UNIVERSITAT POLITÈCNICA DE VALÈNCIA

MIARFID Official Master's Degree in Artificial Intelligence, Pattern Recognition and Digital Imaging

Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València

Detección automática de neumonía para dar soporte al diagnóstico de la COVID-19

TRABAJO FIN DE MÁSTER

Máster Universitario en Inteligencia Artficial, Reconocimento de Formas e Imagen Digital

Autor: Rafael Vicente Sánchez Romero

Tutores: Jon Ander Gómez Adrián
Roberto Paredes Palacios

Curso 2019-2020

Resum

L'anàlisi d'imatge mèdica és, cada vegada més, una eina que s'està convertint en un suport per als metges a l'hora d'emetre un diagnòstic. En aquest treball, abordem el tema de l'anàlisi d'imatge mèdica per tractar de resoldre el problema de la detecció de pneumònia per COVID-19 a diferents tipus d'imatge mèdica, com ara imatges Raigs-X o imatges per resonància magnètica. Per a aconseguir els objectius es proposaràn models ad-hoc i altres que han sigut publicats i es tractará de fer una comparativa entre models i entre tipus de dades per tal de extraure conclusions.

Paraules clau: Inteligència Artificial, Xarxes Neuronals, Imatge mèdica, Pneumònia, COVID-19, Diagnòstic mèdic.

Resumen

El análisis de imagen médica es, cada vez más, una herramienta que se está convirtiendo en un apoyo para el personal médico a la hora de emitir un diagnóstico. En este trabajo, abordamos el tema del análisis de imagen médica para tratar de resolver el problema de la detección de neumonía por COVID-19 en diferentes tipos de imagen médica como por ejemplo imágenes Rayos-X o imágenes por resonancia magnética. Para conseguir los objetivos se propondrán modelos ad-hoc y otros que han sido publicados y se tratará de hacer una comparativa entre modelos y tipos de datos para extraer conclusiones.

Palabras clave: Inteligencia Artificial, Redes Neuronales, Imagen Médica, Neumonía, COVID-19, Diagnóstico médico.

Abstract

Medical imaging analysis is a tool which is increasingly becoming a support for doctors when it comes the time to issue a diagnosis. In this work, we address the medical images analysis topic aiming to solve the problem of detecting COVID-19 pneumonia in different types of medical images, such as X-Ray or magnetic resonance images. To achieve the proposed objectives, ad-hoc models are proposed and already published models are also used. In this work, different models and type of images are compared so that some conclussions are reached.

Key words: Artificial Intelligence, Neural Networks, Medical Image, Pneumonia, COVID-19, Medical diagnosis.

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducción

En los últimos tiempos, una nueva familia de virus ha ganado popularidad como consecuencia de una pandemia mundial con un impacto severo en el mundo. Los coronavirus son una familia de virus que suponen una potencial amenaza para el ser humano. En 2003, los coronavirus causaron una gran pandemia, la pandemia del denominado *severe acute respiratory syndrome* (SARS). Una segunda crisis con origen en la misma familia de virus tuvo lugar unos años más tarde, en 2012, en Arabia Saudí, causando el *middle east respiratory syndrome* (MERS).

Una nueva enfermedad por coronavirus se originó en Wuhan, China en 2019. Esta nueva enfermedad es conocida bajo el nombre de COVID-19 y, según la Organización Mundial de la Salud (OMS), hasta el 27 de julio de 2020 se habían detectado 16.114.449 personas infectadas y 646.641 fallecidas [6]. El 30 de enero de 2020, la OMS declaró la COVID-19 como una emergencia sanitaria pública de interés internacional. A partir del 11 de marzo de 2020 se considera a la COVID-19 como una pandemia [7].

En este contexto, muchos investigadores e investigadoras en los campos de inteligencia artificial y medicina están buscando métodos para prevenir la propagación de esta enfermedad. Como no existe disponibilidad de un tratamiento preventivo de la enfermedad, es muy importante detectar con rapidez a los pacientes de esta enfermedad y aislarlos para romper la cadena de transmisión del virus.

1.1 Motivación

Algunos estudios médicos evidencian que los pacientes de la COVID-19 suelen sufrir de infecciones pulmonares [8]. En la gran mayoría de casos, esta infección se expresa en forma de neumonía.

La inteligencia artificial y, más concretamente, el aprendizaje profundo ayudan al personal médico a reducir el tiempo de diagnóstico de la COVID-19. A parte de las ampliamente conocidas pruebas médicas, como las qPCR o los tests serológicos, los doctores también tienen en cuenta pruebas de imagen médica.

El tipo más común de pruebas de imagen médica son las basadas en Rayos-X y las basadas en resonancia magnética (MRI). El uso de un tipo u otro lleva asociadas ventajas e inconvenientes. Hay tres razones principales por las que usar imágenes Rayos-X sobre imágenes obtenidas por resonancia magnética, especialmente en áreas gravemente afectadas por la pandemia o con falta de recursos, como se indica en [9]:

- **Triaje rápido:** Las imágenes Rayos-X permiten un triaje rápido de los pacientes sospechosos. Este tipo de prueba médica puede realizarse en paralelo con tests bio-

lógicos como qPCR o serológicos, que usualmente requieren de más tiempo para realizarse. Debido a esta rapidez, los sistemas de salud pueden experimentar un alivio, especialmente en aquellas áreas en las que la presión sanitaria excede la capacidad del sistema de salud (Italia, España o Nueva York) o incluso en áreas donde los test biológicos no están disponibles por la alta demanda de los mismos que se puede dar en los picos de contagio de la enfermedad.

- **Disponibilidad y accesibilidad:** Como los equipamientos Rayos-X se consideran básicos para los hospitales en muchos países, la disponibilidad y accesibilidad de este tipo de pruebas es mayor que las de resonancia magnética. Además, los altos costes de equipamiento y mantenimiento que tienen las máquinas de resonancia magnética, hace imposible su disponibilidad en países en vías de desarrollo.

- **Portabilidad:** Los hospitales suelen contar con algunas máquinas de Rayos-X portátiles, haciendo posible realizar las pruebas en la habitación de los pacientes. Con ello se evita el movimiento por dentro del hospital de pacientes que puedan estar infectados, reduciendo, por tanto, la transimisión del virus.

Sin embargo, para el personal médico es muy complejo detectar biomarcadores de neumonía por COVID-19 en imágenes Rayos-X. Por esta razón, cuando tienen la menor duda de que el paciente pueda estar infectado, deben realizar una prueba MRI para dar el diagnóstico final. Así pues, aunque realizar las pruebas MRI es más costoso en término de tiempo y dinero, en muchos casos se hace necesario para poder evaluar la situación de cada sujeto.

Así pues, en este trabajo utilizaremos ambos tipos de datos para comprobar cómo éstos pueden afectar al comportamiento de un clasificador que deba distinguir entre sujetos sanos de aquellos que padecen de neumonía por COVID-19.

1.2 Trabajos previos

El área del análisis de imagen biomédica (segmentación y clasificación) está reconocida como una de las áreas clave que hacen los sistemas de salud más prometedores [10].

Debido a la gran disponibilidad de máquinas de Rayos-X que hay en los hospitales, las imágenes Rayos-X de pulmón son una herramienta ampliamente utilizada por el personal médico. Son más rápidas de obtener que los tests biológicos, por lo que sería de gran utilidad poder detectar neumonía en este tipo de imágenes.

Hoy en día existen varios datasets públicos disponibles con los que los investigadores han realizado sus trabajos sobre el diagnóstico de neumonía y otras enfermedades [11, 12, 13, 14]. Algunos de estos trabajos muestran la necesidad de herramientas automáticas que aceleren el diagnóstico ya que, actualmente, el análisis de imágenes Rayos-X tiene un inconveniente: se necesita mucho tiempo para analizarlas así que no suelen utilizarse por falta de técnicos especialistas en máquinas Rayos-X ni radiólogos.

En esta situación, la aplicación de técnicas de aprendizaje profundo aparece como una alternativa factible. Merece la pena resaltar el trabajo llevado a cabo por Rajpurkar et al. [15]. En él se propone un nuevo modelo llamado CheXNet basado en una DenseNet de 121 capas y entrenado con el dataset ChestX-ray14. Además, en el mismo trabajo se comparan los diagnósticos emitidos por el modelo y por algunos de los mejores radiólogos de Stanford Medicine. Se descubre que la CheXNet emite mejores diagnósticos que los radiólogos en la detección de neumonía y supone, por tanto, una mejora de los resultados que conforman el estado del arte en la materia. Con respecto a la tasa de acierto obtenida

con el subconunto de test para la detección de neumonía no vírica, el modelo mencionado obtiene un 76,80 % varios puntos porcentuales por encima de los mejores modelos hasta el momento, mejorando también en este aspecto los resultados que conformaban el estado del arte en ese problema.

Un año después, en [16], los autores trabajan en la detección de neumonía en imágenes de Rayos-X mediante el uso de transfer learning en una red neuronal preentrenada. Los resultados obtenidos corresponden a varios casos de uso como degeneración macular relativa a la edad, o neumonía pediátrica donde la tasa de acierto en el subconjunto de test alcanzó el 92,80 %. En un trabajo similar [17], los autores muestran una mejora en la tasa de acierto en test alcanzando el 93,73 %. Esto fue posible utilizando una red neuronal convolucional (CNN) con cuatro capas convolucionales y dos densas junto con el uso de algunas técnicas de data augmentation sobre las imágenes.

Otras propuestas han mezclado aprendizaje profundo con modelos estadísticos para mejorar los resultados obtenidos. En [18], los autores usan una ResNet50 [2] preentrenada con el dataset ImageNet [19] con el objetivo de extraer características más profundas de las imágenes de entrada. Una vez extraídas, se introducen en una máquina de vectores soporte (SVM) y se realiza la clasificación. Siguiendo esta aproximación, se mejora la tasa de acierto en test hasta alcanzar el 95,38 %.

La COVID-Net, un nuevo modelo basado en CNN se propone en [9] y es capaz de clasificar pacientes en cuatro clases: control, neumonía viral, no-COVID19 y COVID19. Este modelo, teniendo un número considerablemente menor de parámetros entrenables comparado con modelos como la VGG-19 o la ResNet50, obtiene un resultado muy competitivo llegando a un 93,30 % en tasa de acierto en test.

Existen soluciones más complejas como la propuesta en [20]. Los autores proponen el método DeTraC (descomposición, transfer learning y composición) para resolver un problema de clasificación en tres clases (normal, COVID y SARS). Primero, utilizan una AlexNet [21] preentrenada con ImageNet para extraer características de las imágenes. Una vez extraídas se aplica k-medias con $k = 2$ sobre ellas. Por tanto, cada una de las clases se divide en dos subclases más específicas. Para resolver el nuevo problema de clasificación en seis clases se utiliza una ResNet18 preentrenada. Finalmente, la clasificación se reconvierte al problema inicial de clasificación en tres clases para dar la predicción final. Con este método se consigue una tasa de acierto en test del 95.12 %.

1.3 Resumen de la aproximación propuesta

La mejora de la infraestructura informática en los hospitales ha hecho posisble el desarrollo de técnicas de deep learning que facilitan algunas tareas complejas de análisis de imagen médica. Dentro del campo del aprendizaje profundo, las redes neuronales convolucionales se han impuesto como la técnica más efectiva para tal fin.

Algunas investigaciones recientes descubren que las imágenes Rayos-X de pulmón de pacientes que sufren COVID-19 muestran algunas leves anomalías en la radiografía.

En este trabajo, se propondrán modelos ad-hoc basados en redes neuronales convolucionales y se entrenarán y evaluarán para desubrir las ventajas e inconvenientes que se observan con respecto a los modelos que conforman el estado del arte en la materia. Además, se entrenarán otros modelos con imágenes obtenidas por resonancia magnética para comparar los resultados obtenidos y extraer algunas conclusiones.

1.4 Contexto de aplicación

Como estamos trabajando con un problema médico y el propósito final del trabajo es introducir los modelos en el proceso de diagnóstico médico, necesitaremos seleccionar los mejores modelos basados, no sólo en la tasa de acierto que obtengan, sino en otras métricas como la precisión, el recall, el F1-score, el área sobre la curva ROC (AUC) y el ratio de falsos negativos (FNR).

El objetivo es conseguir modelos no sólo con una tasa de acierto alta sino con un ratio de falsos negativos bajo y un alto F1-score. Con ello, estaremos seguros de que los modelos apenas clasifican pacientes con COVID-19 como sanos. Todas estas métricas nos proporcionarán un mejor conocimiento del estado de los pacientes, facilitando y acelerando el diagnóstico.

La tarea definida con imágenes Rayos-X no es fácil. Esto se debe a que la mayoría de imágenes de la clase COVID-19 fueron tomadas durante la crisis sanitaria de una manera rápida para evitar un colapso del sistema sanitario. Así pues, el preprocesado de las imágenes será importante para poder obtener buenos resultados.

1.5 Objetivos

En este trabajo final de máster se definen tres objetivos:

- Analizar el problema de la detección automática de neumonía por COVID-19 y evaluar su dificultad dependiendo del tipo de imágenes que se utilicen para abordarlo (Rayos-X o MRI)
- Analizar y probar distintas técnicas de preprocesado de datos para decidir, en base a los resultados obtenidos, cuál o cuales son las más apropiadas para el problema que se aborda en este trabajo.
- Diseñar varios modelos de redes neuronales convolucionales y evaluar los resultados obtenidos.
- Utilizar modelos publicados por otros autores previamente y evaluar los resultados obtenidos.

CAPÍTULO 2
Datos y herramientas utilizados

2.1 Datasets

2.1.1. BIMCV COVID-19

El Banco de Imágnes Médicas de la Comunidad Valenciana (BIMCV), hizo público, hace unos años, un dataset llamado PADCHEST [22] compuesto por imágenes Rayos-X de pulmón pertenecientes a sujetos control y pacientes sufriendo de neumonía. Este dataset está compuesto por un total de 160.868 imágenes correspondientes a 67.625 pacientes diferentes. De todas las imágenes, 39.039 fueron etiquetadas manualmente por personal médico.

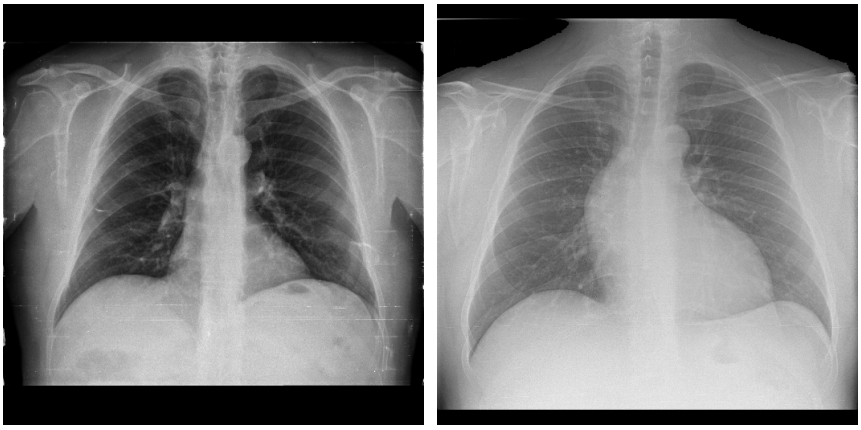
Tras un análisis exploratorio de los datos, sólo 23.244 imágenes resultaron válidas para ser usadas en nuestro problema. De todas ellas sólamente 11.989 fueron declaradas aptas para solucionar nuestro problema, aquellas tomadas siguiendo el *"Postero-Anterior Standard"(PA-AP)*. La razón por las que seleccionar únicamente las imágenes tomadas siguiendo el mencionado estándar es porque, en imágenes Rayos-X, es el estándar dónde mejor se pueden detectar biomarcadores de neumonía por COVID-19. De todas estas imágenes solo se utilizan las 6.871 correspondientes a la clase control. Una muestra de cada clase se puede ver en la Figura 2.1.

Durante la primera ola de la COVID-19 en España, el BIMCV hizo pública una extensión de este dataset bajo el nombre BIMCV PADCHEST COVID-19 [23], introduciendo nuevas imágenes Rayos-X correspondientes a pacientes sufriendo de neumonía causada por la COVID-19. La extensión está formada por 5.587 imágenes de 1.355 pacientes diferentes. Tras un análisis exploratorio del dataset, sólo 1.051 imágenes pertenecen al estándar anteriormente mencionado y a la clase "COVID-19" y son las que se utilizarán junto con las anteriormente seleccionadas.

Mezclando ambos tipos de imágenes, podremos crear experimentos en los que se entrenarán clasificadores cuyo objetivo será distinguir sujetos control de pacientes de COVID-19. La Tabla 2.1 contiene el resumen de las imágenes por partición y por clase.

2.1.2. CC-CCII

Este año, en un esfuerzo por luchar contra la pandemia en el país de origen, China, el China Consortium of Chest CT Image Investigation (CC-CCII) construye un largo dataset con imágenes obtenidas por resonancia magnética de sujetos control, pacientes con neumonía vírica y pacientes con neumonía causada por la COVID-19 [24]. Las imágenes están tomadas de manera axial, puesto que es la perspectiva dónde mejor se ven



(a) Imagen Rayos-X en estándar PA-AP de un sujeto control (sujeto sano)

(b) Imagen Rayos-X en estándar de un paciente que sufre neumonia por COVID-19

Figura 2.1: Imágenes del dataset BIMCV PADCHEST COVID-19.

Tabla 2.1: Distribución de las imágenes del dataset BIMCV PADCHEST COVID-19 en los subconjuntos de entrenamiento, validación y test, separados en las diferentes clases definidas por los radiólogos.

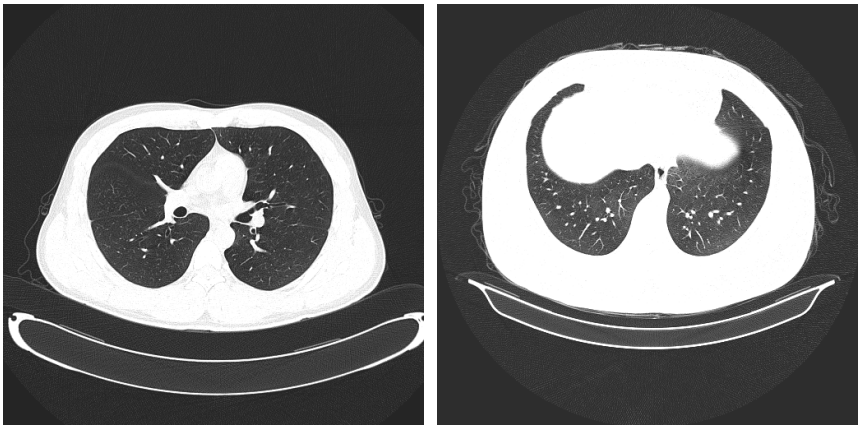
	Partición			
Grupo (id)	Entrenamiento	Validación	Test	Total
Control (C)	6539	166	166	6871
COVID-19 (COV)	657	193	201	1051
Total	7196	359	367	7922

los biomarcadores de la neumonía por COVID-19 en imágenes obtenidas por resonancia magnética.

Este dataset se compone de sesiones de resonancia magnética provenientes de 4.154 pacientes diferentes. En total se cuentan 617.775 imágenes. Tras un análisis exploratorio de los datos, se declaran aptas para nuestro problema 252.827 imágenes correspondientes a las clases COVID-19 y sujeto control. En la Tabla 2.2 se muestra el número de imágenes separadas por clase y partición. Además, en la Figura 2.2 se muestran imágenes de cada clase.

Tabla 2.2: Distribución de las imágenes del dataset CC-CCII en los subconjuntos de entrenamiento, validación y test, separados en las diferentes clases definidas por los radiólogos.

	Partición			
Grupo (id)	Entrenamiento	Validación	Test	Total
Control (C)	57294	18444	20018	95756
COVID-19 (COV)	89762	31855	34454	156071
Total	147056	50299	54472	252827



(a) Corte de una sesión de resonancia magnética en estándar axial de un sujeto control (sujeto sano)

(b) Corte de una sesión de resonancia magnética en estándar axial de un paciente con neumonía por COVID-19

Figura 2.2: Imágenes del dataset CC-CCII.

2.1.3. COVID-CTset

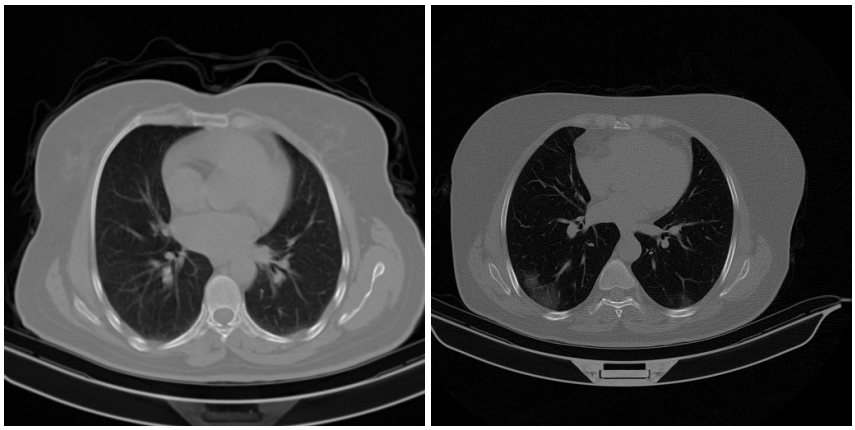
El COVID-CTset se crea expresamente para llevar a cabo la investigación presentada en [1]. Este dataset público se compone de imágenes tomadas del centro Nagin de radiología en Sari, Irán. Las imágenes se tomaron durante la primera ola de la COVID-19 entre Marzo y Abril de 2020. Las imágenes se toman por resonancia magnética con perspectiva axial.

Tabla 2.3: Distribución de las imágenes del dataset COVID-CTset en los subconjuntos de entrenamiento, validación y test, separados en las diferentes clases definidas por los radiólogos.

	Partición		
Grupo (id)	Entrenamiento	Test	Total
Control (C)	7742	2032	9774
COVID-19 (COV)	1931	350	2281
Total	9673	2382	12055

El dataset se compone por 2.281 imágenes de 95 pacientes diferentes de neumonía por COVID-19 y por 9.774 imágenes de 282 sujetos control que hacen un total de 12.055 imágenes. Las imágenes están en formato TIFF en escala de grises de 16 bits. Debido a la escasez de los datos se decide realizar una partición en dos subconjuntos: entrenamiento y test con un 80 % y un 20 % de los datos, respectivamente. La información relativa a las particiones se muestra en la Tabla 2.3.

Además, en este dataset se dispone de información demográfica de los sujetos control y de los pacientes. La distribución de pacientes por rangos de edad, sexo e infección se muestra en la Figura 2.4.



(a) Corte de una sesión de resonancia magnética en estándar axial de un sujeto control (sujeto sano)

(b) Corte de una sesión de resonancia magnética en estándar axial de un paciente con neumonía por COVID-19

Figura 2.3: Imágenes del dataset COVID-CTset.

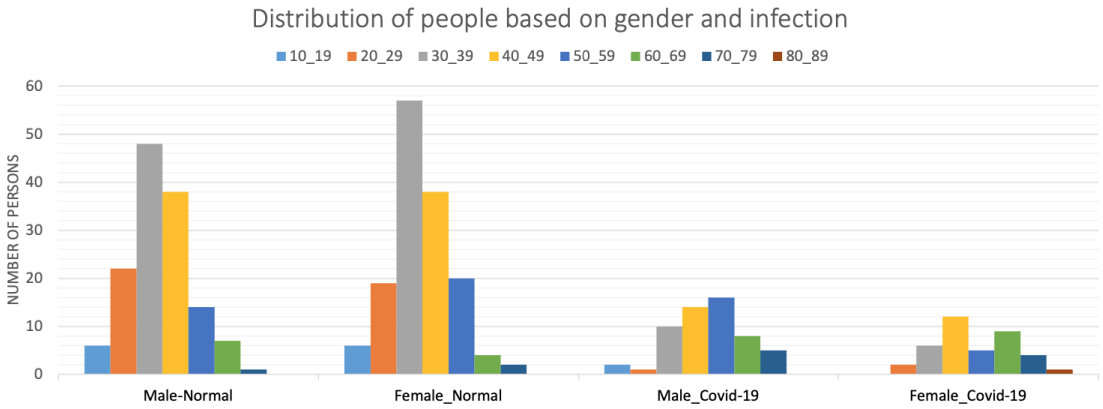


Figura 2.4: Información demográfica del dataset COVID-CTset. Distribución de sujetos por sexo, rangos de edad e infección. Imagen extraída de [1].

CAPÍTULO 3
Metodología

3.1 Definiciones

En esta sección definiremos algunas métricas que se usarán para evaluar el comportamiento de los modelos evaluados. Las siguientes métricas están basadas en la predicción de los clasificadores comparados con la clase real de las muestras. Comparando ambas, podemos obtener la matriz de confusión. Un ejemplo de este tipo de matriz se muestra en la Tabla 3.1.

Profundicemos en cada una de las posibilidades que se dan en la matriz de confusión:

- **Verdadero Positivo (TP):** Una muestra correctamente clasificada. La clase predicha por el clasificador (positivo) coincide con la clase real de la muestra (positivo).
- **Falso Negativo (FN):** Una muestra clasificada incorrectamente. La clase predicha (negativo) no coincide con la clase real de la muestra (positivo).
- **Falso Positivo (FP):** Una muestra clasificada incorrectamente. La clase predicha (positivo) no coincide con la clase real de la muestra (negativo).
- **Verdadero Negativo (TN):** Una muestra clasificada correctamente. La clase predicha (negativo) coincide con la clase real de la muestra (negativo).

En el caso del problema con el que estamos trabajando, los verdaderos positivos serán las muestras de la clase COVID-19 correctamente clasificadas como COVID-19, los verdaderos negativos serán aquellas muestras correctamente clasificadas como control, los falsos positivos serán las muestras control clasificadas como COVID-19 y los falsos negativos, las muestras COVID-19 clasificadas como control.

3.1.1. Precisión

La precisión de un clasificador se calcula como la fracción de verdaderos positivos sobre todas las detecciones hechas por el clasificador (Ecuación 3.1). En el caso particular

Tabla 3.1: Matriz de confusión mostrada para un problema de clasificación en dos clases.

		Predicho	
		Positivo	**Negativo**
Real	**Positive**	TP	FN
	Negativo	FP	TN

de este problema, se calcula como el número de pacientes COVID-19 detectados correctamente sobre el total de pacientes.

$$Precision = \frac{TP}{TP + FP} = \frac{TP}{Todas\ las\ detecciones} \tag{3.1}$$

3.1.2. Recall

El recall de un clasificador se calcula como la fración de verdaderos positivos sobre el total de positivos reales (Ecuación 3.2). En el caso de nuestro problema, se calcula como el número de pacientes correctamente detectados como COVID-19 sobre el total de pacientes COVID-19 reales (detectados correctamente o no por el clasificador).

$$Recall = \frac{TP}{TP + FN} = \frac{TP}{Muestras\ COVID - 19} \tag{3.2}$$

3.1.3. F1-Score

El F1-socre relaciona ambas, precisión y recall, en una sola métrica que es la media harmónica de ellas (Ecuación 3.3).

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3.3}$$

3.1.4. Área bajo la curva ROC

Podemos expresa la capacidad de un clasificador para discriminar entre dos clases mediante la *Receiver Operating Characteristic curve* (ROC curve). Esta curva se crea mostrando el recall contra el ratio de falsos positivos (FPR Ecuación 3.5) estableciendo un umbral y variándolo.

Para medir la capacidad de discriminación de un clasificador, debemos calcular el área bajo la curva ROC (AUROC). Cuanto mayor sea el área, mejor será la capacidad discriminativa del clasificador.

$$FPR = 1 - \frac{FP}{TN + FP} \tag{3.4}$$

3.1.5. Ratio de falsos negativos

El ratio de falsos negativos (FNR) se calcula como la fracción de falsos negativos (pacientes COVID-19 clasificados como control) sobre el total de pacientes.

$$FNR = \frac{FN}{TN + FN} \tag{3.5}$$

3.2 Modelos utilizados

En esta sección se describen los modelos propuestos junto con la explicación de su propuesta para el problema resuelto. Además se profundiza en otros dos modelos de red neuronal propuestos por otros autores y se explican sus características.

De todos los modelos presentados se establecen hipotéticos pros y contras que pueden tener a la hora de resolver el problema que nos atañe y que, más tarde, se comprobarán en el apartado de experimentación.

3.2.1. Modelos propuestos ad-hoc

Los modelos propuestos para este problema están basados en redes neuronales convolucionales (CNN), que son un tipo de redes neuronales profundas (DNN) comunmente utilizadas para analizar imágenes. Se consideran versiones regularizadas del preceptron multicapa (MLP) y se inspiran en el proceso biológico por el cual la conectividad de las neuronas artificiales emula la organización del cortex visual humano.

En la Tabla 3.2 se detallan las topologías de los modelos propuestos. Todos los modelos se identifican con un número y una letra (p. ej. 6b) donde la letra 'b' identifica una variante del modelo 6. Todos los modelos con el mismo número utilizan exactamente el mismo tipo de bloques que realizan la reducción gradual del tamaño de las imágenes y el mismo tipo de capas densas tras la parte convolucional. En general, las variaciones de un modelo tienen que ver con el tamaño del kernel utilizado. Los tamaños de kernel de $7x7$ y $9x9$ en la primera capa convolucional llevan a mejores resultados que otros tamaños como $3x3$ o $11x11$. Se puede observar como algunos modelos tienen más de una capa convolucional aplicada a la entrada de la red con tamaños de kernel diferentes. Los resultados de estas capas iniciales se concatenan antes de la secuencia de bloques de reducción. Es el caso de los modelos *4a*, *4b*, *4c* y *4d*. Por otro lado, el modelo *6a* es un caso especial debido a la concatenación que se realliza después de las dos secuencias de bloques de reducción. Este es un modelo con dos ramas paralelas con las que se tiene como objetivo detectar marcadores de la enfermedad a diferentes escalas. Los resultados de cada rama se concatenan al final para proceder a la clasificación final en la parte densa de la red. La función de activación elegida para las capas de salida es la función Softmax, mientras que para las capas ocultas y la mayoría de las convolucionales se ha utilizado la función ReLU.

3.2.2. MobilenetV2 with Feature Pyramid Network

Hasta 2015 los modelos más potentes como la VGG [25] eran suficientemente profundas como para que en el momento de la retropropagación del error, la derivada parcial del error en las capas más próximas a la entrada fuera cercana a cero, haciendo que la modificación de los pesos en esas capas fuera ínfimo y, por tanto, reduciendo la capacidad de aprendizaje de las mismas. Este problema es lo que se conoce como desvanecimiento del gradiente y, en muchas ocasiones, impide la convergencia de los modelos.

Lo que los autores de [2] propusieron es hacer modelos aún más profundos, pero rompiendo la secuencialidad de los mismos. Esto se consiguió con las denominadas conexiones residuales, que son caminos por los que la información "circula" con mayor facilidad. Con estas conexiones residuales que se "saltan" algunas capas intermedias, se consigue (a) poder construir modelos más profundos y aprovechar esa profundidad para resolver los problemas y (b) evitar el problema del desvanecimiento del gradiente debido a las conexiones residuales que mediante *identity mapping* evitan que durante la retropropagación del error, éste caiga a valores cercanos al cero. Una representación gráfica de este tipo de conexiones se muestra en la Figura 3.1.

Un año después, en 2016, en [3] se propone una mejora sobre la ResNet, dando lugar a la ResNet V2. La diferencia con respecto a la versión inicial tiene que ver con el momento en el que se realizan las activaciones de las neuronas. En la Figura 3.2 se muestra

Tabla 3.2: Topologías detalladas de todos los modelos propuestos y evaluados en la parte experimental de este trabajo.

Model 3a	**Model 4a**	**Model 4b // Model 4c // Model 4d**
$7 \times 7, 32, stride = 2$	$\begin{bmatrix} BN \\ 7 \times 7, 32, \\ stride = 2 \end{bmatrix} \times 2$	$5 \times 5, 64, stride = 2$ // $7 \times 7, 64, stride = 2$ // $9 \times 9, 64, stride = 2$ | $7 \times 7, 64, stride = 2$ // $9 \times 9, 64, stride = 2$ // $11 \times 11, 64, stride = 2$
BlockM3(**32**) = $\begin{bmatrix} 1 \times 1, \mathbf{32} \\ 3 \times 3, \mathbf{32} \\ 1 \times 1, 4 \cdot \mathbf{32} \\ MaxPool 2 \times 2 \end{bmatrix}$	MaxPool 3×3	CONCAT
BlockM3(64) BlockM3(128) BlockM3(256) ×2	BlockM4A = $\begin{bmatrix} BN - 1 \times 1, 32 \\ BN - 3 \times 3, 32 \end{bmatrix} \times 6$ BlockM4B = $\begin{bmatrix} BN - 1 \times 1, 32 \\ MaxPool 2 \times 2, \\ stride = 2 \end{bmatrix}$	BlockM4b(*128*) = $\begin{bmatrix} 1 \times 1, \mathbf{128} \\ 3 \times 3, \mathbf{128} \\ 1 \times 1, \mathbf{128} \\ MaxPool 2 \times 2 \end{bmatrix} \times 2$
GlobalMaxPool - 64-d FC	BlockM4A ×12 - BlockM4B BlockM4A ×24 - BlockM4B	BlockM4b(256) ×3 BlockM4b(512)
FC, softmax	BlockM4A ×16 - AvgPool 7×7 1024-d FC - FC, softmax	$1 \times 1, 512$ - GlobalMaxPool 256-d FC - FC, softmax

Model 5a		**Model 6a**	
(Model 5b)	*(Model 5c)*	*(Model 6b, 6c, 6d)*	
$7 \times 7, 128, stride = 2$	$9 \times 9, 128, stride = 2$	$7 \times 7, 50, stride = 2$ $5 \times 5, 80, stride = 2$	$9 \times 9, 60, stride = 2$ $5 \times 5, 80, stride = 2$
BlockM5(**128**) = $\begin{bmatrix} 1 \times 1, \mathbf{128} \\ 3 \times 3, \mathbf{128} \\ 1 \times 1, \mathbf{128} \\ MaxPool 2 \times 2 \end{bmatrix} \times 2$	BlockM5(128) ×2 BlockM5(256) ×3	BlockM6(**100**) = $\begin{bmatrix} 3 \times 3, \mathbf{100} \\ 1 \times 1, \mathbf{100} \\ MaxPool 2 \times 2 \end{bmatrix}$ BlockM6(150)	BlockM6(100) BlockM6(150)
BlockM5(256) ×3 BlockM5(512) $1 \times 1, 512$	BlockM5(512) $1 \times 1, 512$	BlockM6(200) BlockM6(250) BlockM6(300) BlockM6(350) $1 \times 1, 400$ - GlobalMaxPool	BlockM6(200) BlockM6(250) BlockM6(300) BlockM6(350) $1 \times 1, 400$ - GlobalMaxPool
CONCAT - 256-d FC - FC, softmax		CONCAT - dropout 0.5128-d FC - FC, softmax	

Model 7a // Model 7b // Model 7c // Model 7d	**Model 8a // Model 8b**
$5 \times 5, 64, stride = 2$ // $7 \times 7, 64, stride = 2$ // $9 \times 9, 64, stride = 2$ // $11 \times 11, 64, stride = 2$	$7 \times 7, 64, stride = 2$ BlockM8(**64**) = $\begin{bmatrix} 1 \times 1, \mathbf{64} - BN \\ 3 \times 3, \mathbf{64} - BN \\ 1 \times 1, \mathbf{64} - BN \\ MaxPool 2 \times 2, stride = 2 \end{bmatrix}$
BlockM7(**64**) = $\begin{bmatrix} 1 \times 1, \mathbf{64} - BN \\ 3 \times 3, \mathbf{64} - BN \\ 1 \times 1, \mathbf{64}, -BN \\ MaxPool 2 \times 2 \end{bmatrix}$	BlockM8(128) ×2 Block(256) Block(512)
BlockM7(128) ×2 BlockM7(256) ($3 \times 3, 512$ - $3 \times 3, 1024$ // Only Model 7)	128-d FC - BN (dropout 0.4 // Only Model 8b)
GlobalMaxPool - FC, softmax	FC, softmax

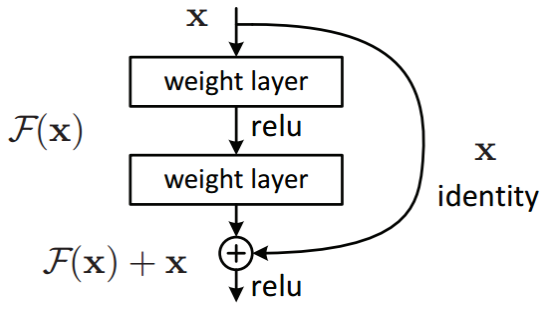
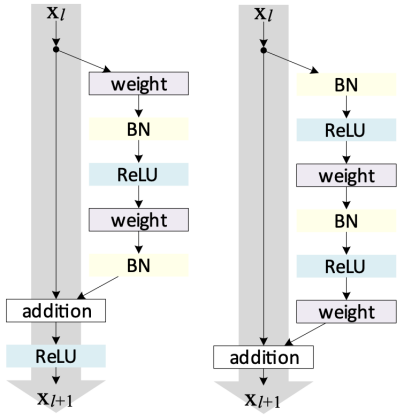


Figura 3.1: Representación gráfica de conexión residual. Imagen extraída de [2].



(a) Bloque residual original

(b) Bloque residual de la versión 2. Se propone una pre-activación completa.

Figura 3.2: Comparación de bloques residuales entre versiones de ResNet. Imágenes extraídas de [3].

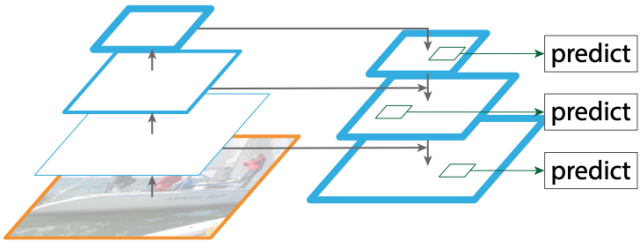


Figura 3.3: Esquema de una Feature Pyramid Network. Imagen extraída de [4].

el cambio de los bloques residuales originales a los de la segunda versión. Mientras que en la versión inicial la activación tenía lugar después de la suma de los resultados obtenidos por ambos caminos (el secuencial y el residual), en la segunda versión los autores demuestran empíricamente que una pre-activación completa antes de la suma mejora los resultados.

En 2017, los autores de [5] proponen un uso más eficiente de las conexiones residuales que se introdujeron con las ResNet y una reducción de la memoria usada por los modelos con el objetivo de que la inferencia pudiera ser utilizada en una amplia variedad de dispositivos. Con ello, nacen las Mobilenets y, posteriormente, su version mejorada, las MobilenetsV2 que redujeron el número de parámetros entrenables de los cerca de 20.000.000 de las ResNets a aproximadamente 800.000 parámetros entrenables con la consecuente reducción de la memoria necesaria para almacenar el modelo.

En la resolución de este problema, se complementa una MobilenetV2 con una *Feature Pyramid Network* (FPN). Las FPN consisten en analizar una misma imagen a diferentes escalas con el objetivo de que la detección de las características relevantes en una imagen sea invariante a la escala de la misma. En la Figura 3.3 se puede observar un esquema de FPN. Esta técnica lleva usándose mucho tiempo, pero hasta hace poco introducía demasiado coste computacional para poderse llevar a cabo. Los autores de [4] proponen utilizar el proceso de *downsampling* propio de las CNN para utilizar una especie de FPN en ellas sin introducir un coste computacional extra excesivo. En nuestro caso esto se

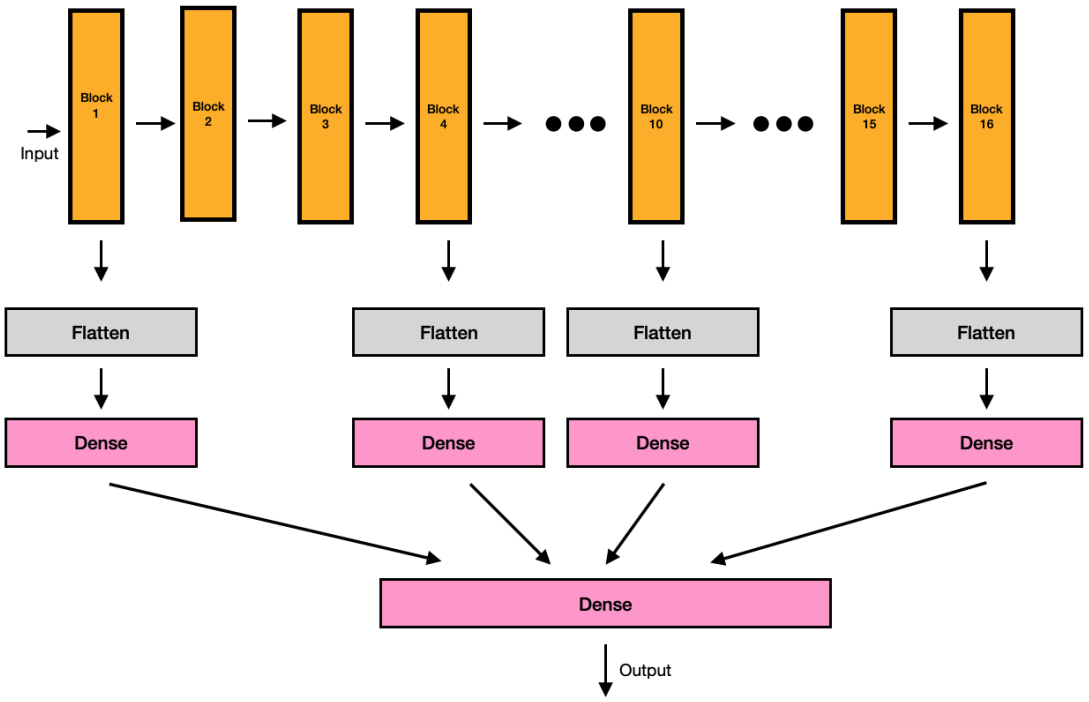


Figura 3.4: Diagrama simplificado del modelo MobilenetV2 con FPN.

obtiene conectando los mapas de salida de los bloques mobilenet 1, 4, 10 y 16 –cuya estructura interna se puede ver en la Figura 3.5– con una capa densa que se encargue de clasificar la imagen a ese tamaño. Con las clasificaciones de cada mapa se crea un nuevo tensor que es el que se introduce finalmente en una capa densa de dos neuronas y función de activación *softmax*. Con todo ello, en la Figura 3.4 se puede observar la topología del modelo descrito.

3.2.3. DarkNet

Este modelo se propone en [26] como el modelo de clasificación de YOLOv2 y que buscaba mejorar el modelo de clasificación de la YOLO original. Este modelo se construye con la experiencia de la primera versión de YOLO así como del conocimiento que se tenía de otros modelos que conformaban el estado del arte en la visón por computador en aquel momento. Así, la DarkNet-19 utiliza, como la VGG, filtros con tamaño $3x3$ y doblando el número de filtros tras cada convolución. El modelo utiliza GlobalMaxPooling para reducir la dimensión de las imágenes y convoluciones con filtros de $1x1$ después de las convoluciones con filtros de tamaño $3x3$ para comprimir mejor la representación de características de una imagen.

En nuestro caso utilizaremos una DarkNet-19 modificada para poder entrenar la cantidad de datos de que disponemos. La red cuenta con 19 capas convolucionales y cinco capas MaxPooling. La topología se muestra con detalle en la Tabla 3.3.

3.2.4. Tensorflow y Keras

Tensorflow es una librería *open source* utilizada en el campo del aprendizaje automático. Esta librería la desarrolló el equipo Google Brain de Google y se lanzó por primera

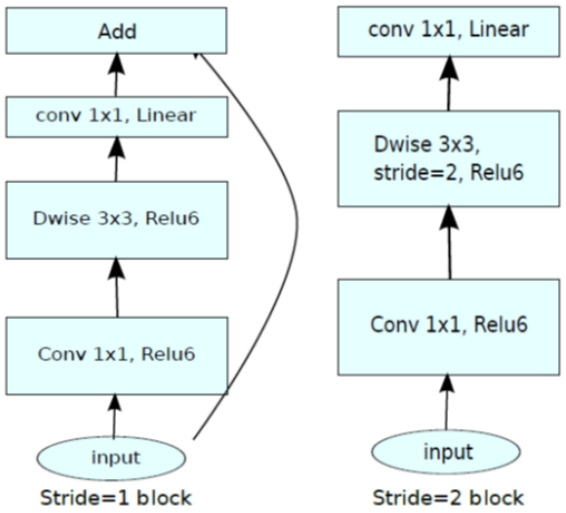


Figura 3.5: Estructura interna de un bloque convolucional del modelo MobilenetV2. Imagen extraída de [5].

Tabla 3.3: Topología de la DarkNet-19 adaptada a este problema.

Tipo de capa	Filtros	Tamaño filtro/Stride	Tamaño salida
Convolucional	8	3x3	256x256
MaxPool		2x2/2	128x128
Convolucional	16	3x3	128x128
Convolucional	16	1x1	128x128
Convolucional	16	3x3	128x128
MaxPool		2x2/2	64x64
Convolucional	32	3x3	64x64
Convolucional	32	1x1	64x64
Convolucional	32	3x3	64x64
MaxPool		2x2/2	32x32
Convolucional	64	3x3	32x32
Convolucional	64	1x1	32x32
Convloucional	64	3x3	32x32
MaxPool		2x2/2	16x16
Convolucional	128	3x3	16x16
Convolucional	128	1x1	16x16
Convolucional	128	3x3	16x16
MaxPool		2x2/2	8x8
Convolucional	256	3x3	8x8
Convolucional	256	1x1	8x8
Convolucional	256	3x3	8x8
Convolucional	128	1x1	8x8
Convolucional	256	3x3	8x8
Convolucional	2	3x3	8x8
Flatten			
Dense	100		
Softmax	2		

vez en 2015 como sucesora de Distbelief –biblioteca para el mismo uso, pero de código cerrado, que Google utilizaba internamente para crear sus propios modelos de redes neuronales–.

Tensorflow tiene disponibles APIs estables para C y Python, pero también para otros lenguajes C++, Swift o JavaScript. En este trabajo hemos escogido Python ya que esto nos permite utilizar Keras.

Keras es una librería escrita en Python, *open source* y que elimina algunas complejidades de Tensorflow consiguiendo que la experimentación sea más rápida. Keras fue diseñado para hacer posible una experimentación rápida con redes neuronales. Esto se consiguie gracias a su modularidad y extensibilidad.

3.2.5. EDDL

La *European Distributed Deep Learning Library* (EDDL) busca ser una librería flexible para el entrenamiento distribuido de modelos de deep learning con soporte para C++ y Python (PyEDDL). Actualmente se encuentra en la etapa final de su desarrollo, bajo la responsabilidad del PRHLT dentro del marco del proyecto europeo *DeepHealth*. Este proyecto está apoyado por el programa de investigación e innovación de la Unión Europea llamado Horizon 2020.

3.2.6. Hardware

Los procesos de entrenamiento y evaluación de todos los modelos presentados en este trabajo se han ejecutado usando un total de 15 máquinas equipadas con un procesador Intel Core i7-7800X CPU @ 3,50GHz (con 6 núcleos multi-hilo), 128 GB de RAM y 2 GPUs NVIDIA GeForce RTX 2080 con 8 GB de RAM cada una.

3.3 Preprocesado de los datos

El preprocesado de los datos es una fase importante que repercute en los resultados obtenidos en la fase de experimentación.

En el caso de este trabajo, los datasets CC-CCII y COVID-CTset ya vienen preparados para poder experimentar con ellos. Las imágenes vienen estandarizadas en términos de tamaño y los valores de las imágenes vienen en escalas listas para poder ser normalizadas con facilidad.

En el caso del dataset BIMCV PADCHEST COVID-19, si que ha sido necesario preprocesar las imágenes. Las imágenes de este conjunto de datos vienen en formato DICOM, un estandar utilizado ampliamente a la hora de trabajar con imágenes médicas. En este formato, los píxeles de la imagen contienen los valores de la intensidad del campo (radiactivo en el caso de los Rayos-X). Por tanto, el primer paso en el preprocesado de los datos es normalizar los valores de intensidad de las imágenes de todo el dataset. Para ello se siguen los siguientes pasos:

- Realizar un análisis exploratorio de los valores de intensidad de todas las imágenes del dataset.
- Detectar posibles outliers. Puede que haya valores excesivamente altos o excesivamente bajos fruto de un error de adquisición de la imagen. Estos valores se deben detectar y corregir. Para su corrección se establecen los valores muy altos al valor de

intensidad máximo dentro de un rango que se establezca como normal. El mismo proceso se repite para los valores excesivamente bajos.

- Una vez realizado el análisis y la corrección de errores de adquisición, se normalizan las imágenes con respecto al valor máximo de cada una de ellas.
- Las imágenes normalizadas se guardan en un formato de imagen de fácil interpretación, en nuestro caso el formato elegido es PNG.

Una vez realizado el preprocesado de las imágenes, deben normalizarse también los tamaños de las mismas. Después de explorar los diferentes tamaños de imagen de que se dispone se decide que el tamaño objetivo al que transformar todas las imágenes es de $200x200$px.

Se aplica *zero padding* a aquellas con un tamaño inferior en cualquiera de sus dimensiones. Así, la imagen original queda lo más centrada posible dentro de un marco con píxeles a cero, que rellena la imagen hasta conseguir el tamaño objetivo. Por otro lado, a las imágenes que son más grandes en cualquiera de sus tamaños se les aplica un recorte centrado en la región de interés del problema (ROI). En nuestro caso esta región está formada por el parénquima pulmonar. Un ejemplo de los resultados de este proceso se puede observar en la Figura 3.6.

3.4 Evaluación

La evaluación de estos sistemas debe hacerse teniendo en cuenta diferentes factores. En el caso de nuestro problema, debemos comprender que una clasificación incorrecta puede poner en peligro la salud e incluso la vida de una persona. Por ello, deberemos prestar atención a las métricas y corrección de las predicciones que establecemos para este problema.

3.4.1. Corrección de las predicciones

Recordemos que en este trabajo utilizaremos dos tipos de datos: imágenes Rayos-X e imágenes tomadas por resonancia magnética.

Realizar un clasificador para el primer tipo de imágenes no tiene complicaciones en cuestión de corrección de las predicciones. Esto se debe a que cada prueba médica consta de una única imágen y, por tanto, la predicción únicamente depende del resultado de clasificar una muestra.

Sin embargo, cuando trabajamos con imágenes obtenidas por resonancia magnética, la dificultad de clasificación de un paciente cambia. Una sesión de resonancia magnética se compone de la toma de varias imágenes a lo largo de un eje longitudinal. Así, una sola sesión de uno de los sujetos puede contener entre 50 y 300 imágenes. Además, debemos tener en cuenta que no todas las imágenes de una sesión serán igualmente válidas. Los primeros y los últimos cortes de una sesión no suelen contener información puesto que no son zonas dónde se concentren los biomarcadores de la pulmonía por COVID-19 y, además, los pulmones aparecen muy cerrados. Se puede observar una comparativa de un corte medio y superior de una misma sesión en la Figura 3.7.

Así, con este tipo de datos, trabajaremos en dos fases. Entrenaremos clasificadores a nivel de imagen que sean capaces de distinguir entre imágenes (cortes) que pertenecen a pacientes con neumonía por COVID-19 y aquellas que pertenecen a sujetos control. En una segunda fase utilizaremos esos clasificadores para que analicen cada uno de los

(a) Imagen original, más pequeña que el tamaño objetivo

(b) Imagen resultante tras aplicar *zero-padding* a la imagen en (a)

(c) Imagen original, más grande que el tamaño objetivo

(d) Imagen resultante tras aplicar el recorte centrado en la ROI a la imagen en (c)

Figura 3.6: Resultados de la normalización de tamaño de las imágenes del dataset.

(a) Primeros/Últimos cortes de una sesión de resonancia magnética. Se observa el pulmón cerrado con lo que no se pueden detectar biomarcadores de neumonía por COVID-19

(b) Corte medio de una sesión de resonancia magnética. Se observa el pulmón abierto con lo que es posible detectar biomarcadores de neumonía por COVID-19 si los hubiese.

Figura 3.7: Comparación entre cortes extremos y medios en una sesión de resonancia magnética de pulmón.

cortes de un paciente y determinaremos si el diagnóstico del mismo debe ser positivo o negativo en neumonía por COVID-19. Para ello estableceremos un umbral que suelen utilizar los propios radiólogos. Si al menos un 10 % de los cortes de una sesión se clasifican como positivos en neumonía por COVID-19, entonces el paciente será diagnosticado como positivo.

3.4.2. Métricas

Las métricas que se ha decidido utilizar para evaluar esta tarea y sus razones son:

- Tasa de acierto. Utilizada para evaluar el comportamiento general de los clasificadores.
- Precisión (sobre la clase COVID-19). Utilizada para ver el porcentaje de detección de casos de COVID-19 sobre el total de casos clasificados.
- Recall (sobre la clase COVID-19). Utilizado para medir el número de muestras de esta clase que se han clasificado correctamente.
- AUROC. Utilizada para comprobar la capacidad discriminativa del clasificador.

CAPÍTULO 4
Experimentación y resultados

En este capítulo expondremos los detalles de la experimentación que hemos llevado a cabo con cada uno de los datasets, los resultados que se han obtenido y se analizarán los mismos.

4.1 Experimentación con el dataset BIMCV PADCHEST COVID-19

El proceso de entrenamiento de cada uno de los modelos se llevó a cabo aplicando técnicas de data augmentation. Las técnicas de *data augmentation* nos permiten aumentar el número de muestras disponibles debido a las transformaciones aleatorias que se introducen sobre las muestras originales. De esta manera conseguimos: (a) introducir variabilidad en el entrenamiento del modelo, lo cual suele conllevar la obtención de mejores resultados y (b) obtener, virtualmente, un número infinito de muestras de entrenamiento y, generalmente, a más muestras, mejores resultados.

Las técnicas de *data augmentation* usadas sobre este dataset son:

- Rotación aleatoria de α grados, con $\alpha \in [-10, 10]$
- Zoom aleatorio en un rango de $[0{,}9, 1{,}1]$

En cuanto al optimizador utilizado durante el entrenamiento, se ha optado por utilizar el algoritmo Stochastic Gradient Descent (SGD) con un Learning Rate de 10^{-3} y un scheduler para ir reduciendo el mismo por un factor constante en las epochs 100, 200, 300 y 400. El número máximo de epochs durante el proceso de entrenamiento se estableció en 500 y el mencionado factor se estableció en 0,7. En algún caso se utilizó un factor de 0,5 pero no se observaron mejoras relevantes.

Se realizaron también entrenamientos con el optimizador Adam con el mismo learning rate, factor de reducción y scheduler, pero tampoco se observaron mejoras significativas.

Durante el entrenamiento se observó overfitting en la gran mayoría de casos a pesar del uso de técnicas de data augmentation. Para la gran mayoría de modelos propuestos la tasa de acierto en entrenamiento alcanzó el 99 % antes de la epoch 200, mientras que en validación la tasa de acierto quedó estancada en el intervalo $[80, 85]$ dependiendo del modelo.

Tabla 4.1: Valores de las métricas obtenidas con el subconjunto de test del dataset PADCHEST COVID-19; precisión, recall y F1-score se muestran para la clase COVID-19 (COV).

Model	Tasa de acierto	Prec.	Recall	F1-Score	AUC	FN ratio
3-	76,4 %	0,98	0,57	0,72	0,925	0,432
4a	**84,7** %	0,97	0,73	0,83	**0,924**	0,271
4b	74,1 %	**0,99**	0,53	0,69	0,942	0,466
4c	79,5 %	0,97	0,63	0,76	0,926	0,374
4d	82,4 %	0,94	0,71	0,81	0,918	0,293
5a	80,1 %	0,88	0,74	0,80	0,904	0,260
5b	77,6 %	0,73	**0,92**	0,81	0,891	**0,084**
5c	83,5 %	0,94	0,74	0,83	0,928	0,255
6a	80,4 %	0,96	0,66	0,79	0,929	0,337
6b	75,0 %	0,72	0,88	0,79	0,866	0,119
6c	84,1 %	0,91	0,78	**0,84**	0,903	0,219
7b	82,4 %	0,96	0,70	0,81	0,911	0,304
7c	83,8 %	0,95	0,74	0,83	0,918	0,262
8a	82,7 %	0,85	0,81	0,83	0,901	0,188
8b	81,0 %	0,97	0,68	0,80	0,914	0,321

4.1.1. Resultados

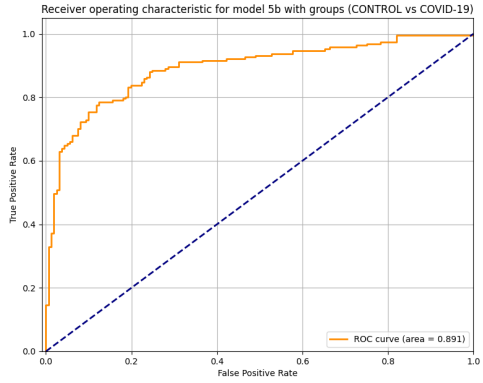
En la Tabla 4.1 se muestran los mejores resultados de entrenamiento para cada modelo propuesto y sus variaciones. Se recogen los resultados de Tasa de acierto, Precisión, Recall, F1-score, AUC y Ratio de falsos negativos

Podemos observar en la Tabla 2.1 que la partición de entrenamiento del dataset está desbalanceada. La clase control representa un 90 % de las muestras del subconjunto mientras que la clase COVID-19 representa tan solo un 10 %. Para mejorar los modelos utilizados, se utilizó una estrategia de igualación a nivel de batch, para garantizar que cada batch tuviera exactamente el mismo número de muestras de cada clase. Esta estrategia es similar a la que los autores de [9] realizan en su experimento.

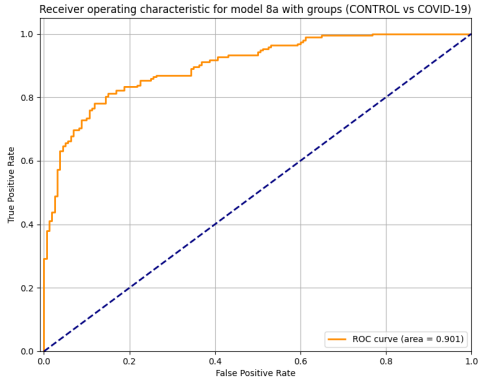
Analizando los resultados de la Tabla 4.1 vemos que hay diferencias relevantes entre modelos en terminos de AUC. En concreto, solo un pequeño número de modelos obtienen un valor de AUC por debajo de 0,90. Por otro lado, en la resolución de este problema buscamos un balance entre los valores de tasa de acierto precisión y recall para la clase COVID-19. En busca de un balance entre estas dos últimas métricas y tal y como se ha explicado anteriormente, buscaremos modelos con altos vares de tasa de acierto y F1-score. Además, también valoraremos positivamente que el clasificador tenga un bajo FN Ratio.

Con todo esto podemos destacar que el modelo 4a es un buen modelo al tener la mayor tasa de acierto de todos y un AUC que consideramos bastante alta. También merece la pena recalcar el modelo 5b con un buen Recall y el FN Ratio más bajo de todos los modelos que se presentan. Sin embargo uno de los modelos más equilibrados que se han obtenido es el modelo 8a con una tasa de acierto a tan solo dos puntos porcentuales de la mayor, un buen F1-score, y un alto AUC.

A pesar de todas las métricas descritas, cuando trabajamos con clasificadores que aspiran a ser incluidos en el proceso de diagnóstico del personal médico debemos profundizar un poco más. La tasa de acierto y el AUROC son buenas métricas, pero para medir la aplicabilidad real de un modelo en casos de usos reales debemos medir el False Negative Ratio (FNR) y buscar que sea lo más bajo posible.

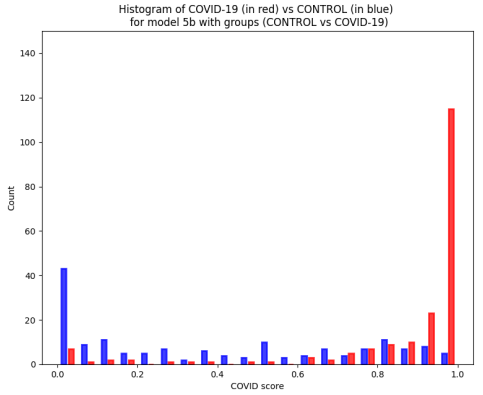


(a) Curva ROC para el modelo 5b (C vs. COV)

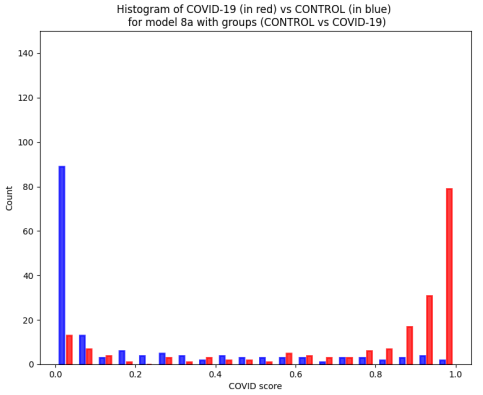


(b) Curva ROC para el modelo 8a (C vs. COV)

Figura 4.1: Curvas ROC para los modelos 5b y 8a en la tarea definida.



(a) (C vs. COV) histograma para el modelo 5b



(b) (C vs. COV) histograma para el modelo 8a

Figura 4.2: (C vs. COV) Distribución de muestras con respecto a la predicción de la clase COVID-19 realizada por los modelos 5b y 8a. En ambas figuras sujetos control se muestran en azul y pacientes COVID-19, en rojo.

El modelo 5b y el 8a se analizan más en profundidad a continuación. El primero por el bajo FN Ratio y el segundo por el equilibrio que se encuentra entre todas sus métricas. En la Figura 4.1 se encuentran las curvas ROC de ambos

El objetivo último con los clasificadores es definir la que la teoría de la decisión denomina como área de rechazo. Para ello se definen dos umbrales th_{low} y th_{high} tal que no existan falsos negativos para $COVID\_score < th_{low}$ y no hayan falsos positivos para $COVID\_score > th_{high}$. En la Figura 4.2 se puede observar lo difícil que es ajustar dichos umbrales para minimizar o incluso eliminar falsos positivos y falsos negativos. Pero observando la Figura 4.2(a) la distribución de las muestras de las clases Control (0) y COVID-19 (1) se comprueba que no es posible definir semejante área. Aún así, si comparamos ambas figuras (Figura 4.2(a) y (b)), podemos decir que la distribución es mucho mejor en (b) mientras que la precisión de detección de la clase COVID-19 (barra roja en 1,0) es mucho mejor en (a). Este inconveniente es un problema complejo de resolver por la idiosincrasia del dataset con el que se ha trabajado en este experimento.

4.2 Experimento con los datasets CC-CCII y COVID-CTset

El proceso de entrenamiento de cada uno de los modelos se llevó a cabo aplicando técnicas de data augmentation. Las técnicas de *data augmentation* nos permiten, como se ha indicado antes, aumentar el número de muestras disponibles debido a las transformaciones aleatorias que se introducen sobre las muestras originales.

Las técnicas de *data augmentation* usadas sobre este dataset son:

- Rotación aleatoria de α grados, con $\alpha \in [-20, 20]$
- Zoom aleatorio en un rango de $[0{,}95, 1{,}05]$
- Flip horizontal y vertical aleatorio
- Shift vertical y horizontal aleatorio del 5 %

En cuanto al optimizador utilizado durante el entrenamiento, se ha optado por utilizar el algoritmo Stochastic Gradient Descent (SGD) con un Learning Rate de 10^{-3} y un scheduler para ir reduciendo el mismo por un factor constante en las epochs 100, 200, 300 y 400. El número máximo de epochs durante el proceso de entrenamiento se estableció en 500 y el mencionado factor se estableció en 0,7. En algún caso se utilizó un factor de 0,5 pero no se observaron mejoras relevantes.

Se realizaron también entrenamientos con el optimizador Adam con el mismo learning rate, factor de reducción y scheduler, pero tampoco se observaron mejoras significativas.

Además, se ha utilizado la técnica de *transfer learning* para un entrenamiento más rápido de los modelos y una mejor convergencia de los modelos. Para ello, en la parte convolucional de ambos modelos utilizados (Mobilenet y Darknet) se han utilizado los pesos entrenados con el conjunto de datos *Imagenet*. La parte densa de los modelos se crea ad-hoc para este problema y el entrenamiento de los modelos tiene lugar en dos etapas. Una primera en la que se aprenden (modifican) únicamente los pesos de la parte densa y una segunda donde, con un learning rate menor, se modifican todos los pesos (parte convolucional y densa). De esta manera se consigue hacer un buen uso de los pesos importados ya que aportan capacidades a los modelos a la hora de detectar bordes, formas simples y otras estructuras básicas en el reconocimiento de entidades en imágenes.

Durante el entrenamiento con el conjunto de datos CC-CCII no se observó *overfitting*. Las técnicas de *data augmentation* seleccionadas y la gran cantidad de datos disponible permitieron una buena generalización del modelo que se traduce, como se verá más tarde, en buenos resultados. En el caso del *dataset* COVID-CTset, se observó *overfitting* en todos los casos a pesar del uso de técnicas de *data augmentation* y del uso de técnicas de regularización. Se discutirá más adelante las causas de los malos resultados de los modelos entrenados con este conjunto de datos.

4.2.1. Resultados

En la Tabla 4.2 se muestran los mejores resultados de entrenamiento para cada modelo propuesto utilizando el *dataset* CC-CCII. En la Tabla 4.3 se muestran los resultados de entrenamiento para cada modelo propuesto utilizando el *dataset* COVID-CTset. En ambas tablas se recogen los resultados de Tasa de acierto a nivel de corte 2D, y la Precisión, Recall, F1-score, Ratio de falsos negativos y Tasa de acierto, a nivel de paciente (Acc. Paciente).

Tabla 4.2: Valores de las métricas obtenidas con el subconjunto de test del dataset CC-CCII; precisión, recall y F1-score se muestran para la clase COVID-19 (COV). Precisión, recall, F1-score, FNR y Acc. Paciente son métricas calculadas a nivel de paciente.

Model	Tasa de acierto	Prec.	Recall	F1-Score	Acc. Paciente	FN ratio
Mobilenet	**94,98 %**	**0,96**	**1,0**	**0,98**	**98 % ± 1**	**0,0**
Darknet	93,95 %	0,89	1,0	0,94	93 % ± 3	0,0

Tabla 4.3: Valores de las métricas obtenidas con el subconjunto de test del dataset COVID-CTset; precisión, recall y F1-score se muestran para la clase COVID-19 (COV). Precisión, recall, F1-score, FNR y Acc. Paciente son métricas calculadas a nivel de paciente.

Model	Tasa de acierto	Prec.	Recall	F1-Score	Acc. Paciente	FN ratio
Mobilenet	90,80 %	0,33	1,0	0,5	33 % ± 13,7	0,0
Darknet	**90,90 %**	**0,0**	**0,0**	**0,0**	**67 % ± 13,7**	**0,33**

Podemos observar en la Tabla 2.2 que la partición de entrenamiento del dataset está desbalanceada. La clase control representa un 40 % aproximadamente de las muestras del subconjunto mientras que la clase COVID-19 representa aproximadamente un 60 %. En nuestro caso, tener más muestras de nuestra clase objetivo permite obtener mejores resultados de clasificación. En el caso de este dataset, no se utilizan técnicas de rebalanceado de los datos, ni a nivel de conjunto de datos, ni a nivel de batch. En el caso del conjunto de datos COVID-CTset no sucede lo mismo. Las muestras de la clase Control representan un 80 % mientras que las de la clase COVID-19 solamente son un 20 % del total. Para mejorar los resultados con este conjunto de datos, se utilizó una estrategia de igualación a nivel de batch, para garantizar que cada batch tuviera exactamente el mismo número de muestras de cada clase. Esta estrategia es similar a la que los autores de [9] realizan en su experimento.

Analizando los resultados de la Tabla 4.2 vemos que en ambos casos, los resultados son muy buenos. Se han obtenido con los mejores modelos obtenidos durante el entrenamiento. En el caso de la Mobilenet los resultados mostrados corresponden al modelo guardado en la *epoch* 4 de 300. En el caso de la Darknet los resultados mostrados corresponden al modelo guardado en *epoch* 5 de 300.

En primer lugar las tasas de acierto a nivel de cortes 2D, son prometedoras y están comprendidas entre el 93 y el 95 porciento. En cualquier caso, es necesario que ese nivel de acierto a nivel de imágenes se traduzca en una buena potencia discriminativa a nivel de paciente. El F1-score nos indica que la clasificación a nivel de paciente es buena. Especial mención a la tasa de falsos negativos que cae en ambos casos a cero, lo cual nos indica que nuestro modelo no comete errores a la hora de clasificar pacientes enfermos. La accuracy a nivel de paciente se obtiene clasificando todas las imágenes 2D que componen una sesión de resonancia mágnética del mismo. Para clasificar un paciente como enfermo, al menos un 10 % de las imágenes de una sesión deben ser clasificadas como COVID-19. Este umbral de decisión se ha determinado empíricamente según lo establecido en [1].

Con todo esto podemos destacar que ambos modelos son buenos, pero, para este problema, podemos determinar que Mobilenet sería mejor que Darknet, por obtener mejores resultados en todas las métricas.

Analizando los resultados de la Tabla 4.3 vemos que en ambos casos, los resultados son malos. Se han obtenido con los mejores modelos obtenidos durante el entrenamiento.

En el caso de la Mobilenet los resultados mostrados corresponden al modelo guardado en la *epoch* 6 de 300. En el caso de la Darknet los resultados mostrados corresponden al modelo guardado en *epoch* 20 de 300.

En primer lugar analizaremos las tasas de acierto a nivel de cortes 2D. En el caso de la Mobilenet la tasa de acierto a nivel de corte 2D se sitúa en torno al 90 %. El F1-score es muy bajo y esto se traduce en una capacidad discriminativa a nivel de paciente que tiene una tasa de acierto que se encontraría en el rango $[20, 46]$ porciento. El elevado intervalo de confianza nos da pistas de por qué los resultados podrían ser tan bajos. En el caso de la Darknet, pese a que la tasa de acierto a nivel de corte 2D puede parecer prometedora, el resto de métricas nos indica que no lo son. Lo que ha sucedido con este modelo es que solamente es capaz clasificar correctamente las muestras de la clase control. A nivel de paciente la tasa de acierto nos indica lo mismo, un 67 % que se corresponde al porcentaje de pacientes en el conjunto de test que pertenecen a la clase control.

CAPÍTULO 5
Conclusiones y trabajo futuro

De este trabajo se pueden extraer varias conclusiones con respecto a los modelos y los datos utilizados.

En cuanto a los modelos, en este trabajo se han creado topologías ad-hoc exclusivamente diseñadas para este problema teniendo en cuenta las topologías de los modelos que conforman el estado del arte en la materia. Los resultados son aceptables, pero estarían algo lejos de poder ser modelos aplicables en un proceso de diagnóstico real. Uno de los motivos de los resultados a la baja de estos modelos presentados es la calidad de los datos utilizados en el entrenamiento. Podemos concluir que, al igual que es difícil para el personal médico detectar neumonía por COVID-19 en imágenes Rayos-X, también lo es para los modelos de visión por computador. Como trabajos futuros en esta línea se propone el refinamiento de los mejores modelos obtenidos y un análisis más profundo de los datos en conjunto con personal médico cualificado que permita comprender mejor las zonas de las imágenes en que se fijan a la hora de elaborar un diagnóstico.

Por otro lado, se ha presentado como válida la opción de utilizar transfer learning en este problema. Los resultados han igualado a los obtenidos en el estado del arte en la materia en el caso del *dataset* CC-CCII. Ante igualdad de modelos, no ha sido posible obtener buenos resultados con el conjunto de datos COVID-CTset. Podemos concluir que tener datos de calidad y en abundancia es un requisito indispensable para que los modelos sean capaces de aprender. Estas dos premisas se cumplen en el primer conjunto de datos, sin embargo en el segundo hay una falta de datos que hace que modelos como Mobilenet (+700.000 parámetros entrenables) o Darknet (+2.000.000 parámetros entrenables) no sean capaces de extraer características de calidad para resolver el problema de clasificación propuesto.

Por último, se puede concluir que, como se podía intuir por el actual proceso de diagnóstico, el problema se vuelve mucho más fácil de solucionar cuando se trabaja con imagen por resonancia magnética que cuando se utiliza imagen por Rayos-X. Como trabajo futuro, se propone seguir analizando el problema de la detección de biomarcadores de la neumonía por COVID-19 en imágenes Rayos-X.

Bibliografía

[1] M. Rahimzadeh, A. Attar, and S. M. Sakhaei, "A fully automated deep learning-based network for detecting covid-19 from a new and large lung ct scan dataset," *medRxiv*, 2020.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*, pp. 630–645, Springer, 2016.

[4] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

[5] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[6] World Health Organization, "Coronavirus disease (covid-19): situation report, 189," 2020.

[7] N. S. Punn, S. K. Sonbhadra, and S. Agarwal, "Covid-19 epidemic analysis using machine learning and deep learning algorithms," *medRxiv*, 2020.

[8] Y.-R. Guo, Q.-D. Cao, Z.-S. Hong, Y.-Y. Tan, S.-D. Chen, H.-J. Jin, K.-S. Tan, D.-Y. Wang, and Y. Yan, "The origin, transmission and clinical therapies on coronavirus disease 2019 (covid-19) outbreak–an update on the status," *Military Medical Research*, vol. 7, no. 1, pp. 1–10, 2020.

[9] L. Wang, Z. Q. Lin, and A. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020.

[10] A. Fourcade and R. Khonsari, "Deep learning in medical image analysis: A third eye for doctors," *Journal of stomatology, oral and maxillofacial surgery*, vol. 120, no. 4, pp. 279–288, 2019.

[11] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, "Covid-19 image data collection: Prospective predictions are the future," *arXiv preprint arXiv:2006.11988*, 2020.

[12] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 590–597, 2019.

[13] A. Stein, "Pneumonia dataset annotation methods. rsna pneumonia detection challenge discussion," 2018.

[14] M. d. l. I. Vayá, J. M. Saborit, J. A. Montell, A. Pertusa, A. Bustos, M. Cazorla, J. Galant, X. Barber, D. Orozco-Beltrán, F. Garcia, *et al.*, "Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients," *arXiv preprint arXiv:2006.01174*, 2020.

[15] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[16] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[17] O. Stephen, M. Sain, U. J. Maduh, and D.-U. Jeong, "An efficient deep learning approach to pneumonia classification in healthcare," *Journal of healthcare engineering*, vol. 2019, 2019.

[18] P. K. Sethy and S. K. Behera, "Detection of coronavirus disease (covid-19) based on deep features," *Preprints*, vol. 2020030300, p. 2020, 2020.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[20] A. Abbas, M. M. Abdelsamea, and M. M. Gaber, "Classification of covid-19 in chest x-ray images using detrac deep convolutional neural network," *arXiv preprint arXiv:2003.13815*, 2020.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[22] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá, "Padchest: A large chest x-ray image dataset with multi-label annotated reports," *arXiv preprint arXiv:1901.07441*, 2019.

[23] M. de la Iglesia Vayá, J. M. Saborit, J. A. Montell, A. Pertusa, A. Bustos, M. Cazorla, J. Galant, X. Barber, D. Orozco-Beltrán, F. García-García, *et al.*, "Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients.," *arXiv preprint arXiv:2006.01174*, 2020.

[24] K. Zhang, X. Liu, J. Shen, Z. Li, Y. Sang, X. Wu, Y. Zha, W. Liang, C. Wang, K. Wang, *et al.*, "Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography," *Cell*, 2020.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[26] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, 2017.