización de extremo a extremo.

El algoritmo RPROP se puede utilizar para determinar simultáneamente los parámetros

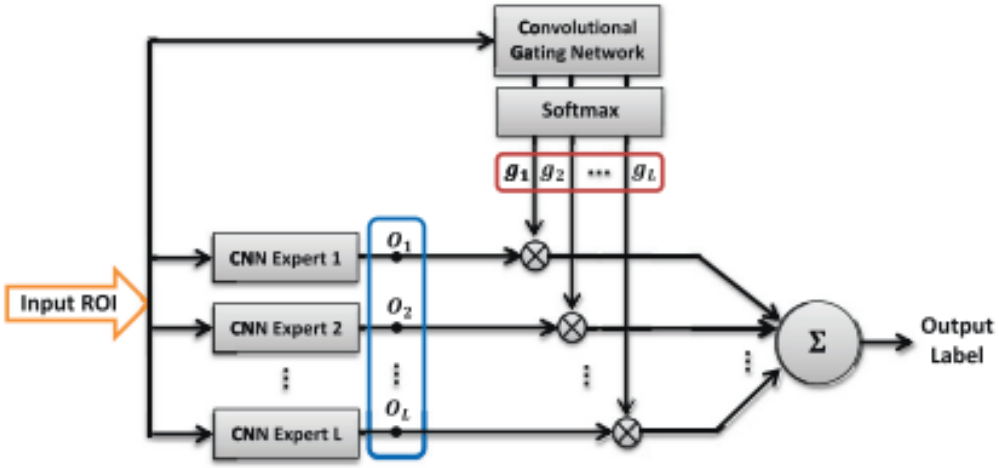

Figura 3.4: Diagrama esquemático del modelo de agrupación conjunta de expertos convolucionales propuesto por Rasti et al. [2017].

libres de las CNNs y CGN. De esta manera, cada experto de CNN está capacitado para especializarse en una región del espacio de entrada de alta dimensión, mientras que la red de compuerta convolucional está capacitada para producir pesos de fusión optimizados.

Dicho esto, se analizó el rendimiento de la CNN en la clasificación de tumores benignos de mama en resonancia magnética frente a malignos y se comparó con otros métodos de clasificación y extracción de características. Así, para todas las ROIs, se extrajeron 10 características escalares. Incluían cinco características de forma y cinco características de textura. En el análisis de referencia, se investigó cuatro clasificadores diferentes: perceptrón multicapa, máquina de vectores de soporte, clasificador de bosques aleatorios y red neuronal convolucional. El clasificador CNN logró una tasa de clasificación del 92,77 %, estadísticamente superior a la alcanzada por los otros tres clasificadores.

Por otro lado, también se evaluó el rendimiento de clasificación del método de *mixture ensemble* propuesto. Para obtener información sobre los rendimientos comparativos de este modelo propuesto también se evaluaron otros dos métodos de conjunto convolucional: los denominados *ave-ensemble* y *soft-ensemble*. La precisión diagnóstica para los diferentes métodos de conjunto fue: modelo propuesto con 96,39 %, *soft-ensemble* con 95,18 % y *ave-ensemble* con 93,98 %.

Finalmente, se pudo llegar a la conclusión de que el modelo propuesto tenía una precisión significativamente mayor que el *ave-ensemble* y los métodos de clasificador único y que el modelo propuesto tiene una precisión estadísticamente similar a la del *soft-ensemble*.

La mejora alcanzada en la precisión de la clasificación del modelo propuesto se puede atribuir al hecho de que los expertos de CNN son entrenados simultáneamente de manera competitiva. La CGN está entrenada para desempeñar el papel de ponderación adaptativa.

Aunque la precisión de clasificación del modelo propuesto y el *soft-ensemble* fue similar, el modelo propuesto presenta varias ventajas. En primer lugar, el modelo propuesto utiliza

significativamente menos parámetros libres que el *soft-ensemble* (703 frente a 1046 parámetros) y, por lo tanto, era menos propenso al sobreajuste. Esto es útil para aplicaciones de diagnóstico médico donde la cantidad de muestras de entrenamiento suele ser pequeña debido a la cantidad limitada de pacientes. En segundo lugar, el modelo propuesto toma menos iteraciones de entrenamiento (épocas) convergentes que el *soft-ensemble*. En tercer lugar, en el experimento realizado, a medida que aumentaba el número de expertos, la precisión de clasificación del modelo propuesto no se deterioraba tan rápido como en el caso del *soft-ensemble*.

En una nueva apuesta por el aprendizaje por transferencia o *transfer learning*, Meng et al. [2022] utilizaron diferentes estrategias para el ajuste fino de la red neuronal DenseNet201 con el objetivo de explorar la eficiencia de identificación de este modelo con respecto a la diferenciación entre lesiones mamarias benignas y malignas en DCE-MRI. DenseNet es una innovación tecnológica que implica la introducción de conexiones de acceso directo para superar los problemas de entrenamiento de redes más profundas [Huang et al., 2017]. El objetivo era encontrar un modelo de *deep transfer learning* (DTL) más preciso para la clasificación y diagnóstico de las lesiones mamarias. Se recuerda que este aprendizaje profundo por transferencia es el proceso de transferir conocimiento de una tarea que ya se aprendió a una nueva tarea.

Las imágenes empleadas en el estudio se obtuvieron en 6 fases (1 fase precontraste y 5 fases poscontraste) y se seleccionó una serie de 12 a 54 imágenes para cada lesión. Para eliminar las señales de interferencia pertenecientes a otros tejidos (como la aorta), las imágenes se recortaron (utilizando Photoshop) y se conservaron los fragmentos de imagen que contenían el tejido mamario. En total, se recolectaron 8400 imágenes DCE-MRI de mama (en promedio, 27 imágenes por paciente), incluidas 4260 imágenes de lesiones benignas y 4140 imágenes de lesiones malignas. Cada grupo de imágenes se dividió aleatoriamente en un conjunto de entrenamiento y un conjunto de prueba. Además, un aumento de datos se realizó antes del entrenamiento del modelo.

DenseNet consta de un bloque denso, una capa de transición y una capa de cuello de botella. El bloque DenseNet revisa la concatenación secuencial de todos los mapas de características en el modelo. DenseNet confiere varias ventajas, como la capacidad de reutilizar características, reducir la explosión de características, además de estar también este modelo asociado con menos problemas de desaparición de gradiente [Huang et al., 2017]. Se eligió DenseNet201 como la columna vertebral para desarrollar un sistema de diagnóstico de lesiones mamarias porque proporciona el mejor rendimiento basado en la tarea de clasificación de ImageNet.

La arquitectura del modelo de DTL se divide en 3 partes: extracción de características, entrenamiento y prueba de datos y validación del modelo. Los principales hiperparámetros se establecieron de la siguiente manera: se utilizó la entropía cruzada binaria como función de pérdida, la optimización se basó en el optimizador de Adam, la tasa de aprendizaje se

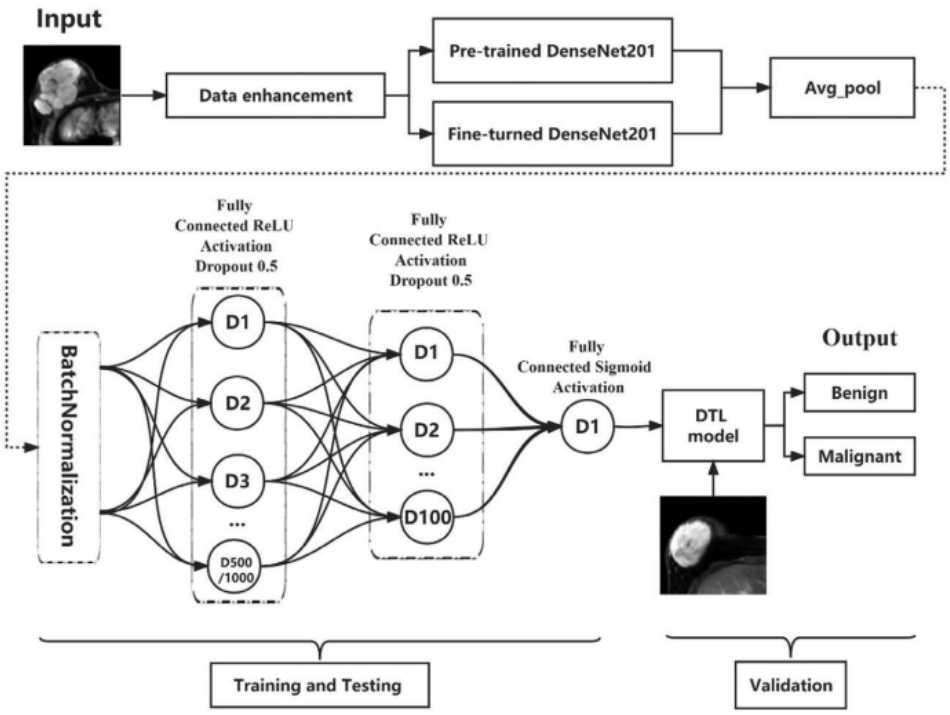

Figura 3.5: Arquitectura del modelo propuesto por Meng et al. [2022].

estableció en 0.0001, el dropout se estableció en 0.5, el número de épocas se fijó en 40, se utilizó ReLu como función de activación y la función sigmoide como función de clasificación. En la Figura 3.5 se expone una visualización de la arquitectura empleada en este estudio.

El rendimiento de DenseNet se puede mejorar mediante un ajuste fino. Así, en este estudio que se está comentando se buscó mejorar el rendimiento del modelo DenseNet201 ideando cuatro estrategias de ajuste fino. Los parámetros de la red neuronal se activaron y utilizaron en el proceso de entrenamiento del modelo, mientras que los parámetros de las capas que se mantuvieron congeladas no participaron en el entrenamiento del modelo. La red de extracción de características no se alteró en las capas de congelación, pero se redujo la cantidad de parámetros que se requerían entrenar. Este hecho puede ahorrar tiempo de entrenamiento y recursos de espacio.

En base a los resultados obtenidos, se llegó a la conclusión de que la segunda estrategia de ajuste fino era la mejor. Tras esta conclusión, se evaluó dicho modelo por medio de una validación cruzada.

Por otro lado, se compuso un mapa de activación de clase combinando la imagen de entrada y el mapa de calor. Dicho mapa puede ayudar a identificar las partes de la imagen en las que se enfocaba el modelo al hacer la predicción final y, por lo tanto, puede proporcionar información sobre el funcionamiento del modelo. El mapa de calor es un mapa de localización aproximado que resalta las regiones importantes para el objetivo de clasificación. Tal análisis puede ayudar aún más en el ajuste de hiperparámetros y ayuda a obtener una comprensión de la razón que subyace a la falla de un modelo. Un ejemplo de este mapa de calor tan

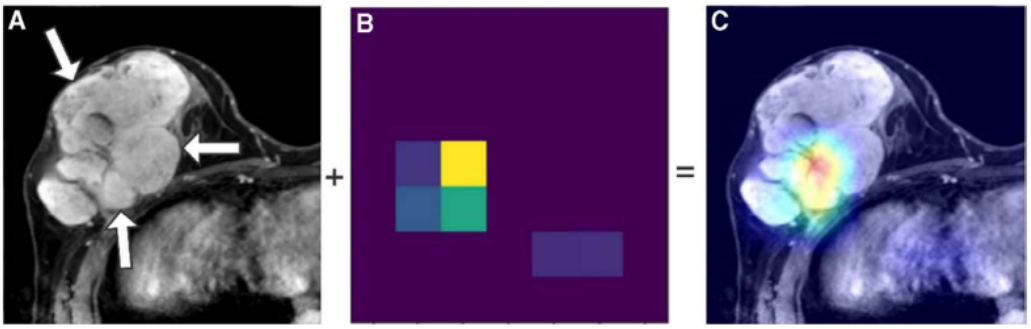

Figura 3.6: Mapa de calor que marca las zonas en las que se enfocaba el modelo propuesto por Meng et al. [2022].

informativo se muestra en la Figura 3.6.

Los resultados logrados por el mejor modelo de los analizados en relación a la clasificación en los grupos benigno y maligno fueron una precisión, exhaustividad, medida F1 y AUROC en el conjunto de validación de 89 %, 80 %, 0.81 y 0.79 respectivamente, siendo éstos superiores a los de los tres modelos restantes. La exactitud para discriminar entre lesiones mamarias benignas y malignas para el mejor modelo fue del 75 %, la más alta de las alcanzadas por los diferentes modelos.

Otro estudio interesante para revisar del estado del arte es el realizado por Hu et al. [2022], debido a las conclusiones que de él se pueden extraer en materia de modelos de aprendizaje profundo para abordar la segmentación en MRI de mama y por el hecho de hacer uso del mismo conjunto de datos que será empleado en el presente proyecto.

Dicho esto, para poner en situación de la tarea resuelta por Hu et al. [2022], es conveniente comentar, como ya es conocido, que la mamografía es el estándar de atención para detectar el cáncer de mama. Sin embargo, la mamografía adolece de varias limitaciones, siendo importante el contraste reducido entre los tumores y el tejido circundante. La superposición de tejido reduce la sensibilidad de la mamografía, especialmente en mamas densas [Freer, 2015]. La resonancia magnética es la modalidad de imagen mamaria con mayor sensibilidad y proporciona una visualización tridimensional reconstruida de los tejidos, evitando así la superposición.

Múltiples estudios han demostrado que la cantidad relativa de tejido fibroglandular (FGT) con respecto al tejido graso en la resonancia magnética de mama es un factor de riesgo de cáncer de mama [McCormack and dos Santos Silva, 2006]. Aunque la cantidad de FGT generalmente la evalúa cualitativamente un radiólogo, las herramientas cuantitativas pueden medir con mayor precisión el volumen de FGT [Alomaim et al., 2020]. Sin embargo, se adquieren cientos de cortes de imágenes por secuencia de resonancia magnética, lo que requiere una segmentación extensa para los métodos cuantitativos. El proceso manual de segmentación es subjetivo y requiere mucho tiempo, lo que lo hace poco práctico para grandes conjuntos de datos. Alternativamente, la cantidad de FGT se puede evaluar a tra-

vés de técnicas interactivas asistidas por el usuario, como la umbralización. Aunque estas técnicas son más rápidas, todavía se ven afectadas por la variabilidad entre lectores y no son lo suficientemente precisas para lograr una cuantificación confiable del volumen de densidad mamaria. Las técnicas de segmentación automatizadas tienen el potencial de reducir la subjetividad y acelerar el procesamiento de datos.

Teniendo todo esto en cuenta, el propósito de este estudio era desarrollar y validar un método de aprendizaje profundo totalmente automatizado basado en la arquitectura U-Net para la segmentación y cuantificación de mamas y FGT.

En relación a los datos empleados, se hizo uso del conjunto de datos de Duke de MRI de cáncer de mama [Saha et al., 2018]. Como ya se mencionó, este conjunto de datos se introducirá en detalle en el siguiente capítulo. Lo más interesante a recalcar de cara a este proyecto es que del conjunto de datos completo se seleccionaron 127 estudios de resonancia magnética precontraste ponderados en T1 con eco de gradiente saturado de grasa. Para cada estudio se eligieron tres cortes de imagen para la segmentación de mama y tres cortes de imagen para la segmentación FGT. Para garantizar la integridad del conjunto de datos de MRI, se aplicaron las siguientes reglas de selección: para seleccionar imágenes para la segmentación de mama, el volumen de MRI se dividió uniformemente en tercios de los cuales se seleccionó aleatoriamente un segmento de imagen de cada tercio del volumen. Para seleccionar imágenes para la segmentación de FGT, el estudio de resonancia magnética se dividió uniformemente en cuartos de los cuales se seleccionaron aleatoriamente dos cortes de imagen de la mitad media (dado que la mayor parte del tejido fibroglandular se concentra en el medio de la mama) y un corte de imagen se seleccionó aleatoriamente de las secciones del primer o último cuarto del estudio de resonancia magnética.

Para cada imagen de resonancia magnética, se trazaron los contornos exteriores de la mama y el FGT. La segmentación manual se realizó en las imágenes seleccionadas para servir como verdad de referencia. Para la segmentación de las mamas se trazó el contorno exterior de cada mama, mientras que para la segmentación de FGT, se aplicó un umbral manual para clasificar los vóxeles de FGT. Todas las segmentaciones manuales fueron revisadas y confirmadas por un radiólogo de mama con siete años de experiencia.

Una descripción general de la población de estudio utilizada en los conjuntos de entrenamiento, validación y prueba se puede encontrar en la Figura 3.7.

En relación a la arquitectura usada, es adecuado mencionar que la arquitectura U-Net se ha utilizado con éxito en múltiples tareas de segmentación de imágenes médicas [Livne et al., 2019]. Esta arquitectura consta de tres secciones: contracción, cuello de botella y expansión. La sección de contracción está hecha de muchos bloques de contracción. Cada bloque toma una entrada, aplica dos capas de convolución de 3×3 y luego usa un *max pooling* de 2×2. El número de filtros o mapas de características después de cada bloque se duplica para que la arquitectura pueda aprender las estructuras complejas de manera efectiva. La capa más inferior media entre la capa de contracción y la capa de expansión y utiliza dos capas CNN

Characteristic	Patients (n = 127)
Age – yrs	53.1 ± 10.8
Race – no (%)	
• White	91 (71.6)
• Black	28 (22.0)
• Hispanic	3 (2.3)
• Other	5 (4.1)
Menopause Status – no (%)	
• Positive	72 (56.7)
• Negative	53 (41.7)
• Not reported	2 (1.6)

Figura 3.7: Población de estudio empleada por Hu et al. [2022].

de 3 × 3 seguidas de una capa de convolución ascendente de 2 × 2.

Antes de entrenar la red, se preprocesó el conjunto de datos en dos pasos. Primero, se normalizó la intensidad de la imagen entre cero y uno en función de los percentiles de intensidad 5 y 95. A continuación, todas las imágenes se redimensionaron a 512 × 512 mediante interpolación bilineal. Se seleccionaron 100 casos para el entrenamiento y 27 casos para prueba. Durante el proceso de entrenamiento, el 10 % de los casos (10 casos) se eligieron aleatoriamente como conjunto de validación. Los modelos para la segmentación de senos y FGT se entrenaron de forma independiente utilizando la misma estructura de red en U.

El rendimiento del modelo de segmentación automatizado se evaluó con el coeficiente de similitud de Dice (DSC), el cual es una métrica de validación estadística que mide la similitud entre dos conjuntos de datos y es un índice de superposición espacial de uso común para evaluar el rendimiento de los modelos de segmentación de imágenes. Los valores de DSC varían entre 0 (indica que no hay superposición espacial entre dos conjuntos de resultados de segmentación binaria) y 1 (indica superposición completa). Para este estudio, los DSCs se calcularon comparando las segmentaciones de imágenes producidas por el modelo con las segmentaciones manuales en las imágenes seleccionadas. Así, se determinó la precisión del modelo en los conjuntos de entrenamiento y prueba.

Los valores promedio de DSC de las segmentaciones de mama y FGT obtenidos en los conjuntos de validación y prueba fueron tales que, en el conjunto de validación, el DSC para las segmentaciones de mama y FGT fue de 0.870 y 0.744 respectivamente mientras que, en el conjunto de datos de prueba, el DSC para las segmentaciones de mama y FGT fue de 0.879 y 0.730 respectivamente.

Capítulo 4

Materiales y metodología

En este capítulo se van a presentar los materiales que han sido empleados en este proyecto y la metodología aplicada para tratar de alcanzar las respuestas a las cuestiones objetivo que se introdujeron en el Capítulo 1 de este documento.

4.1. Materiales

Los materiales utilizados en este estudio vienen dados fundamentalmente por los datos que han sido usados para el entrenamiento y la evaluación de los modelos desarrollados. Los datos constituyen una parte esencial en toda tarea en la que se desee aplicar técnicas de aprendizaje profundo, como es el caso de este proyecto. Aspectos como la cantidad y la variedad suelen ser deseados para el conjunto de datos. El primer factor permite una mejor generalización de los modelos cuando éstos se aplican a datos diferentes de los empleados para entrenarlos y reduce las posibilidades de que aparezca el fenómeno de sobreajuste. La variedad también es requerida para que un modelo no aprenda únicamente a reconocer un tipo concreto de imágenes, sino que pueda ampliar su capacidad de reconocimiento. Por ejemplo, en la tarea de diagnóstico de cáncer de mama, el objetivo suele ser distinguir aquellos tumores que son malignos de los que son benignos. No resultaría muy adecuado si los datos que se le aportasen a un modelo destinado a satisfacer dicha finalidad contuvieran únicamente tumores malignos. Por otro lado, en el caso en que el objetivo fuera clasificar una imagen según la malignidad del tumor acorde a un determinado parámetro (como será el caso de este estudio), no sería bueno que en el conjunto de datos de entrenamiento predominasen las imágenes de una clase concreta, ya que probablemente el modelo no será capaz de clasificar correctamente una imagen del conjunto de prueba que sea perteneciente a una clase diferente bastante menos presente en el conjunto de entrenamiento. Teniendo esto en cuenta, normalmente es conveniente que cada conjunto de datos usado esté balanceado.

En la presente sección se expondrá, en primer lugar, información relativa a la fuente de los datos empleados y la descripción de sus principales características. Seguidamente, se

efectuará un pequeño análisis estadístico que arroje información adicional del conjunto de datos inicial. Por último, se justificarán los criterios tenidos en cuenta para seleccionar un subconjunto de las múltiples imágenes que conforman el conjunto de datos de partida.

4.1.1. Fuente y descripción del conjunto de datos de partida

Como se mencionó al inicio, se han revisado algunas bases de datos públicas con imágenes de resonancias magnéticas. Es conveniente decir que bases de datos públicas con MRI no abundan y si, además se requiere una base de datos destinada a un tipo concreto de tarea, menos aún. Así, por ejemplo, Mahoro and Akhloufi [2022] recogen diferentes bases de datos con imágenes de mamografías, ultrasonidos, termografía y MRI. Para MRI mencionan dos: la base de datos de imágenes de referencia para evaluar la respuesta a la terapia (RIDER) con MRI de mama y la base de datos de Duke con DCE-MRI de mama. No obstante, la primera base de datos aplica a una tarea que no es abordada en este estudio.

Por tanto, los datos que han sido empleados para este proyecto vienen dados por el conjunto de datos de resonancia magnética de cáncer de mama de Duke. Este conjunto de datos es una colección retrospectiva de una sola institución que contiene 922 pacientes reunidas en el Duke Hospital desde el 1 de enero de 2000 hasta el 23 de marzo de 2014 con cáncer de mama invasivo confirmado por biopsia y resonancia magnética preoperatoria disponible en el Duke Hospital. Por tanto, todas las pacientes pertenecientes a este conjunto se caracterizan por tener cáncer de mama confirmado. Los criterios detallados de inclusión/exclusión se describen en [Saha et al., 2018]. Teniendo esto en cuenta, dentro del conjunto introducido se encuentran los siguientes componentes de datos:

- Datos demográficos, clínicos, patológicos, de tratamiento, de resultados y genómicos: recopilados de una variedad de fuentes, incluidas notas clínicas, informes de radiología e informes de patología y han servido como fuente para múltiples artículos publicados sobre radiogenómica, predicción de resultados y otras áreas.

- Resonancia magnética preoperatoria mejorada con contraste dinámico (DCE): descargada de los sistemas PACS y anonimizada para la publicación de The Cancer Imaging Archive (TCIA). Éstos incluyen imágenes de MRI de mama axial adquiridas por escáneres de 1.5T o 3T en las posiciones prona. Las siguientes secuencias de resonancia magnética se comparten en formato DICOM: una secuencia ponderada en T1 sin saturación de grasa, una secuencia precontraste ponderada en T1 con eco de gradiente y saturada de grasa y, en su mayoría, tres o cuatro secuencias poscontraste.

- Ubicaciones de lesiones en DCE-MRI: anotaciones en las imágenes de DCE-MRI por radiólogos.

- Características de imágenes de DCE-MRI: un conjunto de 529 características de imágenes extraídas por computadora por software interno. Estas características representan

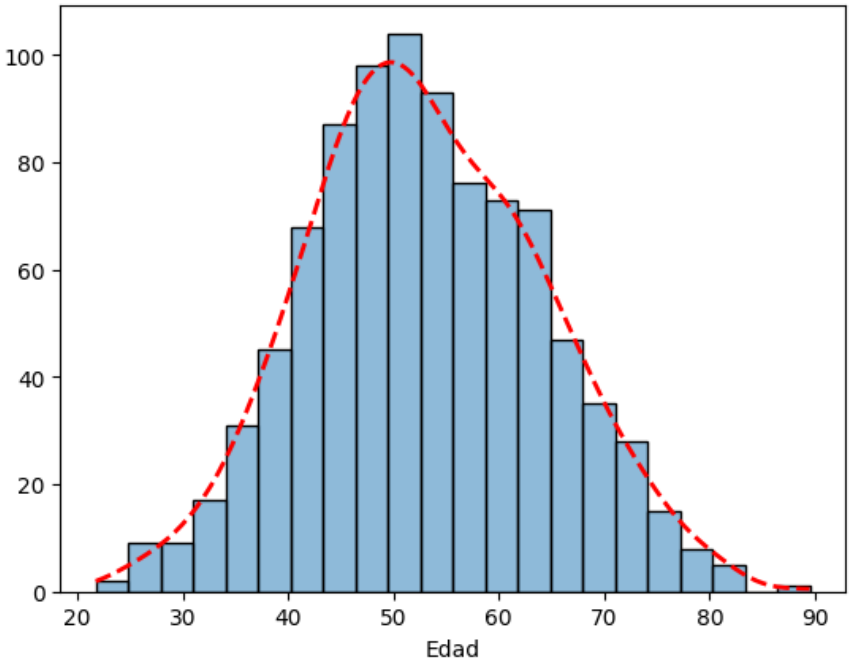

Figura 4.1: Histograma de las pacientes según la edad.

una variedad de características de imágenes que incluyen el tamaño, la forma, la textura y el realce tanto del tumor como del tejido circundante, que se combinan con las características comúnmente publicadas en la literatura, así como las características desarrolladas en laboratorio.

4.1.2. Estadística descriptiva del conjunto de datos de partida

Se desea presentar en esta sección una pequeña estadística descriptiva de los datos empleados en este trabajo exploratorio. El objetivo de este análisis estadístico es identificar posibles tendencias o limitaciones existentes en los datos, lo cual se traducirá en última instancia en restricciones o limitaciones para los modelos desarrollados al estar éstos entrenados con dichos datos. Debido a la gran cantidad de características presentes en el conjunto de datos, sólo se va a comentar acerca de las consideradas como más interesantes para tener una pequeña idea de aspectos que caracterizan a dicho conjunto de datos.

De esta manera, uno de los tipos de características que se han valorado como ilustrativas por la información que aportan es el de características demográficas de las pacientes a las que corresponden las resonancias magnéticas de mama. Éstas vienen dadas por la edad, la menopausia y la raza. En primer lugar, en lo que se refiere a la edad, se desea comprobar si las imágenes tomadas pertenecen o no a pacientes que se encuentren en un grupo concreto de edad. Por ello, se ha obtenido el histograma correspondiente con el objetivo de ver la distribución de éstas por edad. Este histograma aparece representado en la Figura 4.1.

Como se puede comprobar a partir de la Figura 4.1, la distribución por edad de las

Tabla 4.1: Distribución de las pacientes según si presentan o no menopausia en el momento del diagnóstico.

No presentan menopausia	**Presentan menopausia**	**No hay registro**
407	499	16

pacientes se asemeja a una distribución normal que abarca un rango que va desde los 22 a los 90 años aproximadamente. La media de edad se encuentra en torno a los 52 años, por lo que una buena parte de estas personas pertenece a un grupo medio de edad como es el que va desde los 40 a los 65 años.

En lo que respecta a la menopausia, en el conjunto de datos se tienen tres posibles valores: 0 si no tiene menopausia en el momento del diagnóstico, 1 si tiene menopausia en el momento del diagnóstico y 2 si no existe información al respecto para la persona en cuestión. Teniendo esto en cuenta, la Tabla 4.1 muestra la distribución de las pacientes según la menopausia. Como se puede ver, éstas están más o menos bien repartidas entre los grupos que tienen menopausia y no la tienen. Esto es lógico, ya que la edad promedio a la que ocurre la menopausia en los países desarrollados es entre los 51 y 52 años y la media de edad del conjunto de casos considerado para este estudio se sitúa precisamente en los 52 años. No obstante, al haber más pacientes por encima de esos 51-52 años, aparece un mayor número con menopausia que sin ella.

Para cerrar las características demográficas, la raza presenta varias clases en este conjunto de datos (al ser casos tomados en Estados Unidos, estas razas se corresponden con las de dicho país):

- 0 si no hay registro.
- 1 si es blanca (no hispana).
- 2 si es negra.
- 3 si es asiática.
- 4 si es nativa de Alaska.
- 5 si es hispana.
- 6 si es multirracial (dos o más razas).
- 7 si es hawaiana.
- 8 si es indígena americana.

La Figura 4.2 presenta la distribución de las pacientes según su raza. Se puede verificar que las razas blanca y negra son las más predominantes (las dos juntas suponen el 92.6 %

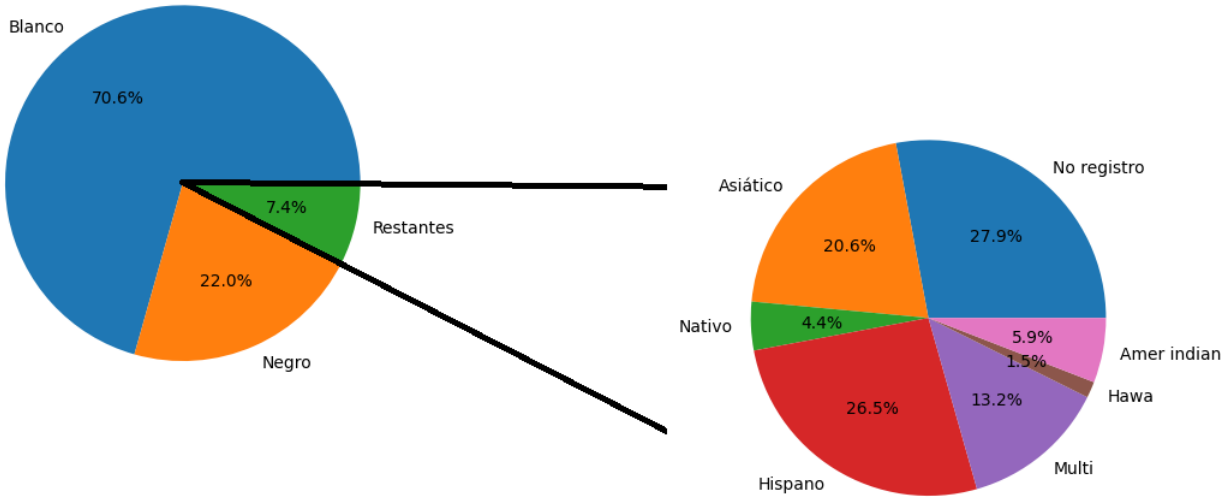

Figura 4.2: Diagrama de sectores de la distribución de las pacientes por su raza.

del total de casos), con la blanca claramente como la ganadora (constituye un 70.6 % de las pacientes). El 7.4 % restante se reparte tal y como aparece representado en la Figura 4.2. Por tanto, queda claro que las imágenes tomadas de casi la totalidad de las pacientes pertenecen a unas razas concretas y, por consiguiente, ello quiere decir que los modelos y resultados alcanzados en este trabajo exploratorio están limitados a una región demográfica específica donde predominan dichas razas, por lo que es posible que la actuación de dichos modelos sea diferente si se le entregan imágenes de personas que sean de una raza distinta (por ejemplo, porque el cáncer de mama en una región diferente se caracterice por algún aspecto que no esté presente en la región estudiada).

Otras características interesantes a analizar del conjunto de datos son las que dan información de si las pacientes han padecido o no metástasis, del tipo histológico, si han requerido o no cirugía y si han sufrido o no recurrencia del tumor. Las conclusiones que se han extraído en relación a estos aspectos son las siguientes:

- El conjunto de pacientes considerado para el presente estudio se caracteriza casi en su totalidad por no presentar metástasis (893 casos de 922).

- En relación al tipo histológico (el cual puede tomar un valor de entre 11 clases diferentes), cabe decir que no hay datos registrados para todas las pacientes. Concretamente, de las 646 personas que tienen tipo histológico registrado, la mayor parte de ellas (575) se corresponden con el tipo ductal y 63 se asocian al tipo lobular. El resto (8) se repar-

Tabla 4.2: Distribución de las pacientes según el tipo histológico.

DCIS	**Ductal**	**Lobular**	**Metaplástico**	**LCIS**	**Tubular**	**Mixto**
1	575	63	1	0	2	0

Micropapilar	Coloide	Mucinoso	Medular
0	0	4	0

ten entre los tipos DCIS, metaplástico, tubular y mucinoso. Esto se puede comprobar a partir de la Tabla 4.2.

- En cuanto a la necesidad o no de cirugía, de entre los 915 casos con registro, 879 requirieron cirugía, mientras que 36 de ellos no la requirieron.
- Según si las pacientes experimentaron o no recurrencia del tumor, de entre las 920 con registro, 833 no presentaron recurrencia, mientras que las 87 restantes sí la presentaron.

Por último, resulta también de gran importancia comprobar cómo se distribuyen los casos estudiados según la característica que se desea predecir. Tal y como se mencionó anteriormente en este documento, la característica que se buscará predecir es el denominado grado de Nottingham, el cual admite tres valores o clases para medir el grado de malignidad del cáncer de mama: 1, 2 y 3. Es interesante ver si estas clases están equilibradas en el conjunto de datos, ya que no sería idóneo tener un conjunto en el que predominase una clase por encima del resto, puesto que entonces el modelo estará especialmente entrenado para reconocer tumores cuya clase se corresponda con ésa más repetida, mientras que el resto de clases las predecirá peor. Así, la Figura 4.3 presenta la distribución de las 638 pacientes que tienen un valor registrado para el grado de Nottingham y la Tabla 4.3 muestra el recuento para cada una de las clases.

4.1.3. Subconjunto seleccionado a partir del conjunto de datos inicial y criterios de selección

Una vez se cuenta con información más detallada acerca del conjunto de datos inicial empleado en este estudio, es adecuado añadir que sólo se ha utilizado una pequeña parte de esos datos. Concretamente, se han usado imágenes de resonancia magnética de mama en formato DICOM y una característica seleccionada para ser predicha por los modelos desarrollados en este trabajo, es decir, una característica elegida para ser el objetivo de predicción en el problema de clasificación abordado.

Además, en relación al gran conjunto de imágenes con el que se cuenta, sólo una pequeña fracción ha sido verdaderamente utilizada. Así, el criterio de selección de imágenes que se ha tenido en cuenta siguiendo las tendencias concluidas de los artículos procedentes del estado

Tabla 4.3: Distribución de las pacientes según el grado de Nottingham.

1	2	3
113	318	207

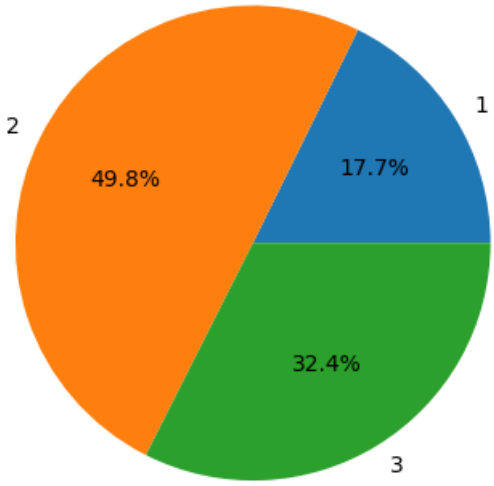

Figura 4.3: Diagrama de sectores de la distribución de las pacientes según el grado de Nottingham de su tumor.

del arte ha sido el de escoger una imagen precontraste y dos imágenes poscontraste para cada paciente. Esto ha dado lugar a un conjunto total de 2766 imágenes con el que se ha alimentado a los diferentes modelos de redes neuronales generados. Se ha considerado que este número de imágenes es un número más que suficiente para poder acometer un correcto entrenamiento y evaluación de los modelos. En lo que se refiere a las dos imágenes poscontraste, se ha valorado adecuado el usar marcos de tiempo (o *time frames*) no consecutivos, por lo que, para una paciente que tenga tres conjuntos de imágenes poscontraste dados para tres marcos de tiempo diferentes, las dos imágenes se han tomado del primer y tercer conjuntos, mientras que, para otra que tenga cuatro conjuntos de imágenes poscontraste para cuatro marcos de tiempo diferentes, las dos imágenes se han tomado del segundo y cuarto conjuntos. A modo de ejemplo, la Figura 4.4 representa una resonancia magnética de mama precontraste y la Figura 4.5 representa una resonancia magnética de mama poscontraste.

Por otro lado, cada imagen precontraste y poscontraste ha sido seleccionada de un conjunto grande de imágenes que representa una evolución dinámica de las mamas. Por lo tanto, no todas las imágenes presentes en ese conjunto grande son adecuadas para la selección, pues pueden ser imágenes donde se vean las mamas con un tamaño reducido o demasiado oscuras. Es por ello, que los criterios que se han considerado a la hora de escoger cada imagen han sido los de tomar aquélla en la que las mamas estén más expandidas o mayor superficie de ellas se vea y que se vean con el mayor contraste posible. Por ejemplo, una imagen que satisface estas premisas aparece en la Figura 4.6, mientras que otra que no las satisface es representada en la Figura 4.7. Estos aspectos comentados acerca de los datos han sido valorados como claves a la hora de ayudar lo máximo posible en el entrenamiento de los modelos. No

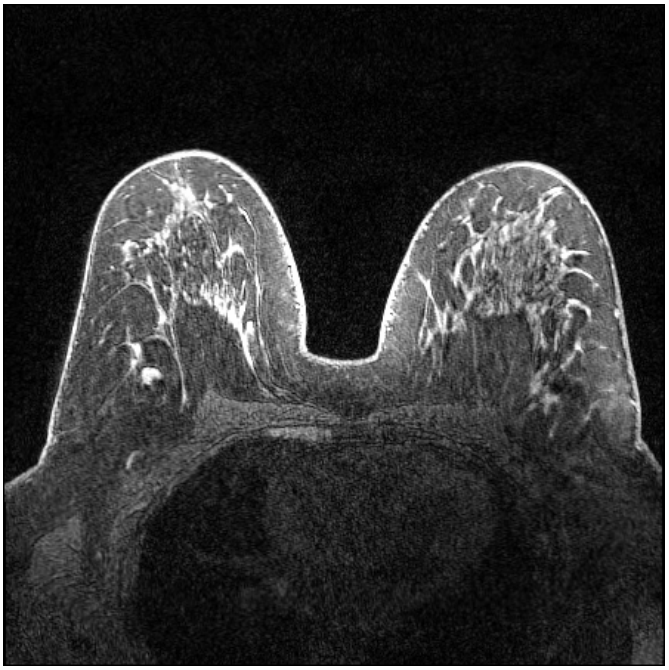

Figura 4.4: Ejemplo de resonancia magnética de mama precontraste.

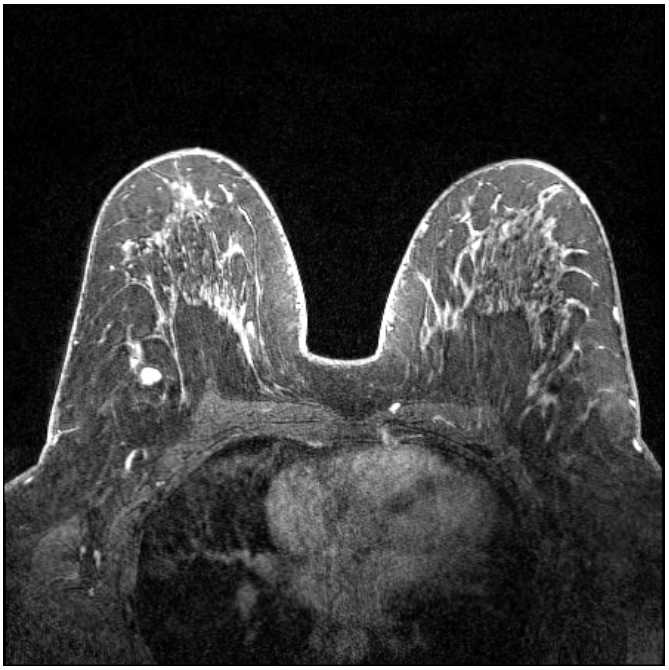

Figura 4.5: Ejemplo de resonancia magnética de mama poscontraste.

Figura 4.6: Ejemplo de resonancia magnética con adecuada extensión y contraste de las mamas.

obstante, como se explicará más adelante, también se ha abordado, por ejemplo, la técnica del aumento de datos y la aplicación de ciertos filtros de imágenes en un preprocesamiento de éstas para analizar la influencia que estas técnicas puedan tener en la actuación de los modelos.

Hasta ahora se ha hablado de los datos con los que se alimenta directamente a los modelos de redes neuronales desarrollados, los cuales vienen dados por las imágenes de resonancias magnéticas de mamas. Sin embargo, es conveniente hablar también de la característica que ha sido seleccionada como objetivo de predicción para el problema de clasificación afrontado. Dado que, tal y como ya se mencionó, las resonancias recogidas en el conjunto de datos corresponden a tumores malignos, se ha considerado interesante clasificar éstas según alguna característica que establezca el grado de malignidad del tumor. De esta forma, se ha escogido concretamente una característica propia del tumor de mama, la cual viene dada por el ya introducido grado de Nottingham. Como ya se comentó anteriormente, en el conjunto de datos de Duke esta característica toma un valor de entre tres clases diferentes que marcan el nivel progresivo de malignidad: 1, 2 ó 3.

4.2. Metodología

En esta sección se pretenden mostrar todos los aspectos relacionados con el desarrollo de modelos de redes neuronales que se ha llevado a cabo en este estudio y que han permitido la correcta consecución de los objetivos marcados al inicio de este documento. De esta forma,

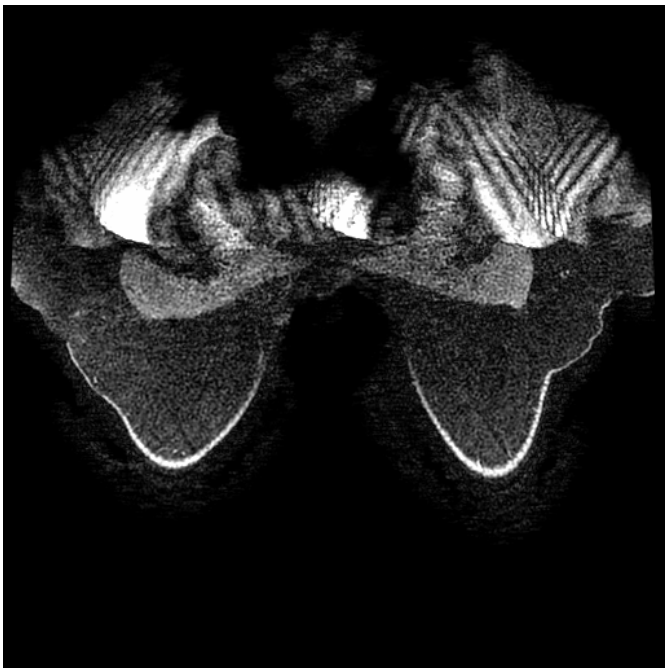

Figura 4.7: Ejemplo de resonancia magnética con inadecuada extensión y contraste de las mamas.

se van a abordar temas que están asociados con la arquitectura asumida para los modelos, el preprocesamiento de las imágenes antes de alimentar a dichos modelos, los parámetros considerados para el entrenamiento de las redes neuronales, etc. Por otro lado, en esta sección se va a exponer también el planteamiento de los diferentes experimentos realizados en este trabajo de exploración. Como se verá más adelante, estos experimentos irán incluyendo pequeñas diferencias entre ellos para comprobar la influencia que ciertos aspectos o técnicas tienen sobre los modelos desarrollados. Así, se revisará, por ejemplo, el efecto que tienen técnicas como la aplicación de filtros de imágenes y el aumento de datos. Siempre la manera empleada para comprobar la influencia que tienen estas técnicas o aspectos sobre los modelos será comparar las correspondientes métricas de evaluación, las cuales miden la actuación de los modelos sobre un determinado conjunto de validación, entre el caso en que se usa la técnica o elemento en cuestión y el caso en que no se utiliza. Estas métricas de evaluación vendrán dadas por la exactitud, la precisión y el valor F (llamado también medida o puntuación F1), entre otras, además de la correspondiente matriz de confusión.

Dicho esto, la presente sección se va a estructurar de la siguiente manera:

- Infraestructuras usadas: Google Colab y Google Drive.
- Modelos desarrollados.
- Procedimiento general de los experimentos: preprocesamiento, creación de modelo, entrenamiento y evaluación.

Figura 4.8: Combinación de herramientas usadas en este trabajo exploratorio.

- Técnicas de preprocesamiento.
- Experimentos realizados.

4.2.1. Infraestructuras usadas: Google Colab y Google Drive

Se va a comenzar introduciendo los marcos o frameworks que han posibilitado la correcta ejecución de los experimentos, ya que han permitido implementar y poner en marcha las diferentes etapas de los experimentos que se explicarán más tarde (estas etapas son fundamentalmente preprocesamiento, creación de modelo, entrenamiento y evaluación). Estas herramientas o frameworks son Google Colab y Google Drive, tal y como se muestra en la Figura 4.8.

Google Colab

Colab, también conocido como "Colaboratory", permite programar y ejecutar Python en el navegador con las siguientes ventajas:

- No requiere configuración.
- Acceso a GPUs sin coste adicional.
- Permite compartir contenido fácilmente.

Google Colab hace uso de un entorno interactivo denominado cuaderno de Colab que permite escribir y ejecutar código. Éstos no son más que cuadernos de Jupyter alojados en Colab.

Los cuadernos de Colab permiten combinar código ejecutable y texto enriquecido en un mismo documento, además de imágenes, HTML, LaTeX y mucho más. Los cuadernos creados en Colab se almacenan en una cuenta de Google Drive y se pueden compartir fácilmente,

permitiendo comentarlos o incluso editarlos. Además, es posible importar los datos propios a los cuadernos de Colab desde la cuenta de Google Drive, incluidas las hojas de cálculo, y también desde GitHub y muchas fuentes más.

Con Colab, se puede importar un conjunto de datos de imágenes, entrenar un clasificador de imágenes con dicho conjunto de datos y evaluar el modelo con tan solo usar unas pocas líneas de código. Los cuadernos de Colab ejecutan código en los servidores en la nube de Google, lo que permite aprovechar la potencia del hardware de Google, incluidas las GPU y TPU, independientemente de la potencia del equipo personal. Lo único que se requiere es un navegador.

Tras esta descripción hecha de Google Colab, cabe decir que este servicio de Google ha constituido la herramienta o marco fundamental que ha posibilitado la creación, entrenamiento y evaluación de los modelos, además de la implementación de técnicas de preprocesamiento de los datos y de la generación de ciertas figuras representativas como son las matrices de confusión o la evolución de la pérdida y la exactitud durante el entrenamiento. Es interesante mencionar también que se ha hecho uso de la potente característica que ofrece Colab de poder vincularse con una cuenta de Google Drive para poder así transferir los datos almacenados en carpetas de Drive a Colab y alimentar en última instancia a los modelos desarrollados. De entre lo comentado anteriormente acerca de Colab, cabe destacar la principal ventaja que otorga en relación al estudio concreto que en este trabajo se lleva a cabo: la posibilidad de uso de máquinas de tipo GPU y TPU. Sin éstas, el entrenamiento de los modelos de redes neuronales habría requerido tiempos de ejecución demasiado elevados e inadmisibles para los tiempos de realización necesarios para este Trabajo Fin de Máster.

Google Drive

Google Drive es un servicio de almacenamiento de datos que son guardados en la nube (plataforma en línea a la que se accede desde cualquier dispositivo con conexión a Internet). El servicio de almacenamiento de Google Drive ofrece 15 GB de capacidad sin costo para cada usuario de Gmail. Además, ofrece planes de pago que disponen de mayor espacio en la nube, entre otros beneficios.

Google Drive permite copiar archivos desde el ordenador para que sean guardados en la nube.

El servicio de Google Drive se caracteriza por:

- Ofrecer una versión sin costo que dispone de hasta 15 GB de almacenamiento.
- Ofrecer opciones con abono que disponen de mayor cantidad de espacio de almacenamiento, entre 25 GB y 100 GB.
- Ser compatible con los sistemas operativos Android, Linux, Mac y Windows.

- Almacenar cualquier tipo de archivo, como fotos, vídeos, archivos de presentación, plantillas, entre otros.
- Mantener todos los archivos en modo privado, excepto que el usuario los designe como públicos o visibles para determinados contactos.
- Compartir documentos de trabajo para que otros usuarios puedan verlos o editarlos en tiempo real.
- Ofrecer un servicio de almacenamiento seguro para los archivos, evitando que resulten afectados ante un posible desperfecto del ordenador del usuario.

Una vez comentadas las principales características de Google Drive, es importante mencionar que éste ha sido empleado con el principal objetivo de constituir un almacén de datos para los análisis efectuados. Concretamente, los datos fueron descargados primeramente de la fuente ya comentada [Saha et al., 2018] y se almacenaron en una carpeta local. Posteriormente, estos datos fueron transferidos a Google Drive, sirviendo éste como puente o conexión entre el almacenamiento local y Google Colab. De esta manera, cuando se llevaban a cabo los experimentos en Google Colab, se procedía a montar Google Drive en Colab (o dicho de otra forma, vincularlos).

No sólo se han almacenado los datos inicialmente obtenidos con los que se deseaba alimentar a los modelos de redes neuronales, sino también datos de imágenes producidas por técnicas como la del aumento de datos para comprobar que las imágenes resultantes fuesen realistas antes de ser aportadas a los modelos para su entrenamiento.

4.2.2. Modelos desarrollados

Tras la presentación de los frameworks utilizados para la implementación y ejecución de los experimentos, el siguiente punto clave a tratar en relación a la metodología puesta en práctica en este estudio es el de las consideraciones tenidas en cuenta a la hora de desarrollar los modelos de predicción.

En primer lugar, una de las conclusiones claramente extraídas del estado del arte es el uso extendido de las redes neuronales convolucionales en el campo del reconocimiento de imágenes de resonancia magnética para cáncer de mama. Es por ello, que la arquitectura asumida en los modelos generados en este proyecto viene dada por la CNN.

Los modelos de redes neuronales convolucionales que se han desarrollado en este trabajo exploratorio se han construido a partir de diferentes tipos de capas, las cuales se van a detallar a continuación. Estas capas usadas se corresponden con las clásicamente empleadas en la construcción de los modelos de redes neuronales convolucionales, entre las que se encuentran las capas convolucional, de pooling, de normalización por lotes y de dropout (estas capas ya fueron introducidas en el marco teórico del Capítulo 2). Todas estas capas usadas se han obtenido en Python del paquete denominado *keras.layers*.

La primera capa utilizada en los modelos se conoce en Keras como *Input* y permite introducir en el modelo un dato de entrada o input de unas determinadas dimensiones. Tal y como ya se comentó en la Sección 4.1 correspondiente a los datos empleados, se ha decidido extraer tres imágenes para cada paciente: una precontraste y dos poscontraste. Por tanto, como se quiere alimentar a la red neuronal con tres imágenes para cada paciente, los modelos desarrollados han sido modelos multi-input. Teniendo esto en cuenta, se ha hecho uso de tres capas *Input* para que cada modelo pueda recibir cada una de las tres imágenes de entrada deseadas.

Otra capa muy importante usada como consecuencia de crear modelos multi-input es la denominada en Keras como *Concatenate*. Como indica su nombre, esta capa permite concatenar varios inputs, por lo que son varias sus entradas y una única su salida.

La capa más característica de las usadas en un modelo de tipo CNN es sin duda la denominada en Keras como *Conv2D*. Como ya se explicó previamente, la arquitectura de la CNN con varias de estas capas permite que la red se concentre en características pequeñas de bajo nivel en la primera capa oculta, luego las ensambla en características más grandes de nivel superior en la siguiente capa oculta y así sucesivamente. Esta estructura jerárquica es común en las imágenes del mundo real, lo cual es una de las razones por las que las CNNs funcionan tan bien para el reconocimiento de imágenes. Los parámetros principales a especificar en relación a esta capa son:

- Número de filtros. Se establece el número de filtros o kernels de convolución a aplicar en esa capa sobre la imagen. Una capa llena de neuronas que usa el mismo filtro genera un mapa de características, el cual resalta las áreas en una imagen que activan más el filtro.
- Tamaño o dimensiones del filtro.
- Función de activación.
- Padding. Para que una capa tenga el mismo alto y ancho que la capa anterior, es común agregar ceros alrededor de las entradas. Esto es lo que se conoce como *zero padding*.

La siguiente capa es la de agrupamiento o pooling. Tal y como ya se comentó, el objetivo de esta capa es el de submuestrear o reducir la imagen de entrada para reducir la carga computacional, el uso de memoria y la cantidad de parámetros (lo que limita el riesgo de sobreajuste). Los parámetros a definir para esta capa son su tamaño, el denominado *stride* y el tipo de *padding*. Una neurona de una capa de pooling no tiene pesos, sino que todo lo que hace es agregar las entradas usando una función de agregación como el máximo o la media. Concretamente, en los modelos se ha usado la capa de pooling cuya función de agregación es el máximo. En Keras esta capa se conoce como *MaxPooling2D*. De esta forma, sólo el valor de entrada máximo en cada campo receptivo pasa a la siguiente capa, mientras que las otras

entradas se eliminan. Además, el *stride* típicamente usado ha sido de 2, por lo que la imagen de salida tiene la mitad de la altura y la mitad del ancho de la imagen de entrada.

Otra capa importante en toda red neuronal es la que representa una capa de neuronas densamente conectadas o, lo que es lo mismo, una capa densa. En Keras ésta se conoce como *Dense*. Los dos parámetros fundamentales a especificar para esta capa son el número de neuronas y el tipo de función de activación. En el modelo global con arquitectura de CNN, las capas de este tipo conforman la parte final del modelo donde destaca concretamente la última de ellas, ya que es donde se aborda concretamente la predicción final. En el presente estudio, la predicción final va asociada a un problema de clasificación, por lo que el número de neuronas de esta última capa densa ha de coincidir con el número de clases de la característica a predecir. Por ejemplo, para los modelos aquí desarrollados que tratan de predecir el grado de Nottingham (el cual se recuerda que tiene tres clases), el número de neuronas requeridas es de tres y, por tanto, la función de activación a usar en este caso es la conocida como *softmax*.

Una capa adicional usada en los modelos es la de aplanamiento, la cual es denominada *Flatten* en Keras. Esta capa permite conectar la parte convolucional del modelo con la parte final formada por capas densas. Se recuerda que esto es debido a que las capas convolucionales y de pooling son capas 2D, mientras que las capas densas son de tipo 1D, por lo que el vector de características ha de ser adaptado al pasar de una parte a la otra del modelo.

En la lucha contra el sobreajuste u *overfitting*, se ha hecho uso también en algunos modelos de la capa conocida en Keras como *Dropout*, definiéndose la correspondiente tasa de desactivación.

Por último, otra capa de gran interés es la denominada en Keras como *BatchNormalization*, la cual aplica la normalización por lotes.

Tras esto, es turno de presentar los modelos concretos que se han propuesto para el análisis exploratorio mostrado en este documento. Los modelos creados se dividen en dos grupos principales: los generados en su totalidad bajo criterio propio y los construidos a partir de modelos ya creados y preentrenados.

En cuanto a los modelos de creación personal, se han propuesto dos modelos principalmente. Ambos son modelos multi-input, ya que, como se comentó, se desea alimentar a la red neuronal con tres imágenes correspondientes a cada paciente. La diferencia entre estos dos modelos radica en el punto del modelo en que se lleva a cabo la concatenación de inputs.

Así, en el primer modelo planteado (denominado de aquí en adelante como modelo con concatenación temprana), la concatenación de inputs se efectúa al inicio del modelo, por lo que la primera capa convolucional recibe ya únicamente una entrada. Un ejemplo de este modelo se muestra en la Figura 4.9.

Por el contrario, en el segundo modelo (denominado de aquí en adelante como modelo con concatenación tardía), la concatenación de inputs se realiza en la parte final del modelo, previo a la capa densa localizada en la parte posterior del modelo. Por tanto, la concatenación

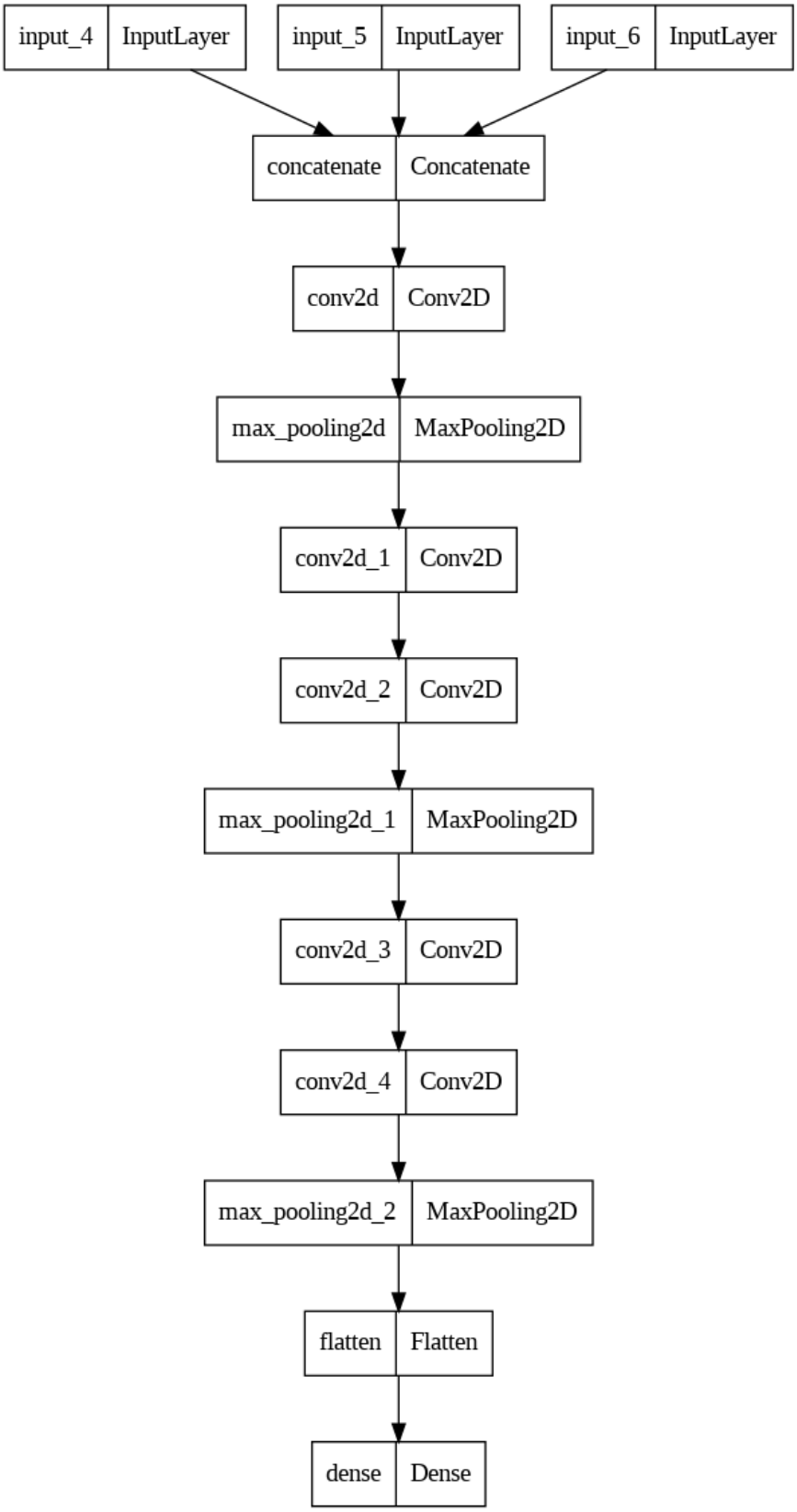

Figura 4.9: Modelo con concatenación temprana.

tiene lugar justo antes de la capa de tipo *Flatten* que conecta la parte de capas convolucionales y de pooling con la parte de capas densas. Un ejemplo de este modelo se muestra en la Figura 4.10.

Además de estos modelos de creación personal, se ha generado también algún modelo a partir de otro ya existente y preentrenado. En este estudio se ha partido concretamente del modelo conocido como MobileNet, el cual es un modelo de clasificación de imágenes que está preentrenado con el conjunto de datos conocido como *ImageNet*. La idea fundamental de este tercer tipo de modelo creado se basa en tomar la base convolucional de la MobileNet y añadirle al final las capas correspondientes encargadas de abordar el problema de clasificación. La Figura 4.11 representa las capas añadidas adicionalmente a la base convolucional tomada del modelo de MobileNet para constituir el modelo final.

4.2.3. Procedimiento general de los experimentos: preprocesamiento, creación de modelo, entrenamiento y evaluación

Hasta ahora se ha aportado información detallada acerca de los frameworks usados y los modelos desarrollados para abordar el problema de clasificación introducido al inicio. No obstante, para cada experimento realizado en este trabajo exploratorio se ha seguido un procedimiento común, el cual consta de una serie de pasos o fases bien diferenciadas. Estas fases son:

- Preprocesamiento de los datos. En este bloque entran aspectos o técnicas que tienen por objetivo adecuar los datos antes de ser alimentados a los modelos.
- Creación del modelo. Con los datos listos, se procede a construir el modelo de aprendizaje profundo, añadiendo las capas que se consideren oportunas de entre las introducidas anteriormente.
- Entrenamiento del modelo. Una vez el modelo está creado, se establecen los parámetros asociados al entrenamiento (número de épocas, algoritmo usado, etc.) y se ejecuta dicho entrenamiento del modelo sobre el conjunto de entrenamiento.
- Evaluación del modelo. Cuando el modelo está ya entrenado, se evalúa éste sobre un conjunto de validación para obtener las correspondientes métricas de evaluación que dan información de la actuación del modelo.

Preprocesamiento de datos

Los datos de partida que se han tomado para este estudio exploratorio son las imágenes seleccionadas del conjunto grande procedente de [Saha et al., 2018] y la característica que

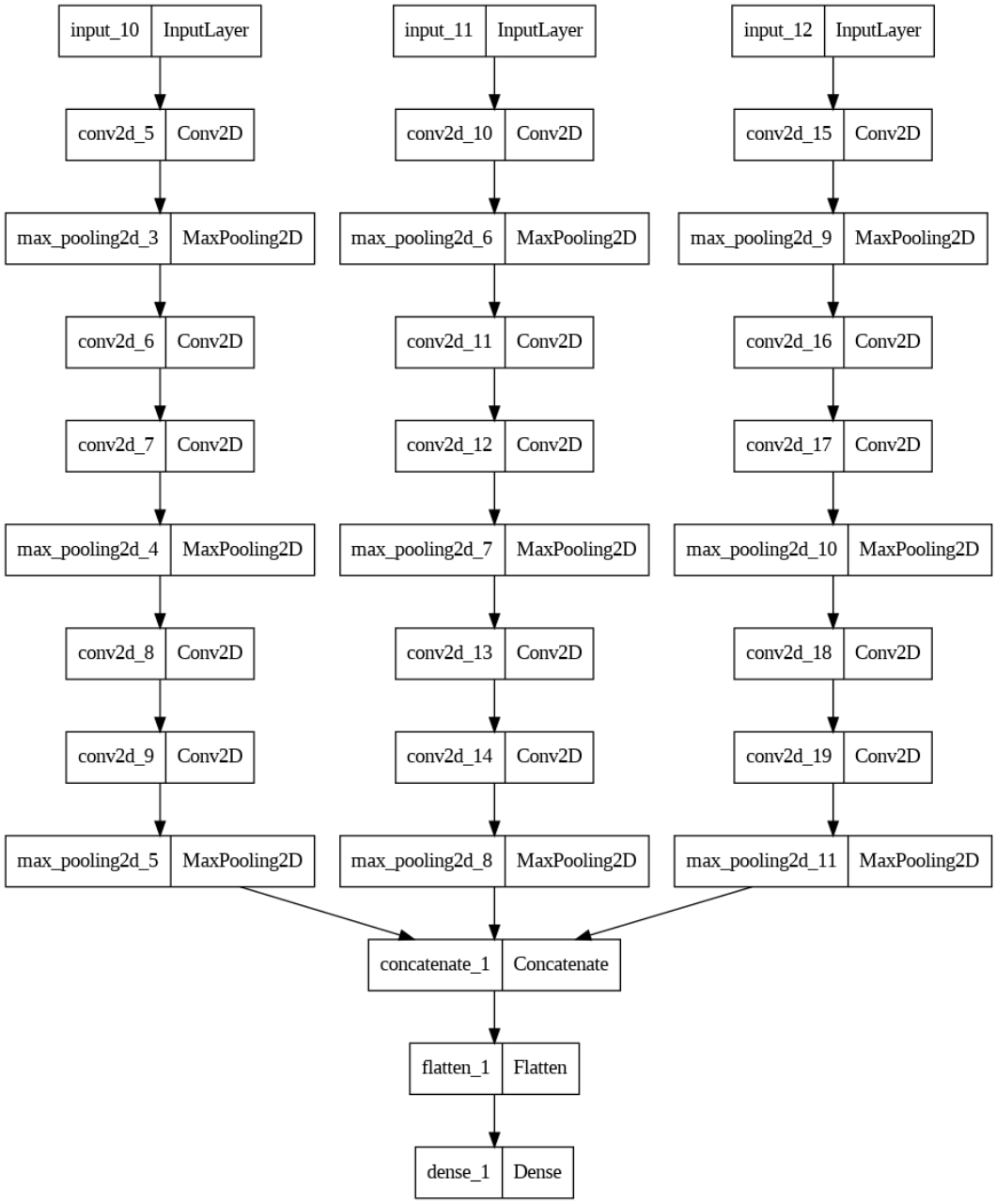

Figura 4.10: Modelo con concatenación tardía.

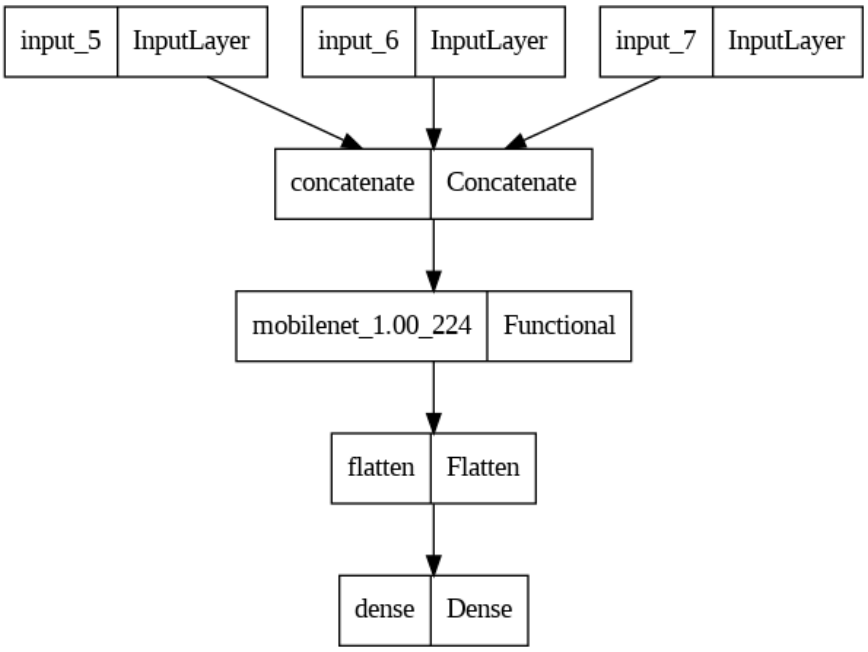

Figura 4.11: Modelo que parte de la base convolucional de una MobileNet preentrenada.

se desea predecir a partir de dichas imágenes. Estos datos no pueden ser entregados directamente a los modelos, sino que requieren un preprocesamiento previo. Por tanto, se van a explicar a continuación algunos de los procesos aplicados en esta línea.

En primer lugar, es importante aclarar que las imágenes de partida vienen en formato DICOM, por lo que se necesita pasar de éste a un formato de imagen como el PNG.

Tanto las imágenes en formato DICOM como en formato PNG han sido almacenadas en carpetas correspondientes de Google Drive. De esta manera, no se requiere efectuar esta conversión cada vez que se realice un experimento diferente.

Debido a que las imágenes obtenidas en formato PNG poseen diferentes dimensiones según la paciente, otro paso acometido que está enmarcado en el campo del preprocesamiento de datos es el de reescalado de las imágenes. Se ha considerado que, dentro del reescalado, lo más ideal es extender la imagen frente a recortarla, ya que en el recorte se podría perder información importante para el modelo. Por tanto, se ha averiguado cuál es el tamaño máximo de las imágenes en formato PNG para extender las que son más pequeñas a dicho tamaño. Estas dimensiones máximas han resultado ser 512×512.

Otro aspecto clave en este preprocesamiento de los datos es la separación del conjunto total de datos en subconjuntos de entrenamiento y validación. El criterio que se ha tomado para acometer esta fragmentación ha sido el de tomar el primer 75 % de los datos para el conjunto de entrenamiento y el 25 % restante para el conjunto de validación. Cabe decir que, una vez se comprobase que se alcanzan modelos con una actuación adecuada, se escogería el mejor de ellos y se añadiría un subconjunto de test a la separación de los datos con el objetivo de realizar un estudio detallado de los hiperparámetros de dicho modelo.

En relación a los datos de las clases a predecir, es interesante mencionar que la red neuronal admite clases que comiencen en 0. Por tanto, en el caso en que se ha deseado predecir el grado de Nottingham por ejemplo, donde las clases vienen dadas por los valores 1, 2 y 3, se ha tenido que restar estos valores en una unidad, pasando las clases a venir dadas por los valores 0, 1 y 2 respectivamente.

Otras técnicas importantes que se han aplicado son las del aumento de datos y los filtros de imágenes. Debido a la mayor extensión de la explicación de estas técnicas dada su importancia, se deja su presentación detallada para una subsección posterior a ésta.

Creación del modelo

Tras el preprocesamiento efectuado de los datos, se está en disposición de crear el modelo deseado. En relación a la creación del modelo, aplica todo lo comentado anteriormente en cuanto a la arquitectura considerada, capas utilizadas, etc. Se ha tenido la costumbre de representar gráficamente cada modelo creado para comprobar que el modelo se ajustaba a lo pretendido.

Entrenamiento del modelo

Con los datos ya preparados de tal forma que los tolere correctamente el modelo y con el modelo ya creado, la siguiente fase se centra en el entrenamiento de dicho modelo. Por tanto, se establecen aquí parámetros de entrenamiento como el tipo de función de pérdida, el algoritmo de optimización y la métrica de evaluación escogida para medir la actuación del modelo durante el entrenamiento. Concretamente, los parámetros de Python escogidos en relación con el entrenamiento para los diferentes experimentos han sido:

- Pérdida: *Sparse categorical cross entropy.*
- Optimizador: *Stochastic gradient descent* (con un determinado valor para el ratio de aprendizaje o *learning rate*).
- Métrica de evaluación: *Accuracy* (al ser un problema de clasificación).

Estos parámetros previos se establecen para el entrenamiento del modelo a través del método *compile()*. Por último, se hace uso también del método *fit()* en el que se indican los conjuntos de datos de entrenamiento y validación, además de poder especificarse el número de épocas y el tamaño del lote, entre otros parámetros.

Durante el entrenamiento del modelo se recibe en Google Colab una serie de resultados para cada época. Estos resultados son la pérdida y la exactitud tanto para el conjunto de entrenamiento como para el de validación. Como es sabido, lo ideal es ver cómo la pérdida va tendiendo a 0 y la exactitud a 1. Además, es importante comprobar que las actuaciones del modelo en ambos conjuntos de datos (entrenamiento y validación) sean parecidas. Si aparece

una divergencia entre ambas actuaciones (es decir, si los valores de pérdida y exactitud para ambos conjuntos tienden a alejarse) esto puede ser un signo claro de que exista un sobreajuste u *overfitting*. Al final del proceso de entrenamiento, resulta muy visual obtener representaciones gráficas de la evolución de la pérdida y la exactitud tanto para el conjunto de entrenamiento como para el de validación.

Evaluación del modelo

Cuando se consigue entrenar correctamente un determinado modelo, es turno de revisar su desempeño en un conjunto de datos de validación. Esto se lleva a cabo en esta última fase del procedimiento ejecutado para cada experimento. En este caso, se han empleado los métodos conocidos en Python como *evaluate()* y *predict()*. Además, se ha hecho uso del paquete *metrics* de la librería *Scikit-Learn* para realizar el cálculo de las diferentes métricas de evaluación deseadas. Las métricas de evaluación escogidas para el problema de clasificación abordado han sido las siguientes:

- Exactitud o *accuracy*. Mide el porcentaje total de casos que el modelo ha clasificado correctamente de entre todas las predicciones realizadas.
- Precisión o *precision*. Es la relación de verdaderos positivos sobre la suma de falsos positivos y verdaderos positivos. Se trata de una métrica que muestra cómo de precisa ha sido la predicción de los casos pronosticados como positivos. Es también llamada valor predictivo positivo.
- Exhaustividad o *recall*. Es la proporción de casos positivos que el modelo es capaz de capturar etiquetándolos como positivos. Es también conocida como sensibilidad.
- Puntuación F1 o *F1 score*. Se utiliza para combinar las medidas de precisión y exhaustividad en un sólo valor. Esto es práctico porque hace más fácil el poder comparar el rendimiento combinado de la precisión y la exhaustividad entre varias soluciones.
- Área bajo la curva ROC o AUC. Mide el área bidimensional completa debajo de la curva de característica operativa del receptor o ROC, la cual es un gráfico que muestra el rendimiento de un modelo de clasificación en todos los umbrales de clasificación. En otras palabras, el AUC proporciona una medida agregada del rendimiento en todos los umbrales de clasificación posibles.

Estas métricas presentadas permiten obtener una idea clara de la verdadera actuación o desempeño del modelo desarrollado.

Por otro lado, resulta muy interesante obtener también la correspondiente matriz de confusión, la cual posibilita el conocimiento de los falsos positivos y negativos y los verdaderos positivos y negativos de una forma bastante visual.

Es importante mencionar que, una vez se comprobase que el modelo en cuestión funcionase bien en un conjunto de validación, sería el turno de acometer un análisis detallado de los hiperparámetros y de evaluar el modelo final en el correspondiente conjunto de datos de prueba.

4.2.4. Técnicas de preprocesamiento

Como se mencionó anteriormente, se exponen ahora con detalle dos técnicas de gran importancia para los modelos de aprendizaje profundo dedicados al reconocimiento de imágenes. Estas técnicas son el aumento de datos y la aplicación de filtros de imágenes.

Aumento de datos

Tal y como se explicó anteriormente, el aumento de datos adopta el enfoque de generar más datos de entrenamiento a partir de los datos disponibles. En el caso de imágenes, esto se consigue aplicando una serie de transformaciones aleatorias a la imagen produciéndose nuevas imágenes de aspecto creíble, de tal forma que durante el entrenamiento el modelo nunca verá exactamente la misma imagen en las diferentes épocas. Esto contribuye a que el modelo se exponga a una mayor variedad de los datos y, en última instancia, a que el modelo generalice mejor.

Se requiere analizar bien la elección de las técnicas específicas de aumento de datos utilizadas, teniendo en cuenta el contexto del conjunto de datos de entrenamiento y el conocimiento del dominio del problema, con el objetivo final de no generar imágenes que nunca podrían darse en la realidad. Es por ello que, de entre la variedad de transformaciones que existen y considerando el dominio dado por resonancias magnéticas de mama, se ha decidido únicamente rotar las imágenes 3 grados, de tal manera que la parte importante de las mismas (las mamas) no se viera oculta al exceder las dimensiones del marco correspondiente de las imágenes.

En Python, esto se puede aplicar de manera sencilla mediante la configuración de las transformaciones que se realizarán en las imágenes leídas por la instancia de la clase *ImageData Generator* de Keras. Posteriormente, se ha hecho uso del método *flow_from_directory* para obtener el generador de datos correspondiente.

Filtros de imágenes

Debido a que la tarea principal abordada en este estudio exploratorio viene dada por la clasificación de imágenes médicas para el diagnóstico del cáncer de mama, se ha considerado que la aplicación de filtros a las imágenes utilizadas puede constituir un factor diferencial en la actuación de los modelos desarrollados. La razón de esto es que los filtros permiten muchas veces ampliar contrastes y facilitar al modelo el aprendizaje de las características más relevantes. Por ello, se ha decidido implementar algunos filtros de imágenes en escala

de grises que están bien contrastados a día de hoy. Estos filtros son: CLAHE (ecualización del histograma) y Sobel/Canny (detección de bordes).

Para aplicar la ecualización de histogramas por medio de Python, se puede hacer uso de una función de *OpenCV* llamada *equalizeHist()*, la cual permite obtener una imagen con un mayor contraste en las intensidades de gris. No obstante, también es cierto que, si bien esta función incrementa el contraste de la imagen, también aumenta el nivel de ruido que pueda tener ésta, además de hacer parecer la imagen como difuminada, especialmente en las regiones más claras. Debido a estos defectos, se ha considerado más conveniente la aplicación de otro algoritmo: el *Contrast Limited Adaptive Histogram Equalization* (o *CLAHE*). Como ya se comentó anteriormente, este algoritmo se enfocará en realizar una redistribución de valores de forma local, dividiendo la imagen en celdas sobre cuyos píxeles se hará el proceso de ecualización y cuyas dimensiones han de definirse buscando un mejor resultado, definiéndose también un umbral límite para dicha transformación. Estas variables son las que se introducen como argumentos de la función *createCLAHE()*, la cual se encarga de la aplicación de dicha técnica.

En lo que se refiere a la detección de bordes, de nuevo la biblioteca OpenCV de Python ha permitido implementar el algoritmo de Sobel usando la función *Sobel()*. Se necesita especificar la función con varios parámetros junto con la imagen. Así, se requiere mencionar la profundidad de la imagen final, especificada con el parámetro *ddepth*. Con el valor de -1, la imagen de salida tendrá la misma profundidad que la imagen de entrada. El orden de las derivadas a utilizar se especifica mediante los parámetros *dx* y *dy*. El tamaño del núcleo Sobel extendido se menciona mediante el parámetro *ksize*. Finalmente, los parámetros *escala* y *delta* son opcionales.

El algoritmo de Canny se ha podido usar por medio de OpenCV mediante la función *cv.Canny*, la cual realiza internamente las etapas que fueron descritas cuando se presentó este algoritmo en el marco teórico del Capítulo 2.

4.2.5. Experimentos realizados

En esta última sección del capítulo dedicado a la metodología aplicada en este trabajo se va a acometer una descripción detallada de cada uno de los experimentos llevados a cabo en este estudio exploratorio. La forma de proceder con cada experimento viene dada por las fases que se presentaron anteriormente: preprocesamiento de los datos, creación del modelo, entrenamiento del modelo y evaluación del mismo. El objetivo de cada uno de estos experimentos es comprobar la influencia o el efecto de algún aspecto o técnica en particular. Este efecto se verá en cómo varían las métricas de evaluación escogidas, ya que uno desea ver si la aplicación o la introducción de algún aspecto novedoso mejora o no la actuación del modelo desarrollado. Los resultados alcanzados para cada uno de los experimentos explicados en esta sección se expondrán en el siguiente capítulo, el cual está enteramente dedicado a los

Tabla 4.4: Parámetros de entrenamiento del modelo del experimento 1.

Parámetros entrenables	4,261,891
Pérdida	Sparse categorical cross entropy
Optimizador	Stochastic Gradient Descent
Ratio de aprendizaje	0.01
Métrica de evaluación	Exactitud
Épocas de entrenamiento	50

resultados obtenidos a lo largo de este estudio. Como las métricas de evaluación seleccionadas han sido la precisión, la exactitud, la matriz de confusión, etc., al tratarse de un problema de clasificación, serán éstas, las que centrarán los resultados expuestos en dicho capítulo.

Dicho esto, se procede a presentar cada uno de los experimentos realizados en el presente trabajo.

Experimento 1. Comprobación de actuación del modelo personal con concatenación temprana de datos

El primer experimento va a tratar de evaluar la actuación o desempeño del primero de los modelos propuestos. Se recuerda que este modelo se caracteriza por ser un modelo multi-input donde la concatenación de entradas se realiza al inicio.

Cabe aclarar que en este primer experimento no se acomete ninguna técnica especial de preprocesamiento de datos como son el aumento de datos o la aplicación de filtros, sino que sólo se han ejecutado los pasos de preprocesamiento esenciales para que los datos puedan ser correctamente entregados al modelo de red neuronal (conversión de imágenes a formato PNG, redimensionamiento de éstas a unas mismas dimensiones, traslación de los valores de las clases de la característica a predecir para que sean bien comprendidos por la red neuronal, etc.).

Es interesante mencionar también que el conjunto de datos de entrenamiento usado en este experimento no está perfectamente balanceado, teniendo la segunda de las tres clases mayor presencia en el mismo.

El modelo creado para este experimento hace uso de las capas que aparecen representadas en la Figura 4.12. Además, los parámetros usados para cada una de estas capas se muestran en la Figura 4.13.

En cuanto al entrenamiento del modelo, el número de parámetros entrenables y los parámetros escogidos se pueden observar en la Tabla 4.4.

Los resultados obtenidos en este experimento tras el entrenamiento del modelo y la evaluación del mismo se presentan en la primera sección del Capítulo 5.

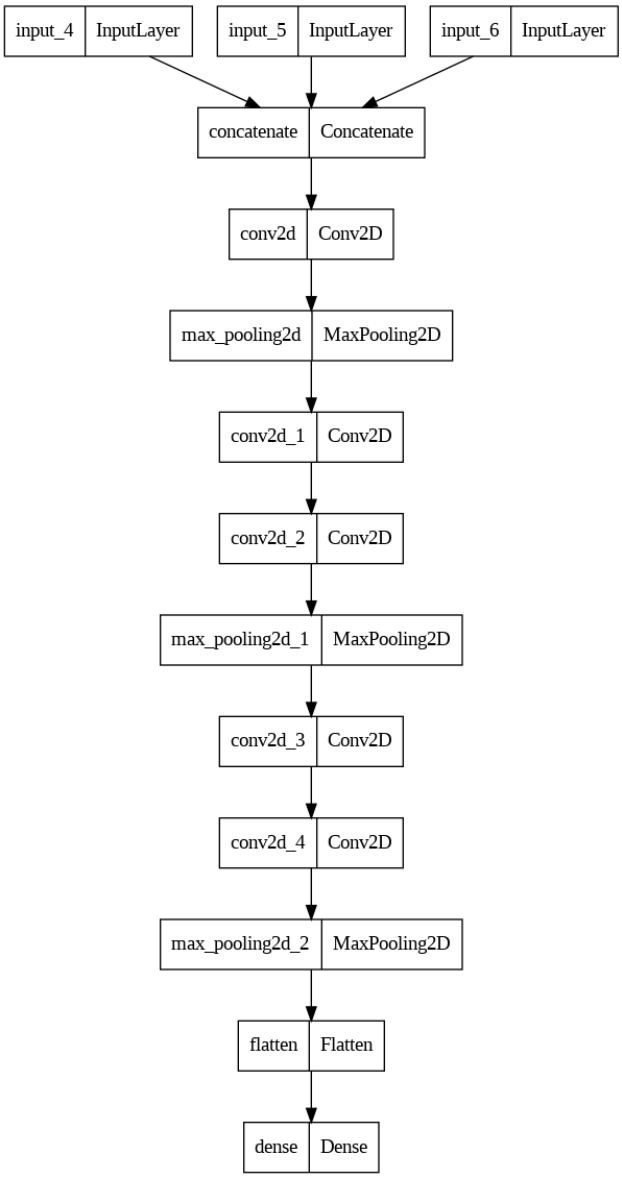

Figura 4.12: Modelo creado para el experimento 1.

Experimento 2. Comprobación de actuación del modelo personal con concatenación tardía de datos

El segundo experimento tiene por objetivo evaluar la actuación o desempeño del segundo de los modelos propuestos. Este modelo se caracteriza por ser un modelo multi-input donde la concatenación de entradas se realiza en la parte final del mismo.

Cabe aclarar de nuevo que en este segundo experimento no se acomete ninguna técnica especial de preprocesamiento de datos como son el aumento de datos o la aplicación de filtros, sino que sólo se han ejecutado los pasos de preprocesamiento esenciales para que los datos puedan ser correctamente entregados al modelo de red neuronal (conversión de imágenes a formato PNG, reescalado de éstas a unas mismas dimensiones, traslación de los valores de las clases de la característica a predecir para que sean bien comprendidos por la red neuronal,

```
input1 = keras.layers.Input(shape=(512,512))
input2 = keras.layers.Input(shape=(512,512))
input3 = keras.layers.Input(shape=(512,512))
input1 = tf.expand_dims(input1, axis=3)
input2 = tf.expand_dims(input2, axis=3)
input3 = tf.expand_dims(input3, axis=3)
concat = keras.layers.Concatenate(axis=3)([input1, input2, input3])
conv1 = keras.layers.Conv2D(64, 7, activation="relu", padding="same")(concat)
pool1 = keras.layers.MaxPooling2D(2, padding='same')(conv1)
conv2 = keras.layers.Conv2D(128, 3, activation="relu", padding="same")(pool1)
conv3 = keras.layers.Conv2D(128, 3, activation="relu", padding="same")(conv2)
pool2 = keras.layers.MaxPooling2D(2, padding='same')(conv3)
conv4 = keras.layers.Conv2D(256, 3, activation="relu", padding="same")(pool2)
conv5 = keras.layers.Conv2D(256, 3, activation="relu", padding="same")(conv4)
pool3 = keras.layers.MaxPooling2D(2, padding='same')(conv5)
flat = keras.layers.Flatten()(pool3)
output = keras.layers.Dense(3, activation="softmax")(flat)
model = keras.Model(inputs=[input1, input2, input3], outputs=[output])
```

Figura 4.13: Parámetros del modelo creado para el experimento 1.

Tabla 4.5: Parámetros de entrenamiento del modelo del experimento 2.

Parámetros entrenables	12,766,851
Pérdida	Sparse categorical cross entropy
Optimizador	Stochastic Gradient Descent
Ratio de aprendizaje	0.01
Métrica de evaluación	Exactitud
Épocas de entrenamiento	20
Pasos para cada época de entrenamiento	100

etc.). Además, se usa el mismo conjunto de datos que en el experimento 1, por lo que el conjunto de datos de entrenamiento usado no está perfectamente balanceado.

El modelo creado para este experimento hace uso de las capas que aparecen representadas en la Figura 4.14. Aparte de esto, los parámetros usados para cada una de estas capas se muestran en la Figura 4.15.

En cuanto al entrenamiento del modelo, el número de parámetros entrenables y los parámetros escogidos han sido aquéllos mostrados en la Tabla 4.5.

Los resultados obtenidos en este experimento tras el entrenamiento del modelo y la evaluación del mismo se presentan en la segunda sección del Capítulo 5.

Experimento 3. Comprobación de actuación del modelo preentrenado

Este tercer experimento tiene como propósito comprobar la actuación o desempeño del tercero de los modelos propuestos. Este modelo se caracteriza por ser un modelo multi-input que parte, a su vez, de un modelo preentrenado como es la MobileNet. Esta MobileNet está preentrenada sobre el conjunto de datos *ImageNet* y se toma concretamente su base convolucional, sobre la cual se añaden las capas requeridas para completar el clasificador. La

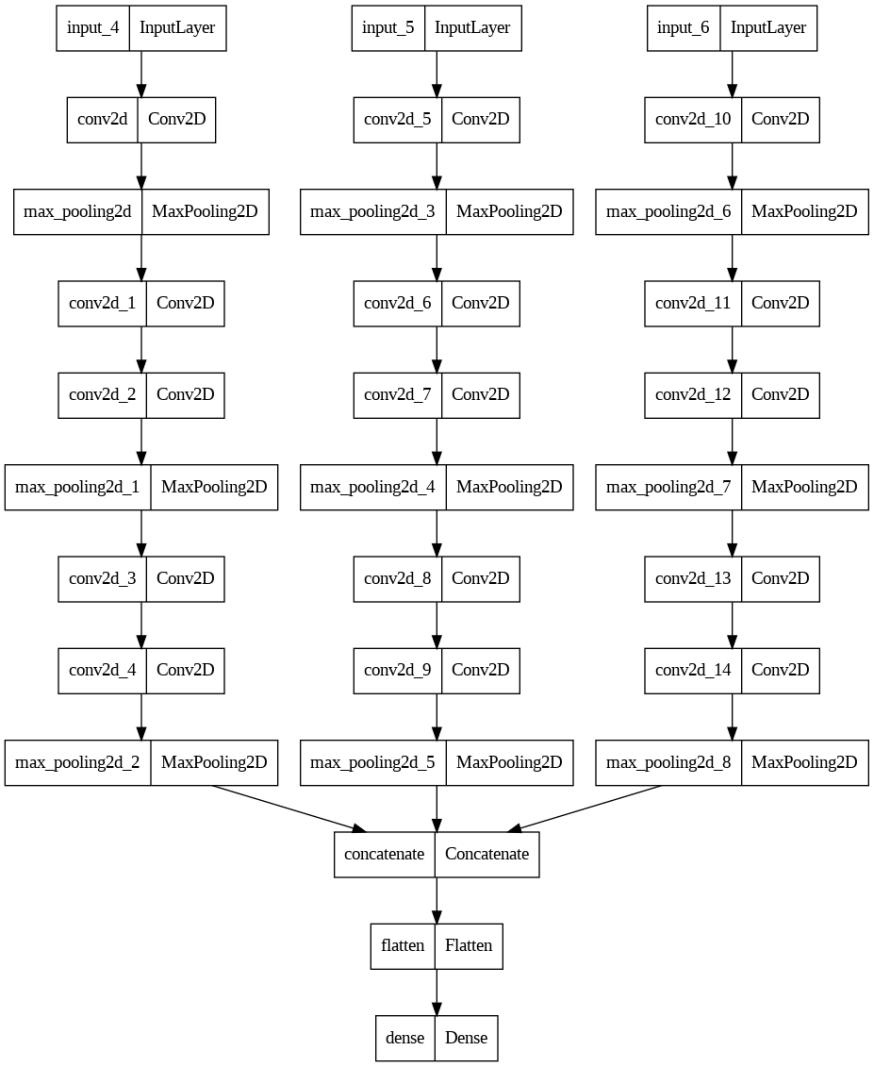

Figura 4.14: Modelo creado para el experimento 2.

concatenación de inputs se acomete al inicio del modelo.

Debido a que en este experimento se hace uso de un modelo que deriva de una MobileNet preentrenada, se requiere de un preprocesamiento de las imágenes de entrada (cada red concreta espera recibir los datos de una manera específica). Así, el uso del método *mobilenet. preprocess_ input* permite escalar los píxeles de las imágenes de entrada entre -1 y 1.

El modelo creado para este experimento hace uso de la base convolucional tomada de la MobileNet y las capas que aparecen representadas en la Figura 4.16. Además, los parámetros usados para cada una de estas capas se muestran en la Figura 4.17.

En cuanto al entrenamiento del modelo, el número de parámetros entrenables y los parámetros escogidos se incluyen en la Tabla 4.6.

Cabe destacar como novedad en relación con la Tabla 4.6 que, de las capas que conforman la base convolucional de la MobileNet, se hacen entrenables a partir de la capa denominada como *conv_ pw_ 9*. El resto de capas anteriores permanecen sin entrenar, es decir, sus pesos

```
input1 = keras.layers.Input(shape=(512,512))
input2 = keras.layers.Input(shape=(512,512))
input3 = keras.layers.Input(shape=(512,512))
input1 = tf.expand_dims(input1, axis=3)
input2 = tf.expand_dims(input2, axis=3)
input3 = tf.expand_dims(input3, axis=3)

# Input1
conv1 = keras.layers.Conv2D(64, 7, activation="relu", padding="same")(input1)
pool1 = keras.layers.MaxPooling2D(2, padding='same')(conv1)
conv2 = keras.layers.Conv2D(128, 3, activation="relu", padding="same")(pool1)
conv3 = keras.layers.Conv2D(128, 3, activation="relu", padding="same")(conv2)
pool2 = keras.layers.MaxPooling2D(2, padding='same')(conv3)
conv4 = keras.layers.Conv2D(256, 3, activation="relu", padding="same")(pool2)
conv5 = keras.layers.Conv2D(256, 3, activation="relu", padding="same")(conv4)
pool3 = keras.layers.MaxPooling2D(2, padding='same')(conv5)

# Input2
conv6 = keras.layers.Conv2D(64, 7, activation="relu", padding="same")(input2)
pool4 = keras.layers.MaxPooling2D(2, padding='same')(conv6)
conv7 = keras.layers.Conv2D(128, 3, activation="relu", padding="same")(pool4)
conv8 = keras.layers.Conv2D(128, 3, activation="relu", padding="same")(conv7)
pool5 = keras.layers.MaxPooling2D(2, padding='same')(conv8)
conv9 = keras.layers.Conv2D(256, 3, activation="relu", padding="same")(pool5)
conv10 = keras.layers.Conv2D(256, 3, activation="relu", padding="same")(conv9)
pool6 = keras.layers.MaxPooling2D(2, padding='same')(conv10)

# Input3
conv11 = keras.layers.Conv2D(64, 7, activation="relu", padding="same")(input3)
pool7 = keras.layers.MaxPooling2D(2, padding='same')(conv11)
conv12 = keras.layers.Conv2D(128, 3, activation="relu", padding="same")(pool7)
conv13 = keras.layers.Conv2D(128, 3, activation="relu", padding="same")(conv12
pool8 = keras.layers.MaxPooling2D(2, padding='same')(conv13)
conv14 = keras.layers.Conv2D(256, 3, activation="relu", padding="same")(pool8)
conv15 = keras.layers.Conv2D(256, 3, activation="relu", padding="same")(conv14
pool9 = keras.layers.MaxPooling2D(2, padding='same')(conv15)

concat = keras.layers.Concatenate(axis=3)([pool3, pool6, pool9])
flat = keras.layers.Flatten()(concat)
output = keras.layers.Dense(3, activation="softmax")(flat)
model2 = keras.Model(inputs=[input1, input2, input3], outputs=[output])
```

Figura 4.15: Parámetros del modelo creado para el experimento 2.

quedan congelados y quedan inalterados durante el proceso de entrenamiento del modelo. Se ha elegido finalmente esta capa concreta tras una serie de pruebas del mismo experimento. Así, se observó que con todas las capas congeladas de la MobileNet la actuación del modelo sobre el conjunto de entrenamiento era bastante mejorable (se alcanzaba una exactitud de en torno a 0.5). Por el contrario, la actuación sobre el conjunto de entrenamiento mejoraba a medida que se le otorgaba al modelo mayor flexibilidad o, lo que es lo mismo, menor número de capas congeladas. En ese camino, se acabó observando que, descongelando a partir de la capa *conv_pw_9*, se conseguía una correcta actuación del modelo sobre el conjunto de entrenamiento con un número no demasiado alto de épocas de entrenamiento. Por tanto, se optó por un punto medio como éste entre las opciones extremas en las que se congelaban todas y ninguna de las capas de la base convolucional de la MobileNet.

Los resultados obtenidos en este experimento tras el entrenamiento del modelo y la evaluación del mismo se presentan en la tercera sección del Capítulo 5.

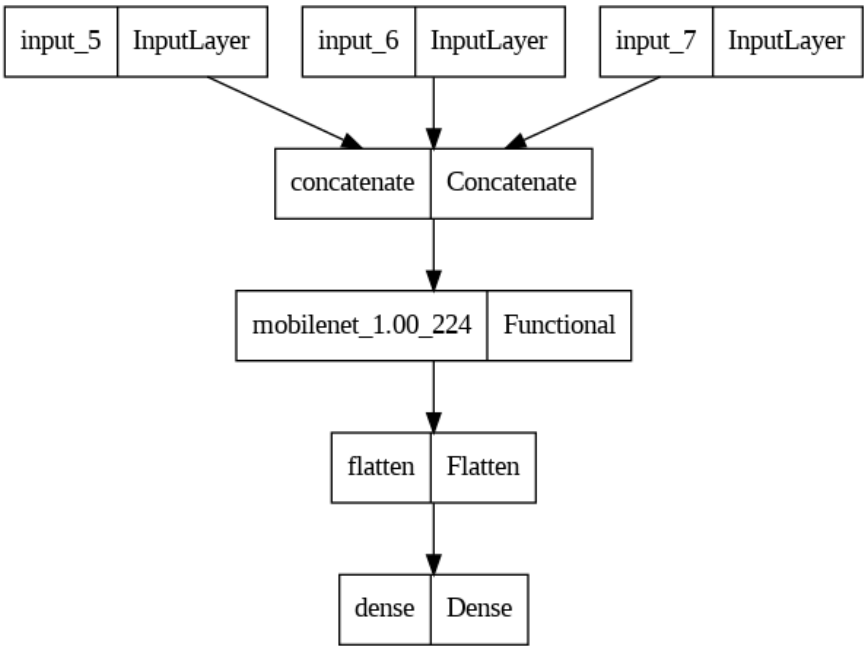

Figura 4.16: Modelo creado para el experimento 3.

```
input1 = keras.layers.Input(shape=(512,512))
input2 = keras.layers.Input(shape=(512,512))
input3 = keras.layers.Input(shape=(512,512))
input1 = tf.expand_dims(input1, axis=3)
input2 = tf.expand_dims(input2, axis=3)
input3 = tf.expand_dims(input3, axis=3)
concat = keras.layers.Concatenate(axis=3)([input1, input2, input3])
conv = conv_base(concat)
flat = keras.layers.Flatten()(conv)
output = keras.layers.Dense(3, activation="softmax")(flat)
model3 = keras.Model(inputs=[input1, input2, input3], outputs=[output])
```

Figura 4.17: Parámetros del modelo creado para el experimento 3.

Tabla 4.6: Parámetros de entrenamiento del modelo del experimento 3.

Parámetros entrenables	4,015,299
Pérdida	Sparse categorical cross entropy
Optimizador	Stochastic Gradient Descent
Ratio de aprendizaje	0.01
Métrica de evaluación	Exactitud
Épocas de entrenamiento	250
1ª capa entrenable de la base convolucional	conv_pw_9

Experimento 4. Comprobación de la actuación del modelo del experimento 1 con un conjunto de datos balanceado

En base a los resultados alcanzados para los tres primeros experimentos, se ha visto necesario plantear este cuarto experimento donde la idea fundamental es ver cómo cambia la actuación del modelo cuando pasa a usarse un conjunto de datos balanceado para el entrenamiento y la validación. Por ello, se han establecido estos conjuntos de datos de tal forma que estén balanceados en términos de clases. Cabe mencionar también que en este experimento se ha hecho uso del modelo del primer experimento, es decir, el modelo personal multi-input con concatenación temprana de las entradas. Se ha decidido proseguir con este modelo frente a los otros debido a que los tres mostraron sufrir el mismo problema de sobreajuste, que el primero permite realizar más ajustes frente al tercero al no partir de un modelo ya creado y que el primero contiene un número muy inferior de parámetros entrenables frente al segundo.

Por tanto, en este experimento aplica lo mismo en cuanto a lo comentado para el preprocesamiento de los datos en el experimento 1, es decir, que no se acomete ninguna técnica especial de preprocesamiento de datos más allá de los pasos esenciales para que los datos puedan ser correctamente entregados al modelo de red neuronal.

Como se utiliza el modelo del experimento 1, se hace uso, por tanto, de las capas que aparecen representadas en la Figura 4.12 y de los parámetros que se muestran en la Figura 4.13.

En cuanto al entrenamiento del modelo, los parámetros escogidos han sido los mismos que aparecen en la Tabla 4.4.

Los resultados obtenidos en este experimento tras el entrenamiento del modelo y la evaluación del mismo se presentan en la cuarta sección del Capítulo 5.

Experimento 5. Comprobación de la actuación del modelo del experimento 1 con conjunto de datos balanceado y aplicación de dropout

Una vez se ha comprobado que el uso de un conjunto de datos balanceado no ha tenido todo el efecto que se podía esperar, se usan los siguientes experimentos como éste para comprobar la influencia que tienen otras técnicas o aspectos. De esta manera, lo primero en lo que cabe pensar cuando se tiene problemas de sobreajuste es en la aplicación de técnicas de regularización, por lo que el objetivo de este experimento es analizar el efecto que ejerce en la actuación del modelo el uso de la técnica de dropout. Dicho esto, en este experimento se va a aplicar exactamente lo mismo que en el experimento anterior, salvo que se va a emplear además la técnica de dropout tal y como se ha comentado.

De nuevo, se siguen las mismas pautas en relación al preprocesamiento de datos que se han considerado en el experimento 4. También conviene decir que el modelo empleado es el mismo que en el experimento 1 con el añadido de las capas de dropout, tal y como se puede

```
input1 = keras.layers.Input(shape=(512,512))
input2 = keras.layers.Input(shape=(512,512))
input3 = keras.layers.Input(shape=(512,512))
input1 = tf.expand_dims(input1, axis=3)
input2 = tf.expand_dims(input2, axis=3)
input3 = tf.expand_dims(input3, axis=3)
concat = keras.layers.Concatenate(axis=3)([input1, input2, input3])
conv1 = keras.layers.Conv2D(64, 7, activation="relu", padding="same")(concat)
pool1 = keras.layers.MaxPooling2D(2, padding='same')(conv1)
drop1 = keras.layers.Dropout(0.2)(pool1)
conv2 = keras.layers.Conv2D(128, 3, activation="relu", padding="same")(drop1)
conv3 = keras.layers.Conv2D(128, 3, activation="relu", padding="same")(conv2)
pool2 = keras.layers.MaxPooling2D(2, padding='same')(conv3)
drop2 = keras.layers.Dropout(0.25)(pool2)
conv4 = keras.layers.Conv2D(256, 3, activation="relu", padding="same")(drop2)
conv5 = keras.layers.Conv2D(256, 3, activation="relu", padding="same")(conv4)
pool3 = keras.layers.MaxPooling2D(2, padding='same')(conv5)
drop3 = keras.layers.Dropout(0.3)(pool3)
flat = keras.layers.Flatten()(drop3)
output = keras.layers.Dense(3, activation="softmax")(flat)
model5 = keras.Model(inputs=[input1, input2, input3], outputs=[output])
```

Figura 4.18: Parámetros del modelo creado para el experimento 5.

Tabla 4.7: Parámetros de entrenamiento del modelo del experimento 5.

Parámetros entrenables	4,261,891
Pérdida	Sparse categorical cross entropy
Optimizador	Stochastic Gradient Descent
Ratio de aprendizaje	0.01
Métrica de evaluación	Exactitud
Épocas de entrenamiento	60

observar en la Figura 4.19. Los parámetros utilizados para las diferentes capas se presentan en la Figura 4.18. En este punto es interesante mencionar que se ha optado por unos valores moderados de la tasa de desactivación (entre 0.2 y 0.35 en vez de 0.5), ya que valores en torno a 0.5 desactivan más neuronas durante cada época de entrenamiento, lo cual dificulta un entrenamiento más rápido de la red, además de conseguir peores resultados globales en relación a la actuación del modelo.

Los parámetros escogidos para el entrenamiento del modelo aparecen en la Tabla 4.7. Se puede comprobar que éstos han cambiado poco con respecto a los empleados en los experimentos 1 y 4. Simplemente, se ha empleado en esta ocasión un mayor número de épocas de entrenamiento ya que, al desactivar neuronas en cada época, se disminuye la capacidad del modelo (hay un menor número de pesos entrenables), lo cual hace que se tarde más en conseguir una buena actuación del mismo.

Los resultados obtenidos en este experimento tras el entrenamiento del modelo y la evaluación del mismo se presentan en la quinta sección del Capítulo 5.

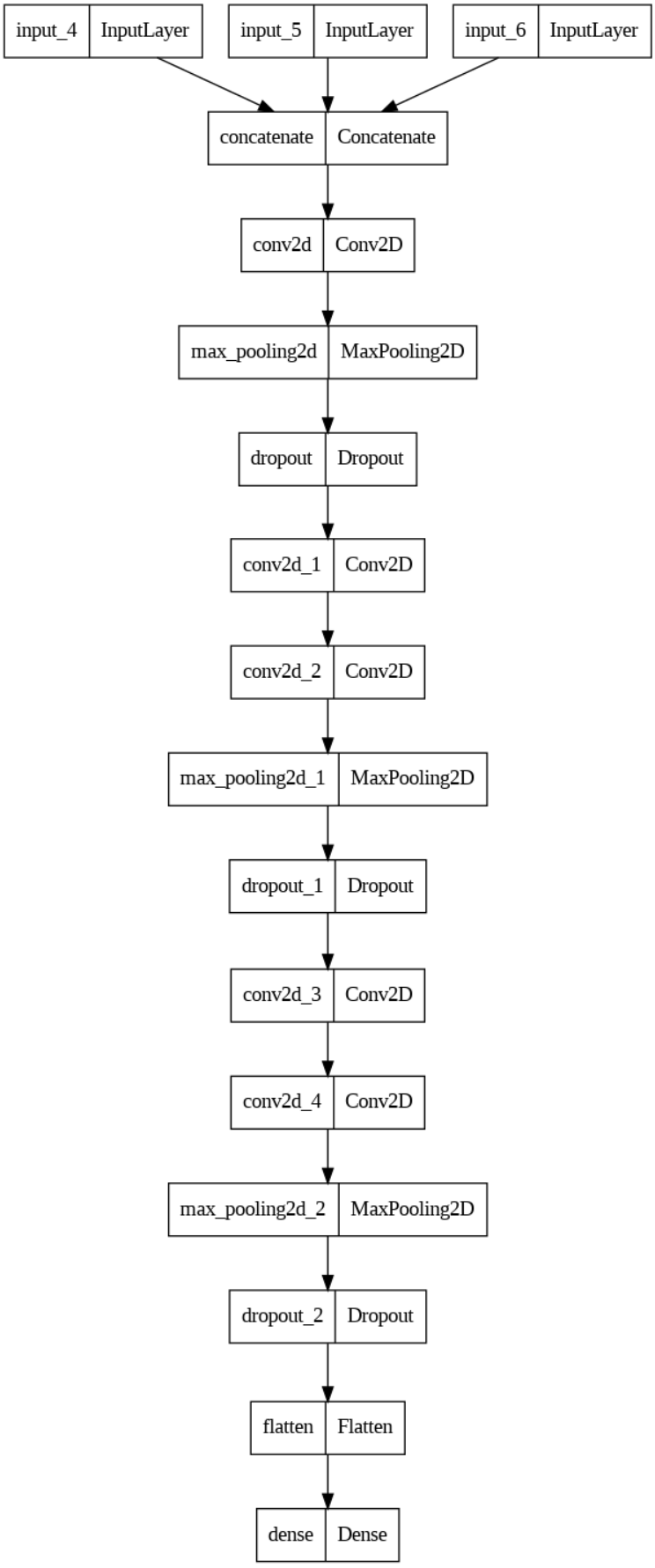

Figura 4.19: Modelo creado para el experimento 5.

Tabla 4.8: Parámetros de entrenamiento del modelo del experimento 6.

Parámetros entrenables	4,262,787
Pérdida	Sparse categorical cross entropy
Optimizador	Stochastic Gradient Descent
Ratio de aprendizaje	0.01
Métrica de evaluación	Exactitud
Épocas de entrenamiento	50

Experimento 6. Comprobación de la actuación del modelo del experimento 1 con conjunto de datos balanceado y aplicación de normalización por lotes

En una línea continuista del experimento 5, se plantea este nuevo experimento, donde se trata de analizar la influencia o el efecto de otra técnica de regularización como es la normalización por lotes.

Se siguen en este experimento los mismos pasos en relación al preprocesamiento de datos que se han considerado en los experimentos 1, 4 y 5. Si bien se usa aquí el mismo modelo base que en los mencionados experimentos 1, 4 y 5, esta vez se añaden capas de normalización por lotes. Este nuevo modelo aparece representado en la Figura 4.20. Los parámetros usados para las distintas capas se muestran en la Figura 4.21. Por otro lado, se vuelve a hacer uso del conjunto de datos balanceado que se utilizó en los dos experimentos anteriores.

La Tabla 4.8 muestra el número de parámetros entrenables y los parámetros escogidos para el entrenamiento del modelo. Es interesante observar que en este modelo se tiene un mayor número de parámetros entrenables en comparación con los que se tenían en los experimentos 1, 4 y 5, ya que aquí se están añadiendo varias capas de normalización por lotes, las cuales introducen un número adicional de parámetros para entrenar.

Los resultados obtenidos en este experimento tras el entrenamiento del modelo y la evaluación del mismo se presentan en la sexta sección del Capítulo 5.

Experimento 7. Comprobación de la actuación del modelo del experimento 1 con conjunto de datos balanceado y algoritmo de optimización diferente

Otro aspecto que se ha considerado interesante para comprobar su influencia en la actuación del modelo es el uso durante el entrenamiento de un algoritmo de optimización diferente al usado hasta ahora. En los experimentos anteriores se ha empleado el algoritmo de *Stochastic Gradient Descent* (SGD) y esta vez se utiliza el conocido como *Root Mean Square Propagation* (RMSprop). Cabe mencionar que, en relación a todo lo demás, se aplica lo mismo que en el experimento 4 (modelo, conjunto de datos y preprocesamiento). Por tanto, el modelo y los parámetros de las capas usados son de nuevo aquéllos que aparecen en la Figura 4.12 y la Figura 4.13 respectivamente.

En lo que se refiere a los parámetros asociados al entrenamiento del modelo, sí que han

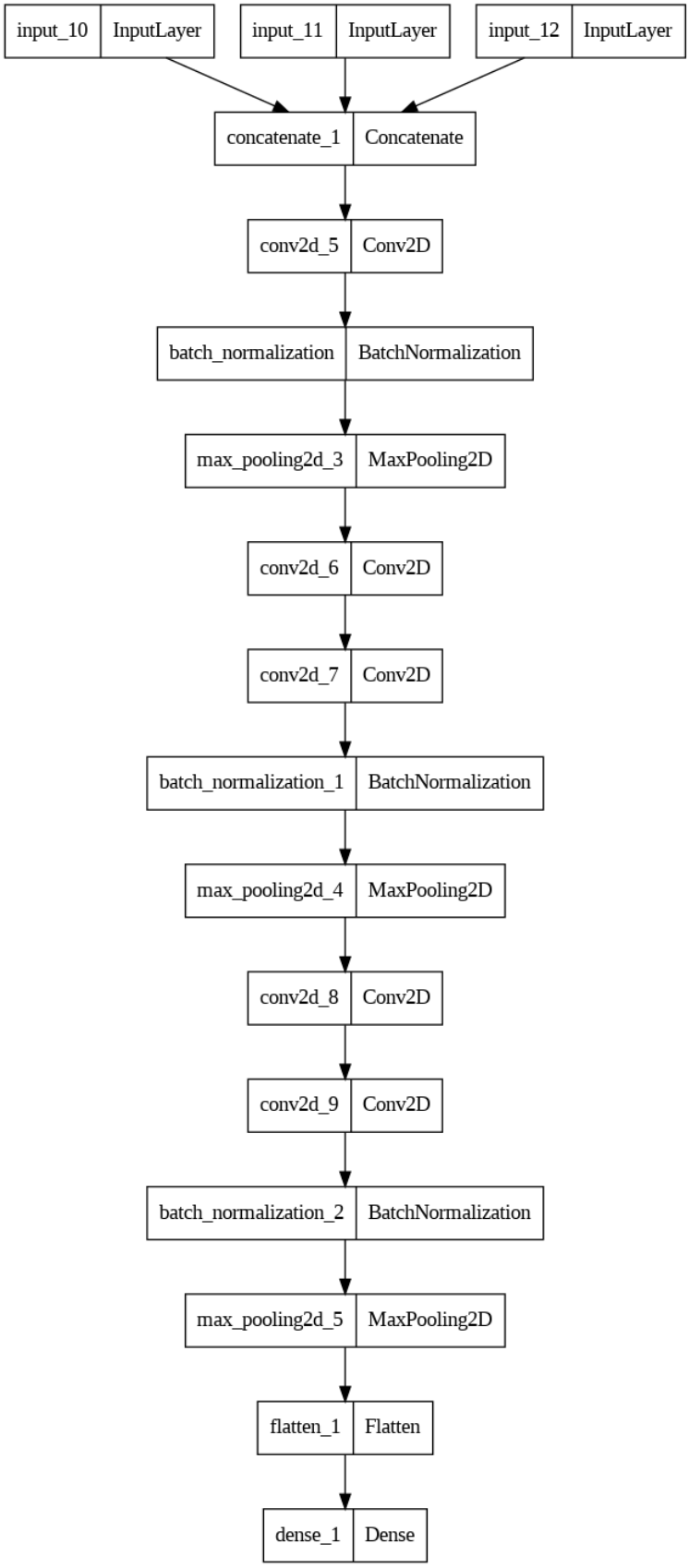

Figura 4.20: Modelo creado para el experimento 6.

```
input1 = keras.layers.Input(shape=(512,512))
input2 = keras.layers.Input(shape=(512,512))
input3 = keras.layers.Input(shape=(512,512))
input1 = tf.expand_dims(input1, axis=3)
input2 = tf.expand_dims(input2, axis=3)
input3 = tf.expand_dims(input3, axis=3)
concat = keras.layers.Concatenate(axis=3)([input1, input2, input3])
conv1 = keras.layers.Conv2D(64, 7, activation="relu", padding="same")(concat)
batch1 = keras.layers.BatchNormalization()(conv1)
pool1 = keras.layers.MaxPooling2D(2, padding='same')(batch1)
conv2 = keras.layers.Conv2D(128, 3, activation="relu", padding="same")(pool1)
conv3 = keras.layers.Conv2D(128, 3, activation="relu", padding="same")(conv2)
batch2 = keras.layers.BatchNormalization()(conv3)
pool2 = keras.layers.MaxPooling2D(2, padding='same')(batch2)
conv4 = keras.layers.Conv2D(256, 3, activation="relu", padding="same")(pool2)
conv5 = keras.layers.Conv2D(256, 3, activation="relu", padding="same")(conv4)
batch3 = keras.layers.BatchNormalization()(conv5)
pool3 = keras.layers.MaxPooling2D(2, padding='same')(batch3)
flat = keras.layers.Flatten()(pool3)
output = keras.layers.Dense(3, activation="softmax")(flat)
model6 = keras.Model(inputs=[input1, input2, input3], outputs=[output])
```

Figura 4.21: Parámetros del modelo creado para el experimento 6.

Tabla 4.9: Parámetros de entrenamiento del modelo del experimento 7.

Parámetros entrenables	4,261,891
Pérdida	Sparse categorical cross entropy
Optimizador	RMSprop
Ratio de aprendizaje	0.00001
Métrica de evaluación	Exactitud
Épocas de entrenamiento	150

cambiado en este caso al utilizar un algoritmo de optimización diferente. Éstos son mostrados en la Tabla 4.9. Se puede comprobar que se ha usado un ratio de aprendizaje y un número de épocas diferentes. Éstos han resultado tras una serie de pruebas realizadas del mismo experimento en las que se buscaba la mejor actuación posible del modelo en el conjunto de datos.

Con todo esto, los resultados alcanzados en este experimento una vez completados el entrenamiento y la evaluación del modelo se exponen en la séptima sección del Capítulo 5.

Experimento 8. Comprobación de la actuación del modelo del experimento 1 con conjunto de datos balanceado y aplicación de la técnica del aumento de datos

Después de comprobar que el problema de sobreajuste continúa tras la realización de los experimentos anteriores, se ha decidido abordar como siguiente paso técnicas analizadas en el estado del arte y que han demostrado tener efectos positivos en relación al sobreajuste,

mejorando la actuación de modelos de aprendizaje profundo en tareas como la asumida en este estudio. Así, la técnica aplicada en este experimento es la del aumento de datos, la cual fue ya introducida en el capítulo correspondiente al marco teórico.

El único aspecto novedoso de este experimento es la introducción del aumento de datos. Por tanto, en relación al resto de características, aplica lo mismo que en el experimento 4. De esta forma, el modelo y las capas empleadas son de nuevo los que se muestran en la Figura 4.12 y la Figura 4.13. Los parámetros asociados al entrenamiento del modelo no han cambiado generalmente, salvo los referentes al número de épocas. Por consiguiente, aplican los mismos que aparecen en la Tabla 4.4, excepto que se han empleado 30 épocas de entrenamiento con 50 pasos por época tanto para el entrenamiento como para la validación. Este límite en el número de pasos por época ha sido necesario, ya que de lo contrario el aumento de datos no cesa en la provisión de imágenes al modelo (se puede aumentar los datos hasta un número indefinido de imágenes).

En lo que se refiere al aumento de datos, se parte del mismo conjunto de datos balanceado que ha sido empleado en el experimento 4. Estos datos balanceados se aportan a los generadores de imágenes característicos del aumento de datos, los cuales efectúan las transformaciones solicitadas a dichas imágenes y, posteriormente, las proveen al modelo durante su entrenamiento. Como ya se comentó anteriormente en este documento, las transformaciones aplicadas a las imágenes han de ser de tal manera que las imágenes resultantes sean realistas. Por ello, la transformación escogida ha sido una rotación de las imágenes con un rango de hasta 3 grados.

Teniendo todo esto en cuenta, los resultados obtenidos en este experimento tras el entrenamiento y la evaluación del modelo se presentan en la octava sección del Capítulo 5.

Experimento 9. Comprobación de la actuación del modelo del experimento 1 con conjunto de datos balanceado y aplicación de filtros de imágenes

Este noveno experimento sigue una línea similar al anterior en el sentido de que hace uso de otra técnica bien contrastada en el estado del arte buscando reducir el sobreajuste del modelo. Esta técnica es la aplicación de filtros de imágenes. Por medio de esta técnica, se busca comprobar si es posible que ésta ayude al modelo a reconocer patrones en las imágenes y, en última instancia, a generalizar mejor. Tal y como ya se introdujo previamente en este documento, se prueban tres filtros diferentes. El primero de ellos es el de ecualización del histograma que se denomina CLAHE y los otros dos están relacionados con la detección de bordes y son los filtros de Sobel y Canny.

La aplicación de filtros de imágenes en escala de grises es el aspecto novedoso de este experimento, por lo que una vez más se aplica lo mismo que en el experimento 4 en relación al modelo y su entrenamiento, el conjunto de datos balanceado y el preprocesamiento de los mismos. De esta forma, el modelo usado es el mostrado en la Figura 4.12 y los parámetros usados para sus capas son los que aparecen en la Figura 4.13. En cuanto al entrenamiento del

Tabla 4.10: Parámetros de entrenamiento del modelo del experimento 10.

Parámetros entrenables	786,435
Pérdida	Sparse categorical cross entropy
Optimizador	Stochastic Gradient Descent
Ratio de aprendizaje	0.01
Métrica de evaluación	Exactitud
Épocas de entrenamiento	50
Capas entrenables de la base convolucional	Ninguna

modelo, aplican los mismos parámetros que en la Tabla 4.4. El punto distintivo de la utilización de filtros de imágenes es que, cuando se construyen los conjuntos de datos balanceados, las imágenes correspondientes a dichos conjuntos y que son alimentadas posteriormente al modelo son las que resultan de aplicarle a las originales el filtro en cuestión.

Con estas consideraciones, los resultados conseguidos en este experimento tras el entrenamiento y la evaluación del modelo se incluyen en la novena sección del Capítulo 5.

Experimento 10. Comprobación de la actuación del modelo del experimento 3 con conjunto de datos balanceado y congelación de todas las capas de la base convolucional

Todo lo probado hasta ahora no ha demostrado ofrecer una mejora importante en relación al problema por sobreajuste que se viene sufriendo desde el inicio. Además, también es cierto que el número de parámetros entrenables existentes en cada experimento puede ser demasiado alto. Por ello, llegados a este punto, se ha visto conveniente reducir todo lo posible dicho número de parámetros entrenables. En la búsqueda de este objetivo, se hace uso en este experimento del modelo del experimento 3, es decir, del modelo preentrenado, en el cual se congelan todas las capas de la base convolucional tomada.

Dicho esto, el modelo empleado en este experimento es el que aparece en la Figura 4.16 y los parámetros usados para las capas que lo conforman son los mostrados en la Figura 4.17. En relación al preprocesamiento de los datos, aplican los mismos comentarios expuestos en el experimento 3. En lo que se refiere al entrenamiento del modelo, éste es caracterizado por los parámetros que aparecen en la Tabla 4.10.

Considerando todo lo comentado, los resultados alcanzados en este experimento tras el entrenamiento y la evaluación del modelo se muestran en la décima sección del Capítulo 5.

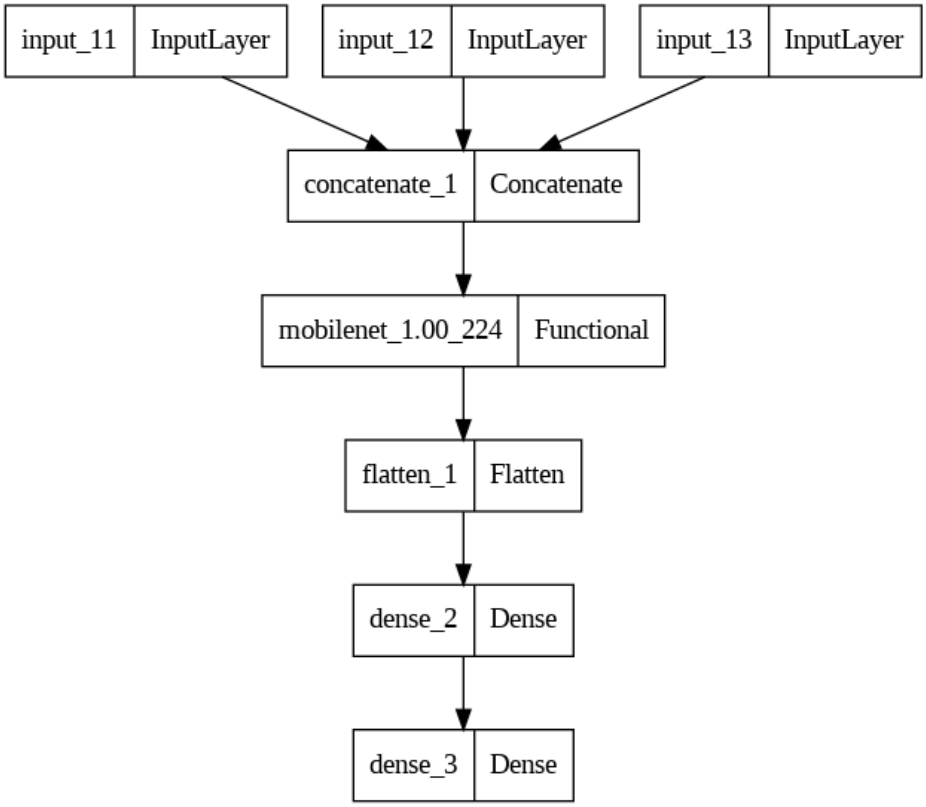

Figura 4.22: Modelo creado para el experimento 11.

Experimento 11. Comprobación de la actuación del modelo del experimento 3 con conjunto de datos balanceado, congelación de todas las capas de la base convolucional e inclusión adicional de una capa densa de neuronas

Observando los resultados del experimento 10, se ha comprobado que se alcanzaban valores excesivamente altos para la función de pérdida, además de una capacidad nula de predicción por parte del modelo. Tratando de mejorar dichos resultados, se ha modificado ligeramente la parte del clasificador del modelo. Así, en el modelo usado en este experimento se ha añadido una capa densa de neuronas adicional.

Teniendo en cuenta lo comentado, el modelo empleado aparece en la Figura 4.22 y los parámetros usados para las capas de dicho modelo se exponen en la Figura 4.23. En relación al preprocesamiento y creación del conjunto de datos, aplican los mismos aspectos empleados en el experimento 10.

En cuanto al entrenamiento del modelo, el número de parámetros entrenables y los parámetros escogidos para este experimento se incluyen en la Tabla 4.11.

Los resultados obtenidos en este experimento tras el entrenamiento y la evaluación del modelo se muestran en la undécima sección del Capítulo 5.

```
input1 = keras.layers.Input(shape=(512,512))
input2 = keras.layers.Input(shape=(512,512))
input3 = keras.layers.Input(shape=(512,512))
input1 = tf.expand_dims(input1, axis=3)
input2 = tf.expand_dims(input2, axis=3)
input3 = tf.expand_dims(input3, axis=3)
concat = keras.layers.Concatenate(axis=3)([input1, input2, input3])
conv = conv_base(concat)
flat = keras.layers.Flatten()(conv)
dense1 = keras.layers.Dense(10, activation="relu")(flat)
output = keras.layers.Dense(3, activation="softmax")(dense1)
model = keras.Model(inputs=[input1, input2, input3], outputs=[output])
```

Figura 4.23: Parámetros del modelo creado para el experimento 11.

Tabla 4.11: Parámetros de entrenamiento del modelo del experimento 11.

Parámetros entrenables	2,621,483
Pérdida	Sparse categorical cross entropy
Optimizador	Stochastic Gradient Descent
Ratio de aprendizaje	0.01
Métrica de evaluación	Exactitud
Épocas de entrenamiento	50
Capas entrenables de la base convolucional	Ninguna

Experimento 12. Comprobación de la actuación del modelo del experimento 1 con reducción de parámetros entrenables y con conjunto de datos balanceado

En este nuevo experimento y, tras los resultados obtenidos para los experimentos 10 y 11, se decide volver a usar el modelo propio con concatenación de datos al inicio, pero tratando de reducir de manera notable el número de parámetros a entrenar. El objetivo buscado con esto es el de intentar reducir el sobreajuste que experimenta el modelo a la vez que tiene capacidad o flexibilidad suficiente para aprender las características claves de las imágenes.

Dicho esto, cabe decir que la estructura empleada para el modelo de este experimento es igual que la mostrada en la Figura 4.12, si bien se han reducido los valores para el número de filtros de cada capa convolucional, de tal manera que se tuviera un número significativamente menor de parámetros para entrenar. Así, los parámetros utilizados para las distintas capas del modelo aparecen en la Figura 4.24.

Los parámetros usados en relación al entrenamiento del modelo coinciden con los presentados en la Tabla 4.4, salvo que el número de parámetros entrenables en este experimento pasa a ser de 858,115. Por tanto, se puede comprobar que este número es significativamente menor que el número de parámetros entrenables usado para este modelo hasta ahora.

Teniendo todo esto en cuenta, los resultados alcanzados para este experimento tras el entrenamiento y la evaluación del modelo se exponen en la duodécima sección del Capítulo

```
input1 = keras.layers.Input(shape=(512,512))
input2 = keras.layers.Input(shape=(512,512))
input3 = keras.layers.Input(shape=(512,512))
input1 = tf.expand_dims(input1, axis=3)
input2 = tf.expand_dims(input2, axis=3)
input3 = tf.expand_dims(input3, axis=3)
concat = keras.layers.Concatenate(axis=3)([input1, input2, input3])
conv1 = keras.layers.Conv2D(16, 7, activation="relu", padding="same")(concat)
pool1 = keras.layers.MaxPooling2D(2, padding='same')(conv1)
conv2 = keras.layers.Conv2D(32, 3, activation="relu", padding="same")(pool1)
conv3 = keras.layers.Conv2D(32, 3, activation="relu", padding="same")(conv2)
pool2 = keras.layers.MaxPooling2D(2, padding='same')(conv3)
conv4 = keras.layers.Conv2D(64, 3, activation="relu", padding="same")(pool2)
conv5 = keras.layers.Conv2D(64, 3, activation="relu", padding="same")(conv4)
pool3 = keras.layers.MaxPooling2D(2, padding='same')(conv5)
flat = keras.layers.Flatten()(pool3)
output = keras.layers.Dense(3, activation="softmax")(flat)
model = keras.Model(inputs=[input1, input2, input3], outputs=[output])
```

Figura 4.24: Parámetros del modelo creado para el experimento 12.

5.

Experimento 13. Comprobación de la actuación del modelo del experimento 1 con diferente característica a predecir

Este experimento surge tras los análisis efectuados de los resultados alcanzados en los experimentos anteriores. Se decide utilizar el mismo modelo que en el experimento 1, pero teniendo ahora como objetivo la predicción de una característica diferente. Ésta se ha escogido de tal forma que esté relacionada con algún aspecto bien observable de la resonancia de mama. Dicha característica elegida ha sido la presencia o no del pezón en la resonancia. Ésta puede adoptar dos valores: 0 si no hay presencia del pezón y 1 si sí la hay. Por consiguiente, el punto diferencial en este experimento es el conjunto de datos. Adicionalmente, es importante mencionar que se trata de una característica fuertemente desbalanceada (existen muchas más pacientes cuyas resonancias no incluyen el pezón que aquéllas para las que sí), por lo que el conjunto de datos usado en este experimento no está balanceado. Si se usara un conjunto balanceado, éste tendría un bajo número de imágenes para entrenar (salvo que se usase una técnica como la del aumento de datos), con el coste negativo que ello supone.

Teniendo en cuenta esto y, como se ha mencionado, el modelo empleado ha sido el del experimento 1 (propio con concatenación inicial de los datos). Por ello, la estructura del mismo coincide con la enseñada en la Figura 4.12. Además, los parámetros usados para las capas que conforman dicho modelo son también iguales a los mostrados en la Figura 4.13, salvo el número de neuronas de la última capa densa que realiza la tarea de clasificación, el cual en este caso ha de ser 2, puesto que se trata de un problema de clasificación binaria.

En lo que se refiere al entrenamiento del modelo, los parámetros empleados son los mismos que en el experimento 1, es decir, que los mostrados en la Tabla 4.4, con la excepción del

Tabla 4.12: Parámetros de entrenamiento del modelo del experimento 14.

Parámetros entrenables	524,290
Pérdida	Sparse categorical cross entropy
Optimizador	Stochastic Gradient Descent
Ratio de aprendizaje	0.01
Métrica de evaluación	Exactitud
Épocas de entrenamiento	20
Capas entrenables de la base convolucional	Ninguna

número de parámetros entrenables y el número de épocas de entrenamiento. Así, el número de parámetros a entrenar pasa a ser en este experimento de 3,213,314, mientras que el número de épocas de entrenamiento ha sido de 15, dado que se alcanzaba un régimen estacionario bastante antes que en otros casos anteriores para las evoluciones de la función de pérdida y la métrica de evaluación de exactitud.

Considerando todo esto, los resultados obtenidos para este experimento tras el entrenamiento y la evaluación del modelo se presentan en la decimotercera sección del Capítulo 5.

Experimento 14. Comprobación de la actuación del modelo del experimento 3 con diferente característica a predecir

En este último experimento llevado a cabo se sigue la misma línea asumida en el experimento anterior, pero se hace uso del modelo del experimento 3, es decir, del modelo preentrenado. Por tanto, en este experimento también se va a tratar de predecir la presencia o no del pezón en las resonancias magnéticas de mama.

Teniendo en cuenta estas consideraciones, el modelo empleado tiene la misma estructura que la mostrada en la Figura 4.16 y utiliza los mismos parámetros para las capas del modelo que aparecen en la Figura 4.17, con la excepción de nuevo del número de neuronas de la última capa densa que acomete la predicción, el cual en este caso pasa a ser igual a 2.

En lo que respecta al entrenamiento del modelo, los parámetros usados son aquéllos que se exponen en la Tabla 4.12. Como puede verse en dicha tabla, se ha decidido congelar los pesos de las capas de la base convolucional de la MobileNet preentrenada, por lo que los parámetros que quedan para entrenar se corresponden con los de la capa densa de neuronas que abordan el problema de clasificación binaria. Además, se ha comprobado que 20 épocas de entrenamiento era un número suficiente para alcanzar un régimen estacionario en relación a las evoluciones de la función de pérdida y la métrica de evaluación con las épocas de entrenamiento.

Teniendo en cuenta todo lo comentado con anterioridad, los resultados alcanzados en este experimento tras el entrenamiento y la evaluación del modelo se presentan en la decimocuarta

sección del Capítulo 5.

Capítulo 5

Resultados

Este capítulo del documento tiene por objetivo el presentar los resultados alcanzados para cada uno de los experimentos realizados que han sido descritos en el capítulo anterior. Estos resultados abarcarán desde valores de métricas de evaluación para calificar la actuación de los modelos hasta las imágenes obtenidas tras la aplicación de, por ejemplo, un filtro de imagen.

Experimento 1. Comprobación de actuación del modelo personal con concatenación temprana de datos

Las evoluciones de la función de pérdida y de la métrica de evaluación escogidas (*sparse categorical cross entropy* como pérdida y exactitud como métrica de evaluación) para la monitorización del entrenamiento del modelo aparecen representadas en la Figura 5.1 tanto para el conjunto de entrenamiento como para el de validación.

A partir de la Figura 5.1, es posible verificar que el modelo consigue actuar muy bien para el conjunto de datos de entrenamiento, puesto que la exactitud acaba tendiendo a 1 y la pérdida tendiendo a 0. No obstante, los resultados alcanzados para el conjunto de datos de validación son bastante pobres, ya que la exactitud se estanca en un valor que está en torno a 0.45 y la pérdida termina divergiendo a partir del momento en que el modelo alcanza la exactitud con valor de 1 en el conjunto de entrenamiento.

Una vez se ha entrenado el modelo, se ha evaluado éste sobre el conjunto de datos de validación. Los valores alcanzados para la función de pérdida y las métricas de evaluación se muestran en la Tabla 5.1. Además, se presenta la correspondiente matriz de confusión en la Figura 5.2.

Como puede comprobarse, los resultados obtenidos en relación a la actuación del modelo de este experimento 1 son bastante deficientes, ya que no demuestra una capacidad predictiva muy sobresaliente en el conjunto de validación. Como es sabido, el objetivo pretendido es que un modelo actúe bien sobre aquellos datos que no ha visto durante el entrenamiento y esto no lo consigue satisfacer este modelo. De la Figura 5.2, se puede comprobar que el modelo

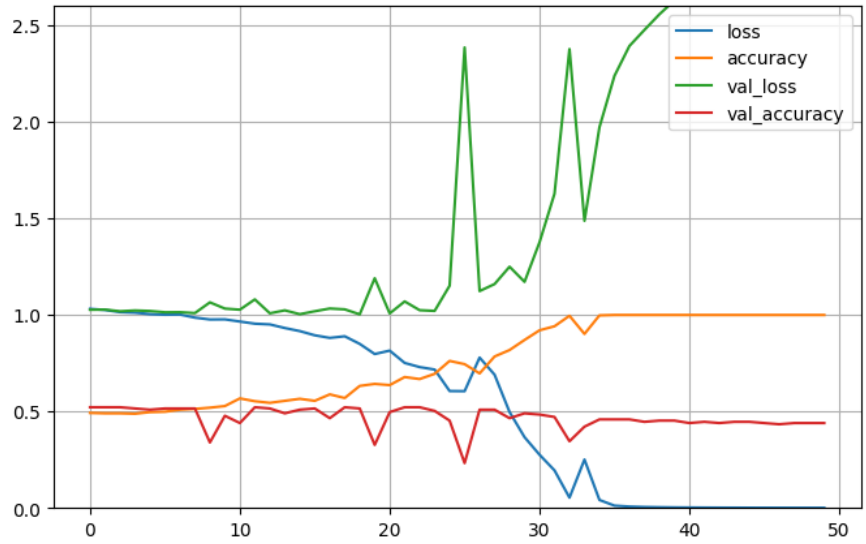

Figura 5.1: Evolución de la función de pérdida y de la métrica de evaluación durante el entrenamiento del modelo del experimento 1.

Tabla 5.1: Resultados alcanzados por el modelo del experimento 1 en el conjunto de datos de validación.

Pérdida	**Exactitud**	**Precisión**	**Exhaustividad**	**Puntuación F1**	**AUC**	**Matriz de confusión**
3.055	0.440	0.402	0.440	0.417	0.557	$\begin{bmatrix} 2 & 20 & 10 \\ 10 & 52 & 21 \\ 4 & 24 & 16 \end{bmatrix}$

no ha sido capaz de predecir bien casi ninguna de las imágenes cuya clase se corresponde con 0 y sólo una pequeña parte de aquéllas que pertenecen a la clase 2.

Todos estos resultados mostrados anteriormente ofrecen como conclusión que el modelo sufre de sobreajuste u overfitting, ya que el modelo consigue actuar correctamente sobre el conjunto de entrenamiento, pero no sobre el conjunto de validación, llegando un momento en que los valores de la métrica de evaluación y la pérdida divergen notablemente entre el conjunto de entrenamiento y el de validación.

Experimento 2. Comprobación de actuación del modelo personal con concatenación tardía de datos

Tras el entrenamiento del modelo propuesto para el segundo experimento, se han obtenido las evoluciones de la función de pérdida y de la métrica de evaluación tanto para el conjunto de entrenamiento como para el de validación, las cuales son representadas en la Figura 5.3.

A partir de la Figura 5.3, es posible corroborar que el modelo consigue actuar bien para el conjunto de datos de entrenamiento, ya que la exactitud acaba tendiendo a 1 y la pérdida tendiendo a 0. Sin embargo, los resultados logrados para el conjunto de datos de validación

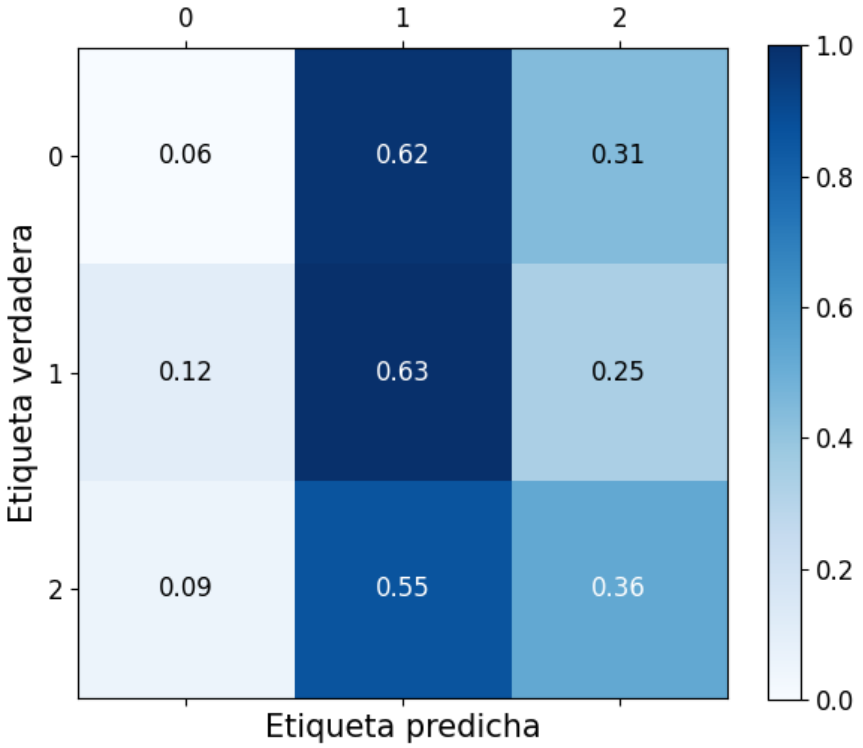

Figura 5.2: Matriz de confusión alcanzada por el modelo del experimento 1 sobre el conjunto de validación.

Tabla 5.2: Resultados alcanzados por el modelo del experimento 2 en el conjunto de datos de validación.

Pérdida	**Exactitud**	**Precisión**	**Exhaustividad**	**Puntuación F1**	**AUC**	**Matriz de confusión**
2.824	0.509	0.471	0.509	0.484	0.592	$\begin{bmatrix} 3 & 21 & 8 \\ 8 & 57 & 18 \\ 4 & 19 & 21 \end{bmatrix}$

son bastante deficientes, puesto que la exactitud se estanca en un valor que está en torno a 0.5 y la pérdida termina divergiendo a partir del momento en que el modelo alcanza la exactitud con valor de 1 en el conjunto de entrenamiento.

Una vez se ha entrenado el modelo, se ha evaluado éste sobre el conjunto de datos de validación. Los valores obtenidos para la función de pérdida y las métricas de evaluación se presentan en la Tabla 5.2. Además, se muestra la correspondiente matriz de confusión en la Figura 5.4.

Como puede verse, los resultados obtenidos en relación a la actuación del modelo de este experimento 2 son de nuevo bastante pobres, puesto que no demuestra una capacidad predictiva buena en el conjunto de validación, si bien es cierto que mejora los resultados alcanzados por el modelo del experimento 1. De la Figura 5.4, se puede comprobar que el modelo no ha sido capaz de predecir bien casi ninguna de las imágenes cuya clase se

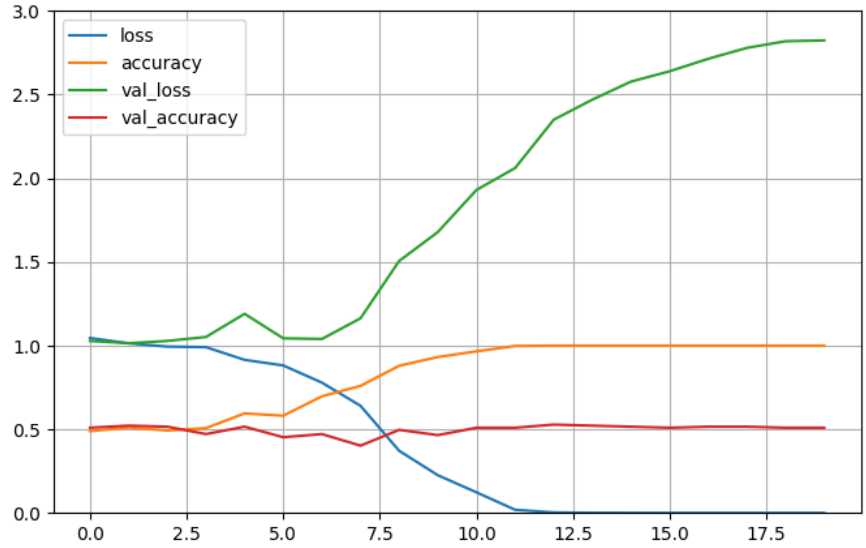

Figura 5.3: Evolución de la función de pérdida y de la métrica de evaluación durante el entrenamiento del modelo del experimento 2.

corresponde con 0. Sin embargo, ha podido acertar la clase correcta en casi el 50 % de las imágenes de clase 2 y en torno al 70 % de las imágenes de clase 1. Estos valores mejoran, por tanto, los conseguidos por el modelo presentado en el experimento 1 para las mismas clases.

Los resultados mostrados anteriormente siguen ofreciendo, no obstante, como conclusión principal que el modelo continúa sufriendo de sobreajuste, ya que el modelo consigue actuar correctamente sobre el conjunto de entrenamiento, pero no sobre el conjunto de validación, llegando un momento en que los valores de la métrica de evaluación y la pérdida divergen notablemente entre el conjunto de entrenamiento y el de validación.

Experimento 3. Comprobación de actuación del modelo preentrenado

Tras el entrenamiento del modelo propuesto para el tercer experimento, se han obtenido las evoluciones de la función de pérdida y de la métrica de evaluación tanto para el conjunto de entrenamiento como para el de validación, las cuales son representadas en la Figura 5.5 y la Figura 5.6 (ésta última constituye un zoom sobre la imagen anterior).

A partir de la Figura 5.5 y 5.6, es posible comprobar una vez más que el modelo consigue actuar bien para el conjunto de datos de entrenamiento, ya que la exactitud acaba tendiendo a 1 y la pérdida tendiendo a 0. No obstante, los resultados logrados para el conjunto de datos de validación son bastante deficientes, con la exactitud estancándose en un valor que está en torno a 0.4 y con la pérdida adoptando valores altos de en torno a 6.7. Es interesante notar también que para este modelo se ha requerido un mayor número de épocas de entrenamiento (250) frente a las usadas en los dos experimentos anteriores. Esto tiene que ver con el hecho de tener buena parte de los pesos de la base convolucional de la MobileNet congelados sin poder modificarse, ya que limita la flexibilidad y capacidad del modelo para aprender patrones y

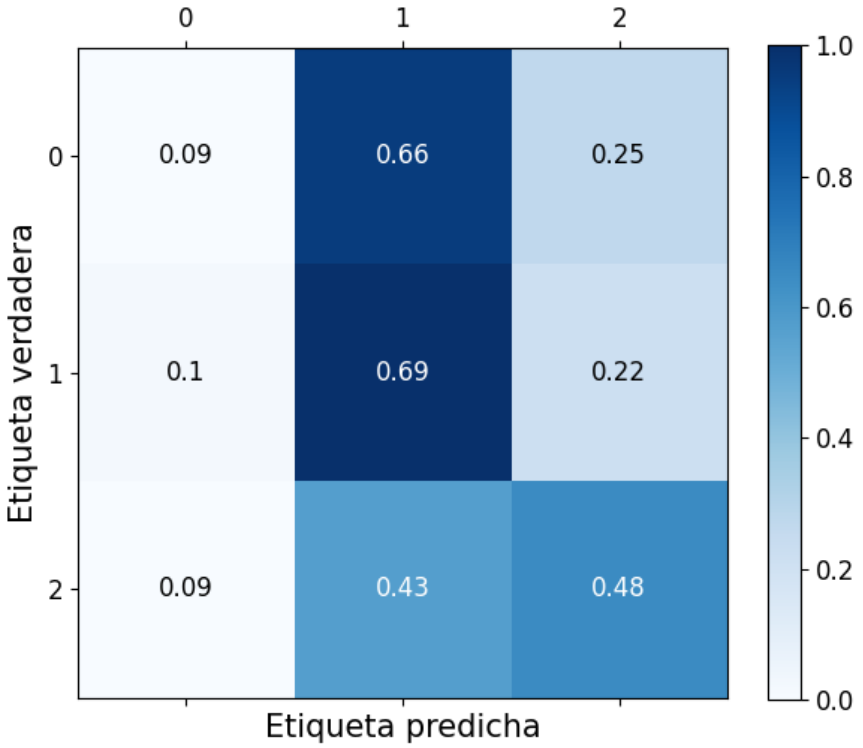

Figura 5.4: Matriz de confusión alcanzada por el modelo del experimento 2 sobre el conjunto de validación.

Tabla 5.3: Resultados alcanzados por el modelo del experimento 3 en el conjunto de datos de validación.

Pérdida	**Exactitud**	**Precisión**	**Exhaustividad**	**Puntuación F1**	**AUC**	**Matriz de confusión**
6.362	0.402	0.393	0.402	0.397	0.546	$\begin{bmatrix} 3 & 17 & 12 \\ 14 & 47 & 22 \\ 9 & 21 & 14 \end{bmatrix}$

características de las imágenes.

Una vez se ha entrenado el modelo, se ha evaluado éste sobre el conjunto de datos de validación. Los valores alcanzados para la función de pérdida y las métricas de evaluación se muestran en la Tabla 5.3. Además, se presenta la correspondiente matriz de confusión en la Figura 5.7.

Como es posible verificar, los resultados obtenidos en relación a la actuación del modelo de este experimento 3 son de nuevo bastante pobres, puesto que no demuestra una capacidad predictiva buena en el conjunto de validación. De la Figura 5.7, se puede ver que el modelo tiene más tendencia a clasificar las imágenes por la clase 1. Esto es algo común en los tres experimentos mostrados hasta ahora, lo cual lleva a pensar que este problema por sobreajuste tiene cierta independencia del modelo (se han probado tres modelos diferentes y todos ellos sufren igualmente dicho problema). Por tanto, se ha considerado que podría ser

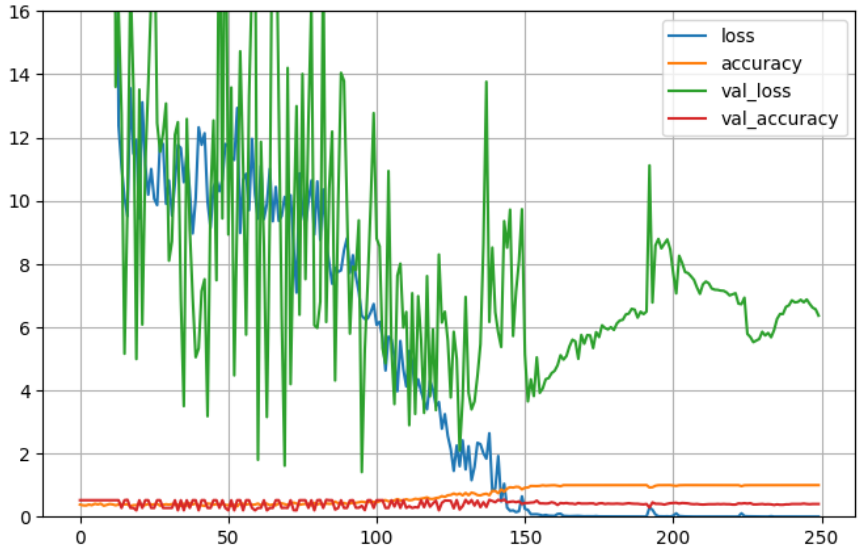

Figura 5.5: Evolución de la función de pérdida y de la métrica de evaluación durante el entrenamiento del modelo del experimento 3.

algo más intrínseco al conjunto de datos usado y su fragmentación en los subconjuntos de entrenamiento y validación. Esto lleva a recuperar la Tabla 4.3 y la Figura 4.3, las cuales permiten comprobar que las clases no están perfectamente balanceadas. Además, el conjunto de datos de entrenamiento empleado en los tres primeros experimentos se caracteriza por contener el primer 75 % de las pacientes que tienen registrado un valor para el grado de Nottingham. Por tanto, este tipo de separación hace que dicho conjunto de entrenamiento contenga claramente más imágenes correspondientes a pacientes cuyo grado de Nottingham se clasifica por 1 (asumiendo como clases 0, 1 y 2). Ello explicaría que el modelo no esté aprendiendo realmente a identificar las características clave de las imágenes, sino más bien a aprender cada imagen específica. Esto conduce a que el modelo acabe tendiendo generalmente en su predicción hacia la clase con mayor presencia en las imágenes. Precisamente, el conjunto de validación contenía exactamente un 52.20 % de imágenes de esta clase, el cual coincide con el valor de la exactitud alcanzada en el conjunto de validación que aparecía para muchas de las épocas de entrenamiento de los modelos de estos tres primeros experimentos.

Todo esto lleva a plantear un claro objetivo de cara al siguiente experimento y es el de establecer un conjunto de datos de entrenamiento que esté balanceado en términos de clases para que el modelo no tenga predilección por una clase en concreto simplemente porque haya mayor presencia de ella en el conjunto de datos.

Experimento 4. Comprobación de la actuación del modelo del experimento 1 con un conjunto de datos balanceado

Las evoluciones de la función de pérdida y de la métrica de evaluación alcanzadas en este cuarto experimento tanto para el conjunto de entrenamiento como para el de validación

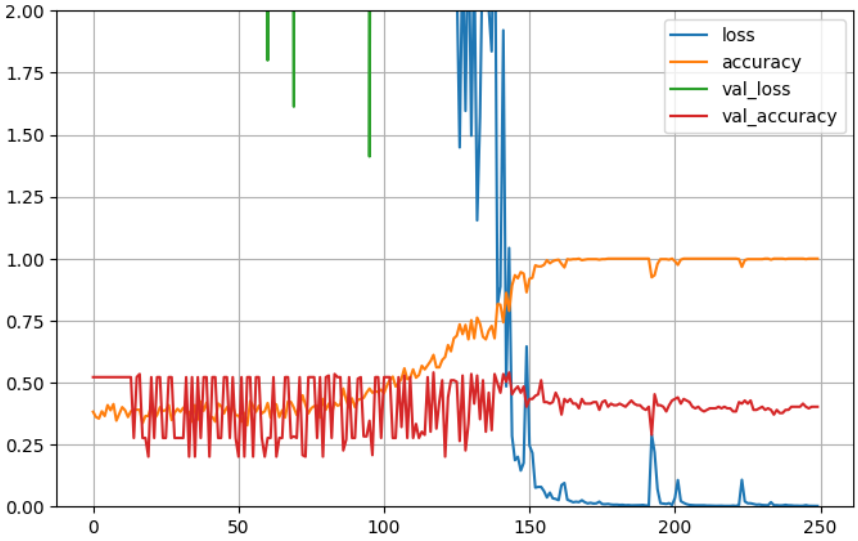

Figura 5.6: Zoom realizado a la Figura 5.5.

aparecen representadas en la Figura 5.8.

Por medio de la Figura 5.8, se puede comprobar que no se mejora la actuación del modelo con respecto a lo obtenido en el experimento 1. De nuevo, el modelo consigue actuar muy bien para el conjunto de datos de entrenamiento, pero no logra actuar igual de bien para el conjunto de datos de validación. Se repite el hecho de que para el conjunto de datos de validación la exactitud se estanca en valores que están por debajo de 0.5 y la pérdida termina divergiendo a partir del momento en que el modelo alcanza la exactitud con valor de 1 en el conjunto de entrenamiento.

Una vez se ha entrenado el modelo, se ha evaluado éste sobre el conjunto de datos de validación. Los valores alcanzados para la función de pérdida y las métricas de evaluación se muestran en la Tabla 5.4. Además, se presenta la correspondiente matriz de confusión en la Figura 5.9.

Como puede comprobarse, los resultados obtenidos en relación a la actuación del modelo en este experimento 4 vuelven a ser bastante mejorables, ya que no demuestra una buena capacidad predictiva en el conjunto de validación. Por tanto, el uso de un conjunto de datos balanceado no parece haber ejercido todo el efecto que se podía esperar. No obstante, de la Figura 5.9, se puede corroborar que el modelo ya no tiene esa tendencia o predilección hacia

Tabla 5.4: Resultados alcanzados por el modelo del experimento 4 en el conjunto de datos de validación.

Pérdida	**Exactitud**	**Precisión**	**Exhaustividad**	**Puntuación F1**	**AUC**	**Matriz de confusión**
2.813	0.407	0.416	0.407	0.405	0.563	$\begin{bmatrix} 6 & 3 & 9 \\ 5 & 7 & 6 \\ 2 & 7 & 9 \end{bmatrix}$

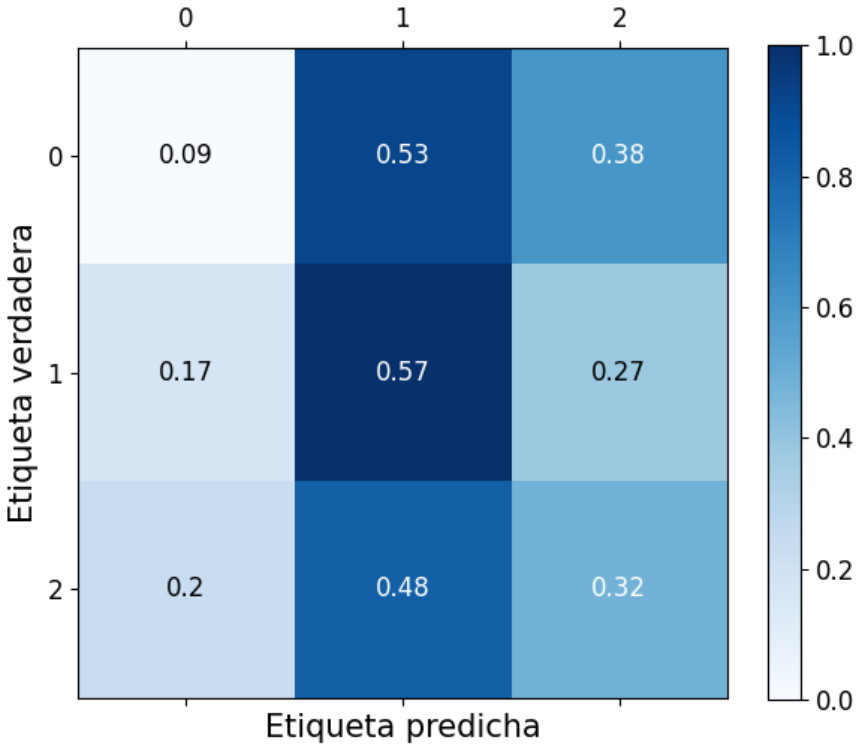

Figura 5.7: Matriz de confusión alcanzada por el modelo del experimento 3 sobre el conjunto de validación.

la clase 1 dado que se tiene un conjunto balanceado de datos. Adicionalmente, se observa una mejora en la capacidad predictiva del modelo hacia la clase 0, ya que consigue acertar la tercera parte de las imágenes enmarcadas dentro de esa clase. En cuanto a la clase 2, también se consigue un mejor resultado, consiguiendo el modelo acertar la mitad de las imágenes que están incluidas en dicha clase.

En definitiva, los resultados mostrados anteriormente ofrecen como conclusión que el modelo sigue sufriendo de sobreajuste a pesar de utilizar esta vez un conjunto de datos balanceado en términos de clases. Sin embargo, se han observado mejoras en la capacidad predictiva del modelo hacia las clases que antes no lograba predecir nada bien (aquéllas que no coincidían con la clase mayoritaria). Esto también es reflejado así si se comparan los resultados de ambos experimentos mostrados respectivamente en la Tabla 5.1 y la Tabla 5.4, los cuales permiten concluir que el modelo del experimento 1 alcanzaba un mayor porcentaje de acierto, pero esto venía dado porque tenía esa tendencia por la clase mayoritaria (la cual ofrece lógicamente mayor probabilidad de acierto) a costa de las clases restantes. Esto se enmienda en cierta manera en este último experimento, consiguiendo aciertos que están más repartidos entre las tres clases, aunque con un peor porcentaje de acierto en términos globales.

Teniendo en cuenta lo comentado arriba, se va a seguir empleando el primer modelo con un conjunto de datos balanceado (es decir, lo usado en este cuarto experimento) de cara

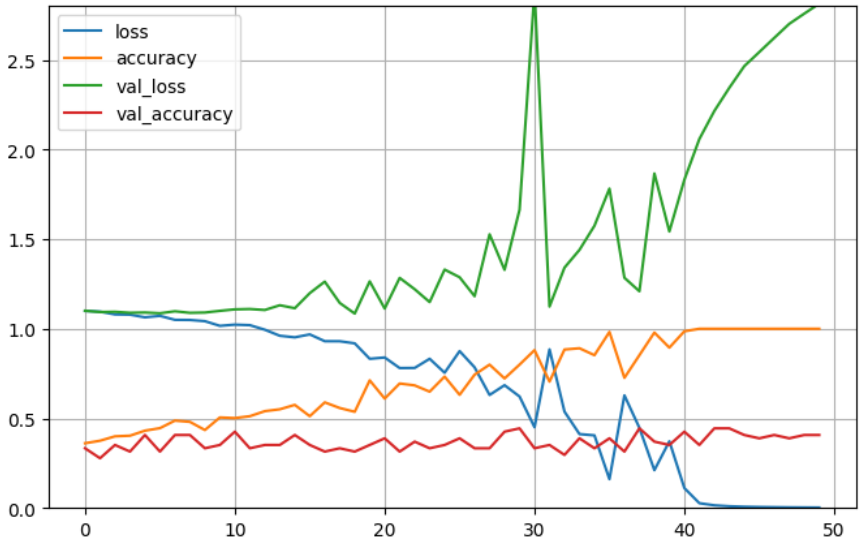

Figura 5.8: Evolución de la función de pérdida y de la métrica de evaluación durante el entrenamiento del modelo del experimento 4.

Tabla 5.5: Resultados alcanzados por el modelo del experimento 5 en el conjunto de datos de validación.

Pérdida	**Exactitud**	**Precisión**	**Exhaustividad**	**Puntuación F1**	**AUC**	**Matriz de confusión**
3.009	0.352	0.361	0.352	0.349	0.523	$\begin{bmatrix} 5 & 4 & 9 \\ 5 & 6 & 7 \\ 2 & 8 & 8 \end{bmatrix}$

a la exploración de otros aspectos para comprobar si es posible solucionar el problema por sobreajuste.

Experimento 5. Comprobación de la actuación del modelo del experimento 1 con conjunto de datos balanceado y aplicación de dropout

Las evoluciones de la función de pérdida y de la métrica de evaluación tanto para el conjunto de entrenamiento como para el de validación aparecen representadas en la Figura 5.10.

A través de la Figura 5.10, se puede verificar que la aplicación del dropout no ha contribuido a mejorar el problema que se viene sufriendo desde el inicio. Otra vez, se consigue una buena actuación en el conjunto de entrenamiento (exactitud tendiendo a 1 y pérdida tendiendo a 0), pero un pobre desempeño en el conjunto de validación. También vuelve a ser una característica común el notable incremento de la pérdida y el estancamiento de la exactitud en un valor bajo para el conjunto de validación una vez se alcanza una exactitud con valor de 1 y una pérdida muy cercana a 0 en el conjunto de entrenamiento.

Tras el entrenamiento del modelo, se ha evaluado éste sobre el conjunto de datos de

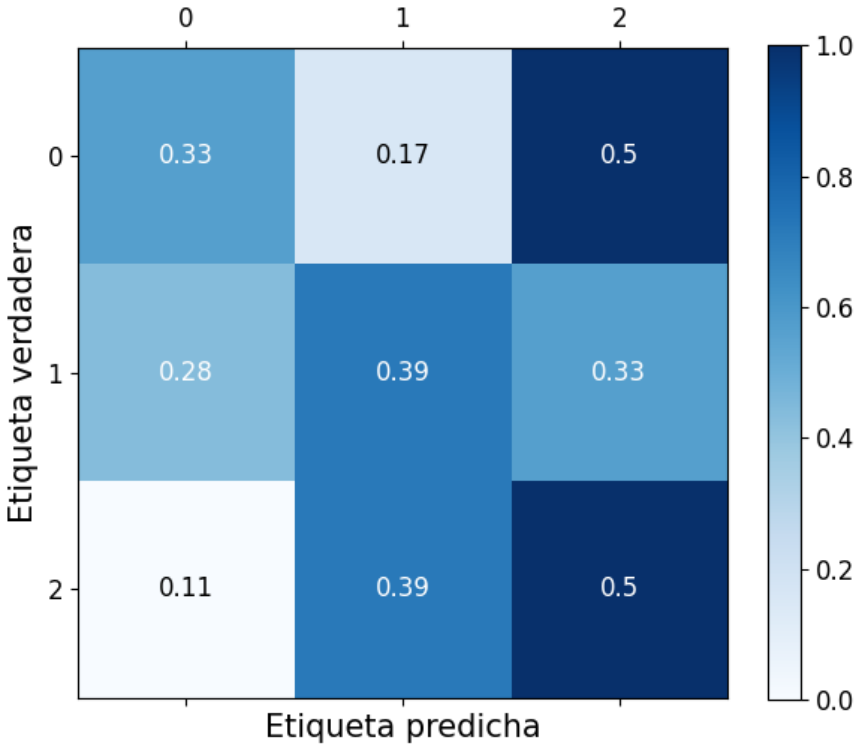

Figura 5.9: Matriz de confusión alcanzada por el modelo del experimento 4 sobre el conjunto de validación.

validación. Los valores que se han obtenido para la función de pérdida y las métricas de evaluación se presentan en la Tabla 5.5. Adicionalmente, se expone la correspondiente matriz de confusión en la Figura 5.11.

Como puede verse, los resultados obtenidos en relación a la actuación del modelo en este experimento 5 no arrojan mejoras con respecto a lo ya visto. Se extrae, por ello, como conclusión que el modelo sigue sufriendo de sobreajuste, a pesar de hacer uso de una técnica de regularización como el dropout.

Experimento 6. Comprobación de la actuación del modelo del experimento 1 con conjunto de datos balanceado y aplicación de normalización por lotes

Las evoluciones de la función de pérdida y de la métrica de evaluación tanto para el conjunto de entrenamiento como para el de validación aparecen representadas en la Figura 5.12.

Por medio de la Figura 5.12, se puede comprobar que la adición de capas de normalización por lotes no ha contribuido a mejorar el problema de sobreajuste que se viene teniendo. Una vez más, se consigue una buena actuación en el conjunto de entrenamiento (exactitud tendiendo a 1 y pérdida tendiendo a 0), pero una mala actuación en el conjunto de validación. Se puede ver, por ejemplo, que incluso se alcanzan generalmente valores demasiado altos para la función de pérdida en el caso del conjunto de validación (estos valores escapan

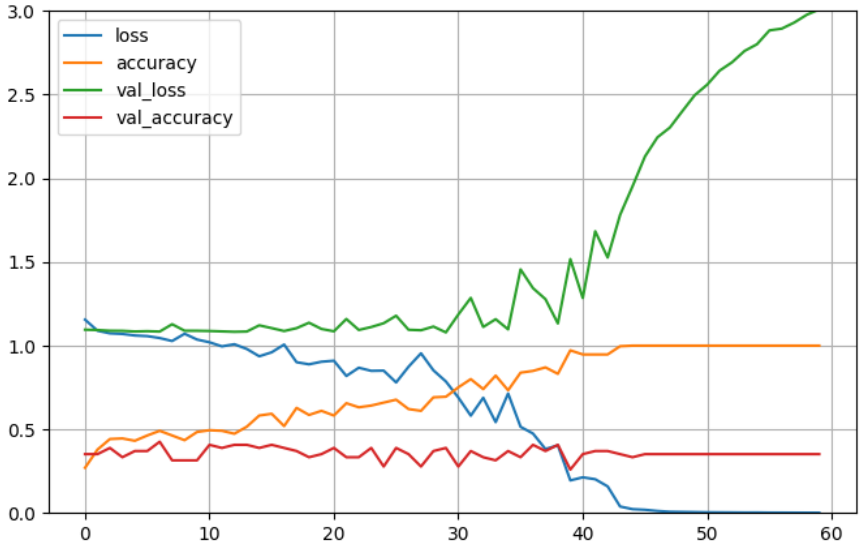

Figura 5.10: Evolución de la función de pérdida y de la métrica de evaluación durante el entrenamiento del modelo del experimento 5.

Tabla 5.6: Resultados alcanzados por el modelo del experimento 6 en el conjunto de datos de validación.

Pérdida	**Exactitud**	**Precisión**	**Exhaustividad**	**Puntuación F1**	**AUC**	**Matriz de confusión**
25.195	0.370	0.393	0.370	0.376	0.526	$\begin{bmatrix} 7 & 2 & 9 \\ 3 & 6 & 9 \\ 4 & 7 & 7 \end{bmatrix}$

sobradamente del límite empleado en el eje vertical).

Tras entrenarse el modelo, se ha evaluado éste sobre el conjunto de datos de validación. Los valores alcanzados para la función de pérdida y las métricas de evaluación se muestran en la Tabla 5.6. Además, se presenta la correspondiente matriz de confusión en la Figura 5.13.

Como puede verificarse, los resultados obtenidos en relación a la actuación del modelo en este experimento 6 siguen sin mostrar mejoras con respecto a lo ya visto. Se obtiene como conclusión, por tanto, que el modelo sigue sufriendo por sobreajuste, a pesar de utilizar capas de normalización por lotes.

Experimento 7. Comprobación de la actuación del modelo del experimento 1 con conjunto de datos balanceado y algoritmo de optimización diferente

Tras la realización de este experimento, se muestra la evolución de la pérdida y la métrica de evaluación con las épocas de entrenamiento para ambos conjuntos de datos (entrenamiento y validación) en la Tabla 5.14.

A partir de la Figura 5.14, se puede ver que el cambio de algoritmo de entrenamiento

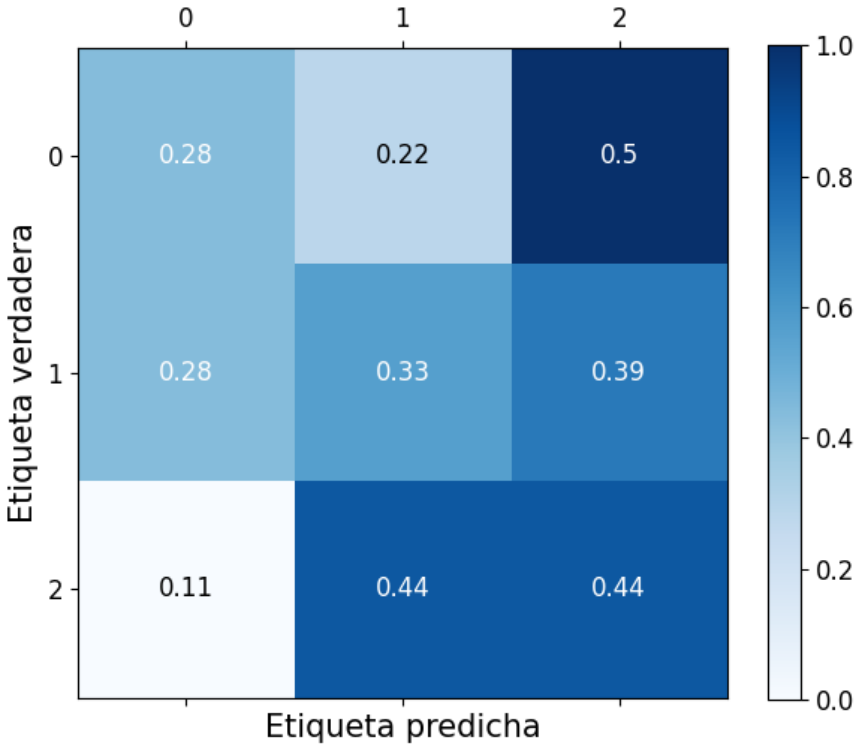

Figura 5.11: Matriz de confusión alcanzada por el modelo del experimento 5 sobre el conjunto de validación.

no provoca ninguna mejora en relación al problema de sobreajuste. Una vez más, se alcanza una buena actuación en el conjunto de entrenamiento, pero un mal desempeño en el conjunto de validación. Además, algo notable que se ha observado con el cambio de algoritmo es el incremento de oscilaciones en las curvas de evolución de la Figura 5.14 durante el entrenamiento.

Una vez llevado a cabo el entrenamiento del modelo, se ha evaluado éste sobre el conjunto de datos de validación. Los valores que se han alcanzado para la función de pérdida y las métricas de evaluación se presentan en la Tabla 5.7. Por otro lado, se expone la correspondiente matriz de confusión en la Figura 5.15. De esta última figura, es posible verificar, por ejemplo, que la capacidad predictiva del modelo en relación a la clase 0 es bastante deficiente.

Como puede comprobarse, los resultados obtenidos en relación a la actuación del modelo

Tabla 5.7: Resultados alcanzados por el modelo del experimento 7 en el conjunto de datos de validación.

Pérdida	**Exactitud**	**Precisión**	**Exhaustividad**	**Puntuación F1**	**AUC**	**Matriz de confusión**
6.762	0.315	0.343	0.315	0.304	0.533	$\begin{bmatrix} 3 & 4 & 11 \\ 3 & 6 & 9 \\ 1 & 9 & 8 \end{bmatrix}$

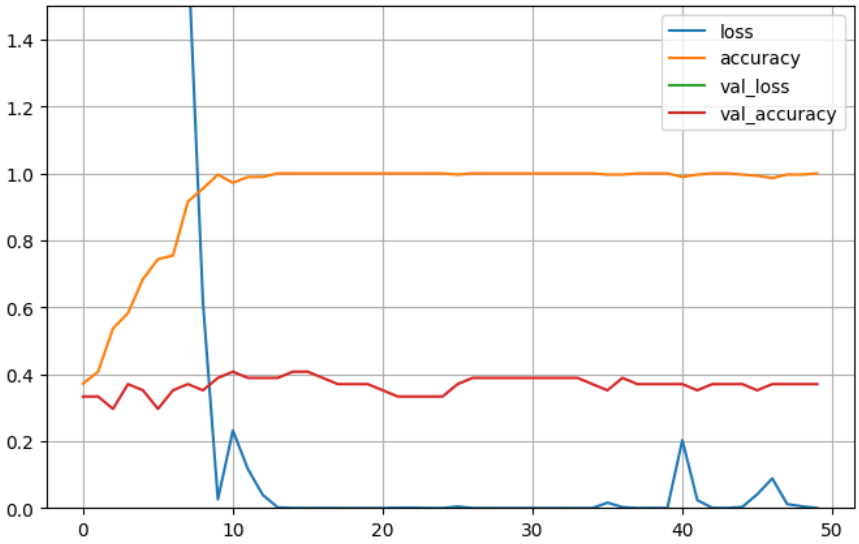

Figura 5.12: Evolución de la función de pérdida y de la métrica de evaluación durante el entrenamiento del modelo del experimento 6.

en este experimento 7 no logran presentar avances con respecto a lo ya mostrado. Por tanto, la conclusión principal es que la actuación del modelo no experimenta ninguna mejora, pese a la utilización de un algoritmo de optimización diferente.

Experimento 8. Comprobación de la actuación del modelo del experimento 1 con conjunto de datos balanceado y aplicación de la técnica del aumento de datos

Tal y como se comentó en el capítulo anterior, este experimento aplica la técnica del aumento de datos tratando de encontrar alguna mejora al problema de sobreajuste. Como también se mencionó, se ha hecho uso de rotaciones de hasta 3 grados para transformar las imágenes originales y obtener así imágenes realistas. Un ejemplo de una imagen original y su correspondiente pareja generada con el aumento de datos se muestra en la Figura 5.16 y en la Figura 5.17 respectivamente.

La Figura 5.18 permite ver la evolución de la pérdida y la métrica de evaluación para ambos conjuntos de datos durante el entrenamiento del modelo. Tras este entrenamiento, se ha evaluado el modelo en el conjunto de validación. Los valores alcanzados para la función de pérdida y las métricas de evaluación se presentan en la Tabla 5.8 y la correspondiente matriz de confusión se muestra en la Figura 5.19.

Las tendencias y conclusiones obtenidas de este experimento son similares a las alcanzadas en experimentos anteriores, en el sentido de que el modelo sigue mostrando una baja capacidad predictiva en el conjunto de validación. Por tanto, la técnica del aumento de datos tampoco demuestra ofrecer una mejora notable en relación a esta cuestión.

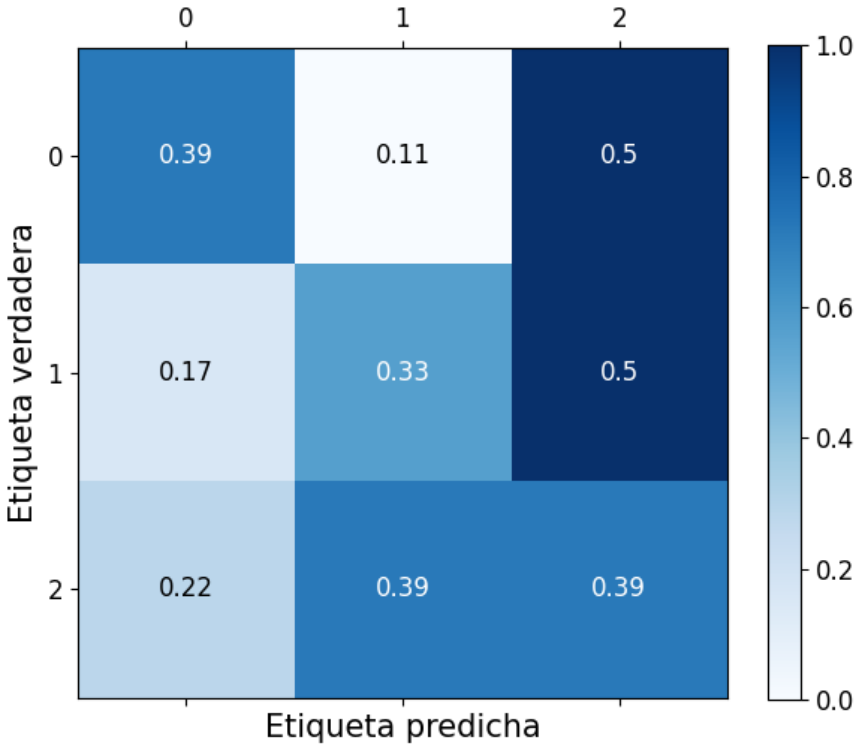

Figura 5.13: Matriz de confusión alcanzada por el modelo del experimento 6 sobre el conjunto de validación.

Tabla 5.8: Resultados alcanzados por el modelo del experimento 8 en el conjunto de datos de validación.

Pérdida	**Exactitud**	**Precisión**	**Exhaustividad**	**Puntuación F1**	**AUC**	**Matriz de confusión**
4.933	0.426	0.444	0.426	0.430	0.556	$\begin{bmatrix} 8 & 0 & 10 \\ 5 & 7 & 6 \\ 3 & 7 & 8 \end{bmatrix}$

Experimento 9. Comprobación de la actuación del modelo del experimento 1 con conjunto de datos balanceado y aplicación de filtros de imágenes

Como ya se explicó, se ha analizado la aplicación de tres filtros propios de imágenes en escala de grises. Esta aplicación se ha llevado a cabo durante la construcción de los conjuntos de datos. Los filtros empleados han sido los denominados CLAHE, Sobel y Canny. De esta forma, la Figura 5.20 expone un ejemplo de resonancia magnética original sin la aplicación de filtros, la Figura 5.21 muestra el empleo del filtro de ecualización del histograma CLAHE, la Figura 5.22 presenta el resultado tras utilizar el filtro de Sobel y la Figura 5.23 permite ver la imagen obtenida tras hacer uso del filtro de Canny.

Con esto, las evoluciones de la función de pérdida y de la métrica de evaluación tanto para el conjunto de entrenamiento como para el de validación aparecen representadas en

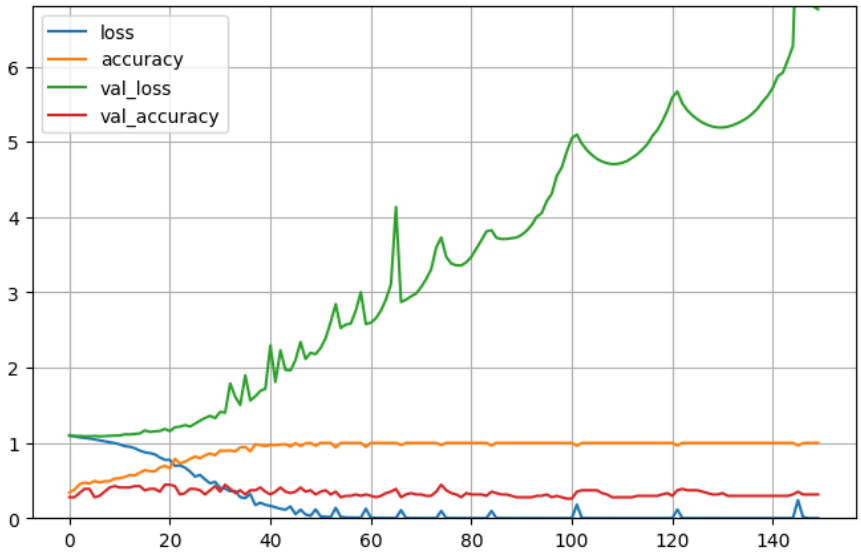

Figura 5.14: Evolución de la función de pérdida y de la métrica de evaluación durante el entrenamiento del modelo del experimento 7.

Tabla 5.9: Resultados alcanzados por el modelo del experimento 9 en el conjunto de datos de validación.

Pérdida	**Exactitud**	**Precisión**	**Exhaustividad**	**Puntuación F1**	**AUC**	**Matriz de confusión**
2.457	0.426	0.428	0.426	0.424	0.543	$\begin{bmatrix} 7 & 3 & 8 \\ 6 & 7 & 5 \\ 3 & 6 & 9 \end{bmatrix}$

la Figura 5.24 en el caso en que el filtro de Sobel es aplicado. No se muestran las figuras análogas para el resto de filtros dado que no ofrecen conclusiones diferentes a las extraídas por medio de la Figura 5.24 ni mejores resultados.

Como de costumbre, se ha evaluado el modelo en el conjunto de validación. Los valores alcanzados para la función de pérdida y las métricas de evaluación se presentan en la Tabla 5.9 y la correspondiente matriz de confusión se expone en la Figura 5.25.

Con los resultados mostrados de este experimento, se ha podido verificar que la aplicación de filtros a las imágenes del conjunto de datos no ha conseguido tampoco proporcionar una mejora sustancial a la capacidad predictiva del modelo, no pudiendo solucionar el problema de sobreajuste.

Experimento 10. Comprobación de la actuación del modelo del experimento 3 con conjunto de datos balanceado y congelación de todas las capas de la base convolucional

La Figura 5.26 presenta la evolución de la función de pérdida y de la métrica de evaluación a lo largo de las épocas de entrenamiento. Se puede observar que las curvas correspondientes

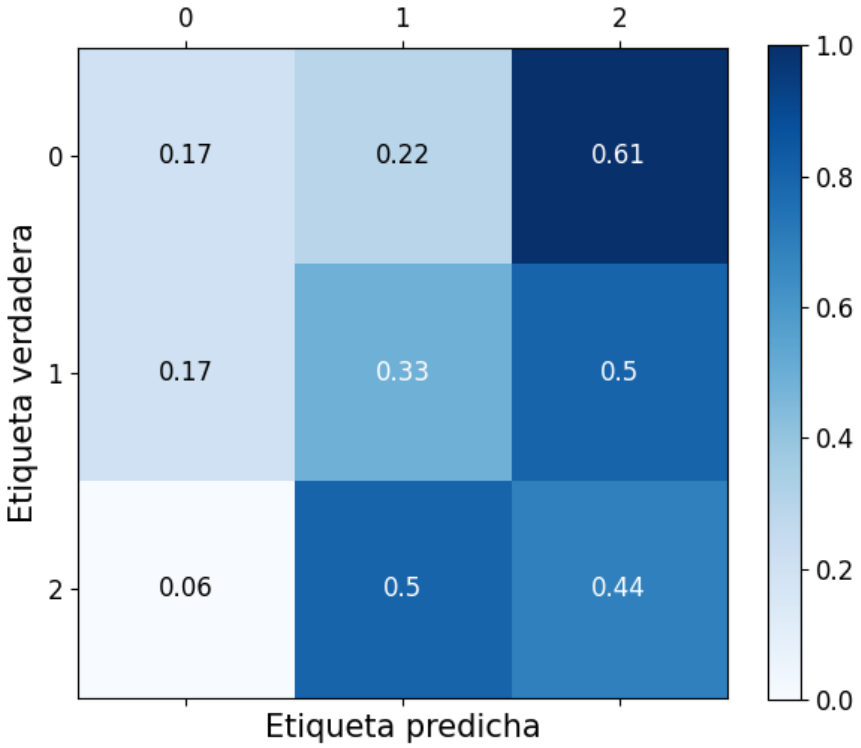

Figura 5.15: Matriz de confusión alcanzada por el modelo del experimento 7 sobre el conjunto de validación.

a la función de pérdida no aparecen. Esto es debido a que se alcanzan valores muy altos para dicha función en ambos conjuntos de datos. Además, puede verse que la métrica de evaluación adquiere valores muy constantes a lo largo del entrenamiento, aunque los del conjunto de entrenamiento experimentan bastantes oscilaciones.

Tras la evaluación del modelo en el conjunto de validación, los valores alcanzados para la función de pérdida y las métricas de evaluación se exponen en la Tabla 5.10 y la correspondiente matriz de confusión se muestra en la Figura 5.27.

Los resultados alcanzados en este experimento son realmente malos, pues se puede corroborar que el modelo no lleva a cabo ningún aprendizaje, ya que se decanta por una clase en concreto y lanza todas las predicciones en base a dicha clase, como se puede ver en la matriz de confusión de la Figura 5.27. Por otro lado, se ha obtenido un AUC de 0.5, el cual significa

Tabla 5.10: Resultados alcanzados por el modelo del experimento 10 en el conjunto de datos de validación.

Pérdida	**Exactitud**	**Precisión**	**Exhaustividad**	**Puntuación F1**	**AUC**	**Matriz de confusión**
371.340	0.333	0.111	0.333	0.167	0.500	$\begin{bmatrix} 18 & 0 & 0 \\ 18 & 0 & 0 \\ 18 & 0 & 0 \end{bmatrix}$

Figura 5.16: Imagen original antes del aumento de datos.

Figura 5.17: Imagen transformada tras la aplicación del aumento de datos.

que el modelo no tiene capacidad de discriminación para distinguir entre la clase positiva y las clases negativas en cada caso. Esto se corresponde con un clasificador aleatorio.

Por ello, la reducción del número de parámetros entrenables tampoco parece ofrecer una mejora al problema de sobreajuste. Ante este empeoramiento observado, se ha decidido como siguiente paso introducir una capa densa adicional de neuronas en el clasificador del modelo para que el modelo recupere la capacidad de aprendizaje, al mismo tiempo que se intenta seguir manteniendo un número bajo de parámetros entrenables.

Experimento 11. Comprobación de la actuación del modelo del experimento 3 con conjunto de datos balanceado, congelación de todas las capas de la base convolucional e inclusión adicional de una capa densa de neuronas

El entrenamiento del modelo de este experimento ha permitido obtener, en primer lugar, la evolución de la función de pérdida y la métrica de evaluación para ambos conjuntos de datos, la cual se ofrece en la Figura 5.28.

Los valores logrados para la función de pérdida y las métricas de evaluación tras la evaluación del modelo en el conjunto de validación se presentan en la Tabla 5.11 y la correspondiente matriz de confusión se enseña en la Figura 5.29.

En base a los resultados conseguidos, se puede comprobar que se ha podido reducir notablemente los valores alcanzados para la función de pérdida. Sin embargo, no se ha conseguido mejorar la capacidad predictiva del modelo, ya que sigue apostando por una única clase al igual que en el experimento anterior, tal y como se puede corroborar por

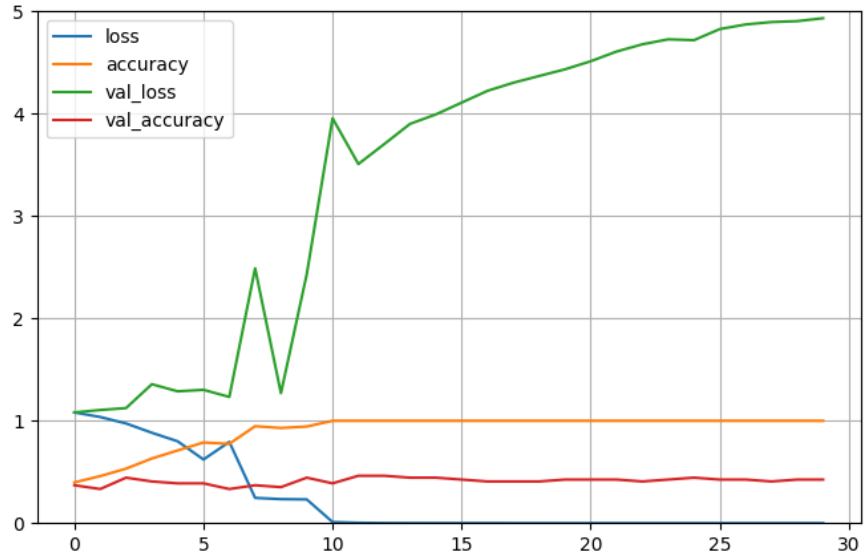

Figura 5.18: Evolución de la función de pérdida y de la métrica de evaluación durante el entrenamiento del modelo del experimento 8.

Tabla 5.11: Resultados alcanzados por el modelo del experimento 11 en el conjunto de datos de validación.

Pérdida	**Exactitud**	**Precisión**	**Exhaustividad**	**Puntuación F1**	**AUC**	**Matriz de confusión**
1.099	0.333	0.111	0.333	0.167	0.500	$\begin{bmatrix} 0 & 0 & 18 \\ 0 & 0 & 18 \\ 0 & 0 & 18 \end{bmatrix}$

medio de la Figura 5.29.

Experimento 12. Comprobación de la actuación del modelo del experimento 1 con reducción de parámetros entrenables y conjunto de datos balanceado

Una vez completado el entrenamiento y la evaluación del modelo empleado en este experimento con el conjunto de datos balanceado que ha sido generado, se han obtenido las evoluciones de la función de pérdida y la métrica de evaluación de exactitud a lo largo de las épocas de entrenamiento. Esto es representado en la Figura 5.30. Como se puede observar de dicha figura, las evoluciones alcanzadas no ofrecen cambios notables en relación a las obtenidas en experimentos anteriores.

Los resultados cuantitativos conseguidos para la función de pérdida y las métricas de evaluación tras la evaluación del modelo en el conjunto de datos de validación se muestran en la Tabla 5.12. Además, se presenta la correspondiente matriz de confusión en la Figura 5.31.

Tras el análisis de los resultados obtenidos en este experimento, se llega a la clara con-

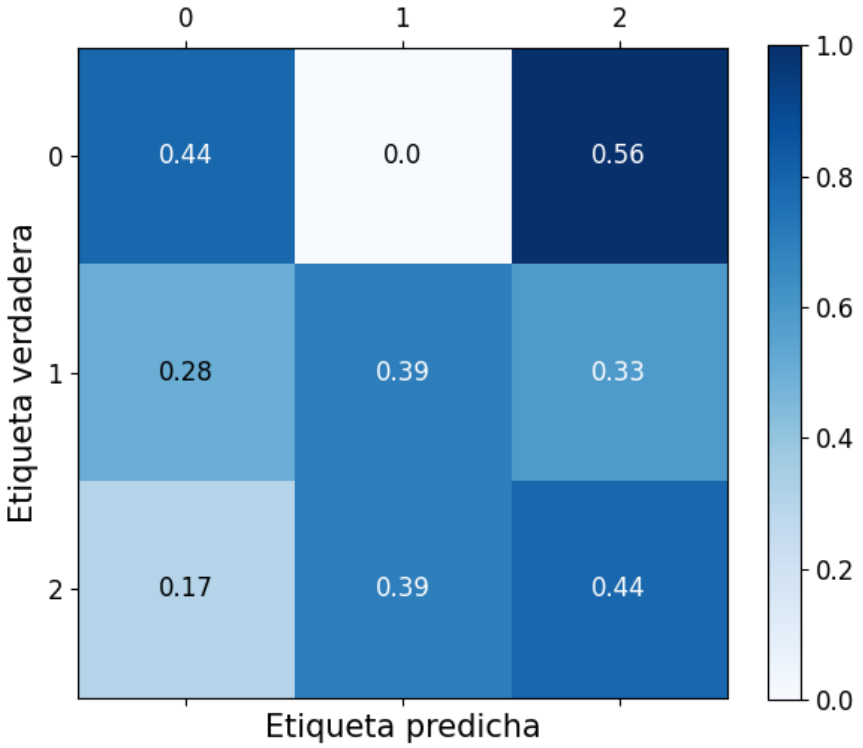

Figura 5.19: Matriz de confusión alcanzada por el modelo del experimento 8 sobre el conjunto de validación.

Tabla 5.12: Resultados alcanzados por el modelo del experimento 12 en el conjunto de datos de validación.

Pérdida	**Exactitud**	**Precisión**	**Exhaustividad**	**Puntuación F1**	**AUC**	**Matriz de confusión**
3.385	0.352	0.371	0.352	0.351	0.503	$\begin{bmatrix} 5 & 1 & 12 \\ 5 & 6 & 7 \\ 2 & 8 & 8 \end{bmatrix}$

clusión de que el sobreajuste sigue estando presente en el modelo, a pesar de haber reducido notablemente el número de parámetros entrenables. Además, visualizando la matriz de confusión de la Figura 5.31, es posible verificar que el modelo muestra cierta predilección en sus predicciones por la última clase. Por tanto, la solución al problema existente no parece estar asociada con una reducción del número de parámetros a entrenar.

En este punto y ante los resultados alcanzados hasta el momento, se decide abordar algún experimento adicional en el que se cambie la característica a predecir. Con los resultados obtenidos en la mano, se considera que el modelo no consigue aprender de forma adecuada aspectos, patrones o características clave de las imágenes y esto pudiera ser debido a que el modelo realmente no puede encontrar en dichas imágenes nada que le permita correlacionar bien con la clase a predecir. Por tanto, debido a que el grado de Nottingham resulta de mediciones hechas a nivel microscópico (pleomorfismo nuclear, tasa mitótica y formación de

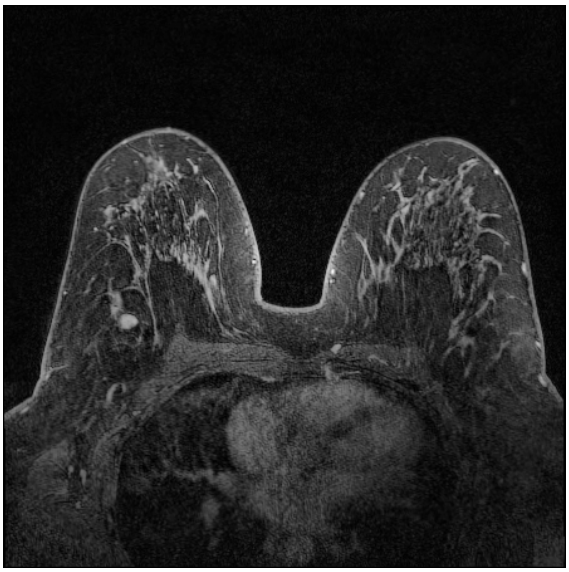

Figura 5.20: Resonancia magnética original sin aplicación de filtro.

Figura 5.21: Resonancia magnética obtenida tras el uso del filtro CLAHE.

túbulos), bien pudiera ser que en la escala macroscópica de las imágenes el modelo no es capaz de encontrar ningún rasgo que pueda estar conectado de forma directa con el grado de Nottingham. Atendiendo a esta hipótesis, se decide cambiar la característica a predecir del grado de Nottingham a otra que a priori debiera estar más directamente relacionada con aspectos observables y visibles en las imágenes de resonancia magnética con las que el modelo es entrenado. Así, la característica escogida ha sido la presencia o no del pezón en las resonancias de mama, ya que se trata de una característica observable de dichas imágenes.

Experimento 13. Comprobación de la actuación del modelo del experimento 1 con diferente característica a predecir

Después del entrenamiento del modelo con el nuevo conjunto de datos, se han obtenido resultados como el mostrado en la Figura 5.32, la cual da información una vez más de las evoluciones de la función de pérdida y la métrica de evaluación con el transcurso de las épocas de entrenamiento. A partir de ella, se puede comprobar que se alcanzan evoluciones prácticamente constantes para la métrica de evaluación tanto para el conjunto de entrenamiento como el de validación y generalmente constantes para la función de pérdida en los mismos conjuntos de datos. Esto ya deja entrever que el modelo no está aprendiendo realmente con el objetivo de mejorar sus predicciones, sino que desde el primer momento al último hace uso de la misma estrategia de predicción.

La Tabla 5.13 permite ver los valores obtenidos para las diferentes métricas de evaluación una vez se ha evaluado el modelo en el conjunto de validación. Por otro lado, la Figura 5.33 representa la matriz de confusión asociada a las predicciones acometidas por el modelo en el mismo conjunto de validación.

Figura 5.22: Resonancia magnética obtenida tras el uso del filtro de Sobel.

Figura 5.23: Resonancia magnética obtenida tras el uso del filtro de Canny.

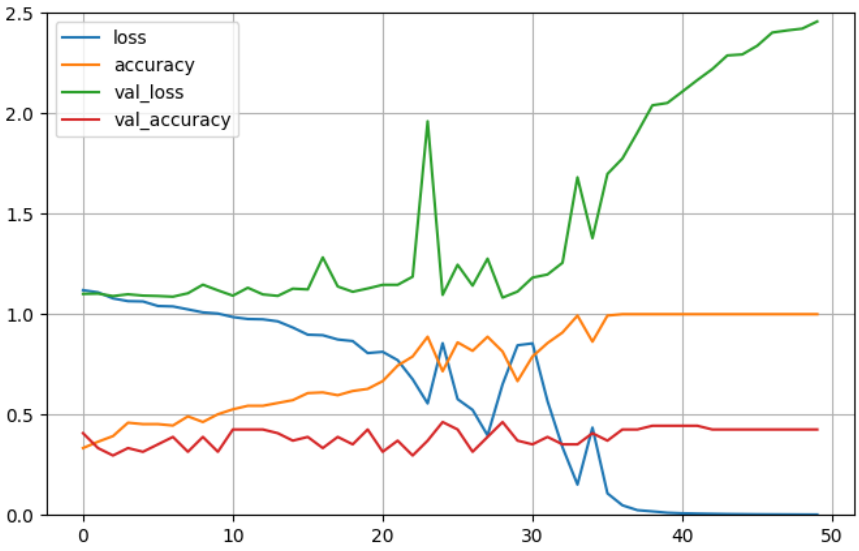

Figura 5.24: Evolución de la función de pérdida y de la métrica de evaluación durante el entrenamiento del modelo del experimento 9.

Tabla 5.13: Resultados alcanzados por el modelo del experimento 13 en el conjunto de datos de validación.

Pérdida	**Exactitud**	**Precisión**	**Exhaustividad**	**Puntuación F1**	**AUC**	**Matriz de confusión**
0.370	0.883	0.779	0.883	0.828	0.596	$\begin{bmatrix} 203 & 0 \\ 27 & 0 \end{bmatrix}$

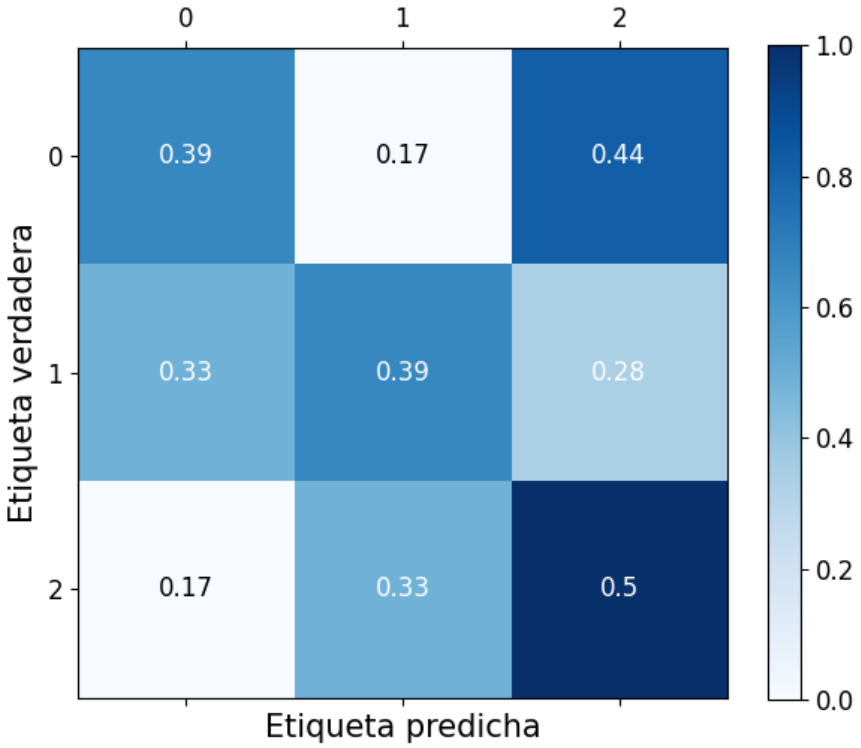

Figura 5.25: Matriz de confusión alcanzada por el modelo del experimento 9 sobre el conjunto de validación.

No hay que confundir los valores obtenidos para las métricas de evaluación de la Tabla 5.13 como muestra de una buena actuación del modelo, ya que la Figura 5.33 permite comprobar lo que realmente ocurre y es que el modelo realiza sus predicciones apostando siempre por la clase con mayor presencia en el conjunto de datos. Esto sumado a lo que se comentó anteriormente en relación a las evoluciones constantes con las épocas de entrenamiento permite concluir claramente que en ningún momento el modelo está aprendiendo ni tratando de mejorar sus predicciones, sino que simplemente apuesta siempre por la clase que se encuentra en el conjunto de datos con mayor abundancia. Por tanto, no se consigue tampoco mejorar la actuación del modelo haciendo uso de una característica que teóricamente está más directamente relacionada con aspectos visuales y observables de la propia resonancia de mama.

Experimento 14. Comprobación de la actuación del modelo del experimento 3 con diferente característica a predecir

La Figura 5.34 presenta las evoluciones de la función de pérdida y la métrica de evaluación tanto para el conjunto de entrenamiento como el de validación. Como se puede verificar, la función de pérdida en el conjunto de validación ha tomado valores demasiado altos hasta tal punto que no aparece en el gráfico. A diferencia del experimento anterior, esta vez sí que se observa un intento de mejora en las predicciones por parte del modelo, alcanzando la

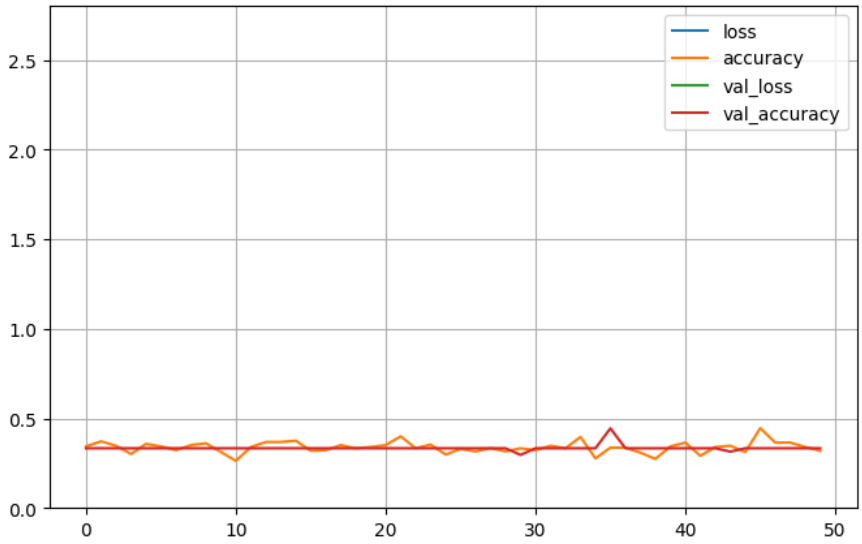

Figura 5.26: Evolución de la función de pérdida y de la métrica de evaluación durante el entrenamiento del modelo del experimento 10.

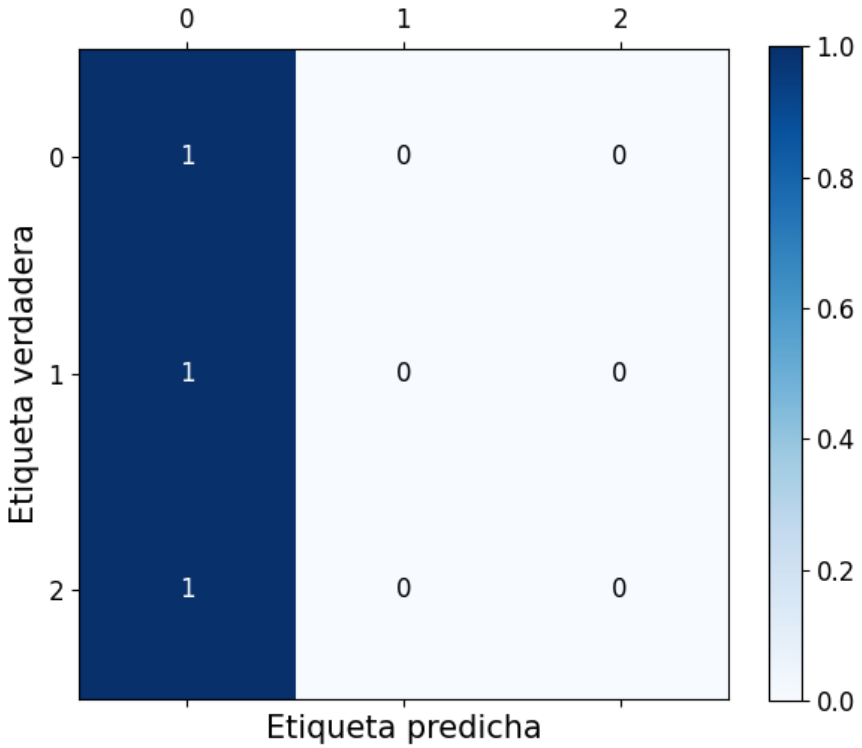

Figura 5.27: Matriz de confusión alcanzada por el modelo del experimento 10 sobre el conjunto de validación.

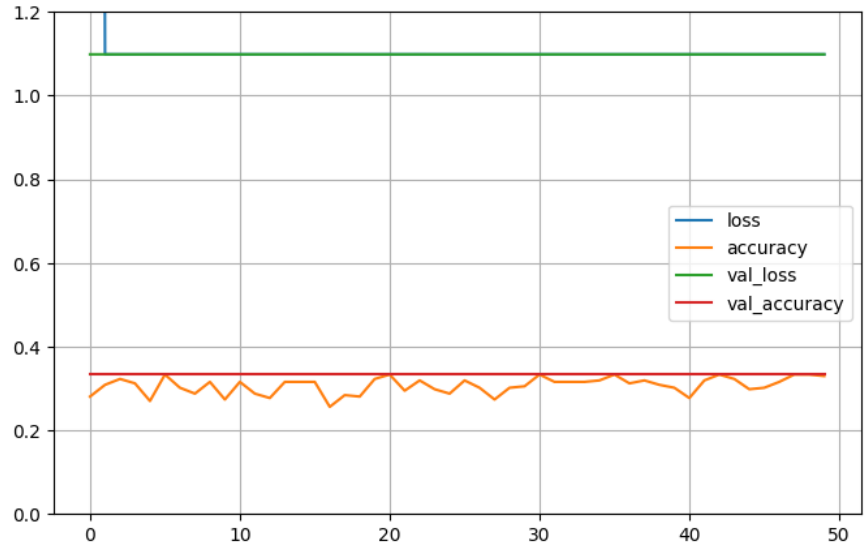

Figura 5.28: Evolución de la función de pérdida y de la métrica de evaluación durante el entrenamiento del modelo del experimento 11.

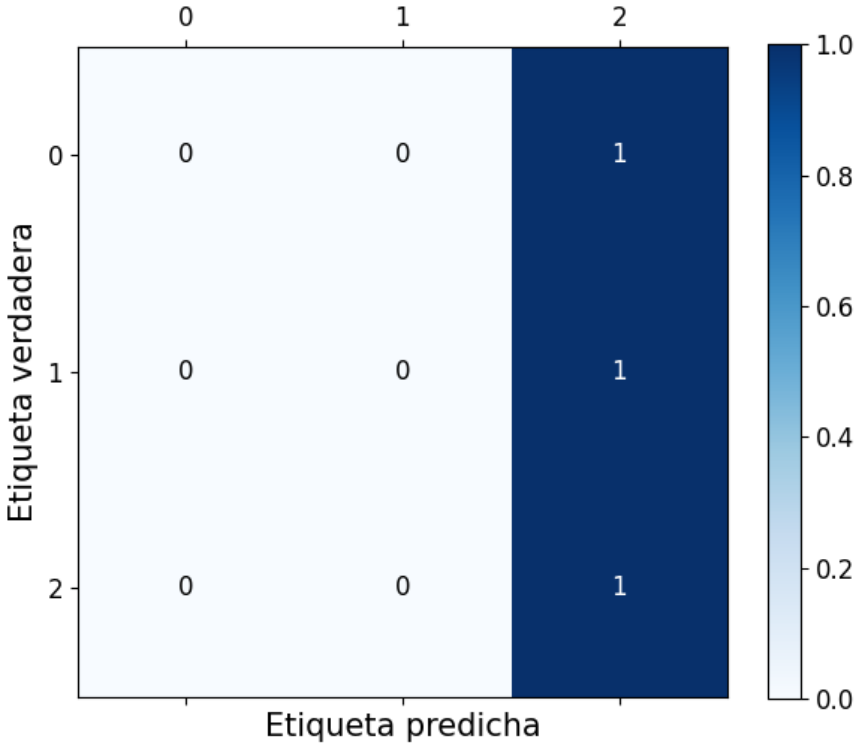

Figura 5.29: Matriz de confusión alcanzada por el modelo del experimento 11 sobre el conjunto de validación.

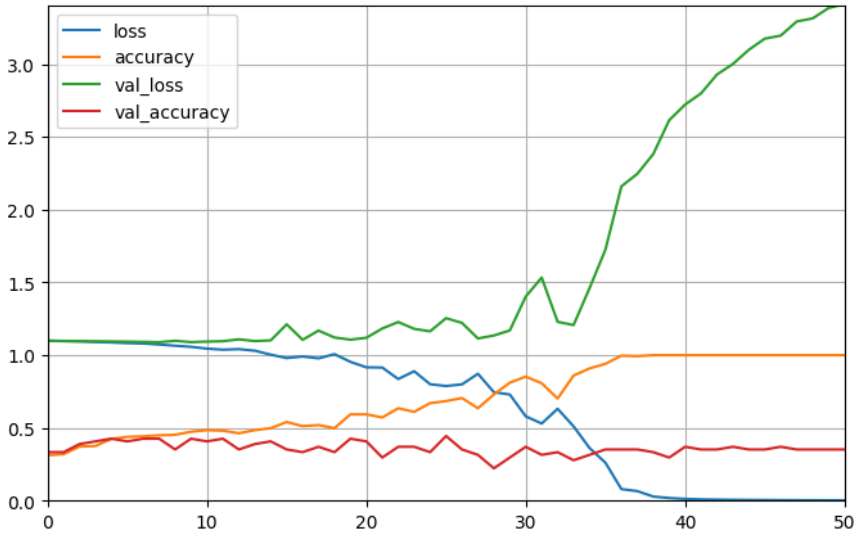

Figura 5.30: Evolución de la función de pérdida y de la métrica de evaluación durante el entrenamiento del modelo del experimento 12.

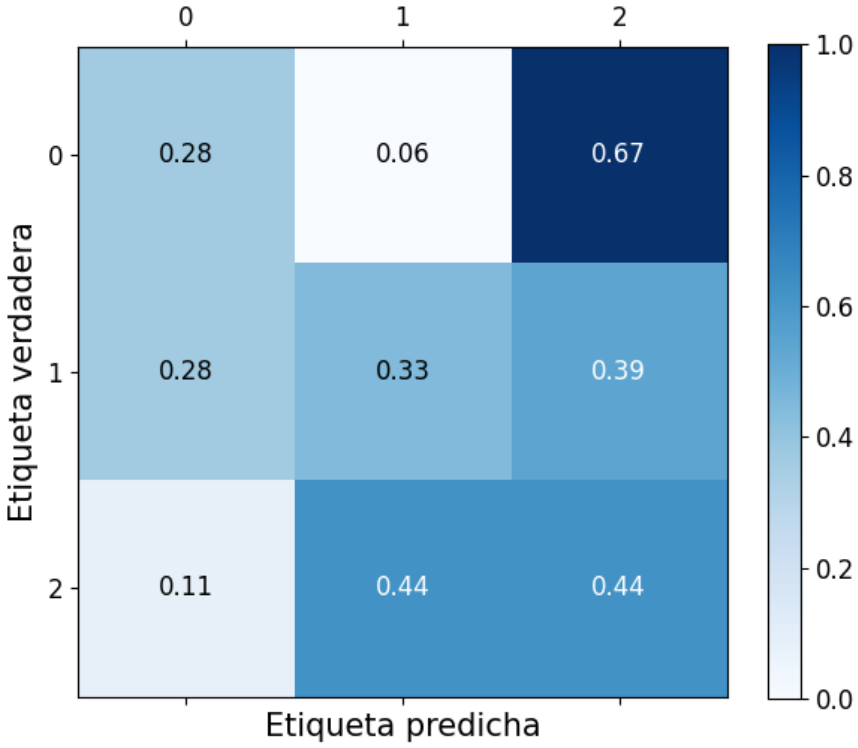

Figura 5.31: Matriz de confusión alcanzada por el modelo del experimento 12 sobre el conjunto de validación.

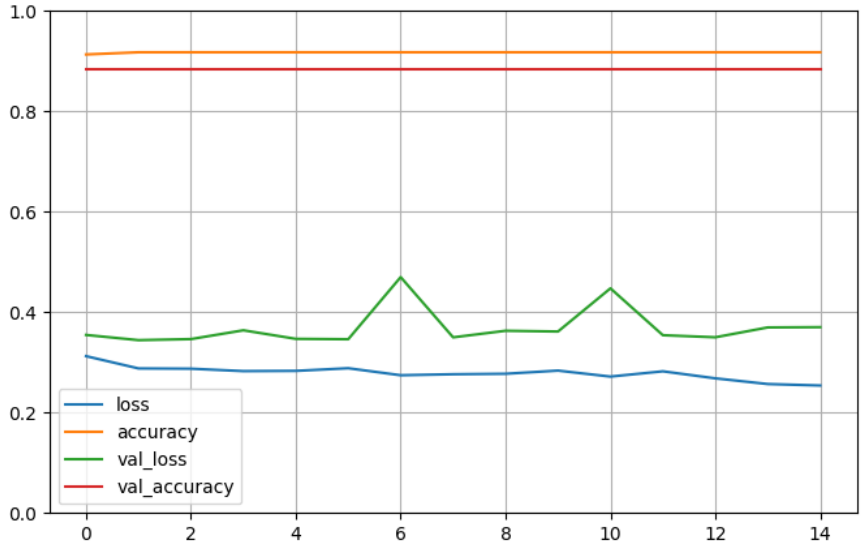

Figura 5.32: Evolución de la función de pérdida y de la métrica de evaluación durante el entrenamiento del modelo del experimento 13.

Tabla 5.14: Resultados alcanzados por el modelo del experimento 14 en el conjunto de datos de validación.

Pérdida	**Exactitud**	**Precisión**	**Exhaustividad**	**Puntuación F1**	**AUC**	**Matriz de confusión**
25.988	0.878	0.778	0.878	0.825	0.505	$\begin{bmatrix} 202 & 1 \\ 27 & 0 \end{bmatrix}$

exactitud de 1 en el conjunto de datos de entrenamiento.

La Tabla 5.14 expone los valores conseguidos para las métricas de evaluación escogidas una vez se ha evaluado el modelo en el conjunto de validación tras el entrenamiento del mismo. Adicionalmente, la Figura 5.35 presenta la correspondiente matriz de confusión en base a las predicciones hechas por el modelo en el mismo conjunto de validación.

A partir de los resultados obtenidos en este experimento, se puede corroborar que una vez más no se consigue mejorar la actuación del modelo. Si bien esta vez se ha observado un intento de hacer mejores predicciones en el conjunto de entrenamiento en comparación con el experimento anterior, en el conjunto de validación se vuelve a demostrar una predilección casi absoluta del modelo hacia la clase con mayor presencia en el conjunto de datos.

Una vez se han presentado todos los experimentos llevados a cabo en este estudio exploratorio y sus resultados, se muestra en última instancia una tabla resumen que recopila la información más importante y concluyente de cada uno de dichos experimentos. Aquellos experimentos que presentan dos filas comparan los dos casos en los que el aspecto analizado no es y sí es aplicado respectivamente. Toda esta información es incluida en la Tabla 5.15 y aparece en las páginas 115-117.

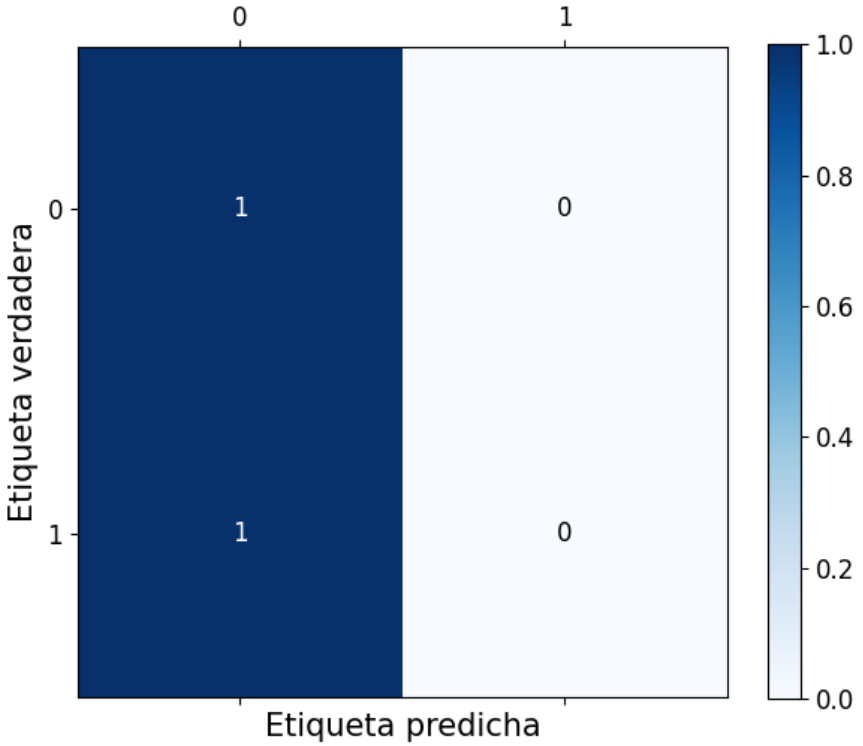

Figura 5.33: Matriz de confusión alcanzada por el modelo del experimento 13 sobre el conjunto de validación.

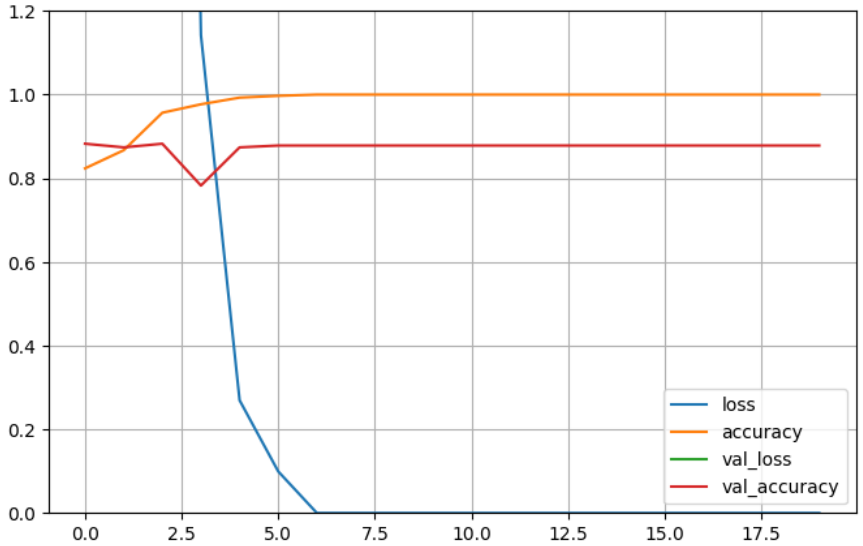

Figura 5.34: Evolución de la función de pérdida y de la métrica de evaluación durante el entrenamiento del modelo del experimento 14.

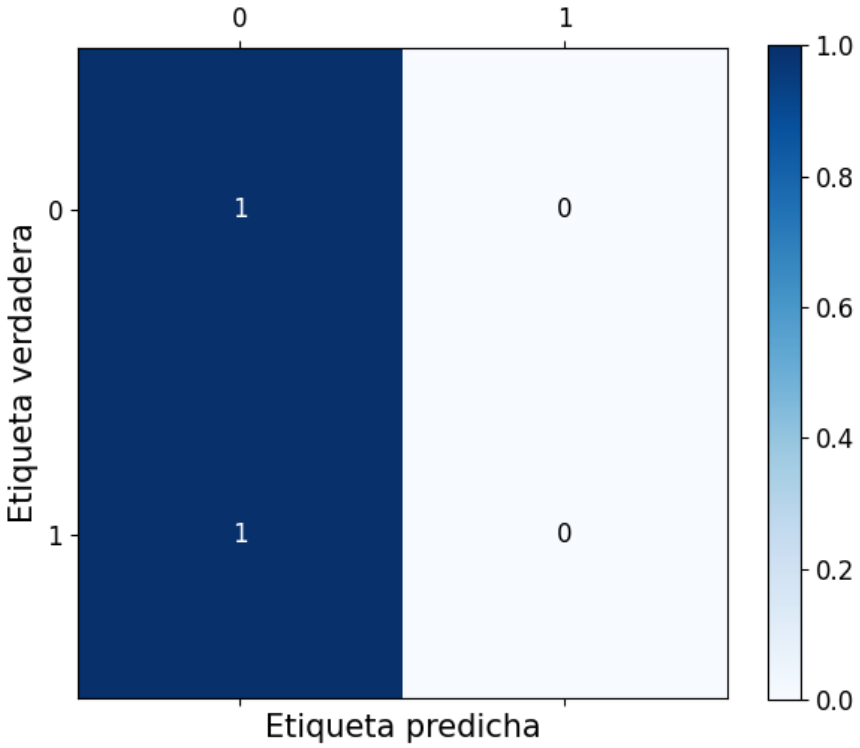

Figura 5.35: Matriz de confusión alcanzada por el modelo del experimento 14 sobre el conjunto de validación.

Tabla 5.15: Tabla resumen de los resultados alcanzados para los diferentes experimentos.

Experimento	**Aspecto analizado**	**Resultados**							**Conclusión**
		Pérdida	Exactitud	Precisión	Exhaustividad	Puntuación F1	AUC	Matriz de confusión	
1	Modelo concatenación temprana	3.055	0.440	0.402	0.440	0.417	0.557	$\begin{bmatrix} 2 & 20 & 10 \\ 10 & 52 & 21 \\ 4 & 24 & 16 \end{bmatrix}$	El modelo padece sobreajuste
2	Modelo concatenación tardía	2.824	0.509	0.471	0.509	0.484	0.592	$\begin{bmatrix} 3 & 21 & 8 \\ 8 & 57 & 18 \\ 4 & 19 & 21 \end{bmatrix}$	El modelo padece sobreajuste
3	Modelo preentrenado	6.362	0.402	0.393	0.402	0.397	0.546	$\begin{bmatrix} 3 & 17 & 12 \\ 14 & 47 & 22 \\ 9 & 21 & 14 \end{bmatrix}$	El modelo padece sobreajuste
4	Balanceo de datos	3.055	0.440	0.402	0.440	0.417	0.557	$\begin{bmatrix} 2 & 20 & 10 \\ 10 & 52 & 21 \\ 4 & 24 & 16 \end{bmatrix}$	No se soluciona el sobreajuste
		2.813	0.407	0.416	0.407	0.405	0.563	$\begin{bmatrix} 6 & 3 & 9 \\ 5 & 7 & 6 \\ 2 & 7 & 9 \end{bmatrix}$	
5	Dropout	2.813	0.407	0.416	0.407	0.405	0.563	$\begin{bmatrix} 6 & 3 & 9 \\ 5 & 7 & 6 \\ 2 & 7 & 9 \end{bmatrix}$	No se soluciona el sobreajuste
		3.009	0.352	0.361	0.352	0.349	0.523	$\begin{bmatrix} 5 & 4 & 9 \\ 5 & 6 & 7 \\ 2 & 8 & 8 \end{bmatrix}$	

Experimento	**Aspecto analizado**	**Resultados**							**Conclusión**
		Pérdida	Exactitud	Precisión	Exhaustividad	Puntuación F1	AUC	Matriz de confusión	
6	Normalización por lotes	2.813	0.407	0.416	0.407	0.405	0.563	$\begin{bmatrix} 6 & 3 & 9 \\ 5 & 7 & 6 \\ 2 & 7 & 9 \end{bmatrix}$	No se soluciona el sobreajuste
		25.195	0.370	0.393	0.370	0.376	0.526	$\begin{bmatrix} 7 & 2 & 9 \\ 3 & 6 & 9 \\ 4 & 7 & 7 \end{bmatrix}$	
7	Algoritmo de optimización	2.813	0.407	0.416	0.407	0.405	0.563	$\begin{bmatrix} 6 & 3 & 9 \\ 5 & 7 & 6 \\ 2 & 7 & 9 \end{bmatrix}$	No se soluciona el sobreajuste
		6.762	0.315	0.343	0.315	0.304	0.533	$\begin{bmatrix} 3 & 4 & 11 \\ 3 & 6 & 9 \\ 1 & 9 & 8 \end{bmatrix}$	
8	Aumento de datos	2.813	0.407	0.416	0.407	0.405	0.563	$\begin{bmatrix} 6 & 3 & 9 \\ 5 & 7 & 6 \\ 2 & 7 & 9 \end{bmatrix}$	No se soluciona el sobreajuste
		4.933	0.426	0.444	0.426	0.430	0.556	$\begin{bmatrix} 8 & 0 & 10 \\ 5 & 7 & 6 \\ 3 & 7 & 8 \end{bmatrix}$	
9	Filtros de imágenes	2.813	0.407	0.416	0.407	0.405	0.563	$\begin{bmatrix} 6 & 3 & 9 \\ 5 & 7 & 6 \\ 2 & 7 & 9 \end{bmatrix}$	No se soluciona el sobreajuste
		2.457	0.426	0.428	0.426	0.424	0.543	$\begin{bmatrix} 7 & 3 & 8 \\ 6 & 7 & 5 \\ 3 & 6 & 9 \end{bmatrix}$	
10	Congelación de capas	371.340	0.333	0.111	0.333	0.167	0.500	$\begin{bmatrix} 18 & 0 & 0 \\ 18 & 0 & 0 \\ 18 & 0 & 0 \end{bmatrix}$	No se soluciona el sobreajuste

Experimento	Aspecto analizado	Resultados							Conclusión
		Pérdida	Exactitud	Precisión	Exhaustividad	Puntuación F1	AUC	Matriz de confusión	
11	Capa densa adicional	371.340	0.333	0.111	0.333	0.167	0.500	$\begin{bmatrix} 18 & 0 & 0 \\ 18 & 0 & 0 \\ 18 & 0 & 0 \end{bmatrix}$	No se soluciona el sobreajuste
		1.099	0.333	0.111	0.333	0.167	0.500	$\begin{bmatrix} 0 & 0 & 18 \\ 0 & 0 & 18 \\ 0 & 0 & 18 \end{bmatrix}$	
12	Reducción de parámetros	2.813	0.407	0.416	0.407	0.405	0.563	$\begin{bmatrix} 6 & 3 & 9 \\ 5 & 7 & 6 \\ 2 & 7 & 9 \end{bmatrix}$	No se soluciona el sobreajuste
		3.385	0.352	0.371	0.352	0.351	0.503	$\begin{bmatrix} 5 & 1 & 12 \\ 5 & 6 & 7 \\ 2 & 8 & 8 \end{bmatrix}$	
13	Característica a predecir	3.055	0.440	0.402	0.440	0.417	0.557	$\begin{bmatrix} 2 & 20 & 10 \\ 10 & 52 & 21 \\ 4 & 24 & 16 \end{bmatrix}$	No se soluciona el sobreajuste
		0.370	0.883	0.779	0.883	0.828	0.596	$\begin{bmatrix} 203 & 0 \\ 27 & 0 \end{bmatrix}$	
14	Característica a predecir	6.362	0.402	0.393	0.402	0.397	0.546	$\begin{bmatrix} 3 & 17 & 12 \\ 14 & 47 & 22 \\ 9 & 21 & 14 \end{bmatrix}$	No se soluciona el sobreajuste
		25.988	0.878	0.778	0.878	0.825	0.505	$\begin{bmatrix} 202 & 1 \\ 27 & 0 \end{bmatrix}$	

Capítulo 6

Discusión, conclusiones y líneas futuras

Tras la presentación de los resultados alcanzados a lo largo de este trabajo exploratorio, se concluye este estudio con la exposición en el presente capítulo de la discusión, las conclusiones extraídas y el planteamiento de líneas futuras.

6.1. Discusión

Se comienza abordando la discusión de los temas más relevantes tratados en el transcurso de este documento. Así, el primer tema a discutir viene dado por los resultados alcanzados en los diferentes experimentos realizados. El denominador común de estos resultados ha sido la consecución de modelos que logran una buena actuación en el conjunto de datos de entrenamiento, pero un mal desempeño, por el contrario, en el conjunto de datos de validación. Esto viene originado por el problema de sobreajuste existente, el cual no ha podido solucionarse a pesar del uso de técnicas de regularización o de otro tipo como el aumento de datos y la aplicación de filtros de imágenes. Otros intentos como el de reducir lo máximo posible el número de parámetros entrenables del modelo o predecir una característica diferente que se correspondiera con un aspecto más visual de las imágenes no aportaron tampoco mejora alguna.

Es interesante comentar que la posibilidad de errores humanos siempre está presente, por lo que, una forma eficaz de comprobar si este problema por sobreajuste era debido a ello, era utilizar el modelo preentrenado, el cual parte de una base convolucional ya creada. No obstante, se pudo verificar que con este modelo el problema no desaparecía. Dignos de recuperar en esta discusión son los experimentos que se realizaron con este modelo preentrenado y congelando todas las capas de la base convolucional para limitar la flexibilidad excesiva de dicho modelo. En este caso, incluso se obtenían resultados peores, como se podía corroborar con las matrices de confusión que demostraban que el modelo no tenía capacidad alguna de discriminación para distinguir entre la clase positiva y las clases negativas en cada caso (el modelo apostaba las predicciones a una única clase). Esto era el máximo exponente del mal

comportamiento mostrado por los modelos en el conjunto de validación. Ante todos estos resultados obtenidos, se analizaron los mismos y se plantearon algunos posibles problemas:

- La primera posibilidad que se pensó, aunque menos probable, tras los experimentos 10 y 11 era que la red preentrenada no sería capaz de dar unos buenos resultados al darle como entrada tres imágenes diferentes cuando ha sido entrenada con color. Al no dejarle aprender las características de la imagen en las capas convolucionales y sólo permitirle clasificar con ellas, lo único que puede hacer es quedarse con la clase que más presencia tiene en el caso de un conjunto de datos no balanceado y cualquiera de ellas en el caso de un conjunto balanceado. Ante esto, surgía la propuesta de volver a usar uno de los modelos propios que fueron presentados en este documento con la inclusión de pocos parámetros para entrenar con el objetivo de comprobar de nuevo si era posible reducir el sobreajuste.

- La otra posibilidad que se valoró como más probable era que los modelos no puedan aprender con los datos que están recibiendo. Revisando los diferentes experimentos y que en ningún momento (salvo en alguna ocasión, lo cual estadísticamente es normal) se supera lo que un modelo que tan solo escogiese una clase podría obtener, es la explicación con más sentido. Cuando se le deja aprender de las características de la imagen con las redes convolucionales, el modelo sólo puede sobreentrenar y quedarse con características individuales de cada una. En el caso en que se usaba un conjunto de datos con presencia mayoritaria de una clase concreta (experimentos 1, 2 y 3), al darle al modelo imágenes nuevas y dado que éste ha aprendido sobre más características de la clase que prevalece, es más probable que escoja esta clase, aunque se reparte igualmente de forma aleatoria (manteniendo esta proporción) entre las tres clases. Si en cambio, no se le permite aprender de las características y tan solo se entrena el clasificador, como en el experimento 10 u 11, el modelo no tiene forma de sobreajustar tan fácilmente sobre imágenes individuales y solamente puede escoger una clase. El hecho de que en la gran cantidad de experimentos no se haya conseguido superar el umbral de la aleatoriedad (se ha podido comprobar que la exactitud alcanzada en el conjunto de validación para los diferentes experimentos era siempre menor que 0.5) es un gran indicativo de este problema. Normalmente, aunque haya sobreajuste, en algún momento durante el entrenamiento se supera el umbral y luego vuelve a bajar. De nuevo, los experimentos 10 y 11 en los que la red escoge tan solo una clase también son una fuerte indicación.

Antes estos posibles problemas, se plantearon también posibles causas. La primera de ellas podía ser que no se estuvieran introduciendo bien los datos a la red y que de alguna forma se tuvieran parejas que no concordasen entre sí, no pudiendo el modelo aprender nada. Esto obviamente se comprobó y se verificó finalmente que no constituía la raíz del problema.

Otra causa que se valoró en un principio era que no se pudiera estimar el grado de Nottingham a partir de las imágenes deseadas, al menos con las técnicas de redes convolucionales. Podría ser que el modelo no pudiera aprender nada porque no tiene forma de estimar este índice de ninguna manera y nuevamente se llegaría a la aleatoriedad. El planteamiento de esta hipótesis llevó a la realización posterior de los experimentos 13 y 14, los cuales abordaban la predicción de una característica diferente. Se consideró interesante escoger esta vez una característica que se relacionase con un aspecto más visual de la propia imagen como era la presencia o no del pezón en la resonancia magnética de mama, al contrario que el grado de Nottingham, el cual se basa más en mediciones a nivel microscópico y no fácilmente visibles en la imagen. Sin embargo, se pudo comprobar también que la predicción de una característica diferente no proporcionaba una mejora en la actuación del modelo, ya que éste seguía apostando por la clase mayoritaria a la hora de efectuar las predicciones.

Además de esta discusión de los resultados obtenidos en el estudio realizado, es conveniente recalcar también que, en el análisis llevado a cabo, no se ha abordado un estudio de los hiperparámetros de los modelos y la correspondiente evaluación en un conjunto de datos de prueba como habría sido normal, debido al hecho de que en ningún momento se ha conseguido un modelo que demostrase una actuación correcta en un conjunto de datos de validación y que no sufriera de sobreajuste. Por ello, no tenía sentido acometer esos pasos comentados en ausencia de un modelo robusto.

Por último, no se debe olvidar que este documento es el resultado de un estudio exploratorio que se enmarca, a su vez, dentro de un proyecto de investigación tal y como se mencionó al inicio. Es por ello, que el objetivo fundamental de este estudio era el de probar diferentes modelos de redes neuronales a los datos deseados, partiendo de lo ya abordado en el estado del arte y analizando en detalle los resultados obtenidos, no habiendo, por tanto, objetivos relacionados con la consecución de resultados concretos o de modelos de un determinado nivel de actuación.

6.2. Conclusiones

Las conclusiones extraídas en este trabajo están directamente relacionadas con los objetivos que se plantearon al comienzo de este documento. Éstos venían dados fundamentalmente por cuestiones que deseaban poder responderse tras completar el presente estudio exploratorio. Teniendo esto en cuenta, las principales conclusiones alcanzadas en dicho estudio exploratorio han sido las siguientes:

- En primer lugar, es adecuado empezar por la principal cuestión que se planteó al inicio, la cual preguntaba por si era posible inferir con un modelo de aprendizaje profundo una característica elegida a partir de las imágenes del conjunto de datos. Tras los resultados alcanzados, se ha podido verificar que esto no ha sido del todo posible,

dado que como ya se ha expuesto hasta ahora, se han obtenido modelos que padecen bastante sobreajuste, actuando muy bien en el conjunto de datos de entrenamiento, pero de manera más deficiente en el conjunto de validación.

- La cuestión que se lanzaba acerca de qué modelo demostraba ser mejor a la hora de inferir, si uno preentrenado o no, no se puede responder otra vez de manera plena por los modelos deficientes que se han alcanzado. No obstante, con los resultados obtenidos, el modelo preentrenado no ha demostrado tener una actuación superior a los otros modelos propuestos.

- Una conclusión clara que sí se ha extraído es que la aplicación de técnicas que son conocidas por su ayuda a solucionar o mejorar el problema de sobreajuste no ha aportado mejoras significativas en cuanto a la actuación de los modelos. Esto incluye las técnicas de preprocesamiento de datos que han sido empleadas como son el aumento de datos y la aplicación de filtros de imágenes.

- La respuesta a la última cuestión que se planteó (si hay presencia de sobreajuste en los modelos y, en caso afirmativo, si ha sido posible corregirlo) viene a resumir la principal conclusión alcanzada en este estudio exploratorio y es la presencia notable confirmada de sobreajuste en los modelos y la imposibilidad comprobada para corregir dicho problema.

6.3. Líneas futuras

Finalmente, como en todo estudio, el presentado en este documento admite posibles líneas futuras de continuación y mejora, las cuales se comentan a continuación:

- El primer trabajo futuro que se podría acometer a partir de este proyecto es el de recolectar y emplear un conjunto de datos más completo. Por ejemplo, podría ser un conjunto que no contenga tantos datos ausentes para la característica que se desea predecir (un cierto número de pacientes no tenía registro del grado de Nottingham y para muchas más no se disponía del dato de presencia o no del pezón en sus correspondientes resonancias magnéticas, reduciéndose el conjunto de casos, de esa forma, a muchos menos de los existentes). Otro punto mejorable de los datos a utilizar sería el de una mayor variedad incluyendo, por ejemplo, resonancias magnéticas de pacientes con tumores benignos, además de las correspondientes a aquéllas que tengan tumores malignos. Esto permitiría obtener modelos que pudieran ser aplicados a ambos tipos de pacientes. Es conveniente recordar que los modelos alcanzados en este estudio se han entrenado únicamente con imágenes pertenecientes a pacientes diagnosticadas con la enfermedad, por lo que ésta podría ser una línea de continuación interesante.

- Otra línea de mejora futura para el proyecto aquí realizado sería contar con un radiólogo experto que pudiera servir de soporte o como verdad de referencia en relación a las características que se desean predecir de las imágenes de resonancia magnética. Por ejemplo, en lo que se refiere a la última característica que se intentó predecir, es decir, la presencia o no del pezón, es cierto que en ciertas ocasiones había un poco de confusión, ya que en la resonancia correspondiente a una paciente para la que supuestamente había presencia del pezón, no se intuía visualmente el mismo. Es por ello, que en este tipo de situaciones sería adecuada la supervisión o el apoyo de un radiólogo experto que pudiera aclarar.

- Un aspecto muy interesante de continuación desde el estudio aquí realizado sería el de crear modelos multi-output, es decir, modelos que acometan la predicción de diferentes características. Es posible que un modelo que abordase la predicción de más de una característica encontrase algún tipo de relación entre las mismas, pudiendo ser esto beneficioso para la actuación global del modelo.

- Por último, otra línea de continuación podría ser la exploración en otro tipo de tarea a la aquí abordada como, por ejemplo, la detección del tumor por medio de las imágenes de resonancia magnética o alguna tarea relacionada con la radiogenómica.

Bibliografía

W. Alomaim, D. O'Leary, and J. Ryan et al. Subjective versus quantitative methods of assessing breast density. *Diagnostics*, 10(5):331, 2020. doi: 10.3390/diagnostics10050331.

N. Antropova, B. Q. Huynh, and M. L. Giger. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. *Medical Physics*, 44 (10):5162–5171, 2017. doi: 10.1002/mp.12453.

N. Antropova, H. Abe, and M. L. Giger. Use of clinical MRI maximum intensity projections for improved breast lesion classification with deep convolutional neural networks. *Journal of Medical Imaging*, 5(1):014503–014503, 2018. doi: 10.1117/1.JMI.5.1.014503.

A. B. Ashraf, D. Daye, and S. Gavenonis et al. Identification of intrinsic imaging phenotypes for breast cancer tumors: preliminary associations with gene expression profiles. *Radiology*, 272(2):374–384, 2014. doi: 10.1148/radiol.14131375.

A. M. Ayalew, A. O. Salau, and Y. Tamyalew et al. X-Ray image-based COVID-19 detection using deep learning. *Multimedia Tools and Applications*, pages 1–19, 2023. doi: 10.1007/s11042-023-15389-8.

T. B. Bevers, B. O. Anderson, and E. Bonaccio et al. Breast cancer screening and diagnosis. *Journal of the National Comprehensive Cancer Network*, 7(10):1060–1096, 2009. doi: 10.6004/jnccn.2009.0070.

N. F. Boyd, J. W. Byng, and R. A. Jong et al. Quantitative classification of mammographic densities and breast cancer risk: results from the Canadian National Breast Screening Study. *Journal of the National Cancer Institute*, 87(9):670–675, 1995. doi: 10.1093/jnci/87.9.670.

N. F. Boyd, L. J. Martin, and M. J. Yaffe et al. Mammographic density and breast cancer risk: current understanding and future prospects. *Breast Cancer Research*, 13:1–12, 2011. doi: 10.1186/bcr2942.

P. Carmeliet and R. Jain. Angiogenesis in cancer and other diseases. *Nature*, 407(6801): 249–257, 2000. doi: 10.1038/35025220.

Y. C. Chang, Y. H. Huang, and C. S. Huang et al. Classification of breast mass lesions using model-based analysis of the characteristic kinetic curve derived from fuzzy c-means clustering. *Magnetic Resonance Imaging*, 30(3):312–322, 2012. doi: 10.1016/j.mri.2011.12.002.

Y. C. Chang, Y. H. Huang, and C. S. Huang et al. Computerized breast lesions detection using kinetic and morphologic analysis for dynamic contrast-enhanced MRI. *Magnetic Resonance Imaging*, 32(5):514–522, 2014. doi: 10.1016/j.mri.2014.01.008.

W. Chen, M. L. Giger, and U. Bick et al. Automatic identification and classification of characteristic kinetic curves of breast lesions on DCE-MRI. *Medical Physics*, 33(8):2878–2887, 2006. doi: 10.1118/1.2210568.

I. M. Coelho, V. N. Coelho, and L. S. Ochi et al. A GPU deep learning metaheuristic based model for time series forecasting. *Applied Energy*, 201:412–418, 2017. doi: 10.1016/j.apenergy.2017.01.003.

J. Costa and J. A. Soria. *Resonancia magnética dirigida a técnicos superiores en imagen para el diagnóstico*. Elsevier, 2nd edition, 2021. ISBN 978-84-9113-646-0.

M. U. Dalmis, A. Gubern-Mérida, and S. Vreemann. A computer-aided diagnosis system for breast DCE-MRI at high spatiotemporal resolution. *Medical Physics*, 43(1):84–94, 2016. doi: 10.1118/1.4937787.

M. U. Dalmis, G. Litjens, and K. Holland et al. Using deep learning to segment breast and fibroglandular tissue in MRI volumes. *Medical Physics*, 44(2):533–546, 2017. doi: 10.1002/mp.12079.

M. U. Dalmis, S. Vreemann, and T. Kooi et al. Fully automated detection of breast cancer in screening MRI using convolutional neural networks. *Journal of Medical Imaging*, 5(1):014502–014502, 2018. doi: 10.1117/1.JMI.5.1.014502.

L. Deng and D. Yu. Deep learning: Methods and applications. *Foundations and Trends in Signal Processing*, 7(3-4):197–387, 2014. doi: 10.1561/2000000039.

B. N. Dontchos, H. Rahbar, and S. C. Partridge et al. Are qualitative assessments of background parenchymal enhancement, amount of fibroglandular tissue on MR images, and mammographic density associated with breast cancer risk? *Radiology*, 276(2):371–380, 2015. doi: 10.1148/radiol.2015142304.

E. R. Dougherty and R. A. Lotufo. *Hands-on morphological image processing*. SPIE Press, 2003.

D. G. Evans, E. F. Harkness, and A. Howell et al. Intensive breast screening in BRCA2 mutation carriers is associated with reduced breast cancer specific and all cause mortality. *Hereditary Cancer in Clinical Practice*, 14(1):8, 2016. doi: 10.1186/s13053-016-0048-3.

P. E. Freer. Mammographic breast density: impact on breast cancer risk and implications for screening. *Radiographics*, 35(2):302–315, 2015. doi: 10.1148/rg.352140106.

C. Garbin, X. Zhu, and O. Marques. Dropout vs. batch normalization: an empirical study of their impact to deep learning. *Multimedia Tools and Applications*, 79:12777–12815, 2020. doi: 10.1007/s11042-019-08453-9.

M. L. Giger. Computerized analysis of images in the detection and diagnosis of breast cancer. *Seminars in Ultrasound, CT and MRI*, 25(5):411–418, 2004. doi: 10.1053/j.sult.2004.07.003.

M. L. Giger, H. P. Chan, and J. Boone. Anniversary paper: History and status of CAD and quantitative image analysis: The role of medical physics and AAPM. *Medical Physics*, 35 (12):5799–5820, 2008. doi: 10.1118/1.3013555.

K. G. Gilhuijs, M. L. Giger, and U. Bick. Computerized analysis of breast lesions in three dimensions using dynamic magnetic-resonance imaging. *Medical Physics*, 25(9):1647–1654, 1998. doi: 10.1118/1.598345.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. ISBN 978-0262035613.

A. Gubern-Mérida, R. Martí, and J. Melendez et al. Automated localization of breast cancer in DCE-MRI. *Medical Image Analysis*, 20(1):265–274, 2015. doi: 10.1016/j.media.2014.12.001.

A. Gubern-Mérida, S. Vreemann, and R. Martí et al. Automated detection of breast cancer in false-negative screening MRI studies from women at increased risk. *European Journal of Radiology*, 85(2):472–479, 2016. doi: 10.1016/j.ejrad.2015.11.031.

A. Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, 2nd edition, 2019. ISBN 978-1-492-03264-9.

S. Haykin. *Neural networks and learning machines*. Pearson Prentice Hall, 3rd edition, 2009.

K. He, X. Zhang, and S. Ren et al. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778. CVPR, 2016.

P. Herent, B. Schmauch, and P. Jehanno et al. Detection and characterization of MRI breast lesions using deep learning. *Diagnostic and Interventional Imaging*, 100(4):219–225, 2019. doi: 10.1016/j.diii.2019.02.008.

S. H. Heywang, D. Hahn, and H. Schmidt et al. MR imaging of the breast using gadolinium-DTPA. *Journal of Computer Assisted Tomography*, 10(2):199–204, 1986. doi: 10.1097/00004728-198603000-00005.

S. Hu, C. Park, and C. O. Lew et al. Fully automated deep learning method for fibroglandular tissue segmentation in breast MRI. *Research Square*, 2022. doi: 10.21203/rs.3.rs-1606703/v1. Versión preliminar.

G. Huang, Z. Liu, and L. Van Der Maaten et al. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 448–456. Proceedings of Machine Learning Research, 2015.

W. A. Kaiser and E. Zeitler. MR imaging of the breast: fast imaging sequences with and without Gd-DTPA. preliminary observations. *Radiology*, 170(3):681–686, 1989. doi: 10.1148/radiology.170.3.2916021.

P. Kang and S. Cho. EUS SVMs: Ensemble of under-sampled SVMs for data imbalance problems. In *Neural Information Processing*, pages 837–846. Springer, 2006. doi: 10.1007/11893028_93.

D. Karras and G. Mertzios. New PDE-based methods for image enhancement using SOM and Bayesian inference in various discretization schemes. *Measurement, Science and Technology*, 20(10):104012, 2009. doi: 10.1088/0957-0233/20/10/104012.

V. King, J. D. Brooks, and J. L. Bernstein et al. Background parenchymal enhancement at breast MR imaging and breast cancer risk. *Radiology*, 260(1):50–60, 2011. doi: 10.1148/radiol.11102156.

K. Kira and L. A. Rendell. A practical approach to feature selection. In *Machine Learning Proceedings 1992*, pages 249–256. 1992. doi: 10.1016/B978-1-55860-247-2.50037-1.

M. V. Knopp, E. Weiss, and H. P. Sinn et al. Pathophysiologic basis of contrast enhancement in breast tumors. *Journal of Magnetic Resonance Imaging*, 10(3):260–266, 1999. doi: 10.1002/(SICI)1522-2586(199909)10:3<260::AID-JMRI6>3.0.CO;2-7.

R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997. doi: 10.1016/S0004-3702(97)00043-X.

N. B. Konyer, E. A. Ramsay, and M. J. Bronskill et al. Comparison of MR imaging breast coils. *Radiology*, 222(3):830–834, 2002. doi: 10.1148/radiol.2223001310.

C. K. Kuhl, S. Schrading, and C. C. Leutner. Mammography, breast ultrasound, and magnetic resonance imaging for surveillance of women at high familial risk for breast cancer. *Journal of Clinical Oncology*, 23(33):8469–8476, 2005. doi: 10.1200/JCO.2004.00.4960.

Y. Lecun, L. Bottou, and Y. Bengio et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi: 10.1038/nature14539.

C. D. Lehman and R. A. Smith. The role of MRI in breast cancer screening. *Journal of the National Comprehensive Cancer Network*, 7(10):1109–1115, 2009. doi: 10.6004/jnccn.2009.0072.

H. Li, Y. Zhu, and E. S. Burnside et al. MR imaging radiomics signatures for predicting the risk of breast cancer recurrence as given by research versions of MammaPrint, Oncotype DX, and PAM50 gene assays. *Radiology*, 281(2):382–391, 2016. doi: 10.1148/radiol.2016152110.

M. Livne, J. Rieger, and O. U. Aydin et al. A U-Net deep learning framework for high performance vessel segmentation in patients with cerebrovascular disease. *Frontiers in Neuroscience*, 13:97, 2019. doi: 10.3389/fnins.2019.00097.

S. J. Lord, W. Lei, and P. Craft et al. A systematic review of the effectiveness of magnetic resonance imaging (MRI) as an addition to mammography and ultrasound in screening young women at high risk of breast cancer. *European Journal of Cancer*, 43(13):1905–1917, 2007. doi: 10.1016/j.ejca.2007.06.007.

W. Lu, Z. Li, and J. Chu. A novel computer-aided diagnosis system for breast MRI based on feature selection and ensemble learning. *Computers in Biology and Medicine*, 83:157–165, 2017. doi: 10.1016/j.compbiomed.2017.03.002.

E. Mahoro and M. A. Akhloufi. Applying deep learning for breast cancer detection in radiology. *Current Oncology*, 29(11):8767–8793, 2022. doi: 10.3390/curroncol29110690.

R. M. Mann, C. K. Kuhl, and K. Kinkel et al. Breast MRI: guidelines from the European Society of Breast Imaging. *European Radiology*, 18:1307–1318, 2008. doi: 10.1007/s00330-008-0863-7.

R. M. Mann, N. Cho, and L. Moy. Breast MRI: State of the art. *Radiology*, 292(3):520–536, 2019. doi: 10.1148/radiol.2019182947.

M. A. Marino, T. Helbich, and P. Baltzer et al. Multiparametric MRI of the breast: A review. *Journal of Magnetic Resonance Imaging*, 47(2):301–315, 2018. doi: 10.1002/jmri.25790.

G. Mariscotti, N. Houssami, and M. Durando et al. Accuracy of mammography, digital breast tomosynthesis, ultrasound and MR imaging in preoperative assessment of breast cancer. *Anticancer Research*, 34(3):1219–1225, 2014. URL `https://ar.iiarjournals.org/content/34/3/1219`.

V. A. McCormack and I. dos Santos Silva. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. *Cancer Epidemiology Biomarkers and Prevention*, 15(6):1159–1169, 2006. doi: 10.1158/1055-9965.EPI-06-0034.

W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(1943):115–133, 1943. doi: 10.1007/BF02478259.

L. A. Meinel, A. H. Stolpen, and K. S. Berbaum et al. Breast MRI lesion classification: Improved performance of human readers with a backpropagation neural network computer-aided diagnosis (CAD) system. *Journal of Magnetic Resonance Imaging*, 25(1):89–95, 2007. doi: 10.1002/jmri.20794.

M. Meng, M. Zhang, and D. Shen et al. Differentiation of breast lesions on dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) using deep transfer learning based on DenseNet201. *Medicine*, 101(45):e31214, 2022. doi: 10.1097/MD.0000000000031214.

O. A. Montesinos, A. Montesinos, and J. Crossa. *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, chapter Fundamentals of Artificial Neural Networks and Deep Learning, page 379–425. Springer, 2022. doi: 10.1007/978-3-030-89010-0_10.

N. R. Mudigonda, R. Rangayyan, and J. L. Desautels. Gradient and texture analysis for the classification of mammographic masses. *IEEE Transactions on Medical Imaging*, 19(10): 1032–1043, 2000. doi: 10.1109/42.887618.

Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(\frac{1}{k^2})$. *Doklady Akademii Nauk SSSR*, 269(3):543–547, 1983.

N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

L. Pérez and J. Wang. The effectiveness of data augmentation in image classification using deep learning. Descargado de arXiv con doi: 10.48550/arXiv.1712.04621, 2017.

Z. Ping, Y. Xia, and T. Shen et al. A microscopic landscape of the invasive breast cancer genome. *Scientific Reports*, 6(1):27545, 2016. doi: 10.1038/srep27545.

D. Ping Tian. A review on image feature extraction and representation techniques. *International Journal of Multimedia and Ubiquitous Engineering*, 8(4):385–396, 2013.

T. Portnoi, A. Yala, and T. Schuster et al. Deep learning model to assess cancer risk on the basis of a breast MR image alone. *American Journal of Roentgenology*, 213(1):227–233, 2019. doi: 10.2214/AJR.18.20813.

A. Prat, J. S. Parker, and C. Fan et al. PAM50 assay and the three-gene model for identifying the major and clinically relevant molecular subtypes of breast cancer. *Breast Cancer Research and Treatment*, 135:301–306, 2012. doi: 10.1007/s10549-012-2143-0.

J. R. Quinlan. *C4.5: programs for machine learning*. Elsevier, 2014.

H. Rahbar and S. C. Partridge. Multiparametric MR imaging of breast cancer. *Magnetic Resonance Imaging Clinics*, 24(1):223–238, 2016. doi: 10.1016/j.mric.2015.08.012.

R. Rasti, M. Teshnehlab, and S. L. Phung. Breast cancer diagnosis in DCE-MRI using mixture ensemble of convolutional neural networks. *Pattern Recognition*, 72:381–390, 2017. doi: 10.1016/j.patcog.2017.08.004.

K. Ravichandran, N. Braman, and A. Janowczyk et al. A deep learning classifier for prediction of pathological complete response to neoadjuvant chemotherapy from baseline breast DCE-MRI. *Medical Imaging*, 10575:79–88, 2018. doi: 10.1117/12.2294056.

A. F. Saftlas, R. N. Hoover, and L. A. Brinton et al. Mammographic densities and risk of breast cancer. *Cancer*, 67(11):2833–2838, 1991. doi: 10.1002/1097-0142(19910601)67:11<2833::AID-CNCR2820671121>3.0.CO;2-U.

A. Saha, M. R. Harowicz, and L. J. Grimm et al. A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. *British Journal of Cancer*, 119(4):508–516, 2018. doi: 10.1038/s41416-018-0185-8.

I. H. Sarker. Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, 2(6):420, 2021a. doi: 10.1007/s42979-021-00815-1.

I. H. Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(160):1–21, 2021b. doi: 10.1007/s42979-021-00592-x.

I. H. Sarker, M. H. Furhad, and R. Nowrozy. AI-driven cybersecurity: An overview, security intelligence modeling and research directions. *SN Computer Science*, 2(173):1–18, 2021. doi: 10.1007/s42979-021-00557-0.

D. Saslow, C. Boetes, and W. Burke et al. American Cancer Society guidelines for breast screening with MRI as an adjunct to mammography. *CA: A Cancer Journal for Clinicians*, 57(2):75–89, 2007. doi: 10.3322/canjclin.57.2.75.

M. Schnall and S. Orel. Breast MR imaging in the diagnostic setting. *Magnetic Resonance Imaging Clinics*, 14(3):329–337, 2006. doi: 10.1016/j.mric.2006.07.004.

D. Sheth and M. L. Giger. Artificial intelligence in the interpretation of breast cancer on MRI. *Journal of Magnetic Resonance Imaging*, 51(5):1310–1324, 2020. doi: 10.1002/jmri.26878.

A. Shimauchi, M. L. Giger, and N. Bhooshan et al. Evaluation of clinical breast MR imaging performed with prototype computer-aided diagnosis breast MR imaging workstation: reader study. *Radiology*, 258(3):696–704, 2011. doi: 10.1148/radiol.10100409.

N. Srivastava, G. Hinton, and A. Krizhevsky et al. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. doi: 10.5555/2627435.2670313.

L. J. Veer, H. Dai, and M. J. Van De Vijver et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, 2002. doi: 10.1038/415530a.

E. Warner. The role of magnetic resonance imaging in screening women at high risk of breast cancer. *Topics in Magnetic Resonance Imaging*, 19(3):163–169, 2008. doi: 10.1097/RMR.0b013e31818bc994.

E. Warner, H. Messersmith, and P. Causer et al. Systematic review: Using magnetic resonance imaging to screen women at high risk for breast cancer. *Annals of Internal Medicine*, 148(9):671–679, 2008. doi: 10.7326/0003-4819-148-9-200805060-00007.

J. Wasserman. Grado histológico de Nottingham, 2022. URL `https://www.mypathologyreport.ca/es/pathology-dictionary/nottingham-histologic-grade/`. *My Pathology Report*. Descargado el 05/07/2023.

J. Wu, X. Sun, and J. Wang et al. Identifying relations between imaging phenotypes and molecular subtypes of breast cancer: model discovery and external validation. *Journal of Magnetic Resonance Imaging*, 46(4):1017–1027, 2017. doi: 10.1002/jmri.25661.

L. M. Wu, J. N. Hu, and H. Y. Gu et al. Can diffusion-weighted MR imaging and contrast-enhanced MR imaging precisely evaluate and predict pathological response to neoadjuvant chemotherapy in patients with breast cancer? *Breast Cancer Research and Treatment*, 135: 17–28, 2012. doi: 10.1007/s10549-012-2033-5.

Y. Xin, L. Kong, and Z. Liu et al. Machine learning and deep learning methods for cybersecurity. *IEEE Access*, 6:35365–35381, 2018. doi: 10.1109/ACCESS.2018.2836950.

E. D. Yeh, D. Georgian-Smith, and S. Raza et al. Positioning in breast MR imaging to optimize image quality. *RadioGraphics*, 34(1):E1–E17, 2014. doi: 10.1148/rg.341125193.

Y. Yuan, X. S. Chen, and S. Y. Liu et al. Accuracy of MRI in prediction of pathologic complete remission in breast cancer after preoperative therapy: a meta-analysis. *American Journal of Roentgenology*, 195(1):260–268, 2010. doi: 10.2214/AJR.09.3908.

Z. H. Zhou. *Machine Learning*. Tsinghua University Press, 2016. ISBN 978-981-15-1966-6.

Y. Zhu, H. Li, and W. Guo et al. Deciphering genomic underpinnings of quantitative MRI-based radiomic phenotypes of invasive breast carcinoma. *Scientific Reports*, 5(1):17787, 2015. doi: 10.1038/srep17787.