Proyecto Final de Carrera

Instituto Tecnológico de Buenos Aires

Clasificación de tumores en mamografías utilizando redes neuronales convolucionales

Autora:

Carolina Mondino

Tutor:

Matías Nazareth Tajerian

Institución:

Hospital Italiano de Buenos Aires

Buenos Aires
26 de Marzo del 2021

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

Según la Agencia Internacional de Investigación sobre Cáncer (*International Agency for Research on Cancer*, IARC por sus siglas en inglés), en el año 2020, 34.332 mujeres murieron en Argentina debido al cáncer. De esos casos, el mayor porcentaje fue causado por cáncer de mama (6.821 muertes, 19,9 %), seguido del de colon (4.037 muertes, 11,8 %) y del de pulmón (3.848 muertes, 11,2 %). En el gráfico a continuación, se observa la distribución de muertes en mujeres por cáncer:

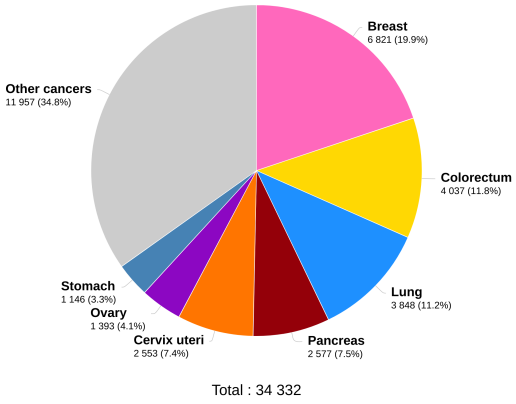

Figura 1.1: Números estimados de muertes de mujeres por cáncer en el año 2020 en Argentina. Fuente: Globocan 2020

Una **tasa estandarizada por edad** (*Age-standardised rate*; ASR por sus siglas en inglés) es una medida resumida de la tasa que tendría una población si tuviera una estructura de edad estándar. La estandarización es necesaria cuando se comparan varias poblaciones que difieren con respecto a la edad debido a que la misma tiene una poderosa influencia en el riesgo de morir de cáncer. El ASR es una media ponderada de las tasas específicas por edad; los pesos se toman de la distribución

poblacional de la población estándar. La población estándar mundial usada en la aplicación es la propuesta por Segi (1960) [1] y modificada por Doll et al. (1966) [2]. Esta tasa se expresa cada 100.000 habitantes.

La **tasa de incidencia de cáncer de mama estandarizada por edad** de Argentina en el año 2020 fue de 73,1 representando una incidencia alta. A continuación, se puede observar un mapa con la incidencia estandarizada de cada país.

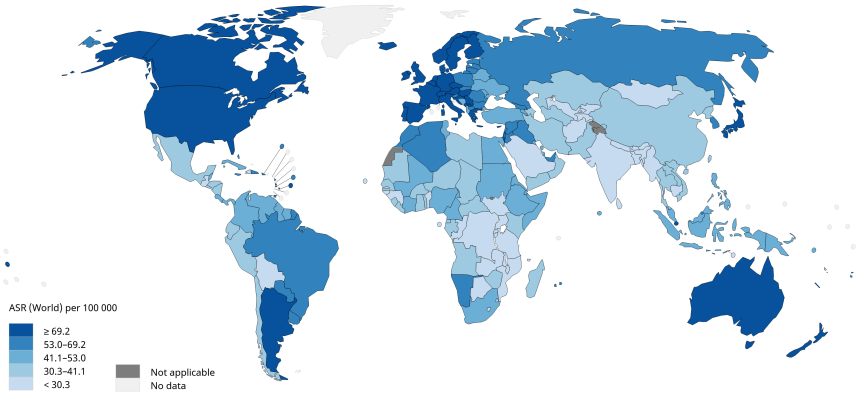

Figura 1.2: Mapa con la tasa estandarizada estimada de incidencia de cáncer de mama en el año 2020. Fuente: Globocan 2020

Además, Argentina presenta una de las tasas más altas de incidencia a nivel mundial de este cáncer con 22.024 nuevos casos en el año 2020.

Por otro lado, la **tasa de mortalidad de cáncer de mama estandarizada por edad** en el año 2020 fue de 18,9 siendo también media-alta. A continuación se observa el mapa con la mortalidad estandarizada de cada país del mundo.

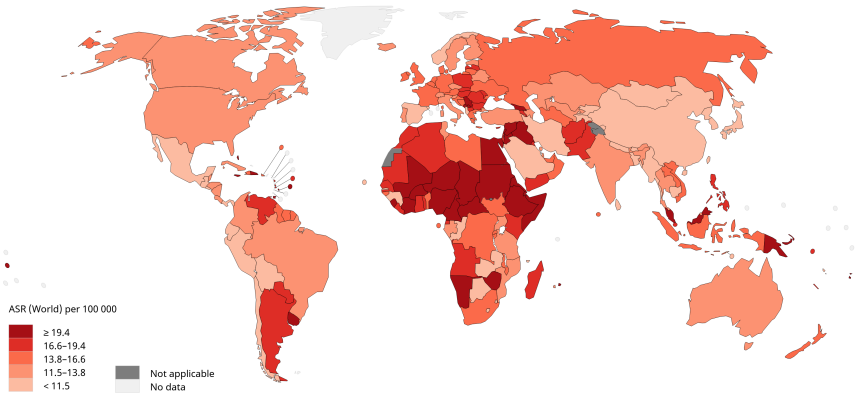

Figura 1.3: Mapa con la tasa estandarizada estimada de mortalidad de cáncer de mama en el año 2020. Fuente: Globocan 2020

Por lo tanto, Argentina posee una tasa elevada tanto de incidencia como de mortalidad para el cáncer de mama comparado con el resto de los países. Cada año, existen más de 22.000 casos nuevos de cáncer de mama; siendo el 32 % del total de los tipos de cáncer presentes en las mujeres.

A nivel mundial, según la IARC, el cáncer de mama es la causa más común de muerte por cáncer en mujeres, con 684.996 muertes en el año 2020. Además, de todos los tipos de cáncer, es el que presenta mayor incidencia en las mujeres, con más de 2.000.000 de nuevos casos en el año 2020. En los gráficos a continuación, se observa la distribución de incidencia y muertes en mujeres por cáncer de mama a nivel mundial.

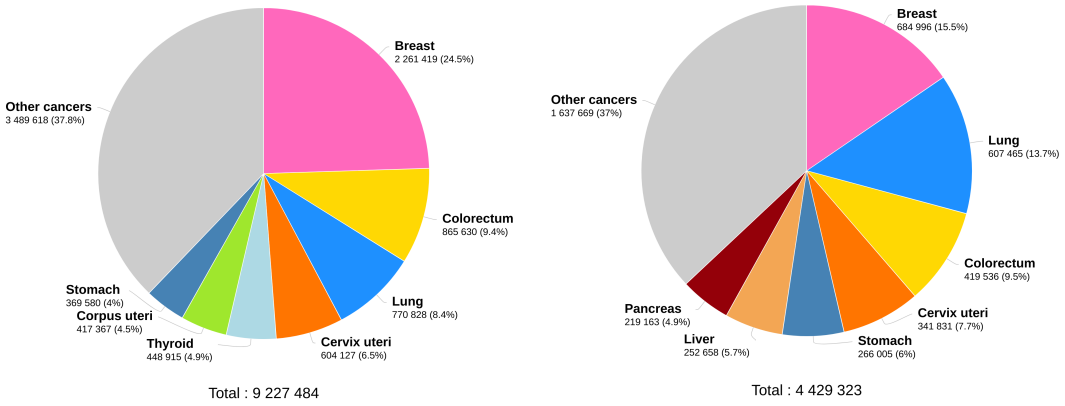

Figura 1.4: Cantidad estimada de nuevos casos (izquierda) y de muertes (derecha) por cáncer a nivel mundial en el año 2020. Fuente: Globocan 2020

Por lo tanto, resulta de gran interés mundial poder reducir la mortalidad del cáncer de mama siendo

el que mayor incidencia y muerte presenta en las mujeres. Es por esto que es importante detectar de forma temprana la presencia y el tipo de tumor, para poder así iniciar el tratamiento a tiempo. Si bien, en los últimos años, según la Sociedad de Cáncer Americana (*American Cancer Society*) se ha reducido de forma considerable la tasa de mortalidad por cáncer de mama en Estados Unidos (Ver Anexo 7.1), es imprescindible detectarlo a tiempo para evitar la propagación del cáncer hacia otros tejidos. Además, según la Sociedad de Cáncer Americana, si el paciente es diagnosticado y tratado a tiempo, tiene hasta un 90 % de probabilidad de sobrevivir en los siguientes 5 años si el cáncer es Estadío I. En un estudio realizado en Argentina [4], de las 636 pacientes diagnosticadas con cáncer de mama entre 1998 y 2012, la tasa de supervivencia relativa a 5 años fue de: 99,2 % al año, de 93,6 % a los 3 años y de 88,0 % a los 5 años. Por estadíos, la tasa de supervivencia a los 5 años fue de 97,75 % en el Estadío I, de 91,2 % en el Estadío II, de 74,4 % en el Estadío III y de 76,2 % en el Estadío IV.

1.1. Estado del arte

Durante un examen de detección mamográfico, se capturan imágenes de ambas mamas incidiendo con rayos X de baja energía al paciente. Estas imágenes son inspeccionadas en busca de lesiones malignas por un radiólogo y en general, los casos sospechosos se vuelven a analizar en una evaluación diagnóstica adicional. Múltiples estudios han demostrado que entre el 20 y el 30 % de los tipos de cáncer diagnosticados se pueden encontrar retrospectivamente en el examen de detección negativo anterior [5], [6]. Se descubrió que la lectura doble mejora el rendimiento de la evaluación mamográfica y se ha implementado en muchos países [7]. Varias lecturas pueden mejorar aún más el rendimiento diagnóstico hasta más de 10 lectores, lo que demuestra que hay margen de mejora en la evaluación de mamografías.

La mamografía es actualmente la herramienta más eficaz para la detección temprana del cáncer de mama; sin embargo, tiene algunas restricciones. La alta **densidad mamaria** es un factor que dificulta el diagnóstico de cáncer de mama [8] debido a que el contraste entre el cáncer y el fondo en la imagen es muy bajo, lo que puede afectar el resultado del diagnóstico [9]. En el examen mamográfico, las lesiones no cancerosas pueden malinterpretarse como cáncer (valor falso positivo), mientras que el cáncer puede pasarse por alto (valor falso negativo). Como resultado, los médicos no detectan entre el 10 % y el 30 % del cáncer de mama [10].

Si en la mamografía se observa que el hallazgo es benigno, en general no se realizan más pruebas médicas. Sin embargo, si el hallazgo es sospechoso o probablemente maligno, se realizan exámenes médicos adicionales. Por ejemplo, se lleva a cabo una biopsia en la que se extrae una pequeña cantidad de tejido del paciente, para que pueda examinarse en un laboratorio. Por lo tanto, el valor falso positivo (el hallazgo se clasifica como cáncer cuando en realidad no lo es) indica el porcentaje de lesiones sospechosas o cancerosas detectadas en la mamografía y posteriormente, sometidas a biopsias. La tasa de fallas en la mamografía ha aumentado en mamas densas donde la probabilidad de cáncer es de cuatro a seis veces mayor que en mamas no densas [11].

Se han propuesto varias soluciones para mejorar la especificidad y sensibilidad de la lectura de la mamografía, así como para disminuir los procedimientos de biopsias innecesarios. **La doble lectura** es una de las soluciones que puede contribuir significativamente a lograr una alta sensibilidad y especificidad [12]. Sin embargo, esto se traduce en tiempos y costos adicionales para el centro de salud.

1.1.1. Sistemas de detección asistida por computadora

Los **sistemas de detección asistida por computadora** (*Computer-aided detection systems*, CAD por sus siglas en inglés) se han desarrollado desde la década de 1960 para superar estas restricciones y en los últimos años, se han estudiado en muchas modalidades de imágenes para la detección del cáncer de mama. En 1998, la Administración de Alimentos y Medicamentos de los Estados Unidos (FDA por sus siglas en inglés) aprobó el primer sistema CAD para mamografías que se extendió rápidamente.

Actualmente, los sistemas CAD ayudan a los radiólogos a encontrar y discriminar entre los tejidos normales y anormales. Estos procedimientos se realizan solo como un lector doble ya que a las decisiones aún las toma el médico. La base de los sistemas CAD generalmente consta de cuatro etapas: preprocesamiento, segmentación, extracción de características y clasificación de las imágenes.

Los beneficios del uso de CAD son controvertidos. Inicialmente, varios estudios han mostrado resultados prometedores con su uso [6], [13], [14]. Un gran ensayo clínico en el Reino Unido ha demostrado que la lectura única con asistencia CAD tiene un rendimiento similar a la lectura doble [15]. Sin embargo, en la última década múltiples estudios concluyeron que las tecnologías CAD actualmente utilizadas no mejoran el desempeño de los radiólogos en la práctica diaria en Estados

Unidos [16], [17], [18].

Por lo tanto, todavía queda un gran margen de mejora para estos sistemas involucrados en la detección del cáncer de mama. Desde el 2010 se ha producido un desarrollo significativo de los algoritmos de reconocimiento de imágenes por computadora. Se han desarrollado nuevas formas de inteligencia artificial (IA) (Ver Anexo 7.4 y 7.5), que superan significativamente a las tecnologías anteriores utilizando arquitecturas de red novedosas y unidades de procesamiento gráfico (GPU, proporcionadas por la tarjeta de video de una computadora) en lugar de unidades de procesamiento (CPU, proporcionadas por el procesador de una computadora). Esto también ha resultado directamente en un rápido aumento en el desarrollo de aplicaciones CAD basadas en IA para tareas médicas.

Un sistema de soporte a la toma de decisiones (*Clinical decision support system*, CDSS por sus siglas en inglés) está destinado a mejorar la prestación de atención médica al ayudar a los médicos con conocimiento clínico específico, información del paciente y otra información de salud. Un CDSS tradicional está compuesto por un software diseñado para ser una ayuda directa para la toma de decisiones clínicas, en el que las características de un paciente individual se comparan con una base de conocimiento clínico y luego se presentan evaluaciones o recomendaciones específicas del paciente al médico para que tome una decisión. Los CDSS hoy en día se utilizan principalmente en el punto de atención, para que el médico pueda combinar sus conocimientos con la información o sugerencias proporcionadas por el sistema. En la actualidad, a menudo se utilizan en aplicaciones web o se integran en las historias clínicas electrónicas.

1.1.2. Redes Neuronales Convolucionales

Recientemente, muchos investigadores trabajaron en la detección del cáncer de mama en mamografías utilizando aprendizaje profundo. En el campo de la visión por computadora, desde 2012, **las redes neuronales convolucionales profundas (CNN, por sus siglas en inglés)** han superado significativamente a los métodos tradicionales [19]. Las CNNs profundas han alcanzado, o incluso superado, el rendimiento humano en la clasificación de imágenes y detección de objetos [20]. Estos modelos tienen un enorme potencial en el análisis de imágenes médicas.

Existen diversas bases de datos de mamografías disponibles públicamente, tales como: *Curated*

Breast Imaging Subset of DDSM (CBIS-DDSM) [1], *Breast Cancer Digital Repository* (BCDR) [2], *INbreast* [3], *Image Retrieval in Medical Applications* (IRMA) [4], *Digital Database for Screening Mammography* (DDSM) [5] y *the Mammographic Image Analysis Society* (MIAS) [6]. Otras bases de datos utilizadas en la literatura son privadas y están restringidas a organizaciones individuales. Las bases de datos públicas presentan una amplia diversidad de casos de pacientes y una variedad de casos normales, benignos y malignos. A continuación se observa una comparación entre las bases de datos públicas existentes, especificando: las vistas (CC, MLO o ambas), el tipo de mamografía (digitales (FFDM) o tradicionales (FSM)), el formato de la imagen, la cantidad de mamografías, el año y el autor de cada base de datos.

Base de datos	Vistas	Tipo	Formato	Cantidad de imágenes	Año	Autor
CBIS-DDSM	Ambas	FFDM	DICOM	3568	2017	Lee et al. [22]
BCDR-F01	Ambas	FSM	TIF	362	2013	Lopez et al. [23]
BCDR-F02	Ambas	FSM	TIF	516	2013	Lopez et al. [23]
BCDR-F03	Ambas	FSM	TIF	736	2013	Lopez et al. [23]
BCDR-D01	Ambas	FFDM	DICOM	143	2013	Lopez et al. [23]
BCDR-D02	Ambas	FFDM	DICOM	456	2013	Lopez et al. [23]
BCDR-DN01	Ambas	FFDM	DICOM	200	2013	Lopez et al. [23]
INBreast	Ambas	FFDM	DICOM	336	2012	Moreira et al. [24]
IRMA	Ambas	Ambos	PNG	3676	2008	Oliveira et al. [25]
DDSM	Ambas	FSM	LJPEG	2620	2006	Rose et al. [26]
MIAS	MLO	FSM	PGM	322	1994	Suckling et al. [27]

Cuadro 1.1: Comparación entre bases de datos públicas ampliamente utilizadas en la literatura con respecto a las vistas (CC, MLO), el tipo de mamografía (digitales (*Full-Field Digital Mammography*, FFDM) o con pantalla de película (*Traditional filmscreen mammography*, FSM)), el formato de las imágenes, la cantidad de imágenes, el año y el autor de la publicación

[1] https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM
[2] https://bcdr.ceta-ciemat.es/
[3] http://medicalresearch.inescporto.pt/breastresearch/index.php/Get$_{I}Nbreast_{D}atabase$
[4] http://www.irma-project.org/index$_{e}n.php$
[5] http://www.eng.usf.edu/cvprg/
[6] https://www.repository.cam.ac.uk/handle/1810/250394

A continuación, se explica brevemente cada una de las bases de datos públicas, listadas anteriormente. **MIAS** es una base de datos antigua que contiene un número limitado de imágenes que son de baja resolución y presentan ruido. A pesar de estos inconvenientes, se ha utilizado ampliamente en la literatura hasta ahora. **DDSM** es un repositorio muy grande que se utiliza en muchos estudios. Las imágenes de DDSM se guardan en archivos de compresión no estándar que requieren el uso de códigos de descompresión. Además, las anotaciones de la Región de Interés (ROI) para las anomalías en las imágenes, indican la posición general de las lesiones, sin una segmentación precisa de las mismas.

El proyecto **IRMA** es una combinación de varias bases de datos de diferentes tamaños y resoluciones. Las anotaciones de las ROI para estas bases de datos son más precisas. La base de datos **INbreast** está ganando más atención en la actualidad. Sus ventajas son la alta resolución y la segmentación precisa de las lesiones. Sin embargo, sus inconvenientes son su pequeño tamaño y las variaciones limitadas de las formas de las masas. **BCDR** es una base de datos prometedora, pero aún se encuentra en su fase de desarrollo. BCDR se subdivide en dos repositorios diferentes: (1) un repositorio basado en mamografías de películas (BCDR-FM) y (2) un repositorio basado en mamografías digitales de campo completo (BCDR-DM). La ventaja es que presentan la posición precisa de las lesiones.

La base de datos **CBIS-DDSM** es una versión actualizada y estandarizada de la base de datos DDSM. CBIS-DDSM es un subconjunto de imágenes del conjunto de datos DDSM original que fue seleccionado y curado por radiólogos expertos. Estas imágenes se han descomprimido y convertido a formato DICOM estandarizado. El conjunto de datos contiene dos vistas de cada mama (es decir, CC y MLO), tipo de anomalía y diagnóstico patológico.

Varios estudios han intentado aplicar *Deep Learning* para analizar mamografías. La siguiente tabla muestra un resumen del estado del arte de los métodos utilizados en la detección del cáncer de mama mediante CNNs.

Autor	Año	Base de datos y cantidad de imágenes	Método	Clasificación	Exactitud	AUC
Huynh et al [28]	2016	Privada (607)	AlexNet y GoogLeNet	TB y TM	-	0,86
Aboutalib et al [29]	2018	Privada (14860)	AlextNet	TB y TM	-	0,78
Hua Li et al. [30]	2019	Privada (2042)	AlexNet, VGGNet y GoogLeNet	TB y TM	0,928	0,804
Cai et al. [31]	2019	Privada (990)	*Scratch based*	TB y TM	0,877	0,934
Agarwal y Carson [32]	2015	DDSM (2620)	*Scratch based*	M y C / TB y TM	0,870 y 0,690	-
Levy y Jai [33]	2016	DDSM (1820)	AlexNet y GoogLeNet	Masa en TM y TB	0,89	-
Pengcheng Xi et. al [34]	2018	CBIS-DDSM (3071)	AlexNet, VGG, GoogLeNet y ResNet	M y C	0,925	-
Khan et al. [35]	2019	CBIS-DDSM y MIAS (3890)	VGG, Inception V3 y Resnet50	M y C / TB y TM	0,896 y 0,754	0,896 y 0,746
Ragab et al. [36]	2019	CBIS-DDSM y DDSM (9368)	AlextNet	TB y TM	0,736	-

Cuadro 1.2: Métodos utilizados en la detección de diferentes lesiones en las mamografías mediante CNNs. Las primeras cuatro publicaciones utilizan bases de datos privadas y las últimas cinco usan bases públicas. Las siglas M, C, TB y TM hacen referencia a masas, calcificaciones, tumor benigno y tumor maligno, respectivamente

En general, el inconveniente de todas las redes neuronales y el aprendizaje profundo es la necesidad de una gran cantidad de muestras de entrenamiento etiquetadas para aprender los patrones en las imágenes para clasificarlas correctamente.

Desafortunadamente, en las imágenes médicas, la cantidad de datos de entrenamiento etiquetados disponibles es limitada. El entrenamiento de un modelo profundo mediante este conjunto da como resultado un sobreajuste ya que el modelo tiende a memorizar el conjunto de entrenamiento. Para

superar el desafío de la insuficiencia de datos, muchos grupos de investigación han ideado diferentes estrategias:

- usando parches 2D en lugar de usar la imagen completa como entrada, lo que también reduce los parámetros del modelo y el sobreajuste. [37]
- introduciendo el aumento de datos utilizando algunas transformaciones (traslación, rotación, etc) y entrenando a la red con los datos aumentados [38].
- transfiriendo el aprendizaje usando pesos previamente entrenados y simplemente reemplazando las últimas capas para la nueva clasificación. Este proceso es conocido como transferencia de aprendizaje (*Transfer Learning*) [39].

Todas éstas alternativas fueron probadas y analizadas en el presente trabajo, tal como se verá en las siguientes secciones.

1.2. Objetivos

En este trabajo se propone diseñar y comparar distintas arquitecturas de redes neuronales convolucionales que puedan clasificar lesiones en las mamografías. El objetivo del trabajo es generar modelos robustos que clasifiquen a las lesiones en la mama tanto **por su tipo** (calcificación o masa) como **por su severidad** (benigna o maligna). Los primeros signos de cáncer de mama no palpable son las calcificaciones, que suelen estar asociadas con el carcinoma ductal in situ (CDIS) [40], pero también pueden estar presentes en tipos de cáncer invasivos (entre el 12,7 y el 41,2 % de las mujeres poseen calcificaciones como único signo de cáncer [41]) (Ver Anexo 7.3).

Por tal motivo, primero se entrenó una CNN que clasifica entre masas o calcificaciones. Luego, se diseñó una segunda CNN cuyo objetivo es el de clasificar a las lesiones en benignas o malignas. Por último, resultó interesante armar un modelo que pueda clasificar de forma categórica en: masa benigna, masa maligna, calcificación benigna o calcificación maligna. Para llevar a cabo este desafío, se tomaron dos caminos diferentes. Por un lado, se unieron las dos CNN previamente entrenadas por lo que la imagen primero es clasificada según el tipo de lesión y luego según su severidad. Por el otro, se diseñó una tercera CNN que realiza la clasificación categórica.

Capítulo 2

Marco Teórico

Existen diversas maneras de capturar imágenes de las mamas, entre las que se encuentran: la tomografía, la resonancia magnética y el ultrasonido. La más ampliamente usada para la detección de cáncer de mama es la mamografía y es la que se va a utilizar en el presente trabajo.

2.1. Mamografía

La mamografía es una imagen de la mama (Ver Anexo 7.2) obtenida con la incidencia de rayos x (radiación ionizante) mediante un aparato llamado mamógrafo. La dificultad en este tipo de imagen es que se detectan cambios sutiles en un material con bajo rango dinámico, ya que los tejidos son radiológicamente similares. Como se pretende identificar en la imagen zonas muy pequeñas que son las calcificaciones, el sistema requiere tener muy alta la resolución espacial. Desde el punto de vista tecnológico, existen dos tipos de mamógrafos según la adquisición de la imagen:

- **Sistemas tradicionales de mamografía con pantalla de película:** (*Traditional film-screen mammography*, FSM, por sus siglas en inglés): obtención de la imagen sobre una película que luego se revela.
- **Sistemas de mamografía digital de campo completo:** (*Full-Field Digital Mammography*, FFDM por sus siglas en inglés) son iguales a los sistemas tradicionales excepto que poseen detectores electrónicos que capturan y facilitan la visualización de las señales de rayos X en la computadora. La matriz de detectores digitales responde a la exposición de rayos X y luego envía una señal para cada ubicación del detector a una computadora, donde se digitaliza, procesa y almacena como una señal. La mamografía digital se divide en:

Mamografía digital directa (DR): posee detectores de radiación que convierten en un solo paso la información en una señal.

Mamografía digital indirecta (CR): se utiliza un equipo analógico que en lugar de la placa radiológica contiene una placa de fósforo fotoestimulable que almacena la información recibida y la mantiene en forma latente hasta que se procesa obteniendo una imagen en formato digital.

La imagen a continuación muestra un esquema de las partes de un **mamógrafo**. Se puede observar el tubo de rayos X (con el ánodo, cátodo y haz de electrones), el filtro, la placa de compresión de la mama y el receptor de la imagen. Los tubos de rayos X en el mamógrafo trabajan a bajos voltajes (25 a 30 kV) comparados con los utilizados en radiografías (50 a 120 kV).

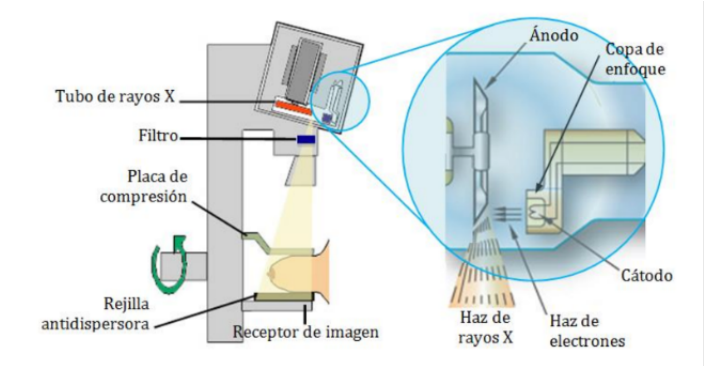

Figura 2.1: Esquema de un mamógrafo

Existen cuatro vistas o proyecciones estandarizadas de la mama en cada mamografía. Por un lado, se encuentran las vistas MLO (*Medio-lateral-oblique*) derecha (RMLO) e izquierda (LMLO). Por el otro, las vistas CC (*Craneocaudal*) derecha (RCC) e izquierda (LCC) tal como se ve a continuación:

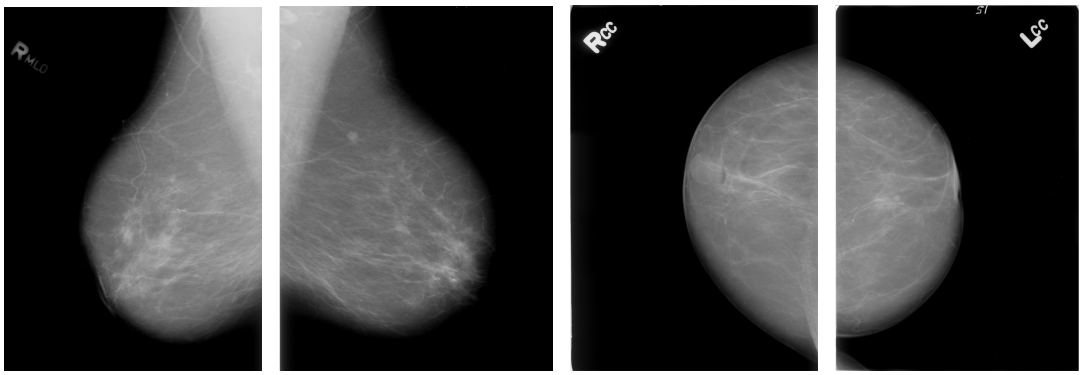

Figura 2.2: Proyecciones estandarizadas de la mamografía. De izquierda a derecha: RMLO, LMLO, RCC y LCC. Fuente: CBIS-DDSM

El nombre de cada vista surge de la localización y dirección de los rayos x incidentes. En el caso de la cráneo-caudal (CC), la dirección del rayo es desde la parte craneal (superior) de la mama hacia la parte caudal (inferior). Por otro lado, en la vista medio-lateral-oblicua (MLO), la mama se comprime en una dirección de 45 grados y el haz de luz viaja desde la parte medial de la mama hacia la lateral. En esta vista, se observa el músculo pectoral.

El sistema de datos e informes de imágenes de mama (*Breast imaging reporting and data system*, BI-RADS por sus siglas en inglés) fue diseñado por el Colegio Americano de Radiología para estandarizar los informes y proporcionar claridad en la interpretación de los estudios de imágenes de mama. BI-RADS creó la siguiente clasificación de la densidad mamaria:

- BI-RADS A: En su mayoría, es tejido graso. En la mamografía, casi no se aprecia tejido glandular (áreas blancas) en donde se pueda esconder un tumor.
- BI-RADS B: Contiene áreas de tejido fibroglandular. Se observan pequeñas áreas blancas que forman parte del tejido glandular.
- BI-RADS C: Contiene tejido denso y heterogéneo. Hay mucho tejido glandular cerca del pezón.
- BI-RADS D: Contiene tejido extremadamente denso. Se ve mucho tejido glandular y se dificulta la detección de un tumor.

Dicha clasificación se observa en la imagen 2.3. En general en la mamografía, tanto el tejido glandular como los tumores en la mama, se ven de color blanco. Por esta razón, en la densidad de

tipo C y D, se puede confundir un tumor en la mama con la densidad del tejido glandular. Por otro lado, el tejido graso se ve de color gris.

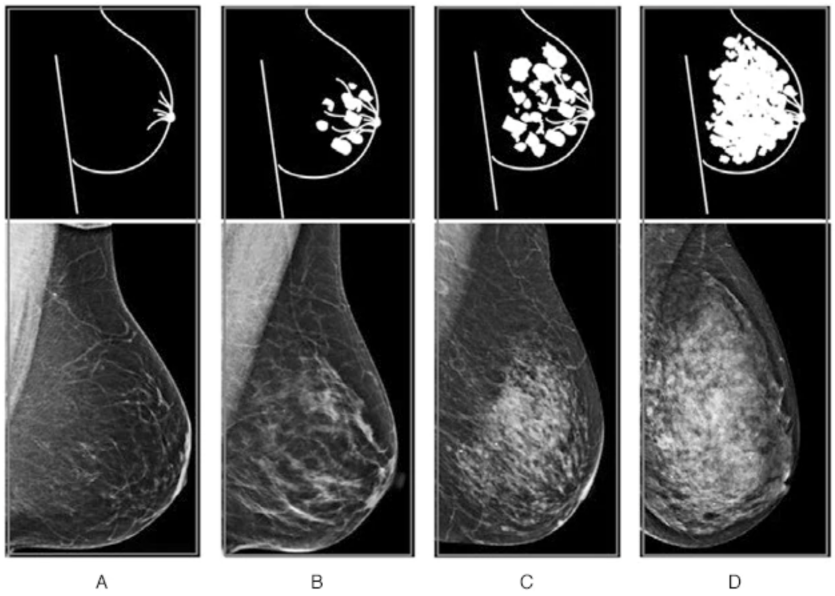

Figura 2.3: Densidad mamaria según BI-RADS, de izquierda a derecha: A, B, C y D. Fuente: *Mayo Foundation for Medical Education and Research*

Cuando se interpreta y analiza una mamografía, se buscan masas y calcificaciones ya que son los dos signos más importantes de malignidad.

2.1.1. Calcificaciones

Las calcificaciones mamarias son pequeñas partículas de calcio que se encuentran dispersas en la mama. En las mamografías, las calcificaciones benignas suelen ser más grandes, gruesas, redondas y con márgenes más lisos que las malignas. Muchas veces es necesario aumentar las vistas de la mamografía (MLO y CC) para poder analizar la morfología y estructura de las calcificaciones.

Para clasificarlas en benignas o malignas, se analizan diferentes propiedades como tamaño, forma, patrón de distribución y densidad. Según BI-RADS se pueden clasificar según: su morfología típicamente benigna (piel, vascular, redonda, con borde, leche de calcio, sutura, etc), su morfología sospechosa (amorfa, gruesa heterogénea, polimórfica fina y lineal fina o lineal ramificada) y su distribución (difusa, regional, agrupada, lineal y segmentaria).

Es importante diferenciar las calcificaciones de origen benigno de las sospechosas, ya que el 55 % del cáncer no palpable se diagnostican por la presencia de calcificaciones [42], y porque las mismas son la principal forma de manifestación del carcinoma ductal in situ (CDIS) [43] (Ver Anexo 7.3.1). Además, como se mencionó anteriormente, entre el 12,7 y el 41,2 % de las mujeres poseen calcificaciones como único signo de cáncer [41].

En general, las calcificaciones benignas y malignas están compuestas por diferentes elementos químicos que las diferencian. Estos no pueden ser determinados a través de la mamografía, siendo necesario realizar estudios químicos. Por un lado, las calcificaciones benignas están compuestas principalmente por oxalato de calcio, mientras que las calcificaciones malignas están compuestas principalmente por fosfato de calcio [44].

Para la correcta clasificación de la calcificación, se debe investigar tanto su morfología como distribución. Los descriptores de la distribución de las calcificaciones en la mama según BI-RADS, especifican la disposición de las mismas en el interior de la mama en relación con la probabilidad de malignidad.

Los distintos tipos de distribuciones según BI-RADS se observan a continuación:

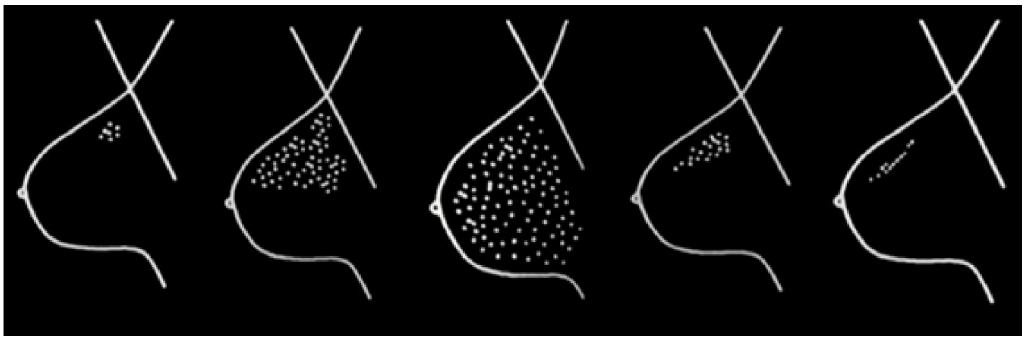

Figura 2.4: Esquema de los descriptores de distribución de calcificaciones según BI-RADS. De izquierda a derecha: Agrupada, Regional, Difusa, Segmentaria y Lineal. Fuente: Arancibia et al.5.

- **Agrupada:** este término se utiliza cuando se encuentran algunas calcificaciones en un área pequeña de tejido. El límite inferior de este descriptor son 5 calcificaciones en 1 cm o cuando hay un patrón definible. El límite superior es cuando hay más microcalcificaciones presentes dentro de los 2 cm. Requieren mayor evaluación con proyecciones magnificadas y deben agruparse en ambas proyecciones (MLO y CC) para considerarlas como tales, ya que, si

solo se agrupan en una proyección, puede corresponder a superposición de calcificaciones en diferentes posiciones. Se consideran benignos o sospechosos según la morfología de cada grupo.

- **Regional:** este patrón describe calcificaciones en un área extensa, mayor de 2 cm en su dimensión más grande. Debido a que pueden cubrir más de un cuadrante, su riesgo de malignidad es bajo, sin embargo, se debe considerar la morfología para establecer el grado de sospecha. Una probabilidad de malignidad se describe en alrededor del 26 % [45].

- **Difusa:** o también llamadas "dispersas", son calcificaciones distribuidas aleatoriamente dentro de la mama. Las calificaciones puntiformes y amorfas en esta distribución suelen ser benignas, sobre todo si son bilaterales.

- **Segmentaria:** este patrón de distribución sugiere el depósito de calcio en los conductos y sus ramas, siguiendo la forma anatómica de un lóbulo mamario, es decir, de forma triangular con la punta dirigida hacia el pezón (Ver Anexo 7.2). Si bien puede ocurrir en patologías benignas, como calcificaciones secretoras, su presentación puede deberse a un cáncer extenso o multifocal. La probabilidad de malignidad se describe en alrededor del 62 % [46], [47].

- **Lineal:** las calcificaciones están dispuestas en una trayectoria lineal que puede ramificarse, lo que sugiere depósitos de calcio dentro de un conducto. Se describe una probabilidad de malignidad de alrededor del 60 % [46], [47]. Cabe señalar que determinadas calcificaciones como las vasculares o lineales gruesas pueden presentar esta distribución, sin embargo, tienen una morfología característicamente benigna.

En la imagen a continuación, se puede observar un resumen de los diferentes tipos de distribuciones; de izquierda a derecha: agrupado, regional, difuso, segmentado y lineal.

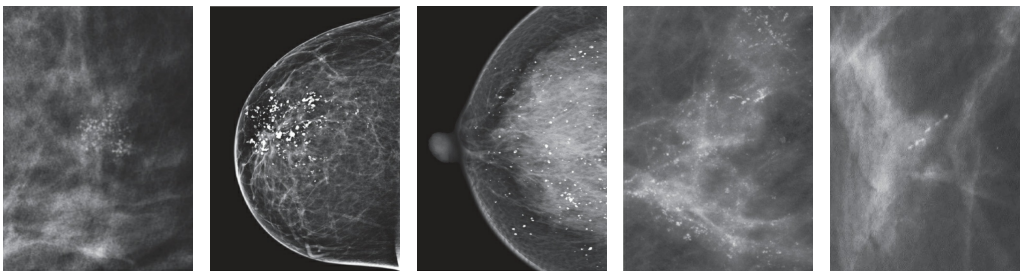

Figura 2.5: Distintas distribuciones observadas en las calcificaciones de las mamografías. De izquierda a derecha: Distribución Agrupada, Regional, Difusa, Segmentaria y Lineal. Fuente: ACR BI-RADS Atlas: *Breast Imaging Reporting and Data System 2013*

2.1.2. Masas

Por otro lado, la otra lesión existente en la mama es la masa que es tridimensional y se ve en dos proyecciones mamográficas diferentes. Tiene bordes total o parcialmente convexos hacia el exterior y parece más denso en el centro que en la periferia. Si una masa potencial se ve solo en una sola proyección, debe llamarse asimetría hasta que se confirme su tridimensionalidad. Según el sistema BI-RADS una masa se caracteriza por:

- **La forma:** redonda, ovalada, lobulada o irregular;
- **El contorno:** circunscrito, microlobulado, enmascarado, indistinto o espiculado.
- **La densidad** con respecto al tejido fibroglandular normal (densidad alta, media o baja) o que contiene grasa.
- **La evolución en el tiempo** cuando se dispone de mamografías anteriores.

El margen es el borde de la lesión. Los descriptores de margen, como los de forma, son predictores importantes de si una masa es benigna o maligna; es decir, el contorno es el criterio morfológico más discriminatorio entre masas benignas y malignas. A continuación se observan los diferentes tipos de bordes que pueden tener las masas según BI-RADS:

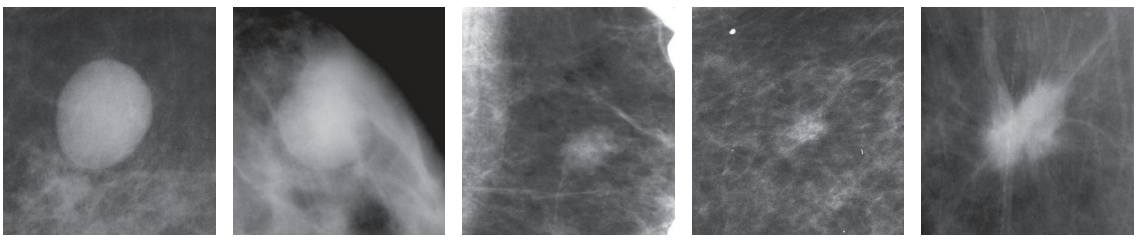

Figura 2.6: Distintos márgenes observados en las masas de las mamografías. De izquierda a derecha: Margen circunscrito, oscurecido, microlobulado, indistinto y espiculado. Fuente: ACR BI-RADS Atlas: *Breast Imaging Reporting and Data System 2013*

El margen de una masa puede ser:

- **Circunscrito**: el contorno está claramente demarcado con una transición abrupta entre la lesión y el tejido circundante. Para la mamografía, si parte del margen está oscurecido, al menos el 75 % del margen debe estar bien definido para que una masa califique como circunscrita. El 25 % restante puede, como máximo, estar enmascarado por la glándula adyacente. Una masa para la cual cualquier porción del margen es indistinta, microbulada o espiculada debe clasificarse sobre la base de este último (el componente más sospechoso).

- **Oscurecido**: es aquel que está oculto por tejido fibroglandular superpuesto o adyacente. Esto se utiliza principalmente cuando parte del margen de la masa está circunscrito, pero el resto (mayor al 25 %) está oculto.
- **Microlobulado**: el margen se caracteriza por ondulaciones de ciclo corto. Para la mamografía y el uso de este descriptor generalmente implica un hallazgo sospechoso.
- **Indistinto**: no hay una demarcación clara de todo el margen o de cualquier parte del margen, del tejido circundante. El uso de este descriptor generalmente implica un hallazgo sospechoso
- **Espiculado**: también llamadas estelares, corresponden a opacidades formadas por un centro denso del que surgen múltiples prolongaciones radiales lineales llamadas espículas. Los aspectos de la mamografía varían y dependen del grosor, la longitud y la distribución de las espículas alrededor de la masa. El uso de este descriptor generalmente implica un hallazgo sospechoso.

En la mamografía, la existencia de un contorno **no circunscrito**, ya sea microlobulado, enmascarado o indistinto, justifica una biopsia para examen histológico. El riesgo de malignidad difiere según la morfología del contorno: el contorno microlobulado se asocia con un 17 % de riesgo de cáncer, el contorno enmascarado con un 33 % de riesgo de cáncer y el contorno indistinto con un 44 % de riesgo de cáncer [48]. La detección de masas es más difícil en comparación con la calcificación debido a la similitud y ambigüedad de sus características con el tejido normal. Las masas se observan generalmente en las regiones densas de la mama con límites más suaves comparados con los de las calcificaciones.

2.2. Redes Neuronales Convolucionales

La red neuronal convolucional (*Convolutional Neural Network*, CNN por sus siglas en inglés), es una red de aprendizaje profundo supervisada que consta de muchas capas convolucionales apiladas. Por lo general, una CNN está compuesta por capas convolucionales, de agrupación y completamente conectadas que se alinean una encima de la otra para formar una red profunda.

Las redes neuronales (Ver Anexo 7.6) completamente conectadas (*Fully connected neural networks*) normalmente no funcionan bien con imágenes. Esto se debe a que si cada píxel es una entrada, a medida que agregamos más capas, la cantidad de parámetros aumenta exponencialmente. Por ejemplo, si se tiene una imagen de tamaño 32 x 32 x 3, una sola neurona completamente conectada

en la primera capa oculta tendría 3.072 pesos. Por otro lado, para una imagen de tamaño 200 x 200 x 3, una neurona completamente conectada en la primera capa oculta tendría 120.000 pesos. El otro desafío es que una cantidad de parámetros tan grande puede conducir rápidamente a un ajuste excesivo (*overfitting*). Una solución es usar imágenes más pequeñas, pero claramente se perderá información. Lo importante es que lo que distingue a una imagen de otra es su estructura espacial, aspecto que se usa en las CNNs. A continuación, se van a analizar las diferentes capas presentes en una CNN.

2.2.1. Capa de Convolución

El objetivo de una capa convolucional es el filtrado que implica aplicar un kernel por una imagen para encontrar patrones. Por ejemplo, consideremos una imagen representada por una matriz de tamaño 5x5 y un filtro de 3x3 tal como se observa a continuación.

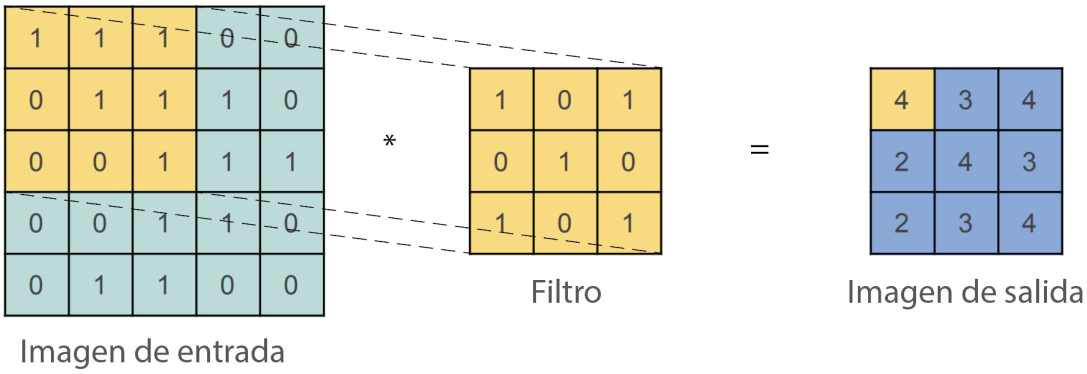

Figura 2.7: Esquema de la Convolución. De izquierda a derecha, se encuentran: la imagen de entrada, el kernel y la imagen de salida (mapa de características)

El procedimiento se basa en mover el filtro por sobre toda la imagen y se va multiplicando el número de la imagen por la del filtro para luego hacer la sumatoria de todos los valores obtenidos. Por cada posición que toma el kernel sobre la imagen, se obtiene un valor que se ve reflejado en la imagen de salida (mapa de características). En la figura 2.7, se observa el proceso de convolución. Sea I la imagen de entrada, K el kernel y S(i,j) el mapa de características, entonces:

$$S(i,j) = (I * K)(i,j) = \sum_{u=0}^{m} \sum_{l=0}^{n} K(u,l) \cdot I(i-u, j-l) \tag{2.1}$$

Por ejemplo, para obtener el número 4 resaltado en color amarillo en la imagen de salida, se mul-

tiplica cada número del filtro por cada número de la imagen de entrada y luego se suman todos los valores. Por lo tanto, el cálculo que se hace es: $1 \cdot 1 + 1 \cdot 0 + 1 \cdot 1 + 0 \cdot 0 + 1 \cdot 1 + 1 \cdot 0 + 0 \cdot 1 + 0 \cdot 0 + 1 \cdot 1 = 4$; es decir, se toma el 1 de la esquina superior izquierda del filtro de 3x3 y se lo multiplica por el 1 de la esquina superior izquierda de la imagen. Luego se multiplica el 0 en el filtro con la ubicación correspondiente en la imagen que es un 1, y así sucesivamente. Se hace para las 9 entradas del kernel y se suman todos los cálculos del producto como se observa a continuación.

Por lo tanto, las capas convolucionales aplican sistemáticamente filtros a las imágenes de entrada para crear mapas de características (*feature maps*) que resuman la presencia de las características de la imagen entrada. Cuando se entrena una imagen, los pesos cambian y por lo tanto cuando es el momento de evaluar a una imagen, estos pesos devuelven valores altos si está viendo un patrón que ha visto antes. Las combinaciones de pesos altos de varios filtros permiten a la red predecir el contenido de una imagen.

2.2.2. Zero Padding

Uno de los grandes inconvenientes de la convolución es que se pierden datos de la imagen original tal como se observa en la figura 2.8.

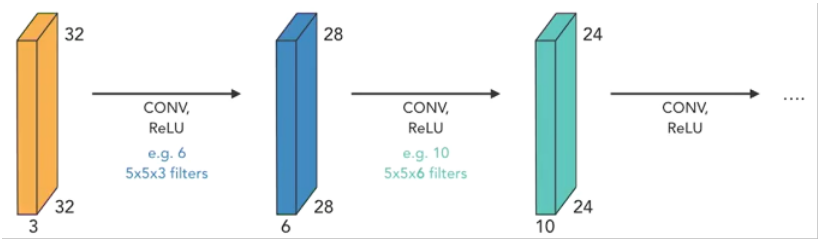

Figura 2.8: Ejemplo de reducción de la dimensión de una imagen luego de realizar sucesivas convoluciones

En la figura se observa que la imagen inicial es a color ya que posee 3 canales (RGB, un canal para el color rojo, otro para el verde y otro para el azul). Por lo tanto, su tamaño es 32 x 32 x 3 en donde la altura y el ancho de la imagen son iguales a 32 y la cantidad de canales es 3. Para calcular el tamaño de la imagen después de la convolución se hace el siguiente cálculo:

$$n' = n - f + 1 \tag{2.2}$$

en donde n' es el tamaño de la salida; n es el tamaño de la imagen de entrada y f es el tamaño

del filtro. En la figura 2.8, se observa que a la imagen de entrada, se le aplica una primera capa de convolución con 6 filtros de 5 x 5 x 3; es decir, se le aplica el filtro de 5 x 5 a cada canal. Como la imagen posee el tamaño de 32 x 32 x 3 y se le aplican 6 filtros de 5 x 5 x 3, entonces el tamaño de la salida es 28 x 28 x 6 ya que la cantidad de canales que posee la salida, es igual a la cantidad de filtros aplicados; es decir, en este caso son 6 filtros. A continuación se observa el cálculo realizado aplicando la ecuación 2.2:

$$n' = 32 - 5 + 1 = 28 \tag{2.3}$$

Si luego se le vuelve a hacer otra convolución con 10 filtros de tamaño 5 x 5 x 6 entonces el tamaño de la salida resulta de 24 x 24 x 10 ya que aplicando la ecuación 2.2 queda:

$$n' = 28 - 5 + 1 = 24 \tag{2.4}$$

Si queremos preservar el tamaño espacial del volumen de entrada, de modo que el ancho y la altura de entrada y salida sean iguales, entonces el *zero padding* es una técnica útil. Se define a F como el tamaño del filtro, S como el paso (*stride*), n como el tamaño de la imagen y P como la cantidad de relleno que se necesita (*padding*), entonces el tamaño de salida de la imagen viene dado por la siguiente ecuación:

$$n' = \frac{n - F + 2P}{S} + 1 \tag{2.5}$$

El *zero padding* es un hiperparámetro que nos permite controlar el tamaño espacial de la imagen de salida. Por ejemplo, se comienza el proceso con una imagen de 5 x 5 (N=5), el tamaño del filtro es 3 x 3, (F=3), el stride es 1 (S=1) y la cantidad de relleno también es 1 (P=1). Entonces el tamaño de salida permanece como 5 x 5. Por lo tanto, se puede observar que al usar *zero padding*, se ha conservado el tamaño de la imagen original. Por último, a continuación se observa la figura 2.7 con *Zero Padding*.

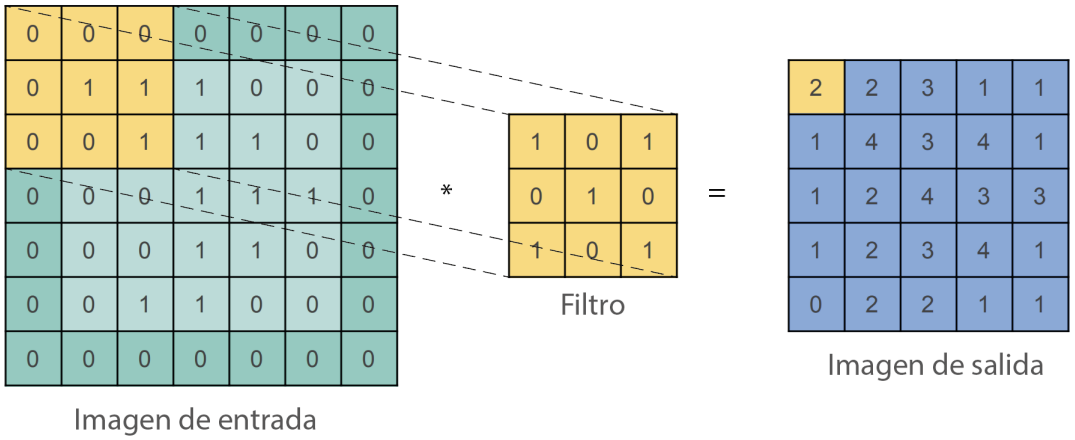

Figura 2.9: Esquema de la Convolución con *Zero Padding*. De izquierda a derecha, se encuentran: la imagen de entrada, el kernel y la imagen de salida

2.2.3. Pooling Layer

La capa que se utiliza luego de la convolucional es principalmente la capa de agrupación (*Pooling Layer*), la cual divide a la imagen de entrada en un conjunto de rectángulos no superpuestos y, para cada subregión, genera un valor.

Una limitación de los mapa de características (*feature maps*) generados, es que registran la posición precisa de las características en la imagen de entrada. Esto significa que pequeños movimientos en la posición de la característica en la imagen (recortes, rotaciones, etc) darán como resultado un mapa de características diferente (se genera un sobreajuste). Un enfoque común para abordar este problema es el submuestreo de la imagen, que consiste en crear una versión de menor resolución que contiene los elementos estructurales importantes de la imagen, pero sin los detalles finos. La capa de agrupación es la que realiza dicho submuestreo y opera sobre cada mapa de características por separado. El resultado de usar una capa de agrupación y crear mapas de características submuestreados es crear una versión resumida de las características detectadas en la entrada.

Por lo tanto, el objetivo de realizar el agrupamiento es reducir progresivamente el tamaño espacial de la representación para reducir la cantidad de parámetros y cálculos en la red. Un gran beneficio de esto es que reducirá la posibilidad de que el modelo se ajuste demasiado. En otras palabras, al tener menos información espacial, el modelo tiene menos parámetros para entrenar, lo que reduce las posibilidades de sobreajuste. Las dos capas principales de agrupación son agrupación máxima

(*Max Pooling*) y agrupación media (*Average Pooling*).

- **Agrupación Máxima**: se captura el valor máximo de cada subregión y se lo coloca en la imagen de salida. En general, se trabaja con bloques de 2 x 2 con un tamaño del paso de 2. Esto significa que se toma una entrada que tenga un tamaño de 4 x 4 y se la reduce a un tamaño de 2 x 2 tal como se observa en la figura 2.10.
- **Agrupación Promedio**: genera el valor promedio de la subregión. Es similar al anterior, excepto que esta vez se toma el promedio de cada uno de los bloques en lugar del valor máximo para cada uno de los cuatro cuadrados.

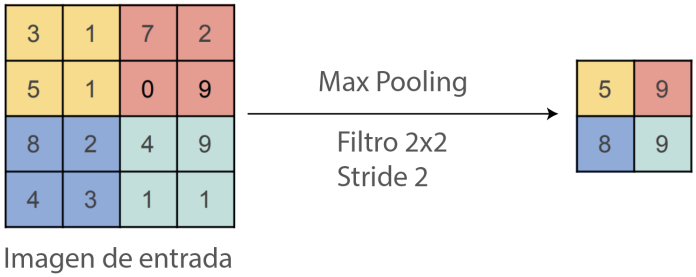

Figura 2.10: Ejemplo de aplicar *Max Pooling* a una imagen

Para ilustrar todas las capas analizadas hasta el momento, a continuación se puede observar un ejemplo de arquitectura de una CNN:

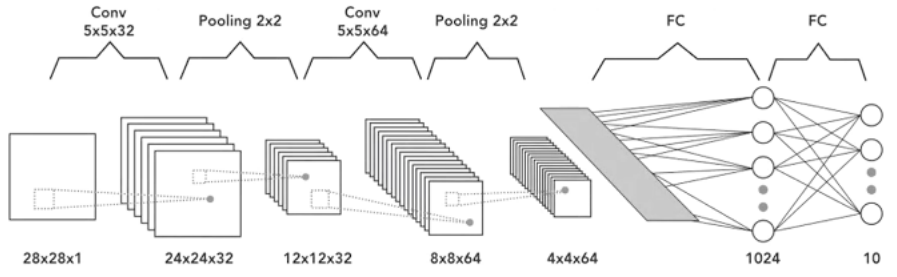

Figura 2.11: Ejemplo de una Red Neuronal Convolucional

La entrada de la red es una imagen de tamaño 28 x 28 x 1; es decir, es una imagen en escala de grises porque solo posee un canal. Los datos de entrada pasan por dos capas de convolución que tienen un tamaño de kernel de 5 x 5. La primera y la segunda convolución tienen 32 y 64 mapas

de características (*features maps*) de salida respectivamente. A cada capa de convolución le sigue un submuestreo (*pooling*), y podemos ver que después de esas capas, las dimensiones se reducen a la mitad.

En resumen, la imagen de entrada es de 28 x 28 x 1, luego de pasar por una convolución con 32 filtros cada uno con tamaño de 5 x 5, se obtienen 32 mapas de características de tamaño 24 x 24, tal como se observa a continuación, aplicando la ecuación 2.2.

$$n' = 28 - 5 + 1 = 24 \tag{2.6}$$

Luego hay una capa de *Pooling* con un bloque de 2 x 2 en la que se obtienen 32 mapas de característica de tamaño 12 x 12. Le sigue una convolución con 64 filtros de 5 x 5 por lo que el tamaño final resulta de 8 x 8 x 64, aplicando la ecuación 2.2 queda:

$$n' = 12 - 5 + 1 = 8 \tag{2.7}$$

Luego, se le aplica la última capa de Pooling con un bloque de 2 x 2 y el resultado final son 64 mapas de característica de tamaño 4 x 4. Para finalizar con la arquitectura de la red neuronal, se le hace un *flatten* al resultado obtenido para que pase de dimensión 2D a 1D y se le aplican dos capas completamente conectadas. La última de ellas, posee 10 neuronas que son las que finalmente clasifican entre las clases.

Como se observa en la figura 2.11, la red neuronal realiza la clasificación en 10 clases distintas. La última capa completamente conectada posee tantas neuronas como clases (en este caso son 10) y en cada una de esas neuronas se presenta la probabilidad de que la imagen de entrada pertenezca a esa clase. Por lo tanto, la suma de las probabilidades que se presentan en las neurona, da uno. Finalmente, se elije a la clase que posee mayor probabilidad.

2.2.4. Mejoras en las CNN

Existen distintos tipos de mejoras que se les puede aplicar a las CNN. Sin embargo, las dos que se describirán en el presente trabajo son el aumento de los datos y las capas de *dropout*.

El propósito del **Dropout** es evitar un ajuste excesivo y consta de un abandono aleatorio de neuronas que obliga a la red a aprender una representación redundante de los datos. Por lo tanto, el

dropout elimina aleatoriamente a neuronas en una capa de un conjunto de entrenamiento con probabilidad p. Esta probabilidad de abandono la determina el usuario siendo la opción más común: p = 0,5; es decir, la mitad de las neuronas de una capa se eliminan durante el entrenamiento. Por lo tanto, la red no puede depender de la activación de ningún conjunto de unidades ocultas, ya que pueden apagarse en cualquier momento durante el entrenamiento. Por lo tanto, el modelo se ve obligado a aprender patrones más generales y sólidos a partir de los datos.

El **aumento de imágenes** es el proceso de tomar imágenes que ya están en un conjunto de datos de entrenamiento y manipularlas para crear muchas versiones alteradas de la misma imagen. Esto no solo proporciona más imágenes para entrenar, sino que también puede ayudar a exponer a nuestro modelo a una variedad más amplia de manipulaciones de imágenes, como la reflexión y, por lo tanto, hacer que el modelo sea más robusto. El objetivo al aplicar el aumento de datos es incrementar la generalización del modelo. Dado que que la red está constantemente viendo versiones nuevas y ligeramente modificadas de los datos de entrada, puede aprender características más robustas. Existen principalmente dos tipos de método de aumento de datos, tales como:

- **Estático**: Generación y expansión de conjuntos de datos a través del aumento de datos. En este método, se aplica la transformación a cada imagen y se guarda la imagen transformada en el disco. El modelo de aprendizaje profundo está entrenado tanto en imágenes originales como en imágenes transformadas.
- **Dinámico**: Aumento de datos en el lugar. En este método, las transformaciones de las imágenes se realizan mientras se entrena el modelo. *Image Data Generator* de Keras usa este tipo de aumento de datos.

2.2.5. Arquitecturas

ImageNet [1] es un proyecto que tiene como objetivo proporcionar una gran base de datos de imágenes para fines de investigación y fue creada en 2009 por Fei-Fei Li. Contiene más de 14 millones de imágenes que pertenecen a más de 20.000 clases. *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC [2], por sus siglas en inglés) es una competencia anual organizada por el equipo de ImageNet desde el año 2010, donde los equipos de investigación evalúan sus algoritmos de visión por computadora en diversas tareas de reconocimiento visual, como la clasificación y localización de objetos. Los datos de entrenamiento son un subconjunto de ImageNet con 1.2 millones de imágenes

[1] http://www.image-net.org/
[2] http://www.image-net.org/challenges/LSVRC/index

pertenecientes a 1000 clases.

Deep Learning llegó al centro de atención en 2012 cuando Alex Krizhevsky y su equipo ganaron la competencia por un margen de 11 % (con la red denominada: AlexNet). A partir de dicho año, comenzaron a ganar el desafío diferentes redes neuronales convolucionales como: en el año 2013 ZedFNet (8 capas), en el año 2014 VGG (19 capas) y GoogleLeNet (22 capas) y en el año 2015 ResNet (152 capas). Las categorías van desde gran tiburón blanco hasta glaciares y el ganador es quien tenía la menor cantidad de errores posible. A lo largo de los años, el porcentaje de error para las cinco categorías principales se ha reducido de poco más del 28 % para el modelo ganador de ImageNet de 2010 a poco más del 3.5 % para el modelo ResNet. Curiosamente, el número de capas en el modelo de red neuronal de convolución ha aumentado de ocho capas en 2012 y 2013, a 19 capas con VGG en 2014, 22 capas de GoogLeNet y finalmente 152 capas con ResNet.

Los ganadores de ILSVRC han lanzado sus modelos a la comunidad de código abierto. Además muchos grupos de investigación también comparten sus modelos que han entrenado para tareas similares, por ejemplo, MobileNet, SqueezeNet, etc. Keras [3] contiene muchos modelos ya entrenados que poseen tanto la arquitectura como los pesos.

Como se mencionó anteriormente, una de las primeras redes profundas fue **AlexNet** [49], que consta de 5 capas convolucionales seguidas de tres capas completamente conectadas y que terminan con una capa softmax. Cada una de las dos primeras capas convolucionales es seguida por la normalización y las capas de agrupación máxima, y una capa de agrupación máxima sigue a la última capa convolucional. AlexNet utilizó la función de activación de ReLU (Ver Anexo 7.6.4) y su arquitectura se observa a continuación.

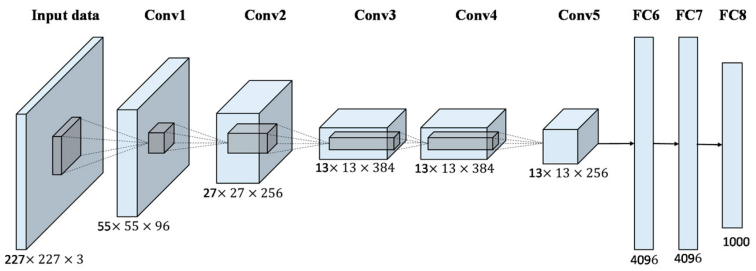

Figura 2.12: Arquitectura AlexNet

[3] https://keras.io/api/applications/

El Grupo de Geometría Visual de la Universidad de Oxford mejoró AlexNet reemplazando el gran tamaño de kernel de los filtros en AlexNet por múltiples filtros de tamaño de kernel de 3 x 3. Esta arquitectura se conoce como **VGG** [50], que significa *Visual Geometry Group*. VGG requiere una alta potencia computacional, ya que necesita una gran cantidad de memoria de almacenamiento y un alto tiempo de computación. La arquitectura de VGG-16 consta de 16 capas de la siguiente manera: 13 capas convolucionales, 5 capas de agrupación máxima y 3 capas densas. A continuación se observa la red VGG16:

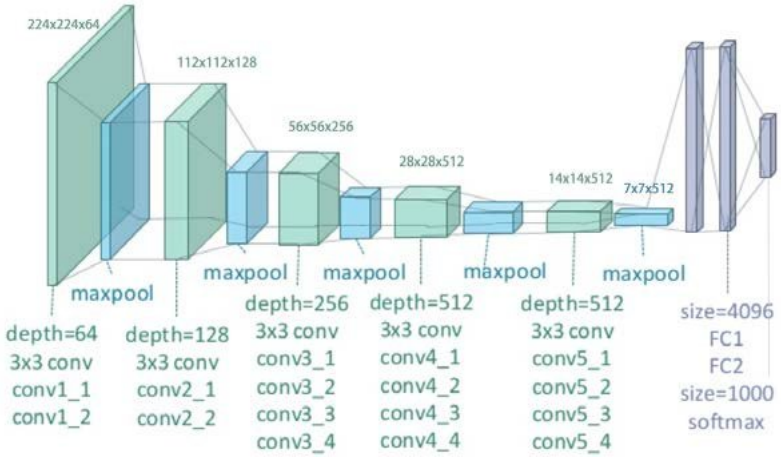

Figura 2.13: Diseño de la red VGG16

GoogLeNet [51] introdujo el modelo inicial, ya que sugiere que la mayoría de las conexiones en la arquitectura densa están correlacionadas y, por lo tanto, pueden eliminarse. Usó tres tamaños de convoluciones diferentes, 5 x 5, 3 x 3, y 1 x 1, para reducir los requisitos computacionales y comprender mejor los pequeños detalles. más redujo el número total de parámetros. Introdujo una capa convolucional de agrupación promedio global como su última capa convolucional para promediar los valores del canal en el mapa de características 2D.

A diferencia de GoogLeNet, AlexNet y VGG, la red residual (**ResNet** [52]) no es una arquitectura de red secuencial, sino una arquitectura en red. Utiliza microarquitecturas (bloques de construcción junto con capas de agrupación, convolución, etc.) para construir una macroarquitectura. Además, introdujo conexiones de salto en bloque en capas convolucionales para construir un módulo residual. Canziani et al. [53] presentan una revisión exhaustiva del desempeño de modelos previamente entrenados en problemas de visión por computadora utilizando datos del desafío ImageNet.

2.2.6. Transfer Learning y Fine Tuning

Como se mencionó anteriormente, muchas de las arquitecturas ganadoras de la competencia ILSVRC están disponibles para su uso, por ejemplo, en Keras. Estos modelos fueron entrenados con un gran conjunto de datos, generalmente en una tarea de clasificación de imágenes a gran escala. Para realizar nuevas clasificaciones, se pueden utilizar los modelos pre-entrenados (de ahora en adelante, también llamados 'modelo base') realizando aprendizaje por transferencia (*Transfer Learning*) o ajuste fino (*Fine Tuning*) con el fin de personalizar este modelo para una tarea determinada.

El aprendizaje por transferencia y el ajuste fino son métodos populares en la visión por computadora porque nos permiten construir modelos precisos ahorrando mucho tiempo. Con el aprendizaje por transferencia, en lugar de comenzar el proceso de aprendizaje desde cero, se parte de patrones que se han aprendido al resolver un problema diferente. Debido al costo computacional de entrenar modelos desde cero, es una práctica común importar y usar modelos de la literatura publicada (por ejemplo, VGG, Inception, MobileNet). Existen dos formas de personalizar un modelo previamente entrenado:

- **Transferencia por aprendizaje**: Las últimas capas de un modelo pre-entrenado son específicas de la tarea de clasificación original y del conjunto de clases en las que se entrenó el modelo. Por lo tanto, para realizar transferencia de aprendizaje, se elije un modelo pre-entrenado y se le elimina la última capa completamente conectada que es la que realiza la clasificación en las categorías. Luego se le agrega un nuevo clasificador sobre el modelo base que se entrenará desde cero. De esa manera, se puedan reutilizar los mapas de características aprendidos previamente por la red y no es necesario reentrenar todo el modelo.
- **Ajuste fino**: Se basa no solo en reemplazar y volver a entrenar al clasificador colocado en la parte superior de la red pre-entrenada, sino que también en ajustar algunos de los pesos de las últimas capas de la misma. Es posible ajustar los pesos de todas las capas del modelo base o mantener algunas de las capas anteriores fijas (debido a problemas de sobreajuste) y solo ajustar los pesos de las últimas capas. Esto está motivado por la observación de que las primeras capas de la red contienen características más genéricas (por ejemplo, detectores de bordes o detectores de manchas de color) que deberían ser útiles para muchas tareas, pero las últimas capas de la red se vuelven progresivamente más específicas para los detalles de las clases del conjunto de datos original.

Por lo tanto, en otras palabras, se pueden usar las capas convolucionales como un extractor de

características (*Transfer Learning*) o se pueden ajustar las capas ya entrenadas para adaptarse a un problema en cuestión (*Fine-tuning*).

2.3. Evaluación del modelo

2.3.1. Métricas de evaluación para clasificación

Existen diversas métricas para la evaluación del rendimiento del modelo, las cuales son: exactitud, precisión, recall, f1-score, la matriz de confusión, la curva ROC y el AUC (*Area under the curve*). Para entender el significado de cada uno, primero hay que analizar los siguiente conceptos:

- Verdadero Positivo (VP): es la cantidad de observaciones clasificadas como positivas cuando en realidad lo eran.
- Verdadero Negativo (VN): es la cantidad de observaciones clasificadas como negativas cuando en realidad lo eran.
- Falso Positivo (FP): es la cantidad de observaciones clasificadas como positivas cuando en realidad eran negativas.
- Falso Negativo (FN): es la cantidad de observaciones clasificadas como negativas cuando en realidad eran positivas.

Por lo tanto, cada métrica se calcula de la siguiente manera:

$$Exactitud = \frac{VP + VN}{VP + FN + FP + VN} \tag{2.8}$$

$$Precisión = \frac{VP}{VP + FP} \tag{2.9}$$

$$Recall = TVP = \frac{VP}{VP + FN} \tag{2.10}$$

$$F1 - score = \frac{2 * Precisión * Recall}{Precisión + Recall} \tag{2.11}$$

$$TFP = \frac{FP}{FP + TN} \tag{2.12}$$

2.3.2. Matriz de confusión

La matriz de confusión es una tabla que permite evaluar el desempeño de un algoritmo de clasificación. Cada columna de la matriz presenta la cantidad de instancias en la clase predicha mientras que cada fila representa la cantidad de instancias en la clase verdadera. El nombre *confusión* proviene del hecho de que esta matriz permite determinar si el sistema está confundiendo dos clases. A continuación se observa la matriz:

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Figura 2.14: Matriz de confusión

El **error tipo I** se da cuando se rechaza a la hipótesis nula cuando es verdadera y es el falso positivo. Por otro lado, el **error tipo II** es cuando no se rechaza a la hipótesis nula siendo esta falsa y es el falso negativo. Por lo tanto, si la clasificación binaria es entre tumor benigno o maligno, el FP se da si se le comunica al paciente que presenta un tumor maligno cuando en realidad es benigno. Esto se traduce en estudios adicionales (como la biopsia) al paciente para determinar la morfología y la estructura del tumor. Por último, el caso del FN se da cuando se le comunica al paciente que posee un tumor benigno en la mama cuando en realidad es maligno. Este último caso (error de tipo II) se tiene que reducir al máximo ya que la paciente tiene cáncer de mama y no va a iniciar el tratamiento pertinente debido a que se consideró que el tumor es benigno. Por lo tanto, el error de tipo II es el que se va a minimizar.

2.3.3. Curva ROC

La curva ROC (*Receiver Operating Characteristics*) es un gráfico que ilustra la capacidad de diagnóstico de un clasificador binario a medida que varía un umbral que discrimina entre las clases; es decir, visualiza el efecto de un umbral (por ejemplo de probabilidad) elegido sobre la eficiencia de clasificación. La curva ROC se crea trazando la tasa de verdaderos positivos frente a la tasa de

falsos positivos en varios valores de umbral. Por lo que es posible analizar el error de la clasificación midiendo el área debajo de la curva. El AUC (*Area under the curve*) representa el grado o medida de separabilidad entre las clases; es decir, indica cuánto es capaz el modelo de distinguir entre clases. Cuanto más cerca de 1 es este área, mejor será la clasificación. Un clasificador se representa en el espacio ROC como (Tasa FP, Tasa VP). El espacio ROC es un cuadrado de [0,1] x [0,1], tal que en el eje x están los valores de la Tasa FP y en el eje y los valores de la Tasa de VP. A continuación se observa el gráfico de la curva ROC:

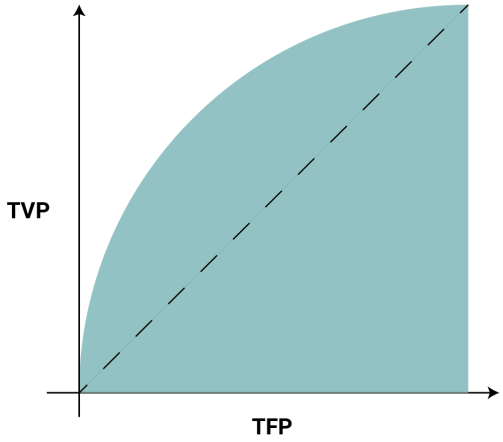

Figura 2.15: Esquema de la curva ROC. En el eje x se ubica la tasa de falsos positivos (TFP) y en el eje y la tasa de verdaderos positivos (TVP). En color celeste se observa el área debajo de la curva.

Por ejemplo, el punto (TFP1, TVP1) es mejor que el punto (TFP2, TVP2) si:

- Tasa FP1 $<$ Tasa FP2
- Tasa VP1 $>$ Tasa VP2

En el gráfico el punto (Tasa FP1, Tasa VP1) debe estar más a la izquierda y arriba del (Tasa FP2, Tasa VP2).

Un clasificador se representa con un solo punto en la curva ROC. Por lo que para construir una curva ROC, se debe usar alguna estrategia que depende del problema y del clasificador. Por ejemplo, si el clasificador provee una probabilidad, se la utiliza como el umbral para calcular la TVP y la TFP.

El área debajo de la curva ROC (AUC) es una métrica que puede ser usada para comparar diferentes test de clasificación y es una medida de la precisión del mismo. Tal como fue mencionado anteriormente, el TVP se calcula como los VP divido la suma de los VP y los FN; es decir, es el ratio de las observaciones positivas que fueron correctamente clasificadas (la cantidad de observaciones clasificadas como positivas divido la cantidad de observaciones realmente positivas). Esta tasa describe qué tan bueno es el modelo para predecir la clase positiva cuando el resultado real es positivo. Por lo tanto:

$$TVP = \frac{VP}{VP + FN} \tag{2.13}$$

La TFP se calcula como los FP sobre la suma de los VN y los FP; es decir, es el ratio de las observaciones negativas que fueron incorrectamente clasificadas (la cantidad de observaciones negativas que fueron predichas como positivas divido todas las observaciones que eran realmente negativas). La tasa se calcula de la siguiente manera:

$$TFP = \frac{FP}{VN + FP} \tag{2.14}$$

Un modelo que es excelente tiene un AUC cercano a 1, lo que significa que tiene una buena medida de separabilidad entre clases. Por otro lado, un modelo deficiente tiene un AUC cercano a 0, lo que significa que tiene la peor medida de separabilidad entre clases; es decir, predice 0 como 1 y 1 como 0. Por último, cuando el AUC es 0,5 significa que el modelo no tiene capacidad de separación de clases en absoluto. A continuación se observa una comparación realizada por Ferraris [21] entre los distintos tipos de curvas ROC que se pueden obtener:

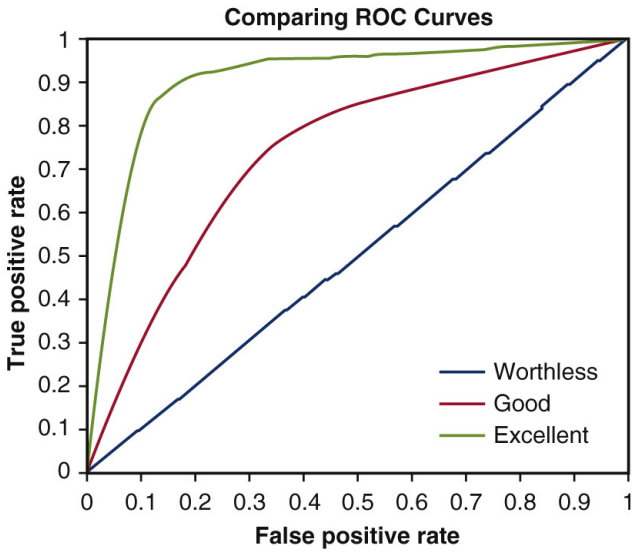

Figura 2.16: Comparación entre los distintos tipos de curvas ROC que se pueden obtener: excelente (verde), buena (roja) y sin valor (azul).

Muchos algoritmos de clasificación utilizan la probabilidad para distribuir muestras en clases y, en la mayoría de los casos, el umbral de probabilidad tiene un valor predeterminado de 0,5. Lo que significa que el algoritmo clasifica una muestra como positiva si la probabilidad de que esa muestra sea positiva es superior a 0.5 y clasifica una muestra como negativa si la probabilidad de que esa muestra sea positiva es menor de 0.5.

Este umbral predeterminado puede no ser suficiente. Por ejemplo, al diagnosticar una enfermedad puede ser prudente elegir un umbral de probabilidad más bajo para evitar cualquier posibilidad de que la enfermedad se clasifique erróneamente. Para construir la curva ROC, se va variando el umbral de clasificación y graficando los valores de la TVP y la TFP. Por ejemplo, primero el umbral vale 0 por lo que todas las observaciones se clasifican como clase 1. Luego, el umbral se modifica a 0.1 por lo que todas las observaciones con probabilidad menor a 0.1 se predicen como clase 0 y las mayor a 0.1 como clase 1. Luego, se mueve el umbral a 0.2 y se repite el proceso anterior. Así se va construyendo la curva ROC que se observa en la imagen 2.15. El clasificador perfecto, se encuentra donde la TFP es cero y la TVP es uno.

Capítulo 3

Materiales y Métodos

3.1. Datos

3.1.1. Base de datos: CBIS-DDSM

En el año 2017, la Universidad de Stanford creó la base de datos ***Curated Breast Imaging Subset of DDSM*** (CBIS-DDSM, por sus siglas en inglés), una versión actualizada de la base *Digital Database for Screening Mammography* (DDSM) que proporciona mamografías fácilmente accesibles y una segmentación de ROI mejorada [22]. Además, contiene a las imágenes de DDSM convertidas al formato DICOM estándar.

Por lo tanto, para desarrollar el presente trabajo se eligió la base de datos CBIS-DDSM ya que no solo posee la categorización del tumor en benigno o maligno, sino que también posee la distinción entre masas y calcificaciones. Otras razones de su elección son que la base de datos contiene las imágenes con las regiones de interés (ROI), es una de las que posee mayor cantidad de mamografías y es la más utilizada en las publicaciones actuales. A continuación, se muestran imágenes de la base de datos CBIS-DDSM de cada clase que son las que se van a utilizar para el entrenamiento de las CNN del proyecto. Por un lado, las imágenes de las masas son:

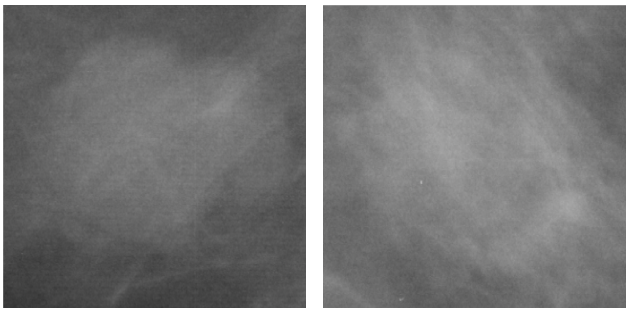

Figura 3.1: Imagen de la ROI de una masa benigna (izquierda) y maligna (derecha). Fuente: CBIS-DDSM

Por el otro, las que contienen a las calcificaciones son:

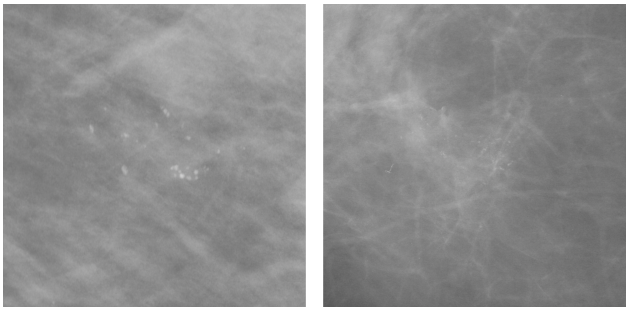

Figura 3.2: Imagen de la ROI de una calcificación benigna (izquierda) y maligna (derecha). Fuente: CBIS-DDSM

Se puede observar en las imágenes que la discriminación entre benigno o maligno tanto en las masas como en las calcificaciones, es compleja a simple vista. Por otro lado, es más fácil distinguir para una persona no experta en la materia, las masas de las calcificaciones.

El conjunto de datos se descargó del sitio web de CBIS-DDSM [1] e incluyó vistas craneocaudales (CC) y oblicua mediolateral (MLO) para la mayoría de los exámenes. La base de datos CBIS-DDSM contiene 1566 pacientes en total y tiene un peso de 163.5 GB (cada imagen pesa alrededor 25 MB). Por otra parte, la base de datos provee archivos con extensión .cvs en el que se dispone la siguiente información:

- ID del paciente
- Categoría de la densidad: 1, 2, 3 o 4

[1] https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM

- Mama: derecha o izquierda
- Vista: CC o MLO
- Cantidad de anormalidades en la imagen: 1 a 7. Esto es necesario ya que hay algunos casos que contienen múltiples anormalidades.
- Forma de la masa (cuando aplica): distorsión arquitectónica, tejido mamario asimétrico, densidad asimétrica focal, irregular, lobulado, ganglio linfático, ovalado y redondo.
- Margen de la masa (cuando aplica): circunscrito, mal definido, microlobulado, oscurecido y espiculado.
- Tipo de calcificación: amorfa, gruesa, distrófica, cáscara de huevo, ramificación lineal fina, gran varilla, centro lúcido, pleomórfica, punteada, redonda y regular, cutánea y vascular.
- Distribución de la calcificación (cuando aplica): agrupados, difusa, lineales, regionales y segmentarios.
- Asistencia BI-RADS: 0 a 5
- Patología: Benigno o Maligno
- Calificación de sutileza: 1 a 5. Grado de dificultad para los radiólogos en detectar la anormalidad en la imagen
- Ruta a los archivos: de la imagen, de la imagen segmentada y de la máscara ROI.

Se descargaron las imágenes en formato DICOM desde la página web de la base CBIS-DDSM usando el programa *NBIA Data Retriever*. Como se mencionó anteriormente, la información de dicha ruta se encontraba en los archivos .cvs. Una vez que se descargaron desde la base de datos a una memoria extraíble de un terabyte, se subieron a Google Drive. Luego, se les cambió tanto el nombre como la ubicación del archivo. Todo el proceso de descarga y carga duró aproximadamente 96 horas. Finalmente, las imágenes quedaron organizadas en cuatro carpetas de la siguiente forma: masas para entrenamiento, masas para prueba, calcificaciones para entrenamiento y calcificaciones para prueba.

En el presente trabajo se van a probar arquitecturas creadas desde cero y redes pre-entrenadas tales como la VGG16 y la Resnet50 que fueron entrenadas con millones de imágenes de tamaño 224 x 224 x 3 (imágenes RGB de tamaño 224 x 224). Sin embargo, las mamografías de la base de

datos, poseen un tamaño mucho mayor (5000x3000) y si se les hace el cambio de tamaño a 224 x 224, la red no aprende correctamente ya que no se aprecian las regiones de interés (las masas y calcificaciones) en una imagen tan pequeña con poca resolución y tanto ruido. Por tal motivo, se decidió entrenar a la red neuronal con las ROI de las mamografías en las que solo se ven la masa o calcificación, con un tamaño aproximado de 300x400 cada una. Dichas imágenes de las ROI, estaban incluidas en la base de datos descargada junto con las mamografías.

Un problema enfrentado cuando se descargaron las imágenes de la base de datos fue que en la carpeta se encontraban tanto las imágenes de las ROI como las máscaras de los tumores segmentados. Como solo se desea trabajar con las ROIs, se las separó según el peso de cada imagen ya que las ROI tenían un tamaño aproximado de 300x400 mientras que las máscaras de 3000x5000.

3.1.2. División de datos en entrenamiento y prueba

La base de datos cuenta con un conjunto de imágenes de las regiones de interés para el entrenamiento de la red y otro conjunto para la prueba. La suma de las mamografías de entrenamiento y prueba, arrojan un total de 2961 imágenes tal como se observa en la siguiente tabla:

	Entrenamiento	Prueba	Total
Masas	1158	354	1512
Calcificaciones	1176	273	1449
Total	2334	627	**2961**

Cuadro 3.1: Distribución de las imágenes de las regiones de interés (ROI) de la base de datos CBIS-DDSM

En el trabajo, se mantuvo la división realizada por los autores de la base de datos. Por lo tanto, la distribución final fue de 2.334 imágenes de entrenamiento (equivalen al 78,8 %) y 627 imágenes de prueba (equivalen al 21,17 %). Posteriormente, el conjunto de validación se extrajo del de entrenamiento con una relación 20:80.

3.1.3. Armado de tensores

Un tensor es la estructura de datos principal que utilizan las redes neuronales y es una generalización de vectores y matrices; es decir, es una matriz multidimensional. Por ejemplo, un vector es un tensor unidimensional o de primer orden y una matriz es un tensor bidimensional o de segundo orden.

La clase *Image Data Generator* de Keras, presenta tres formas de cargar los datos en memoria: *flow*, *flow from directory* y *flow from dataframe*. Cada una de estas funciones ejecuta la misma tarea de cargar el conjunto de datos de imágenes en la memoria, solo que de diferente forma. Por lo tanto, se decidió usar al método *.flow()* que asume que X e y son matrices NumPy; es decir, las imágenes y sus respectivas clases deben estar en formato matriz de Numpy por lo que se hicieron tensores para las imágenes y sus etiquetas.

Debido a que las imágenes descargadas estaban en formato .dcm (DICOM), se convirtieron a .png, se estandarizaron los tamaños de las imágenes en 150 x 150 y se armaron los tensores de Numpy, uno con las imágenes y otro con las clases (tanto para el entrenamiento como para la prueba) cuyos tamaños quedaron:

- Entrenamiento:

  Tensor con imágenes: (2.334, 150, 150)

  Tensor con clases: (2.334)

- Prueba:

  Tensor con imágenes: (627, 150, 150)

  Tensor con clases: (627)

Además, se armaron los tensores con las etiquetas de cada imagen según cada clasificación probada; es decir, existen tres tensores de etiquetas de la siguiente forma:

- Clasificación 1: Masas (clase 0) o Calcificaciones (clase 1)
- Clasificación 2: Tumores Benignos (clase 0) o Malignos (clase 1)
- Clasificación 3: Masa Benigna (clase 0), Masa Maligna (clase 1), Calcificación Benigna (clase 2) o Calcificación Maligna (clase 3)

De esta manera, se guardaron los tensores en formato *.npy* y se procesaron los datos para poder utilizar el método *.flow()* provisto por Keras. Luego, se armaron dos funciones que cargan los datos en la notebook.

3.1.4. Procesamiento de los datos

Debido a que la cantidad de imágenes que entran a la red neuronal son escasas, se realizó el **aumento de la cantidad de datos** usando la clase *Image Data Generator* de Keras. Por último,

se verificó que los **datos estén balanceados** entre las clases. En el caso de la clasificación entre masas o calcificaciones, se tienen 1158 masas (49,61 %) y 1176 calcificaciones (50,3 %). Por otro lado, para la clasificación entre tumores benignos o malignos, la distribución de mamografías quedó: 1008 tumores benignos (50.02 %) y 1007 tumores malignos (49,97 %), lo que arrojó un total de 2015 imágenes de entrenamiento.

Como se mencionó anteriormente, tanto la red VGG16 como la Resnet50 fueron entrenadas con imágenes RGB de tamaño 224 x 224. Por lo tanto, se decidió usar las imágenes de las ROI únicamente y se les hizo el *reshape* con el tamaño 150 x 150. El procesamiento que recibieron las imágenes fue:

- Se importaron los datos de entrenamiento y prueba en formato de matrices Numpy (utilizando el método *flow*)
- Se importaron las etiquetas para el problema de clasificación
- Se normalizaron los píxeles para que se encuentren en el rango (0-1)
- Se mezcló el conjunto de datos
- Se dividieron los datos del conjunto de entrenamiento en subconjuntos de 'entrenamiento' y 'validación'
- Se construyeron los generadores de Keras para entrenamiento y validación de datos

3.1.5. Ambiente de desarrollo

Se utilizó la librería Keras usando Tensorflow en Python en el entorno Google Colaboratory. Todo el trabajo se realizó en una MacBook Pro con macOS Catalina. Se eligió Keras debido a que provee diversas redes neuronales tales como VGG16 o Resnet50 y se pueden usar tanto las arquitecturas como los pesos pre-entrenados de las mismas. Se utilizó la librería pydicom para leer las imágenes en formato DICOM; y pandas para leer y utilizar la información presente en los archivos Excel.

3.2. Red Neuronal Convolucional

El objetivo primario de la red es poder discriminar entre masas y calcificaciones. Luego, el segundo objetivo es clasificar entre tumor benigno o maligno a cada masa y calcificación. Por último, se

desea realizar la clasificación categórica (masa benigna, masa maligna, calcificación benigna o calcificación maligna), para la cuál se tomaron dos caminos diferentes:

- **Primer camino:** Se diseñaron dos redes neuronales, una que clasifica entre masas o calcificaciones y la otra que discrimina entre tumores benignos o malignos. Luego, se unieron ambas redes en un solo modelo para que pueda clasificar entre las 4 clases; es decir, masa benigna, masa maligna, calcificación benigna o calcificación maligna.
- **Segundo camino:** se diseñó una única CNN que clasifica entre las cuatro clases posibles.

Por lo tanto, se realizaron varios experimentos para poder analizar con cuáles se obtienen los mejores resultados de clasificación entre dos y cuatro clases. Primero, se diseñaron redes neuronales convolucionales con distintas capas, se entrenaron y testearon. Luego, se procedió a realizar *Transfer Learning* y *Fine tuning* con dos arquitecturas distintas. Las comparaciones entre los resultados se encuentran en la siguiente sección. Por lo tanto, para cada una de las tres clasificaciones mencionadas, se probara:

- CNN desde cero con dos arquitecturas diferentes
- Transferencia de aprendizaje con VGG16
- Transferencia de aprendizaje con Resnet50

Para realizar cada unos de los modelos y luego poder compararlos, se tomó como:

- **Métrica:** Exactitud. Se eligió esta métrica para poder comparar los resultados obtenidos con los publicados, que utilizan como métrica en su gran mayoría a la exactitud y en menor medida al AUC (*Area under de curve*).
- **Función de costo** (Ver Anexo 7.6.5): *Binary Cross-Entropy* (para las clasificaciones binarias) y *Categorical Cross-Entropy* (para la clasificación categórica)
- **Función de activación de la última capa densa:** Sigmoidea (para las clasificaciones binarias) y softmax (para la clasificación categórica)
- **Optimizador** (Ver Anexo 7.6.6): RMSprop y Adam
- ***Batch Size*:** 128
- ***Early stopping*:** Ajustado con el valor de la pérdida del conjunto de validación

3.2.1. Primer caso: CNN desde cero

Primero, se decidió diseñar y construir una red neuronal convolucional desde cero para probar la incidencia de las distintas capas y los métodos de mejora (como aumentar los datos) en los resultados. Lo primero que se hizo fue descubrir qué tan grande debe ser el modelo ya que una red demasiado pequeña no podrá generalizar bien; por otro lado, un modelo con demasiados parámetros puede aprender lentamente y sobreajustarse.

Para ello, se diseñaron diversas arquitecturas diferentes de las cuáles se eligieron dos de ellas. Con cada una, se realizaron distintos experimentos que consistieron en agregar capas como *dropout* o convolucionales, en cambiar el tamaño de los filtros, entre otros. A continuación, se analizan ambas arquitecturas.

Primera arquitectura

En esta sección del trabajo, se hicieron varios experimentos para analizar cuál es la arquitectura que mejor predice. Para lograr eso, se arrancó con un modelo sencillo y se fue complejizando a medida que se hacían las pruebas. Considerando que el tamaño de la imagen de entrada es de 150x150x1, a continuación se muestra la cantidad de parámetros de cada capa y el tamaño de salida de la imagen.

Capa	**Tamaño de la salida**	**Cantidad de parámetros**
Conv2D (32, (3, 3), relu)	148 x 148 x 32	320
MaxPooling2D ((2, 2))	74 x 74 x 32	0
Conv2D (64, (3, 3), relu)	72 x 72 x 64	18.496
MaxPooling2D((2, 2))	36 x 36 x 64	0
Flatten()	82.944	0
Dense(16, relu)	16	1.327.120
Dense(1, sigmoid)	1	17

Cuadro 3.2: Cantidad de parámetros y tamaño de salida luego de cada capa de la primera arquitectura diseñada

Se observa en la tabla que la cantidad de parámetros totales y entrenables en ésta arquitectura son **1.345.953**.

Segunda arquitectura

Esta arquitectura resulta ser más profunda y compleja que la anterior. El cálculo de parámetros es de la siguiente forma:

Capa	**Tamaño de la salida**	**Cantidad de parámetros**
Conv2D (32, (3, 3), relu)	148 x 148 x 32	320
Conv2D (32, (3, 3), relu)	146 x 146 x 32	9.248
MaxPooling2D ((2, 2))	73 x 73 x 32	0
Conv2D (64, (3, 3), relu)	71 x 71 x 64	18.496
Conv2D (64, (3, 3), relu)	69 x 69 x 64	36.928
MaxPooling2D((2, 2))	34 x 34 x 64	0
Conv2D (128, (3, 3), relu)	32 x 32 x 128	73.856
Conv2D (128, (3, 3), relu)	30 x 30 x 128	147.584
MaxPooling2D((2, 2))	15 x 15 x 128	0
Flatten()	28.800	0
Dense(32, relu)	32	921.632
Dropout(0,5)	32	0
Dense(1, sigmoid)	1	33

Cuadro 3.3: Cantidad de parámetros y tamaño de salida luego de cada capa de la segunda arquitectura diseñada

En este caso, son **1.208.097 parámetros** totales y entrenables del modelo. Tanto en la arquitectura 1 como en la 2, se realizaron diversos experimentos para analizar con cual se obtenían mejores resultados. A continuación se listan algunos de los modelos evaluados:

- Agregar una capa de *dropout*
- Realizar aumento de datos
- Agregar más bloques de capa convolucional y *Max Pooling*.
- Cambiar el tamaño a la capa densa
- Agregar el decaimiento de la tasa de aprendizaje (*Learning Rate Decay*) al optimizador
- Hacer más grande el kernel de la primera capa (cambiar de 3x3 a 5x5).
- Probar con ambos optimizadores (Adam y RMSprop)
- Agregar regularización L2
- Hacer *Batch Normalization*

A continuación se observan ambas arquitecturas:

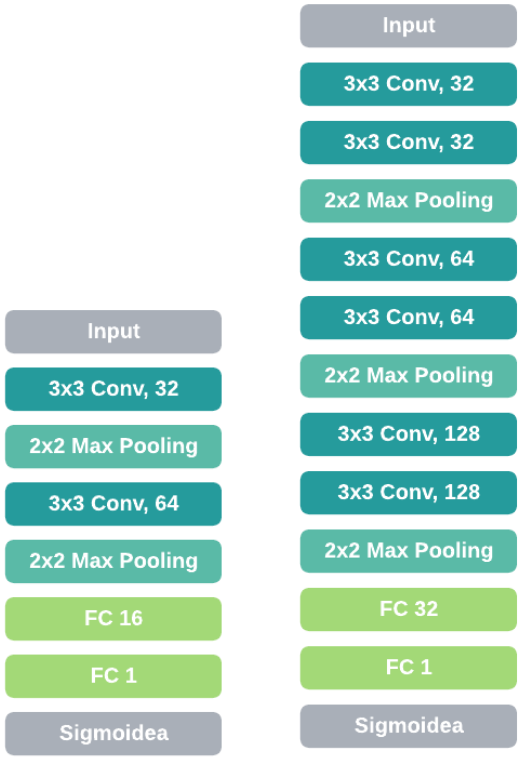

Figura 3.3: Diseño básico de la primera (izquierda) y la segunda (derecha) arquitectura entrenada. Posteriormente se le agregaron más capas convolucionales, *dropout*, entre otras.

3.2.2. Segundo caso: VGG 16

Primero se importó el modelo VGG16 en Keras junto con los pesos de ImageNet pero sin incluir las últimas tres capas densas del modelo ya que eran las que clasificaban entre las 1000 categorías de la competencia. De esa manera, se importó el modelo con 14.714.688 de parámetros entrenables cuyos pesos posteriormente se congelaron. Para realizar *Transfer Learning*, se creó un modelo secuencial al que se le agregó el modelo VGG16 recientemente importado y luego se le agregaron capas que son las que realizan la clasificación. Primero, se le hace un *flatten* a la salida del modelo pre-entrenado y luego se le agregan dos capas completamente conectadas: una con 256 neuronas y la otra con 1 neurona. De esa manera, quedan solo **2.097.665 de parámetros entrenables** ya que solo se actualizan los pesos del clasificador agregado. A continuación, se observan las capas agregadas al modelo.

Capa	Tamaño de salida	Cantidad de Parámetros
VGG 16 (Modelo)	4 x 4 x 512	14.714.688
Flatten	8.192	0
Dense (256)	256	2.097.408
Dense (1)	1	257

Cuadro 3.4: Cantidad de parámetros y tamaño de salida luego de cada capa realizando transferencia de aprendizaje de la red VGG16

A continuación se observa las capas agregadas al modelo para realizar la transferencia de aprendizaje.

Figura 3.4: Capas agregadas al modelo VGG16 para realizar la transferencia de aprendizaje

3.2.3. Tercer caso: Resnet 50

Se utilizó la arquitectura ResNet50 pre-entrenada, de la misma manera que la anterior, disponible en Keras por lo que hubo un ahorro de tiempo y recursos en el entrenamiento de la misma. Al igual que con la VGG16, se importaron tanto la arquitectura de la Resnet50 como los pesos de ImageNet y se le eliminaron las últimas capas densas teniendo así 23.587.712 de parámetros. Luego, se creó el modelo secuencial, se le agregó la Resnet recientemente importada y se le congelaron todos los pesos para que no se actualicen. Por último, se le agregó una capa de flatten, una capa densa con 256 neuronas y otra capa densa con 1 neurona. De esta manera, quedaron **13.107.713 de parámetros entrenables**, un número considerablemente mayor que el de la VGG que eran 2.097.665 parámetros entrenables. El final de la red quedó de la siguiente manera:

Capa	Tamaño de salida	Cantidad de Parámetros
Resnet 50 (Modelo)	5 x 5 x 2048	23.587.712
Flatten	51.200	0
Dense (256)	256	13.107.456
Dense (1)	1	257

Cuadro 3.5: Cantidad de parámetros y tamaño de salida luego de cada capa realizando transferencia de aprendizaje de la red Resnet50

3.2.4. Desarrollo de la red

Las arquitecturas disponibles en Keras están entrenadas en problemas de clasificación de imágenes a gran escala. Las capas convolucionales actúan como extractor de características y las *fully connected layers* actúan como clasificadores. Dado que estos modelos son muy grandes y han visto una gran cantidad de imágenes, tienden a aprender características muy buenas y discriminatorias.

Como ya se mencionó anteriormente, se realizan tres clasificaciones distintas: dos binarias (entre masas o calcificaciones y entre tumor benigno o maligno) y una categórica (entre masa benigna, masa maligna, calcificación benigna y calcificación maligna). Para cada una de esas clasificaciones, el desarrollo de las redes neuronales consta de tres arquitecturas: desde cero, VGG16 y Resnet50. Para la VGG16 y Resnet50, se realiza tanto *Fine Tuning* como *Transfer Learning*. Para el *Fine Tuning*, se descongelan los pesos de algunas de las capas superiores de la red pre-entrenada y se entrenan conjuntamente tanto las capas del clasificador recientemente agregadas como las últimas capas del modelo pre-entrenado. La métrica elegida para el trabajo es la exactitud del conjunto de prueba. Por lo tanto, cuando se habla de "mejor resultado", se hace referencia a una exactitud del modelo más alta.

Para cada clasificación, se probaron diversas modificaciones de la arquitectura original y en cada caso, primero se arrancó con una arquitectura simple que se la fue complejizando. Si con el cambio realizado a la red, se obtenían mejores resultados, el cambio se mantenía para los siguientes modelos. Sin embargo, si con el cambio realizado no se obtenía una mejor exactitud, dicha modificación se descartaba. Por ejemplo, en el caso de la arquitectura diseñada desde cero, primero se agregó una capa de *dropout* y luego se hizo aumento de datos. Si con la capa de *dropout*, se obtiene una exactitud más alta que la del modelo anterior, entonces se mantiene el cambio y se hace aumento de datos con la capa de *dropout* incorporada en la red. De lo contrario, se descarta el cambio y se hace aumento de los datos del modelo anterior (sin capa de *dropout*).

En todos los modelos se usó ***early stopping*** (según el valor de la pérdida del grupo de validación) con una paciencia definida por el usuario. Durante el entrenamiento, cada vez que se encontraba un modelo mejor que el anterior, se guardaban los pesos. De esa manera, cuando el entrenamiento finalizaba, ya sea porque se cumplían la cantidad de *epochs* establecidos o porque se cumplía la paciencia definida, siempre se tenía guardado el mejor modelo. Teniendo en cuenta, que el mejor modelo en este caso era el que presentaba menor valor de pérdida del grupo de validación.

Por otro lado, se fue variando la cantidad de *epochs* de cada entrenamiento. A los modelos más básicos, se los entrenó por 100 o 500 *epochs*. Si en las curvas obtenidas se veía que la pérdida iba en descenso o que la exactitud iba en aumento y se cortaba el entrenamiento, se volvía a correr el modelo pero con más *epochs*. Por lo contrario, si en las curvas se observaba cierto estancamiento o sobreajuste, no se volvía a entrenar al modelo con más *epochs*.

Hiperparámetros

En el aprendizaje automático, un hiperparámetro es un parámetro cuyo valor se establece antes de que comience el proceso de aprendizaje. Hay varios de estos valores que debemos especificar como por ejemplo: la función de activación que se desea usar para las neuronas, la tasa de aprendizaje, la cantidad de neuronas en cada capa, la cantidad de capas ocultas, etc. **Para las arquitecturas diseñadas desde cero**, los hiperparámetros a establecer fueron:

- La función de activación: para cada capa de la red neuronal se usó la función ReLU y para la última capa densa se usó tanto la sigmoidea (para las clasificaciones binarias) como la softmax (para la clasificación categórica).
- La tasa de aprendizaje del optimizador: se fue modificando entre 0,01 y 0,0001.
- Cantidad de neuronas en cada capa: se fue modificando la cantidad de neuronas de la ante última capa densa. Se probaron 16, 32, y 64 neuronas.
- Tamaño de los filtros de las capas convolucionales: todos son de tamaño 3x3. Se probó a modificar el tamaño del primer filtro a 5x5.
- Cantidad de capas ocultas: la primera arquitectura posee 6 capas ocultas mientras que la segunda arquitectura 11 capas ocultas.
- *Epochs*: en los modelos más básicos se comenzó con 100 *epochs* pero a medida que se iban complejizando se aumentó a 1000.
- *Batch Size*: en general todos los modelos se entrenaron con el tamaño de 128. Se probó con 32, 64 y 128.

Para las arquitecturas en las que se hicieron transferencia de aprendizaje, los hiperparámetros a establecer fueron:

- Tasa de aprendizaje y decaimiento: se modificaron tanto la tasa de aprendizaje entre 0,001 a 0,0001, como el decaimiento entre 0,01 a 0,001.
- *Epochs*: en los modelos más básicos se comenzó con 100 *epochs* pero a medida que se iban complejizando se aumentó a 1000. Para algunos casos puntuales, se entrenó por 2000 o 3000 *epochs* (solo en los casos en los que las curvas de exactitud y pérdida mostraban potencial de crecimiento y decrecimiento respectivamente).
- *Batch Size*: se probó con 32, 64 y 128. En general, todos los modelos se entrenaron con el tamaño de 128.
- Cantidad de capas agregadas para realizar la nueva clasificación: se agregaron dos capas densas luego del modelo pre-entrenado
- Cantidad de neuronas y función de activación de cada capa densa agregada: se fue variando la cantidad de neuronas de la ante última capa densa agregada con 64 a 256 neuronas

Capítulo 4

Resultados

A continuación se muestran los resultados de las tres clasificaciones distintas analizadas, dos binarias y una categórica. Para cada una, se va a analizar el desempeño con las tres arquitecturas explicadas anteriormente (desde cero, VGG16 y Resnet 50) y luego, se va a elegir el mejor modelo de cada clasificación. El proceso es tal como se observa en la siguiente imagen:

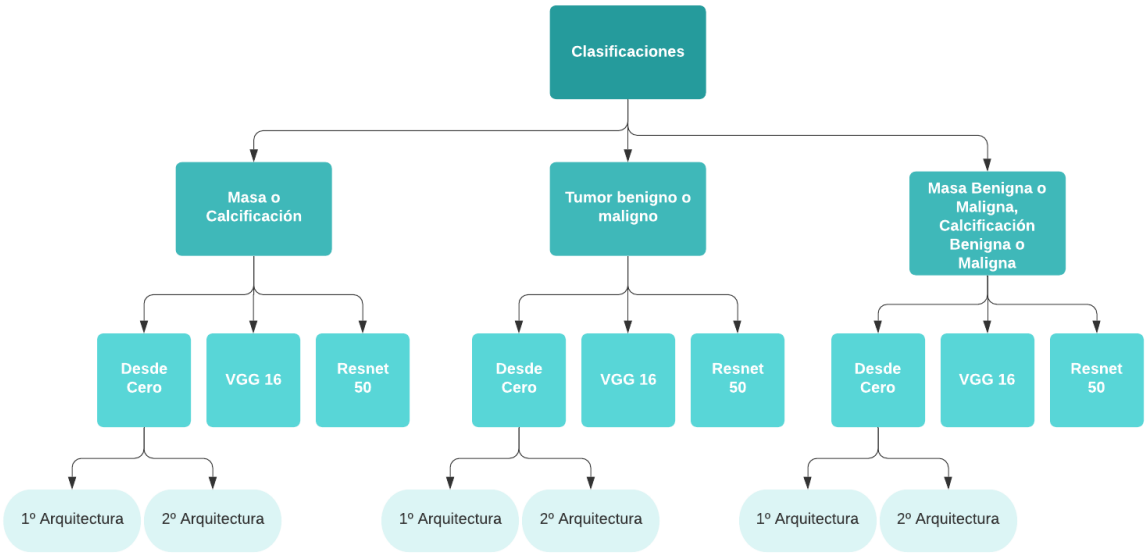

Figura 4.1: Esquema de las clasificaciones realizadas

Por último, a continuación se muestran los resultados tal cual explicados en el diagrama de la figura 4.1. Primero, se muestran los de la clasificación de masas o calcificaciones para la red diseñada desde cero (con sus dos arquitecturas), luego para la VGG16 y por último para la Resnet50. Una

vez analizados todos esos modelos, en la sección Discusión, se hace una comparación entre los mejores y se muestran las curvas ROC superpuestas y las matrices de confusión normalizadas. Luego, se hace lo mismo con las siguientes clasificaciones que son la de tumor benigno o maligno y la de masa benigna, masa maligna, calcificación benigna o clasificación maligna.

4.1. Clasificación: Masas o calcificaciones

4.1.1. CNN desde cero: Arquitectura 1

Como se mencionó anteriormente, primero se diseñó el Modelo 0 que es básico y se obtuvo una exactitud en el grupo de prueba de 0,8421, lo que resultó en un buen comienzo. Sin embargo, se observaba cierto *overfitting* en la curva de la pérdida. Por lo tanto, se le agregó una capa de *dropout* (Modelo 1) y se realizó el aumento de datos (Modelo 2) para poder contrarrestarlo. De esa forma, se mejoró la exactitud. A continuación, se hizo más profunda a la red lo que dio como resultado un mejor modelo (Modelo 3).

Finalmente, se realizaron diversos cambios a la arquitectura de la red para analizar si era posible, aumentar aún más el valor de la métrica. Sin embargo, ninguna de las pruebas realizadas pudo superar la exactitud del conjunto de prueba obtenido con el Modelo 3. Se hizo más grande a la capa densa con 48 neuronas (Modelo 4), se fue variando la tasa de aprendizaje a medida que el modelo se acercaba al mínimo local (Modelo 5), se hizo aún más profunda a la red (Modelo 6), se hizo más grande el kernel de la primera convolución (Modelo 7), se cambió el optimizador por el de Adam (Modelo 8), se hizo regularización (Modelo 9) y *batch normalization* (Modelo 10).

A continuación se puede observar una tabla en donde se muestra el número de modelo, el *epoch* óptimo y la exactitud y el error del grupo de prueba:

Modelo	***Epoch* óptimo**	**Exactitud Prueba**	**Pérdida Prueba**
0	46	0,8421	0,4992
1	52	0,8405	0,4271
2	84	0,8644	0,3649
3	**341**	**0,8931**	**0,3032**
4	80	0,8484	0,3687
5	157	0,8548	0,3687
6	160	0,8708	0,3451
7	136	0,8389	0,3759
8	434	0,8484	0,3583
9	378	0,8532	0,3671
10	61	0,8580	0,3477

Cuadro 4.1: Tabla de resultados de la primera arquitectura para la clasificación de masas o calcificaciones. El mejor modelo es el número 3

A continuación, se explica detalladamente como es la arquitectura de la red neuronal de cada modelo:

- Modelo 0: Se utiliza la red básica explicada anteriormente. Se usa el optimizador RMSprop, *Batch Size* de 32 y 100 *epochs*.
- Modelo 1: Al modelo 0, se le agrega una capa de *dropout* (con p = 0,5) entre las dos capas densas de la red.
- Modelo 2: Al modelo 1, se le hace el aumento de datos.
- Modelo 3: Al modelo 2, se le agrega antes de la primera capa densa, una capa convolucional con 128 neuronas (kernel 3x3) y luego un *Max Pooling* de 2x2. Además, a la capa densa que tenia 16 neuronas, se la modifica a 32. Por último, se aumenta el *batch size* a 128 y la cantidad de *epochs* a 500.
- Modelo 4: Al modelo 3, se le aumenta el tamaño de la primera capa densa a 48 neuronas.
- Modelo 5: Al modelo 3, se le agrega el decaimiento a la tasa de aprendizaje del optimizador con un valor de 0,001.
- Modelo 6: Se prueba a hacer aún más profundo al modelo 3, agregándole otra capa convolucional de 256 filtros (de 3x3) y un *Max Pooling* (de 2x2) previos a la primera capa densa.
- Modelo 7: Se agranda el filtro de la primera capa convolucional del modelo 3 a un tamaño de 5x5.

- Modelo 8: Se entrena el mejor modelo hasta ahora (modelo 3) con el optimizador Adam.
- Modelo 9: Se le agrega a cada capa convolucional del modelo 3, el regularizador L2 con un parámetro de 0,00005.
- Modelo 10: Se agrega *Batch Normalization* al modelo 3, después de cada capa convolucional, seguido de una función de activación no lineal (en este caso, una Relu).

La arquitectura del mejor modelo es:

```
Conv2D(32, (3,3), relu)
MaxPool2D((2,2))
Conv2D(64, (3,3), relu)
MaxPool2D((2,2))
Conv2D(128, (3,3), relu)
MaxPool2D((2,2))
Flatten()(x)
Dense(32, relu)
Dropout(0.5)
Dense(1, sigmoid)
```

A continuación se observan las curvas de exactitud y pérdida para el conjunto de entrenamiento y validación del Modelo 3 (mejor obtenido) con una exactitud de 0,8931.

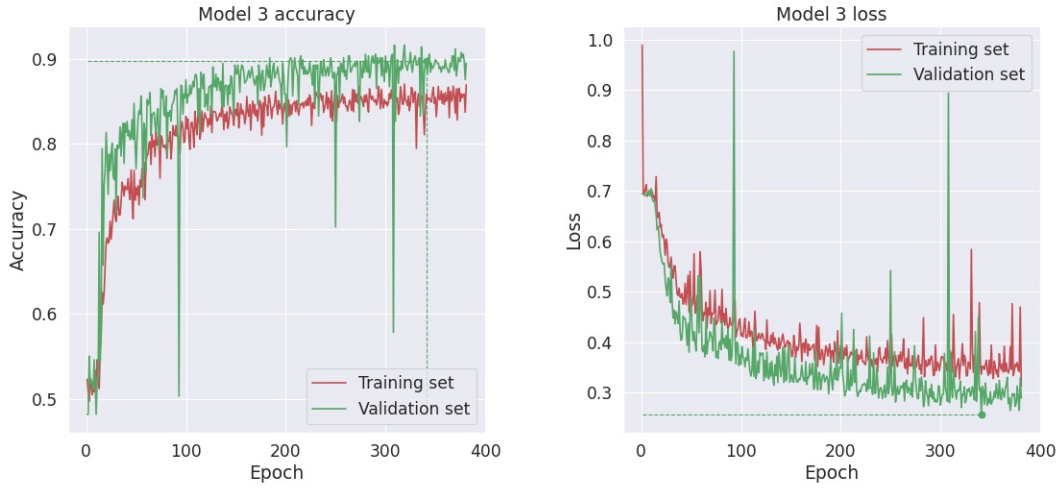

Figura 4.2: Curva de la exactitud (izquierda) y de la pérdida (derecha) para los grupos de entrenamiento y validación para el mejor modelo de la primera arquitectura para la clasificación de masas o calcificaciones. Con una línea punteada verde, se marca el *epoch* óptimo y su valor de exactitud y de pérdida

A continuación se observa la matriz de confusión (obtenida con el conjunto de prueba) del mejor modelo de esta arquitectura:

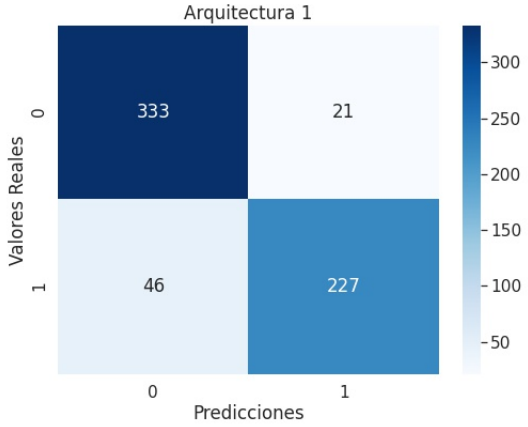

Figura 4.3: Matriz de confusión del mejor modelo de la primera arquitectura (Modelo 3) para la clasificación de masas (clase 0) o calcificaciones (clase 1) obtenida con el conjunto de prueba

Se observa que la diagonal está oscura, lo que se corresponde con el alto valor de la exactitud obtenida para el conjunto de prueba. Por lo tanto, con la primera arquitectura se obtuvo una exactitud alta con solo 21 casos de falsos positivos (la predicción fue calcificación cuando en realidad

era una masa) y 46 falsos negativos (la predicción fue una masa cuando en realidad era una calcificación).

4.1.2. CNN desde cero: Arquitectura 2

Para esta arquitectura, se continuó con la numeración de los modelos de la anterior; razón por la cuál, el primer modelo es el número 11. En este caso, primero se agregó la capa de *dropout* (Modelo 11) y luego se aumentaron los datos (Modelo 12) con lo que se obtuvieron mejores resultados. Posteriormente, se probaron diversas modificaciones pero ninguna logró una exactitud mayor. Se agregó el regularizador L2 y se cambió al optimizador por el de Adam (Modelo 13), se hizo a la red más profunda (Modelo 14) y se probó a realizar *Batch Normalization* (Modelo 15).

Por último, se entrenó nuevamente el Modelo 13 con una diferencia en la tasa de aprendizaje del optimizador (Modelo 16), obteniendo de esta manera, el mejor modelo hasta el momento. A continuación se observa la tabla con los resultados:

Modelo	***Epoch* óptimo**	**Exactitud Prueba**	**Pérdida Prueba**
11	43	0,8373	0,4189
12	49	0,8883	0,3077
13	490	0,8851	0,3173
14	875	0,8851	0,3012
15	481	0,8803	0,3987
16	**1.441**	**0,9011**	**0,3242**

Cuadro 4.2: Tabla de resultados de la segunda arquitectura para la clasificación de masas o calcificaciones. El mejor modelo es el número 16.

- Modelo 11: Se empieza con el modelo básico explicado anteriormente, agregándole la capa de *dropout* entre las dos capas densas. Se utiliza el optimizador RMSprop (tasa de aprendizaje de 0,001), un *batch size* de 128 y 500 *epochs*.
- Modelo 12: Al modelo 11, se le hace el aumento de datos
- Modelo 13: Al modelo 12, se le cambia el optimizador por el de Adam (tasa de aprendizaje 0,001) y a cada capa convolucional se le agrega un regularizador L2 con parámetro 0,0001.
- Modelo 14: Se hace más profundo al modelo 13, agregando dos capas convolucionales con 256 filtros de 3x3 y una capa de *Max Pooling* de 2x2 antes de la primera capa densa. Se entrena por 1000 *epochs* pero no se obtienen mejores resultados.

- Modelo 15: Al modelo 13, se le agrega *Batch Normalization* y una activación no lineal (Relu) luego de cada capa convolucional del modelo. No se obtienen mejores resultados.
- Modelo 16: Se diseña igual que el modelo 13, solo que reduciendo la tasa de aprendizaje del optimizador a 0,0001. Además, se entrena por 1500 *epochs*.

La arquitectura del mejor modelo es:

```
Conv2D(32, (3,3), relu, l2(0.0001))
Conv2D(32, (3,3), relu, l2(0.0001))
MaxPool2D((2,2))
Conv2D(64, (3,3), relu, l2(0.0001))
Conv2D(64, (3,3), relu, l2(0.0001))
MaxPool2D((2,2))
Conv2D(128, (3,3), relu, l2(0.0001))
Conv2D(128, (3,3), relu, l2(0.0001))
MaxPool2D((2,2))
Flatten()(x)
Dense(32, relu)
Dropout(0.5)
Dense(1, sigmoid)
```

A continuación se aprecian las curvas del mejor modelo obtenido:

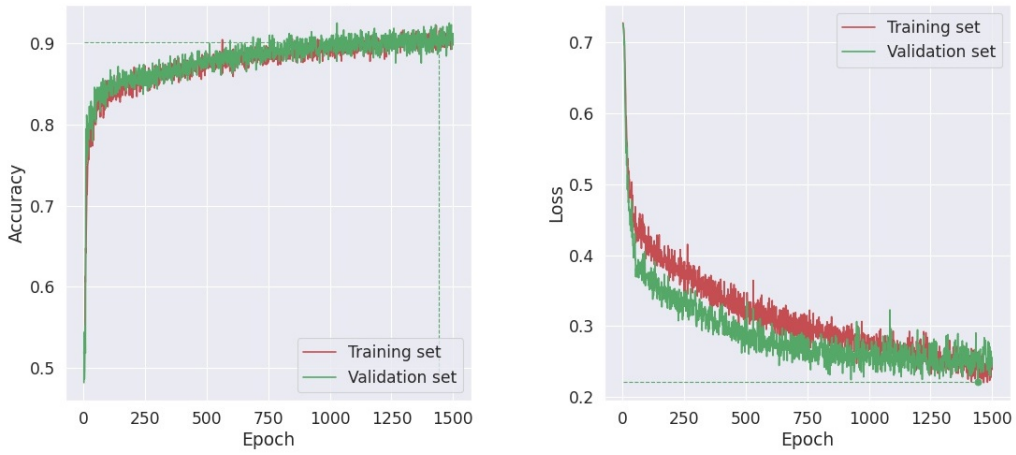

Figura 4.4: Curva de la exactitud (izquierda) y de la pérdida (derecha) para los grupos de entrenamiento y validación para el mejor modelo de la segunda arquitectura para la clasificación de masas o calcificaciones

A continuación, se observa la matriz de confusión del mejor modelo obtenida con el conjunto de prueba:

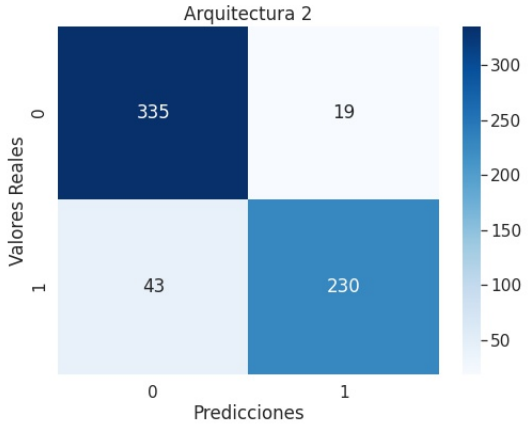

Figura 4.5: Matriz de confusión del mejor modelo de la segunda arquitectura para la clasificación de masas (clase 0) o calcificaciones (clase 1)

Cómo se puede apreciar, la exactitud del mejor modelo de esta arquitectura, supera al de la anterior. Además, es la que presenta mayor valor de verdaderos positivos y verdaderos negativos. Sin embargo, se observa la misma confusión en cuanto a la clasificación de las calcificaciones ya que existen 43 calcificaciones que las clasifica como masas; es decir, se obtuvieron 19 casos de falsos positivos y 43 falsos negativos.

4.1.3. VGG16

Se realizó *Transfer Learning* y eligió el mejor modelo obtenido para luego hacer *Fine Tuning*. Al igual que en la arquitectura anterior, primero se probó el modelo más básico posible (Modelo 1), obteniendo una exactitud de 0,7974, siendo este un buen comienzo. Luego, se agregó una capa de *dropout* (Modelo 2) y se disminuyó la cantidad de neuronas de la capa densa (Modelo 3), obteniendo una leve mejoría de la exactitud. A continuación, se probó potenciar aún más el aumento de los datos (Modelo 4), no obteniendo mejores resultados. Finalmente, se agregó el decaimiento de la tasa de aprendizaje del optimizador (Modelo 5) y luego, se probó a cambiar el optimizador por el de Adam (Modelo 6). A continuación, se muestran los resultados de realizar la transferencia de aprendizaje:

Modelo	***Epoch* óptimo**	**Exactitud Prueba**	**Pérdida Prueba**
1	29	0,7974	0,8112
2	32	0,7974	0,6618
3	**32**	**0,8198**	**0,5625**
4	68	0,7847	0,6161
5	181	0,8118	0,5574
6	311	0,8054	0,5711

Cuadro 4.3: Tabla de resultados luego de realizar la transferencia de aprendizaje de la VGG16 para la clasificación de masas o calcificaciones. El mejor modelo es el número 3

A continuación, se explica detalladamente que se hizo en cada modelo para realizar la transferencia de aprendizaje para la extracción de características.

- Modelo 1: Se entrenan las capas agregadas que realizan la clasificación, tal como se explicó anteriormente. Se utilizó el optimizador RMSprop, *batch size* de 128 y 500 *epochs*.
- Modelo 2: Al modelo 1, se le agrega una capa de *dropout* (p=0,5) entre las dos capas densas y se entrena por 200 *epochs*.
- Modelo 3: Al modelo 2, se le disminuye la cantidad de neuronas de la ante última capa densa a 128.
- Modelo 4: Al modelo 3, se le hace aún más el aumento de datos.
- Modelo 5: Al modelo 3, se le agrega el decaimiento de la tasa de aprendizaje al optimizador con un valor de 0,001.
- Modelo 6: Al modelo 3, se le cambia el optimizador por el de Adam con un decaimiento de la tasa de aprendizaje de 0,001.

La mejor arquitectura de la transferencia de aprendizaje es:

```
Modelo VGG16
Flatten()(x)
Dropout(0.5)
Dense(128, relu)
Dense(1, sigmoid)
```

Posteriormente, se realiza *Fine Tuning* del mejor modelo obtenido con la transferencia de aprendizaje. Se eligió el Modelo 3 para dicha tarea ya que posee el valor más alto de exactitud. Primero,

se realiza el ajuste fino de solo una capa de la red, obteniendo un modelo considerablemente mejor (Modelo 7). Luego, se hace *Fine Tuning* de una segunda capa (Modelo 8). A continuación, se observa la tabla con los resultados obtenidos luego de realizar el ajuste fino:

Modelo	***Epoch* óptimo**	**Exactitud Prueba**	**Pérdida Prueba**
7	**37**	**0,8740**	**0,3325**
8	15	0,8628	0,4382

Cuadro 4.4: Tabla de resultados luego de realizar ajuste fino de la VGG16 para la clasificación de masas o calcificaciones. El mejor modelo es el número 7.

A continuación se explica cada modelo detalladamente:

- Modelo 7: Se descongelan los pesos de la última capa convolucional de la VGG16 denominada *block5_conv3* del bloque 5 que posee 2.359.808 de parámetros. Luego, se vuelve a entrenar la parte del clasificador junto con esta nueva capa. Se usa el optimizador RMSprop y *batch size* de 128.

- Modelo 8: Se descongelan también los pesos de la ante última capa convolucional de la VGG16 denominada *block5_conv2* que también posee 2.359.808 de parámetros. Por lo que, en esta etapa, se suman 4.719.616 de parámetros entrenables a los del clasificador. Luego, se vuelve a entrenar al clasificador junto con las dos capas con los pesos descongelados (*block5_conv3* y *block5_conv2*)

Como se puede observar, al realizar el ajuste fino, se ve que aumenta la exactitud del modelo. De tal forma, que la mejor arquitectura, resulta ser la del Modelo 7. A continuación se muestran sus curvas:

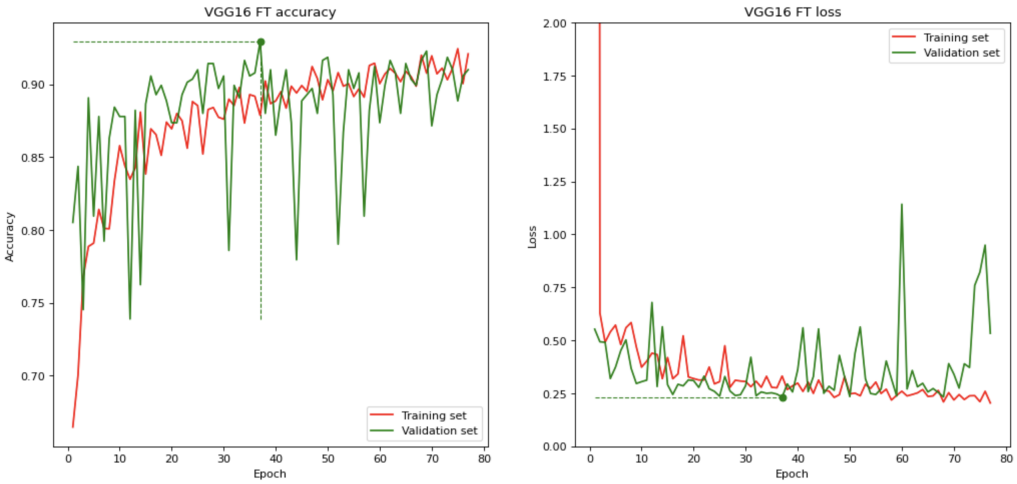

Figura 4.6: Curva de la exactitud (izquierda) y de la pérdida (derecha) para los grupos de entrenamiento y validación luego de realizar ajuste fino de la VGG16 para la clasificación de masas o calcificaciones

A continuación, se observa la matriz de confusión del mejor modelo, obtenida con el conjunto de prueba:

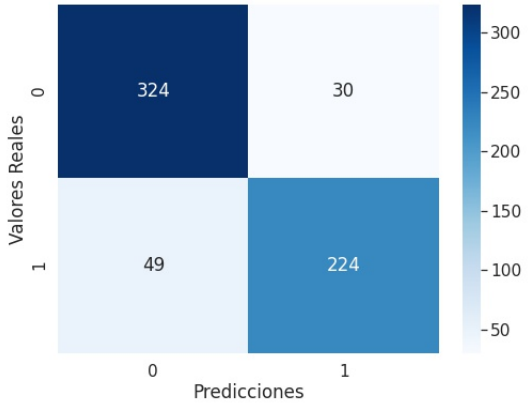

Figura 4.7: Matriz de confusión del mejor modelo de la VGG16 obtenido con el ajuste fino para la clasificación de masas (clase 0) o calcificaciones (clase 1)

Esta red presenta la misma confusión que las dos anteriores (las arquitecturas entrenadas desde cero) para la clasificación de las calcificaciones. Se observa que se obtuvieron 30 casos de falsos positivos y 49 falsos negativos.

4.1.4. Resnet 50

Por último pero no menos importante, se realizó *Transfer Learning* con la Resnet 50. Se obtuvieron resultados con un nivel de exactitud del conjunto de prueba significativamente menor respecto tanto a las redes neuronales anteriores como a las métricas publicadas en otras investigaciones. Razón por la cuál, se decidió no incluir la tabla con los resultados.

Primero, se entrenó la red como fue explicada anteriormente (Modelo 1, exactitud = 0,6220) en el que se utilizó la capa densa con 256 neuronas. Luego, se agregó la capa de *dropout* (Modelo 2, exactitud = 0,5136) y se probó a disminuir la cantidad de neuronas en la capa densa a 128 (Modelo 3, exactitud = 0,5646). Por último, se aumentó más los datos aplicando transformaciones a las imágenes (Modelo 4, exactitud = 0,6140). Por lo tanto, el mejor modelo en la etapa de *Transfer Learning*, resultó ser el primero. En consecuencia, se aplicó *Fine Tuning* al Modelo 1, obteniendo así el Modelo 5 con mejores resultados con una exactitud de 0,6699.

Existen diversas causas posibles por las que la Resnet 50 no funciona correctamente con las imágenes provistas, las cuáles pueden ser:

- La cantidad de imágenes disponibles para el entrenamiento es muy baja y la cantidad de parámetros a actualizar de la red es muy alta (más de 13.000.000)
- Posee muy pocas capas densas al final de la red con las que se hace la clasificación y por lo tanto, las que eliminamos para hacer *Transfer Learning*. Solo tiene una capa de *Average Max Pooling* y una *softmax* lo que provoca que la red no tenga tanta versatilidad en las imágenes a clasificar.
- La red fue entrenada con 1000 categorías de objetos distintos que son muy diferentes a las mamografías. Con lo que la extracción de características de por ejemplo, un barco puede no servir para la masa de una mamografía (esto se da también con la VGG16).

4.2. Clasificación: Tumores benignos o malignos

4.2.1. CNN desde cero: Arquitectura 1

Para esta clasificación binaria, se realizaron los mismos experimentos que en el caso anterior. Sin embargo, se obtuvieron valores de la exactitud más bajos. Primero se diseñó el Modelo 0 que es

básico y se obtuvo una exactitud en el grupo de prueba de 0,5423, lo que resultó muy bajo. Con el objetivo de mejorar dicho valor, se le agregó una capa de *dropout* (Modelo 1) y se realizó el aumento de datos (Modelo 2) para poder contrarrestarlo. De esa forma, se obtuvo una exactitud levemente mejor que la anterior.

A continuación, se realizaron diversos cambios a la arquitectura de la red para analizar si era posible aumentar el valor de la exactitud. Se hizo más profunda a la red (Modelo 3), más grande a la capa densa (Modelo 4), se agregó el decaimiento a la tasa de aprendizaje del optimizador (Modelo 5), se hizo aún más profunda a la red (Modelo 6), se hizo más grande el kernel de la primera convolución (Modelo 7), se cambió el optimizador por el de Adam (Modelo 8), se hizo regularización (Modelo 9) y *batch normalization* (Modelo 10).

Por lo tanto, el mejor modelo fue el número 4. A continuación se pueden observar todos los resultados obtenidos:

Modelo	***Epoch* óptimo**	**Exactitud Prueba**	**Pérdida Prueba**
0	33	0,5390	0,7186
1	70	0,3955	0,6947
2	24	0,6076	0,6923
3	46	0,4976	0,7341
4	**391**	**0,6379**	**0,6892**
5	151	0,5183	0,7043
6	312	0,5087	0,7151
7	164	0,5311	0,7112
8	96	0,5263	0,7045
9	346	0,4976	0,7087
10	182	0,5311	0,7154

Cuadro 4.5: Tabla de resultados de la primera arquitectura para la clasificación de tumores benignos o malignos. El mejor modelo es el número 4

En esta etapa, se hicieron los mismos experimentos que en la clasificación anterior, solo que se obtuvieron diferentes resultados. La explicación de cada modelo es:

- Modelo 0: Se utiliza la red básica con el optimizador RMSprop (tasa de aprendizaje = 0,001), *batch size* de 32 y 500 *epochs*.
- Modelo 1: Al modelo 0, se le agrega una capa de *dropout* (con p = 0,5).
- Modelo 2: Al modelo 0, se le agrega el aumento de datos.

- Modelo 3: Al modelo 2, se lo hace más profundo, agregando una capa convolucional con 128 neuronas y luego un *Max Pooling* de 2x2. Además, se aumenta el *batch size* a 128 y la cantidad de neuronas de la primera capa densa a 32. Se entrena por 500 *epochs* y no se obtienen mejores resultados.
- Modelo 4: Se mantiene la profundidad del modelo anterior solo que se le aumenta el tamaño de la primera capa densa a 48 neuronas y se entrena por 500 *epochs*.
- Modelo 5: Se le agrega al modelo 4, el decaimiento a la tasa de aprendizaje del optimizador con un valor de 0,001.
- Modelo 6: Al modelo 4, se lo prueba a hacer aún más profundo, agregándole otra capa convolucional de 256 filtros y un *Max Pooling* de 2x2.
- Modelo 7: Al modelo 4, se le agranda el filtro de la primera capa convolucional a un tamaño de 5x5.
- Modelo 8: Se entrena el mejor modelo hasta ahora (modelo 4) con el optimizador Adam.
- Modelo 9: Se le agrega a cada capa convolucional de la red (modelo 4), el regularizador L2 con un parámetro de 0,00005.
- Modelo 10: Se hace *Batch Normalization* después de cada capa convolucional, seguido de una función de activación no lineal (en este caso, una ReLU).

La arquitectura del mejor modelo es:

```
Conv2D(32, (3,3), relu)
Conv2D(32, (3,3), relu)
MaxPool2D((2,2))
Conv2D(64, (3,3), relu)
Conv2D(64, (3,3), relu)
MaxPool2D((2,2))
Conv2D(128, (3,3), relu)
Conv2D(128, (3,3), relu)
MaxPool2D((2,2))
Flatten()(x)
Dense(48, relu)
Dense(1, sigmoid)
```

A continuación se observan los gráficos del mejor modelo obtenido:

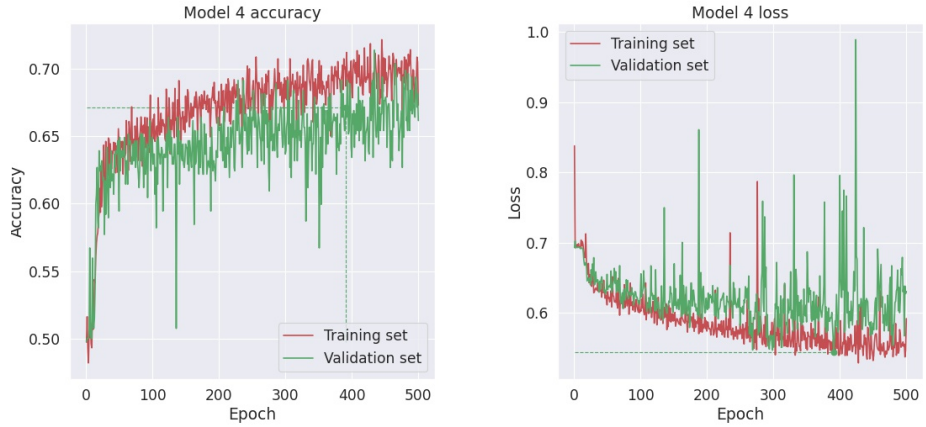

Figura 4.8: Curva de la exactitud (izquierda) y de la pérdida (derecha) para los grupos de entrenamiento y validación para el mejor modelo de la primera arquitectura para la clasificación de tumores benignos o malignos

Además, a continuación se observa la matriz de confusión de dicho modelo, obtenida con el conjunto de prueba:

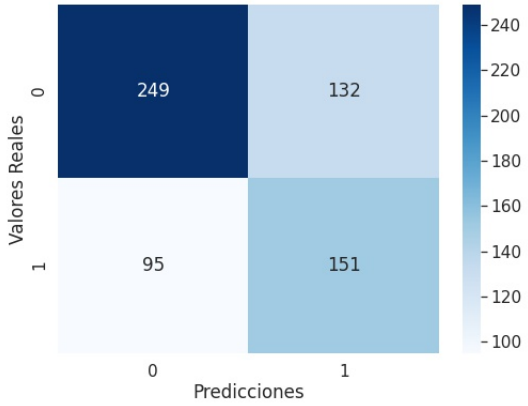

Figura 4.9: Matriz de confusión del mejor modelo de la primera arquitectura para la clasificación de tumores benignos (clase 0) o malignos (clase 1)

Se observa que el modelo confunde entre las clases ya que predice como clase 1 (tumor maligno) a muchas imágenes que son clase 0 (tumor benigno); es decir, existen 132 imágenes que tienen tumores benignos pero son clasificados como tumores malignos. Por lo tanto, se obtuvieron 132

casos de falsos positivos y 95 de falsos negativos. Un aspecto positivo de este modelo es que entre los dos errores que se pueden cometer, en este caso, el error de tipo II (FN) es menor que el error de tipo I (FP).

4.2.2. CNN desde cero: Arquitectura 2

Primero, se realizó el modelo con *dropout* (Modelo 11), luego se le hizo el aumento de datos con lo que se obtuvo una mejor exactitud (Modelo 12). Luego, se agregó el regulizador L2 (Modelo 13), se agregaron más capas convolucionales con *Max Pooling* (Modelo 14) y se hizo *Batch Normalization* (Modelo 15). Luego, se entrenó nuevamente el Modelo 14 con una diferencia en la tasa de aprendizaje del optimizador (Modelo 16). Por último, se entrenó el modelo 16 por más *epochs* (Modelo 17). A continuación se observa la tabla con los resultados:

Modelo	***Epoch* óptimo**	**Exactitud Prueba**	**Pérdida Prueba**
11	119	0,4864	0,7437
12	346	0,5120	1,2164
13	339	0,3923	0,6967
14	247	0,3923	0,6969
15	317	0,5578	0,6852
16	934	0,5470	0,7220
17	**1.967**	**0,5582**	**0,7204**

Cuadro 4.6: Tabla de resultados de la segunda arquitectura para la clasificación de tumores benignos o malignos. El mejor modelo es el número 17.

A continuación se muestran la explicación de cada modelo de forma detallada:

- Modelo 11: Se empieza con el modelo básico explicado anteriormente, agregándole la capa de *dropout* entre las dos capas densas. Se utiliza el optimizador RMSprop (tasa de aprendizaje = 0.001), un *batch size* de 128 y 500 *epochs*.
- Modelo 12: Al modelo 11, se le hace el aumento de datos
- Modelo 13: Al modelo 12, se le cambia el optimizador por el de Adam (tasa de aprendizaje = 0,001) y a cada capa convolucional se le agrega un regularizador L2 con parámetro 0.0001.
- Modelo 14: Se hace más profundo al modelo 13, agregando dos capas convolucionales con 256 filtros de 3x3 y una capa de *Max Pooling* de 2x2 antes de la primera capa densa. Se entrena por 1000 *epochs* pero no se obtienen mejores resultados.

- Modelo 15: Al modelo 13, se le agrega *Batch Normalization* y una activación no lineal (ReLU) luego de cada capa convolucional del modelo. No se obtienen mejores resultados.
- Modelo 16: Se diseña igual que el modelo 14, solo que reduciendo la tasa de aprendizaje del optimizador a 0,0001. Además, se entrena por 1500 *epochs*.
- Modelo 17: Es igual que el modelo anterior solo que se entrena por más *epochs*.

Por lo tanto, el mejor modelo resultó ser el 17. A continuación se observan las curvas del modelo.

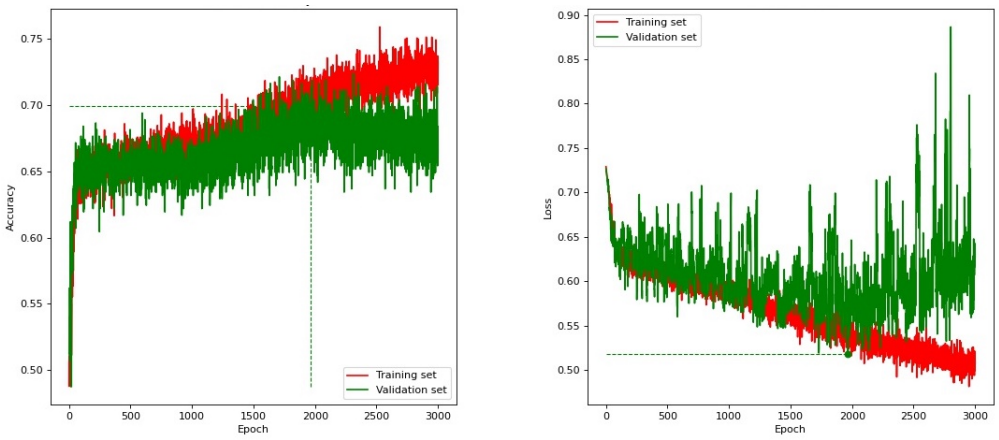

Figura 4.10: Curva de la exactitud (izquierda) y de la pérdida (derecha) para los grupos de entrenamiento y validación para el mejor modelo de la segunda arquitectura para la clasificación de tumores benignos o malignos

Además, a continuación se observa la matriz de confusión de dicho modelo, obtenida con el conjunto de prueba:

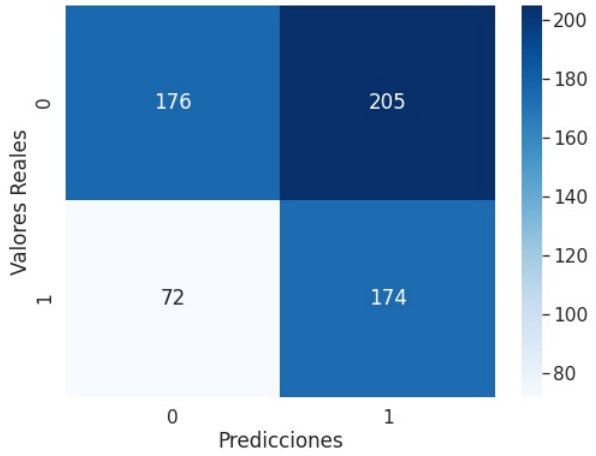

Figura 4.11: Matriz de confusión del mejor modelo de la segunda arquitectura para la clasificación de tumores benignos (clase 0) o malignos (clase 1)

Se obtuvieron 205 casos de falsos positivos y 72 de falsos negativos; lo que arroja un total de 277 imágenes con la predicción errónea. Como se puede observar, con este modelo se obtuvo un valor de falso negativo de 72, lo que resulta inferior que el falso negativo obtenido con la primera arquitectura que tenía un valor de 95. Sin embargo, este modelo presenta un menor valor de exactitud.

4.2.3. VGG16

Al igual que en la arquitectura anterior, primero se probó el modelo más básico posible (Modelo 1), obteniendo una exactitud de 0,6411 que resultó ser una de las más altas. Se hicieron cambios en la arquitectura suponiendo que iban a mejorar la exactitud del modelo pero no fue así. Se agregó una capa de *dropout* (Modelo 2) y se disminuyó la cantidad de neuronas de la capa densa a 128 (Modelo 3), no obteniendo mejoras. A continuación, se probó potenciar aún más el aumento de los datos (Modelo 4).

Como los resultados no eran los esperados, se cambió la tasa de aprendizaje y el decaimiento del optimizador para analizar si esto traería una mejora. Primero, se probó con el optimizador RMS-prop con una tasa de aprendizaje de 0,001 y un decaimiento de 0,001 (Modelo 5). Luego, se cambio la tasa de aprendizaje a 0,0001 (Modelo 6) obteniéndose una exactitud de 0,6602, la mejor hasta el momento. Posteriormente, se probó a cambiar el optimizador a Adam con una tasa de aprendizaje

de 0,001 (Modelo 7), luego se le cambió la tasa de aprendizaje y se le agregó el decaimiento (Modelo 8), no obteniendo mejores resultados. Por último, se le cambió la cantidad neuronas de la primera capa densa a 256 (Modelo 9), se modificó la tasa de aprendizaje al optimizador Adam (Modelo 10) y se probaron nuevas capas de transferencia de aprendizaje (Modelo 11). A continuación, se muestran los resultados obtenidos realizando transferencia de aprendizaje:

Modelo	***Epoch* óptimo**	**Exactitud Prueba**	**Pérdida Prueba**
1	9	0,6411	0,6904
2	60	0,6283	0,7566
3	31	0,5694	0,8249
4	346	0,6396	0,7100
5	164	0,6124	0,6754
6	**201**	**0,6602**	**0,6373**
7	285	0,5885	0,8627
8	301	0,6061	0,6861
9	355	0,6427	0,6609
10	164	0,6491	0,6185
11	49	0,5646	1,0514

Cuadro 4.7: Tabla de resultados luego de realizar transferencia de aprendizaje de la VGG16 para la clasificación de tumores benignos o malignos. El mejor modelo es el número 6.

El modelo básico con el que se comienza es:

```
Modelo VGG16
Dense(256, relu)
Dense(1, sigmoid)
```

Los modelos explicados detalladamente son:

- Modelo 1: Se hace la transferencia tal cuál fue explicada anteriormente. Se usa a RMSprop como optimizador y el *batch size* es 128.
- Modelo 2: Se le agrega capa de *dropout* antes de la primera capa densa con p=0.5.
- Modelo 3: Se hace más pequeña a la primera capa densa con 128 neuronas.
- Modelo 4: Se hace aún más aumento de datos; es decir, se modifica a la imágenes de manera más significativa. Se usa el optimizador RMSprop con tasa de aprendizaje de 0,0001.
- Modelo 5: Se le agrega el decaimiento de la tasa de aprendizaje al optimizador RMSprop (tasa de aprendizaje = 0,001 y decaimiento=0,001).

- Modelo 6: Se cambia la tasa de aprendizaje del optimizador RMSprop (tasa de aprendizaje = 0,0001 y decaimiento = 0,001).
- Modelo 7: Se utiliza el optimizador Adam (tasa de aprendizaje = 0,001).
- Modelo 8: Se le agrega el decaimiento al optimizador Adam y se cambia la tasa de aprendizaje (tasa de aprendizaje = 0,0001 y decaimiento = 0,001). La capa densa presenta 128 neuronas.
- Modelo 9: Se le cambia la cantidad de neuronas a la primera capa densa a 256.
- Modelo 10: Se varía la tasa de aprendizaje del optimizador Adam (tasa de aprendizaje = 0,01 y decaimiento = 0,001).
- Modelo 11: Como no se estaban obteniendo los resultados esperados, se probó a cambiar las capas agregadas al modelo que realizan la clasificación. Las nuevas capas para la transferencia de aprendizaje son: *GlobalAveragePooling2D*, capa densa (1024 unidades), otra capa densa (512 unidades) y una ultima capa densa (1 unidad). Se usó el optimizador RMSprop y *batch size* de 128.

La arquitectura del mejor modelo es:

```
Modelo VGG16
Flatten()
Dropout(0.5)
Dense(128, relu)
Dense(1, sigmoid)
```

En esta clasificación se obtuvieron valores de exactitud significativamente inferiores respecto a la clasificación anterior (masas o calcificaciones) y al estado del arte. Por lo tanto, en lugar de elegir sólo el mejor modelo de la tabla 4.2.3 para realizar el ajuste fino; se procedió a realizar el ajuste fino de todos los modelos obtenidos de la transferencia de aprendizaje. El objetivo fue ampliar la cantidad de modelos a los que se le realiza el *fine tuning* para poder tener más probabilidad de encontrar un modelo con mayor exactitud.

El mejor modelo obtenido luego de la transferencia de aprendizaje fue el número 6. Sin embargo, luego de realizar ajuste fino del modelo 4, se obtuvo el modelo 12 que presentó el mayor valor de exactitud. Por lo tanto, el modelo 12 es el que se eligió como el mejor modelo para la clasificación

de tumores benignos o malignos con la VGG16. A continuación, se muestra el resultado obtenido para el mejor modelo:

Modelo	***Epoch* óptimo**	**Exactitud Prueba**	**Pérdida Prueba**
12	8	**0,6715**	0,6873

Cuadro 4.8: Tabla con resultados luego de realizar ajuste fino de una capa del modelo 4 de la VGG16 para la clasificación de tumores benignos o malignos

La explicación del modelo es:

- Modelo 12: Igual que en el ajuste fino realizado en la clasificación anterior, se descongelaron los pesos de la capa llamada *block5_conv3* de la VGG.

Por lo tanto, el mejor modelo resulto ser el número 12 del ajuste fino. A continuación se muestran sus curvas:

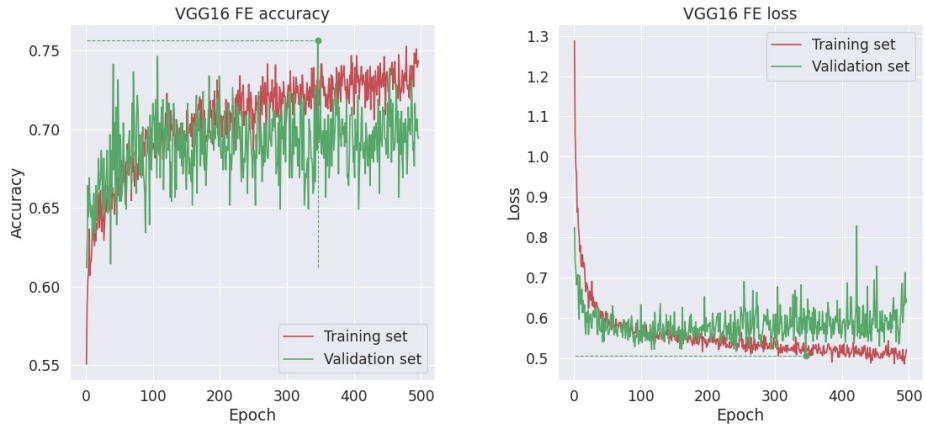

Figura 4.12: Curva de la exactitud (izquierda) y de la pérdida (derecha) para los grupos de entrenamiento y validación de la transferencia de aprendizaje de la VGG16 del modelo 4 para la clasificación de tumores benignos o malignos

Figura 4.13: Curva de la exactitud (izquierda) y de la pérdida (derecha) para los grupos de entrenamiento y validación del modelo 12 que es obtenido realizando el ajuste fino del modelo 4 para la clasificación de tumores benignos o malignos

Por lo tanto, el mejor modelo es el número 12. A continuación se muestra su matriz de confusión, obtenida con el conjunto de prueba.

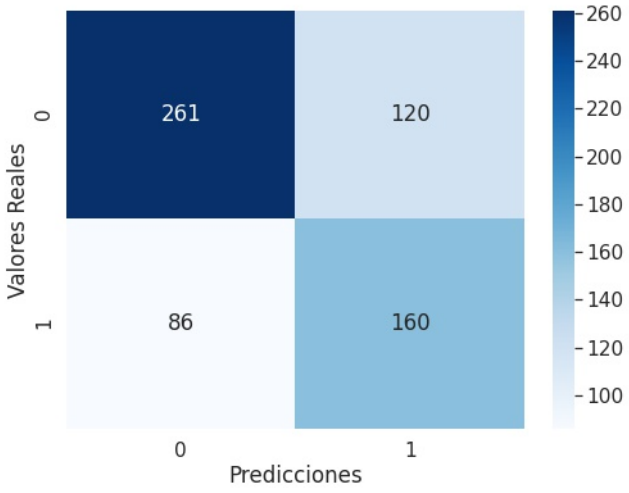

Figura 4.14: Matriz de confusión del mejor modelo con el ajuste fino de la VGG16 para la clasificación de tumores benignos (clase 0) o malignos (clase 1)

Como se puede ver en la matriz, existen 120 imágenes que poseen tumores benignos que fueron predichos como tumores malignos; es decir, se obtuvieron 120 casos de falsos positivos y 86 de falsos negativos.

4.2.4. Resnet 50

Al igual que en la clasificación anterior, con la Resnet 50 se obtuvieron resultados significativamente menores respecto a los modelos anteriores y al estado del arte. La mayor exactitud obtenida fue de 0,4450, razón por la cual no se incluyeron los resultados obtenidos de esta clasificación en el presente trabajo.

4.3. Clasificación: Masas benignas, Masas malignas, Calcificaciones benignas o Calcificaciones malignas

En esta clasificación, quedó un total de 1917 imágenes con la siguiente distribución:

- Masa Benigna es la clase 0 y representa un 26,08 %
- Masa Maligna es la clase 1 y representa 26,08 %
- Calcificación Benigna es la clase 2 y representa 26,08 %
- Calcificación Maligna es la clase 3 y representa 21,75 %

4.3.1. Camino 1

Se tomó el mejor modelo de cada clasificación binaria anterior para combinarlos y realizar la clasificación categórica. Se predijo la etiqueta con cada una de dichas redes; es decir, por un lado se clasificó a la mamografía en masa o calcificación y por el otro en benigna o maligna. Posteriormente, se unieron las etiquetas para la clasificación categórica. A continuación, se muestran los valores de las exactitudes de los mejores modelos de las dos clasificaciones binarias:

	Masa o Calcificación	**Tumor Benigno o Maligno**
Mejor Arquitectura	2º Arquitectura	VGG16
Mejor Modelo	16	12
Exactitud del mejor modelo	0,901	0,6715

Cuadro 4.9: Mejores modelos para cada clasificación binaria que se combinan para realizar la clasificación categórica.

El hecho de combinar los dos modelos binarios anteriormente entrenados, arroja una exactitud muy baja, de **0,2679**. Esto sucede porque para que la predicción realizada por el modelo categórico sea considerada correcta, las predicciones de los dos modelos binarios deben ser correctas. Como los

resultados no son los esperados, se diseñó una tercera CNN que realiza la clasificación entre las cuatro clases tal como se analiza a continuación.

4.3.2. Camino 2

Para este caso, se realizó lo mismo que en las clasificaciones binarias solo que cambiando a la función de activación de la última capa densa (sigmoidea) por una softmax con 4 unidades. Se entrenaron las dos arquitecturas diseñadas desde cero y se realizó la transferencia de aprendizaje con la VGG16 y la Resnet50, obteniendo como resultado una exactitud de 0,2743 y de 0,2855 respectivamente. Como sólo se obtuvieron resultados aceptables con las arquitecturas diseñadas desde cero, son los únicos resultados que se presentan a continuación.

CNN desde cero: Arquitectura 1

Para esta arquitectura, se realizaron los mismos experimentos que en las clasificaciones binarias. La mejor CNN fue la siguiente:

```
Conv(64, (3,3), relu)
MaxPooling2D((2, 2))
Conv(128, (3,3), relu)
MaxPooling2D((2, 2))
Conv(256, (3,3), relu)
MaxPooling2D((2, 2))
Conv(512, (3,3), relu)
MaxPooling2D((2, 2))
Flatten()
Dense(64, relu)
Dropout(0.5)
Dense(4, softmax)
```

Obteniendo como mejor resultado:

***Epoch* óptimo**	**Exactitud Prueba**	**Pérdida Prueba**
379	0,5502	0,9516

Cuadro 4.10: Tabla con el mejor resultado de la primera arquitectura para la clasificación categórica

Se observa que el valor de la exactitud del modelo categórico es extremadamente baja. Para poder

comparar los resultados obtenidos en esta clasificación con los de las binarias, se obtuvo el valor de la exactitud para la clasificación según tipo y según severidad de la lesión del modelo categórico, tal como se ve en la siguiente tabla.

Modelo	Categórico	Masa o calcificación	Tumor benigno o maligno
Exactitud	0,5502	0,8947	0,6156

Cuadro 4.11: Tabla con la exactitud obtenida en la clasificación categórica con la primera arquitectura y su posterior discriminación en las dos clasificaciones binarias

Se observa como la exactitud de ambas clasificaciones binarias son similares a los valores obtenidos en los dos análisis previos. A continuación, se ven las curvas de pérdida y exactitud del mejor modelo que clasifica entre las cuatro clases.

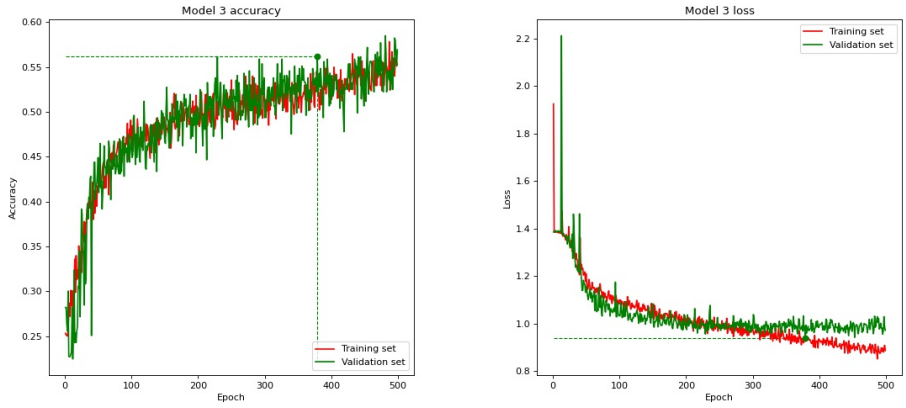

Figura 4.15: Curva de la exactitud (izquierda) y de la pérdida (derecha) para los grupos de entrenamiento y validación de la primera arquitectura para la clasificación categórica.

La matriz de confusión del mejor modelo, obtenida con el conjunto de prueba es:

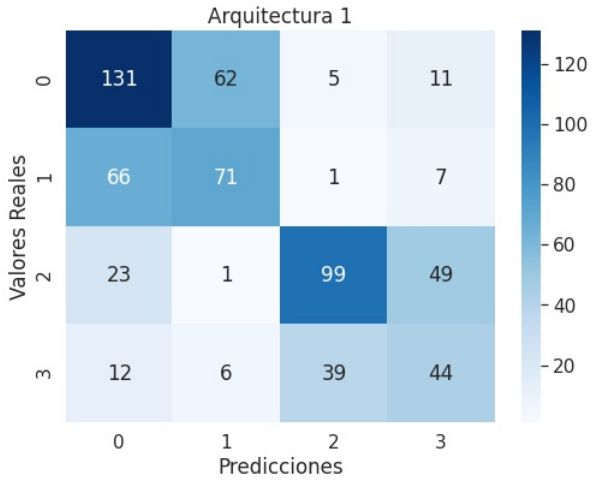

Figura 4.16: Matriz de confusión del mejor modelo de la primera arquitectura para la clasificación categórica: masa benigna (clase 0), masa maligna (clase 1), calcificación benigna (clase 2) y calcificación maligna (clase 3)

Como se puede ver en la matriz, el modelo presenta mucha confusión entre las clases 0 y 1, 2 y 3. Por ejemplo, predice que existen 66 imágenes que posen masas benignas cuando en realidad son masas malignas. Por lo tanto, se observa que la mayor confusión del modelo se presenta cuando debe discriminar entre tumor benigno o maligno.

CNN desde cero: Arquitectura 2

La mejor CNN de la segunda arquitectura fue la siguiente:

```
Conv(32, (3,3), relu)
Conv(32, (3,3), relu)
MaxPooling2D((2, 2))
Conv(64, (3,3), relu)
Conv(64, (3,3), relu)
MaxPooling2D((2, 2))
Conv(128, (3,3), relu)
Conv(128, (3,3), relu)
MaxPooling2D((2, 2))
Flatten()
Dense(32, relu)
Dropout(0.5)
```

```
Dense(4, softmax)
```

Obteniendo como mejor resultado:

***Epoch* óptimo**	**Exactitud Prueba**	**Pérdida Prueba**
1889	0,6012	0,9557

Cuadro 4.12: Tabla con el mejor resultado de la segunda arquitectura para la clasificación categórica

Se obtuvo una exactitud del modelo en su conjunto que clasifica en las cuatro categorías pero también se discriminaron las exactitudes para cada clasificación binaria en pos de poder compararla con los modelos anteriores tal como se observa en la siguiente tabla.

Modelo	Categórico	Masa o calcificación	Tumor benigno o maligno
Exactitud	0,6012	0,8883	0,6618

Cuadro 4.13: Tabla con la exactitud obtenida en la clasificación categórica con la segunda arquitectura y su posterior discriminación en las dos clasificaciones binarias

Las curvas del modelo son:

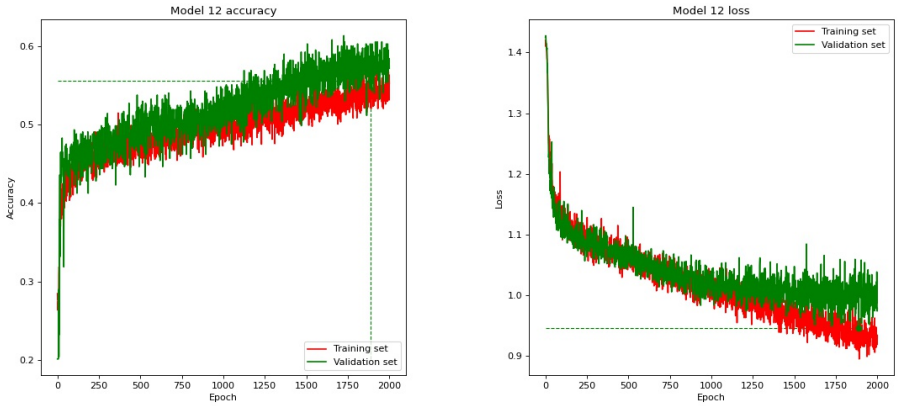

Figura 4.17: Curva de la exactitud (izquierda) y de la pérdida (derecha) para los grupos de entrenamiento y validación de la segunda arquitectura para la clasificación categórica.

La matriz de confusión del modelo, obtenida con el conjunto de prueba, es:

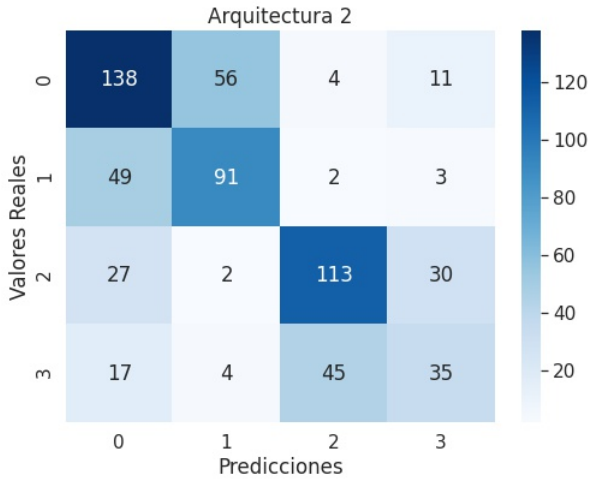

Figura 4.18: Matriz de confusión del mejor modelo de la segunda arquitectura para la clasificación categórica: masa benigna (clase 0), masa maligna (clase 1), calcificación benigna (clase 2) y calcificación maligna (clase 3)

Como se puede observar, la confusión en el modelo surge cuando el tumor es benigno o maligno. Por ejemplo, se ve que el modelo confunde la clase 0 (masa benigna) con la clase 1 (masa maligna). Por lo que, el modelo presenta alta exactitud en la clasificación entre masas y calcificaciones pero no puede discriminar con el mismo valor de la métrica entre tumores benignos y malignos.

Capítulo 5

Discusión

En la actualidad, cada vez más se utiliza inteligencia artificial en el campo de la salud con diferentes objetivos tales como para: la detección temprana de enfermedades, generar diagnósticos y crear sistemas de soporte a la decisión.

El objetivo de este trabajo es realizar una comparación entre arquitecturas de CNNs para las distintas clasificaciones de las lesiones en las mamografías. De esa forma, se busca sentar las bases para que en un futuro se pueda crear un sistema de soporte a la decisión con alta exactitud que pueda detectar el cáncer de mama de forma temprana. Como se mencionó anteriormente, durante muchos años se estudió cómo mejorar la sensibilidad y especificidad de las lecturas de las mamografías. Se hicieron numerosos estudios en los que se comparó el rendimiento de un solo lector respecto de varios lectores y se demostró que la doble lectura mejoraba el rendimiento. Sin embargo, en la práctica, no siempre se puede aplicar la doble lectura. Toda esta situación, motivó al desarrollo del presente trabajo que clasifica lesiones en las mamografías y le sirve al médico como un segundo lector. De esa forma, se busca mejorar la exactitud de la lectura y optimizar la realización de biopsias lo que se traduce en una reducción en el impacto al paciente y una mayor eficiencia para el hospital.

A continuación, con el fin de guiar la discusión, se presentan ciertos resultados tales como las curvas ROC superpuestas de los mejores modelos y las matrices de confusión normalizadas para cada clasificación.

5.1. Clasificación de masas o calcificaciones

Todas las predicciones realizadas en la sección de Resultados, se hicieron con un umbral de 0,5; es decir, los valores de la salida de la función sigmoidea menores a 0,5 se consideraron como clase 0 (masa) y los que eran superiores a 0,5 se consideraron como clase 1 (calcificaciones). Una vez obtenidos los mejores modelos para cada arquitectura, se analizó cuál era el umbral que mejor separaba a las clases. Por lo que, para la comparación que se realiza entre los modelos a continuación, se especifica tanto la exactitud con umbral 0,5 como la exactitud máxima que es la obtenida con el umbral que posee mayor separación entre clases.

	Arquitectura 1	**Arquitectura 2**	**VGG16**
Modelo	3	16	7
Exactitud con umbral 0,5	0,8931	0,9011	0,8740
Umbral que maximiza a la exactitud	0,50	0,40	0,65
Exactitud máxima	0,8931	**0,9074**	0,8787

Cuadro 5.1: Comparación de los mejores modelos de cada arquitectura para la clasificación de masas y calcificaciones. Se especifica el número de modelo, la exactitud (con umbral 0,5), el umbral que la maximiza y la exactitud en dicho umbral (máxima)

En la tabla se observa que el mejor resultado lo obtuvo el modelo de la segunda arquitectura con una exactitud de 0,907. A continuación, se observan las curvas ROC superpuestas de los mejores modelos obtenidos con cada arquitectura para la clasificación de masas o calcificaciones.

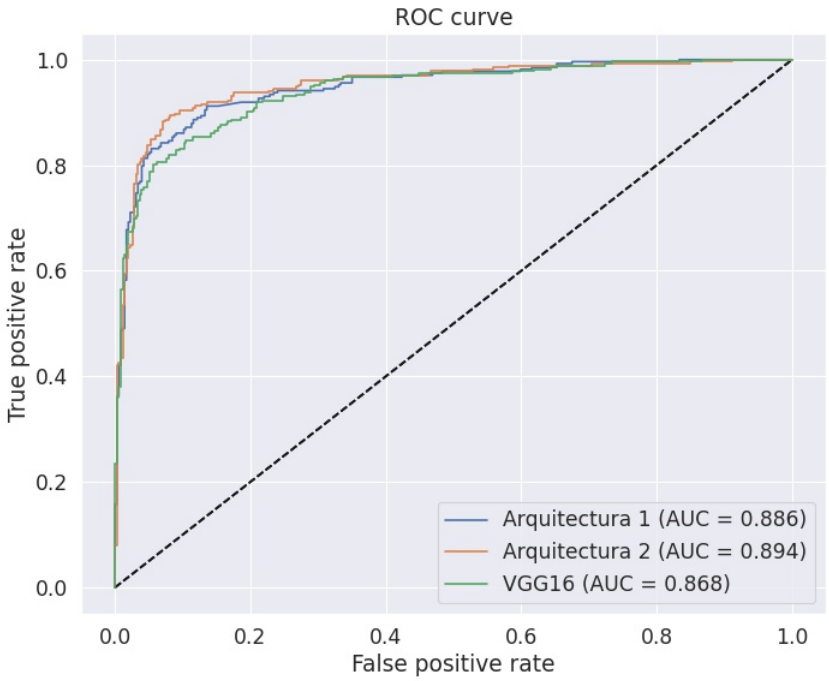

Figura 5.1: Curva ROC y valores de AUC de los mejores modelos para la clasificación de masas y calcificaciones

Como se puede apreciar, el modelo que presenta mayor exactitud y AUC para la clasificación de masas y calcificaciones, es la segunda arquitectura. Además, se observa que dicho modelo tiene una muy buena curva ROC, tal como se puede comparar en la figura 2.16. Por último, se muestran las matrices de confusión (normalizadas según los valores reales) de cada modelo:

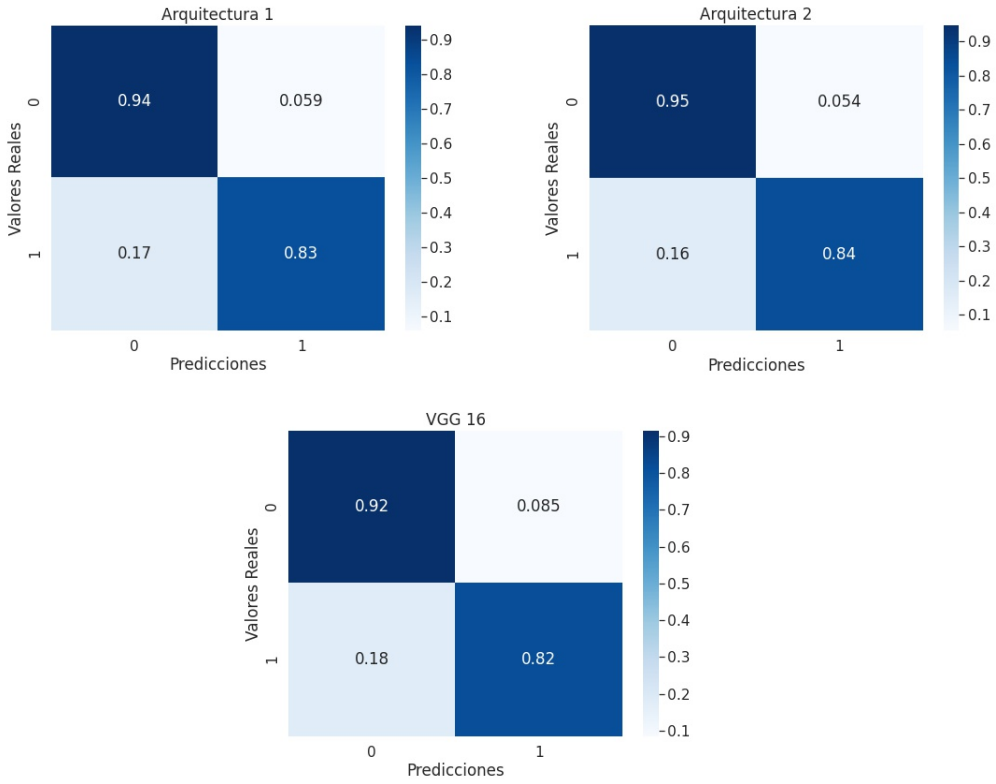

Figura 5.2: Matrices de confusión (normalizadas según las filas) de los mejores modelos obtenidos para la clasificación de masas (clase 0) y calcificaciones (clase 1) con un umbral de 0,5

Con la **primera arquitectura** se obtuvo una exactitud alta. En la figura 5.2 se puede apreciar la matriz de confusión con los valores normalizados por fila; es decir, según los valores reales. En la misma se observa cómo el modelo predice correctamente al 94 % de las masas y al 83 % de las calcificaciones.

Luego, se realizó la clasificación según tipo de lesión con la **segunda arquitectura**. Al igual que en el caso anterior, la matriz de confusión normalizada según las filas se ve en la figura 5.2 y se observa que el modelo predice correctamente al 95 % de las masas y al 84 % de las calcificaciones, siendo mejor que el modelo anterior.

Por último, se realizó la clasificación con el modelo obtenido de la **transferencia de aprendizaje con la VGG16**. En la figura 5.2 se observa que este último modelo es el que tiene más falsos negativos (error de tipo II, valor que se desea minimizar).

Por último, en la figura 5.2, se puede observar que todas las matrices de confusión son buenas ya que tienen la diagonal de color azul oscuro. Sin embargo, se ve que la segunda arquitectura es la que presenta mayor valor de VN y VP y menor valor de FN y FP. Por lo tanto, es la elegida como mejor arquitectura. A continuación, se observa una tabla que resume el valor de exactitud para la clasificación solo de las masas, solo de las calcificaciones y la total. Se observa que la segunda arquitectura es la que presenta mayor exactitud en cada una. Además, en los tres modelos para la clasificación de masas o calcificaciones existe una mayor exactitud al predecir a las masas.

Modelo	**Exactitud Masa**	**Exactitud Calcificación**	**Exactitud Total**
Arquitectura 1	0,9406	0,8315	0,8931
Arquitectura 2	**0,9463**	**0,8424**	**0,9011**
VGG16	0,9152	0,8205	0,874

Cuadro 5.2: Exactitud de la clasificación de masas y calcificaciones junto con la total con umbral en 0,5

Como conclusión, el mejor modelo para la clasificación entre masas y calcificaciones es el de la segunda arquitectura con una exactitud de 0,9074 (con un umbral de 0,4) o de 0,9011 (con un umbral de 0,5). Esto es así porque es el modelo que presenta mayor exactitud, mayor AUC y mejor matriz de confusión. Los resultados obtenidos, son comparables e incluso mejores que el estado del arte tal como se observa en la siguiente tabla.

Autor	Base de datos	Cantidad de imágenes	Método	Exactitud	AUC
Agarwal y Carson (2015) [32]	DDSM	2620	*Scratch based*	0,87	-
Pengcheng Xi et. al. (2018) [34]	CBIS-DDSM	3071	AlexNet,VGG, GoogLeNet y ResNet	0,925	-
Khan et al. (2019) [35]	CBIS-DDSM y MIAS	3890	VGG, Inception V3 y Resnet50	0,896	0,896
Modelo propuesto	**CBIS-DDSM**	**2960**	***Scratch based***	**0,9074**	**0,894**

Cuadro 5.3: Comparación del mejor modelo propuesto en el presente trabajo respecto de los métodos publicados en la actualidad, para la clasificación entre masas y calcificaciones.

La tabla 5.3 solo toma a las publicaciones que hacen referencia a la clasificación de masas y calcificaciones de la tabla 1.2 presentada en el Estado del Arte. Se puede apreciar que las publicaciones en la tabla 5.3, al igual que en el presente trabajo, utilizaron las bases de datos públicas (CBIS-DDSM, DDSM y MIAS) para realizar la clasificación entre masas y calcificaciones. Por lo tanto, la cantidad de imágenes para el entrenamiento de las redes neuronales, es acotada. Respecto a las arquitectura utilizadas en la tabla 5.3, se encuentran las CNNs: VGG, Resnet y arquitecturas entrenadas desde cero, tal como en este trabajo.

Por último, respecto a los resultados, se observa que el modelo de Pengcheng Xi et. al. [34] es el que presenta la mayor exactitud (0,925), seguido por el modelo propuesto en este trabajo con una exactitud de 0,9074. Luego, se encuentran las otras dos publicaciones, con una exactitud inferior a 0,90. Por lo tanto, el modelo presentado en este trabajo, es el segundo mejor modelo, tomando como métrica a la exactitud.

5.2. Clasificación de tumores benignos o malignos

En la tabla que se observa a continuación, se encuentra la comparación entre los mejores modelos para la clasificación de tumores benignos (clase 0) y tumores malignos (clase 1). En la misma se encuentra la exactitud con umbral 0,5, el umbral que posee mayor separación entre clases y la exactitud en dicho umbral.

Modelo	**Arquitectura 1**	**Arquitectura 2**	**VGG16**
Modelo	4	17	12
Exactitud con umbral 0,5	0,6379	0,5582	0,6715
Umbral que maximiza a la exactitud	0,70	0,75	0,65
Exactitud máxima	0,6682	0,6826	**0,6905**

Cuadro 5.4: Comparación de los mejores modelos de cada arquitectura para la clasificacion de tumores benignos o malignos. Se especifica el número de modelo, la exactitud (con umbral 0,5), el umbral que la maximiza y la exactitud en dicho umbral (máxima).

En la tabla se observa que la exactitud más alta se obtuvo con el modelo de la VGG16. Además, se ve que tanto los umbrales que maximizan a la exactitud como la diferencia entre las exactitudes (modificando el umbral), son mayores que en la clasificación anterior. A continuación, se observan las curvas ROC superpuestas de los tres modelos.

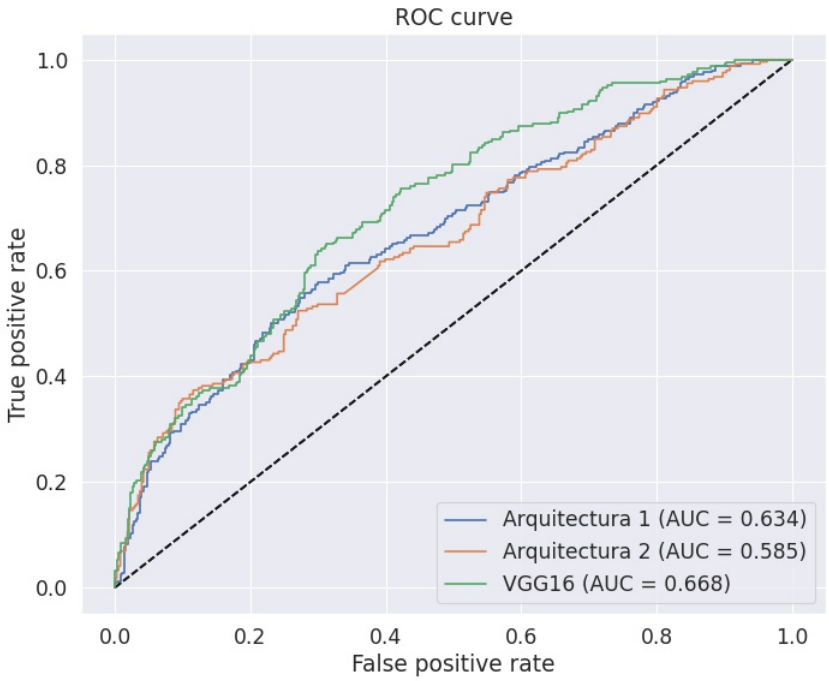

Figura 5.3: Curva ROC y valores de AUC de los mejores modelos para la clasificación de tumores benignos o malignos

Se observa que las áreas debajo de las curvas (AUC) son significativamente inferiores a las de la clasificación anterior y las curvas ROC no son tan buenas comparadas con la figura 2.16. Además, la arquitectura que mejor separa entre clases y la que posee mayor valor de exactitud, es la VGG16. A continuación, se observan las matrices de confusión normalizadas según las filas:

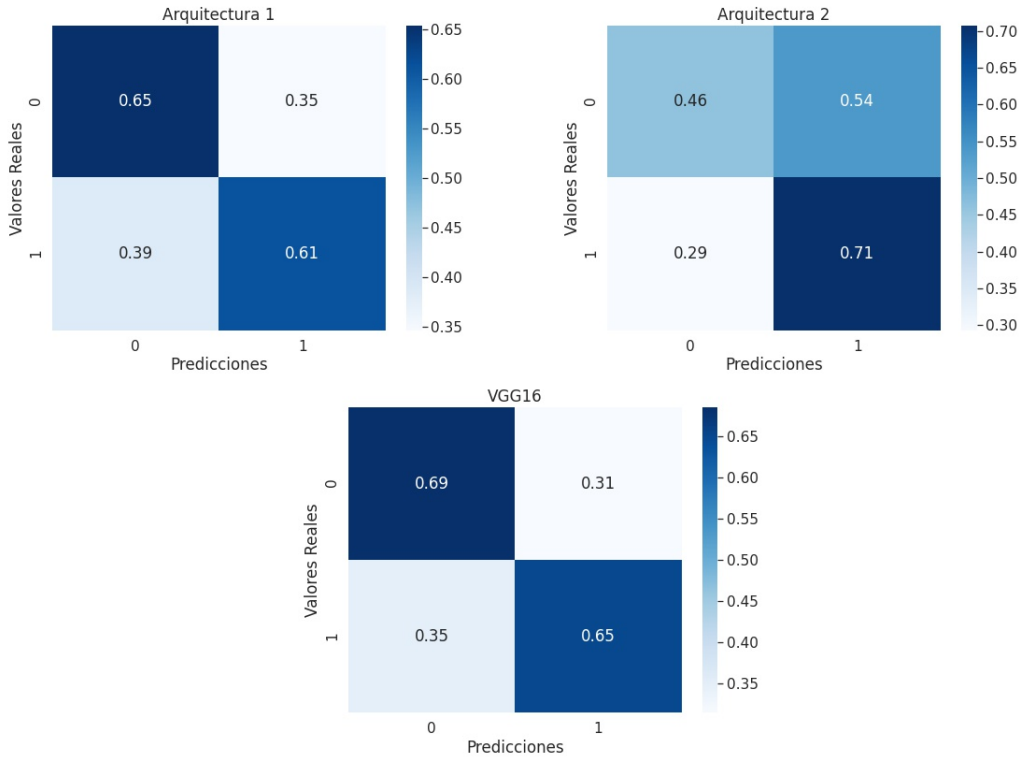

Figura 5.4: Matrices de confusión (normalizadas según las filas) de los mejores modelos obtenidos para la clasificación de tumores benignos (clase 0) y malignos (clase 1) con un umbral de 0,5

Para la **primera arquitectura**, al igual que en el caso anterior, en la figura 5.4 se observan las matrices de confusión, normalizadas según las filas, de los mejores modelos. En la misma, se aprecia que el modelo predice correctamente al 65 % de los tumores benignos y al 61 % de los tumores malignos.

Luego, para la **segunda arquitectura** ocurre algo particular: la cantidad de falsos positivos es mayor que la de verdaderos negativos. Por lo tanto, el modelo predice como malignos al 54 % de los tumores que realmente son benignos. Por lo que, el modelo confunde a las lesiones benignas. Sin embargo, al 71 % de las lesiones malignas las clasifica como tal y es el modelo que menor FN presenta. Por lo tanto, tiene dos características muy deseadas.

Por último, realizando **transferencia de aprendizaje con la VGG16**, se observa en la figura 5.4 que el 69 % de los tumores benignos y el 65 % de los malignos, son clasificados correctamente. Estos porcentajes son mayores comparados con los de la primera arquitectura. A continuación se

observa una tabla con los valores de exactitudes obtenidos para solo tumor benigno, solo tumor maligno y el total.

Modelo	**Exactitud Tumor Benigno**	**Exactitud Tumor Maligno**	**Exactitud Total**
Arquitectura 1	0,6535	0,6138	0,6379
Arquitectura 2	0,4619	**0,7073**	0,5582
VGG16	**0,6850**	0,6504	**0,6715**

Cuadro 5.5: Exactitud de la clasificación de tumores benignos y malignos junto con la total para cada arquitectura con umbral 0,5

En la tabla se observa como la red VGG16 es la que presenta mayor exactitud tanto en la total como en la clasificación de tumores benignos. Sin embargo, la segunda arquitectura es la que presenta mayor exactitud en la clasificación de los tumores malignos.

Como conclusión, el mejor modelo para esta clasificación es el de la transferencia de aprendizaje de la VGG16 ya que es el que presenta mayor exactitud total con un valor de 0,6905 (con umbral 0,65) o 0,6715 (con umbral 0,5). Además, este modelo es el que presenta mayor AUC y mejor matriz de confusión. Por último, se observa que el valor de exactitud de 0,6905 es comparable con algunos de los valores publicados en el estado del arte e inferiores a otros, tal como se observa en la tabla 5.6.

Autor	Base de datos	Cantidad de imágenes	Método	Exactitud	AUC
Huynh et al. (2016) [28]	Privada	607	AlexNet y GoogLeNet	-	0,86
Aboutalib et al. (2018) [29]	Privada y DDSM	14860	AlextNet	-	0,78
Hua Li et al. (2019) [30]	Privada	2042	AlexNet, VGGNet y GooLeNet	0,928	0,804
Cai et al. (2019) [31]	Privada	990	*Scratch based*	0,877	0,934
Agarwal y Carson (2015) [32]	DDSM	8752	*Scratch based*	0,69	-
Levy y Jai (2016) [33]	DDSM	1820	AlexNet y GoogLeNet	0,89	-
Khan et al. (2019) [35]	CBIS-DDSM y MIAS	3890	VGG, Inception V3 y Resnet50	0,754	0,746
Ragab et al. (2019) [36]	DDSM y CBIS-DDSM	7528	AlextNet	0,736	-
Modelo propuesto	**CBIS-DDSM**	**2642**	**VGG16**	**0,6905**	**0,727**

Cuadro 5.6: Comparación del mejor modelo propuesto en el presente trabajo respecto de los métodos publicados en la actualidad, para la clasificación entre tumores benignos y malignos.

La tabla 5.6 toma solo a las publicaciones que hacen referencia a la clasificación entre tumor benigno y maligno de la tabla 1.2 (presentada en el Estado del Arte). Para la clasificación entre tumor benigno o maligno, las investigaciones publicadas en la actualidad, se realizaron con diversas bases de datos tal como se observa en la tabla 5.6. Algunos investigadores utilizan bases de datos privadas y otros usan públicas tales como la CBIS-DDSM, MIAS o DDSM. Por tal motivo, la cantidad de imágenes para el entrenamiento del modelo, es variada. Por ejemplo, Aboutalib et al. [29] utiliza 14.860 imágenes mientras que Huynh et al. [28] usa 607 imágenes. Respecto a las arquitecturas utilizadas, la más popular es AlexNet, seguida de GoogLeNet. Sin embargo, también hay publicaciones que utilizan VGG y arquitecturas entrenadas desde cero, tal como en este trabajo.

Por último, respecto a los resultados, existe más diversidad comparado con la clasificación entre masas y calcificaciones. De las cuatro publicaciones que utilizan bases de datos públicas, dos utilizaron la base DDSM y las otras dos la CBIS-DDSM. La mayor exactitud (0,89) fue reportada por Levy y Jai [33] y la menor (0,69) fue reportada por Agarwal y Carson [32] (ambos usaron DDSM). Las dos publicaciones restantes, utilizaron la misma base de datos que el presente trabajo y obtuvieron exactitudes de 0,754 y 0,736. Por lo tanto, comparando nuestro trabajo con las investigaciones que utilizaron la base CBIS-DDSM, la exactitud obtenida en este trabajo es inferior. Sin embargo, los resultados obtenidos son comparables con los de Agarwal y Carson [32].

5.3. Clasificación categórica

Por último, se realizó la clasificación categórica según tipo y severidad de la lesión en la mama. Se tomaron dos caminos diferentes. En el **primer camino**, se combinaron las dos mejores CNN de las clasificaciones binarias, para realizar la clasificación categórica pero se obtuvo una exactitud extremadamente baja. Como los resultados no eran los esperados, se buscó una alternativa para esta clasificación.

Primero, para aumentar la exactitud del modelo, en lugar de combinar las dos mejores CNNs obtenidas en las clasificaciones anteriores, se entrenaron nuevas CNNs. La particularidad de estos nuevos entrenamientos es que se modificó el conjunto de datos ya que se dividió el conjunto de entrenamiento en dos: un grupo con 1513 imágenes solo de masas (1159 imágenes para el entrenamiento y 354 para la prueba) y otro grupo con 1130 imágenes solo de calcificaciones (857 imágenes de entrenamiento y 273 de prueba). Luego, se entrenaron dos redes neuronales, una para cada grupo. De esa forma, se pretendía que una CNN se centre solo en discriminar entre tumor benigno o maligno para las masas y la otra CNN que se centre en las calcificaciones. De esa manera, se buscaba hacer lo mismo que se hizo en la combinación de modelos, solo que el dataset para cada red, quedaba más pequeño. No se obtuvieron buenos resultados pero es posible que sea debido a la poca cantidad de imágenes para el entrenamiento que quedaron para cada red. Por lo tanto, si en el futuro se publica una base de datos con más imágenes, sería un buen primer camino, probar a realizar lo explicado anteriormente con el fin de mejorar la exactitud en la clasificación de la severidad de la lesión. Como no se obtuvieron resultados comparables con el estado del arte, se busco otra alternativa que se presenta a continuación.

Por lo tanto, en el **segundo camino**, se entrenó una tercera CNN que realiza la clasificación categórica. Para este red neuronal, se probaron las mismas arquitecturas que en las clasificaciones anteriores; es decir, las redes desde cero y transferencia de aprendizaje con la VGG16 y la Resnet50. Tanto con la VGG como con la Resnet50, se obtuvieron los mismos resultados que en el primer camino, razón por la cuál no se incluyeron los resultados en el trabajo. Sin embargo, con las redes entrenadas desde cero, se obtuvieron valores de exactitud más altos.

A continuación, se observan las matrices de confusión (normalizadas según las filas) de los mejores modelos para cada arquitectura entrenada en la clasificación categórica.

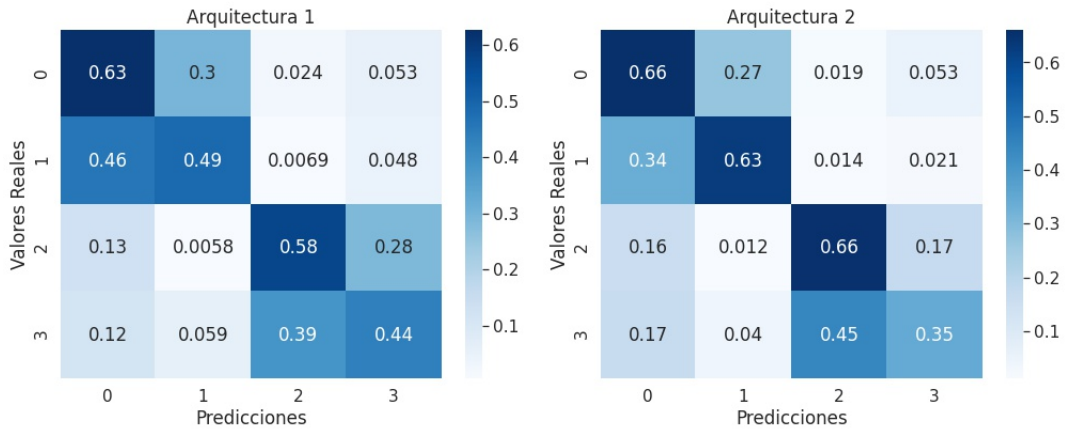

Figura 5.5: Matrices de confusión normalizadas de los mejores modelos obtenidos para la clasificación categórica: masa benigna (clase 0), masa maligna (clase 1), calcificación benigna (clase 2) y calcificación maligna (clase 3)

Por un lado, con la **primera arquitectura**, se obtuvo una exactitud del modelo categórico de 0,5502 lo que representa una exactitud de 0,8947 para la clasificación entre masas o calcificaciones y un 0,6156 para la clasificación entre tumor benigno o maligno. La matriz de confusión se puede ver en la figura 5.5. En la misma, se observa que son correctamente categorizadas el 63 % de las masas benignas (clase 0), el 58 % de las calificaciones benignas (clase 2), el 49 % de las masas malignas (clase 1), y el 44 % de las calcificaciones malignas (clase 3). Se observa que el modelo predice con mayor exactitud a las lesiones que son benignas respecto de las malignas. Además, se ve que el modelo confunde a la severidad de cada lesión. Por ejemplo, clasifica al 39 % de las calcificaciones malignas como benignas y al 46 % de las masas malignas como benignas, situación que se desea corregir.

Por otro lado, para la **segunda arquitectura**, se obtuvo una exactitud del modelo categórico de 0,6012 lo que representa una exactitud de 0,8883 para la clasificación entre masas o calcificaciones y un 0,6618 para la clasificación entre tumor benigno o maligno. Se concluye que disminuyó levemente la exactitud para la clasificación según el tipo de lesión pero aumentó significativamente para la clasificación según la severidad, respecto de la primera arquitectura. Como se puede ver en la figura 5.5, son correctamente categorizadas el 66 % de las masas benignas (clase 0) y de las calificaciones benignas (clase 2), el 63 % de las masas malignas (clase 1) y el 35 % de las calcificaciones malignas (clase 3). Se observa que el modelo clasifica mejor a las clases benignas que a las clases malignas. Además, al igual que el modelo anterior, confunde a las imágenes según la severidad ya que clasifica al 45 % de las calcificaciones malignas como benignas y al 34 % de las masas malignas como benignas.

Por último, a continuación se muestra una tabla en la que se especifica la exactitud obtenida con la clasificación categórica y además, se discrimina en la exactitud para masa benigna, masa maligna, calcificación benigna y calcificación maligna.

Modelo / Exactitud	**Masa benigna**	**Masa maligna**	**Calcificación benigna**	**Calcificación maligna**	**Exactitud Total**
Arquitectura 1	0,6267	0,4896	0,5755	**0,4356**	0,5502
Arquitectura 2	**0,6602**	**0,6275**	**0,6569**	0,3465	**0,6012**

Cuadro 5.7: Exactitud de la clasificación de masa benigna, masa maligna, calcificación benigna y calcificación maligna junto con la total del modelo que realiza la clasificación categórica

Se observa que la exactitud para las masas benignas, masas malignas y calcificaciones benignas es mayor la obtenida con la segunda arquitectura. Sin embargo, para la clasificación de las calcificaciones malignas la primera arquitectura presenta mayor exactitud. Como la mayor exactitud total es obtenida con la segunda arquitectura, es la red elegida como mejor modelo para la clasificación categórica con una exactitud de 0,6012.

Por último, para poder comparar los resultados obtenidos de los mejores modelos que realizan la clasificación binaria respecto de la categórica, se presenta la tabla 5.8. Para eso, se discriminó a la exactitud del mejor modelo categórico (exactitud de 0,6012) en las clasificaciones: masa o calcificación y tumor benigno o maligno. Luego, se agrega la exactitud del mejor modelo obtenido para la clasificación según tipo de lesión (exactitud de 0,9011) y según su severidad (exactitud

0,6715). De esa manera, se puede comparar la exactitud obtenida con los modelos que realizan las clasificaciones binarias respecto de la clasificación categórica.

Modelo / Clasificación	Masa o calcificación	Tumor benigno o maligno
Binario	**0,9011**	**0,6715**
Categórico	0,8883	0,6618

Cuadro 5.8: Comparación de la exactitud obtenida con el mejor modelo que realiza la clasificación categórica respecto de la obtenida con los mejores modelos de cada clasificación binaria

Se aprecia que tanto para la clasificación según el tipo (masa o calcificación) como para la severidad de la lesión (tumor maligno o benigno), los modelos con mayor exactitud son los que realizan la clasificación binaria. Además, comparando con el estado del arte, se observa que la exactitud obtenida del modelo categórico para la clasificación según tipo de lesión, es comparable con las presentadas en al tabla 5.3. Por último, se observa que la exactitud obtenida del modelo categórico para la clasificación según severidad de la lesión, es inferior a las presentadas en al tabla 5.6.

5.4. Comparación de todos los modelos

A modo de resumen, a continuación se presenta una tabla que compara las exactitudes obtenidas para cada una de las clasificaciones (tanto binarias como la categórica) según el modelo.

Clasificación	1º Arquitectura	2º Arquitectura	VGG16
Masa o calcificación	0,8931	**0,9011**	0,8740
Tumor benigno o maligno	0,6379	0,5582	**0,6715**
Categórica	0,5502	**0,6012**	0,2743

Cuadro 5.9: Tabla con los valores de exactitud obtenidos para cada clasificación realizada. Las exactitudes para las clasificaciones binarias, se expresan con umbral de 0,5

Se observa que para la clasificación según tipo de lesión (masa o calcificación) y para la categórica, los modelos con los que se obtuvo mejor exactitud son los que se diseñaron desde cero; en particular, con la segunda arquitectura. Además, se aprecia que se obtuvieron modelos con alta exactitud para la clasificación de masas o calcificaciones, comparables con resultados publicados tal como se muestra en la figura 1.2.

A continuación se observan diversas métricas obtenidas de los mejores modelos de cada una de las

clasificaciones realizadas.

Clasificación	Exactitud	Precisión	Recall	F1-score	AUC
Masas y calcificaciones	0,9011	0,9236	0,8424	0,8812	0,8944
Categórica (Masas y calcificaciones)	0,8883	0,9176	0,8168	0,8643	0,8801
Tumores benignos y malignos	0,6715	0,5714	0,6504	0,6083	0,6677
Categórica (Tumores benignos y malignos)	0,6618	0,5732	0,5406	0,5564	0,6404

Cuadro 5.10: Tabla con los valores de exactitud, precisión, recall, f1-score y AUC, obtenidas con los mejores modelos para cada clasificación realizada. Todas las métricas se expresan con umbral de 0,5

5.5. Limitaciones

Algunas de las limitaciones enfrentadas durante el desarrollo del proyecto fueron:

- Inexistencia de base de datos pública de mamografías que tenga un tamaño considerable para el uso de *Deep Learning*. Por consecuencia, la red en algunas clasificaciones, no termina de aprender y generalizar.
- Procesamiento en GPU. Como se ha mencionado anteriormente, el presente proyecto se realizó en Google Collaboratory que provee un uso gratuito limitado de la GPU a cada usuario. En muchas oportunidades se excedía el límite de uso gratuito y no se podía continuar con el entrenamiento de las redes neuronales. Por tal motivo, se tenía que esperar unas horas o incluso unos días para reanudar los entrenamientos de las CNNs lo que se traducía en retrasos en el proyecto.
- Tiempos de entrenamiento. Los tiempos de entrenamiento de las CNNs eran largos por lo que se demoraba mucho cada nueva modificación que se le hacía a un modelo para ver si se reflejaba en un aumento de la exactitud. Por ejemplo, cada modelo de las arquitecturas desde cero para la clasificación categórica tardó mínimo 2 horas en entrenarse y se corrieron 21 modelos lo que arroja un total de 42 horas de entrenamiento seguidas.

- Redes neuronales poco robustas debido a que solo fueron expuestas a pocas imágenes y de la misma base de datos.

5.6. Desafíos a futuro

El algoritmo diseñado en este trabajo, fue entrenado con parches de mamografías que contienen masas o calcificaciones. Por lo que, para que el médico pueda hacer uso del mismo, debe seleccionar un área sospechosa de la mamografía y recortarla para que luego sea evaluado por la red neuronal. Un desafío a futuro puede ser la automatización de la selección de las regiones de interés, para que el algoritmo detecte cuáles son las posibles áreas en la mamografía con lesiones y no lo tenga que hacer manualmente el médico.

Además, otro desafío a futuro podría ser generar más datos con la misma base de datos. Lo que se podría hacer es ubicar la región de interés en la mamografía y obtener el parche de la ROI. Luego, obtener cinco o seis parches más de tejido mamario que se solapen con la ROI pero no en un 100 %. De esa manera se podría obtener de una sola mamografía, un parche con la ROI (la que se usa actualmente en el presente trabajo) y cinco o seis parches más que contienen parte de la ROI y parte del fondo. De esa manera, se estaría generando aumento de datos y esto haría al modelo más robusto.

Por último, si se llegara a publicar una base de datos con más imágenes, se podría entrenar una CNN que reconozca tumor benigno o maligno solo en las calcificaciones y otra que reconozca lo mismo pero sólo en las masas. De esa manera, el algoritmo primero poseería una CNN que clasifica a la imagen en masa o calcificación. Posteriormente, existirán dos CNN adicionales: una que clasifica a la masa en benigna o maligna y otra que haga lo mismo pero con la calcificación. Esto no se pudo implementar en el trabajo debido a que se tenían muy pocas imágenes y el modelo no alcanzaba a generalizar.

Capítulo 6

Conclusiones

En este proyecto se crearon tres modelos distintos que clasifican lesiones en las mamografías: calcificaciones o masas; tumores benignos o malignos; y masa benigna, masa maligna, calcificación benigna o calcificación maligna. Se diseñaron dos arquitecturas de redes neuronales con un profundo análisis sobre su construcción y desempeño. Además, se realizó *transfer learning* de dos arquitecturas muy estudiadas que son la VGG16 y la Resnet50 para posteriormente realizar *fine tuning* de ambas.

El trabajo se realizó con la intención de generar un modelo robusto que clasifique a las lesiones en la mama tanto por su tipo (calcificación o masa) como por su severidad (benigna o maligna). Para la clasificación de tumores benignos o malignos, existe un amplio espectro de valores de exactitud publicados en la actualidad. El valor de exactitud obtenido en el presente proyecto (69 %) se encuentra dentro de dicho rango; es decir, es comparable con los resultados obtenidos por Agarwal y Carson [32] pero es significativamente más bajo que los publicados por Hua Li et al [30] tal como se observa en la tabla 5.6.

En la clasificación de calcificaciones o masas, se obtuvieron resultados prometedores ya que se alcanzó el 90 % de exactitud, valor que es comparable con las investigaciones publicadas por Pengcheng Xi et. al [34]. Incluso, la exactitud obtenida es más alta que algunas de las investigaciones publicadas en la actualidad, tal como se observa en la tabla 5.3.

Por lo tanto, en trabajos futuros, se propone incursionar en nuevos modelos o métodos de *Machine*

Learning para poder mejorar la exactitud de la clasificación de la severidad de la lesión en la mama. De esa manera, se podría construir un modelo final robusto que sea de gran ayuda para los médicos porque no solo actúa como un Sistema de Soporte a la Toma de Decisión (CDSS); sino que también, pueden llegar a detectar lesiones (como microcalcificaciones) que a simple vista no se puedan observar. El sistema siempre tiene como objetivo poder detectar el cáncer de mama en el estadío más temprano posible para poder tener el diagnostico y el tratamiento. De esa manera, poder reducir la tasa de mortalidad por cáncer de mama en las mujeres.

Capítulo 7

Anexo

7.1. Curva de tendencia de mortalidad por cáncer

A continuación se observa la tendencia de disminución de la muerte por cáncer de mama en los Estados Unidos.

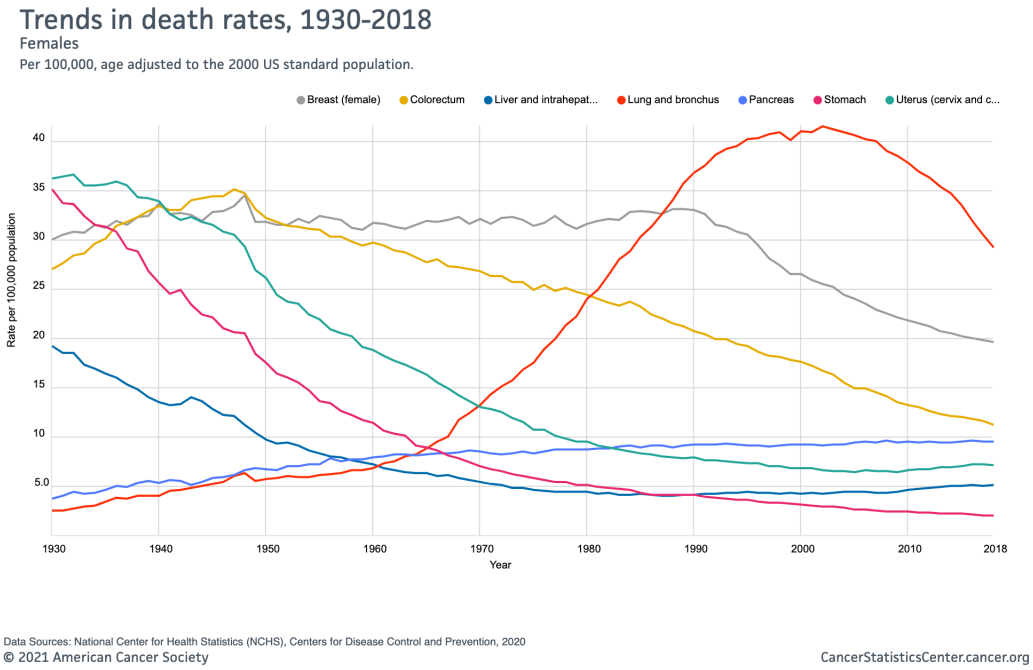

Figura 7.1: Tendencia de la mortalidad por cáncer en mujeres en Estados Unidos. Fuente: *American Cancer Society*

7.2. Anatomía y fisiología de la mama

La anatomía de la mama es como se puede observar en la siguiente imagen.

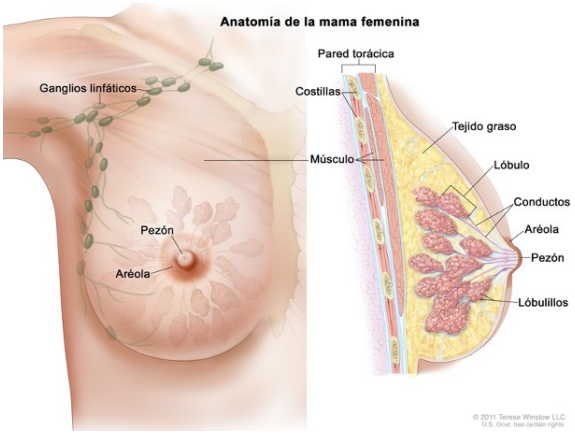

Figura 7.2: Anatomía de la mama

En el lado izquierdo, se encuentran las costillas, el músculo intercostal y los músculos pectorales (mayor y menor) quienes están conectados a la mama a través de tejido conectivo. Luego, en el centro de la mama, se ubica el tejido glandular que rodea a la glándula mamaria compuesta por 15 o 20 lóbulos que a la vez, están formado por lobulillos llamados alvéolos (la unidad funcional de la mama). Cada lobulillo está compuesto por una membrana, células mioepiteliales, células estromales y el ductulo. Además, se encuentran los ductos galactóforos cuya función es transportar la leche desde los lóbulos hasta el pezón.

Rodeando al tejido glandular, se encuentra el graso o adiposo llamado estroma (es la gran mayoría de la mama) y a los ligamentos de cooper que hacen de sostén a la mama. Por otro lado, también se encuentran los vasos linfáticos que transportan los desperdicios de las células hacia los nodos que se ubican en la zona de la axila.

En la imagen de la mama a continuación, se puede observar como los conductos se van ramificando hasta llegar a la unidad funcional de la mama llamada alveolo o acino. En los mismos, se produce la leche que luego llega a los pezones a través de los conductos.

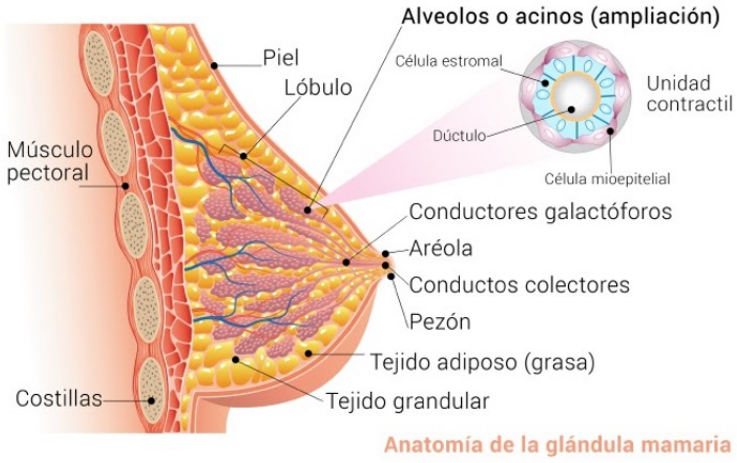

Figura 7.3: Anatomía de la mama junto con su unidad funcional

7.3. Cáncer de mama

Según el Instituto Nacional de Cáncer de Estados Unidos, el cáncer es un conjunto de enfermedades en las que células anormales se multiplican sin control y pueden tanto invadir los tejidos cercanos como diseminarse hacia otras partes del cuerpo a través de los sistemas sanguíneos y linfáticos. Los tres tipos principales de cáncer son: carcinomas, sarcomas y leucemias o linfomas.

El origen de las células cancerígenas es una mutación en el ADN; es decir, un error en la duplicación del ADN previo a la división celular. Dichas células se dividen y multiplican hasta formar el tumor que puede ser benigno (las células cancerígenas están rodeadas por tejido conectivo y no se diseminan a otros órganos) o maligno (las células anormales no se restringen a un área y se pueden propagar por todo el cuerpo). El cáncer de mama es un carcinoma con crecimiento no controlado de las células epiteliales de la mama. El tumor que se genera puede ser invasivo (cáncer) o no invasivo (in-situ). En general, se considera que si el tumor es in-situ (no invade otros órganos), no es cáncer.

7.3.1. In situ (no invasivo)

Dentro cáncer de mama no invasivo, existen dos grandes tipos de tumores el carcinoma ductal in situ (DCIS) y carcinoma lobulillar in situ (LCIS), también conocido como neoplasia lobulillar:

- **Carcinoma ductal in-situ (DCIS):** es la proliferación de células epiteliales mamarias en

los conductos. Es un precursor del cáncer invasivo, aunque no todos los DCIS progresan; de hecho a veces crece tan lentamente que incluso sin tratamiento no afectaría la salud de la mujer. Los estudios a largo plazo han encontrado que solo entre el 20 % y el 53 % de las mujeres con DCIS no tratado son finalmente diagnosticadas con cáncer de mama invasivo [55], [56].

- **Carcinoma lobulillar in-situ (LCIS):** también llamado neoplasia lobular, es la proliferación de células malignas en los lobulillos. Es una afección benigna asociada con un mayor riesgo de cáncer de mama, pero sin el potencial de progresar a un cáncer invasivo, por lo que se eliminó de la última edición del sistema de estadificación del cáncer de mama del AJCC . Aproximadamente el 30 % de las pacientes sometidas a escisión local adecuada de la lesión, desarrollan cáncer de mama en los siguiente 15 o 20 años. Por lo tanto, la neoplasia lobulillar es una lesión premaligna que sugiere un riesgo elevado de cáncer mamario subsiguiente, más que un cáncer en sí mismo.

7.3.2. Invasivo

La gran mayoría de los tipos de cáncer de mama son invasivos (80 %) lo que significa que las células anormales traspasaron las membranas de las glándulas o ductos en donde se originaron para crecer en el tejido circundante de la mama. La clasificación del cáncer de mama respecto a su biología molecular que hace referencia a la presencia o ausencia de expresión de ciertos receptores es (ER: receptor de estrógenos y PR: receptor de progesterona):

- **Luminal A (ER+/PR+/HER2-):** Este es el tipo más común de cáncer de mama y tiende a ser de crecimiento más lento y menos agresivo que otros subtipos. Los tumores luminales A se asocian con el pronóstico más favorable, en parte porque suelen responder a la terapia hormonal.
- **Luminal B (ER+/PR-/HER2+):** tienden a ser de mayor grado y más agresivos.
- **HER2 - amplificado (ER-/PR-/HER2+):** amplificación del gen HER2 en el cromosoma 17q. En el pasado, este subtipo tenía un mal pronóstico, pero el uso de terapias dirigidas para el cáncer HER2+ ha mejorado los resultados para estos pacientes.
- **Basal (ER-/PR-/HER2-):** llamados triple negativo, se caracterizan por la falta de expresión del receptor de estrógeno, progesterona y HER2 y tienden a ser de alto grado y son poco comunes.

7.3.3. Estadificación

Las etapas del cáncer de mama son:

Etapa	Tipo	Tasa de supervivencia a 5 años
0	Carcinoma ductal in situ o carcinoma lobulillar in situ	92 %
I	Carcinoma invasivo de 2 cm o menos de tamaño sin afectación ganglionar y sin metástasis a distancia	87 %
II	Carcinoma invasivo menor a 5 cm sin afectación ganglionar pero con ganglios axilares móviles y sin metástasis a distancia	75 %
III	Carcinoma invasivo de tamaño menor a 5 cm con afectación ganglionar y ganglios axilares fijos	46 %
IV	Cualquier forma de cáncer de mama con metástasis a distancia	13 %

Cuadro 7.1: Etapas del cáncer de mama con al supervivencia a 5 años en Estados Unidos. Fuente: *American Cancer Society*

7.4. Inteligencia Artificial

La Inteligencia Artificial (IA), es la capacidad de una computadora digital o un robot controlado por computadora para realizar tareas comúnmente asociadas con seres inteligentes [57]. El término se aplica frecuentemente al proyecto de desarrollar sistemas que posean los procesos intelectuales característicos de los seres humanos, como por ejemplo la capacidad de razonar, de descubrir significados, de generalizar o de aprender de experiencias pasadas. La IA es un área de la computación fundada oficialmente como una disciplina en el año 1956 en la Universidad de Dartmouth, en Nuevo Hampshire, Estados Unidos.

La IA se basa en programas con la habilidad de aprender y razonar como humanos. Por otro lado, el Aprendizaje Automático es un subcampo de la IA y son algoritmos con la habilidad de aprender sin ser específicamente programados para ello. A la vez, el Aprendizaje Profundo es un subcampo del Aprendizaje Automático en el que las redes neuronales artificiales se adaptan y aprenden de grandes cantidades de datos.

7.4.1. IA en el Hospital Italiano de Buenos Aires

En el área de Diagnóstico e Intervencionismo Mamario se realizan todas las prácticas y exámenes requeridos para el seguimiento y control de pacientes sin síntomas (screening) o para evaluar la

presencia de patología en la mama y/o axila (diagnóstico). Los estudios mamográficos se asignan diariamente a los médicos radiólogos para su informe, recibiendo cada uno entre 200 y 400 casos por mes. De esta forma, el hospital informa un promedio de 30.000 mamografías digitales anuales.

Además, el hospital cuenta con un sistema integrado RIS/PACS desde el año 2010, el cual les permite realizar análisis comparativos con estudios previos. Por otro lado, los médicos cuentan con la asistencia de un sistema de desarrollo propio de Inteligencia Artificial llamado "Artemisia" para la evaluación de la Densidad Mamográfica.

Una vez redactados los reportes de mamografía, se somete a revisión de pares el 10 % de los estudios informados por especialistas (aproximadamente 300 estudios mensuales). Además, se realizan auditorías de calidad de informes por parte del médico que solicitó el estudio.

El hospital realiza al año 25.000 estudios ecográficos, 500 resonancias magnéticas de mama y 600 biopsias mamarias bajo ecografía. Trabajan de manera integral junto a mastólogos, anatomopatólogos, radioterapeutas, oncólogos y cirujanos plásticos para lograr el enfoque interdisciplinario que requiere la subespecialidad.

7.4.2. Aplicaciones de IA en salud

Una de las aplicaciones más importantes en salud, es la **detección temprana de enfermedades** tal como un sistema de detección ocular llamado *EyeArt* [58] que detecta de forma temprana la retinopatía diabética; es decir, la ceguera diabética. Se envían las imágenes de los fondos de ojo realizados a los pacientes a la nube en dónde las evalúa un algoritmo de IA que informa si el paciente necesita o no, ser derivado para seguimiento.

Otra aplicación de la IA en salud es en el **dignóstico de enfermedades**. Un ejemplo es *Watson for health* de IBM [59], lanzado en el año 2015, con el objetivo de ayudar a los profesionales de la salud a resolver los desafíos más grandes del mundo en materia de salud con el uso de datos e inteligencia artificial. *Watson* puede revisar y almacenar mucha información médica exponencialmente más rápido que cualquier ser humano. Con la implementación de la computación cognitiva, Watson es capaz de analizar decenas de millones de piezas de datos en cuestión de segundos e identificar 300 terapias alternativas, lo que puede parecer un desafío para un equipo de médicos.

Por otro lado, otro campo importante de la IA es en los **sistemas de soporte a la toma de decisiones** que realizan un análisis predictivo que busca respaldar la toma de decisiones y acciones clínicas. El uso del reconocimiento de patrones para identificar a los pacientes en riesgo de desarrollar una afección, o ver cómo se deteriora debido al estilo de vida, factores ambientales, genómicos u otros, es otra área en la que la IA comienza a afianzarse en la atención médica. Por ejemplo, existen algoritmos de IA que clasifican la densidad mamaria según BI-RADS en A, B, C y D.

Por último, otra aplicación de la IA es en los **tratamientos**. Por un lado, se pueden escanear los registros de los pacientes para ayudar a los médicos a identificar a las personas con enfermedades crónicas que pueden estar en riesgo de sufrir un episodio adverso. Por otro lado, la IA puede ayudar a los médicos a coordinar los planes de atención y ayudar a los pacientes a manejar sus programas de tratamiento.

7.5. Machine Learning

El aprendizaje automático o *Machine Learning* es un subconjunto de la disciplina de inteligencia artificial y según Tom M. Mitchell se define como:

> "Es la disciplina que se dedica a la construcción de programas de computación que automáticamente mejoren con la experiencia"

Otra definición según Pedro M. Domingos es:

> "Es la disciplina que se dedica a investigar la forma de realizar tareas generalizando a partir de ejemplos"

Por lo tanto, se podría decir que es la capacidad que tienen los algoritmos de recibir un conjunto de datos y aprender del mismo por sí solos. De esa manera, cambian o ajustan parámetros a medida que procesan la información. Además, pueden hacer predicciones, por ejemplo, recomendarnos productos de forma personalizada tras analizar nuestras compras. Por su parte, el aprendizaje automático está formado por distintos tipos dependiendo de la respuesta o aprendizaje que se vaya a proporcionar.

El **aprendizaje supervisado** (*supervised learning*) se da cuando un algoritmo aprende a partir de variables que ya tienen un valor de verdad o etiqueta proporcionado por los seres humanos. Un ejemplo de este tipo de aprendizaje es cuando existen datos de un conjunto de pacientes que tienen

una enfermedad y se les suministró una droga. En este caso, los datos de entrada al algoritmo son: paciente, peso, altura, avance de la enfermad, etc; y la etiqueta es si se curó o no se curó. Por lo que es posible inferir si otro paciente que no está en el conjunto de entrenamiento se va a curar o no. Por lo tanto, el algoritmo es capaz de generalizar y clasificar de manera automática, sin nuestra intervención. Otro ejemplo de aprendizaje supervisado es el filtrado automático que realiza el mail, basado en información anterior que el usuario le haya ofrecido al sistema, que le permite categorizar correos como "bandeja de entrada" o "spam" (no deseado).

En el caso del **aprendizaje no supervisado** (*unsupervised machine learning*), no existe el supervisor; es decir, solamente hay datos de entrada y se aprende a través de evaluaciones alternativas. Un ejemplo son las asociaciones entre las compras de los clientes de una librería; es decir, si un cliente compra un libro se puede inferir que otros libros le podrían interesar. Por lo tanto, en este caso, el algoritmo no necesita que los datos estén etiquetados ya que su objetivo es encontrar relaciones por si mismo y así poder clasificarlos o agruparlos en categorías.

Por último, en el **aprendizaje por refuerzo** (*reinforcement machine learning*), el sistema recibe recompensas en función de las decisiones que toma. El objetivo es aprender a cómo relacionar situaciones con decisiones de modo tal de maximizar una función de recompensa.

7.6. Redes Neuronales

7.6.1. Redes neuronales biológicas

La neurona es la unidad funcional del cerebro y es una célula especialmente diseñada para transmitir información a otras células nerviosas, musculares o glándulas. La mayoría de las neuronas tienen un cuerpo celular, un axón y dendritas, tal como se observa en la imagen a continuación.

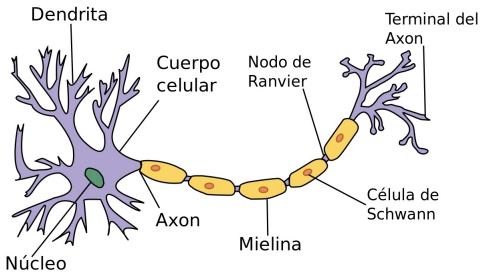

Figura 7.4: Esquema de las partes de una neurona

Las neuronas son células excitables; es decir, producen eventos eléctricos llamados potenciales de acción que les permiten comunicarse entre sí para procesar información. La función de cada una de éstas células es la de integrar toda la información que reciben de otras neuronas y por lo tanto, se empiezan a excitar. Si dicha excitación supera un umbral, la neurona empieza a disparar información en el cono axónico que viaja por el axón; evento conocido como despolarización. Cuando llega a la terminal del axón, se empieza a ramificar y llegar a árboles dendríticos de otras neuronas. Por lo tanto, las neuronas se comunican entre sí, enviando una señal eléctrica lo que hace que comience una reacción en cadena. Finalmente, una vez que el mensaje llega a su objetivo como un músculo o una glándula, el neurotransmisor es estimulado y actúa.

7.6.2. Perceptrón Simple

En 1957, Rosenblatt comenzó a trabajar en la primera red neuronal artificial conocida como el Perceptrón. Él estaba interesado en el funcionamiento del ojo de una mosca ya que se había descubierto que la mayor parte de la operación que le dice a dicho insecto si debe huir, se realiza en el ojo. La retina de una mosca contiene varios sensores de luz dispuestos como una matriz cuyas salidas están conectadas a un conjunto de elementos de procesamiento que reconocen patrones particulares. Las salidas de estos elementos de procesamiento van a una unidad lógica de umbral que luego se excita o dispara en función de un cierto nivel y tipo de entrada. Esto luego determina si una mosca huye o se queda quieta.

Por lo tanto, el Perceptrón o neurona artificial que creó Rosenblatt calcula una suma ponderada de las entradas. Realizando un paralelismo con el funcionamiento de una neurona fisiológica, las partes del perceptrón son las siguientes:

- **Entradas**. La información que llega a cada neurona; es decir, las entradas de la red, se

denominan X_i. Como existen muchas neuronas que transmiten información, se tiene el vector de entradas: $X = (X_1, X_2, X_3, ..., X_n)$

- **Pesos**. Como las neuronas pueden inhibir o excitar a otras, se le asignan pesos a las entradas que son los siguientes: $W = (W_1, W_2, W_3, ..., W_n)$. Si la sinapsis es excitatoria, el W es alto y positivo. Por lo contrario, si la sinapsis es inhibitoria, el W es alto y negativo.
- **Estado de excitación**. En el soma de la neurona, se suman todas las entradas y multiplican por su peso, tal como se muestra a continuación.

$$h = X \cdot W = \sum_{i=1}^{N} X_i * W_i \tag{7.1}$$

- **Estado de activación**. Cuando la excitación es mayor a un umbral, entonces la función de activación genera una salida.

Por lo tanto, a continuación se observa la arquitectura de un Perceptrón. Se puede apreciar el peso W1 multiplicado por la entrada X1 con W2 multiplicado por la entrada X2 y así sucesivamente. Luego se le agrega un sesgo b y se realiza la suma ponderada que alimenta a una función de activación para determinar si se dispara o no la neurona. Es un clasificador binario porque si supera cierto umbral, devuelve uno, de lo contrario devuelve un cero. Es necesario que la red aprenda los valores de b y W y cuantifique el error.

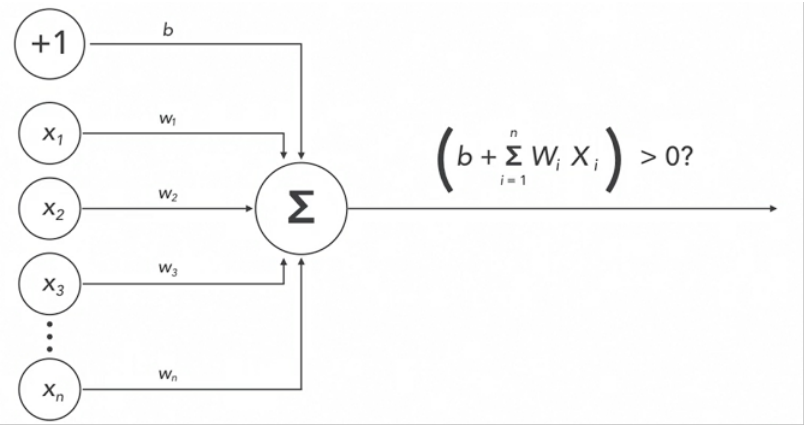

Figura 7.5: Esquema del Perceptrón Simple.

Con un problema de clasificación, para una entrada determinada sabemos cuál es la salida real y cuando pasamos la misma entrada a nuestro modelo obtenemos una salida predicha. A la diferencia

entre estas dos salidas la llamamos pérdida (*loss*) y, para mejorar la capacidad de predicción de nuestro modelo, queremos minimizarla.

7.6.3. Perceptrón Multicapa

Como se hizo en el caso del Perceptrón Simple, se desean cambiar las predicciones actualizando los pesos dentro del modelo. Existen limitaciones al usar un solo perceptrón, ya que la salida solo puede ser una combinación lineal de las entradas por lo que se necesita introducir la no linealidad a la red neuronal. Para hacer la extención de una sola neurona a una red neuronal, se define un perceptrón multicapa (*Multilayer Perceptron*), como aquel en el que los perceptrones se dividen en varias capas; es decir, la salida de un perceptrón será la entrada a otro perceptrón. Esto se conoce como una capa completamente conectada (*Fully connected layer*) en la que cada capa procesa todas las salidas de la capa anterior.

Cada uno de los perceptrones del diagrama a continuación, se comporta como un Perceptrón Simple. En la imagen se observa una red neuronal en la que la entrada a este perceptrón es una combinación lineal de todas las salidas multiplicadas por el peso, indicado por Wij, de los perceptrones de la capa anterior. La salida de este perceptrón está conectada a todos los perceptrones en la siguiente capa. En este caso, se introduce la no linealidad asegurándose de que la función de activación que se usa (la función que determina si la neurona se activa o no), no sea lineal. Por último, se observa una capa de entrada y una capa de salida, y si tiene capas intermedias, como se muestra en este diagrama, esto se conoce como capa oculta. En la imagen, solo se ve una capa oculta, pero en realidad podría haber múltiples. Una red neuronal profunda es entonces una red neuronal con una o más capas ocultas.

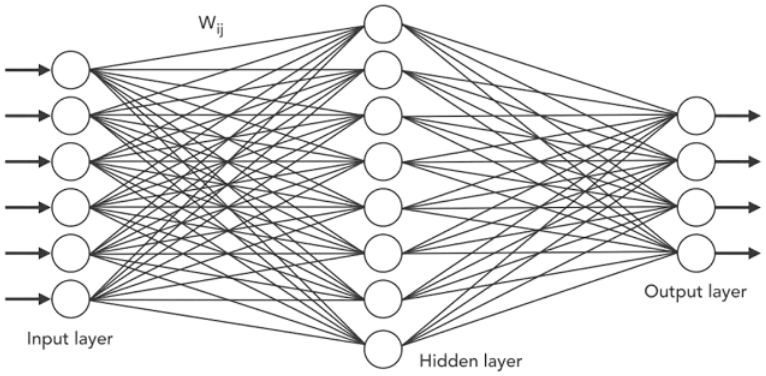

Figura 7.6: Esquema del Perceptrón Multicapa.

Por lo tanto, realizando la analogía con las neuronas biológicas, una red neuronal está formada por nodos separados llamados neuronas que a su vez, están organizadas en una serie de grupos llamados capas. Los nodos de cada capa están conectados a los nodos de la siguiente capa. Los datos fluyen de la entrada a la salida a lo largo de estas conexiones. Cada nodo individual está capacitado para realizar un cálculo matemático simple y luego enviar su resultado a todos los nodos a los que está conectado. Por lo tanto, primero la red neuronal toma un conjunto de valores de entrada en la capa de entrada y luego esos valores pasan por todas las capas siguientes. Finalmente, cada nodo ajusta ligeramente el valor que recibe y pasa su resultado al siguiente nodo.

7.6.4. Funciones de activación

Como se discutió en el caso del Perceptrón, la neurona artificial hace los cálculos de sumas ponderadas y determina si se dispara o no. La función de activación se usa para determinar la salida de la red neuronal y se dividen en lineales y no lineales. Para asegurarnos de que exista la no linealidad en la red, debemos incorporar funciones de activación no lineales. Existen diversas funciones como las que se observan a continuación:

- **La función escalón** que proporciona una salida de 0 o 1. Si la salida está por encima de cierto umbral, entonces se dispara la neurona y se obtiene un uno. Si el valor de la salida es menor que el umbral, entonces no se dispara y se obtiene un cero.

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$$

- **La función sigmoidea** o también llamada la función logística, la salida está entre 0 y 1, lo

que resulta útil para clasificaciones binarias. El problema es lo que se conoce como gradiente de fuga (*Vanishing gradient*) que cerca de los límites, la red no aprende rápidamente y esto se debe a que la pendiente es casi nula en ambos extremos. La ecuación es:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{7.2}$$

- **La función softmax** es muy similar a la sigmoidea. Mientras la softmax opera sobre un vector, la sigmoidea lo hace sobre un escalar. De hecho, la función sigmoidea es un caso especial de la función softmax para un clasificador con solo dos clases de entrada. La ecuación es la siguiente, en donde $\vec{z}$ es el vector de entrada y C es la cantidad de clases.

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \tag{7.3}$$

- **La tangente hiperbólica** es otra posible función de activación no lineal. La tanh se define en términos de exponentes y tiene una salida entre -1 y 1. Como una función tanh puede estar formada por sigmoideas, también está presente el problema del gradiente de fuga. La ecuación es:

$$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} = \frac{2}{1 + e^{-2x}} - 1 = 2 \cdot sigmoid(2x) - 1 \tag{7.4}$$

- **La ReLU** (*Rectified Linear Unit*) devuelve un 0 para cualquier valor de x que sea menor que cero y x para el caso contrario tal como se ve a continuación.

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases}$$

A continuación se aprecian las curvas de las funciones de activación.

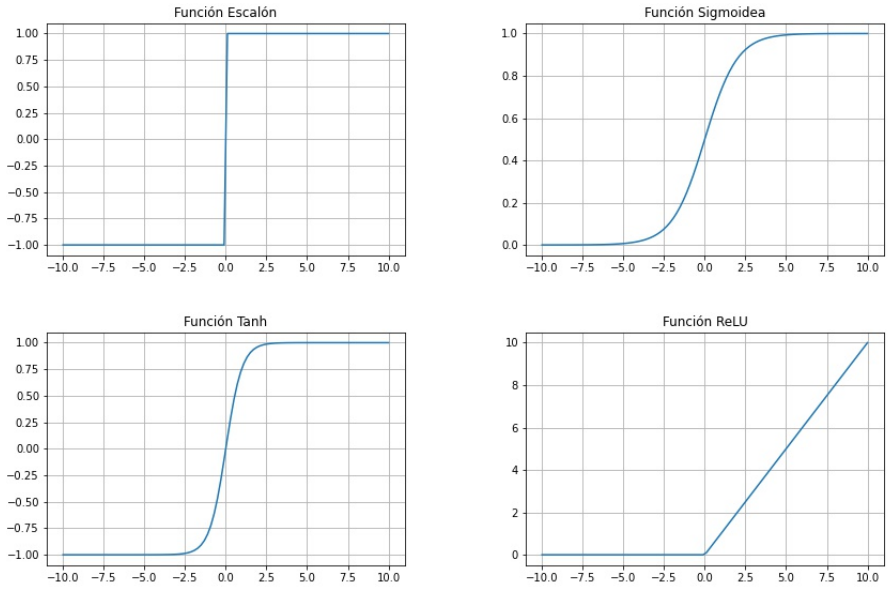

Figura 7.7: Distintas funciones de activación: escalón, sigmoidea, tangente hiperbólica y ReLU.

7.6.5. Función de pérdida: Entropía Cruzada

Una de las funciones de pérdida utilizada en el presente proyecto es la ***Binary Cross-Entropy*** que se utiliza para las clasificaciones binarias. Solo se necesita un nodo de salida para clasificar los datos en dos clases. El valor de salida debe pasar a través de una función de activación sigmoidea por lo que el rango de salida es entre cero y uno. La función de pérdida es:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(\hat{y_i}) + (1 - y_i) \cdot log(1 - \hat{y_i}) \tag{7.5}$$

donde $\hat{y_i}$ es el i-ésimo valor escalar en la salida del modelo, y_i es la clase real correspondiente y N es el número de elementos en el conjunto de entrenamiento.

Por otro lado, si se tienen más de dos clases, se usa la ***Categorical Cross-Entropy*** cuya función es:

$$Loss = -\sum_{i=1}^{N} y_i \cdot log(\hat{y_i}) \tag{7.6}$$

donde $\hat{y_i}$ es el i-ésimo valor escalar en la salida del modelo, y_i es la clase real correspondiente y N es el número de valores escalares en la salida del modelo.

7.6.6. Método del descenso del gradiente

El descenso de gradiente (*Gradient Descent*) es un algoritmo de optimización que se utiliza para iterar a través de diferentes combinaciones de pesos para encontrar las mejores tal que minimicen el error. Esto permite encontrar el mínimo de la función de pérdida y luego obtener mejores predicciones. Para este método, se definen dos parámetros: la dirección para obtener el mínimo de la función de pérdida y la tasa de aprendizaje (tasa de aprendizaje), que nos dice el tamaño del paso para llegar al mismo.

Si se quiere aplicar el descenso del gradiente a la función escalón, se tiene que poder tomar gradiente de la misma. Sin embargo, el gradiente no es diferenciable en ciertos puntos cuando se toma a esta función. Una solución es utilizar la función sigmoidea que dará un valor de cero o uno pero con la diferencia que se puede derivar para encontrar el gradiente.

Matemáticamente, el método de descenso es un algoritmo iterativo en el que se comienza desde un punto aleatorio y se va acercando al mínimo de la función (se llama "descenso" porque los pasos son hacia abajo). La longitud del paso no debe ser tan pequeña para que haga lenta la convergencia; ni tan grande que pueda pasar por alto a un mínimo. Se pueden caer en mínimos locales por lo que se recomienda probar con diferentes puntos iniciales. Aunque el descenso de gradiente se limita a la optimización en espacios continuos, el concepto general de realizar repetidamente un pequeño movimiento hacia mejores configuraciones se puede generalizar a espacios discretos. La ecuación del método es:

$$X_{k+1} = X_k + \alpha_k \cdot d_k \tag{7.7}$$

en donde $d_k = -\nabla f(X_k)$ es la dirección del descenso y α_k es la longitud del paso. La iteración del método es de la siguiente manera:

- Se elige la dirección del descenso. Queremos movernos en la dirección de máximo decaimiento de la función lo que equivale a la pendiente más negativa. $d_k = -\nabla f(X_k)$
- Se busca la longitud del paso, proceso que es costoso y se puede reemplazar por el mínimo aproximado del polinomio interpolador $\alpha_k = argminf(X_k + \alpha \cdot d_k)$
- Luego, se calcula:

$$X_{k+1} = X_k + \alpha_k \cdot d_k \tag{7.8}$$

Un refinamiento necesario es cambiar el tamaño de los pasos que se dan para evitar sobrepasar al mínimo y rebotar infinitamente alrededor de él. Si se modera el tamaño del paso para que sea proporcional al tamaño del gradiente, cuando estemos cerca, daremos pasos más pequeños. Esto supone que a medida que nos acercamos al mínimo, la pendiente se vuelve menos profunda.

7.6.7. Retropropagación

La retropropagación (*Backpropagation*) fue introducida en 1985 por Hinton, Rumelhart y Williams. Como ya se mencionó anteriormente, se desea comparar el valor predicho por la red neuronal con el valor de la salida esperada o real. Por lo tanto, primero calculamos la función de pérdida y luego su gradiente. La parte *back* (al revés) del nombre se debe a que el cálculo del gradiente avanza hacia atrás en toda la red, con el gradiente de la capa final de pesos que se calcula primero y el gradiente de la primera capa de pesos se calcula en último lugar. Los cálculos parciales del gradiente de una capa se reutilizan en el cálculo del gradiente de la capa anterior. Este flujo hacia atrás de la información de error permite un cálculo eficiente del gradiente en cada capa.

Bibliografía

[1] Segi, M. (1960) Cancer Mortality for Selected Sites in 24 Countries (1950–57). Department of Public Health, Tohoku University of Medicine, Sendai, Japan.

[2] Doll, R., Payne, P., Waterhouse, J.A.H. eds (1966). Cancer Incidence in Five Continents, Vol. I. Union Internationale Contre le Cancer, Geneva.

[3] Schneider, A. P., Zainer, C. M., Kubat, C. K., Mullen, N. K., Windisch, A. K. (2014). The breast cancer epidemic: 10 facts. The Linacre Quarterly, 81(3), 244-277.

[4] Dalla Fontana, F., Seiref, S., Costa, L., Pizzi, J., Schiaffino, R., Bernardi, S. A. (2019). Análisis de supervivencia y causa de muerte en pacientes con cáncer de mama.

[5] Bae, M. S., Moon, W. K., Chang, J. M., Koo, H. R., Kim, W. H., Cho, N., ... Seo, M. (2014). Breast cancer detected with screening US: reasons for nondetection at mammography. Radiology, 270(2), 369-377.

[6] Birdwell, R. L., Ikeda, D. M., O'Shaughnessy, K. F., Sickles, E. A. (2001). Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection. Radiology, 219(1), 192-202.

[7] Blanks, R. G., Wallis, M. G., Moss, S. M. (1998). A comparison of cancer detection rates achieved by breast cancer screening programmes by number of readers, for one and two view mammography: results from the UK National Health Service breast screening programme. Journal of Medical screening, 5(4), 195-201.

[8] Ertosun, M. G., Rubin, D. L. (2015, November). Probabilistic visual search for masses within mammography images using deep learning. In 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 1310-1315). IEEE.

[9] Longo, R., Tonutti, M., Rigon, L., Arfelli, F., Dreossi, D., Quai, E., ... Cova, M. A. (2014). Clinical study in phase-contrast mammography: image-quality analysis. Philosophical Transactions of the Royal Society A: Mathematical, physical and engineering sciences, 372(2010), 20130025.

[10] Bird, R. E., Wallace, T. W., Yankaskas, B. C. (1992). Analysis of cancers missed at screening mammography. Radiology, 184(3), 613-617.

[11] Boyd, N. F., Guo, H., Martin, L. J., Sun, L., Stone, J., Fishell, E., ... Yaffe, M. J. (2007). Mammographic density and the risk and detection of breast cancer. New England journal of medicine, 356(3), 227-236.

[12] Dinnes, J., Moss, S., Melia, J., Blanks, R., Song, F., Kleijnen, J. (2001). Effectiveness and cost-effectiveness of double reading of mammograms in breast cancer screening: findings of a systematic review. The Breast, 10(6), 455-463.

[13] Brem, R. F., Baum, J., Lechner, M., Kaplan, S., Souders, S., Naul, L. G., Hoffmeister, J. (2003). Improvement in sensitivity of screening mammography with computer-aided detection: a multiinstitutional trial. American Journal of Roentgenology, 181(3), 687-693.

[14] Ciatto, S., Del Turco, M. R., Risso, G., Catarzi, S., Bonardi, R., Viterbo, V., ... Indovina, P. L. (2003). Comparison of standard reading and computer aided detection (CAD) on a national proficiency test of screening mammography. European journal of radiology, 45(2), 135-138.

[15] Gilbert, F. J., Astley, S. M., Gillan, M. G., Agbaje, O. F., Wallis, M. G., James, J., ... Duffy, S. W. (2008). Single reading with computer-aided detection for screening mammography. New England Journal of Medicine, 359(16), 1675-1684.

[16] Lehman, C. D., Wellman, R. D., Buist, D. S., Kerlikowske, K., Tosteson, A. N., Miglioretti, D. L., Breast Cancer Surveillance Consortium. (2015). Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA internal medicine, 175(11), 1828-1837.

[17] Fenton, J. J., Taplin, S. H., Carney, P. A., Abraham, L., Sickles, E. A., D'Orsi, C., ... Elmore, J. G. (2007). Influence of computer-aided detection on performance of screening mammography. New England Journal of Medicine, 356(14), 1399-1409.

[18] Fenton, J. J., Abraham, L., Taplin, S. H., Geller, B. M., Carney, P. A., D'Orsi, C., ... Breast Cancer Surveillance Consortium. (2011). Effectiveness of computer-aided detection in community mammography practice. Journal of the National Cancer institute, 103(15), 1152-1161.

[19] Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.

[20] He, K., Zhang, X., Ren, S., Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision (pp. 1026-1034).

[21] Ferraris, V. A. (2019). Commentary: should we rely on receiver operating characteristic curves? From submarines to medical tests, the answer is a definite maybe!. The Journal of thoracic and cardiovascular surgery, 157(6), 2354-2355.

[22] Lee, R. S., Gimenez, F., Hoogi, A., Miyake, K. K., Gorovoy, M., Rubin, D. L. (2017). A curated mammography data set for use in computer-aided detection and diagnosis research. Scientific data, 4(1), 1-9.

[23] Lopez, M. G., Posada, N., Moura, D. C., Pollán, R. R., Valiente, J. M. F., Ortega, C. S., ... Araújo, B. M. F. (2012, July). BCDR: a breast cancer digital repository. In 15th International conference on experimental mechanics (Vol. 1215).

[24] Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., Cardoso, J. S. (2012). Inbreast: toward a full-field digital mammographic database. Academic radiology, 19(2), 236-248.

[25] Oliveira JE, Guelda MO, Araújob AdA, Ottc B, Desernoa TM. Towards a standard reference database for computer-aided mammography. In: Proc SPIE vol. 2008. p. 69151Y

[26] Chris Rose DT, Williams A, Wolstencroft K, Taylor C (2006) DDSM: digital database for screening mammography.

[27] J, P. (1994). The mammographic image analysis society digital mammogram database. Digital Mammo, 375-386.

[28] Huynh, B. Q., Li, H., Giger, M. L. (2016). Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. Journal of Medical Imaging, 3(3), 034501.

[29] Aboutalib, S. S., Mohamed, A. A., Berg, W. A., Zuley, M. L., Sumkin, J. H., Wu, S. (2018). Deep learning to distinguish recalled but benign mammography images in breast cancer screening. Clinical Cancer Research, 24(23), 5902-5909.

[30] Li, H., Zhuang, S., Li, D. A., Zhao, J., Ma, Y. (2019). Benign and malignant classification of mammogram images based on deep learning. Biomedical Signal Processing and Control, 51, 347-354.

[31] Cai, H., Huang, Q., Rong, W., Song, Y., Li, J., Wang, J., ... Li, L. (2019). Breast microcalcification diagnosis using deep convolutional neural network from digital mammograms. Computational and mathematical methods in medicine, 2019.

[32] Agarwal, V., Carson, C. (2015). Using Deep Convolutional Neural Networks to predict semantic features of lesions in mammograms. C231n Course Project Reports.

[33] Lévy, D., Jain, A. (2016). Breast mass classification from mammograms using deep convolutional neural networks. arXiv preprint arXiv:1612.00542.

[34] Xi, P., Shu, C., Goubran, R. (2018, June). Abnormality detection in mammography using deep convolutional neural networks. In 2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA) (pp. 1-6). IEEE.

[35] Khan, H. N., Shahid, A. R., Raza, B., Dar, A. H., Alquhayz, H. (2019). Multi-view feature fusion based four views model for mammogram classification using convolutional neural network. IEEE Access, 7, 165724-165733.

[36] Ragab, D. A., Sharkas, M., Marshall, S., Ren, J. (2019). Breast cancer detection using deep convolutional neural networks and support vector machines. PeerJ, 7, e6201.

[37] Suk, H. I., Lee, S. W., Shen, D., Alzheimer's Disease Neuroimaging Initiative. (2014). Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. NeuroImage, 101, 569-582.

[38] Ting, F. F., Tan, Y. J., Sim, K. S. (2019). Convolutional neural network improvement for breast cancer classification. Expert Systems with Applications, 120, 103-115.

[39] Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., ... Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging, 35(5), 1285-1298.

[40] Stomper, P. C., Geradts, J., Edge, S. B., Levine, E. G. (2003). Mammographic predictors of the presence and size of invasive carcinomas associated with malignant microcalcification lesions without a mass. American Journal of Roentgenology, 181(6), 1679-1684.

[41] Del Turco, M. R., Mantellini, P., Ciatto, S., Bonardi, R., Martinelli, F., Lazzari, B., Houssami, N. (2007). Full-field digital versus screen-film mammography: comparative accuracy in concurrent screening cohorts. American Journal of Roentgenology, 189(4), 860-866.

[42] Gajdos, C., Tartter, P. I., Bleiweiss, I. J., Hermann, G., De Csepel, J., Estabrook, A., Rademaker, A. W. (2002). Mammographic appearance of nonpalpable breast cancer reflects pathologic characteristics. Annals of surgery, 235(2), 246.

[43] Holland, R., Hendriks, J. H. (1994, August). Microcalcifications associated with ductal carcinoma in situ: mammographic-pathologic correlation. In Seminars in diagnostic pathology (Vol. 11, No. 3, pp. 181-192).

[44] Hernández, P. A., Estrada, T. T., Pizarro, A. L., Cisternas, M. L. D., Tapia, C. S. (2016). Breast calcifications: description and classification according to bi-rads 5th edition. Rev. Chil. Radiol., 22, 80-91.

[45] American College of Radiology, D'Orsi, C. J. (2013). ACR BI-RADS Atlas: Breast Imaging Reporting and Data System; Mammography, Ultrasound, Magnetic Resonance Imaging, Follow-up and Outcome Monitoring, Data Dictionary. ACR, American College of Radiology.

[46] Bent, C. K., Bassett, L. W., D'Orsi, C. J., Sayre, J. W. (2010). The positive predictive value of BIRADS microcalcification descriptors and final assessment categories. American Journal of Roentgenology, 194(5), 1378-1383.

[47] Uematsu, T., Kasami, M., Yuen, S. (2008). Usefulness and limitations of the Japan Mammography Guidelines for the categorization of microcalcifications. Breast Cancer, 15(4), 291-297.

[48] Liberman, L., Abramson, A. F., Squires, F. B., Glassman, J. R., Morris, E. A., Dershaw, D. D. (1998). The breast imaging reporting and data system: positive predictive value of mammographic features and final assessment categories. AJR. American journal of roentgenology, 171(1), 35-40.

[49] Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.

[50] Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[51] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," arXiv:1409.4842, 2014.

[52] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[53] Canziani, A., Paszke, A., Culurciello, E. (2016). An analysis of deep neural network models for practical applications. arXiv preprint arXiv:1605.07678.

[54] Shu, X., Zhang, L., Wang, Z., Lv, Q., Yi, Z. (2020). Deep neural networks with region-based pooling structures for mammographic image classification. IEEE transactions on medical imaging, 39(6), 2246-2255.

[55] Giuliano AE, Connolly JL, Edge SB, et al. Breast Cancer-Major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. CA Cancer J Clin. 2017;67(4):290-303.

[56] Sanders ME, Schuyler PA, Simpson JF, Page DL, Dupont WD. Continued observation of the natural history of low-grade ductal carcinoma in situ reaffirms proclivity for local recurrence even after more than 30 years of follow-up. Mod Pathol. 2015;28(5):662-669

[57] Copeland, B. (2020, August 11). Artificial intelligence. Encyclopedia Britannica. https://www.britannica.com/technology/artificial-intelligence

[58] https://www.eyenuk.com/en/products/eyeart/

[59] https://www.ibm.com/watson-health