TRABAJO FIN DE MÁSTER (TFM)

Máster Big Data y Business Analytics

Curso 2021-2022

Facultad de Estudios Estadísticos

Universidad Complutense de Madrid

DESARROLLO DE UNA RED NEURONAL PARA LA DETECCIÓN Y CLASIFICACIÓN DE ENFERMEDADES TORÁCICAS

Jordi GARCÍA PABLOS

Álvaro GUERÍN ALMAZOR

Pablo JIMENO DOMÍNGUEZ

Javier LÓPEZ BAHÓN

José NAFRÍA FERNÁNDEZ

Santiago MOTA HERCE

Carlos ORTEGA FERNÁNDEZ

Contenido

Tabla de figuras

1. Resumen

Las enfermedades torácicas son causantes, sólo en los servicios de Urgencias, de un 4-6% de las consultas totales. Es por eso que, gracias a las nuevas tecnologías y los avances en la inteligencia artificial y el aprendizaje de las máquinas, el buen uso de estos puede tener un gran valor para el sector sanitario y la humanidad en general. El proyecto tratará de, mediante algoritmos de inteligencia artificial y redes neuronales, enseñar a un ordenador para ser capaz de detectar la presencia de quince enfermedades a partir de radiografías.

2. Introducción

2.1. Enfermedades torácicas

Las enfermedades torácicas son causantes, sólo en los servicios de Urgencias de un 4-6% de las consultas totales. Estas afectan a órganos tan vitales como el corazón, los pulmones y el esófago entre otros, teniendo posibles implicaciones con patologías potencialmente graves e incluso mortales.

En España se mantienen como la tercera causa de muerte, con 51.476 defunciones lo que implica un 10,4% del total en el año 2020 según los últimos datos publicados por el INE (Estadística, s.f.).

Por estas razones y para agilizar a los médicos sus tareas a la hora de tratar a los pacientes se ha visto la necesidad de realizar este estudio. En el Anexo, en el punto 1, se explican las enfermedades más relevantes del conjunto de datos elegido.

2.2. Métodos de diagnóstico actuales

En la actualidad la mayoría de los procesos del tórax precisan del apoyo radiológico. Este es el método principal debido a su bajo coste, fácil realización y relativa buena sensibilidad. A partir de esta técnica se puede decidir si es necesaria otras técnicas radiológicas.

Las radiografías son interpretadas por radiólogo, un médico especializado en interpretar los exámenes radiológicos. Esto implica una limitación a la percepción del ojo humano, con hasta un 30% de error en la interpretación (Brady, 2016). El desarrollo de una inteligencia artificial (IA) en este ámbito podría conseguir mejorar los resultados en la detección de este tipo de enfermedades, aparte de reducir los tiempos.

2.3. Redes neuronales

La inteligencia artificial (IA) es la combinación de algoritmos planteados con el propósito de crear máquinas que presenten las mismas capacidades que el ser humano. Ésta se puede usar en diferentes ámbitos de la industria y de la vida una vez que se consigue resolver el problema que se encontró.

Las redes neuronales artificiales son modelos computacionales inspirados en el comportamiento observado en su homólogo biológico (Marcel van Gerven, 2019). El motivo

por el que usar redes neuronales es porque se trata de una técnica que intenta imitar la capacidad de procesamiento de un cerebro, aunque a una escala mucho menor. Actualmente no se espera que una red neuronal pueda llegar a la altura de un cerebro humano, pero pueden ser una solución muy potente para aplicar en algunos problemas específicos.

2.4. Objetivos

El objetivo principal de este proyecto es desarrollar una red neuronal (Redes neuronales) que sepa detectar y clasificar diferentes enfermedades torácicas. Esto se pretende conseguir entrenando la red neuronal con un dataset[1] de radiografías torácicas.

Con esta red neuronal se le dará al médico una diagnosis sobre el estado del paciente, de esta manera el médico podrá ayudarse de un sistema que ha estudiado miles de imágenes previamente. El objetivo no es eliminar la decisión del médico si no aportarle una nueva herramienta que le ayude a diagnosticar al paciente.

Como ya se ha podido observar en otros desarrollos, las redes neuronales pueden ser de gran ayuda en Medicina para la clasificación y detección de patrones. Desde detección de cánceres, insuficiencias cardiacas o enfermedades oftalmológicas. Los registros médicos pueden ser usados para el entrenamiento de estas redes neuronales y aportar en el diagnóstico del médico.

2.5. Dataset

Se parte de un dataset publicado en Kaggle compartido por el NIH (National Institutes of Health) con la esperanza de que las instituciones académicas y de investigación puedan crear una IA capaz de procesar grandes cantidades de radiologías, con dos objetivos; confirmar los resultados que los radiólogos han encontrado e identificar otros hallazgos que pueden haber sido pasado por alto.

El dataset cuenta con 112.120 imágenes de parte de 30.805 pacientes del NIH. Aunque actualmente es el mayor conjunto de datos público sobre radiografías de tórax, ciento doce mil imágenes no son tantas imágenes como para realizar una red neuronal extremadamente precisa.

Con el objetivo de guardar la privacidad de los pacientes únicamente se han compartido la imagen del tórax, la enfermedad, edad y sexo del paciente. Para el tratamiento de los datos se ha utilizado un dataset con las imágenes en formato .TFREC lo que facilita el trabajo de limpieza y modelaje[2].

Se tienen dos conjuntos de datos, uno con las radiografías y la enfermedad que el médico ha diagnosticado y otro que incluye información adicional sobre las radiografías, como el id del paciente, sexo y edad. El primer conjunto de datos es el que se utiliza para crear la red neuronal, mientras que el segundo se utiliza para realizar un análisis sobre los pacientes.

[1] En el apartado 2.5 se explica en profundidad el dataset o conjunto de datos empleado.

[2] En el apartado 5.3 se trata en profundidad el tratamiento de los datos.

3. Metodología

3.1. Organización del equipo

Para la planificación del proyecto y la correcta organización de las tareas es fundamental la creación de un Diagrama de Gantt donde se precise la dedicación necesaria para realizar cada tarea. Estas se pueden clasificar en dos grupos: análisis y desarrollo de los modelos.

3.2. Entorno de trabajo

Elegir un entorno de trabajo que permita conocer en cualquier momento los avances realizados es una cuestión importante. Con el servicio de Google Drive se pueden solventar diferentes problemáticas como el almacenaje de los datos o poder trabajar con ellos usando Google Colab. Además, para poder ver los desarrollos del equipo se usa una hoja de cálculo donde cada modelo que se desarrolle se apunte con diferentes características.

4. Análisis descriptivo

4.1. Introducción

El objetivo de realizar un análisis descriptivo del conjunto de datos es poder sacar más información sobre las distintas enfermedades, intentar encontrar patrones y relaciones sobre las enfermedades en diferentes pacientes. El análisis se ha dividido en dos partes.

Se comenzará analizando a los pacientes que han participado en el muestreo. Se analiza la edad y el sexo de los pacientes. En una segunda parte se analizan las imágenes que se van a estudiar: número de imágenes con una misma enfermedad, número de enfermedades que presenta cada imagen y concurrencia entre enfermedades.

4.2. Radiografía por paciente

Cada imagen pertenece a un único paciente, pero hay imágenes que son de varios pacientes. Por eso se comprueba el número de radiografías por paciente. Para hacerlo más sencillo se agrupan los pacientes por grupos.

Número radiografías	Porcentaje de pacientes
1	56.8
<= 5	84.3
<= 10	92.7

Se observa que la mayor parte del peso se concentra en pacientes con 5 o menos radiografías, y casi la totalidad está en 10 o menos. Hay algún caso excepcional en el que se tienen 184 en un mismo paciente, seguramente se deba a algún error.

4.3. Pacientes

4.3.1. Edades

Se realiza un primer análisis general sobre los pacientes en el que se descubre que el valor máximo de edad es de 414 años. Estos valores son erróneos y además alteraran los resultados del análisis, se eliminan los (16) pacientes con más de 99 años.

A continuación, se realiza un diagrama y un histograma. Las gráficas revelan un mayor número de pacientes en el rango de los 40 a los 60 años. Además, hay franjas de edades en las que el número de pacientes es inusualmente bajo comparado con edad más próxima.

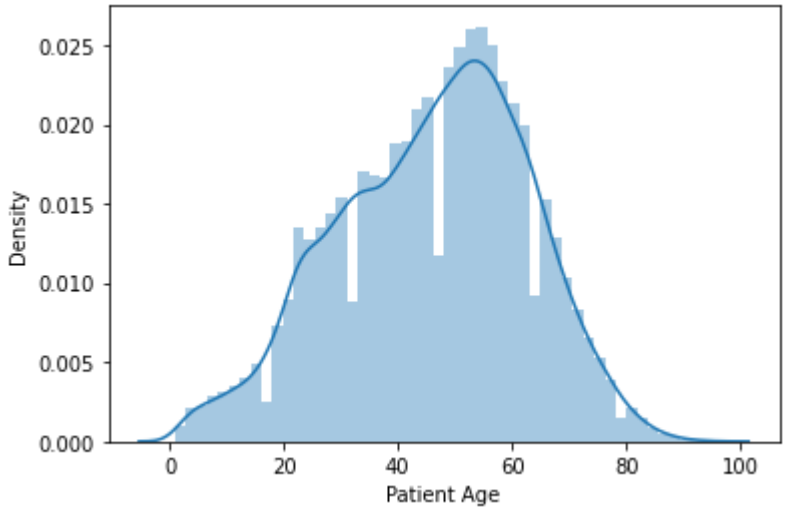

Fig. 1 Gráfico de densidad de los pacientes por edad

Para continuar con el estudio, se crea una pirámide de población con los pacientes del estudio. Se compara con la pirámide de EEUU. Se observa un desbalance entre la población de EEUU (izquierda) y el dataset (derecha). La edad media del estudio no representa a la población real, lo que nos hace pensar si la desigualdad de edad se debe a que la edad influye en padecer una enfermedad torácica.

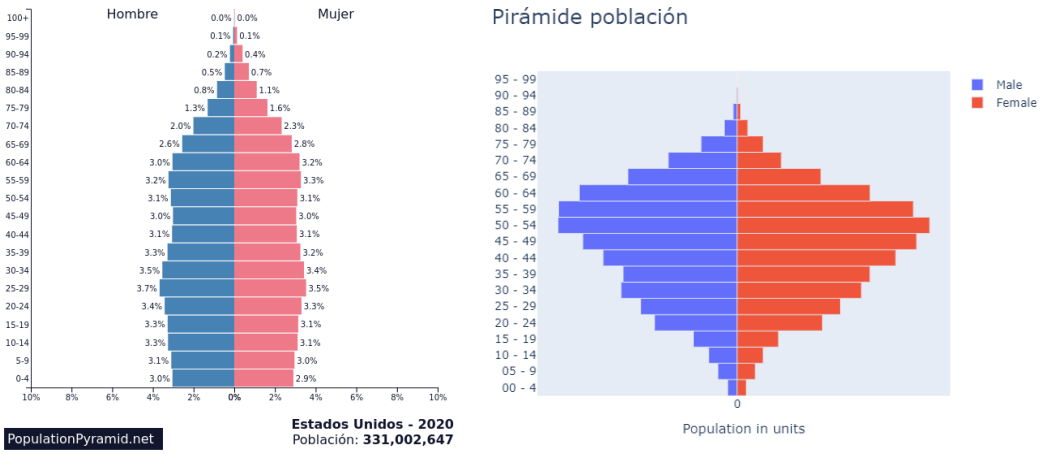

Fig. 2 Pirámide poblacional de EEUU (izquierda) y de los paciente (derecha)

Para finalizar, se comparan dos diagramas de dispersión el de sanos (verde) sobre enfermos (rojo). El dato que arrojan ambos gráficos es que la edad sí influye en las enfermedades torácicas, se observa que la curva verde es notablemente mayor en los rangos de baja edad y a partir de los 50 años la curva de enfermos toma la delantera. Además, el pico de enfermos se concentra entre los 50 y 65 años.

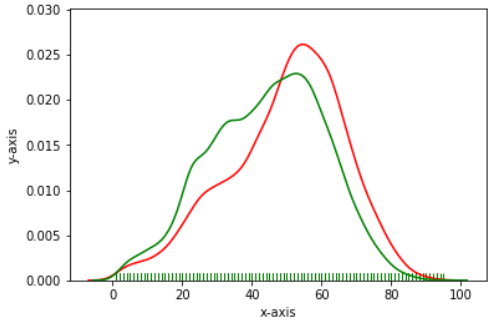

Fig. 3 Gráfico de densidad de los pacientes por edad diferenciando entre sanos (verde) y enfermos (rojo)

4.3.2. Género

Para finalizar el análisis se hace una distinción en las radiografías por sexo, se ha realizado un estudio general, otro para los casos enfermos y un tercero para los sanos. Se ve un mayor número de pacientes masculinos. La proporción entre hombres y mujeres se mantiene prácticamente constante en los casos sanos y los enfermos. Comparando los 3 casos, se ve una ligera tendencia a que los hombres tengan enfermedades torácicas.

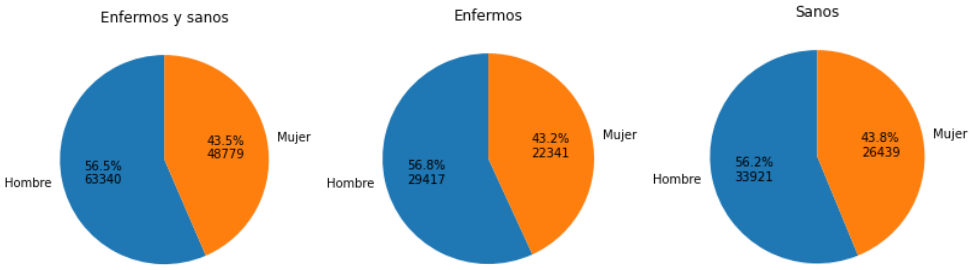

Fig. 4 Gráficos circulares por sexo

Para confirmar que los hombres son más propensos a tener enfermedades torácicas que las mujeres. Se realiza una tabla de contingencia, el odd ratio de la tabla es de 1.0263. Esto indica que los hombres son ligeramente más propensos a caer enfermos que las mujeres, puede que este sea el motivo por el que el número de hombres sea mayor.

Sexo	Sano	Enfermo
Hombre	33922	29418
Mujer	26439	22341

Para finalizar el estudio se realiza una hipótesis, que resulta en un chi-cuadrado de 4,6 y un p-valor de 0,032. Con esta información se rechaza la hipótesis nula de que el sexo y la enfermedad son independientes.

4.4. Imágenes

4.4.1. Enfermedades

Se comienza con un análisis general de las radiografías. Las ciento doce mil imágenes se han recogido de forma que la mitad son de pacientes con enfermedades torácicas (46%) y la otra mitad no presenta síntomas de ninguna enfermedad (54%).

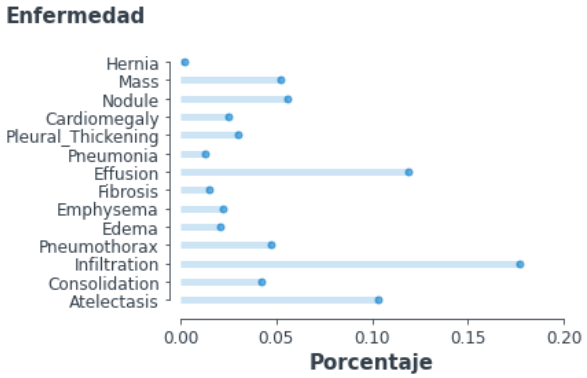

Fig. 5 Porcentaje de aparición por enfermedad

Dentro de las imágenes con enfermedades encontradas el número de imágenes por enfermedad es muy dispar. De las enfermedades effusion, infiltración, atelectasis se tienen un peso importante sobre el conjunto (> 10%), mientras que del resto se tiene un menor peso (entre 2% y 6%) a excepción de Hernia que apenas se tienen datos. Este desbalance en los datos perjudica la predictibilidad del modelo.

4.4.2. Posición de la radiografía

En la toma de radiografía no siempre se puede disponer del paciente de frente a la máquina (PA, Posterior Anterior). En ocasiones por motivos de invalidez del paciente o por otras causas la radiografía se toma de perfil al paciente (AP, Anterior Posterior).

Las imágenes que se han tomado en la postura PA, abarcan el 60% de los pacientes, más de la mitad de las imágenes. Como en el caso de la edad se crea una tabla de contingencia que nos pueda aportar información.

Tipo de toma	Sano	Enfermo
PA	28008	39302
AP	21059	23751

La ratio odds arroja un valor de 0.63, lo que es una fuerte correlación entre los pacientes en postura AP y caer enfermo. Esta relación confirma que los pacientes en mal estado realizan la AP por no estar en condiciones para realizar la PA. Los datos dicen que se rechaza la hipótesis nula de que la posición y la enfermedad son independientes. Los valores obtenidos fueron chi-cuadrado: 316.8 y p-valor:7^-71.

4.4.3. Número de enfermedades por imagen

Se descubre que hay imágenes con más de una enfermedad diagnosticada, lo cual puede aportar relación entre enfermedades[3]. El número de imágenes con una sola enfermedad es notablemente mayor al de imágenes con más enfermedades.

Núm. enferm.	1	2	3	4	5	6	7	8	9
Imágenes	91324	14306	1247	4856	301	67	16	1	2

4.4.4. Concurrencia

Como se ve en el apartado anterior, casi el 20% de las imágenes tienen diagnosticadas dos o más enfermedades. Para verificar si existe alguna relación entre las distintas enfermedades, se estudia la concurrencia entre enfermedades.

Los datos objeto de estudios son aquellas imágenes que tienen algún tipo de enfermedad. A continuación, se muestra la matriz de concurrencia, cada columna corresponde a una enfermedad y en cada fila se indica el número de imágenes que comparten ambas enfermedades. Ej: hay 40 casos de Hernia y Atelectasis.

Para hacer esta representación más gráfica se ha realizado un diagrama de Chor con la matriz de correlaciones anterior. Cada enfermedad tiene un color, de donde sale una flecha que indica el número de imágenes que tiene ambas enfermedades en función del espesor.

De este estudio se concluye que aparentemente sí hay relación entre algunas enfermedades, especialmente Infiltration. Se observa en los casos de Effusion que el 30% de los pacientes tiene Infiltration. También en el 40% de los casos de Neumonía el paciente presenta Infiltration.

	Atelectasis	**Cardiomegaly**	**Consolidation**	**Edema**	**Effusion**	**Emphysema**	**Fibrosis**	**Hernia**	**Infiltration**	**Mass**	**Nodule**	**Pleural_Thickening**	**Pneumonia**	**Pneumothorax**
Atelectasis	11559	370	1223	221	3275	424	220	40	3264	739	590	496	262	774
Cardiomegaly	370	2776	169	127	1063	44	52	7	587	102	108	111	41	49
Consolidation	1223	169	4667	162	1287	103	79	4	1221	610	428	251	123	223
Edema	221	127	162	2303	593	30	9	3	981	129	131	64	340	33
Effusion	3275	1063	1287	593	13317	359	188	21	4000	1254	912	849	269	996
Emphysema	424	44	103	30	359	2516	36	4	449	215	115	151	23	747
Fibrosis	220	52	79	9	188	36	1686	8	345	25	10	8	3	9
Hernia	40	7	4	3	21	4	8	227	33	25	10	8	3	9
Infiltration	3264	587	1221	981	4000	449	345	33	19894	1159	1546	750	605	946
Mass	739	102	610	129	1254	215	117	25	1159	5782	906	411	70	341
Nodule	590	108	428	131	912	115	166	10	1546	906	6331	411	70	341
Pleural_Thickening	496	111	251	64	849	151	176	8	750	452	411	3385	48	289
Pneumonia	262	41	123	340	269	23	11	3	605	71	70	48	1431	41
Pneumothorax	774	49	223	33	996	747	80	9	946	431	341	289	41	5302

[3] En el apartado de concurrencia se explica en profundidad la relación entre enfermedades.

	Atelectasis	Cardiomegaly	Consolidation	Edema	Effusion	Emphysema	Fibrosis	Hernia	Infiltration	Mass	Nodule	Pleural_Thickening	Pneumonia	Pneumothorax
Atelectasis	100%	3,2%	10,6%	1,9%	28,3%	3,7%	1,9%	0,3%	28,2%	6,4%	5,1%	4,3%	2,3%	6,7%
Cardiomegaly	13,3%	100%	6,1%	4,6%	38,3%	1,6%	1,9%	0,3%	21,1%	3,7%	3,9%	4,0%	1,5%	1,8%
Consolidation	26,2%	3,6%	100%	3,5%	27,6%	2,2%	1,7%	0,1%	26,2%	13,1%	9,2%	5,4%	2,6%	4,8%
Edema	9,6%	5,5%	7,0%	100%	25,7%	1,3%	0,4%	0,1%	42,6%	5,6%	5,7%	2,8%	14,8%	1,4%
Effusion	24,6%	8,0%	9,7%	4,5%	100%	2,7%	1,4%	0,2%	30,0%	9,4%	6,8%	6,4%	2,0%	7,5%
Emphysema	16,9%	1,7%	4,1%	1,2%	14,3%	100%	1,4%	0,2%	17,8%	8,5%	4,6%	6,0%	0,9%	29,7%
Fibrosis	13,0%	3,1%	4,7%	0,5%	11,2%	2,1%	100%	0,5%	20,5%	6,9%	9,8%	10,4%	0,7%	4,7%
Hernia	17,6%	3,1%	1,8%	1,3%	9,3%	1,8%	3,5%	100%	14,5%	11,0%	4,4%	3,5%	1,3%	4,0%
Infiltration	16,4%	3,0%	6,1%	4,9%	20,1%	2,3%	1,7%	0,2%	100%	5,8%	7,8%	3,8%	3,0%	4,8%
Mass	12,8%	1,8%	10,5%	2,2%	21,7%	3,7%	2,0%	0,4%	20,0%	100%	15,7%	7,8%	1,2%	7,5%
Nodule	9,3%	1,7%	6,8%	2,1%	14,4%	1,8%	2,6%	0,2%	24,4%	14,3%	100%	6,5%	1,1%	5,4%
Pleural_Thickening	14,7%	3,3%	7,4%	1,9%	25,1%	4,5%	5,2%	0,2%	22,2%	13,4%	12,1%	100%	1,4%	8,5%
Pneumonia	18,3%	2,9%	8,6%	23,8%	18,8%	1,6%	0,8%	0,2%	42,3%	5,0%	4,9%	3,4%	100%	2,9%
Pneumothorax	14,6%	0,9%	4,2%	0,6%	18,8%	14,1%	1,5%	0,2%	17,8%	8,1%	6,4%	5,5%	0,8%	100%

Fig. 6 Tabla de frecuencias de enfermedades (arriba) y Tabla de frecuencias relativas respecto a la enfermedad principal o fila (abajo)

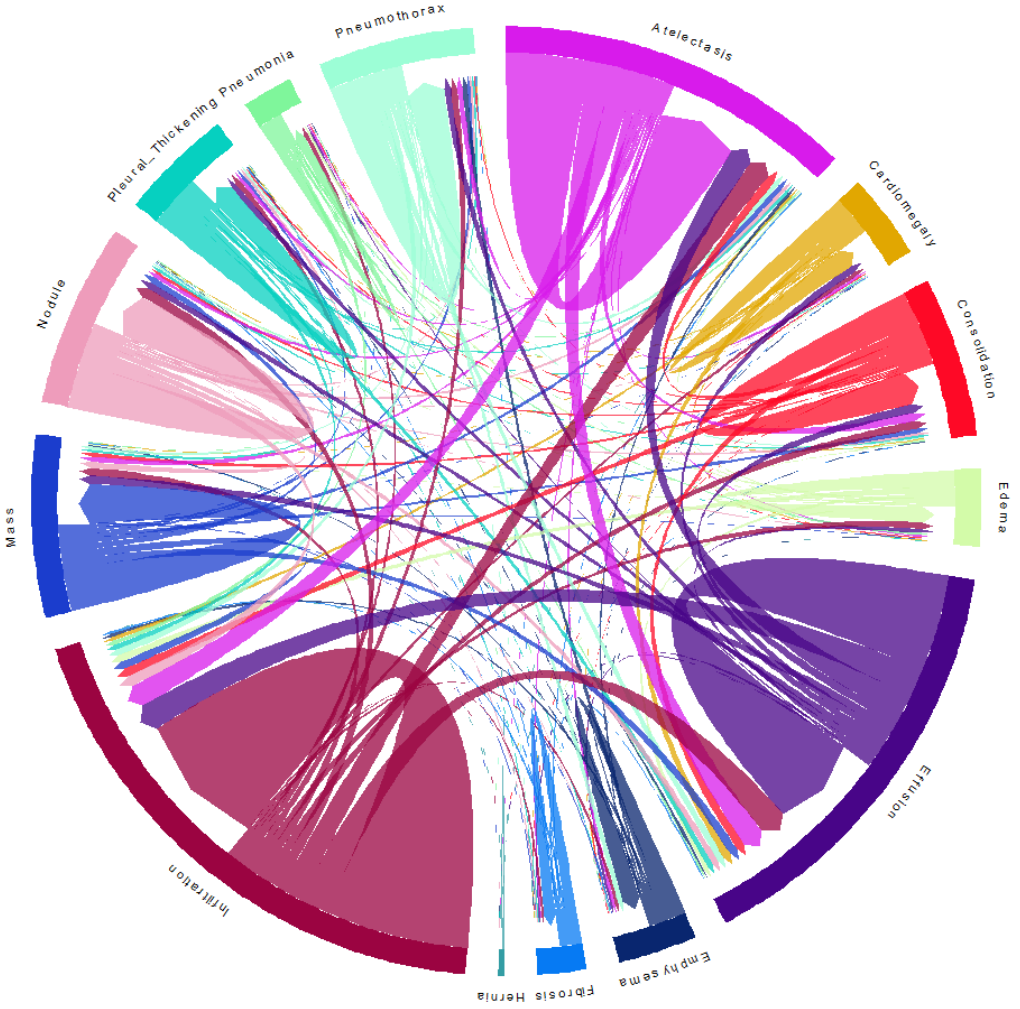

Fig. 7 Gráfico chord de las enfermedades para analizar la concurrencia

5. Desarrollo del modelo

5.1. Entorno de ejecución: Google Colab

Como entorno de ejecución se tomó la decisión de utilizar Google Collab. Un entorno de ejecución que te permite la utilización de GPU (Graphic Processing Units) muy útiles para el entrenamiento y utilización de modelos y tecnologías de Deep Learning. Además para que la sesión fuera alojada en máquinas más potentes (mayor RAM y mejores GPUs) se decidió pagar la suscripción para la herramienta Google Collab Pro. Google Colab también permite el control de versiones, el compartido de código entre otras ventajas.

5.2. Sistema de almacenamiento: Google Drive

Para almacenar todo lo relacionado al proyecto se utilizó el sistema de almacenamiento en la nube: Google Drive. Además de la capacidad de compartir el contenido entre todos los integrantes del equipo, esta herramienta nos permite "mapear" (mount) la carpeta donde se almacenan los datos sobre la máquina en la que se ejecuta la sesión del cuaderno de Python en Google Colab utilizando una librería de Python, que crea un enlace lógico (mapeo) sobre la máquina que aloja el cuaderno. Esta operación es muy sencilla de hacer desde el propio código Python y nos permitió la convergencia de los distintos flujos de trabajo de los distintos integrantes.

5.3. Datos iniciales

El dataset original en el que se recopila la información de los pacientes evaluados fue realizada por el Instituto Nacional de Salud de EEUU y se pueden encontrar en su web oficial. Sin embargo, para la tarea de entrenamiento de redes neuronales se decidió partir de una entrada de Kaggle donde los datos ya habían pasado por un proceso de organización y ligero preprocesado previo.

5.4. Preprocesamiento de los datos

5.4.1. Archivos TFRec

Los datos utilizados para el entrenamiento del modelo vienen contenidos en archivos de tipo TFRec (TensorFlow Record). Este tipo asociado a la librería de Python, Tensorflow, es un un formato simple, eficiente y comprimido para el almacenamiento de datos estructurados (binarios). El total de 256 TFRec contienen el total de las imágenes junto con sus respectivas 15 etiquetas ya analizadas en esta memoria en apartados previos. La proporción de clases en cada uno de los ficheros es la misma que la de la población completa de todas las muestras por lo que durante el desarrollo del modelo no ha habido necesidad de encargarse de esta parte al realizar las particiones de datos.

5.4.2. Conversión de TFRec a TF Datasets

Para el desarrollo se ha utilizado la librería Tensorflow, y más en concreto Keras. Para la utilización eficiente de estas librerías se decidió utilizar el objeto Tensorflow Dataset el cual gestiona el almacenamiento de la gran cantidad de datos pesados utilizados de una manera

adecuada. Para poder declarar este dataset se necesita un "mapa" (diccionario de Python) que decodifica los TFRec. Este mapa ya viene dado por el autor de la fuente de datos preprocesada y autor de estos archivos desde Kaggle.

Se realizó una partición de datos en:

- Datos de entrenamiento (training): - 80% - Imágenes y etiquetas utilizadas para el entrenamiento de la red neuronal.
- Datos de validación (validation): - 10% - Utilizadas para cuantificar la capacidad de generalización de la red durante el entrenamiento.
- Datos de prueba (test): - 10% - Utilizadas para medir la calidad del modelo con datos "nuevos" nunca antes visto por la red (apartados del proceso de entrenamiento)

En el proceso de decodificación de los ficheros TFR a TF Datasets se tomaron las siguientes decisiones de pre-procesamiento y optimización:

1. **Eliminación de todas las muestras de pacientes sin enfermedad.** Estos representaban el 50% de las muestras y alteraba mucho el entrenamiento de la red. Dado que el objetivo del modelo es la ayuda a la toma de decisión del profesional médico en cuanto a la decisión del tipo de enfermedad torácica que puede sufrir un paciente, el hecho de que más de la mitad de las muestras no mostrasen ninguna patología aplicaba un sesgo muy importante a la red dando siempre una probabilidad muy alta a la etiqueta "No finding" para cualquier muestra. Esto puede ser debido a la complejidad de la tarea ya que una radiografía de una persona no enferma frente a que, si lo está, en cuestión de lo que atañe a este modelo, es mínima ya que se representa en forma de muy pocos píxeles en la radiografía y sin formas ni ejes delimitantes muy definidos. Los primeros modelos (que no son mencionados en esta memoria) tenían en cuenta esta etiqueta y estas muestras lo cual hacía que alrededor del 90% de las muestras dieran como enfermedad más probable: ninguna enfermedad; lo cual no tiene ningún sentido si analizamos el objetivo de este modelo de ayuda al diagnóstico.
2. **Procesamiento de imagen.** Se le realiza un cambio de tamaño a formato cuadrado (NxN) debido a que es más recomendado para el uso de redes convolucionales. El valor de N variará dependiendo del modelo implementado. Se transforma a un objeto de tipo Tensor de 3 canales (NxNx3). Cada píxel (en todos los canales) es codificado como un número de coma flotante (float).
3. **Determinación del tamaño del lote (batch size).** Este es el valor que define el número de muestras que el dataset extrae y preprocesa (definido en el punto anterior) en bloque desde los archivos y que alimenta a la red neuronal para cualquier interacción con esta (entrenamiento, evaluación, etc.). Este también es lo que almacenará en disco mientras la red neuronal lo utiliza y que será limpiado posteriormente, esto ayuda a no desbordar la capacidad de la RAM y agilizar los procesos.
4. **Activación del modo de captación previa (prefetch) del dataset.** Este tipo de objetos (TF Dataset) te permite activar este modo que permite llevar a cabo la extracción y preparación (o preprocesado) del siguiente bloque de datos con el que se va a alimentar a la red neuronal mientras esta está usando otro bloque. Esto permite paralelizar estos dos procesos de tal manera que nada más acabar con un bloque (o batch) la red no tiene que esperar a que se prepare el siguiente, sino que ya se tiene preparado. Aunque este proceso lleve un poco más de coste computacional (la RAM tiene que ser capaz de

almacenar dos bloques de datos) agiliza mucho el proceso si la potencia de la máquina lo permite, y en el caso de este proyecto, así lo hizo.

Así se ven alguno de los elementos de este dataset debidamente formateado y tras el proceso mencionado:

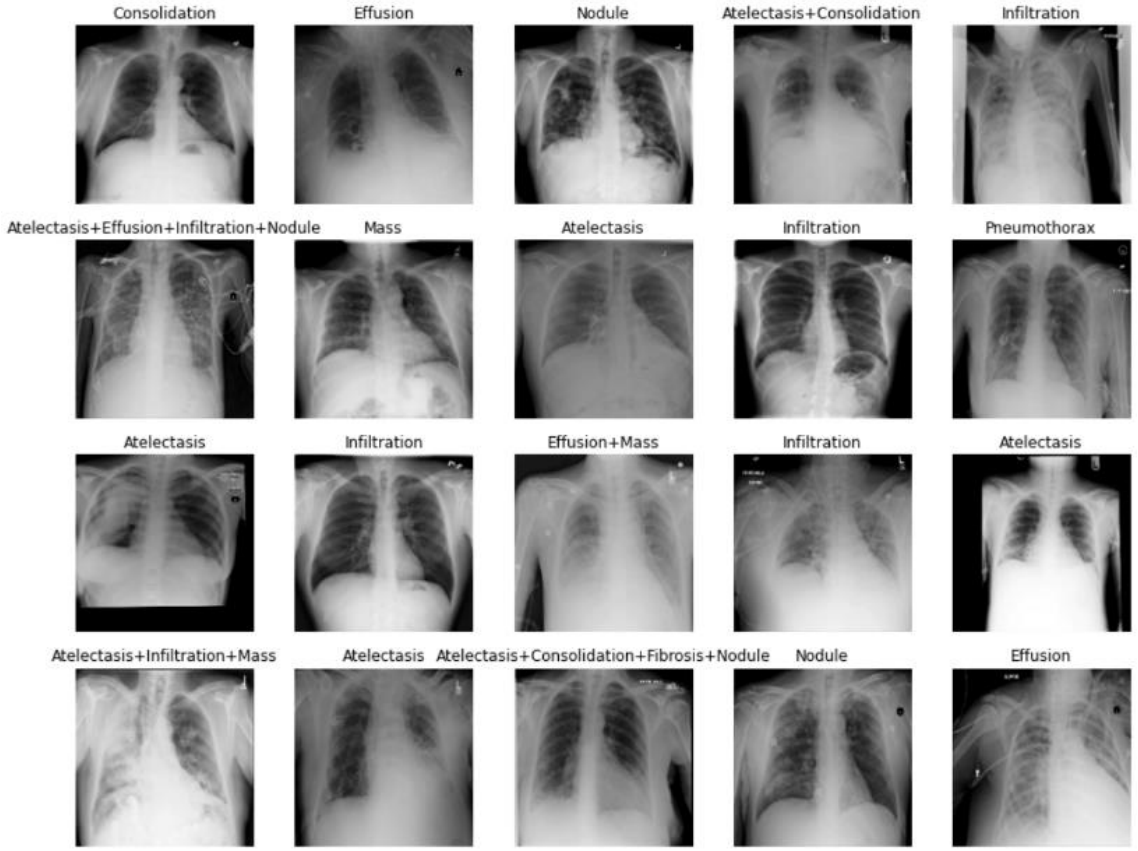

Fig. 8 Ejemplo de radiografías con la patología diagnosticada

5.5. Construcción de los modelos

5.5.1. Salida del modelo

En el momento de construir una red neuronal hay distintas decisiones que tomar y parámetros que ajustar para que se ajuste al objetivo deseado: devolver las probabilidades de que un paciente tenga cierta patología o enfermedad.

Debido a que se va a construir un modelo multiclase y multietiqueta (un paciente puede sufrir más de una enfermedad) se decidió que la salida de la red neuronal fuera un vector con 14 elementos donde se representan probabilidades independientes de sufrir cada una de las enfermedades en lugar de que las probabilidades interrelacionadas y que sumarán 1 entre sí (softmax). Para conseguir esto, en la última capa densa de la red es necesario colocar 14 neuronas con función de activación sigmoide, lo cual fue un factor común a lo largo de todos los modelos probados

En segundo lugar, se debe elegir la función de pérdida por la cual los pesos de cada neurona se van actualizar para ser capaces de aproximarse a la salida deseada para cada muestra: maximizar la probabilidad para las enfermedades que sufre el paciente donde el valor 1 es el objetivo; y minimizar las probabilidades de las enfermedades que no sufre donde el valor

deseado es negativo. Debido a que la variable objetivo está codificada como un vector de 14 elementos binarios junto con la decisión de la función de activación, la función de pérdida adecuada en este caso es entropía cruzada binaria (binary cross entropy) que se define como:

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log\big(p(y_i)\big) + (1 - y_i) \cdot \log\,(1 - p(y_i))$$

También se muestra a continuación un pequeño ejemplo del formato de las salidas de la red con un escenario más sencillo en el que solo hubiera 3 enfermedades:

Orden de etiquetas	Neumonía	Infiltración	Hernia
Salida objetivo	0	1	1
Salida de la red (predicción)	0.10	0.85	0.92

5.5.2. Arquitectura

La siguiente figura muestra la arquitectura general de los distintos modelos implementados:

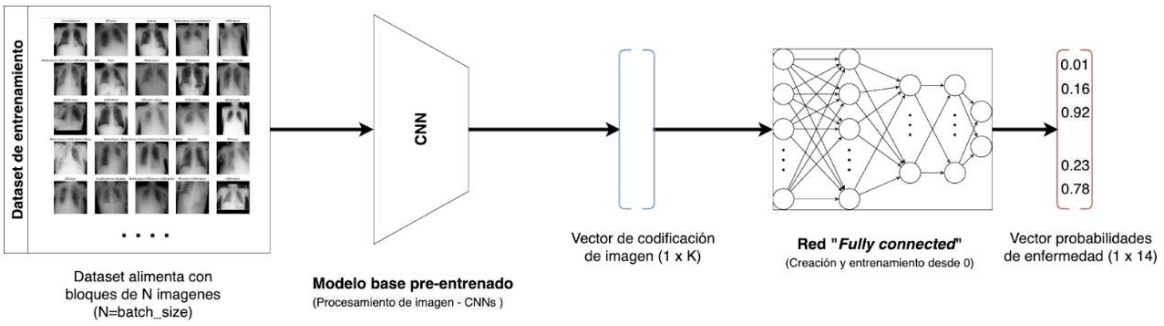

Fig. 9 Esquema básico de la red entrenada para los diferentes modelos

Donde los dos bloques principales son:

1. **Modelo de imagen pre-entrenado.** En la mayoría de los casos (exceptuando el modelo baseline) se utilizaron redes pre-entrenadas disponibles desde la librería de Keras como "codificadores" imagen-vector. Estas redes fueron entrenadas con el dataset de clasificación Imagenet. En el ámbito del Deep Learning inicialización (con pesos ya actualizados para dicho dataset) se usan en la mayoría de los casos para alcanzar el mínimo de la función de coste (converger) en muchas menos epochs (o número de iteraciones a través del dataset completo) ya que el modelo "ya lleva algo aprendido".
2. **Red fully connected.** A la salida del modelo de imagen se añadió una red neuronal fully-connected (red neuronal común) cuya salida es el vector de probabilidades ya descrito. En los distintos modelos se intentó jugar con el número de capas ocultas, el número de neuronas por capas ocultas, métodos de normalización (batch normalization, dropout, kernel initialization .etc) en base a los resultados del entrenamiento.

5.6. Entrenamiento

El entrenamiento de la red es el proceso mediante el cual se ajustan los pesos de cada capa oculta de la red para ajustarse a la salida deseada para cada una de las muestras del dataset de entrenamiento para los datos de entrada (imagen). Hay distintos hiperparametros y decisiones que tomar respecto al entrenamiento:

1. **Tasa de aprendizaje (learning rate).** Este es el parámetro que indica cuánto va variar los pesos para ajustarse al dato que está analizando. suelo tomar valores múltiplos de 10 con 0.1 como valor casi máximo. Es uno de los parámetros más importantes ya que un valor muy alto hará que no alcancemos el mínimo de la función de pérdida por saltos muy grandes que puede nunca ajustarse este mínimo. Un valor muy pequeño también puede llegar a no converger en el mínimo de la función de pérdida o tardar muchas epochs en llegar. Se puede observar gráficamente en la siguiente figura:

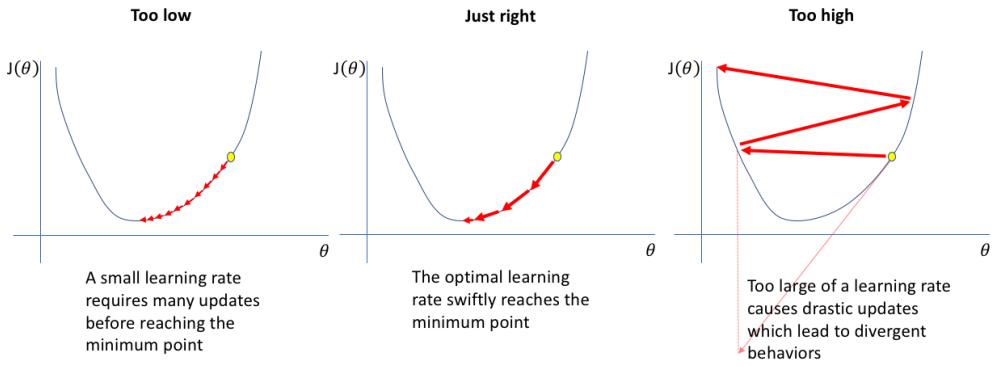

Fig. 10 Representación de evolución del learning rate en función de su valor

2. **Optimizador.** Este tipo de algoritmos ayudan al descenso por gradiente a ajustarse mejor optimizando este proceso de distinta manera. Keras tiene más de 10 optimizadores implementados dentro de su librería, alguno de los más conocidos son Adam, RMSProp, SGD...
3. **Transfer Learning.** Los modelos de imagen ya son cargados con los pesos entrenado con Imagenet pero en algunos de los modelos se decidió poner ciertos bloques o capas de estas redes tan complejas como entrenables para que esta red fuera capaz de ajustarse un poco más al tipo de imágenes de nuestro proyecto a la vez que aprovecha todo lo "aprendido" al iniciar el entrenamiento con unos pesos ya ajustados.
4. **Técnicas para evitar el sobreajuste.** El sobreajuste es un fenómeno muy común en el mundo del deep learning y se produce cuando los pesos se ajustan demasiado a los datos que ha visto durante el entrenamiento perdiendo la capacidad de generalizar a datos nunca vistos por el modelo antes (datos de test). Para reducir este fenómeno se utilizaron distintas técnicas a lo largo de los distintos modelos:
    - *Dropout*: Esta técnica ignora la salida de cierto porcentaje de neuronas distintas en cada iteración en la capa en la que se aplica. Esto evita que ciertas neuronas se ajusten demasiado a cierta característica de las imágenes que ha visto y sea incapaz de generalizar.
    - *Batch Normalization*: Estandarización de las entradas de cada capa para cada lote (batch). Esto estabiliza el proceso de aprendizaje y reduce drásticamente el

número de epochs que necesita la red para converger por lo que también evita el sobreajuste

- *Early Stopping*: Mediante esta técnica se permite para el entrenamiento de la red si durante k epochs las métricas de validación no mejoran evitando que la red se siga ajustando a los datos de entrenamiento mientras con los de validación empeora o se estanca.

5.7. Resultados y validaciones

En la siguiente tabla se pueden observar la mayoría de los modelos que el equipo ha sido capaz de implementar y probar con éxito:

Nombre del Modelo	Batch size	Tama. Imag.[4]	Paramet. entrenab.	Epochs (es)[5]	Tiempo entren.	Train Loss	Train AUC	Val Loss	Val AUC	Test LRAP	Test MAP@ 3
baseline_model	128	300	11.236.174	3	0h6m	0,2859	0,6711	0,2884	0,6466	0,5486	0,5958
InceptionV3_01	128	150	28.686	10	22m18s	0,2853	0,7218	0,3038	0,6571	0,5275	0,5875
InceptionV3_03	128	150	2.789.006	10	19m14s	0,2900	0,6539	0,2858	0,6626	0,5615	0,6060
InceptionV3_03b	128	150	2.796.686	10	26m01s	0,2529	0,7903	0,2916	0,6850	0,5613	0,6135
InceptionV3_04	128	150	2.789.006	20	41m18s	0,2894	0,6544	0,2859	0,6665	0,5594	0,6079
InceptionV3_05	128	300	2.789.006	10	30m22s	0,2965	0,6068	0,2917	0,6343	0,5411	0,5826
InceptionV3_06b	128	224	15.548.878	25 (21)	103m5s	0,2572	0,7653	0,2667	0,7339	0,6203	0,6740
InceptionV3_06d	128	224	15.548.878	14 (10)	37m01s	0,2688	0,0735	0,2685	0,7359	0,6168	0,6641
InceptionV3_06e	128	299	1.556.558	14 (10)	61m56s	0,2653	0,7406	0,2644	0,7549	0,6369	0,6941
InceptionV3_07	128	224	15.548.878	15 (11)	37m18s	0,0296	0,6732	0,2867	0,6772	0,5913	0,6277
VGG16_01	128	224	200.347.894	10	26m02s	0,2925	0,6402	0,2860	0,6706	0,5594	0,5995
effNet_model	128	300	43.031.342	20	68m	0.6566x	0.4501	0.5323x	0.5040	0,2624	0,2497
denseNet	128	224	15.146.702	20 (8)	22m	0,2670	0,7293	0,2670	0,7349	0,6162	0,6746
denseNet_64bs	64	224	15.146.703	20 (8)	40m	0.2669	0.7506	0.2702	0.7191	0,6156	0,6717

Es importante explicar las métricas para evaluar los modelos y el porqué de la elección de estas antes de explicar un poco más en profundidad el modelo ganador:

- **AUC (*Area Under the curve*)**: Esta métrica da valor al área debajo de la curva ROC (Sensibilidad vs. Especificidad). Por lo tanto, evalúa el poder de un clasificador binario tanto para los casos positivos (1 o enfermedad) como para los negativos (0 o sano). Además, nuestro clasificador multiclase se puede ver como 14 clasificadores binarios por lo que el número final es obtenido promediando los 14 valores de AUC.
- **MAP@3 (*Mean Average Precision at 3)* - precisión**: Esta métrica permite evaluar la precisión para las K primeras clases (las K con probabilidades más altas) con valores entre 0 y 1: en otras palabras, simboliza el porcentaje de las enfermedades (que sufre el paciente) que han sido encontradas entre los K primeros resultados. En este caso se eligió K = 3 ya que las personas con hasta 3 enfermedades ya representan más del 90%

[4] Tamaño de las imágenes: Todas las imágenes de entradas fueron convertidas a imágenes cuadradas de 3 canales, por lo tanto 300 significa que la imagen era: (300,300,3)

[5] Epochs (early stopping): El valor entre paréntesis indica en que epoch fue parado el entrenamiento gracias a la técnica de Early Stopping explicad anteriormente.

del dataset. Con un ejemplo: si el enfermo sufre Neumonía y Nódulos, y nuestro modelo devuelve como probabilidades más altas, por orden: 1-Infiltración, 2-Nodulos, 3-Infiltración; MAP@3 = 1 ya que fue capaz de encontrar todas las clases positivas entre los 3 primeros resultados, pero no tiene en cuenta el **ranking** ya que, aunque no fueron encontradas como las dos más probables, esta métrica devuelve 1 ya que evalúa el grupo de K clases como un conjunto sin importar el orden. Por eso se analizó también **LRAP.** Aun así, esta métrica es muy útil ya que puede que el ranking no sea lo más importante debido a que es una modelo de ayuda al profesional médico y puede que el orden no sea decisivo mientras estas probabilidades estén en el top que se muestre al profesional y junto con su juicio y pruebas adicionales ayuden a tomar las decisiones adecuadas.

- **LRAP *(Label Ranking Average Precision)* - ranking**: Esta métrica nos aporta algo más de información sobre el ranking de las clases, es decir el orden en el que aparecen (de mayor a menor probabilidad) comparando con la salida deseada: las N primeras probabilidades deberían ser la N clases positivas, por lo tanto, las N enfermedades sufridas por el paciente. Para el ejemplo que se mostraba en el párrafo anterior (1-Infiltración, **2-Nodulos, 3-Infiltración**) donde conseguíamos una MAP@3 = 1, LRAP tendría un valor de 0.58 debido a que el que el score otorgado a Nódulos sería 1/2 y para infiltración seria 2/3 y por lo tanto: $\frac{\frac{1}{2}+\frac{2}{3}}{2} = 0.58$. Esto permite evaluar como de arriba están apareciendo las enfermedades sufridas por el paciente en la predicción.

Una vez elegidas las premisas para la comparación evaluación de los distintos modelos y tras un largo periodo de entrenamiento y validación, se pudo concluir que el mejor modelo:

Nombre del Modelo	Batch size	Tama. Imag.*	Paramet. entrenab.	Epochs (es)**	Ti empo entren.	Train Loss	Train AUC	Val Loss	Val AUC	Test LRAP	Test MAP@3
InceptionV3_06e	128	299	1.556.558	14 (10)	61m56s	0,2653	0,7406	0,2644	0,7549	0,6369	0,6941

Dicho modelo usa como codificador de la imagen la famosa red creada por Google: **InceptionV3** la cual recibe su nombre por la famosa película con el mismo nombre donde se pronuncia la frase *"We need to go deeper"*. Ir más profundo es lo que consiguieron con esta red paralelizando diferentes conjuntos de capas convolucionales y promediando sus salidas haciendo que estas funcionaran como una sola capa, pero mucho más compleja que se demostró que era capaz de capturar algunas características de los dataset que antes no se podían. Podemos ver esto en su arquitectura:

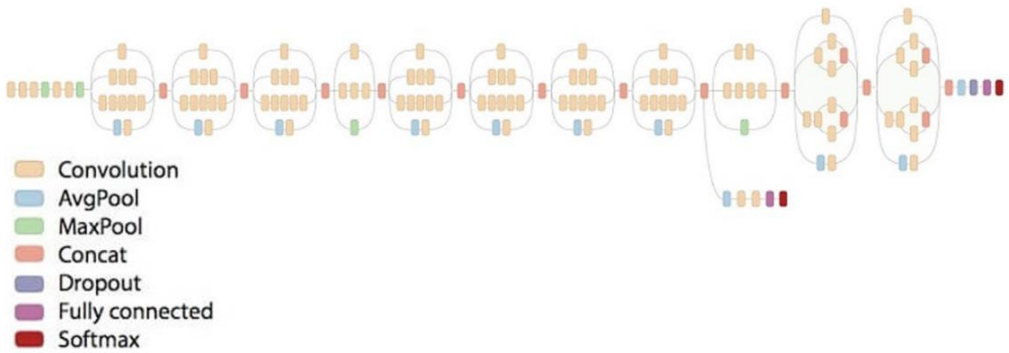

Fig. 11 Esquema de la red del modelo InceptionV3

Además, se sustituyó la última capa de clasificación softmax por una capa de *Global Max Pooling* con lo cual convertimos la salida de la antepenúltima capa (8,8,2048) en un vector de 2048 elementos que es será la representación vectorial de nuestra imagen. Como se mencionó anteriormente esta red esta entrenada con el dataset *Imagenet* por lo que se tomó la decisión de poner algunas de las capas convolucionales como entrenables para que los pesos (ya inicializados con un valor del entrenamiento previo) aprendieran sobre nuestro tipo de imágenes que difiera mucho del utilizado en su preentrenamiento y así conseguir moldear un poco la codificación de las imágenes al caso de clasificación objetivo de este proyecto.

A esta red se le concatenó una red *Fully Connected* con la siguiente estructura:

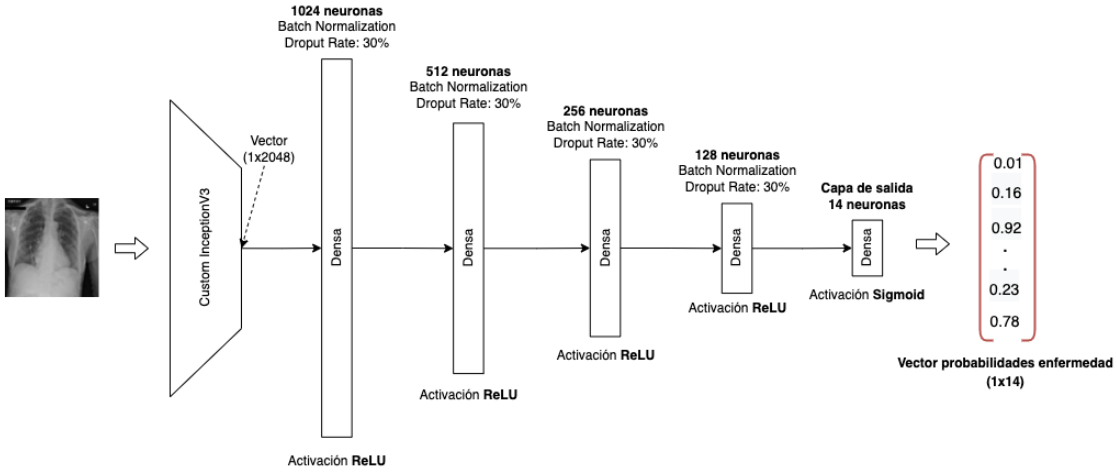

Fig. 12 Esquema de la red entrenada para el modelo ganador

Se puede observar como todas las capas densas tienen varias técnicas pare reducir el sobreajuste como Batch Normalization o Dropout.

Además, el entrenamiento del modelo se realizó con:

- Early Stopping fijándose en la Validation Loss. Como esta no mejoró desde la epoch 10 hasta la 14, el entrenamiento se paró y se guardaron los pesos de la epoch 10 donde se consiguió el mejor valor para el conjunto de validación (antes de que se produjera sobreajuste)
- El optimizador elegido fue *Adam*.
- La tasa de aprendizaje se fijó en 0.0001

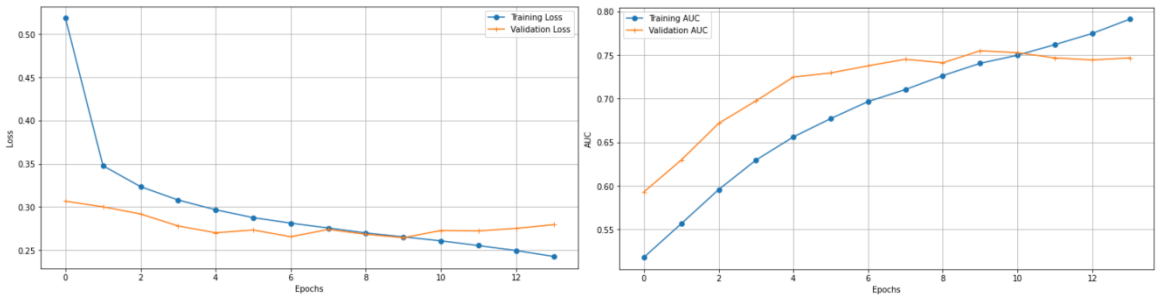

Fig. 13 Gráficas de pérdidas (izquierda) y AUC (derecha)

En las anteriores graficas se puede observar como la perdida y el AUC fueron evolucionando durante las epochs y como se paró el entrenamiento debido a que las métricas de entrenamiento y validación se empezaban a separar y se puede observar el punto donde se empezaba a producir el sobreajuste alrededor de la epoch 10-11.

Una vez realizado el entrenamiento se procedió a guardar el modelo y probarlo con datos no vistos anteriormente por la red: El dataset de prueba (*Test Dataset*)

Las dos métricas que utilizamos para evaluar el modelo (MAP@3 y LRAP) fueron calculadas para el conjunto de test ya que lo más importante es observar cómo se comporta el modelo con imágenes "nuevas".

Los valores de estas dos métricas para las 5429 muestras de test:

- MAP@3 = 0.6941
- LRAP = 0.6369

Esto quiere decir que el modelo es capaz de encontrar de media el 70% de las enfermedades en el top 3 de probabilidades. Este número no es ideal, pero por ejemplo MAP@5 = 0.85 y como en la salida del modelo de cara al usuario está planeado dar las 5 primeras probabilidades deja mucho menos margen de error. El LRAP es bastante parecido al MAP por lo que los rankings de las probabilidades no son muy malos de media, aunque esta métrica no tiene una explicación fácil promediada a lo largo de todos los miles de muestras por lo que se ha utilizado su valor para ver como mejoraba la capacidad de "rankear" las enfermedades mejor y en el caso de este modelo obtuvo su máximo valor.

5.8. Simulación de salida del modelo

Dentro de los límites de tiempo el equipo decidió crear una simulación de lo que, en una línea futura, podría ser la salida que podría ver el personal médico que usara esta herramienta. Esta funcionalidad recibiría las radiografías de los pacientes a analizar y sus etiquetas reales si ya han sido diagnosticados y se quiere comprobar la efectividad del modelo lo cual sería útil en sus primeros meses de implementación.

Con esto el modelo devolvería el top K enfermedades (5 en este ejemplo, K es elegido por el usuario) junto con la probabilidad asociada a cada una. En caso de haber dado también las enfermedades que realmente sufre cada paciente, se calcularía las métricas descritas como información para los técnicos en fases más adelantadas del proyecto.

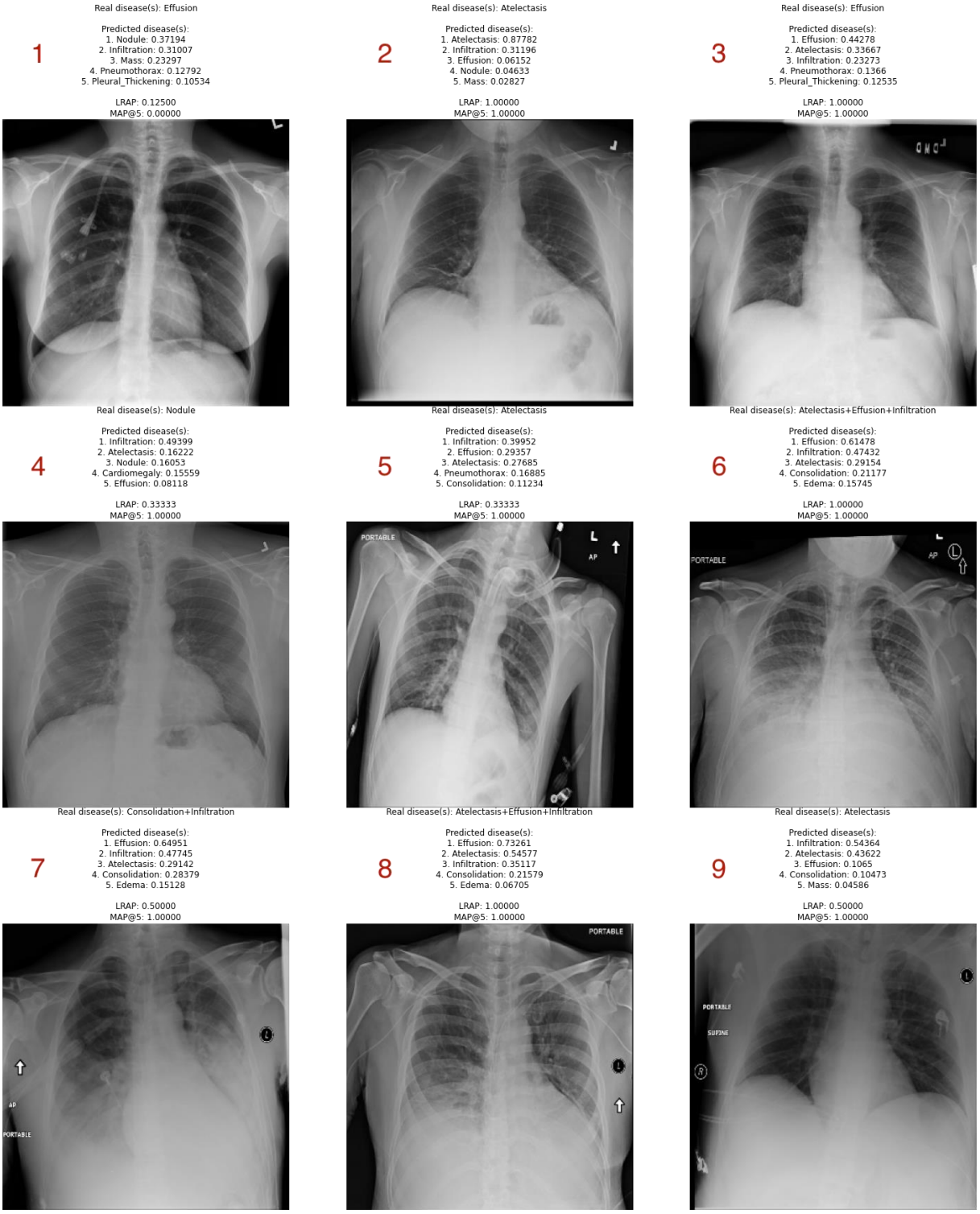

Fig. 14 Ejemplo de salida del modelo entrenado, en el que se detallan la enfermedad real, las 5 enfermedades más probables y las tasas de LRAP y MAP@5

Estas radiografías fueron cogidas de manera aleatoria del dataset de prueba y el texto en la parte superior de cada una de ellas define la información que se tiene sobre las patologías del paciente y las predicciones de la red para cada uno de ellos. Aquí podemos analizar distintos casos y sacar ciertas conclusiones rápidas:

- **Paciente 2**: se detecta a la perfección la única enfermedad que sufre el paciente con un 87% de probabilidad en el ranking 1 y con una gran diferencia respecto a los demás. MAP@5 = LRAP = 1.
- **Paciente 8**: sufre de 3 enfermedades que son detectadas en el ranking 1, 2 y 3 aunque en el caso de la patología "Infiltración" no recibe una gran probabilidad: 35%. MAP@5 = LRAP = 1.
- **Paciente 9**: la enfermedad que padece es detectada en segundo lugar, pero se puede apreciar que las dos primeras enfermedades "rankeadas" tienen una probabilidad muy superior al resto de patologías. Estos son los casos donde la experiencia y sabiduría de un/a médico/a ayudará a discernir cuál de las dos es la que padece realmente el paciente.
- **Paciente 1**: también es destacable que el modelo no es capaz de detectar la enfermedad correcta. Además, por su valor de LRAP, ya que este caso solo hay una enfermedad objetivo, es fácil ver que esta ha sido predicha en el ranking 8 ya que 1/0.125 = 8. Vemos como el modelo, obviamente falla de manera muy grande con algunos pacientes por lo que, por lo delicado del tema a tratar, este modelo debería de ser mejorado para poder ponerse en producción.

También es importante remarcar que la versión de este modelo no funcionaría bien para pacientes sin ninguna patología ya que fueron descartados a la hora del procesamiento del dataset de entrenamiento por lo que el modelo tendrá siempre un sesgo hacía dar probabilidades relativamente altas ya que nunca "vio" imagines de pacientes donde el vector objetivo está formado por todo "0"s. Una buena mejora sería cambiar el entrenamiento para que fuera capaz de leer este tipo de vectores. Por el momento esta herramienta (en esta versión) podría ser usada para una vez que el médico ya ha hecho las pruebas necesarias y puede que tenga alguna duda se pueda apoyar en la información dada por nuestra red.

6. Líneas futuras

6.1. Productivización del modelo en un Hospital

"Ninguno de nosotros es tan bueno como todos nosotros juntos." - Ray Kroc

Con el modelo el personal sanitario recibirá una orientación y una detección más certera que implicarán menos pruebas sobre el paciente y más tiempo disponible de ellos para atender y llegar a más personas.

6.2. Beneficiar a pacientes en países subdesarrollados

El proyecto tiene la posibilidad de mostrar un lado más 'humano' y que llegue a aportar un granito más hacia un mundo mejor. Es una gran oportunidad para personas que en unas condiciones precarias tengan, o bien: la opción de detectar la enfermedad de forma precoz y conseguir tratarla o poder tener la opción a detectarla en primer lugar.

6.3. Crear una radiología virtual

Hay un gran abanico de enfermedades, las de corazón, por ejemplo, que se detectan mediante otras pruebas, véase: imágenes de perfusión miocárdica (MPI), resonancias magnéticas (IRM) y un largo etcétera. Expandiendo tanto el campo de las enfermedades a detectar como el de las pruebas de entrada, se pueden lograr resultados muy positivos en la mortalidad de las personas.

6.4. Ensamblado con modelo de datos cuantitativos y cualitativos

La predictibilidad se podría mejorar si además de aportar la radiografía se aportasen otros datos como como la edad, sexo o hábitos del paciente. De esta manera, podríamos mejorar el modelo creado en este trabajo.

7. Conclusiones

7.1. Análisis

Se tiene un punto de partida complicado para la generación de modelos, pues el número de imágenes es muy ajustado, son de diferentes posturas corporales y hay enfermedades de las que no se tiene apenas imágenes.

La población de estudio no se correlaciona con la población real, de este análisis se concluye que entre los 40 y los 60 están el mayor número de enfermos de tórax. Si vemos los datos (gráfica de densidades) la edad influye en caer enfermo. Son las edades comprendidas entre 45-65 años las que más enfermas están. Por esta razón la muestra tomada para la red no debería estar sesgada en relación a la edad sobre la población a la que va dirigida.

Se concluye con que el sexo y la posición de la radiografía influyen en la diagnosis del paciente, los hombres son más propensos a enfermar que las mujeres y las radiografías AP es más probable que tengan alguna enfermedad.

7.2. Modelos

El equipo se ha esforzado al máximo y el mayor LRAP conseguido ha sido 0.6369, para una herramienta médica no se considera un valor suficientemente elevado como para hacer diagnósticos de forma independiente, el uso de estos resultados debería estar supervisado por un experto. Creemos que con un mayor número de imágenes se podrían haber alcanzado mayor precisión.

8. Bibliografía

Brady, A. P. (7 de Dec de 2016). *Error and discrepancy in radiology: inevitable or avoidable?* Obtenido de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5265198/#:~:text=Errors%20and%20discrepancies%20in%20radiology,reported%20in%20many%20targeted%20studies.

Estadística, I. N. (s.f.). *INE*. Obtenido de https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176780&menu=ultiDatos&idp=1254735573175

J.D., G. J., & F.L., G. M. (s.f.). *TAC, RMN y PET en enfermedades torácicas.* Obtenido de https://www.neumosur.net/files/publicaciones/ebook/3-TAC-ENFERMEDADES-Neumologia-3_ed.pdf

Marcel van Gerven, S. B. (19 de December de 2019). *Artificial Neural Networks as Models of Neural Information Processing.* Obtenido de https://www.frontiersin.org/research-topics/4817/artificial-neural-networks-as-models-of-neural-information-processing

Navarra, C. U. (s.f.). *Clínica Universidad de Navarra*. Obtenido de https://www.cun.es/enfermedades-tratamientos/enfermedades/dolor-toracico

https://www.topdoctors.es/

https://www.mayoclinic.org/es-es

https://www.cancer.org/