UNIVERSIDAD DE OVIEDO

Programa de doctorado en Ciencias de la Salud

EL PROBLEMA DE LA GENERALIZACIÓN DE LOS ALGORITMOS DE APRENDIZAJE PROFUNDO PARA EL ANÁLISIS DE IMAGEN MÉDICA: investigación utilizando imágenes radiográficas

GENERALIZATION OF DEEP LEARNING ALGORITHMS FOR MEDICAL IMAGING INTERPRETATION: an exploratory analysis using radiographic images

Pablo Menéndez Fernández-Miranda

Oviedo 2022

UNIVERSIDAD DE OVIEDO

Programa de doctorado en Ciencias de la Salud

EL PROBLEMA DE LA GENERALIZACIÓN DE LOS ALGORITMOS DE APRENDIZAJE PROFUNDO PARA EL ANÁLISIS DE IMAGEN MÉDICA: investigación utilizando imágenes radiográficas

GENERALIZATION OF DEEP LEARNING ALGORITHMS FOR MEDICAL IMAGING INTERPRETATION: an exploratory analysis using radiographic images

Pablo Menéndez Fernández-Miranda

Directores

José A. Vega Álvarez
Lara Lloret Iglesias

Oviedo 2022

Universidad de Oviedo

RESUMEN DEL CONTENIDO DE TESIS DOCTORAL

1.- Título de la Tesis	
Español/Otro Idioma: EL PROBLEMA DE LA GENERALIZACIÓN DE LOS ALGORITMOS DE APRENDIZAJE PROFUNDO PARA EL ANÁLISIS DE IMAGEN MÉDICA: investigación utilizando imágenes radiográficas.	Inglés: GENERALIZATION OF DEEP LEARNING ALGORITHMS FOR MEDICAL IMAGING INTERPRETATION: an exploratory analysis using radiographic images.

2.- Autor	
Nombre: MENÉNDEZ FERNÁNDEZ-MIRANDA	DNI/Pasaporte/NIE:
Programa de Doctorado: CIENCIAS DE LA SALUD	
Órgano responsable: CENTRO INTERNACIONAL DE POSTGRADO	

RESUMEN (en español)

En la literatura se ha reportado que los algoritmos de Deep Learning (DL), traducido como aprendizaje profundo, sufren, con frecuencia, problemas de generalización cuando son implementados en entornos diferentes de aquellos donde han sido entrenados. A pesar de que esta cuestión ha sido ampliamente discutida en otros campos, son pocos los autores que han abordado la problemática en el área de la imagen médica. Teniendo en cuenta que generalizar hace referencia a la capacidad de mantener el rendimiento en cualquier entorno externo al lugar de entrenamiento, resolver el problema de la generalización es imprescindible si se desea lograr que los modelos de inteligencia artificial alcancen la práctica médica diaria.

Las escasas investigaciones que han abordado esta problemática coinciden en que, al igual que ocurre en otras áreas de la inteligencia artificial, como el desarrollo de coches autónomos, la deficiencia de generalización que sufren los algoritmos de DL se produce debido a desemejanzas existentes entre los datos con los que se entrenan los algoritmos, y los datos sobre los que se desean utilizar los modelos. Estas disimilitudes parecen responder a diferencias, a menudo imperceptibles, en la distribución de ciertos atributos de los propios datos, dando lugar a lo que se ha denominado *distribution shifts* (DS).

La importancia de desvelar los factores que ocasionan los DS entre los conjuntos de datos de entrenamiento y los conjuntos de datos de aplicación, es un asunto cuya relevancia ha sido ampliamente subrayada en la literatura. Sin embargo, hasta la fecha, la naturaleza de estos factores permanecía incierta.

Este trabajo ha tratado de arrojar luz a esta problemática, proponiéndose identificar y caracterizar los principales factores responsables de la citada deficiencia de generalización de los algoritmos de DL en el campo de la imagen médica. Con este objetivo, se diseñaron tres experimentos cuya meta fue la de evaluar el efecto que tenían sobre la validez interna y externa de un algoritmo de DL los siguientes factores: el centro del que se obtuvieron los datos de entrenamiento, lo que comprendió el estudio conjunto de otros agentes como las diferencias en el criterio de etiquetado o en la demografía poblacional; el protocolo de adquisición de las imágenes; el modelo de equipo de adquisición de rayos X; el tipo de función de respuesta del detector; y el procesado de imagen aplicado por el fabricante.

En la experimentación, se compararon los resultados obtenidos de tres entrenamientos diferentes de una red neuronal artificial para clasificación de radiografías de tórax de pacientes COVID-19 y pacientes control. Estos tres entrenamientos se realizaron con los mismos hiperparámetros y se diferenciaron únicamente en los conjuntos de datos utilizados, que incluyeron combinaciones de imágenes adquiridas en dos instituciones y por tres modelos diferentes de equipos de rayos X. A través del análisis de las diferencias de rendimiento obtenidas sobre diferentes conjuntos de test que alternaron dos instituciones y tres modelos de equipos de adquisición, se pudo inferir el efecto aislado de los factores presentes en los mismos. Finalmente, se realizó un análisis de las *features* que extraía la red neuronal utilizando un algoritmo de *clustering* jerarquizado, con la finalidad de esclarecer si la influencia de estos factores afectaba también al valor de las *features*.

Los resultados mostraron que el tipo de función de respuesta del detector fue el factor más relevante en términos de generalización, pues fue el único que le impidió generalizar al algoritmo. En segundo lugar, el procesado de imagen aplicado por el equipo fue el segundo factor en importancia, ya que aunque no impidió la generalización, condicionó una reducción significativa del rendimiento en validez externa. Sorprendentemente, estos dos factores resultaron ser más influyentes sobre el valor de las *features* que la propia presencia o ausencia de patología. En contraposición, los factores asociados con el centro de entrenamiento indujeron diferencias salvables, pues el algoritmo logró generalizar sin pérdida de rendimiento a un centro externo que utilizaba el mismo modelo de equipo de rayos X que aquel que había sido utilizado para adquirir las imágenes de entrenamiento.

RESUMEN (en Inglés)

As it has been extensively reported in the literature, Deep Learning (DL) algorithms often show a lack of generalization when they are deployed in environments different from those where they have been trained. Although this issue has been widely discussed in other fields, in the medical image domain only a handful of authors have addressed it. Considering that generalization refers to the ability of the algorithms to maintain performance in environments different from those used to train them, solving the generalization deficiency is essential to achieve an artificial intelligence which can reach the medical practice.

The scarce previous works that have already addressed this issue in the medical field agree with other artificial intelligence areas, such as the development of autonomous cars, that generalization deficiency of DL algorithms is produced due to the existence of differences between the training data and the data on which the models are intended to be applied. These dissimilarities are probably secondary to differences in the distribution of certain attributes of the data itself which are often imperceptible, giving rise to what have been called distribution shifts (DS).

The importance of identifying and controlling the factors that cause the DS has been widely emphasized in the literature. However, previously to this research these factors remained uncertain. This work has attempted to shed light on this issue trying to identify and characterize the main factors behind the aforementioned deficiency of generalization of DL algorithms in the medical image field. To achieve this aim, three experiments were designed to evaluate the effect of the following factors on the internal and external validation of a DL algorithm: the institution where the training data were obtained, which included the analysis of other agents such as differences in labeling criteria or population demographics; the image acquisition protocol; the model of X-ray machine used to acquire the images; the type of response function of the detector; and the image processing applied by the manufacturer.

The experimentation consisted in the comparison of the results obtained by three different trainings of an artificial neural network to classify chest X-rays belonging to COVID-19 and control patients. These three trainings were performed with the same hyperparameters so they differed only in the datasets used. These datasets included combinations of images acquired at two different institutions and by three different models of X-ray machine. Through the analysis of the performance differences obtained on different test sets that alternated two institutions and three models of acquisition equipment, it was possible to study the isolated effect of the aforementioned factors on the algorithm's performance. Finally, an analysis of the features extracted by the neural network was performed using a hierarchical clustering algorithm in order to explore whether the influence of these factors also affected the value of the features.

The results showed that the type of response function of the detector was the most relevant factor in terms of generalization, as it was the only factor that impeded the algorithm to generalize. Secondly, the image processing applied by the equipment was the second most important factor, since it did not impede generalization but it produced a significant performance decrease. Moreover, these two factors proved to have more influence on the values of the features than the presence or absence of pathology itself. By contrast, the factors related to the training institution produced surmountable differences, since the algorithm was able to generalize without performance loss to an external institution which used the same model of X-ray machine as the one used to acquire the training images.

SRA. PRESIDENTA DE LA COMISIÓN ACADÉMICA DEL PROGRAMA DE DOCTORADO EN CIENCIAS DE LA SALUD.

DEDICATORIA

Esta tesis está dedicada a mi familia, el lugar donde se encuentra mi origen y donde, de acuerdo con mis aspiraciones, se encontrará mi final. El lugar que me transmitió los valores que soportan mi identidad y que guiarán e impregnarán este y todos los trabajos que afronte. El lugar que me regalo el cariño, el ejemplo y el conocimiento que me han convertido en una persona.

De esta forma, y respetando la jerarquía que otorga la edad, dedico esta tesis especialmente a:

a *mi abuelo Pepe*, por haberme dejado la herencia de un ejemplo cuyo recuerdo logra resumir todo en lo que anhelo convertirme algún día;

a *mi abuela Carmen*, por haberme enseñado que la verdadera grandeza se esconde detrás de la humildad;

a *mis abuelos Pilar y Aurelio*, por haberme mostrado el poder del cariño;

a *mi tía-madre Carmen*, por haberme demostrado con el ejemplo de su vida que la duración de la adversidad es inversamente proporcional a la lucha propia;

a *mis padres José y Candela*, por haberse esforzado todo cuanto han podido en construir mi personalidad, por haber sabido levantarme las veces que me he caído, y por haberme dado siempre, todo;

a *mi prima-hermana Carmen*, por enseñarme cada día a través de su ejemplo a afrontar la dificultad demostrandome que la felicidad está en uno mismo;

a *mi hermana Candela*, por haberme transmitido el valor de la familia anteponiéndola siempre a sí misma;

a *mi novia Marta*, por ser la luz verde y mágica que puede ser oída, la luz que ilumina mis noches frías;

a *mis sobrinos Pepe, Pedro y José Luis*, por recordarme todos los días la importancia de ser siempre un niño.

AGRADECIMIENTOS

Expreso mi más sincero agradecimiento a:

a el director y tutor de esta tesis doctoral, el *profesor José Antonio Vega Álvarez*, por haberme abierto las puertas de la Universidad antes de conocerme, por haberme inculcado la afición por la curiosidad, y haberme trasmitido ilusión en cada conversación que hemos mantenido estos cuatro años;

a la co-directora de esta tesis doctoral, la *doctora Lara Lloret Iglesias*, por haberme acercado el conocimiento de una disciplina apasionante de la que, antes de realizar esta tesis, me encontraba muy alejado, por haberme prestado ayuda inmediata a pesar de su apretada agenda siempre que lo necesite, y por haberme regalado un ejemplo de verdadera investigadora;

a *Enrique Marqués Fraguela*, por el apoyo, el soporte, y sus grandes contribuciones a esta tesis doctoral con discusiones de las que surgieron las ideas principales que han guiado la investigación;

A *David Rodríguez González*, por haberme prestado tutela y supervisión a lo largo de estos años;

A *Enrique Marco de Lucas*, por haber sido una de las semillas del proyecto;

A mis *compañeros Amaia Pérez del Barrio y Pablo Sanz Bellón*, porque han sido siempre un apoyo a lo largo de la duración de este trabajo;

Al *grupo de Computación Avanzada y e-Ciencia del Instituto de Física de Cantabria*, por prestar soporte técnico y ceder, de manera desinteresada, los recursos de computación necesarios para llevar a cabo esta investigación.

Índice general

Índice de tablas

Índice de figuras

Índice de abreviaturas

AUC *Area Under the Receiver Operating Characteristic Curve.*
CNN *convolutional neural networks.*
COVID-19 *Coronavirus Disease* 2019.
DICOM *Digital Imaging and Communication In Medicine.*
DL *Deep Learning.*
DS *distribution shifts.*
EI exposure index.
GB *Gigabytes.*
GPUs Unidades de Procesamiento Gráfico.
Grad-CAM *Gradient-weighted Class Activation Mapping.*
IA Inteligencia Artificial.
IC intervalos de confianza.
IE índices de exposición.
KS *Kolmogorov-Smirnov.*
KSL *Kolmogorov-Smirnov-Lilliefors.*
ML *Machine Learning.*
MWU *U-Mann-Whitney-Wilcoxon.*
PA postero-anterior.
ROC *Receiver Operating Characteristic Curve.*
RXT radiografías de tórax.
SGD *Stochastic Gradient Descent.*

1

Introducción

1. Introducción

Las publicaciones en medicina referentes a sistemas de Inteligencia Artificial (IA) se han multiplicado de forma exponencial en las últimas dos décadas y, especialmente, en el último lustro [1]. En sus orígenes, esta disciplina, definida como la capacidad de las máquinas de simular funciones cognitivas humanas [2], consistía en la programación de algoritmos que simulaban la toma de decisiones de un ser humano. Este primer abordaje dio lugar a lo que se denominó la IA simbólica o guiada por el conocimiento, ya que los patrones programados eran el resultado del conocimiento adquirido por el ser humano a través de su experiencia [3].

Posteriormente, surgió una nueva aproximación que impulsó de forma notoria el desarrollo de la IA y que marcaría, desde sus origines, el camino más prometedor para el crecimiento de esta doctrina: la llamada IA guiada por datos, también conocida como *Machine Learning* (ML) o aprendizaje automático [3].

A diferencia de la IA simbólica, que se centraba en la programación de patrones para la resolución de tareas específicas, el ML consiste en la programación de algoritmos de aprendizaje, es decir, en la programación de algoritmos capaces de aprender por sí mismos los patrones necesarios para la resolución de una tarea específica tras un proceso de exposición a los datos denominado entrenamiento [4].

Dentro de este numeroso grupo de algoritmos, existe un subgrupo que se encuentra conformado por las llamadas redes neuronales artificiales, y que se denomina comúnmente *Deep Learning* (DL) o aprendizaje profundo [3]. Estos algoritmos son considerados en la actualidad el estado del arte en múltiples campos de la IA, entre los que se encuentra el campo de la visión artificial. Es por esta razón que un tipo concreto de redes neuronales, las llamadas redes neuronales convolucionales o *convolutional neural networks* (CNN), son consideradas actualmente el estado del arte en clasificación de imagen médica [5].

Paralelamente, en los últimos años se ha observado una tendencia creciente en la realización de estudios de imagen en medicina [6][7], tendencia que ha sido más acusada desde el origen de la pandemia por *Coronavirus Disease* 2019 (COVID-19) en diciembre de 2019 [8], [9]. Este incremento en la demanda de estudios radiológicos no se ha acompasado de un incremento en el número de radiólogos [7], lo que ha ido progresivamente en detrimento de la calidad diagnóstica [10]. En este contexto, ha emergido con fuerza el DL como una tecnología que ya ha demostrado ser capaz de asistir en la tarea de interpretación de imágenes médicas, consiguiendo una mejora sustancial en la calidad de los diagnósticos [11].

Sin embargo, publicaciones previas han advertido de que, a menudo, los algoritmos de DL para análisis de imagen médica presentan buenos resultados sobre los subconjuntos de entrenamiento, validación y test utilizados en las fases de desarrollo y validación del algoritmo, si bien fracasan posteriormente cuando son utilizados en entornos ajenos al de desarrollo [12]-[18]. Este fenómeno responde a una incapacidad

de los algoritmos para generalizar, es decir, para mantener el rendimiento observado sobre los subconjuntos de entrenamiento y validación, sobre datos externos.

Esta incapacidad para generalizar es frecuentemente obviada durante el proceso de desarrollo de los modelos de DL, ya que si bien la validez interna es evaluada correctamente sobre un subconjunto de validación, la validez externa o generalización es habitualmente estimada utilizando un subconjunto de test que proviene de la misma población de la que se obtuvieron los subconjuntos de entrenamiento y validación, y que por tanto brinda nuevamente resultados de validez interna y no de validez externa [15].

De acuerdo con lo anterior, y tal y como ha sido publicado por algunos autores, el rendimiento de los algoritmos de DL debería de ser evaluado de forma sistemática utilizando datos procedentes de fuentes diferentes de las que se obtuvieron los subconjuntos de entrenamiento y validación, lo cual no constituye la práctica habitual en el campo de la imagen médica [12], [15]-[18]. Este hecho es de gran importancia, ya que un requisito fundamental para que el uso de esta tecnología pueda extenderse y alcanzar la práctica médica, es que los algoritmos sean capaces de generalizar.

Sorprendentemente, esta deficiencia en la capacidad de generalización ha sido muy poco estudiada en el campo de la imagen médica, a pesar de que en otras áreas de la IA, como en el desarrollo de coches autónomos, ha cobrado una relevancia importante [19].

En este trabajo se abordará la problemática de la deficiencia de generalización que presentan, a menudo, los algoritmos de DL aplicados al análisis de imagen médica. Para ello, se utilizarán CNN para realizar una tarea de clasificación o diagnóstico de radiografías de tórax (RXT) de pacientes COVID-19 y pacientes sanos, denominados en esta investigación como pacientes control. Con ello, se pretenderá arrojar luz al problema de la generalización planteado, con el objetivo de poder caminar hacia la resolución del mismo y, en último término, hacia la aproximación de IA y del DL a la práctica médica habitual.

2

Estado actual del problema

2. Estado actual del problema

La primera controversia en torno al problema de la generalización de los algoritmos de DL para imagen médica, radica en la cuestión de si la generalización puede ser lograda o no por los algoritmos actuales. A pesar de que algunos autores sugieren que un cierto grado de generalización a centros externos se puede alcanzar con caída del rendimiento del algoritmo [15], [20]; otros defienden que un algoritmo de DL para clasificación de RXT no puede ser capaz de generalizar a centros ajenos al de entrenamiento [12], [16] o a centros que presenten diferencias significativas con aquel de donde provienen las imágenes utilizadas para desarrollar el modelo [18].

La segunda controversia se encuentra en torno al origen o causa de la falta de generalización reportada. La escasa literatura que ha abordado esta cuestión apunta a que el origen se encuentra en la existencia de desemejanzas entre los datos de entrenamiento y los datos sobre los que se aplicará el modelo final entrenado [16], [18]. Estas disimilitudes parecen responder a diferencias, a menudo imperceptibles, en la distribución de ciertos atributos de los propios datos, dando lugar a lo que se ha denominado *distribution shifts* (DS) [16], [18].

Sin embargo, los atributos o causas que se encuentran tras los mencionados DS y, en último término, tras la falta de generalización, permanecen aún desconocidas en el campo de la imagen médica [16], [18]. La hipótesis actual ampliamente más aceptada afirma que se deben a diferencias entre los centros de entrenamiento y los centros de aplicación [12], [21]; diferencias que han sido poco descritas [21].

A día de hoy, la mayoría de los autores que han tratado esta cuestión se han limitado al estudio de la generalización a través de la evaluación del rendimiento de diferentes algoritmos en subconjuntos de test provenientes de instituciones externas a las de entrenamiento, no prestando atención a factores que podrían ser relevantes, como el modelo de equipo utilizado para adquirir las imágenes, el protocolo de adquisición, o el número de centros que aportaron las imágenes de entrenamiento [12], [15], [16], [20].

Esta investigación, a diferencia de los trabajos previamente publicados, se ha propuesto estudiar el efecto aislado de los factores que, de acuerdo con la literatura y nuestro conocimiento, podrían afectar a la generalización de los algoritmos de DL desarrollados para el análisis de imagen médica, prestando atención, no solo a la existencia o no del efecto, sino también a la magnitud del mismo. Además, en este trabajo se analiza por separado la influencia de estos factores sobre la validez interna y externa o generalización del algoritmo, así como su capacidad para aprender relaciones causales a través de análisis de explicabilidad, dos aspectos también diferenciales con publicaciones anteriores.

Los factores estudiados han sido divididos en dos categorías, y son los siguientes:

- Factores asociados a la institución de la que provienen las imágenes de entrenamiento: corresponden a diferentes agentes que no modifican los valores de los píxeles de

las imágenes y que vienen condicionados por el centro en el que se adquieren las imágenes. Estos factores son, principalmente, la propia institución de la que provienen las imágenes, la demografía poblacional, la epidemiología de la enfermedad, o el criterio de etiquetado de las imágenes del centro.

- Factores asociados al equipo de adquisición de las imágenes: corresponden con aquellos que sí modifican el valor de los píxeles de las imágenes, y son el protocolo de adquisición, y el modelo de equipo, que comprende el tipo de función de respuesta del detector, y el procesado de imagen aplicado por el mismo y específico del fabricante [22].

3

Hipótesis y objetivos

3. Hipótesis y objetivos

3.1. Hipótesis

Teniendo en cuenta el contexto descrito previamente en el Apartado 2, este trabajo hipotetizó que un algoritmo de inteligencia artificial de tipo DL, en concreto una red neuronal convolucional, podría lograr clasificar correctamente radiografías de pacientes COVID-19, y que además, podría hacerlo en centros externos al lugar donde el algoritmo había sido entrenado, es decir, que sería capaz de generalizar a entornos ajenos.

Sin embargo, también se planteó que probablemente los factores descritos en la literatura y detallados en el citado Apartado 2, podrían afectar a la capacidad de actuación del algoritmo, tanto en validez interna como en validez externa o generalización.

De entre estos factores, se propuso que los relacionados con el equipo de adquisición de rayos X serían, probablemente, los más influyentes. En concreto, se postuló que el modelo de equipo de adquisición podría ser el elemento más condicionante a la hora de conseguir que el algoritmo desarrollado fuera capaz de generalizar.

De acuerdo con esta hipótesis, las imágenes radiológicas podrían tener diferentes texturas dependiendo del equipo de adquisición de las mismas y, en menor medida, también del protocolo de adquisición. Estas diferencias texturales podrían ser las principales responsables de que existiera un *distribution shift* entre los datos de entrenamiento y los datos de test, responsable de la deficiencia de robustez o generalización de los algoritmos de DL.

De ser cierta esta hipótesis, sí sería posible lograr un alto nivel de generalización a un centro externo con una red neuronal para una tarea de clasificación de imagen médica, siempre y cuando el algoritmo fuera utilizado con imágenes adquiridas por un equipo del mismo modelo que el que obtuvo las imágenes de entrenamiento. En otras palabras, controlando el factor equipo, se reduciría el *distribution shift* lo suficiente como para poder generalizar.

3.2. Enumeración de las hipótesis específicas

Las hipótesis específicas de este trabajo se enumeran a continuación:

1. Un algoritmo de inteligencia artificial de tipo aprendizaje profundo puede lograr un alto nivel de precisión en la interpretación de radiografías.
2. Un algoritmo de aprendizaje profundo para la interpretación de radiografías, puede generalizar a radiografías provenientes de poblaciones diferentes de aquellas que dieron lugar a los datos utilizados en su entrenamiento.
3. Existen algunos factores que pueden influir en la capacidad de generalización de los algoritmos de aprendizaje profundo para la interpretación de radiografías.

4. El equipo de adquisición de las imágenes médicas, es el factor más influyente en la capacidad de generalización de los algoritmos de aprendizaje profundo para la interpretación de radiografías.

3.3. Objetivo principal

El objetivo principal de esta investigación es evaluar la capacidad de generalización de un algoritmo de DL de tipo CNN entrenado para realizar diagnóstico médico, esclareciendo en el proceso los factores que determinan dicha capacidad.

3.4. Objetivos específicos

Con el fin de acometer el objetivo principal, se plantean los objetivos específicos que se enumeran a continuación, y cuya consecución cronológica, permitirá alcanzar el objetivo principal de la investigación.

1. Desarrollar un algoritmo de inteligencia artificial de tipo aprendizaje profundo para la interpretación de radiografías.
2. Evaluar la capacidad de generalización de un algoritmo de aprendizaje profundo para la interpretación de radiografías.
3. Detectar los factores más relevantes que pudieran influir en la capacidad de generalización de los algoritmos de aprendizaje profundo para la interpretación de radiografías.
4. Cuantificar la influencia del equipo de adquisición de las imágenes médicas en la capacidad de generalización de los algoritmos de aprendizaje profundo para la interpretación de radiografías.

4

Materiales y métodos

4. Materiales y métodos

4.1. Introducción

Se diseñaron tres experimentos que permitieron evaluar la capacidad diagnóstica de neumonía por COVID-19 de una CNN en RXT, atendiendo tanto a su rendimiento en validez interna, como a su rendimiento en validez externa o generalización. Este diseño experimental compartimentado en tres ensayos diferentes, permitió estudiar el efecto que ejercían sobre la capacidad de aprendizaje y sobre la capacidad de generalización de un algoritmo de DL los factores descritos en el Apartado 2: aquellos asociados con la institución, y aquellos asociados con el equipo de adquisición de rayos X.

En concreto, se estudió la influencia de los siguientes factores sobre la validez interna y la generalización: el centro de donde se obtuvieron los datos de entrenamiento, incluyendo el conjunto de elementos asociados al mismo, como las diferencias en el criterio de etiquetado de las imágenes, o las diferencias en la demografía poblacional; el protocolo de adquisición de las imágenes; y el equipo de adquisición de rayos X, lo que incluyó el estudio por separado del efecto del tipo de función de respuesta del detector, y el efecto inducido por el procesado de imagen aplicado por el fabricante.

Los tres experimentos se llevaron a cabo utilizando la misma arquitectura, una CNN conformada por la parte convolucional de una VGG16 y un clasificador conectado mediante una capa de *Max Pooling*. Los hiperparámetros escogidos, que se detallarán más adelante, tampoco variaron entre experimentos, lo que permitió atribuir las diferencias de rendimiento diagnóstico observadas entre los diferentes entrenamientos desarrollados a factores externos al propio algoritmo.

4.2. Aprobación del Comité de Ética

Esta investigación ha involucrado pacientes de dos instituciones: del Hospital Universitario Marqués de Valdecilla, en Santander, denominado en adelante como Institución 1; y del Hospital de Sierrallana y Tres Mares, en Torrelavega, denominado en adelante como Institución 2. De acuerdo con lo establecido en las guías de buenas prácticas y calidad científica, este trabajo ha recibido la aprobación del comité de ética correspondiente al ámbito de aplicación, el Comité de Ética de la Investigación con Medicamentos de Cantabria. El certificado de aprobación emitido por este comité, se adjunta en el Sección 9.1.

4.3. Reclutamiento de los pacientes

Los pacientes fueron reclutados aleatoriamente desde cuatro bases de datos (Figura 4.1). La primera base de datos contenía a todos los pacientes a los que se les

habían realizado RXT en la Institución 1 adquiridas por el equipo *Fujifilm FDR Smart FGX*, entre el 15 de septiembre de 2019 y el 25 de noviembre de 2020. La segunda base de datos recogía a los pacientes con RXT adquiridas en la Institución 2 por un equipo del mismo modelo, un *Fujifilm FDR Smart FGX*, durante el mismo período. La tercera base de datos incluyó a los pacientes con RXT adquiridas en la Institución 2 por un equipo *Revolution XRD de General Electric*, durante el período comprendido entre el 1 de enero de 2020 y el 25 de noviembre de 2020. Por último, la cuarta base de datos contenía a todos los pacientes con RXT adquiridos en la Institución 2 por un *Carestream DRX Evolution Plus*, entre el 1 de enero de 2018 y el 25 de noviembre de 2020.

Posteriormente, las imágenes fueron filtradas, de forma que solo se preservaron las proyecciones postero-anterior (PA), y en caso de que a un paciente se le hubieran realizado varias RXT, solo se preservó la radiografía cronológicamente más antigua. El objetivo de esta medida fue eliminar los posibles sesgos que pudieran derivarse de incluir varias imágenes pertenecientes al mismo paciente.

Es importante destacar que de los equipos utilizados, los dos *Fujifilm* y el *Carestream*, emplearon una función de respuesta logarítmica, mientras que el equipo *General Electric* utilizó una función de respuesta de tipo lineal [23].

4.4. Etiquetado de las imágenes

Las imágenes seleccionadas fueron posteriormente etiquetadas manualmente por radiólogos expertos en dos clases: COVID-19 y control. Los criterios de inclusión para cada una de las dos categorías se resumen en la Tabla 4.1. Las imágenes que no cumplieron todos los criterios de inclusión de alguna clase, fueron excluidas de la investigación. En este texto, estas dos categorías se denominarán clases objetivo.

Tabla 4.1: Criterios de inclusión para las clases objetivo.

Clase objetivo	Historia clínica	Hallazgos de imagen
COVID-19	Test RT-PCR positivo	Informe de al menos tres radiólogos reportando hallazgos compatibles con COVID-19 en la RXT.
Control	Sin síntomas de COVID-19	Informe de al menos tres radiólogos reportando ausencia de hallazgos patológicos en la RXT.

RT-PCR = reacción en cadena de la polimerasa con transcriptasa inversa.
Se requirió el cumplimiento de todos los criterios de inclusión para incluir una imagen en una de las dos clases.

4.5. Conjunto y subconjuntos de datos utilizados

Como se indicó previamente, se seleccionaron RXT en proyección PA adquiridas en dos instituciones y por cuatro equipos radiológicos diferentes, y se etiquetaron manualmente como COVID-19 o control (Figura 4.1).

Una vez que las imágenes habían sido recopiladas y seleccionadas, se realizó un muestreo aleatorio con estratificación por clase objetivo y por equipo para construir el conjunto de datos principal. Este conjunto contenía:

- 394 RXT adquiridas en la Institución 1 por el dispositivo *Fujifilm FDR Smart FGX*.
- 244 RXT adquiridas en la Institución 2 por el dispositivo *Fujifilm FDR Smart FGX*.
- 192 RXT adquiridas en la Institución 2 por el dispositivo *General Electric Revolution XRD*.
- 44 RXT adquiridas en la Institución 2 por el dispositivo *Carestream DRX Evolution Plus*.

Los tamaños muestrales fueron escogidos en base al número de imágenes disponibles de cada dispositivo, una vez realizado el proceso de limpieza y filtrado.

Posteriormente, el conjunto de datos principal se dividió en ocho subconjuntos basándose en el equipo de adquisición de las imágenes y en la institución donde habían sido adquiridas. Los subconjuntos resultantes fueron los siguientes:

- F1: 150 imágenes adquiridas en la Institución 1 por el equipo *Fujifilm FDR Smart FGX*.
- F2: 150 imágenes adquiridas en la Institución 1 por el equipo *Fujifilm FDR Smart FGX*.
- F3: 150 imágenes adquiridas en la Institución 2 por el equipo *Fujifilm FDR Smart FGX*.
- S2: 98 imágenes adquiridas en la Institución 2 por el equipo *General Electric Revolution XRD*.
- T1: 94 imágenes adquiridas en la Institución 1 por el equipo *Fujifilm FDR Smart FGX*.
- T2: 94 imágenes adquiridas en la Institución 2 por el equipo *Fujifilm FDR Smart FGX*.
- T3: 94 imágenes adquiridas en la Institución 2 por el equipo *General Electric Revolution XRD*.
- T4: 44 imágenes adquiridas en la Institución 2 por el equipo *Carestream DRX Evolution Plus*.

Los subconjuntos F1, F2, F3 y S2 se utilizaron para entrenar los algoritmos de DL, mientras que los subconjuntos T1, T2, T3 y T4 se utilizaron para evaluar el rendimiento de los modelos entrenados (Figura 4.2).

Es importante destacar que todos los subconjuntos se estratificaron por clase objetivo, por tanto todos ellos contenían el mismo número de imágenes COVID-19 y control.

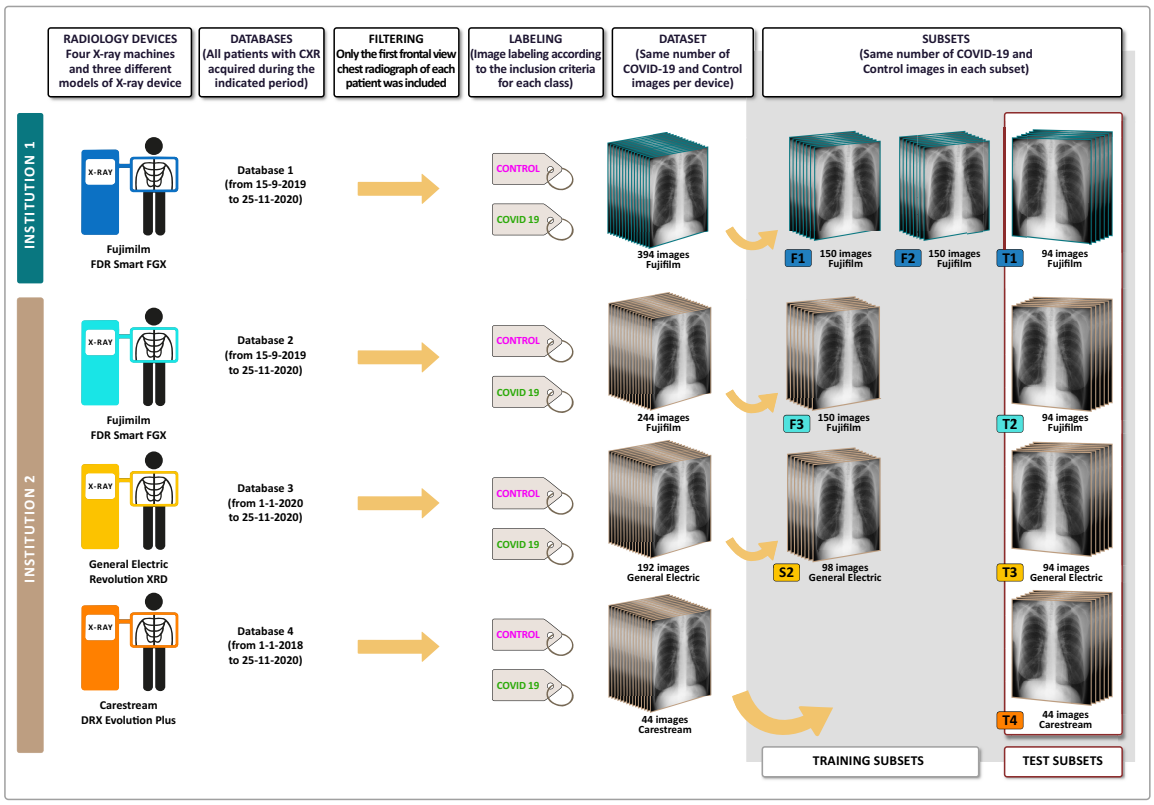


Figura 4.1: Procesos de reclutamiento de los pacientes y de elaboración de los conjuntos de datos de entrenamiento y test.

4.6. Preprocesamiento de las imágenes

Las imágenes fueron recopiladas como píxeles monocromos de 16 bits enteros sin signo, en formato *Digital Imaging and Communication In Medicine* (DICOM). Tras la recopilación de las mismas, se aplicó un preprocesamiento. El primer paso del preprocesamiento consistió en la asignación de una ventana adecuada. A continuación, se invirtieron los valores de los píxeles de las imágenes cuando procedió. Posteriormente, las imágenes fueron redimensionadas a 512 x 512 píxeles utilizando una interpolación *spline* cúbica [24]. Se decidió utilizar este tipo de técnica para realizar la interpolación, dado que emplea polinomios de bajo grado, lo que evita las oscilaciones indeseadas producto de emplear otras técnicas que utilizan polinomios de alto grado. Por último, los valores de los píxeles se re-escalaron de forma que el valor más pequeño posible fuera el 0 y el más alto el 1. Finalmente, las imágenes se apilaron en tres canales, ya que este es un requisito necesario para poder utilizar los modelos pre-entrenados sobre *ImageNet*.

4.7. Protocolos de adquisición de las imágenes

Los protocolos de adquisición utilizados se detallan en la Tabla 4.2.

Para valorar si existían diferencias entre los protocolos de adquisición de los equipos *Fujifilm FDR Smart FGX* de ambas instituciones, se compararon los índices de exposición (IE). De esta forma se pudo conocer si existían diferencias estadísticamente significativas entre los protocolos de adquisición de ambos equipos.

Se seleccionó la métrica IE como métrica representativa y resumen del protocolo, ya que mide el *kerma* en aire en la superficie del detector [25], lo que significa que cuantifica la dosis de radiación que alcanza el panel. Es importante notar, no obstante, que cada fabricante tiene su propio IE, el cual es calculado de manera diferente [26], no siendo por tanto una métrica que permita la comparación entre dispositivos de fabricantes diferentes.

Tabla 4.2: Estadística descriptiva de los parámetros de adquisición utilizados más relevantes.

Equipo	IE / Exposición relativa*	Kilovoltaje pico (kvp)	Corriente de rayos X en el tubo (mA)	Exposición expresada (µAs)	Distancia fuente-detector (mm)	Duración de la exposición (ns)
***Fujifilm*[1]**						
Mediana	130.00	120.00	200.00	2500.00	1000.00	12.00
Media	132.84	119.99	200.66	2655.01	1000.00	13.29
SD	36.00	0.09	4.39	1160.39	0.00	5.86
Min	37.00	119.00	192.00	600.00	1000.00	3.00
Max	330.00	120.00	212.00	11700.00	1000.00	59.00
***Fujifilm*[2]**						
Mediana	114.00	119.00	192.00	2500.00	1000.00	13.00
Media	120.52	119.99	192.60	2680.74	1000.00	14.00
SD	41.17	1.11	2.09	1252.42	0.00	6.37
Min	69.00	104.00	188.00	900.00	1000.00	5.00
Max	547.00	124.00	210.00	12900.00	1000.00	63.00
***General Electric*[2]**						
Mediana	64.50	120.00	200.00	1360.00	1800.00	6.00
Media	70.88	119.90	202.08	1485.94	1780.10	6.89
SD	32.88	1.79	20.36	688.48	116.27	3.44
Min	26.00	108.00	200.00	540.00	1000.00	2.00
Max	309.00	130.00	400.00	6360.00	1930.00	31.00
***Carestream*[2]**						
Mediana	148.57	125.00	125.00	1650.00	1795.50	13.50
Media	146.81	124.89	125.80	1734.09	1796.89	13.84
SD	36.00	0.75	5.28	674.78	5.94	5.19
Min	7.65	120.00	125.00	100.00	1792.00	1.00
Max	244.57	125.00	169.00	3500.00	1820.00	25.00

IE = índice de exposición.

*IE en el caso de los equipos *Fujifilm* y del equipo *Carestream*; Exposición relativa en el caso del equipo *General Electric*.

[1]Equipo de la Institución 1; [2]Equipo de la Institución 2.

4.8. Entrenamiento de los modelos

Se entrenó una CNN para clasificar RXT de pacientes con COVID-19 y de pacientes control. La arquitectura de la CNN se conformó de la parte convolucional de una VGG16, a la que se le acopló una capa de *Max Pooling*, y de un clasificador consistente en una neurona de salida con una función de activación sigmoide. La arquitectura convolucional escogida fue una VGG16, ya que esta red había demostrado buenos resultados detectando neumonía RXT en trabajos previos [27]-[29].

En cuanto a los hiperparámetros, la VGG16 se importó con los pesos preentrenados para *ImageNet* siguiendo la técnica de *transfer learning*. Posteriormente, se realizó un *fine tuning* descongelando todas las capas durante el entrenamiento y utilizando un *learning rate* de 0.000075. Además, se utilizó una inicialización *Glorot Normal* [30] para ajustar el clasificador.

Como optimizador, se optó por un *Stochastic Gradient Descent* (SGD), y como función de pérdida, por una *binary cross-entropy*.

Con esta configuración, se entrenaron tres modelos utilizando diferentes subconjuntos de imágenes (Figura 4.2). La técnica de entrenamiento utilizada fue una validación cruzada, en concreto una *k-fold Cross Validation* con estratificación, definiendo un valor de k = 5. El número de épocas fijado fue de 500, programando un *early stopping* con una paciencia de 50 épocas. El tamaño de *batch* empleado fue de 16 imágenes.

Para los tres entrenamientos se emplearon 300 RXT y los subconjuntos descritos en la la Sección 4.5, obteniendo como resultado los siguientes modelos:

- Modelo-F1F2: entrenado con los subconjuntos F1 y F2, es decir, con imágenes adquiridas en una sola institución y por un solo equipo.
- Modelo-F1F3: entrenado con los subconjuntos F1 y F3, es decir, con imágenes adquiridas en dos instituciones diferentes, pero por un solo modelo de equipo.
- Modelo-F1'F3'S2: entrenado con 101 imágenes obtenidas aleatoriamente del subconjunto F1, 101 imágenes obtenidas aleatoriamente del subconjunto F3, y las 98 imágenes del subconjunto S2. Por tanto, este modelo fue entrenado con imágenes adquiridas en dos instituciones diferentes y por dos modelos de equipos distintos.

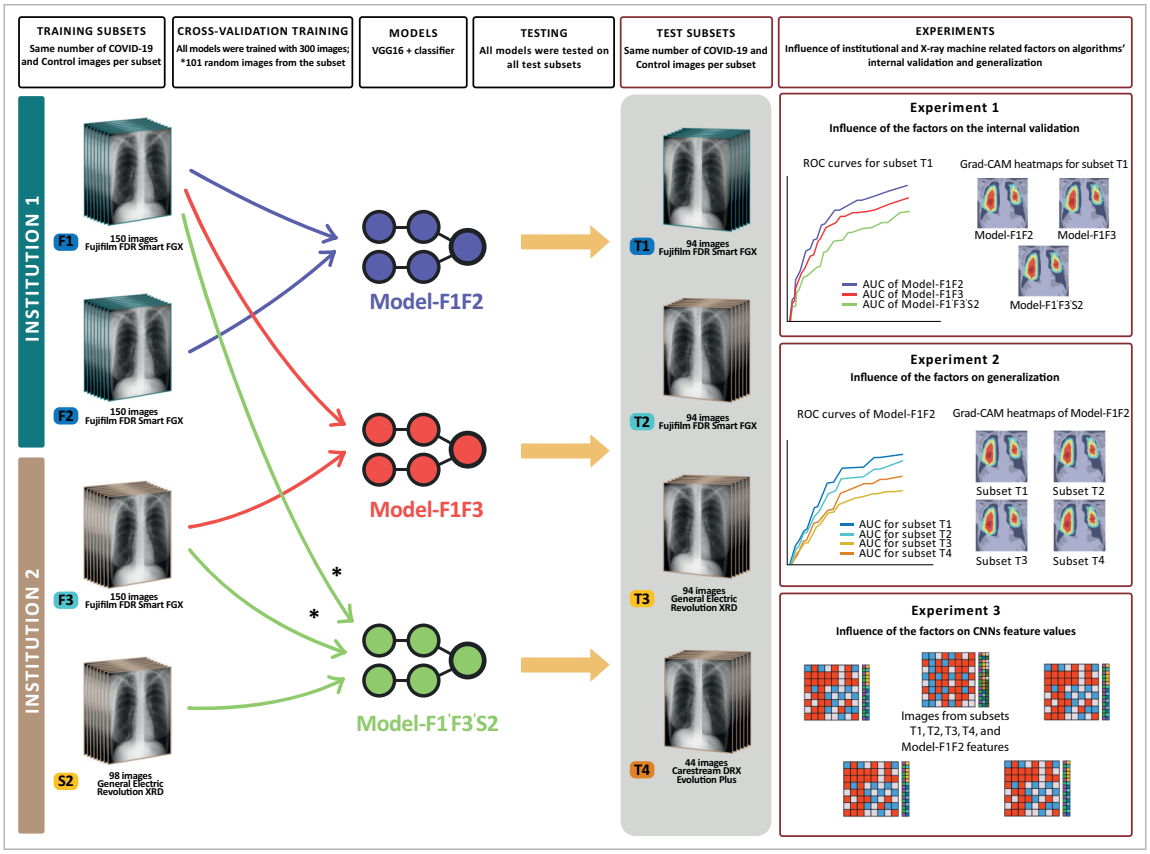


Figura 4.2: Diseño experimental. Se entrenó un algoritmo de *Deep Learning* de tipo CNN de tres formas diferentes. En primer lugar se realizó un entrenamiento utilizando los conjuntos F1 y F2, lo que dio como resultado un modelo que se denominó Modelo-F1F2. Posteriormente se realizaron otros dos entrenamientos, uno utilizando los subconjuntos F1 y F3, del que se obtuvo el modelo Modelo-F1F3, y otro utilizando 101 imágenes aleatorias de los subconjuntos F1 y F3, y el subconjunto S2, obteniendo el modelo Modelo-F1'F3'S2.

4.9. Evaluación de los modelos

La evaluación de los modelos entrenados se llevó a cabo desde tres enfoques diferentes con el objetivo de estudiar, no solo el rendimiento de los modelos, sino también los elementos aprendidos por el algoritmo, y la influencia de los factores propuestos en el Apartado 2 sobre los valores de las *features* del mismo. Estos tres enfoques fueron los siguientes:

- Rendimiento: consistió en la evaluación de la capacidad diagnóstica de los modelos o capacidad para clasificar correctamente las RXT.
- Explicabilidad: el análisis de la explicabilidad se centró en tratar de entender los elementos en base a los cuales el algoritmo tomaba las decisiones.

- *Features*: se llevó a cabo un análisis de la influencia directa que tenían los factores estudiados en este trabajo sobre el valor de las *features* extraídas por la parte convolucional del algoritmo y utilizadas por el clasificador para confeccionar una predicción.

Para los tres enfoques y los tres modelos desarrollados, se utilizaron los subconjuntos de test previamente explicados en la Sección 4.5.

4.9.1. Análisis del rendimiento

Para el análisis del rendimiento se emplearon curvas de *Receiver Operating Characteristic Curve* (ROC), que son gráficos que representan visualmente la capacidad diagnóstica de un clasificador binario para todos los puntos de corte [31]. Este tipo de curvas se construye representando la sensibilidad o tasa de verdaderos positivos, frente a (1 – especificidad) o tasa de falsos positivos. Dado que realmente son curvas de probabilidad, el *Area Under the Receiver Operating Characteristic Curve* (AUC) representa la probabilidad global de acierto del clasificador. De esta forma, cuanto mayor sea el valor del AUC, mayor será la capacidad diagnóstica del modelo, con un máximo posible de 1, o del 100 % expresado en términos porcentuales [31].

Por todo ello, y teniendo en cuenta que todos los subconjuntos de test utilizados en este trabajo estaban balanceados por clase, es decir, contaban con el mismo número de pacientes COVID-19 y pacientes control, se optó por el empleo de curvas de ROC y de su AUC para estimar la capacidad diagnóstica de los diferentes modelos desarrollados y hacer comparaciones del rendimiento entre los mismos.

4.9.2. Análisis de la explicabilidad

Con el objetivo de tratar de entender los elementos que habían sido aprendido por los modelos, es decir, los elementos de la imagen que utilizaba el algoritmo para confeccionar una predicción y clasificar una RXT en una de las dos clases objetivo, COVID-19 o control, se utilizaron los mapas de calor generados por el algoritmo *Gradient-weighted Class Activation Mapping* (Grad-CAM) [32]. Esta técnica, convertida en la actualidad en una de las más populares dentro del mundo de la explicabilidad de las redes neuronales de visión artificial, consiste en generar mapas de calor que destacan las regiones de la imagen que han sido consideradas como más importantes por la red neuronal para llevar a cabo la clasificación. Para ello, utilizan los gradientes de la clase objetivo de la última capa convolucional [32].

La aplicación del algoritmo Grad-CAM permitió identificar los puntos de la imagen en los que los modelos entrenados se centraban para tomar una decisión, por lo que gracias a él se hizo posible estudiar si los modelos decidían en base a características clínicas propias de la enfermedad COVID-19, como las opacidades pulmonares, o en base a características texturales sin significado clínico. En otras palabras, si las predicciones de los modelos se basaban en relaciones espurias o en relaciones causales, lo cual es esencial para lograr la generalización.

4.9.3. Análisis de *clustering* jerarquizado

Para estudiar si los factores propuestos en el Apartado 2 tenían influencia sobre el valor de las *features* extraídas y utilizadas por el algoritmo para realizar el diagnóstico de las RXT, se empleó un algoritmo de ML no supervisado, el *clustering* jerarquizado [33].

Este algoritmo realiza agrupaciones de los datos en *clusters* basándose en las similitudes intrínsecas existentes en los propios datos. Además, lo hace mediante un abordaje no supervisado, lo cual reduce la posibilidad de que existan sesgos humanos en el análisis [33]. Por esta razón, se consideró que este tipo de técnica podía ser un método útil para estudiar si los valores que tomaban las *features* extraías por la red neuronal y en base a las cuales confeccionaba su predicción, dependían de los factores objeto de estudio en este trabajo. Se decidió denominar como clases ocultas de alto nivel a aquellas derivadas de factores ajenos a las clases objetivo, COVID-19 y control en este trabajo, con mayor poder discriminativo para separar las imágenes en clases, que las propias clases objetivo.

De acuerdo con estos términos, se consideraron el centro del que provenían los datos, el modelo de equipo de rayos X utilizado para adquirir las imágenes, el procesado de imagen aplicado por el fabricante, y el tipo de función de respuesta del detector, como posibles clases ocultas.

Para realizar este análisis, se comenzó con la extracción de los mapas de *features* de la última capa convolucional de los modelos Modelo-F1F2 y Modelo-F1'F3'S2. Posteriormente, se aplicó un *Global Max Pooling* a los mapas de *features* con el fin de obtener un valor único y representativo de cada mapa, que se consideró el valor de la *feature*. A continuación, se realizó una estandarización de los valores obtenidos y se aplicó el algoritmo de *clustering* jerarquizado utilizando el lenguaje de programación *Python* (versión 3.6.8) [34] y la librería *Seaborn* (versión 0.11.1) [35]. Finalmente, se examinaron los *clusters* de imágenes resultantes hallados por el algoritmo intentando identificar si correspondían con las clases objetivo o con alguna de las clases ocultas propuestas.

Los *clusters* fueron fácilmente reconocibles de forma visual ya que se utilizaron mapas de calor en los que se representaron tanto la clase objetivo como las clases ocultas de cada imagen mediante un código de colores.

La agrupación jerárquica se ejecutó cuatro veces. En primer lugar, el algoritmo de *clustering* jerarquizado se ejecutó con las *features* extraídas por el Modelo-F1F2 de los subconjuntos T1, T2 y T3. En segundo lugar, las imágenes de los subconjuntos T1, T2 y T3 se recortaron preservando únicamente las zonas centrales de la imagen, libres de etiquetas metálicas (Figure 4.3), y se ejecutó de nuevo el algoritmo de *clustering* jerarquizado con las *features* extraídas por el Modelo-F1F2 de las imágenes recortadas. El objetivo de este proceso fue el de comprobar que las etiquetas metálicas que indicaban la lateralidad en las RXT no estaban participando en la formación de los *clusters*. En tercer lugar, el algoritmo de *clustering* se ejecutó con las *features* extraídas por el Modelo-F1F2 de 44 imágenes muestreadas aleatoriamente

de los subconjuntos T1, T2 y T3 y de las 44 imágenes del T4. Por último, se ejecutó el algoritmo de *clustering* con las *features* extraídas por el Modelo-F1'F3'S2 antes y después de ser entrenado de las imágenes de los subconjuntos T1, T2 y T3. El propósito de este último paso fue el de poder evaluar si el proceso de entrenamiento y el de *fine tuning* modificaban la influencia de estos factores sobre los valores de las *features*.

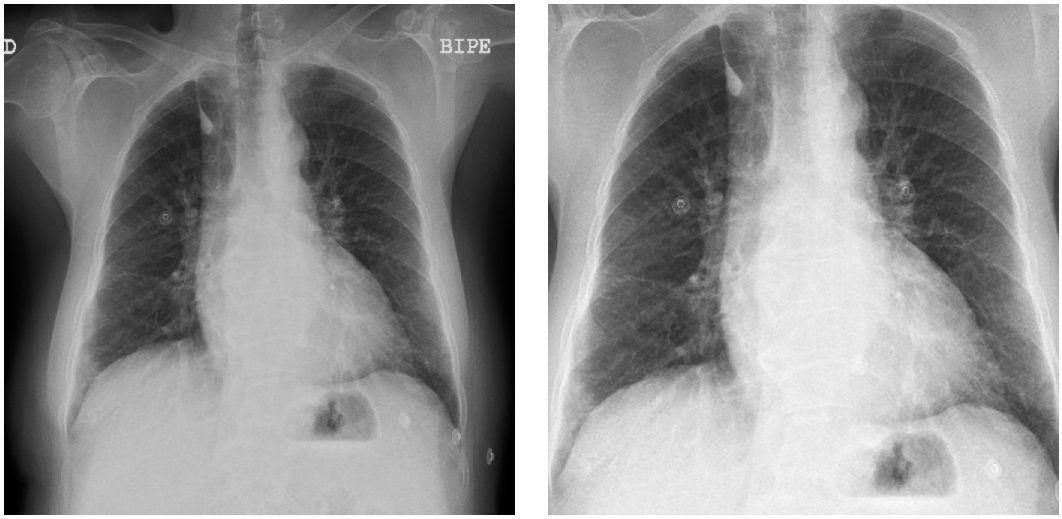


Figura 4.3: Ejemplo de cómo fueron recortadas las imágenes para eliminar las etiquetas metálicas. A la izquierda se puede observar la imagen original, mientras que a la derecha se presenta su versión recortada.

4.10. Análisis estadístico

Para establecer si existían diferencias estadísticamente significativas entre los IE, se comenzó realizando un test de normalidad, en concreto el test de *Kolmogorov-Smirnov-Lilliefors* (KSL) [36]-[38]. Posteriormente, conocido el resultado del test anterior, se realizaron un test de *U-Mann-Whitney-Wilcoxon* (MWU) [39], [40] y un test de *Kolmogorov-Smirnov* (KS) para muestras independientes [38], con el objetivo de esclarecer si existían diferencias estadísticamente significativas entre las distribuciones de los IE. Estos test fueron realizados utilizando el lenguaje de programación *Python* (version 3.6.8) [34] y las librerías *statsmodels* version(0.12.2) [41] y *SciPy* (versión 1.5.4) [42].

Los intervalos de confianza (IC) al 95 % de las AUC calculadas por *k-fold cross validation* se calcularon utilizando el lenguaje de programación *R* (versión 4.1.0) [43] con el paquete *cvAUC* (versión 1.1.0) [44]. Para calcular las diferencias entre las AUC con sus IC al 95 %, se utilizó la técnica del *bootstrapping* [45], considerando

cualquier diferencia en la que el IC al 95 % no incluyera el 0, como estadísticamente significativa con un *p-valor* $< 0{,}05$. Por último, se realizó también el test *DeLong* para muestras emparejadas [46] como segundo método de comparación. Para ello, se utilizó el paquete de programación, disponible en R, *pROC* 3.6 [47].

Teniendo en cuenta que el test de *DeLong* es a veces excepcionalmente conservador [48], se consideró el *bootstrapping* como la técnica de elección en caso de discrepancias entre los resultados obtenidos por ambas técnicas. El *p-valor* considerado para establecer la significación estadística fue, en todos los casos, de 0.05 a dos colas.

4.11. Recursos de programación empleados

En esta investigación se utilizaron dos lenguajes de programación: *Python* (versión 3.6.8) [34] y *R* (versión 4.1.0) [43]. Las principales librerías de *Python* empleadas fueron:

- *Pydicom* (versión 2.1.1) [49]: *pydicom* es una librería de *Python* diseñada para trabajar con ficheros DICOM, tales como imágenes médicas, informes y objetos. Esta librería facilita la lectura de este tipo complejo de archivos y los convierte a estructuras naturales *Python*, permitiendo su fácil manipulación utilizando este lenguaje.
- *TensorFlow* (versión 2.0.0) [50]: *TensorFlow* es una librería de código abierto desarrollada por investigadores de *Google* para desarrollar herramientas de ML y DL, además permite otras funcionalidades como análisis estadístico y predictivo.
- *Keras* (versión 2.2.4) [51]: Keras es una librería de código abierto especializada en la programación de todo tipo de redes neuronales artificiales.
- *Seaborn* (versión 0.11.1) [35]: Seaborn es una librería de visualización de datos basada en *matplotlib*. Con respecto a esta, implementa clases pre-confeccionadas que permiten la fácil representación de realidades complejas.
- *Statsmodels* (versión 0.12.2) [41]: librería que implementa funciones para análisis estadístico con amplia experiencia de uso.
- *SciPy* (versión 1.5.4) [42]: *SciPy* es otra librería que cuenta con múltiples funciones para análisis estadístico complejo.

Por su parte, los principales paquetes de *R* utilizados fueron:

- *cvAUC* (versión 1.1.0) [44]: paquete diseñado para estimar el valor de las AUC obtenidas por métodos de validación cruzada.
- *pROC* (versión 3.6) [47]: Paquete que permite confeccionar, visualizar, manipular y comparar las curvas de ROC.

Como plataforma de programación se utilizaron cuadernos *Jupyter* [52] que corrieron en una computadora del Grupo de Computación Avanzada y e-Ciencia del Instituto de Física de Cantabria dotada de dos Unidades de Procesamiento Gráfico (GPUs) *Tesla* V100-PCIE-32 *Gigabytes* (GB).

4.12. Experimentos

Se realizaron tres experimentos para acometer la consecución de los objetivos de este trabajo. Estos tres experimentos se detallan a continuación:

- Experimento 1: se centro en estudiar el grado de influencia de los factores propuestos en el Apartado 2 sobre el rendimiento y el aprendizaje de un algoritmo de DL en validez interna.
- Experimento 2: analizó el grado de influencia de los factores propuestos en el Apartado 2 sobre el rendimiento y aprendizaje de un algoritmo de DL en validez externa o generalización.
- Experimento 3: trató de esclarecer si los factores objeto de estudio afectaban de forma significativa a los valores de las *features* extraídas por una red neuronal convolucional.

4.12.1. Experimento 1

Se diseñó un primer experimento con el objetivo de evaluar la influencia que tenían los factores propuestos en el Apartado 2 sobre el rendimiento en validez interna de un algoritmo de DL. En concreto, se estudió en qué medida el centro del que provenían las imágenes, el protocolo de adquisición de las imágenes, el modelo de aparato de rayos X, el tipo de función de respuesta del detector, y el procesado de imagen aplicado por el fabricante, podían afectar al rendimiento y al aprendizaje de un algoritmo de DL, en validez interna.

Con este fin, se comenzó comparando el rendimiento y los mapas de calor Grad-CAM obtenidos por los modelos Modelo-F1F2 y Modelo-F1F3 sobre el subconjunto de test T1. Dado que ambos modelos habían sido entrenados con 300 imágenes obtenidas por el mismo modelo de equipo *Fujifilm*, en el caso del Modelo-F1F2 en la Institución 1, y en el caso del Modelo-F1F3 la mitad en la Institución 1 y la mitad en la Institución 2, y con diferente protocolo de adquisición, el subconjunto de test T1 brindó resultados referentes a la validez interna del algoritmo, y cuya comparación permitió conocer la influencia que el centro del que provenían las imágenes y que el protocolo de adquisición, tenían sobre el rendimiento en validez interna y sobre la capacidad de aprender relaciones causales del algoritmo.

Posteriormente, se compararon los resultados de rendimiento y los mapas de calor obtenidos por los modelos anteriores, el Modelo-F1F2 y el Modelo-F1F3, sobre el subconjunto T1, con los resultados obtenidos por el Modelo-F1'F3'S2 sobre el mismo subconjunto de test. Este tercer modelo había sido entrenado utilizando el mismo número de imágenes que las empleadas para entrenar los otros dos modelos, pero incluyendo, además de imágenes obtenidas por el equipo *Fujifilm*, imágenes obtenidas por el equipo *General Electric*, por lo que las diferencias en el aprendizaje pudieron ser achacadas al efecto del factor modelo de equipo de rayos X, es decir, al efecto conjunto del procesado de imagen aplicado por el fabricante y del tipo de función de

respuesta del detector (Figura 4.2).

4.12.2. Experimento 2

Para este experimento, se utilizó el Modelo-F1F2, ya que era el único que había sido entrenado con imágenes adquiridas por un solo modelo de equipo, el *Fujifilm*, y provenientes de una sola institución, la Institución 1.

Este experimento consistió en la comparación del rendimiento de este modelo en validez interna, calculado utilizando el subconjunto de test T1, con el rendimiento en generalización sobre diferentes subconjuntos de test:

- Subconjunto T2: contenía imágenes adquiridas por el mismo modelo de equipo de rayos X que el subconjunto T1, pero adquiridas con un protocolo de adquisición diferente y en otra institución. Por lo tanto, la comparación de los resultados obtenidos con el test T1 y el T2 permitió conocer el efecto de los factores asociados a la institución y del protocolo de adquisición sobre la generalización del algoritmo.
- Subconjunto T3: contenía imágenes adquiridas en la misma institución que el subconjunto T2, pero por un modelo de equipo con un procesado de imagen y un tipo de función de respuesta diferentes a las de los modelos de equipo utilizados para adquirir las imágenes de los subconjuntos T1 y T2. La diferencia de los resultados obtenidos sobre el subconjunto T2 y el T3 permitió conocer el efecto conjunto del modelo de equipo de rayos X, es decir, el efecto sumatorio del procesado de imagen y del tipo de función de respuesta.
- Subconjunto T4: contenía imágenes adquiridas por un modelo de equipo de la misma institución que los subconjuntos T2 y T3, que utilizaba el mismo tipo de función de respuesta que el equipo que adquirió las imágenes del subconjunto T2, y diferente del que adquirió las imágenes del T3. Su procesado de imagen era diferente al del resto de equipos. La comparación de los resultados obtenidos sobre los subconjuntos T2 y T4 permitió esclarecer la influencia del procesado de imagen; y la diferencia de los resultados obtenidos sobre los subconjuntos T2, T3 y T4 la del tipo de función de respuesta del detector.

Además de comparar el rendimiento, también se comparó la actuación del modelo sobre los diferentes subconjuntos de test en términos de explicabilidad, comparando los mapas de calor resultado de aplicar el algoritmo Grad-CAM.

4.12.3. Experimento 3

El tercer experimento trató de esclarecer si los factores propuestos en el Apartado 2 tenían influencia sobre el valor que tomaban las *features* extraídas por un algoritmo de DL, en concreto por una CNN pre-entrenada para *ImageNet*. Estas *features*, de cuya extracción se encarga la parte convolucional de la CNN, son los elementos finales en base a los cuales el algoritmo toma la decisión final de clasificar una imagen en una clase objetivo o en otra. Si los factores objeto de estudio tuvieran mayor influencia

sobre los valores de las *features* que la propia presencia o ausencia de patología, es decir, que las clases objetivo, una generalización del algoritmo a imágenes de una clase oculta de alto nivel no vista durante el entrenamiento podría no ser posible. Tal y como se explicó en la Subsección 4.9.3, en este trabajo se define como clase oculta de alto nivel a aquella que, teniendo mayor poder discriminativo que la propia clase objetivo, no es identificada ni controlada durante el proceso de entrenamiento.

En este caso, se aplicó un algoritmo de *clustering* jerarquizado para realizar una agrupación de las imágenes de los subconjuntos de test en base al valor de sus *features*, los cuales habían sido extraídos por la CNN. Posteriormente, se analizaron los grupos de imágenes encontrados, tratando de visualizar si las imágenes pertenecientes a cada una de las clases objetivo habían sido agrupadas en *clusters* diferentes, o si por el contrario las imágenes pertenecientes a alguna de las clases ocultas propuestas habían sido separadas del resto. Estos resultados darían información tanto de si los factores propuestos en el Apartado 2 tenían influencia sobre el valor de las *features*, como de si, de haber influencia, esta sería mayor que la de las clases objetivo, lo cual confirmaría que verdaderamente se trataba de clases ocultas de alto nivel.

Este experimento constó de cuatro ensayos. En el primero se realizó la agrupación de las imágenes de los subconjuntos de test T1, T2 y T3 en base al valor de las *features* extraídas por el Modelo-F1F2. En el segundo se realizó la agrupación de las mismas imágenes en base a las *features* extraídas por el mismo modelo, pero previamente las imágenes fueron recortadas preservando la región central, los pulmones y el mediastino, para eliminar las etiquetas metálicas y asegurarse de que estas no condicionaban, en ninguna medida, la agrupación. El tercer ensayo consistió en la agrupación de 44 imágenes aleatorias de los subconjuntos T1, T2 y T3, y de las 44 del subconjunto T4, en base a los valores de las *features* extraídas por el Modelo-F1F2. Finalmente, el cuarto ensayo agrupó las imágenes de los subconjuntos T1, T2 y T3 en base a las *features* extraídas por el Modelo-F1'F3'S2 antes y después de ser entrenado, con el objetivo de evaluar si el grado de influencia de los factores variaba tras el entrenamiento y el proceso de *fine tuning*.

5

Resultados - Results

5. Resultados - Results

5.1. Patients

This research included 874 patients, 45.08 % (394) from Institution-1 and the remaining from Institution-2. The study sample comprised of 42.56 % (372) females and 57.44 % (502) males. The median age was 62 years, while the average age was 60.21 ± 17.14 years (5-96). Population descriptive statistics for all the subsets of the dataset are summarized in Table 5.1.

Tabla 5.1: Descriptive Statistics of Population Age and Gender Distribution.

	Age					**Gender**		
Subsets of the Dataset	Median	Mean	SD	Min	Max	Female % (n)	Male % (n)	N
Fujifilm[1]	59.00	58.01	16.92	16	93	44.16 (174)	55.84 (220)	394
Subset F1	59.50	58.53	17.32	16	93	40.67 (61)	59.33 (89)	150
Subset F2	56.50	56.34	17.10	17	89	54.00 (81)	46.00 (69)	150
Subset T1	61.00	59.83	15.87	24	92	34.04 (32)	65.96 (62)	94
Fujifilm[2]	66.00	63.42	16.50	5	96	38.52 (94)	61.48 (150)	244
Subset F3	61.00	61.89	16.29	20	92	38.00 (57)	62.00 (93)	150
Subset T2	68.00	65.85	16.63	5	96	39.36 (37)	60.64 (57)	94
General Electric[2]	63.00	60.09	17.78	18	91	43.75 (84)	56.25 (108)	192
Subset S2	63.00	59.79	17.59	18	87	42.86 (42)	57.14 (56)	98
Subset T3	63.00	60.41	18.06	19	91	44.68 (42)	55.31 (52)	94
Carestream[2]	65.00	62.57	17.33	20	86	45.45 (20)	54.55 (24)	44
Subset T4	65.00	62.57	17.33	20	86	45.45 (20)	54.55 (24)	44
TOTAL	62.00	60.21	17.14	5	96	42.56 (372)	57.44 (502)	874

Subsets F1, F2, F3 and S2 were used to train the models, while subsets T1, T2, T3, and T4 were used to test the models.
[1]X-ray machine from Institution-1; [2]X-ray machine from Institution-2.

5.2. Image Acquisition Protocols

The exposure index (EI) of the Fujifilm devices from both institutions were compared to determine whether a statistically significant difference existed between the acquisition protocols of both institutions. The distribution of the exposure indexes

did not follow a normal distribution ($p < 0.001$), and a statistically significant difference was detected among the exposure indexes of the two Fujifilm devices (MWU $p = 0.001$, KS $p = 0.002$) (Figure 5.1) (Table 5.2). Therefore, the acquisition protocol of the Fujifilm from Institution-1 and the acquisition protocol of the Fujifilm from Institution-2 were different.

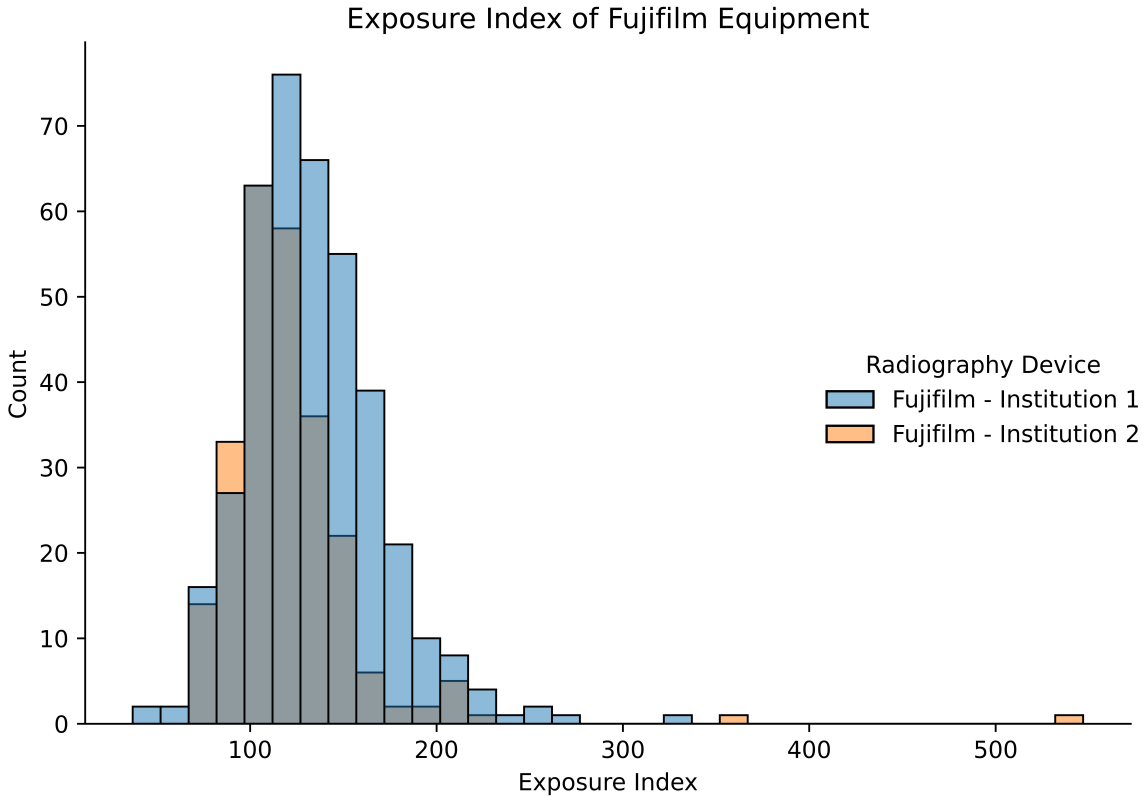


Figura 5.1: Exposure indexes of Fujifilm equipment from both institutions. The grey area represents the overlapping of the two distributions.

Tabla 5.2: P-values for the Statistical Analysis of both Fujifilm Exposure Index.

Radiography device	KSL	MWU	KS
Fujifilm from Institution-1	< 0.001	< 0.001	< 0.002
Fujifilm from Institution-2	< 0.001		

KSL = Kolmogorov-Smirnov-Lilliefors; MWU = U-Mann-Whit
-ney-Wilcoxon; KS = Kolmogorov-Smirnov for two samples.
A two-tailed *p-value* < 0.5 was considered statistically significant.
Descriptive statistics of both Fujifilm exposure indexes are in Table 4.2.

5.3. Experiments

5.3.1. Experiment 1

There were no significant differences between the internal validation performances of Model-F1F2 and Model-F1F3 (Figure 5.2) (Table 5.3 and Table 5.4). Thus, the addition of images to the training sample which were acquired in different institutions with different image protocols, did not have a significant impact on the algorithm's internal validation.

Tabla 5.3: Models' AUCs with 95 %CI*

Tests	Model-F1F2	Model-F1F3	Model-F1'F3'S2
Subset T1	0.878 (0.847, 0.909)	0.836 (0.799, 0.873)	0.785 (0.745, 0.826)
Subset T2	0.780 (0.738, 0.822)	0.728 (0.682, 0.773)	0.679 (0.631, 0.727)
Subset T3	0.544 (0.491, 0.596)	0.555 (0.502, 0.607)	0.751 (0.708, 0.795)
Subset T4	0.689 (0.618, 0.760)	0.622 (0.548, 0.700)	0.590 (0.514, 0.664)

*95 % Confidence Intervals (CI) are reported in parentheses.

Tabla 5.4: Bootstrapping Differences between the Models' AUCs with 95 %CI.

Tests	Model-F1F2 & Model-F1F3	Model-F1F2 & Model-F1'F3'S2	Model-F1F3 & Model-F1'F3'S2
T1	0.029 (−0.008, 0.068)	0.080 (0.024, 0.141)**	0.052 (0.008, 0.101)*
T2	0.040 (-0.021, 0.107)	0.096 (0.010, 0.184)*	0.055 (-0.024, 0.134)
T3	-0.008 (-0.049, 0.032)	-0.234 (-0.361, -0.095)***	-0.226 (-0.352, -0.092)***
T4	0.039 (-0.068, 0.161)	0.092 (-0.052, 0.234)	0.053 (-0.054, 0.160)

95 % Confidence Intervals (CI) are reported in parentheses.
P-value for two-tailed DeLong tests: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Grad-CAM heatmaps based on the predictions of Model-F1F2 and Model-F1F3 for images from subset T1 showed similar activation patterns among each other. The heatmaps for COVID-19 images depicted activations exclusively inside the lungs, while control images lacked activations within any region of the image (Figure 5.3). Seemingly, both Model-F1F2 and Model-F1F3 were able to learn the radiological findings of COVID-19. Therefore, both models made predictions based on causal relationships instead of spurious relationships.

By contrast, the internal validation performances of Model-F1F2 and Model-F1F3 on subset T1 were, respectively, 8 % ($p < 0.01$) and 5.2 % ($p < 0.05$) higher than the internal validation performance of Model-F1'F3'S2 on the same subset (Figure 5.2) (Table 5.4). As it was previously described, Model-F1F2 and Model-F1F3 were both

trained exclusively with 300 images from Fujifilm devices. While Model-F1'F3'S2 was also trained with 300 images, rather than only including images from the Fujifilm devices, Model-F1'F3'S2 incorporated images from the General Electric device in the training as well.

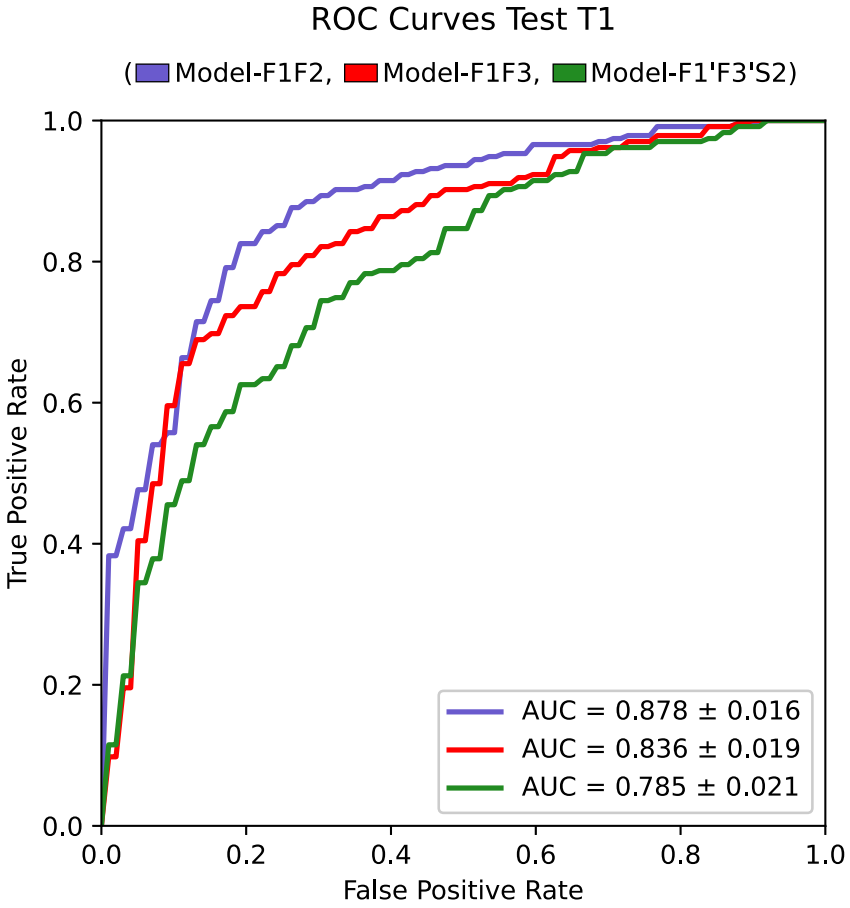


Figura 5.2: Models' ROC for subset T1. Model-F1F2 and Model-F1F3 were trained with images acquired by only one model of X-ray machine, while Model-F1'F3'S2 was trained with images acquired by two different X-ray device models. Model-F1F2 and Model-F1F3 achieved a better internal validation performance than Model-F1'F3'S2. This performance was evaluated using subset T1, which contained 94 images acquired by the Fujifilm device from Institution-1.

In addition, Grad-CAM heatmaps obtained from the predictions of Model-F1'F3'S2 showed more activation areas on both COVID-19 and control images than those observed for Model-F1F2 and Model-F1F3, many of which were located outside the lungs and did not have clinical or radiological meaning (Figure 5.3). The addition of images acquired by multiple models of X-ray device to the training sample decreased the algorithm's internal validation performance and led the algorithm to

learn spurious relationships (confounding factors) instead of causal relationships.

5.3.2. Experiment 2

Experiment 2 results showed that Model-F1F2, which was trained with images from only one institution and acquired by only one X-ray device model (Fujifilm - Institution 1), managed to generalize across hospitals. This being said, Model-F1F2 suffered a variable decrease in performance when having to perform on images from external institutions or acquired by external X-ray device models (Figure 5.4). Particularly, Model-F1F2 generalized to Fujifilm images from Institution-2 (subset T2) with a loss in the AUC of 9.8 % ($p = 0.06$), and to Carestream images from Institution-2 (subset T4) with a loss in the AUC of 18.9 % ($p < 0.05$) (Table 5.5). Conversely, Model-F1F2 did not generalize to General Electric images from Institution-2 (subset T3), as it showed a loss in the AUC of 33.5 % ($p < 0.001$) which led the model to perform randomly (Table 5.3 and Table 5.5). Ultimately, Model-F1F2 generalized across institutions and across different makes and models of X-ray device with the same type of response function, however, not across X-ray devices with a different type of response function.

Tabla 5.5: The Influence of Institutional and X-ray Device Related Factors on the Generalization Performance of Model-F1F2.

Tests	Performance difference	Factors that might hinder generalization
Subset T1 - Subset T2	0.098 (0.045, 0.154)$^{+}$	Institution
Subset T1 - Subset T3	0.335 (0.291, 0.368)***	Institution + DIP + DRF
Subset T1 - Subset T4	0.189 (0.141, 0.236) *	Institution + DIP
Subset T2 - Subset T3	0.237 (0.192, 0.266)***	DIP + DRF
Subset T2 - Subset T4	0.091 (0.042, 0.135)	DIP
Subset T3 - Subset T4	-0.146 (-0.171, -0.111)*	DRF

DIP = X-ray device's image processing, DRF = X-ray device's type of response function.
95 % Confidence Intervals are reported in parentheses.
$P - values$ for two-tailed DeLong tests: $^{+}p = 0.06$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Additionally, Grad-CAM heatmaps obtained from Model-F1F2 predictions for test images acquired by the Fujifilm device from Institution-1 (subset T1) and by the Fujifilm device from Institution-2 (subset T2) were similar to each other. This being said, Grad-CAM heatmaps based on predictions for test images from the Carestream device (subset T4) showed a significant decrease in the model's sensitivity to detect the radiological findings of COVID-19. Considering that the Carestream and the Fujifilm devices had the same type of response function, the decrease in the model's

sensitivity to detect COVID-19 could be secondary to differences in the image processing applied by the X-ray machine. Finally, Grad-CAM heatmaps obtained from predictions for test images from the General Electric device (subset T3) did not show any activation areas, probably due to the fact that this device had a different type of response function than the Fujifilm and Carestream devices'.

5.3.3. Experiment 3

The hierarchical clustering algorithm grouped images from subsets T1 (Fujifilm - Institution-1), T2 (Fujifilm - Institution-2), and T3 (General Electric - Institution-2) into two well-defined clusters based on the features extracted by Model-F1F2. These two clusters corresponded to the Fujifilm and General Electric device models which acquired the images from the subsets (Figura 9.1). The algorithm did not separate the images from the two Fujifilm devices, despite these having a different acquisition protocol and belonging to different institutions. In other words, the clustering algorithm successfully separated images from different makes and models of X-ray device. Additionally, images belonging to the different target classes (COVID-19 and control) were not separated.

The same result was also observed when clustering was run with test images excluding metallic tokens (Figura 9.2). Moreover, the addition of images acquired by a third model of X-ray machine, subset T4 (Carestream - Institution 2), resulted in the grouping of images from the three X-ray device models into separated clusters, mixing images from the two Fujifilm devices (Figure 5.5).

The clustering algorithm also managed to group images from subsets T1, T2, T3, and T4 into two clusters. These clusters coincided with the different types of response functions used by the X-ray devices which acquired the subset images (Figure 5.5). Therefore, images acquired by X-ray devices with the same type of response function had feature values which were more similar among each other, than feature values from images acquired by X-ray devices with different type of response function.

Finally, the clustering algorithm was also run with features extracted from subsets T1, T2 and T3 by Model-F1'F3'S2 before and after being trained. The algorithm found evident clusters of images that corresponded with images acquired by each model of X-ray machine. This result was observed both when image features were extracted by the pre-trained version of Model-F1'F3'S2 and when features were extracted by the trained version of Model-F1'F3'S2 (Figura 9.3). Thus, fine-tuning did not overcome feature values differences found among images acquired by different X-ray device models.

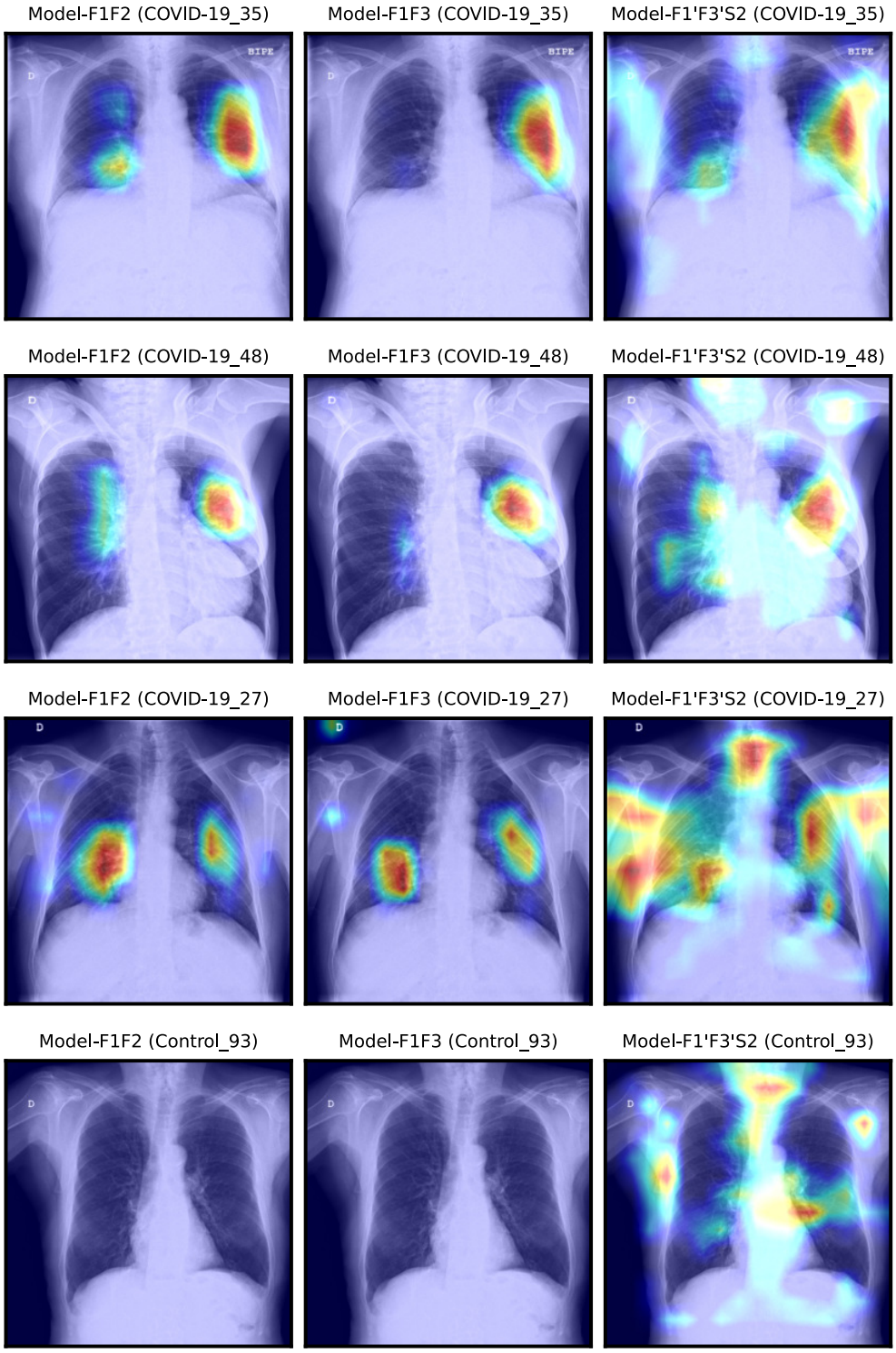


Figura 5.3: Example of five radiographs with their predicted Grad-CAM heatmaps generated by the three DL models. The addition of images acquired in different institutions with different image protocols, did not have a significant impact on the algorithm's internal validation. On the other hand, the addition of images acquired by multiple X-ray device models to the training sample led the algorithm to learn spurious relationships instead of causal relationships.

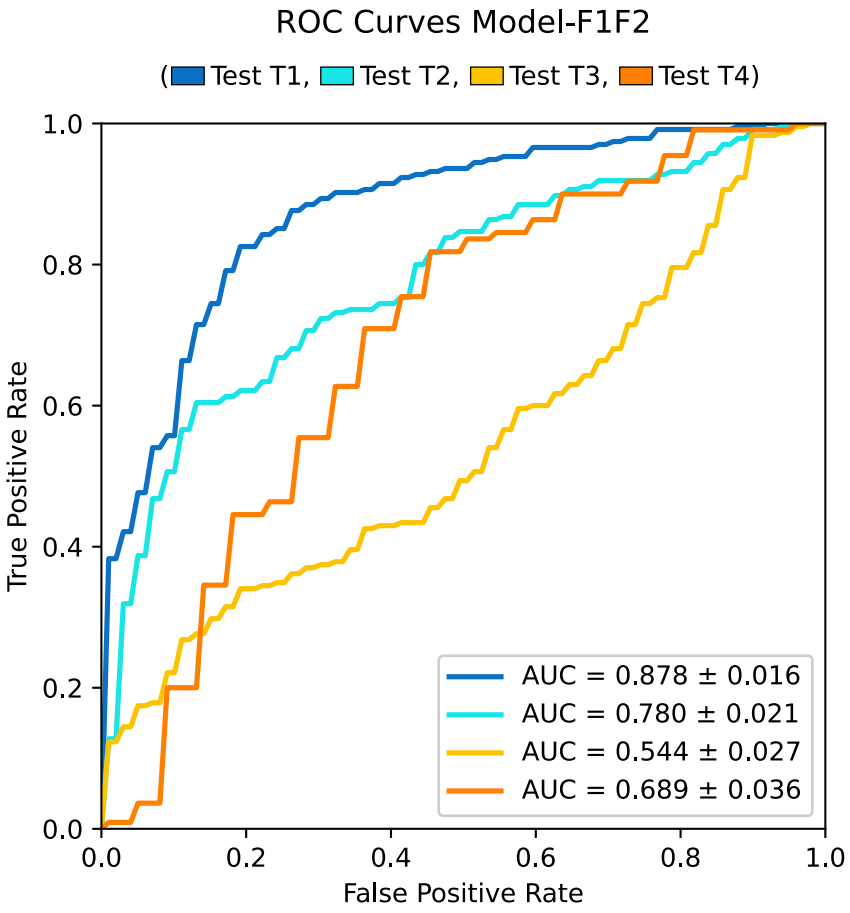


Figura 5.4: Model-F1F2 ROC curves for the different tests. Model-F1F2 was trained with images acquired by the Fujifilm device from Institution-1. This model generalized to images acquired by the Fujifilm device from Institution-2 (subtest T2), and also to images acquired by the Carestream device from Institution-2 (subtest T4). By contrast, Model-F1F2 did not generalize to images acquired by the General Electric device from Institution-2 (subtest T3). Subset T1 contained images from the Fujifilm device from Institution-1.

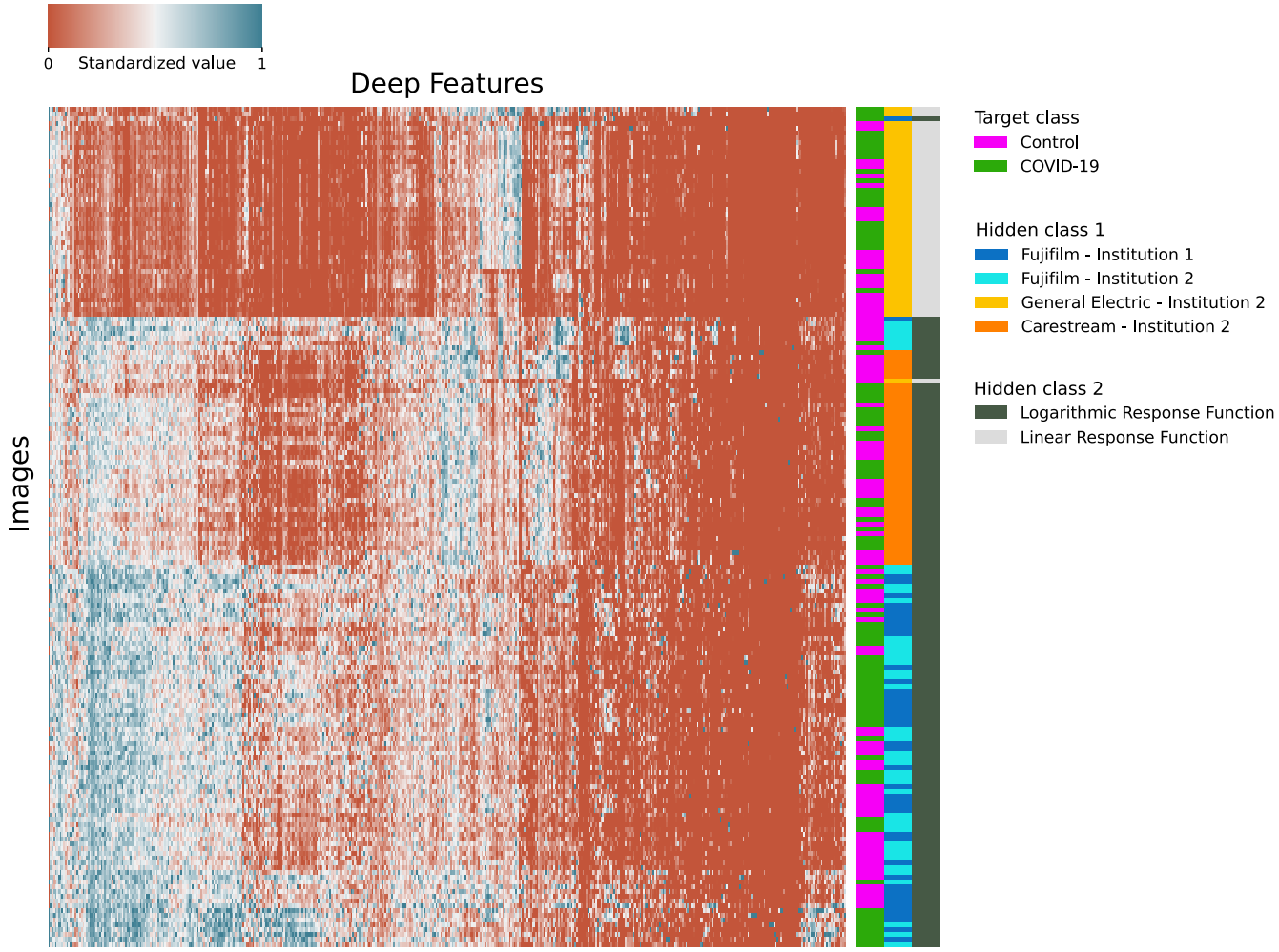


Figura 5.5: Clusterization of images from subsets T1, T2, T3, and T4 based on Model-F1F2 features. Images clusterization from subsets T1, T2, T3, and T4 using a hierarchical clustering algorithm based on Model-F1F2 features resulted in three clusters. These three clusters corresponded with images from each of the three models of X-ray machine that acquired the images. The radiographs from the two Fujifilm machines were grouped together, despite being from different institutions. Additionally, images were also clustered into two groups based on the type of response function that the X-ray machines which acquired the images had. Finally, images belonging to the different target classes (COVID-19 and control) were not separated.

6

Discusión - Discussion

6. Discusión - Discussion

6.1. Experiment 1 – Factors Influence on the Internal Validation

Through this investigation, we observed that device related factors significantly affect DL algorithms' internal validation performance. Specifically, algorithms' internal validation performance decreases as more X-ray machine models acquire the training images. According to this, Model-F1F2 and Model-F1F3, which were trained with images from only one make and model of X-ray machine, achieved a better performance in internal validation than Model-F1'F3'S2, which was trained with images from two different make and models of X-ray device.

Taking this into account, the decrease in the algorithms' performance is not the only consequence encountered when training with images from multiple device models. Grad-CAM heatmaps showed that as more models of X-ray machine acquired the training images, more textural and non-radiological activation areas appeared in the heatmaps. Particularly, Model-F1F2 and Model-F1F3 showed similar activation patterns in the Grad-CAM heatmaps, which included activation areas inside the lungs exclusively for COVID-19 patients, and absence of activation areas for control patients. By contrast, Model-F1'F3'S2 showed several activation areas without radiological meaning, including activation areas outside the lungs in COVID-19 patients, and activation areas inside the lungs in control patients. In other words, Model-F1F2 and Model-F1F3 seemed to predict based on the detection of the radiological findings of COVID-19 rather than based on other image findings without radiological meaning, as Model-F1'F3'S2 did. This issue is of vital importance when discussing the generalization power of DL algorithms. To build robust DL models that can generalize, a key requirement is to be sure that the algorithm predicts based on causal relationships [53], such as pathological radiographic findings, rather than based on spurious relationships, unrelated to the disease.

Unlike device related factors, institutional related factors and the image acquisition protocols do not significantly affect algorithms' internal validation. Model-F1F2 and Model-F1F3 did not show differences in performance, despite Model-F1F2 being trained exclusively with images from Fujifilm Institution 1 and Model-F1F3 being trained with images from both Fujifilm devices (Institution 1 and Institution 2). Grad-CAM heatmaps were very similar for both DL models.

To our knowledge, this investigation is the first one of its kind which studies the effects of training a DL algorithm for medical image classification with chest radiographs from multiple institutions and acquired by different device models on the algorithm's internal validation performance.

In light of the results accomplished in Experiment 1, we propose an alternative training technique to develop DL algorithms for medical image classification. This

technique consists in training different algorithms using, for each one, images acquired by only one X-ray device model. It is important to highlight that within this technique, it does not matter if images are acquired by different devices or in different institutions, as long as all devices are the same model of X-ray machine. Through this technique, it is possible to achieve high performance algorithms that are able to learn causal relationships with a smaller training sample.

6.2. Experiment 2 - Factors Influence on the Generalization

This research attempts to quantified the influence that institutional and device related factors might have on DL algorithms' generalization for radiography classification. Our results found that DL algorithms can generalize across institutions and X-ray devices with the same type of response function. However, this being said, algorithms may suffer a variable decrease in their performance when deployed on external datasets. On the other hand, generalization across X-ray devices with a different type of response function was not observed.

When it comes to institutional related factors, Model-F1F2 was able to generalize across institutions with a decrease in its performance of 9.8 %. This measure was computed as the difference between the model's performance on subset T1 and subset T2. Both subsets contained 94 images acquired by the same make and model of X-ray device (a Fujifilm FDR Smart FGX), however, while images from subset T1 were acquired in Institution-1, images from T2 were acquired in Institution-2, which used a different image acquisition protocol. Therefore, the 9.8 % reduction in the generalization performance was attributed to institutional related factors. Experiment 2 results suggest that institutional related factors do not impede the generalization of DL algorithms, however, they can decrease the algorithm's performance.

The acquisition protocol probably had a non-significant influence on the model's generalization performance, as clustering algorithms in Experiment 3 did not separate images from both Fujifilm devices into different clusters, despite having different acquisition protocols.

Model-F1F2 also generalized across X-ray devices with the same type of response function. Specifically, the performance of Model-F1F2 on subset T2 (images from Fujifilm - Institution 2) was 9.1 % higher than the performance of Model-F1F2 on subset T4 (images from Carestream - Institution 2). Images from subsets T2 and T4 were all acquired in Institution 2 by two devices which both had a logarithm response function. Thus, the main difference between the images from subsets T2 and T4 was the image processing applied by the X-ray device. Therefore, we assumed that differences in the image processing applied by the X-ray machine reduced the algorithm's generalization performance by 9.1 %. In other words, Model-F1F2 generalized across X-ray devices with the same type of response function with a loss in its performance of 9.1 %. By contrast, Model-F1F2 performed randomly on subset T3 (images from

General Electric – Institution 2), which contained images acquired in Institution 2 by a device which had a linear response function. This result indicates that the algorithm was not able to generalize across devices with a different type of response function.

The previously mentioned results were also supported by Grad-CAM heatmaps. Heatmaps based on the predictions of Model-F1F2 for subset T1 images showed activation areas in locations where radiological findings of COVID-19 could be found. Although heatmaps based on predictions for subset T2 were similar, the model slightly reduced its sensitivity to detect lung opacities in the images from this subset. A higher reduction in the model's sensitivity was observed for heatmaps based on predictions for subset T4, however, these heatmaps still showed activation areas on several lung opacities. By contrast, heatmaps based on predictions for subset T3 images did not show any activation areas, as the algorithm did not generalize to this subset.

The generalization of DL algorithms for radiography classification to external datasets has been argued by a handfulf of uthors. Pooch et al. [18] conclude that state-of-the-art DL algorithms do not generalize to external data which have differences with the training data. Similar to this, Zech et al. [12] and Sathitratanacheewin et al. [16] defend that CNNs do not generalize to external sites. Additionally, Zech et al. [12] and Maguolo et al. [17] warn that neural networks can often distinguish the dataset or the hospital where the images come from. For Maguolo et al. [17], this issue is very important since most papers obtain images of each class to predict from different datasets. Trying to understand how CNNs distinguish the source of the dataset, Cohen et al. [21] propose discrepancies in image labeling criteria among medical centers to be the potentially cause. On the other hand, for Rajpurkar et al. [20] and Pan et al. [15] DL algorithms for radiography classification can generalize to datasets from external institutions with a decrease in their performance.

Our research can shed light on the controversy surrounding DL algorithms' generalization, as this work separately analyzed the influence of multiple factors on this issue. We observed that the X-ray device's response function is probably the most important factor for DL algorithms' generalization, as it can impede it. The second most relevant factor might be the device's image processing. Although this factor does not impede the algorithm to generalize, it may significantly decrease its performance. Finally, institutional related factors and the image acquisition protocol can also reduce the algorithm's performance, however, probably less than the device's image processing. Taking all this into account, we propose a hierarchy of factors that might affect the generalization of DL algorithms for medical image classification. This hierarchy is shown in Figure 6.1.

In summary, although institutional and device related factors may reduce algorithm's generalization, DL algorithms can generalize across institutions and X-ray devices with the same type of response function. By contrast, these algorithms might not generalize across devices with a different type of response function.

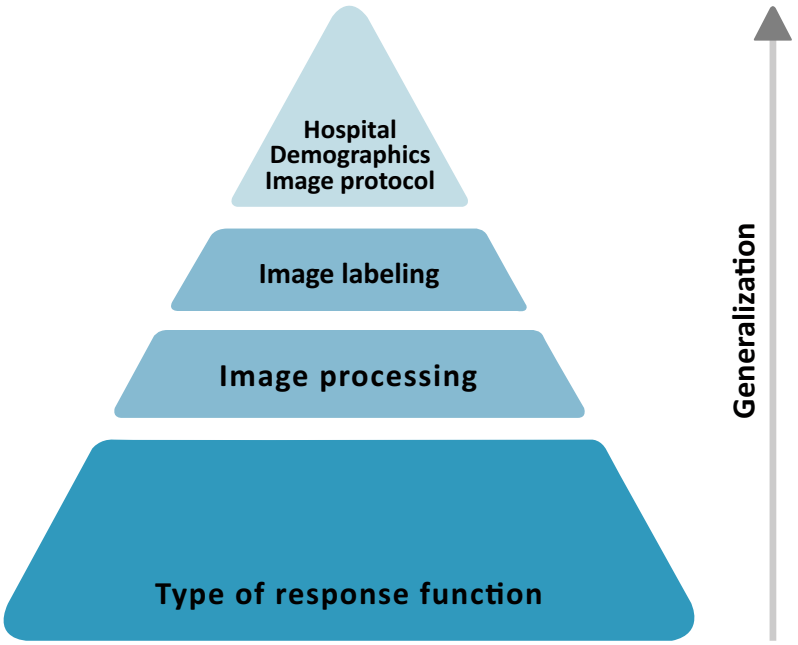


Figura 6.1: Hierarchy of factors affecting the generalization of Deep Learning algorithms for medical image classification. This pyramid shows the hierarchy of factors affecting the generalization of DL algorithms for medical image classification. The type of response function of the radiography device is the most relevant factor since it is the only one that impedes generalization. The second most important factor is the image processing applied by the X-ray machine. The image processing hinders generalization, however, it does not impede it. Finally, institutional related factors, including the image labeling, the institution, and population demographics, have a surmountable influence on algorithm's generalization.

6.3. Experiment 3 – Factors Influence on CNN Feature Values

The unsupervised clustering algorithm demonstrated that feature values extracted by CNNs from images might be highly dependent on the model of X-ray device that acquires the images. This happens since each model of X-ray device applies a characteristic image processing and has a specific response function. Therefore, radiological images have different textures depending on the model of X-ray device that acquires the images. This leads to disparities in CNN-feature values among images from different X-ray devices that could hinder generalization.

The clustering algorithm in Experiment 3 separated images from each model of X-ray device into different clusters. Images acquired by devices with a logarithmic response function and images acquired by devices with a linear response function were

also separated from each other into two different clusters. Additionally, the algorithm did not separate images from the two Fujifilm devices into separated clusters, despite them belonging to different institutions and having different acquisition protocols. Therefore, the model of X-ray device and ultimately, its image processing and its type of response function, had an important influence on CNN-feature values. In contrast, the institution and the acquisition protocol did not significantly affect CNN-feature values.

Similar results were observed when metallic tokens were removed, thus confirming their lack of effect on the clusterization process. Clusterization was also not affected after fine-tunning was applied.

On the other hand, the algorithm did not clearly separate the target classes (COVID-19 and control) into different clusters. Hence, following our terminology, the equipment, its image processing, and its type of response function were considered high-level hidden classes. We use this term to refer to imaging categories in the dataset which are even more evident for CNNs than the target class. In other words, in the eyes of the VGG16, images from different models of X-ray device were more different from each other than images from COVID-19 and control classes were from each other.

Experiment 3 results suggest that the model of X-ray device has a strong influence on CNN-feature values, which can be even higher than the influence of target classes when faced with complex problems, such as the classification of COVID-19 radiographs. The feature values susceptibility to the X-ray machine, indicates that features extracted by VGG16 and probably by other ImageNet CNNs, are mainly based on textures instead of shapes. This issue is extremely relevant for DL algorithms' generalization, as features based on shapes are potentially more robust and invariant than features based on textures. Consequently, neural networks able to predict based on shape features rather than on textural features could potentially solve the generalization issue.

Outside of the medical field, Geirhos et al. [14] have proposed that ImageNet-trained CNNs are biased towards recognizing textures rather than shapes. These authors also suggest that shape biased networks are inherently more robust than texture biased networks [14], something that our research also puts forward in the medical field.

According to us, our investigation is the first which explores the effect of X-ray devices on the feature values extracted by ImageNet-trained CNNs. We conclude that features extracted by ImageNet-trained CNNs from radiographs are mainly based on textures that depend on the X-ray device which acquired the images. Therefore, Transfer Learning from ImageNet-trained CNNs might not be the best solution to achieve robust DL algorithms for medical image classification, when training with small or medium datasets. However, an exception to this rule might be tasks where geometric differences among classes are particularly evident, like the classification of anatomical areas.

Moreover, we introduce the use of hierarchical clustering as a useful technique

to graphically visualize the fine-tunning process of CNNs, and to detect high-level hidden classes in datasets. In our opinion, finding these hidden classes could be advantageous, since rather than training only one algorithm using all training images, it is possible to train different algorithms using for each one, images from only one hidden class. This method could achieve high performing algorithms with a smaller training sample. Additionally, this technique could also facilitate the algorithm to learn causal relationships instead of spurious relationships, thus leading to more reliable Grad-CAM heatmaps, as Experiment 1 showed.

7

Conclusiones - Conclusions

7. Conclusiones - Conclusions

7.1. Conclusiones (en castellano)

Las conclusiones de este trabajo se detallan a continuación:

1. Las CNNs pueden llevar a cabo la tarea de diagnóstico de COVID-19 en radiografías de tórax con un alto nivel de precisión, como apuntaba la literatura previa a este trabajo.
2. Los factores que afectan a la validez interna y a la generalización de los algoritmos de DL aplicados a imagen médica pueden ser clasificados en dos grupos: factores asociados al centro del que provienen las imágenes de entrenamiento, los cuales no afectan al valor de los píxeles de las imágenes; y factores asociados al equipo de adquisición de las imágenes, los cuales sí afectan al valor de los píxeles de las mismas.
3. Los factores asociados con el centro del que provienen las imágenes de entrenamiento probablemente no afecten a la validez interna de los algoritmos de DL, sin embargo, son capaces de afectar a su capacidad de generalización, reduciendo el rendimiento de los algoritmos cuando se utilizan en centros externos.
4. Los factores asociados con el equipo de adquisición de las imágenes son capaces de afectar de forma significativa al rendimiento de los algoritmos de DL, tanto en validez interna, como en generalización.
5. Diferencias en el tipo de función de respuesta del detector de los aparatos de rayos X pueden impedir la generalización de los algoritmos de DL, mientras que diferencias en otros factores asociados al equipo de adquisición pueden afectar pero no impedir la generalización.
6. Las texturas de una imagen radiográfica dependen del modelo de equipo de rayos X que haya adquirido dicha imagen.
7. Las CNNs entrenadas para *ImageNet* probablemente estén sesgadas a la identificación de texturas en lugar de a la identificación de formas.
8. Los valores de las *features* extraídas por las CNNs dependen del modelo de equipo de rayos X que haya adquirido la imagen.
9. El modelo de equipo de rayos X que adquiere las imágenes de entrenamiento y de test es un factor crucial en el rendimiento final de un algoritmo de DL, siendo probablemente, una de las causas principales que subyacen a la deficiencia de generalización que presentan, con frecuencia, los algoritmos de DL para análisis de imagen médica.
10. Entrenar un algoritmo de DL diferente para cada clase oculta de alto nivel en lugar de un solo algoritmo para todo el conjunto de entrenamiento, podría lograr modelos de DL que predijesen en base a relaciones causales y que lograsen altos niveles de precisión, utilizando un conjunto de entrenamiento de menor tamaño.
11. Entrenar un algoritmo de DL diferente para cada clase oculta de alto nivel podría

ser una solución a la deficiencia de generalización que muestran, a menudo, los algoritmos de DL para imagen médica.

12. Redes neuronales que extrajesen *features* basadas en formas en vez de *features* basadas en texturas, podrían constituir una solución a la deficiencia de generalización que muestran los algoritmos de DL para imagen médica utilizados actualmente.

7.2. Conclusions (en inglés)

The conclusions of this work are listed below:

1. Classification of COVID-19 chest radiographs can be accurately solved by CNNs, as previous literature already supports.
2. Factors that affect the internal validation and the generalization of DL algorithms for medical image analysis can be classified into two groups: institutional related factors, which are those which do not modify the image pixel values; and device related factors, which are those which modify the image pixel values.
3. Institutional related factors do not affect the DL algorithm's internal validation, however, they decreased DL algorithm's generalization performance when deployed to an external institution.
4. Device related factors significantly reduce both DL algorithm's internal validation and generalization performances.
5. The type of response function of the radiography device may impede the generalization of DL algorithms, while other device related factors hindered, but not impede, the algorithm's generalization.
6. Image textures are different for each model of X-machine.
7. ImageNet-trained CNNs are probably biased towards identifying textures rather than shapes.
8. Feature values extracted by ImageNet-trained CNNs depend on the X-ray machine which acquires the image.
9. The model of X-ray machine which acquires the training and test images is a potentially crucial factor in the performance of DL algorithms, and it is maybe one of the primary causes behind DL algorithms' generalization deficiency for radiography classification.
10. Training a different algorithm for each high-level hidden class can lead to DL models which predicts based on causal relationships, and that achieve high performance with a small training sample.
11. Training a different algorithm for each high-level hidden class can be one solution to the DL algorithms' lack of generalization.
12. Neural networks able to predict based on shape features rather than on textural features can be a possible solution for the DL algorithms' lack of generalization.

8

Bibliografía

8. Bibliografía

[1] A. Bohr y K. Memarzadeh, «The rise of artificial intelligence in healthcare applications,» *Artificial Intelligence in Healthcare*, A. Bohr y K. Memarzadeh, eds., págs. 25-60, 2020. DOI: 10.1016/B978-0-12-818438-7.00002-2. dirección: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7325854/.

[2] E. Neri, N. de Souza, A. Brady y col., «What the radiologist should know about artificial intelligence –an ESR white paper,» *Insights into Imaging*, vol. 10, n.º 1, pág. 44, 2019. DOI: 10.1186/s13244-019-0738-2. dirección: https://doi.org/10.1186/s13244-019-0738-2.

[3] F. Chollet, *Deep Learning with Python*. Manning, nov. de 2017, ISBN: 9781617294433.

[4] A. Géron, *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems*. Sebastopol, CA: O'Reilly Media, 2017, ISBN: 978-1491962299.

[5] H. Yu, L. T. Yang, Q. Zhang, D. Armstrong y M. J. Deen, «Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives,» *Neurocomputing*, vol. 444, págs. 92-110, 2021, ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2020.04.157. dirección: https://www.sciencedirect.com/science/article/pii/S0925231221001314.

[6] R. Smith-Bindman, M. L. Kwan, E. C. Marlow y col., «Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000-2016,» *JAMA*, vol. 322, n.º 9, págs. 843-856, mayo de 2019. DOI: 10.1001/jama.2019.11456. dirección: https://doi.org/10.1001/jama.2019.11456.

[7] R. J. M. Bruls y R. M. Kwee, «Workload for radiologists during on-call hours: dramatic increase in the past 15 years,» *Insights into Imaging*, vol. 11, n.º 1, pág. 121, 2020. DOI: 10.1186/s13244-020-00925-z. dirección: https://doi.org/10.1186/s13244-020-00925-z.

[8] N. Zhu, D. Zhang, W. Wang y col., «A Novel Coronavirus from Patients with Pneumonia in China, 2019,» *New England Journal of Medicine*, vol. 382, n.º 8, págs. 727-733, 2020.

[9] A. Borghesi y M. Roberto, «COVID-19 outbreak in Italy: experimental chest X-ray scoring system for quantifying and monitoring disease progression,» *La radiologia medica*, vol. 125, págs. 509-513, 2020.

[10] A. L. Chetlen, T. L. Chan, D. H. Ballard y col., «Addressing Burnout in Radiologists,» *Academic Radiology*, vol. 26, n.º 4, págs. 526-533, 2019, ISSN: 1076-6332. DOI: https://doi.org/10.1016/j.acra.2018.07.001. dirección: https://www.sciencedirect.com/science/article/pii/S1076633218303416.

[11] E. J. Hwang, J. G. Nam, W. H. Lim y col., «Deep Learning for Chest Radiograph Diagnosis in the Emergency Department,» *Radiology*, vol. 293, n.º 3, págs. 573-580, 2019.

[12] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano y E. K. Oermann, «Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study,» *PLOS Medicine*, vol. 15, págs. 1-17, nov. de 2018.

[13] A. Subbaswamy y S. Saria, «Counterfactual Normalization: Proactively Addressing Dataset Shift and Improving Reliability Using Causal Mechanisms,» en *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, R. Silva, A. Globerson y A. Globerson, eds., vol. 2, Association For Uncertainty in Artificial Intelligence (AUAI), ene. de 2018, págs. 947-957.

[14] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann y W. Brendel, «ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.,» en *International Conference on Learning Representations*, 2019. dirección: `https://openreview.net/forum?id=Bygh9j09KX`.

[15] I. Pan, S. Agarwal y D. Merck, «Generalizable Inter-Institutional Classification of Abnormal Chest Radiographs Using Efficient Convolutional Neural Networks,» *Journal of Digital Imaging*, vol. 32, n.º 5, págs. 888-896, 2019.

[16] S. Sathitratanacheewin, P. Sunanta y K. Pongpirul, «Deep learning for automated classification of tuberculosis-related chest X-Ray: dataset distribution shift limits diagnostic performance generalizability,» *Heliyon*, vol. 6, n.º 8, e04614, 2020.

[17] G. Maguolo y L. Nanni, «A critic evaluation of methods for COVID-19 automatic detection from X-ray images,» *Information Fusion*, vol. 76, págs. 1-7, 2021.

[18] E. H. P. Pooch, P. Ballester y R. C. Barro, «Can We Trust Deep Learning Based Diagnosis? The Impact of Domain Shift in Chest Radiograph Classification,» en *Thoracic Image Analysis*, Springer International Publishing, 2020, págs. 74-83.

[19] K. Kurzer, P. Schörner, A. Albers, H. Thomsen, K. Daaboul y J. Zöllner, *Generalizing Decision Making for Automated Driving with an Invariant Environment Representation using Deep Reinforcement Learning*, feb. de 2021.

[20] P. Rajpurkar, A. Joshi, A. Pareek y col., *CheXpedition: Investigating Generalization Challenges for Translation of Chest X-Ray Algorithms to the Clinical Setting*, 2020. arXiv: `2002.11379 [eess.IV]`.

[21] J. P. Cohen, M. Hashir, R. Brooks y H. Bertrand, «On the limits of cross-domain generalization in automated X-ray prediction,» en *Medical Imaging with Deep Learning*, 2020. dirección: `https://openreview.net/forum?id=VB2MOuOKyq`.

[22] L. Lanca y A. Silva, *Digital Imaging Systems for Plain Radiography*. Springer-Verlag New York, 2013.

[23] KCARE, «Quantitative evaluation of digital detectors for general radiography,» KCARE, inf. téc. 05078, 2005.

[24] G. Wolberg e I. Alfy, «Monotonic Cubic Spline Interpolation,» ép. CGI '99, USA: IEEE Computer Society, 1999, pág. 188, ISBN: 0769501850.

[25] M. L. Butler, L. Rainford, J. Last y P. C. Brennan, «Are exposure index values consistent in clinical practice? A multi-manufacturer investigation,» *Radiation Protection Dosimetry*, vol. 139, n.º 1-3, págs. 371-374, mar. de 2010.

[26] J. A. Seibert y R. L. Morin, «The standardized exposure index for digital radiography: an opportunity for optimization of radiation dose to the pediatric population,» *Pediatric Radiology*, vol. 41, n.º 5, págs. 573-581, 2011.

[27] A. K. Das, S. Kalam, C. Kumar y D. Sinha, «TLCoV- An automated Covid-19 screening model using Transfer Learning from chest X-ray images,» *Chaos Solitons Fractals*, vol. 144, pág. 110 713, mar. de 2021.

[28] A. Shelke, M. Inamdar, V. Shah y col., «Chest X-ray Classification Using Deep Learning for Automated COVID-19 Screening,» *SN computer science*, vol. 2, n.º 4, pág. 300, 2021.

[29] C. Sitaula y M. B. Hossain, «Attention-based VGG-16 model for COVID-19 chest X-ray image classification,» *Applied Intelligence*, vol. 17, págs. 1-14, nov. de 2020.

[30] X. Glorot e Y. Bengio, «Understanding the difficulty of training deep feedforward neural networks,» en *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Y. W. Teh y M. Titterington, eds., ép. Proceedings of Machine Learning Research, vol. 9, Chia Laguna Resort, Sardinia, Italy: PMLR, mayo de 2010, págs. 249-256. dirección: `https://proceedings.mlr.press/v9/glorot10a.html`.

[31] S. H. Park, J. M. Goo y C.-H. Jo, «Receiver operating characteristic (ROC) curve: practical review for radiologists,» *Korean journal of radiology*, vol. 5, n.º 1, págs. 11-18, ene. de 2004. DOI: `10.3348/kjr.2004.5.1.11`.

[32] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh y D. Batra, «Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization,» en *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, oct. de 2017.

[33] G. MJ y S. Wilson, «Introduction to hierarchical clustering.,» *Journal of clinical neurophysiology: official publication of the American Electroencephalographic Society*, vol. 19, n.º 2, págs. 144-151, abr. de 2002.

[34] G. Van Rossum y F. L. Drake Jr, *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[35] M. Waskom, O. Botvinnik, D. O'Kane y col., *mwaskom/seaborn: v0.8.1 (September 2017)*, ver. v0.8.1, sep. de 2017. DOI: `10.5281/zenodo.883859`. dirección: `https://doi.org/10.5281/zenodo.883859`.

[36] G. E. Dallal y L. Wilkinson, «An Analytic Approximation to the Distribution of Lilliefors's Test Statistic for Normality,» *The American Statistician*, vol. 40, n.º 4, págs. 294-296, 1986.

[37] H. W. Lilliefors, «On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown,» *Journal of the American Statistical Association*, vol. 62, n.º 318, págs. 399-402, 1967.

[38] Y. Dodge, «The Concise Encyclopedia of Statistics,» en New York, NY: Springer New York, 2008, cap. Kolmogorov-Smirnov Test, págs. 283-287.

[39] H. Mann y D. Whitney, «On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other.,» *Annals of Mathematical Statistics*, vol. 18, n.º 1, págs. 50-60, 1947.

[40] F. Wilcoxon, «Individual Comparisons by Ranking Methods,» *Biometrics Bulletin*, vol. 1, n.º 6, págs. 80-83, 1945.

[41] S. Seabold y J. Perktold, «statsmodels: Econometric and statistical modeling with python,» en *9th Python in Science Conference*, 2010.

[42] P. Virtanen, R. Gommers, T. E. Oliphant y col., «SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python,» *Nature Methods*, vol. 17, págs. 261-272, 2020.

[43] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020.

[44] E. LeDell, M. Petersen y M. van der Laan, *cvAUC: Cross-Validated Area Under the ROC Curve Confidence Intervals*, dic. de 2014.

[45] B. Efron y R. Tibshirani, «Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy,» *Statistical Science*, vol. 1, n.º 1, págs. 54-75, 1986.

[46] E. R. DeLong, D. M. DeLong y D. L. Clarke-Pearson, «Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach,» *Biometrics*, vol. 44, n.º 3, págs. 837-845, sep. de 1988.

[47] X. Robin, N. TurcK, A. HainarD y col., «pROC: an open-source package for R and S+ to analyze and compare ROC curves,» *BMC Bioinformatics*, vol. 12, pág. 77, 2011.

[48] A. J. Vickers, A. M. Cronin y C. B. Begg, «One statistical test is sufficient for assessing new predictive markers,» *BMC Medical Research Methodology*, vol. 11, n.º 13, ene. de 2011.

[49] D. Mason, «SU-E-T-33: pydicom: an open source DICOM library,» *Medical Physics*, vol. 38, n.º 6Part10, págs. 3493-3493, 2011.

[50] Martın Abadi, Ashish Agarwal, Paul Barham y col., *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*, 2015.

[51] F. Chollet y col., *Keras*, 2015. dirección: `https://github.com/fchollet/keras`.

[52] T. Kluyver, B. Ragan-Kelley, F. Pérez y col., «Jupyter Notebooks – a publishing format for reproducible computational workflows,» en *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, F. Loizides y B. Schmidt, eds., IOS Press, 2016, págs. 87-90.

[53] Y. Luo, J. Peng y J. Ma, «When causal inference meets deep learning,» *Nature Machine Intelligence*, vol. 2, n.º 8, págs. 426-427, 2020.

9

Anexos

9. Anexos

9.1. Anexo 1: Aprobación del Comité de Ética de la Investigación con Medicamentos de Cantabria

T. CONCEPCION SOLANAS GUERRERO, Secretaria del **COMITÉ DE ÉTICA DE LA INVESTIGACIÓN CON MEDICAMENTOS DE CANTABRIA**

CERTIFICA

Que este Comité ha evaluado la propuesta de los Investigadores Principales del estudio:

TÍTULO**: Desarrollo de aplicaciones de inteligencia artificial para la asistencia al diagnóstico de neumonía viral con radiografías de tórax.**

TIPO DE ESTUDIO: **Proyecto de Investigación (Código interno: 2020.187)**

y considera que:

- Se cumplen los requisitos necesarios de idoneidad del protocolo en relación con los objetivos del estudio y están justificados los riesgos y molestias previsibles para el sujeto, teniendo en cuenta los beneficios esperados.
- Es adecuado el procedimiento para obtener el consentimiento informado.
- La capacidad del investigador y sus colaboradores, y las instalaciones y medios disponibles, tal y como ha sido informado, son apropiados para llevar a cabo el estudio.

Este CEIm, emite un informe **FAVORABLE** para que dicho Estudio sea realizado en **HOSPITAL UNIVERSITARIO MARQUÉS DE VALDECILLA, HOSPITAL DE SIERRALLANA y HOSPITAL DE LAREDO**, actuando como investigadores principales los Dres. **LARA LLORET IGLESIAS y JOSÉ ANTONIO PARRA BLANCO.**

Como queda reflejado en el Acta: **11/2020 de 30 de abril de 2020**.

Lo que firmo en Santander, a **21 de mayo de 2020**

T. CONCEPCION SOLANAS GUERRERO
Secretaria del CEIm

Av. Cardenal Herrera Oria s/n
39011 Santander - España
www.idival.org
Tel. +34 942 31 55 15
Fundación Instituto de Investigación Marqués de Valdecilla
CIF: G 39788773

9.2. Anexo 2: Mapas de *clustering* jerarquizado

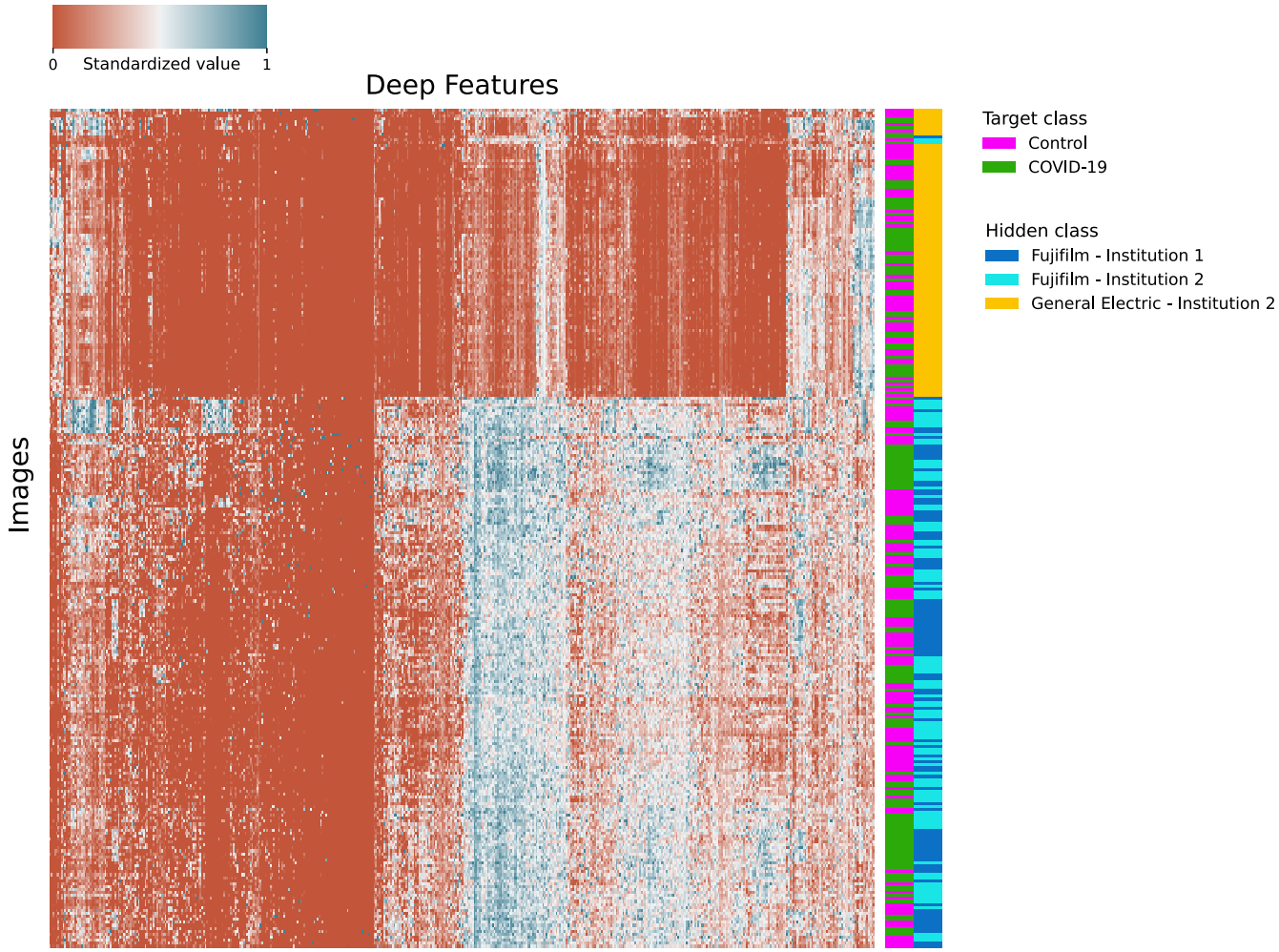


Figura 9.1: Clusterization of images from subsets T1, T2, and T3 based on Model-F1F2 features. The clusterization of images using hierarchical clustering resulted in two clusters. One cluster contained the images acquired by the Carestream device and the other cluster contained the images acquired by the Fujifilm devices. Therefore, images were separated based on the device which acquired them. Additionally, images acquired by the two Fujifilm devices were not separated, despite being from different institutions and having different acquisition protocols.

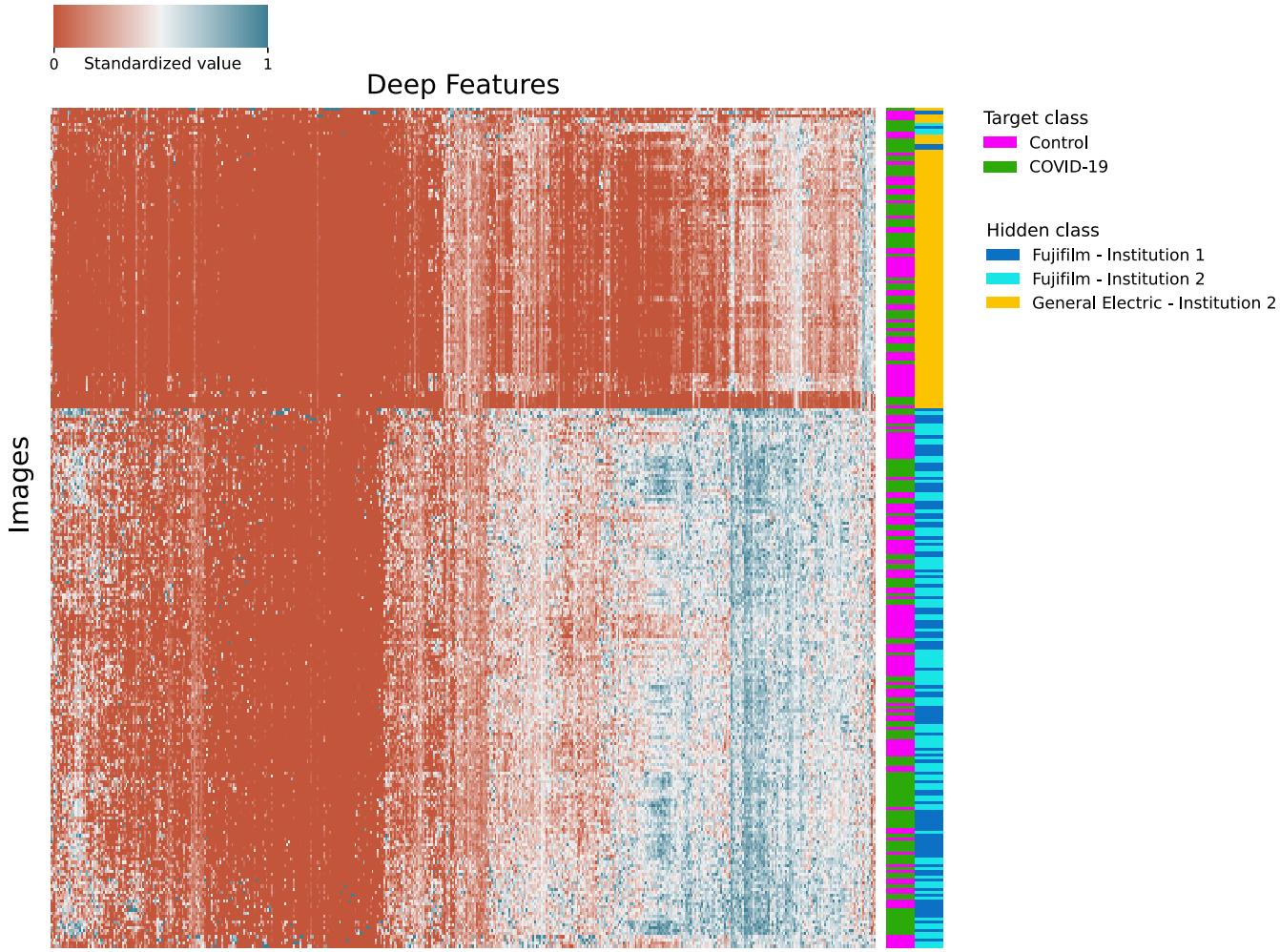


Figura 9.2: Clusterization of cropped images from subsets T1, T2, and T3 based on Model-F1F2 features. When images were cropped to eliminate metallic tokens, no relevant changes were observed in regards to the results shown in **??**

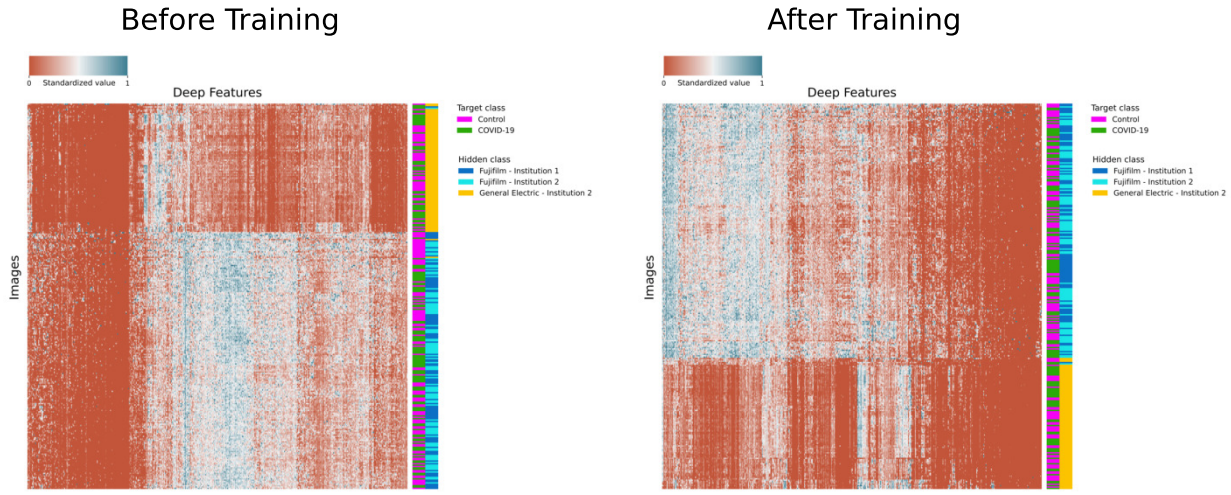
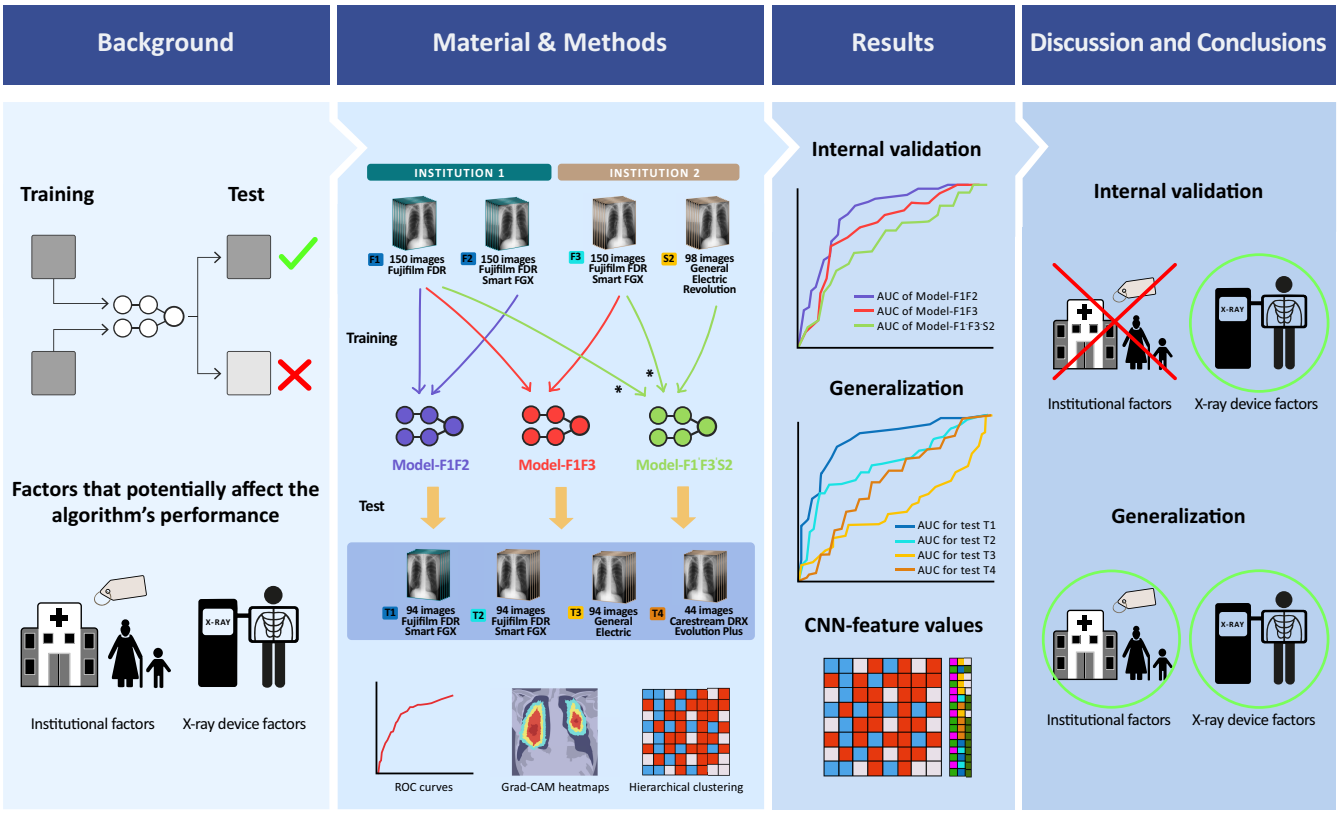


Figura 9.3: Clusterization of images from subsets T1, T2, and T3 based on features extracted by Model-F1'F3'S2 before and after being trained. Images were clustered based on features extracted by the pre-trained and trained versions of Model-F1'F3'S2. In both cases, images were successfully grouped into different device clusters. Thus, fine-tunning did not reduce differences in feature values between images from different X-ray machines.

9.3. Anexo 3: Resumen gráfico del trabajo



9.4. Anexo 4: *Curriculum vitae*

CURRÍCULUM VÍTAE NORMALIZADO

Pablo Menéndez Fernández-Miranda.

Generado desde: Editor CVN de FECYT

Fecha del documento: 16/05/2022

v 1.4.3

e1f2a8e94b7b00b1619c0d7eaddf7de1

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN

FECYT FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA

Pablo Menéndez Fernández-Miranda.

Apellidos:	**Menéndez Fernández-Miranda.**
Nombre:	**Pablo**
DNI:	**71901048N**
ORCID:	**0000-0003-4742-8955**
Fecha de nacimiento:	**14/03/1993**
Sexo:	**Hombre**
Nacionalidad:	**España**
País de nacimiento:	**España**
C. Autón./Reg. de nacimiento:	**Principado de Asturias**
Provincia de contacto:	**Asturias**
Ciudad de nacimiento:	**Avilés**
Dirección de contacto:	**Bernardo Álvarez Galán, 80**
Código postal:	**33405**
País de contacto:	**España**
C. Autón./Reg. de contacto:	**Principado de Asturias**
Ciudad de contacto:	**Salinas**
Teléfono fijo:	**(0034) 985518961**
Correo electrónico:	**pablomfm@telecable.es / pmfm@hotmail.es**
Teléfono móvil:	**(0034) 676095980**

Situación profesional actual

Entidad empleadora: Hospital Universitario Marqués de Valdecilla
Categoría profesional: Médico Residente
Fecha de inicio: 25/05/2018
Funciones desempeñadas: Residente de Radiología.

Cargos y actividades desempeñados con anterioridad

	Entidad empleadora	Categoría profesional	Fecha de inicio
	Infinitargo S.L.	Socio fundador	01/04/2016

Entidad empleadora: Infinitargo S.L.
Tipo de entidad: Entidad Empresarial
Categoría profesional: Socio fundador
Fecha de inicio-fin: 01/04/2016 - 01/03/2017

Formación académica recibida

Titulación universitaria

Estudios de 1º y 2º ciclo, y antiguos ciclos (Licenciados, Diplomados, Ingenieros Superiores, Ingenieros Técnicos, Arquitectos)

Titulación universitaria: Titulado Superior
Nombre del título: Graduado en Medicina
Entidad de titulación: Universidad de Oviedo **Tipo de entidad:** Universidad
Fecha de titulación: 12/07/2017

Doctorados

Programa de doctorado: Estudiante de doctorado en Medicina - Inteligencia Artificial en imagen médica
Entidad de titulación: Universidad de Oviedo **Tipo de entidad:** Universidad

Otra formación universitaria de posgrado

Titulación de posgrado: Estudiante Máster Oficial Interuniversitario en Data Science
Entidad de titulación: Universidad de Cantabria y Universidad Internacional Menéndez Pelayo **Tipo de entidad:** Universidad
Facultad, instituto, centro: Facultad de Ciencias
Fecha de titulación: 22/06/2022

Formación especializada, continuada, técnica, profesionalizada, de reciclaje y actualización (distinta a la formación académica reglada y a la sanitaria)

1 **Título de la formación:** HelloAI Professional Edition - Artificial Intelligence Online Training
Entidad de titulación: GE Healthcare, KTH Royal Institute of Technology, and LEITAT Technological Center
Fecha de finalización: 30/11/2021 **Duración en horas:** 187 horas

2 **Título de la formación:** DermoPracticAP
Entidad de titulación: Comisión de Formación Continuada de las Profesiones Sanitarias **Tipo de entidad:** Agencia Estatal
Fecha de finalización: 08/12/2020 **Duración en horas:** 200 horas

3 **Título de la formación:** Aprende a programar con Python
Entidad de titulación: Universidad Nacional de Educación a Distancia (UNED) **Tipo de entidad:** Universidad
Fecha de finalización: 09/11/2020 **Duración en horas:** 20 horas

4 **Título de la formación:** Gestiona la EPOC
Entidad de titulación: Sociedad Española de Médicos de Atención Primaria (SEMERGEN) **Tipo de entidad:** Asociaciones y Agrupaciones
Fecha de finalización: 30/10/2020 **Duración en horas:** 100 horas

5 **Título de la formación:** Essential Math for Machine Learning - Python Edition
Entidad de titulación: Microsoft y edX **Tipo de entidad:** Entidad Empresarial
Fecha de finalización: 29/05/2020 **Duración en horas:** 6 horas

6 **Título de la formación:** Desmitificando Radiomics: Nuevos Informes Radiológicos Avanzados 2
Entidad de titulación: Sociedad Española de Radiología Médica (SERAM) e Instituto de Estudios de Ciencias de la Salud de Castilla y León (IECSCYL) **Tipo de entidad:** Asociaciones y Agrupaciones
Fecha de finalización: 04/05/2020 **Duración en horas:** 25 horas

7 **Título de la formación:** Desmitificando Radiomics: Nuevos Informes Radiológicos Avanzados 1
Entidad de titulación: Sociedad Española de Radiología Médica (SERAM) e Instituto de Estudios de Ciencias de la Salud de Castilla y León (IECSCYL) **Tipo de entidad:** Asociaciones y Agrupaciones
Fecha de finalización: 22/12/2019 **Duración en horas:** 35 horas

8 **Título de la formación:** Introducción a la Inteligencia Artificial en Radiología
Entidad de titulación: Sociedad Española de Radiología Médica (SERAM) e Instituto de Estudios de Ciencias de la Salud de Castilla y León (IECSCYL) **Tipo de entidad:** Asociaciones y Agrupaciones
Fecha de finalización: 16/10/2019 **Duración en horas:** 20 horas

9 **Título de la formación:** Programa de Educación Continuada en Tratamiento con Anticoagulantes Orales 2018-2019
Entidad de titulación: Comisión de Formación Continuada de las Profesiones Sanitarias de la Comunidad de Madrid-Sistema Nacional de Salud
Fecha de finalización: 31/08/2019 **Duración en horas:** 50 horas

10 **Título de la formación:** Actualización en microbiota autóctona, probióticos y prebióticos módulo I: microbiota, probióticos y prebióticos
Entidad de titulación: Consejo Catalán de Formación Continuada de las Profesiones Sanitarias y la Comisión de Formación Continuada del Sistema Nacional de Salud **Tipo de entidad:** Entidad Gestora del Sistema Nacional de Salud
Fecha de finalización: 10/06/2019 **Duración en horas:** 20 horas

11 **Título de la formación:** Actualización en microbiota autóctona, probióticos y prebióticos módulo II: aplicaciones clínicas
Entidad de titulación: Consejo Catalán de Formación Continuada de las Profesiones Sanitarias y la Comisión de Formación Continuada del Sistema Nacional de Salud **Tipo de entidad:** Entidad Gestora del Sistema Nacional de Salud
Fecha de finalización: 10/06/2019 **Duración en horas:** 20 horas

12 **Título de la formación:** Revisión de Artículos en la Revista Radiología
Entidad de titulación: Sociedad Española de Radiología Médica (SERAM) – organiza; Instituto de Estudios de Ciencias de la Salud de Castilla y León (ECSCYL) y la Unión Europea de Especialistas Médicos (UEMS) – acreditan

Fecha de finalización: 04/06/2019 **Duración en horas:** 20 horas

13 **Título de la formación:** Statistics for Medical Professionals
Entidad de titulación: Universidad de Stanford **Tipo de entidad:** Universidad
Fecha de finalización: 02/06/2019 **Duración en horas:** 24 horas

14 **Título de la formación:** Módulo de Prescripción Electrónica Online
Entidad de titulación: Subdirección de Desarrollo y Calidad Asistencial del Servicio Cántabro de Salud (SCS) - (SOFOS)
Fecha de finalización: 21/02/2019 **Duración en horas:** 8 horas

15 **Título de la formación:** Introducción al manejo del SPSS
Entidad de titulación: Colegio Oficial de Médicos de Cantabria
Fecha de finalización: 07/02/2019 **Duración en horas:** 12 horas

16 **Título de la formación:** Paving a Path to Relief of Irritable Bowel Syndrome and Chronic Idiopathic Constipation: Live Patient Simulations Demostrating Innovations in Diagnosis and Treatment
Entidad de titulación: Paradigm Medical Communications
Fecha de finalización: 22/12/2018 **Duración en horas:** 15 horas

17 **Título de la formación:** Redefining Breast Cancer Treatment with Personalized Medicine: HR+ Cases
Entidad de titulación: The France Foundation
Fecha de finalización: 08/12/2018 **Duración en horas:** 10 horas

18 **Título de la formación:** Curso especializado en Motores de Búsquedas Bibliográficas
Entidad de titulación: Hospital Universitario Marqués de Valdecilla **Tipo de entidad:** Instituciones Sanitarias
Fecha de finalización: 19/11/2018 **Duración en horas:** 2 horas

19 **Título de la formación:** Curso de Radiología Básica para Residentes HUMV (1 EDICION)
Entidad de titulación: Hospital Universitario Marqués de Valdecilla **Tipo de entidad:** Instituciones Sanitarias
Fecha de finalización: 06/11/2018 **Duración en horas:** 3 horas

20 **Título de la formación:** R.C.P. básica
Entidad de titulación: Hospital Universitario Marqués de Valdecilla **Tipo de entidad:** Instituciones Sanitarias
Fecha de finalización: 12/07/2018 **Duración en horas:** 3 horas

21 **Título de la formación:** Introducción en la Biblioteca Marquesa de Pelayo
Entidad de titulación: Hospital Universitario Marqués de Valdecilla **Tipo de entidad:** Instituciones Sanitarias
Fecha de finalización: 06/07/2018 **Duración en horas:** 2 horas

22 **Título de la formación:** Curso extensivo de Triaje
Entidad de titulación: Hospital Universitario Marqués de Valdecilla **Tipo de entidad:** Instituciones Sanitarias
Fecha de finalización: 30/06/2018 **Duración en horas:** 30 horas

23 **Título de la formación:** Curso de Urgencias XXIII para Médicos Residentes
Entidad de titulación: Hospital Universitario Marqués de Valdecilla **Tipo de entidad:** Instituciones Sanitarias
Fecha de finalización: 29/06/2018 **Duración en horas:** 100 horas

24 **Título de la formación:** Sistema de Notificación y Aprendizaje para la Seguridad del Paciente para Hospitales (SiNASP)
Entidad de titulación: Instituto Universitario Avedis Donabedian-UAB **Tipo de entidad:** Fundación
Fecha de finalización: 14/06/2018 **Duración en horas:** 4 horas

25 **Título de la formación:** Protección Radiológica R-1
Entidad de titulación: Hospital Universitario Marqués de Valdecilla **Tipo de entidad:** Instituciones Sanitarias
Fecha de finalización: 01/06/2018 **Duración en horas:** 4 horas

26 **Título de la formación:** Prescripción Electrónica Asistida
Entidad de titulación: Hospital Universitario Marqués de Valdecilla **Tipo de entidad:** Instituciones Sanitarias
Fecha de finalización: 31/05/2018 **Duración en horas:** 2 horas

27 **Título de la formación:** Curso Intensivo MIR Asturias de Preparación del Examen MIR 2017
Entidad de titulación: Curso Intensivo MIRAsturias **Tipo de entidad:** Entidad Empresarial
Fecha de finalización: 09/02/2018 **Duración en horas:** 672 horas

28 **Título de la formación:** Atención Materno-Infantil
Entidad de titulación: Curso Intensivo MIRAsturias **Tipo de entidad:** Entidad Empresarial
Fecha de finalización: 11/01/2018 **Duración en horas:** 60 horas

29 **Título de la formación:** Bases Quirúrgicas de la Asistencia Sanitaria
Entidad de titulación: Curso Intensivo MIRAsturias **Tipo de entidad:** Entidad Empresarial
Fecha de finalización: 08/01/2018 **Duración en horas:** 100 horas

30 **Título de la formación:** Actualización en Patología Clínica General
Entidad de titulación: Curso Intensivo MIRAsturias **Tipo de entidad:** Entidad Empresarial
Fecha de finalización: 19/12/2017 **Duración en horas:** 100 horas

31 **Título de la formación:** Actualización en Patología Onco-hematológica
Entidad de titulación: Curso Intensivo MIRAsturias **Tipo de entidad:** Entidad Empresarial
Fecha de finalización: 15/12/2017 **Duración en horas:** 84 horas

32 **Título de la formación:** Medicina Preventiva, Bioestadística y Metodología Aplicada a la Asistencia Médica.
Entidad de titulación: Curso Intensivo MIRAsturias **Tipo de entidad:** Entidad Empresarial
Fecha de finalización: 09/12/2017 **Duración en horas:** 80 horas

Formación sanitaria especializada

Título de la especialidad: Residente de Radiología
Entidad de titulación: MINISTERIO DE EDUCACION Y CIENCIA Y DEPORTE **Tipo de entidad:** Agencia Estatal
Ciudad entidad titulación: España
Fecha de inicio-fin: 23/05/2018 - 23/05/2022 **Duración:** 4 años

Conocimiento de idiomas

Idioma	Comprensión auditiva	Comprensión de lectura	Interacción oral	Expresión oral	Expresión escrita
Inglés	B2	B2	B2	B2	B2

Actividad docente

Formación académica impartida

Nombre de la asignatura/curso: Aprendizaje automático II
Titulación universitaria: MÁSTER UNIVERSITARIO EN CIENCIA DE DATOS / MASTER IN DATA SCIENCE
Fecha de inicio: 12/03/2021 **Fecha de finalización:** 12/03/2021
Entidad de realización: Universidad de Cantabria y Universidad Menéndez Pelayo **Tipo de entidad:** Centros y Estructuras Universitarios y Asimilados
Facultad, instituto, centro: Facultad de Ciencias de la Universidad de Cantabria

Formación sanitaria especializada impartida

Título de la especialidad: Radiodiagnóstico **Título de la subespecialidad:** Neurorradiología
Entidad de titulación: Hospital Universitario Marqués de Valdecilla **Tipo de entidad:** Instituciones Sanitarias
Fecha de inicio-fin: 22/10/2020 - 22/10/2020

Formación sanitaria en I+D, y/o posformación sanitaria especializada en I+D impartida

1 **Título de la especialidad:** Curso de Tendencias en Investigación Clínica - Inteligencia Artificial y Datos Ómicos: Medicina de Precisión Personalizada
Entidad de titulación: Instituto de investigación sanitaria Valdecilla (IDIVAL)
Fecha de inicio-fin: 10/05/2021 - 10/05/2021

2 **Título de la especialidad:** Curso de Tendencias en Investigación Clínica - Inteligencia Artificial y Datos Ómicos: Resolución de un caso práctico
Entidad de titulación: Instituto de investigación sanitaria Valdecilla (IDIVAL)
Fecha de inicio-fin: 10/05/2021 - 10/05/2021

Dirección de tesis doctorales y/o proyectos fin de carrera

1 **Título del trabajo:** Estudio radiómico de hamartomas y tumores carcinoides de pulmón
Entidad de realización: Universidad de Oviedo **Tipo de entidad:** Universidad
Alumno/a: David Corral Fontecha
Fecha de defensa: 08/06/2022

2 **Título del trabajo:** Estudio radiómico de tumores malignos del pulmón
Fecha de defensa: 06/06/2022

3 **Título del trabajo:** Biopsia pulmonar con aguja gruesa: complicaciones, rentabilidad diagnóstica y hallazgos anatomopatológicos en el periodo 2017-2019. ¿A partir de los hallazgos de la TAC, puede la Inteligencia Artificial (IA) predecir el tipo hidrológico?
Entidad de realización: Universidad de Cantabria **Tipo de entidad:** Universidad
Alumno/a: Alba Fernández Rodríguez
Calificación obtenida: 10 - Matrícula de Honor
Fecha de defensa: 15/05/2021

Actividad sanitaria

Cursos y seminarios impartidos orientados a la mejora de la atención de salud para profesionales sanitarios

1 **Nombre del curso:** Curso de Tendencias en Investigación Clínica
Ciudad entidad organizadora: Santander, Cantabria, España
Entidad de realización: Instituto de Investigación Sanitaria de Valdecilla **Tipo de entidad:** Fundación
Ciudad entidad realización: Santander, Cantabria, España
Inteligencia Artificial y Datos Ómicos: Medicina de Precisión Personalizada.
Fecha de inicio-fin: 12/04/2021 - 13/05/2021 **Horas impartidas:** 2

2 **Nombre del curso:** Curso de Urgencias XXIV para Médicos Residentes
Ciudad entidad organizadora: Santander, Cantabria, España
Tipo de participación: Participativo - Ponencia invitada/ Keynote
Autor de correspondencia: Si
Entidad de realización: Hospital Universitario Marqués de Valdecilla **Tipo de entidad:** Instituciones Sanitarias
Ciudad entidad realización: Santander, Cantabria, España
Fecha de inicio-fin: 01/10/2020 - 30/10/2020 **Horas impartidas:** 2

Congresos, cursos y seminarios orientados a la atención de salud

1 **Tipo del evento:** Congreso
Ciudad entidad realización: London, Inner London, Reino Unido
Entidad organizadora: The British Institute of Radiology **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Londres, Inner London, Reino Unido
Ámbito geográfico: Otros organismos internacionales
Idioma en que se impartió: Inglés **Fecha de presentación:** 04/11/2020
Congreso - BIR Annual Congress 2020.

2 **Tipo del evento:** Congreso
Ciudad entidad realización: Viena, Austria
Entidad organizadora: European Society of Radiology (ESR) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Viena, Austria
Ámbito geográfico: Otros organismos internacionales **Tipo de participación:** Participativo - Ponencia oral (comunicación oral)
Idioma en que se impartió: Inglés **Fecha de presentación:** 15/07/2020
Congreso - European Congress of Radiology (ECR) 2020.

3 **Tipo del evento:** Congreso
Ciudad entidad realización: Santander, Cantabria, España
Entidad organizadora: Sociedad Española de Médicos de Atención Primaria (SEMERGEN) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Madrid, Comunidad de Madrid, España
Ámbito geográfico: Autonómica
Fecha de presentación: 30/01/2020
Congreso - XV Congreso Autonómico de Cantabria. SEMERGEN Cantabria 2020.

4 **Tipo del evento:** Congreso
Ciudad entidad realización: Chicago, Estados Unidos de América
Entidad organizadora: Radiological Society of North America (RSNA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Chicago, Estados Unidos de América
Ámbito geográfico: Otros organismos internacionales
Idioma en que se impartió: Inglés **Fecha de presentación:** 01/12/2019
Congreso - 105th Scientific Assembly and Annual Meeting of the Radiological Society of North America (2019).

5 **Tipo del evento:** Congreso
Ciudad entidad realización: Valladolid, Castilla y León, España
Entidad organizadora: Sociedad Española de Neurorradiología (SENR) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Madrid, Comunidad de Madrid, España
Ámbito geográfico: Nacional
Idioma en que se impartió: Español **Fecha de presentación:** 24/10/2019
Congreso - XLVIII Reunión Anual Sociedad Española de Neurorradiología.

6 **Tipo del evento:** Congreso
Ciudad entidad realización: Valencia, Comunidad Valenciana, España
Entidad organizadora: European Society of Medical Imaging Informatics (EuSoMII)
Ciudad entidad organizadora: Viena, Austria
Ámbito geográfico: Otros organismos internacionales **Tipo de participación:** Participativo - Ponencia oral (comunicación oral)
Idioma en que se impartió: Inglés **Fecha de presentación:** 18/10/2019
Congreso - EuSoMII ANNUAL MEETING Valencia 18-19 October 2019 “Medical Imaging Informatics – AI, Clinical Applications and more”.

7 **Tipo del evento:** Congreso
Ciudad entidad realización: Gijón, Principado de Asturias, España
Entidad organizadora: Sociedad Española de Médicos de Atención Primaria (SEMERGEN) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Madrid, Comunidad de Madrid, España
Ámbito geográfico: Nacional
Idioma en que se impartió: Español **Fecha de presentación:** 16/10/2019
Congreso - 41º Congreso Nacional SEMERGEN 2019.

8 **Tipo del evento:** Congreso
Ciudad entidad realización: Logroño, La Rioja, España
Entidad organizadora: Sociedad Centro Norte de Radiología (CENORA)
Ámbito geográfico: Nacional **Tipo de participación:** Participativo - Ponencia oral (comunicación oral)
Fecha de presentación: 04/10/2019
Congreso - X Congreso CENORA 2019.

9 **Tipo del evento:** Congreso
Ciudad entidad realización: Sevilla, Andalucía, España
Entidad organizadora: Sociedad Española de Radiología de Urgencias (SERAU) y European Society of Emergency Radiology (ESER)
Ámbito geográfico: Otros organismos internacionales
Idioma en que se impartió: Inglés **Fecha de presentación:** 16/05/2019
Congreso - SERAU-ESER Annual Scientific Meeting 2019.

10 **Tipo del evento:** Congreso
Ciudad entidad realización: Valladolid, Castilla y León, España
Entidad organizadora: Sociedad Española de Radiología (SERAM) y American Roentgen Ray Society (ARRS)
Ámbito geográfico: Otros organismos internacionales **Tipo de participación:** Participativo - Ponencia oral (comunicación oral)
Fecha de presentación: 04/04/2019
Congreso - II Jornadas SERAM-ARRS.

Experiencia científica y tecnológica

Actividad científica o tecnológica

Proyectos de I+D+i financiados en convocatorias competitivas de Administraciones o entidades públicas y privadas

Nombre del proyecto: Desarrollo de aplicaciones de inteligencia artificial para la asistencia al diagnóstico de neumonía con radiografías de tórax
Entidad de realización: Consejo Superior de Investigaciones Científicas (CSIC)
Ciudad entidad realización: Santander, Cantabria, España
Nombres investigadores principales (IP, Co-IP,...): Lara Lloret Iglesias; David Rodríguez González
Nº de investigadores/as: 5
Fecha de inicio: 02/09/2020
Cuantía total: 6.000 €

Actividades científicas y tecnológicas

Producción científica

Publicaciones, documentos científicos y técnicos

1 Amaia Pérez del Barrio; Pablo Menéndez Fernández-Miranda; Pablo Sanz Bellón; Lara Lloret Iglesias; David Rodríguez González. Inteligencia artificial en Radiología: introducción a los conceptos más importantes. Radiología. Elsevier, 30/04/2022.
Tipo de producción: Artículo científico **Tipo de soporte:** Revista
Autor de correspondencia: No

2 Lara Lloret Iglesias; Pablo Sanz Bellón; Amaia Pérez del Barrio; Pablo Menéndez Fernández-Miranda; David Rodríguez González; José Antonio Álvarez Vega; Andrés Antonio González Mandly; José Antonio Parra Blanco. A primer on deep learning and convolutional neural networks for clinicians. Insights into Imaging. 12 - 1, Springer, 12/08/2021.
Tipo de producción: Artículo científico **Tipo de soporte:** Revista
Autor de correspondencia: No

3 Javier Azcona Sáez; Darío Herrán de la Gala; Phoebe Phuong-Boi Bui; Javier Arnáiz García; Pablo Menéndez Fernández-Miranda; Yasmina Lamprecht; Marta Drake Pérez; Enrique Marco de Lucas. Instagram's Influence on Radiology Today: Reviewing the Evolving Educational Journey from a Hospital to National Societies. Journal of Digital Imaging. Springer, 08/07/2021.
Tipo de producción: Artículo científico **Tipo de soporte:** Revista
Autor de correspondencia: No

4 Juan Miranda Bautista; Javier Fernández Jara; Santiago Miranda Bautista; Pablo Menéndez Fernández-Miranda; Maria Valencia Mora; Begoña Gutierrez San Jose; Mateo González Estevez; Blanca Mur Molina; Patricia Patilla Vázquez. Infraspinatus atrophy due to Bennett lesion causing suprascapular nerve palsy. Journal of Ultrasonography. 21 - 85, pp. 177 - 181. 07/06/2021.
Tipo de producción: Artículo científico **Tipo de soporte:** Revista
Autor de correspondencia: No

5 Pablo Menéndez Fernández-Miranda; Pablo Sanz Bellón; Amaia Pérez del Barrio; Lara Lloret Iglesias; Pedro Solís García; Fernando Aguilar Gómez; David Rodríguez González; José Antonio Álvarez Vega. Developing a Training Web Application for Improving the COVID-19 Diagnostic Accuracy on Chest X-ray. Journal of Digital Imaging. Springer, 08/03/2021.
Tipo de producción: Artículo científico **Tipo de soporte:** Revista
Autor de correspondencia: Si

6 Prevalence of D. fragilis infection in the household contacts of a group of infected patients. Enfermedades Infecciosas y Microbiología Clínica.36 - 7, pp. 395 - 462. Elsevier, 01/10/2017.
Tipo de producción: Artículo científico **Tipo de soporte:** Revista

7 Radiologia Vascular. Protocolos urgentes de radiodiagnóstico: TC, RM e intervencionismo básico. Valnera Medicina, 01/02/2022.
Tipo de producción: Capítulo de libro **Tipo de soporte:** Libro

8 Pablo Menéndez Fernández-Miranda; Enrique Marqués Fraguela; Pablo Sanz Bellón; Marta Drake Pérez; Andrés González Mandly. Inteligencia Artificial en Medicina. Tendencias en Investigación Clínica 2021.
Tipo de producción: Capítulo de libro **Tipo de soporte:** Libro
Autor de correspondencia: Si

Trabajos presentados en congresos nacionales o internacionales

1 **Título del trabajo:** Aortoenteric fistula: Keys to the diagnosis of an increasingly less prevalent but fatal entity.
Nombre del congreso: European Congress of Radiology (ECR) 2022
Ciudad de celebración: Córdoba, España
Fecha de celebración: 02/03/2022
Fecha de finalización: 06/03/2022
Entidad organizadora: European Society of Radiology (ESR)

2 **Título del trabajo:** Evaluation of medical devices on critically ill patients using portable chest radiography.
Nombre del congreso: European Congress of Radiology (ECR) 2022
Ciudad de celebración: Córdoba, España
Fecha de celebración: 02/03/2022
Fecha de finalización: 06/03/2022
Entidad organizadora: European Society of Radiology (ESR)

3 **Título del trabajo:** Radical Cystectomy with Urinary Diversion: normal CT findings and main complications
Nombre del congreso: European Congress of Radiology (ECR) 2022
Ciudad de celebración: Córdoba, España
Fecha de celebración: 02/03/2022
Fecha de finalización: 06/03/2022
Entidad organizadora: European Society of Radiology (ESR)

4 **Título del trabajo:** Spontaneous pneumomediastinum: Who? Why? How can radiologists help?
Nombre del congreso: European Congress of Radiology (ECR) 2022
Ciudad de celebración: Córdoba, España
Fecha de celebración: 02/03/2022
Fecha de finalización: 06/03/2022
Entidad organizadora: European Society of Radiology (ESR)

5 **Título del trabajo:** Vascular complications in renal transplant: Doppler ultrasound evaluation and the potential therapeutic role of interventional radiology.
Nombre del congreso: European Congress of Radiology (ECR) 2022
Ciudad de celebración: Córdoba, España
Fecha de celebración: 02/03/2022
Fecha de finalización: 06/03/2022
Entidad organizadora: European Society of Radiology (ESR)

6 **Título del trabajo:** Nutcracker Syndrome: A Silent Enemy. Diagnosis, Endovascular Treatment And Its Complications.
Nombre del congreso: 107th Scientific Assembly and Annual Meeting of the Radiological Society of North America (2021)
Ciudad de celebración: Córdoba, España
Fecha de celebración: 28/11/2021
Fecha de finalización: 02/12/2021
Entidad organizadora: Radiological Society of North America (RSNA)

7 **Título del trabajo:** The Leading Role Of The Radiologist In The Assessment Of Retroperitoneal Haematoma: From Diagnosis To Endovascular Treatment
Nombre del congreso: 107th Scientific Assembly and Annual Meeting of the Radiological Society of North America (2021)
Ciudad de celebración: Córdoba, España
Fecha de celebración: 28/11/2021
Fecha de finalización: 02/12/2021
Entidad organizadora: Radiological Society of North America (RSNA)

8 **Título del trabajo:** Urinary Diversions: Main Normal And Abnormal Findings With Special Focus
Nombre del congreso: 107th Scientific Assembly and Annual Meeting of the Radiological Society of North America (2021)
Ciudad de celebración: Córdoba, España
Fecha de celebración: 28/11/2021
Fecha de finalización: 02/12/2021
Entidad organizadora: Radiological Society of North America (RSNA)

9 **Título del trabajo:** "Mi caso favorito de la urgencia": Síndrome de May-Thurner
Nombre del congreso: VIII Congreso SERAU
Ciudad de celebración: Córdoba, España
Fecha de celebración: 08/10/2021
Fecha de finalización: 09/10/2021
Entidad organizadora: Sociedad Española de Radiología de Urgencias (SERAU)

10 **Título del trabajo:** Actinomicosis pélvica que simula por imagen una neoplasia ginecológica avanzada.
Nombre del congreso: VIII Congreso SERAU
Ciudad de celebración: Córdoba, España

Fecha de celebración: 08/10/2021
Fecha de finalización: 09/10/2021
Entidad organizadora: Sociedad Española de Radiología de Urgencias (SERAU)

11 **Título del trabajo:** Cómo realizar e interpretar una arteriografía de miembros inferiores
Nombre del congreso: 35 Congreso Nacional SERAM
Ciudad de celebración: Zaragoza, España
Fecha de celebración: 19/05/2021
Fecha de finalización: 26/05/2021
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM)

12 **Título del trabajo:** Degeneración mucoide del ligamento cruzado anterior. Su diagnóstico, ¿es relevante?
Nombre del congreso: 35 Congreso Nacional SERAM
Ciudad de celebración: Zaragoza, España
Fecha de celebración: 19/05/2021
Fecha de finalización: 26/05/2021
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM)

13 **Título del trabajo:** El cráneo postquirúrgico: lo que el radiólogo debe saber.
Nombre del congreso: 35 Congreso Nacional SERAM
Ciudad de celebración: Zaragoza, España
Fecha de celebración: 19/05/2021
Fecha de finalización: 26/05/2021
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM)

14 **Título del trabajo:** Síndromes aórticos agudos en la urgencia radiológica: definición, fisiopatología y hallazgos en imagen.
Nombre del congreso: 35 Congreso Nacional SERAM
Ciudad de celebración: Zaragoza, España
Fecha de celebración: 19/05/2021
Fecha de finalización: 26/05/2021
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM)

15 **Título del trabajo:** ¿Son adecuadas las peticiones? Evaluación de las solicitudes urgentes de ecografía y TC durante un mes en hospital de segundo nivel.
Nombre del congreso: 35 Congreso Nacional SERAM
Ciudad de celebración: Zaragoza, España
Fecha de celebración: 19/05/2021
Fecha de finalización: 26/05/2021
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM)

16 **Título del trabajo:** Ablación de nódulos tiroideos benignos guiada por ecografía: estudio de cohortes prospectivo.
Nombre del congreso: 35 Congreso Nacional SERAM
Ciudad de celebración: Zaragoza, España
Fecha de celebración: 19/05/2021
Fecha de finalización: 26/05/2021
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM)

17 **Título del trabajo:** Claves radiológicas para el diagnóstico de pielitis incrustante: lo que el radiólogo debe saber.
Nombre del congreso: 35 Congreso Nacional SERAM
Ciudad de celebración: Zaragoza, España
Fecha de celebración: 19/05/2021
Fecha de finalización: 26/05/2021
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM)

18 **Título del trabajo:** Complicaciones Osteomusculares de los Aneurismas Micóticos en el Esqueleto Axial.
Nombre del congreso: 35 Congreso Nacional SERAM
Ciudad de celebración: Zaragoza, España
Fecha de celebración: 19/05/2021
Fecha de finalización: 26/05/2021
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM)

19 **Título del trabajo:** Dolor agudo en fosa ilíaca izquierda: más allá de la diverticulitis.
Nombre del congreso: 35 Congreso Nacional SERAM
Ciudad de celebración: Zaragoza, España
Fecha de celebración: 19/05/2021
Fecha de finalización: 26/05/2021
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM)

20 **Título del trabajo:** EL PACIENTE CON TRASPLANTE DE PULMÓN: COMPLICACIONES Y CLAVES DIAGNÓSTICAS
Nombre del congreso: 35 Congreso Nacional SERAM
Ciudad de celebración: Zaragoza, España
Fecha de celebración: 19/05/2021
Fecha de finalización: 26/05/2021
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM)

21 **Título del trabajo:** El cuello: patología en los diferentes espacios
Nombre del congreso: 35 Congreso Nacional SERAM
Ciudad de celebración: Zaragoza, España
Fecha de celebración: 19/05/2021
Fecha de finalización: 26/05/2021
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM)

22 **Título del trabajo:** El valor de la ecografía dinámica en la exploración del hombro
Nombre del congreso: 35 Congreso Nacional SERAM
Ciudad de celebración: Zaragoza, España
Fecha de celebración: 19/05/2021
Fecha de finalización: 26/05/2021
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM)

23 **Título del trabajo:** Encefalitis por Enterovirus en la edad pediátrica. Experiencia en nuestro hospital.
Nombre del congreso: 35 Congreso Nacional SERAM
Ciudad de celebración: Zaragoza, España
Fecha de celebración: 19/05/2021
Fecha de finalización: 26/05/2021
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM)

24 **Título del trabajo:** Ictus de la arteria de Percheron: actualización de los hallazgos radiológicos.
Nombre del congreso: 35 Congreso Nacional SERAM
Ciudad de celebración: Zaragoza, España
Fecha de celebración: 19/05/2021
Fecha de finalización: 26/05/2021
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM)

25 **Título del trabajo:** Integración del sistema de gestión de dosis en el Servicio Cántabro de Salud (SCS): DISCA
Nombre del congreso: 35 Congreso Nacional SERAM
Ciudad de celebración: Zaragoza, España
Fecha de celebración: 19/05/2021
Fecha de finalización: 26/05/2021
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM)

26 **Título del trabajo:** La ecografía en el estudio de la metatarsalgia central: mucho más que la neuralgia de Morton
Nombre del congreso: 35 Congreso Nacional SERAM
Ciudad de celebración: Zaragoza, España
Fecha de celebración: 19/05/2021
Fecha de finalización: 26/05/2021
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM)

27 **Título del trabajo:** Malformaciones vasculares cerebrales: ¿cuál es el papel de cada prueba de imagen?.
Nombre del congreso: 35 Congreso Nacional SERAM
Ciudad de celebración: Zaragoza, España
Fecha de celebración: 19/05/2021
Fecha de finalización: 26/05/2021
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM)

28 **Título del trabajo:** Más allá del vólvulo de sigma
Nombre del congreso: 35 Congreso Nacional SERAM
Ciudad de celebración: Zaragoza, España
Fecha de celebración: 19/05/2021
Fecha de finalización: 26/05/2021
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM)

29 **Título del trabajo:** Nuevas estrategias para la optimización del control dosimétrico en los servicios de radiología utilizando programas de Business Intelligence.
Nombre del congreso: 35 Congreso Nacional SERAM
Ciudad de celebración: Zaragoza, España
Fecha de celebración: 19/05/2021
Fecha de finalización: 26/05/2021
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM)

30 **Título del trabajo:** Patrones radiológicos de las infecciones pulmonares en el TC de tórax.
Nombre del congreso: 35 Congreso Nacional SERAM
Ciudad de celebración: Zaragoza, España
Fecha de celebración: 19/05/2021
Fecha de finalización: 26/05/2021
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM)

31 **Título del trabajo:** Problemas éticos y legales que surgirán de la implantación de sistemas de Inteligencia Artificial en Radiología.
Nombre del congreso: 35 Congreso Nacional SERAM
Ciudad de celebración: Zaragoza, España
Fecha de celebración: 19/05/2021
Fecha de finalización: 26/05/2021
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM)

32 **Título del trabajo:** Sospecha de colecistitis aguda: otras entidades que hay que buscar
Nombre del congreso: 35 Congreso Nacional SERAM
Ciudad de celebración: Zaragoza, España
Fecha de celebración: 19/05/2021
Fecha de finalización: 26/05/2021
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM)

33 **Título del trabajo:** Síndrome de obstrucción de la unión pieloureteral: todo lo que el radiólogo necesita saber.
Nombre del congreso: 35 Congreso Nacional SERAM
Ciudad de celebración: Zaragoza, España
Fecha de celebración: 19/05/2021
Fecha de finalización: 26/05/2021
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM)

34 **Título del trabajo:** Tejido esplénico ectópico, un gran imitador de masas abdominales: esplenosis y bazos accesorios.
Nombre del congreso: 35 Congreso Nacional SERAM
Ciudad de celebración: Zaragoza, España
Fecha de celebración: 19/05/2021
Fecha de finalización: 26/05/2021
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM)

35 **Título del trabajo:** ¿Que ha aportado el Ácido Gadoxético a nuestro servicio de Radiodiagnóstico en estos 10 años?
Nombre del congreso: 35 Congreso Nacional SERAM
Ciudad de celebración: Zaragoza, España
Fecha de celebración: 19/05/2021
Fecha de finalización: 26/05/2021
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM)

36 **Título del trabajo:** Telorragia en el varón: una condición infrecuente
Nombre del congreso: 1º Certamen de Casos Clínicos convocado por el Colegio de Médicos de Cantabria
Ciudad de celebración: Santander, España
Fecha de celebración: 05/05/2021
Fecha de finalización: 05/05/2021
Entidad organizadora: Colegio de Medicos de Cantabria (CGCOM)

37 **Título del trabajo:** Retroperitoneal haematoma: from the CT room to the interventional radiology suite
Nombre del congreso: European Congress of Radiology (ECR) 2021
Ciudad de celebración: Viena, Austria
Fecha de celebración: 02/03/2021

Fecha de finalización: 07/03/2021
Entidad organizadora: European Society of Radiology (ESR)

38 **Título del trabajo:** Reviewing Idiopathic Intersticial Penumonia's Radiographic Features According to the Latest American Thoracic Society / European Resiratory Society Updates
Nombre del congreso: 106th Scientific Assembly and Annual Meeting of the Radiological Society of North America (2020)
Ciudad de celebración: Chicago, Estados Unidos de América
Fecha de celebración: 20/11/2020
Fecha de finalización: 05/12/2020
Entidad organizadora: Radiological Society of North America (RSNA)

39 **Título del trabajo:** Spinal epidural hematoma. Every imaging technique counts
Nombre del congreso: BIR Annual Congress 2020
Autor de correspondencia: No
Ciudad de celebración: Londres, Reino Unido
Fecha de celebración: 04/11/2020
Fecha de finalización: 06/11/2020
Entidad organizadora: The British Institute of Radiology (BIR) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Londres, Reino Unido
Dario Herrán de la Gala; Otros; Pablo Menéndez Fernández-Miranda.

40 **Título del trabajo:** When the duodenum reaches a crossroad: Wilkie's syndrome
Nombre del congreso: BIR Annual Congress 2020
Autor de correspondencia: No
Ciudad de celebración: Londres, Reino Unido
Fecha de celebración: 04/11/2020
Fecha de finalización: 06/11/2020
Entidad organizadora: The British Institute of Radiology (BIR) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Londres, Reino Unido
Dario Herrán de la Gala; Otros; Pablo Menéndez Fernández-Miranda.

41 **Título del trabajo:** Endovascular treatment of the superior mesenteric artery
Nombre del congreso: European Congress of Radiology (ECR) 2020
Autor de correspondencia: No
Ciudad de celebración: Viena, Austria
Fecha de celebración: 15/07/2020
Fecha de finalización: 19/07/2020
Entidad organizadora: European Society of Radiology (ESR) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Viena, Austria
Carmen González-Carreró Sixto; Otros; Pablo Menéndez Fernández-Miranda.

42 **Título del trabajo:** Radiologists role in the diagnosis, management and prognosis of gastrointestinal volvulus
Nombre del congreso: European Congress of Radiology (ECR) 2020
Autor de correspondencia: No
Ciudad de celebración: Viena, Austria
Fecha de celebración: 15/07/2020

Fecha de finalización: 19/07/2020
Entidad organizadora: European Society of Radiology (ESR) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Viena, Austria
Carmen González-Carreró Sixto; Otros; Pablo Menéndez Fernández-Miranda.

43 **Título del trabajo:** Shoulder pain: the role of dynamic ultrasound and its correlation with the physical examination
Nombre del congreso: European Congress of Radiology (ECR) 2020
Autor de correspondencia: No
Ciudad de celebración: Viena, Austria
Fecha de celebración: 15/07/2020
Fecha de finalización: 19/07/2020
Entidad organizadora: European Society of Radiology (ESR) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Viena, Austria
Amaia Pérez del Barrio; Pablo Sanz Bellón; Pablo Menéndez Fernández-Miranda.

44 **Título del trabajo:** A comprehensive guide to understand Convolutional Neural Networks: the main Artificial Intelligence architectures for medical imaging
Nombre del congreso: European Congress of Radiology (ECR) 2020
Autor de correspondencia: Si
Ciudad de celebración: Viena, Austria
Fecha de celebración: 15/07/2020
Fecha de finalización: 19/07/2020
Entidad organizadora: European Society of Radiology (ESR) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Viena, Austria
Pablo Menéndez Fernández-Miranda.

45 **Título del trabajo:** Complications and diagnostic effectiveness of transthoracic lung core-needle biopsy: a retrospective observational study
Nombre del congreso: European Congress of Radiology (ECR) 2020
Autor de correspondencia: Si
Ciudad de celebración: Viena, Austria
Fecha de celebración: 15/07/2020
Fecha de finalización: 19/07/2020
Entidad organizadora: European Society of Radiology (ESR) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Viena, Austria
Pablo Menéndez Fernández-Miranda.

46 **Título del trabajo:** Diagnosis of mucoid degeneration of the anterior cruciate ligament. Is it relevant?
Nombre del congreso: European Congress of Radiology (ECR) 2020
Autor de correspondencia: Si
Ciudad de celebración: Viena, Austria
Fecha de celebración: 15/07/2020
Fecha de finalización: 19/07/2020
Entidad organizadora: European Society of Radiology (ESR) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Viena, Austria
Pablo Menéndez Fernández-Miranda.

47 **Título del trabajo:** Dosimetric information system of Cantabria (DISCA): Integration of the dose management system in the cantabrian health service
Nombre del congreso: European Congress of Radiology (ECR) 2020
Autor de correspondencia: No
Ciudad de celebración: Viena, Austria
Fecha de celebración: 15/07/2020
Fecha de finalización: 19/07/2020
Entidad organizadora: European Society of Radiology (ESR) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Viena, Austria
Pablo Sanz Bellón; Amaia Pérez del Barrio; Pablo Menéndez Fernández-Miranda.

48 **Título del trabajo:** Encrusted pyelitis: an important cause of acute renal failure with a clearly increasing incidence
Nombre del congreso: European Congress of Radiology (ECR) 2020
Autor de correspondencia: Si
Ciudad de celebración: Viena, Austria
Fecha de celebración: 15/07/2020
Fecha de finalización: 19/07/2020
Entidad organizadora: European Society of Radiology (ESR) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Viena, Austria
Pablo Menéndez Fernández-Miranda.

49 **Título del trabajo:** How Business Intelligence will revolutionise radiation protection in radiology departments
Nombre del congreso: European Congress of Radiology (ECR) 2020
Autor de correspondencia: Si
Ciudad de celebración: Viena, Austria
Fecha de celebración: 15/07/2020
Fecha de finalización: 19/07/2020
Entidad organizadora: European Society of Radiology (ESR) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Viena, Austria
Pablo Menéndez Fernández-Miranda.

50 **Título del trabajo:** Liver cystic lesions: beyond the abscess
Nombre del congreso: European Congress of Radiology (ECR) 2020
Autor de correspondencia: No
Ciudad de celebración: Viena, Austria
Fecha de celebración: 15/07/2020
Fecha de finalización: 19/07/2020
Entidad organizadora: European Society of Radiology (ESR) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Viena, Austria
Amaia Pérez del Barrio; Pablo Sanz Bellón; Pablo Menéndez Fernández-Miranda.

51 **Título del trabajo:** Micotic aneurisms: Its possible complications in the human spine
Nombre del congreso: European Congress of Radiology (ECR) 2020
Autor de correspondencia: No
Ciudad de celebración: Viena, Austria

Fecha de celebración: 15/07/2020
Fecha de finalización: 19/07/2020
Entidad organizadora: European Society of Radiology (ESR) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Viena, Austria
Pablo Sanz Bellón; Amaia Pérez del Barrio; Pablo Menéndez Fernández-Miranda.

52 **Título del trabajo:** Pediatric enterovirus encephalitis: A serious pathology you have to be prepared for as a radiologist
Nombre del congreso: European Congress of Radiology (ECR) 2020
Autor de correspondencia: No
Ciudad de celebración: Viena, Austria
Fecha de celebración: 15/07/2020
Fecha de finalización: 19/07/2020
Entidad organizadora: European Society of Radiology (ESR) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Viena, Austria
Pablo Sanz Bellón; Pablo Menéndez Fernández-Miranda.

53 **Título del trabajo:** Radiologic Evaluation of Childhood Asymmetrical Labium Majus Enlargement (CALME)
Nombre del congreso: European Congress of Radiology (ECR) 2020
Autor de correspondencia: No
Ciudad de celebración: Viena, Austria
Fecha de celebración: 15/07/2020
Fecha de finalización: 19/07/2020
Entidad organizadora: European Society of Radiology (ESR) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Viena, Austria
Amaia Pérez del Barrio; Otros; Pablo Menéndez Fernández-Miranda.

54 **Título del trabajo:** Radiological keys for a correct assessment of lung transplant complications
Nombre del congreso: European Congress of Radiology (ECR) 2020
Autor de correspondencia: Si
Ciudad de celebración: Viena, Austria
Fecha de celebración: 15/07/2020
Fecha de finalización: 19/07/2020
Entidad organizadora: European Society of Radiology (ESR) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Viena, Austria
Pablo Menéndez Fernández-Miranda.

55 **Título del trabajo:** Radiologists in the regulation of the responsibilities of artificial intelligence applied to radiodiagnosis: Let's make it real
Nombre del congreso: European Congress of Radiology (ECR) 2020
Autor de correspondencia: No
Ciudad de celebración: Viena, Austria
Fecha de celebración: 15/07/2020
Fecha de finalización: 19/07/2020
Entidad organizadora: European Society of Radiology (ESR) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Viena, Austria
Pablo Sanz Bellón; Amaia Pérez del Barrio; Pablo Menéndez Fernández-Miranda.

56 **Título del trabajo:** Stroke Code: How to avoid missing an Artery of Percheron Infarction
Nombre del congreso: European Congress of Radiology (ECR) 2020
Autor de correspondencia: Si
Ciudad de celebración: Viena, Austria
Fecha de celebración: 15/07/2020
Fecha de finalización: 19/07/2020
Entidad organizadora: European Society of Radiology (ESR) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Viena, Austria
Pablo Menéndez Fernández-Miranda.

57 **Título del trabajo:** The ethical and legal issues arising from the implantation of AI systems in the radiological practice
Nombre del congreso: European Congress of Radiology (ECR) 2020
Autor de correspondencia: Si
Ciudad de celebración: Viena, Austria
Fecha de celebración: 15/07/2020
Fecha de finalización: 19/07/2020
Entidad organizadora: European Society of Radiology (ESR) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Viena, Austria
Pablo Menéndez Fernández-Miranda.

58 **Título del trabajo:** The study description in dose information systems: the problem and the proposed solution
Nombre del congreso: European Congress of Radiology (ECR) 2020
Autor de correspondencia: No
Ciudad de celebración: Viena, Austria
Fecha de celebración: 15/07/2020
Fecha de finalización: 19/07/2020
Entidad organizadora: European Society of Radiology (ESR) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Viena, Austria
Amaia Pérez del Barrio; Pablo Menéndez Fernández-Miranda.

59 **Título del trabajo:** Upper urinary tract infections and how radiologists can help in their management
Nombre del congreso: European Congress of Radiology (ECR) 2020
Autor de correspondencia: No
Ciudad de celebración: Viena, Austria
Fecha de celebración: 15/07/2020
Fecha de finalización: 19/07/2020
Entidad organizadora: European Society of Radiology (ESR) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: Viena, Austria
Carmen González-Carreró Sixto; Otros; Pablo Menéndez Fernández-Miranda.

60 **Título del trabajo:** La importancia de no subestimar síntomas leves en pacientes con buen estado general.
Nombre del congreso: XV Congreso Autonómico de Cantabria. SEMERGEN Cantabria 2020.
Ciudad de celebración: Santander, Cantabria, España
Fecha de celebración: 30/01/2020

Fecha de finalización: 01/02/2020
Entidad organizadora: Sociedad Española de Médicos de Atención Primaria (SEMERGEN). **Tipo de entidad:** Asociaciones y Agrupaciones

61 **Título del trabajo:** Atypical Bacterial Infections of the Central Nervous System: An Unnkonw Treath.
Nombre del congreso: 105th Scientific Assembly and Annual Meeting of the Radiological Society of North America (2019).
Ciudad de celebración: Chicago, Estados Unidos de América
Fecha de celebración: 01/12/2019
Fecha de finalización: 12/12/2019
Entidad organizadora: Radiological Society of North America (RSNA).
Ciudad entidad organizadora: Estados Unidos de América

62 **Título del trabajo:** Encefalitis por Enterovirus en pediatría: una patología grave para la que hay que estar preparados como radiólogos.
Nombre del congreso: XLVIII Reunión Anual Sociedad Española de Neurorradiología
Ciudad de celebración: Santander, Cantabria, España
Fecha de celebración: 24/10/2019
Fecha de finalización: 26/10/2019
Entidad organizadora: Sociedad Española de Neurorradiología **Tipo de entidad:** Asociaciones y Agrupaciones

63 **Título del trabajo:** El cráneo postquirúrgico: cómo sobrevivir en las guardias.
Nombre del congreso: XLVIII Reunión Anual Sociedad Española de Neurorradiología
Ciudad de celebración: Santander, Cantabria, España
Fecha de celebración: 24/10/2019
Fecha de finalización: 26/10/2019
Entidad organizadora: Sociedad Española de Neurorradiología **Tipo de entidad:** Asociaciones y Agrupaciones

64 **Título del trabajo:** Hallazgos de imagen de las complicaciones neurológicas en los pacientes trasplantados de pulmón: revisión de una cohorte de 8 años.
Nombre del congreso: XLVIII Reunión Anual Sociedad Española de Neurorradiología
Ciudad de celebración: Santander, Cantabria, España
Fecha de celebración: 24/10/2019
Fecha de finalización: 26/10/2019
Entidad organizadora: Sociedad Española de Neurorradiología **Tipo de entidad:** Asociaciones y Agrupaciones

65 **Título del trabajo:** Ictus de la Arteria de Percheron: un diagnóstico vital que puede pasar desapercibido en el TC multimodal.
Nombre del congreso: XLVIII Reunión Anual Sociedad Española de Neurorradiología
Ciudad de celebración: Santander, Cantabria, España
Fecha de celebración: 24/10/2019
Fecha de finalización: 26/10/2019
Entidad organizadora: Sociedad Española de Neurorradiología **Tipo de entidad:** Asociaciones y Agrupaciones
Pablo Menéndez Fernández-Miranda.

66 **Título del trabajo:** Manifestaciones neurológicas del Lupus Eritematoso Sistémico.
Nombre del congreso: XLVIII Reunión Anual Sociedad Española de Neurorradiología
Ciudad de celebración: Santander, Cantabria, España

Fecha de celebración: 24/10/2019
Fecha de finalización: 26/10/2019
Entidad organizadora: Sociedad Española de Neurorradiología **Tipo de entidad:** Asociaciones y Agrupaciones

67 **Título del trabajo:** Patrones de señal de M.O. en la RM.
Nombre del congreso: XLVIII Reunión Anual Sociedad Española de Neurorradiología
Ciudad de celebración: Santander, Cantabria, España
Fecha de celebración: 24/10/2019
Fecha de finalización: 26/10/2019
Entidad organizadora: Sociedad Española de Neurorradiología **Tipo de entidad:** Asociaciones y Agrupaciones

68 **Título del trabajo:** Convolutional Neural Networks: the state-of-the-art of Artificial Intelligence for medical imaging.
Nombre del congreso: EuSoMII ANNUAL MEETING - “Medical Imaging Informatics – AI, Clinical Applications and more”
Ciudad de celebración: Valencia, Comunidad Valenciana, España
Fecha de celebración: 18/10/2019
Fecha de finalización: 19/10/2019
Entidad organizadora: European Society of Medical Imaging Informatics (EuSoMII). **Tipo de entidad:** Asociaciones y Agrupaciones
Pablo Menéndez Fernández-Miranda.

69 **Título del trabajo:** What kind of legal and ethical issues will arise from using AI systems in the medical practice?
Nombre del congreso: EuSoMII ANNUAL MEETING - “Medical Imaging Informatics – AI, Clinical Applications and more”
Ciudad de celebración: Valencia, Comunidad Valenciana, España
Fecha de celebración: 18/10/2019
Fecha de finalización: 19/10/2019
Entidad organizadora: European Society of Medical Imaging Informatics (EuSoMII). **Tipo de entidad:** Asociaciones y Agrupaciones
Pablo Menéndez Fernández-Miranda.

70 **Título del trabajo:** Trampas en la radiografía de tórax en las que un médico de atención primaria no deberia caer.
Nombre del congreso: 41º Congreso Nacional SEMERGEN 2019
Ciudad de celebración: Gijón, Principado de Asturias, España
Fecha de celebración: 16/10/2019
Fecha de finalización: 19/10/2019
Entidad organizadora: Sociedad Española de Médicos de Atención Primaria (SEMERGEN) **Tipo de entidad:** Asociaciones y Agrupaciones

71 **Título del trabajo:** Hallazgos radiológicos de las principales malformaciones vasculares cerebrales: malformación arteriovenosa, telangiectasia capilar, angioma cavernoso y angioma venoso
Nombre del congreso: X Congreso CENORA 2019
Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones

Pablo Menéndez Fernández-Miranda.

72 **Título del trabajo:** Patología pulmonar en el trasplantado de pulmón
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si
Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: España
Pablo Menéndez Fernández-Miranda.

73 **Título del trabajo:** Aneurismas de las arterias coronarias una patología todavía poco conocida.
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si
Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: España

74 **Título del trabajo:** Aneurismas micóticos: sus posibles complicaciones.
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si
Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: España

75 **Título del trabajo:** Angiomiolipoma renal con trombosis venosa tumoral hacia la VCI.
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si
Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: España

76 **Título del trabajo:** Aspectos ético-legales de la aplicación de sistemas de inteligencia artificial a la radiología.
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si
Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Tipo de entidad: Asociaciones y Agrupaciones

Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA)
Ciudad entidad organizadora: España

77 **Título del trabajo:** CAD-RAS nuestra experiencia.
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si
Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: España

78 **Título del trabajo:** Calcificación de la vía urinaria alta y fracaso renal agudo en pacientes con antecedentes de manipulación urológica
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si
Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: España
Pablo Menéndez Fernández-Miranda.

79 **Título del trabajo:** Colecciones hepáticas más allá del absceso.
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si
Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: España

80 **Título del trabajo:** Complicaciones y rentabilidad diagnóstica de la biopsia pulmonar con BAG
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si
Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: España
Pablo Menéndez Fernández-Miranda.

81 **Título del trabajo:** Cómo hacer e interpretar una arteriografía de miembros inferiores.
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si
Ciudad de celebración: Logroño, La Rioja, España

Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: España

82 **Título del trabajo:** Descripción de una muestra de 24 pacientes sometidos a ablación de nódulos tiroideos benignos guiada por ecografía en el Hospital Universitario Marqués de Valdecilla
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si
Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: España
Pablo Menéndez Fernández-Miranda.

83 **Título del trabajo:** Dolor en el cuadrante superior derecho más allá de la colecistitis aguda.
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si
Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: España

84 **Título del trabajo:** Dolor en el hombro el papel de la ecografía dinámica y su correlación con la exploración física
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si
Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: España

85 **Título del trabajo:** El papel del radiólogo en el síndrome de May-Thurner del diagnóstico al tratamiento.
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si
Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: España

86 **Título del trabajo:** Espacios del cuello: la anatomía para entender la patología.
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si

Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: España

87 **Título del trabajo:** Evaluación radiológica de la obstrucción de la unión pieloureteral
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si
Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: España

88 **Título del trabajo:** Evaluación radiológica del traumatismo facial guía para el residente.
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si
Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: España

89 **Título del trabajo:** Neuroimagen en maltrato infantil, una gran responsabilidad.
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si
Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: España

90 **Título del trabajo:** Patología aórtica en la urgencia radiológica
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si
Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: España
Pablo Menéndez Fernández-Miranda.

91 **Título del trabajo:** Patrones radiológicos de las infecciones pulmonares.
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si

Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: España
Pablo Menéndez Fernández-Miranda.

92 **Título del trabajo:** Pielonefritis aguda complicada. Cómo podemos ayudar en su manejo clínico.
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si
Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: España

93 **Título del trabajo:** Redes Convolucionales: state of art en Inteligencia Artificial para imagen médica
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si
Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: España
Pablo Menéndez Fernández-Miranda.

94 **Título del trabajo:** Sistema de Información Dosimétrica de Cantabria (DISCA) integración del sistema de gestión de dosis Radimetrics en el Servicio Cántabro de Salud (SCS).
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si
Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: España

95 **Título del trabajo:** Tejido esplénico de localización ectópica. Una entidad benigna que puede ser origen de graves errores diagnósticos
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si
Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA)
Ciudad entidad organizadora: España

96 **Título del trabajo:** Tumor mucinoso intraductal vs pancreatitis crónica.
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si
Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: España

97 **Título del trabajo:** Vólvulos del tracto gastrointestinal. Un diagnóstico radiológico.
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si
Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: España

98 **Título del trabajo:** Ácido gadoxético. Conclusiones tras 10 años introducción en el servicio.
Nombre del congreso: X Congreso CENORA 2019
Autor de correspondencia: Si
Ciudad de celebración: Logroño, La Rioja, España
Fecha de celebración: 04/10/2019
Fecha de finalización: 05/10/2019
Entidad organizadora: Sociedad Centro-Norte de Radiología (CENORA) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad organizadora: España

99 **Título del trabajo:** ARTERY OF PERCHERON INFARCTION
Nombre del congreso: SERAU-ESER Annual Scientific Meeting 2019
Ciudad de celebración: Sevilla, Andalucía, España
Fecha de celebración: 16/05/2019
Fecha de finalización: 17/05/2019
Entidad organizadora: Sociedad Española de Radiología de Urgencias (SERAU) y European Society of Emergency Radiology (ESER)
Pablo Menéndez Fernández-Miranda; Pablo Sanz Bellón; Amaia Pérez del Barrio.

100 **Título del trabajo:** Encrusted pyelitis: a cause of acute renal failure which is commonly missed diagnosis.
Nombre del congreso: SERAU-ESER Annual Scientific Meeting 2019
Ciudad de celebración: Sevilla, Andalucía, España
Fecha de celebración: 16/05/2019
Fecha de finalización: 17/05/2019
Entidad organizadora: Sociedad Española de Radiología de Urgencias (SERAU) y European Society of Emergency Radiology (ESER)
Pablo Menéndez Fernández-Miranda; Amaia Pérez del Barrio; Pablo Sanz Bellón.

101 **Título del trabajo:** Enterovirus encephalitis in the pediatric age, a serious pathology and more frequent than it seems to be
Nombre del congreso: SERAU-ESER Annual Scientific Meeting 2019

Ciudad de celebración: Sevilla, Andalucía, España
Fecha de celebración: 16/05/2019
Fecha de finalización: 17/05/2019
Entidad organizadora: Sociedad Española de Radiología de Urgencias (SERAU) y European Society of Emergency Radiology (ESER)
Pablo Sanz Bellón; Pablo Menéndez Fernández-Miranda; Amaia Pérez del Barrio.

102 **Título del trabajo:** Pyogenic Liver Abscess: what the radiologist should know
Nombre del congreso: SERAU-ESER Annual Scientific Meeting 2019
Ciudad de celebración: Sevilla, Andalucía, España
Fecha de celebración: 16/05/2019
Fecha de finalización: 17/05/2019
Entidad organizadora: Sociedad Española de Radiología de Urgencias (SERAU) y European Society of Emergency Radiology (ESER)
Amaia Pérez del Barrio; Pablo Menéndez Fernández-Miranda; Pablo Sanz Bellón.

103 **Título del trabajo:** El absceso hepático: claves para su diagnóstico diferencial
Nombre del congreso: II Jornadas SERAM-ARRS
Ciudad de celebración: Valladolid, Castilla y León, España
Fecha de celebración: 04/04/2019
Fecha de finalización: 06/04/2019
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM) y American Roentgen Ray Society (ARRS)
Amaia Pérez del Barrio; Pablo Menéndez Fernández-Miranda; Pablo Sanz Bellón.

104 **Título del trabajo:** Encefalitis por Enterovirus en la edad pediátrica, una patología grave y más frecuente de lo que parece
Nombre del congreso: II Jornadas SERAM-ARRS
Ciudad de celebración: Valladolid, Castilla y León, España
Fecha de celebración: 04/04/2019
Fecha de finalización: 06/04/2019
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM) y American Roentgen Ray Society (ARRS) **Tipo de entidad:** Asociaciones y Agrupaciones
Pablo Sanz Bellón; Pablo Menéndez Fernández-Miranda; Amaia Pérez del Barrio.

105 **Título del trabajo:** Uropatía incrustante: un cuadro causado por Corynebacterium urealyticum cada vez más frecuente
Nombre del congreso: II Jornadas SERAM-ARRS
Ciudad de celebración: Valladolid, Castilla y León, España
Fecha de celebración: 04/04/2019
Fecha de finalización: 06/04/2019
Entidad organizadora: Sociedad Española de Radiología Médica (SERAM) y American Roentgen Ray Society (ARRS)
Pablo Menéndez Fernández-Miranda; Amaia Pérez del Barrio; Pablo Sanz Bellón.

Gestión de I+D+i y participación en comités científicos

Comités científicos, técnicos y/o asesores

1 **Título del comité:** Club Bibliográfico de la Sociedad Española de Radiología Médica (SERAM)
Primaria (Cód. Unesco): 320111 - Radiología

2 **Título del comité:** EuSoMII Young Committee
Primaria (Cód. Unesco): 320111 - Radiología
Secundaria (Cód. Unesco): 331400 - Tecnología médica
Entidad de afiliación: European Society of Medical Imaging Informatics (EuSoMII) **Tipo de entidad:** Asociaciones y Agrupaciones
Ciudad entidad afiliación: Viena, Austria

Organización de actividades de I+D+i

Título de la actividad: Desarrollo de una aplicación de entrenamiento en el diagnóstico de COVID-19 en radiografías de tórax
Tipo de actividad: Innovación tecnológica **Ámbito geográfico:** Internacional no UE
Ciudad entidad convocante: Santander, Cantabria, España
Fecha de inicio: 11/04/2020

Otros méritos

Estancias en centros de I+D+i públicos o privados

1 **Entidad de realización:** Netherlands Cancer Institute - NKI
Ciudad entidad realización: Amsterdam, Zuid-Holland, Holanda
Fecha de inicio-fin: 31/01/2022 - 31/03/2022 **Duración:** 2 meses
Objetivos de la estancia: Doctorado/a

2 **Entidad de realización:** Leiden University Medical Centre
Ciudad entidad realización: Leiden, Zuid-Holland, Holanda
Objetivos de la estancia: Invitado/a
Tareas contrastables: Revisión de las lineas de investigación y de los métodos científicos, clínicos y asistenciales del centro

Sociedades científicas y asociaciones profesionales

1 **Nombre de la sociedad:** Sociedad Española de Radiología Musculoesquelética (SERME)
Ciudad entidad afiliación: España
Fecha de inicio-fin: 01/12/2018 - 01/01/2019

2 **Nombre de la sociedad:** European Society of Medical Imaging Informatics (EuSoMII)
Ciudad entidad afiliación: Austria
Fecha de inicio: 01/08/2020

3 **Nombre de la sociedad:** European Society of Radiology (ESR)
Ciudad entidad afiliación: Austria
Fecha de inicio: 01/06/2018

4 **Nombre de la sociedad:** Radiological Society of North America (RSNA)
Ciudad entidad afiliación: Austria
Fecha de inicio: 01/06/2018

5 **Nombre de la sociedad:** Sociedad Española de Radiología Médica (SERAM)
Ciudad entidad afiliación: España
Fecha de inicio: 05/04/2018

9.5. Anexo 5: Difusión de los resultados

9.5.1. Artículo científico en la revista *Journal of Digital Imaging*: *Developing a Training Web Application for Improving the COVID-19 Diagnostic Accuracy on Chest X-ray*

Journal of Digital Imaging
https://doi.org/10.1007/s10278-021-00424-7

Developing a Training Web Application for Improving the COVID-19 Diagnostic Accuracy on Chest X-ray

P. Menéndez Fernández-Miranda[1,6] · P. Sanz Bellón[1,6] · A. Pérez del Barrio[1,6] · L. Lloret Iglesias[2] · P. Solís García[3] · F. Aguilar-Gómez[2] · D. Rodríguez González[2] · J. A. Vega[4,5]

Received: 19 May 2020 / Revised: 6 November 2020 / Accepted: 11 January 2021

Abstract

In December 2019, a new coronavirus known as 2019-nCoV emerged in Wuhan, China. The virus has spread globally and the infection was declared pandemic in March 2020. Although most cases of coronavirus disease 2019 (COVID-19) are mild, some of them rapidly develop acute respiratory distress syndrome. In the clinical management, chest X-rays (CXR) are essential, but the evaluation of COVID-19 CXR could be a challenge. In this context, we developed *COVID-19 TRAINING*, a free Web application for training on the evaluation of COVID-19 CXR. The application included 196 CXR belonging to three categories: *non-pathological*, *pathological compatible with COVID-19*, and *pathological non-compatible with COVID-19*. On the training screen, images were shown to the users and they chose a diagnosis among those three possibilities. At any time, users could finish the training session and be evaluated through the estimation of their diagnostic accuracy values: sensitivity, specificity, predictive values, and global accuracy. Images were hand-labeled by four thoracic radiologists. Average values for sensitivity, specificity, and global accuracy were .72, .64, and .68. Users who achieved better sensitivity registered less specificity ($p < .0001$) and those with higher specificity decreased their sensitivity ($p < .0001$). Users who sent more answers achieved better accuracy ($p = .0002$). The application *COVID-19 TRAINING* provides a revolutionary tool to learn the necessary skills to evaluate COVID-19 on CXR. Diagnosis training applications could provide a new original manner of evaluation for medical professionals based on their diagnostic accuracy values, and an efficient method to collect valuable data for research purposes.

Keywords Chest X-ray · COVID-19 · Diagnostic accuracy values · Medical application · Medical education · Training on diagnosis

Background

In December 2019, a new coronavirus named 2019-nCoV, also known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was isolated in the airway epithelial cells of a cluster of patients with pneumonia of unknown cause in Wuhan, China [1]. Since then, the infection caused by 2019-nCoV has rapidly spread globally [2] affecting more than 4 million people in about 215 countries with more than 280.000 reported deaths to date [3]. On the 11th of March 2020, the World Health Organization (WHO) declared the novel coronavirus outbreak a global pandemic [4].

Most patients with coronavirus disease 2019 (COVID-19) are mild cases, whose symptoms are usually self-limiting and recover within 2 weeks [5]. However, others progress rapidly and develop acute respiratory distress syndrome (ARDS) and septic shock, eventually resulting in multiple

✉ P. Menéndez Fernández-Miranda
pablomenendezfernandezmiranda@gmail.com

✉ J. A. Vega
javega@uniovi.es

1 Departamento de Radiología, Hospital Universitario "Marqués de Valdecilla", Santander, Spain

2 Grupo de Computación Avanzada y e-Ciencia, Instituto de Física de Cantabria, (IFCA), Consejo Superior de Investigaciones Científicas (CSIC), Santander, Spain

3 Wroclaw, Poland

4 Departamento de Morfología y Biología Celular, Universidad de Oviedo, Oviedo, Spain

5 Facultad de Ciencias de La Salud, Universidad Autónoma de Chile, Santiago, Chile

6 Departamento Morfología y Biología Celular, Universidad de Oviedo, Oviedo, Spain

Published online: 08 March 2021

organ failure [6]. At the time of this writing, fever and cough were the main clinical manifestations, followed by dyspnea, myalgia or weakness, and chest tightness [7]. Furthermore, a variable percentage of patients report decreased smell function or even anosmia and dysgeusia [8]. Currently, COVID-19 is being diagnosed using molecular detection methods such as the reverse-transcriptase polymerase chain reaction (RT-PCR) test, regarded as the standard of reference [9–11]. Nevertheless, despite RT-PCR COVID-19 testing has a specificity of 100% [11], it shows a potentially high false negative rate deteriorating the sensitivity, and then it is not a definitive diagnostic method [10, 12]. Other techniques for the detection of COVID-19 are being used to increase the efficiency in the diagnosis, especially medical imaging modalities such as chest X-rays (CXR) and computed tomography (CT) scans [9, 10, 13].

In patients with a high level of clinical suspicion of COVID-19 and negative RT-PCR, CXR can be key to identify false negatives to RT-PCR COVID-19 testing as CXR abnormalities may appear before eventually testing positive on RT-PCR [13]. Moreover, in populations around the world with limited access to reliable real-time molecular diagnostic methods, the utilization of CXR for early disease detection also plays a capital role [14]. Additionally, imaging is also critical in assessing the severity and disease progression in a COVID-19 infection [15]. Thus, physicians evaluating COVID-19 on medical images should be aware of the imaging manifestations and radiological features that have been well-described [16].

However, the interpretation of chest radiographs is a challenging task, requiring experience and expertise [17–19]. The American College of Radiology (ACR) recommends that qualified radiologists be available to interpret all radiographs obtained in the Emergency Departments (ED) [20], and previous studies have reported suboptimal performance in the interpretation of CXR by ED physicians compared with expert radiologists [21–23]. Importantly, the number of chest radiographs per ED visit have increased during the last decades [24] which is a practical limitation with regard to the full-time availability of expert radiologists [25]. This situation has deteriorated since the pandemic outbreak appeared in the worldwide clinical scenario.

In the context of COVID-19 pandemic and the high demand of CXR reporting, many CXR are being interpreted by nonexpert physicians who have been forced to acquire the competence of detecting and evaluating the radiological features of COVID-19. Being aware of it, we have developed a free Web application to ease the learning process of interpreting COVID-19 CXR for physicians, residents, students, and anybody else interested in acquiring this competence. As far as we know, this application called *COVID-19 TRAINING* (https://xray.covid.ifca.es/en) is the first available tool that allows a single user to calculate their diagnostic accuracy values: sensitivity, specificity, positive predictive value, negative predictive value, and global accuracy. The *COVID-19 TRAINING* application can set a precedent for further applications because this is the first time that a physician is considered as a single diagnostic tool by themselves. Usually, diagnostic accuracy values are calculated for diagnostic techniques or collectives of physicians but not for a single professional. In our opinion, providing this kind of tool to physicians can help them to evaluate performance by setting specific metrics, and to follow up their progression in their diagnostic efficiency.

Methods

Patients Recruitment

We collected 196 CXR in our Institution belonging to 33 females (35.48%) and 60 males (64.51%). The average age of these patients was 61.43 years with a standard deviation (SD) of 17.20 years and a range from 19 to 88 (Table 1).

The chest radiographs were all hand-labeled and classified by four expert thoracic radiologists into 3 categories: *non-pathological*, *pathological compatible with COVID-19*, and *pathological non-compatible with COVID-19*. Classification and inclusion criteria for each category were as follows: (1) *non-pathological:* (a) the CXR was obtained before the appearance of SARS-CoV-2 virus; and (b) four reports made

Table 1 Characteristics of the patients included in the application and number of X-rays for each category

Attribute	Non-pathological	Pathological compatible with COVID-19	Pathological non compatible with COVID-19	Total
Gender, *n* (%)				
Female	9 (42.86)	10 (23.81)	14 (46.67)	33 (35.48)
Male	12 (57.14)	32 (76.19)	16 (53.33)	60 (64.51)
Age, mean ± SD	61.04 ± 16.74	64.64 ± 15.16	54.07 ± 17.69	61.43 ± 17.20
Range	31–84	30–88	19–82	19–88
CXR, *n* (%)	45 (22.96)	121 (61.73)	30 (15.31)	196 (100)

CXR chest X-ray

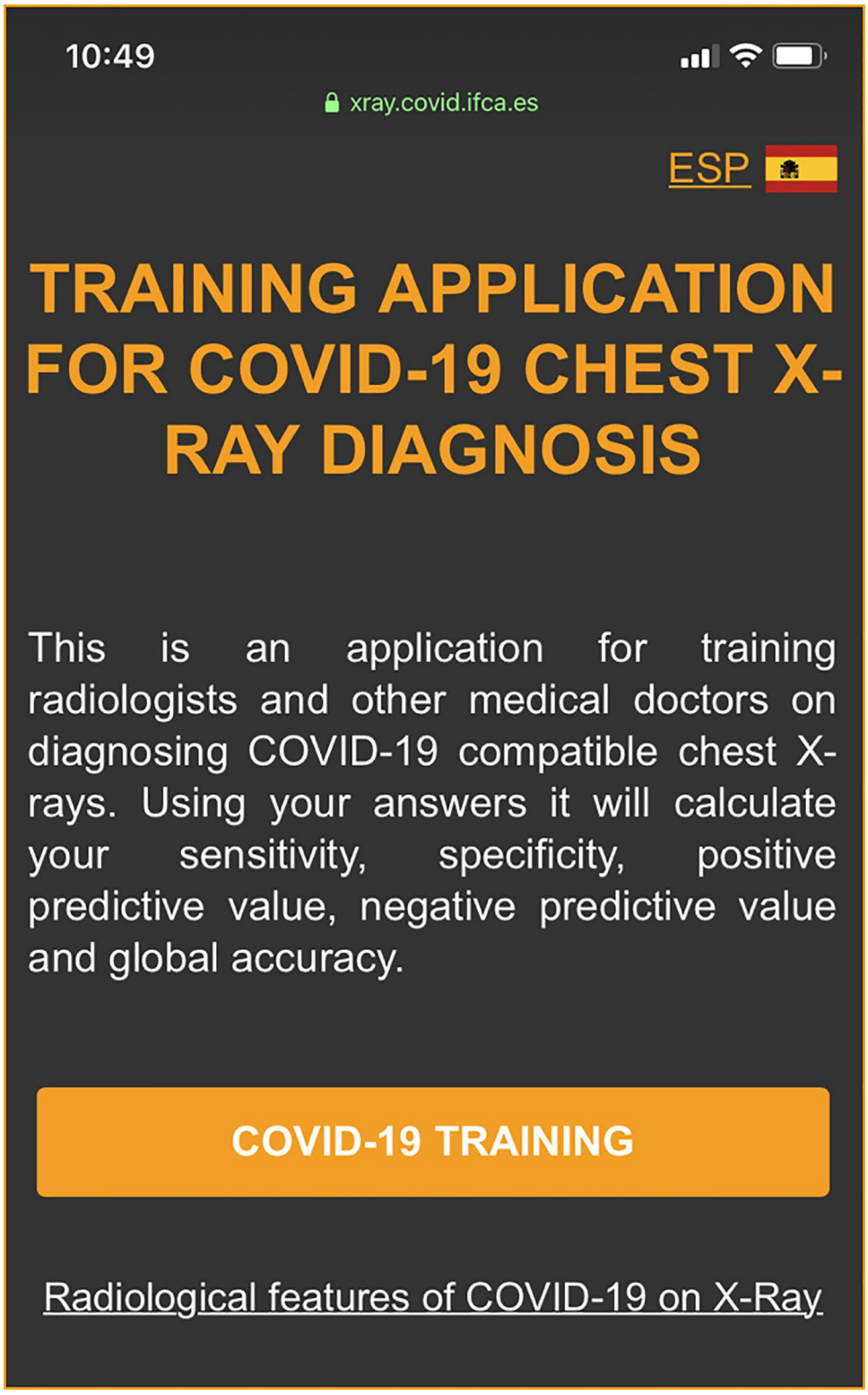


Fig. 1 Welcome page screen, as viewed from a mobile device

by the four expert thoracic radiologists confirm the absence of any pathological findings on the CXR; (2) *pathological compatible with COVID-19*: (a) the CXR has four reports made by the four expert thoracic radiologists indicating pathological findings compatible with COVID-19; (b) the patients have *a subsequent confirmation* of the disease by a positive RT-PCR COVID-19 testing; and (c) progression of the radiological findings reported previously were observed

on further CXR; (3) *pathological non-compatible with COVID-19:* (a) the CXR was obtained before the appearance of SARS-CoV-2 virus; and (b) the CXR has four reports made by the four expert thoracic radiologists confirming pathological findings.

The exclusion criteria for the three categories were as follows: (a) low-quality image, (b) improper alignment of the X-ray tube to the film, and (c) did not meet any of the other inclusion criteria.

The chest radiographs were obtained with different equipment (portable and conventional machines), in different projections (posterior-anterior and anterior–posterior) and with different patient positions (standing and supine decubitus), in order to reproduce the most realistic clinical scenario. Data about patient distribution for each category is shown in Table 1. This distribution tries to guarantee that the application shows the user a proportional number of images from each category with a slight predominance of the category *pathological compatible with COVID-19*, which are the most interesting images for the users. In addition, the application also resembles the real clinical context where females are less frequently infected than males [26].

Data Access and Anonymization

This work was approved by the Ethical Committee of our Institution, and all the data collected to develop the application were fully de-identified before transfer to the development team.

Application Programming and Launching

Subject information and links to the corresponding images are stored in a SQLite database. The application has been implemented using Python 3 [27] for the backend and Flask, a Python framework, for generating the client side or frontend. Some external libraries were used in order to expand the possibilities of the frontend, achieving multilanguage (currently supporting Spanish and English) or more structured styles hierarchy. Some other good practices based on the Clean Code principles are being applied. All the code is publicly available in an open access repository [28]. User management is supported by an Authentication and Authorization Infrastructure based on Open ID Connect standard.

A public beta version was published on the 11th of April 2020, and the 24th of April we launched the final version.

Application Usability and Requirements

The *COVID-19 TRAINING* application was designed to help all doctors, residents, medical students, and anybody else interested in acquiring the skills to recognize the radiological findings of COVID-19 on CXR. It includes CXR with different levels of difficulty so everyone will be able to join. The application is available for free and adapted both for mobile and tablet (Fig. 1) as well as for computer screens (Fig. 2) in Spanish and English languages.

Access to the Application

The access and navigation on the application is shown on a video added in the Appendix 1.

First step: The first screen that the user finds when opening the application is the welcome page (Figs. 1 and 2). It serves as the starting point of the application and shows a headline and a paragraph describing briefly its functionalities. On this interface two buttons are displayed: "*COVID-19 TRAINING*" that provides easy access to the login screen, and "Radiological features of COVID-19 on

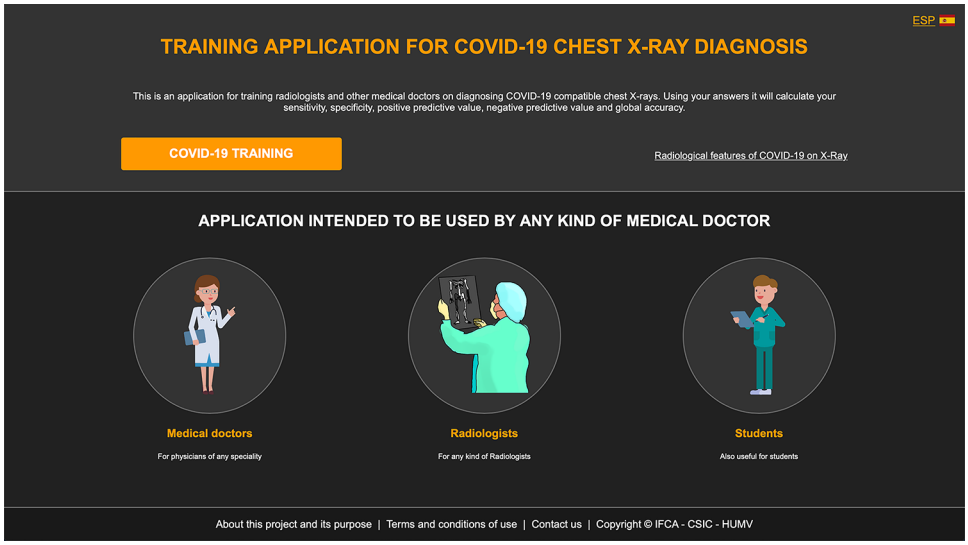


Fig. 2 Welcome page screen, as viewed from a desktop device

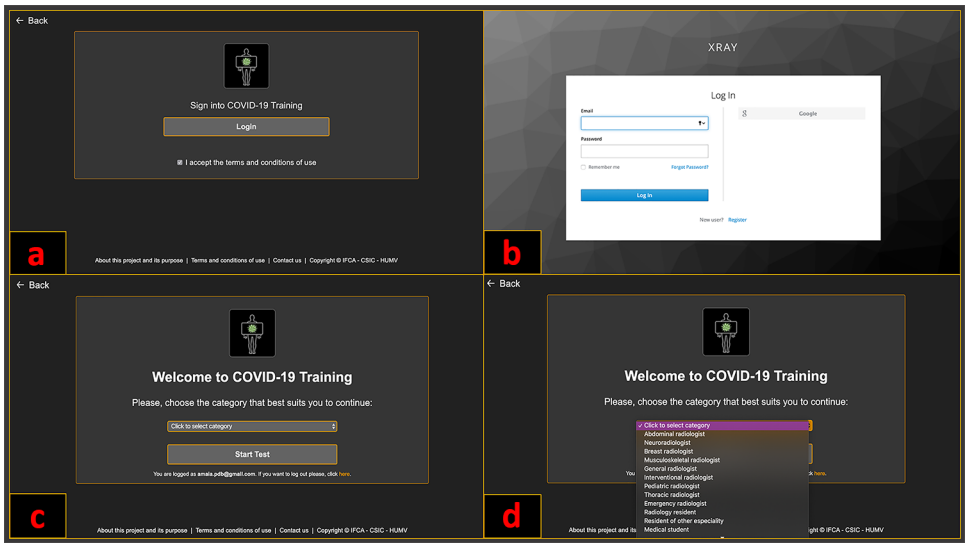


Fig. 3 **a** Login and **b** authentication screens, and **c**, **d** drop-down menu to select the specialty or professional category of the user, as viewed from a desktop device

X-Ray," which opens a summary of the radiological COVID-19 findings that the user should be able to identify on CXR before starting the training.

Second Step: Login screen requires the user to accept the terms and conditions and to press the "Login" button to access the authentication interface (Fig. 3a). Authentication can be easily done either just entering a Google Gmail or creating a new account by pressing on the "Register" button in the lowest part of the screen (Fig. 3b). If the user's preferred option is the second, they will only need to provide their name, surname, email address, and password. An email will be automatically sent with a link to verify the account.

Third Step: Finally, a drop-down menu will allow the user to select their specialty or professional category (Fig. 3c, d). Available options are shown on Table 2.

Training and Obtaining the Diagnostic Accuracy Values

Once the authentication is done and the profile category is selected, a new screen with CXR will be open. Besides the CXR, patient's age and gender are also provided (Fig. 4a).

On the left side of the screen, the user can choose in a drop-down menu one diagnosis among the three possible categories described before: *non-pathological*, *pathological compatible with COVID-19*, and *pathological non-compatible with COVID-19* (Fig. 4b). After selecting the desired option, by pressing the "Next" button, the answer will be registered and a new case will be charged. At any time, the user can finish the training session by pressing the "End Test" button (Fig. 4). Each time the user takes a test, a new set of training images is selected.

A virtual magnifying glass is also implemented. Users simply need to press on the image to summon the magnifying glass, and they will see a zoomed image within its radius, without disturbing the rest of the page (Fig. 5). To remove it, they only need to press on the image again.

The "End Test" button links to the results interface, where the user's sensitivity, specificity, predictive values and global accuracy are provided after each training (Figs. 6 and 7). If the users scroll down on this interface, those questions that they answered incorrectly will be displayed, indicating the given response and the correct answer, so they will learn from their mistakes. *Pathological non compatible with COVID-19* images also include a reference to the real pathology of the patient.

Table 2 Specialties and professional categories available to select on the drop-down menu

Specialty/category
Abdominal radiologist
Neuroradiologist
Breast radiologist
Musculoskeletal radiologist
General radiologist
Interventional radiologist
Pediatric radiologist
Thoracic radiologist
Emergency radiologists
Radiology resident
Resident of other specialty
Medical student
Pulmonology physician
Internal medicine physician
Emergency physician
Intensive care physician
Other kind of physician
Diagnostic Imaging Technician (radiographer)
Other category

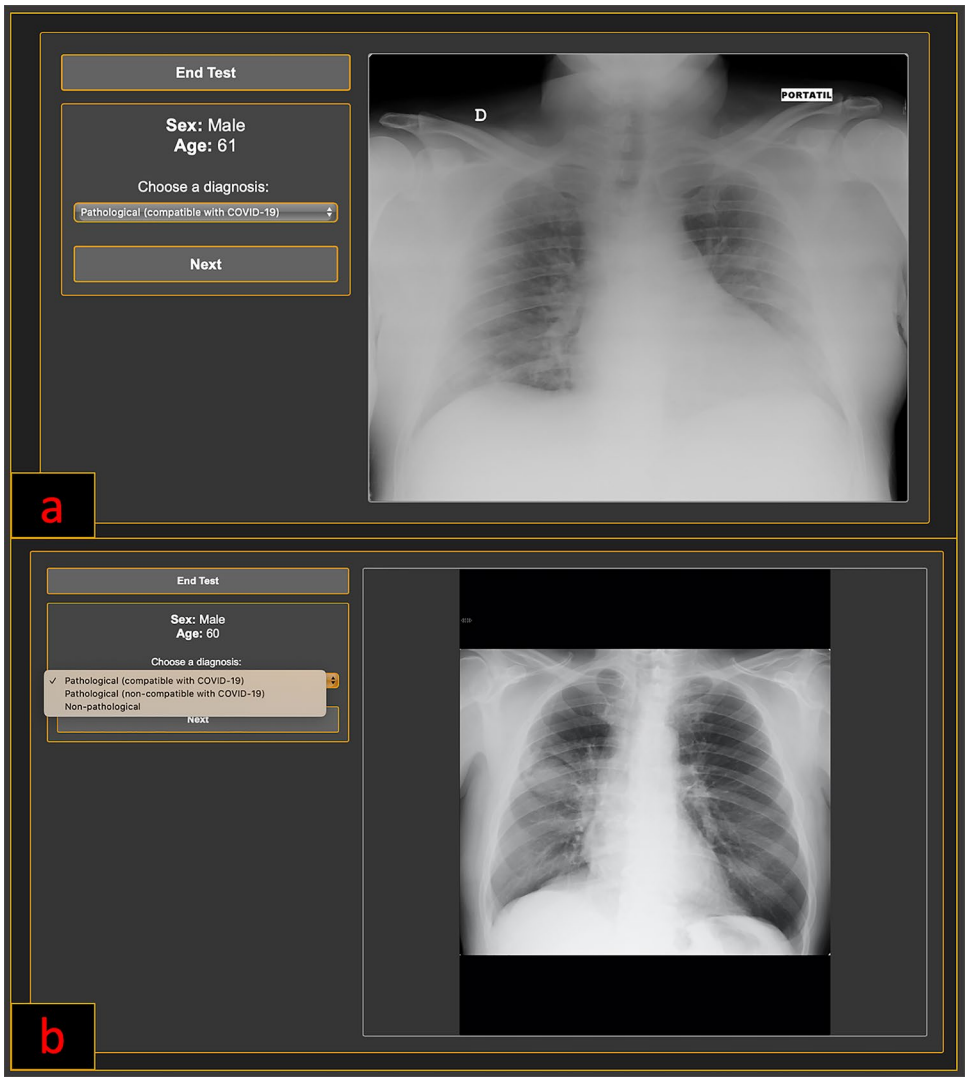


Fig. 4 **a** Training interface and **b** the diagnoses available on the drop-down menu, as viewed from a desktop device

Data Collection and Analysis of the User's Results

The *COVID-19 TRAINING* application allows the user to evaluate themselves through the estimation of their diagnostic accuracy values: sensitivity, specificity, positive predictive value, negative predictive value, and global accuracy. Diagnostic accuracy values are calculated for COVID-19 diagnosis. The application registers the answers of the users and classifies them into the following four conventional categories [29, 30]: (1) true positive (TP): the user classifies properly a CXR which belongs to the category of *pathological compatible with COVID-19*; (2) false positive (FP): the user classifies improperly a CXR which does not belong to the category of *pathological compatible with COVID-19,* in this category; (3) true negative (TN): the user classifies properly a CXR which does not belong to the category of *pathological compatible with COVID-19* in another category; (4) false negative (FN): the user classifies improperly a CXR which belongs to the category of *pathological compatible with COVID-19,* in another category.

The sensitivity represents the user's ability to determine the COVID-19 cases correctly. It accounts for the proportion of true positives in patient-cases: TP/(TP + FN) [29]. The specificity shows the user's capacity to rule out COVID-19 correctly. To estimate specificity, the proportion of true negatives in healthy cases should be calculated: TN/(TN + FP) [29].

Positive predictive value (PPV) defines the probability of having COVID-19 when the user classifies the CXR into the category *pathological compatible with COVID-19*. Therefore, it represents the proportion of COVID-19 patients within the patients with positive CXR for COVID-19 according to the user's criteria: TP/(TP + FP) [30]. By contrast, negative predictive value (NPV) describes the probability of not having COVID-19 when the user does not

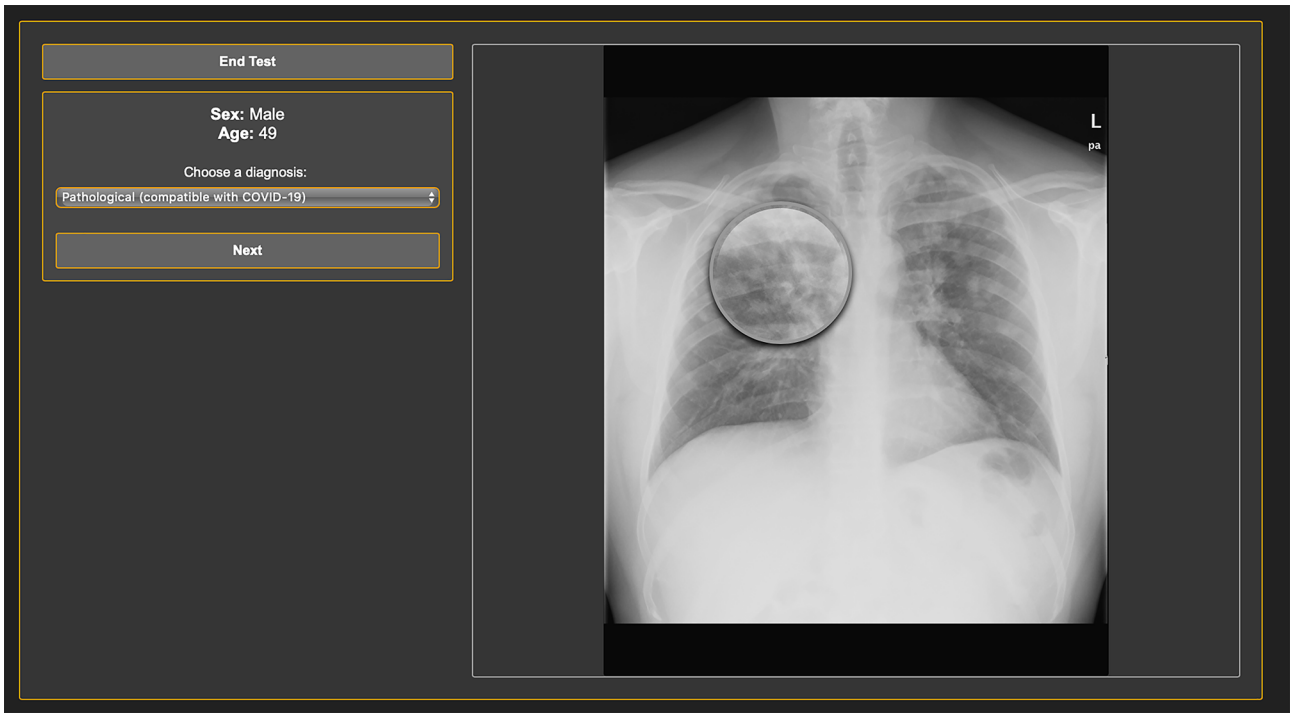


Fig. 5 Virtual magnifying glass, as viewed from a desktop device

classify the CXR into *pathological compatible with COVID-19* category. It is defined as the proportion of subjects without COVID-19 within the patients with a negative CXR according to the user's criteria: TN/(TN + FN) [30].

Finally, global accuracy depicts the ability of the user to differentiate COVID-19 patients and non-COVID-19 patients. It is the proportion of true positives and true negatives in all evaluated cases: (TP + TN)/(TP + TN + FP + FN) [29].

The information gathered by the application includes TP, TN, FP, FN, sensitivity, specificity, PPV, NPV, and global accuracy for each user and each time they take the test. The specialty or professional category selected after the authentication and the answers given to each CXR are also stored. Nevertheless, the specialty or category chosen before starting the training have only been recorded since the launch of the last version on the 24th of April. Beta test did not record it.

Engagement with the application is tracked using a customized data capture system and validated with Google Analytics [31]. This program is an effective resource for measuring the diffusion and understanding the geodemographics of users [32].

Statistical analysis was performed with IBM SPSS Statistics [33]. Averages for diagnostic accuracy values were computed following two different approaches: firstly, population diagnostic values were calculated from the totals of TP, TN, FP, and FN collected; differently, the means for the diagnostic values obtained by the users were also estimated. The test used to assess differences in the performances between two groups was chi-squared test. To indicate a statistically significant difference, a $p < 0.05$ was considered.

Results

After the beta launch on the 11th of April, the application had 431 users within the first 3 days and 704 within the first week, with a total number of answers of 23,130. This version had users in more than 20 countries, according to Google Analytics reports. Figure 8 shows a map with users' locations.

The application registered 16,360 (70.73%) correct answers (TP + TN) and 6770 (29.27%) incorrect answers (FP + FN). Consequently, the odds of answering correctly was 2.42. Positive likelihood ratio was 2.26, 95% CI [2.20, 2.33], and negative likelihood ratio was 0.39, 95% CI [0.38, 0.40]. Each user sent a mean of 32.86 (SD 39.21) answers, ranging from 1 to 434 (Figs. 9 and 10): an average of 23.24 (SD 28.09) of them were correct, and an average of 9.62 (SD 12.66) were incorrect.

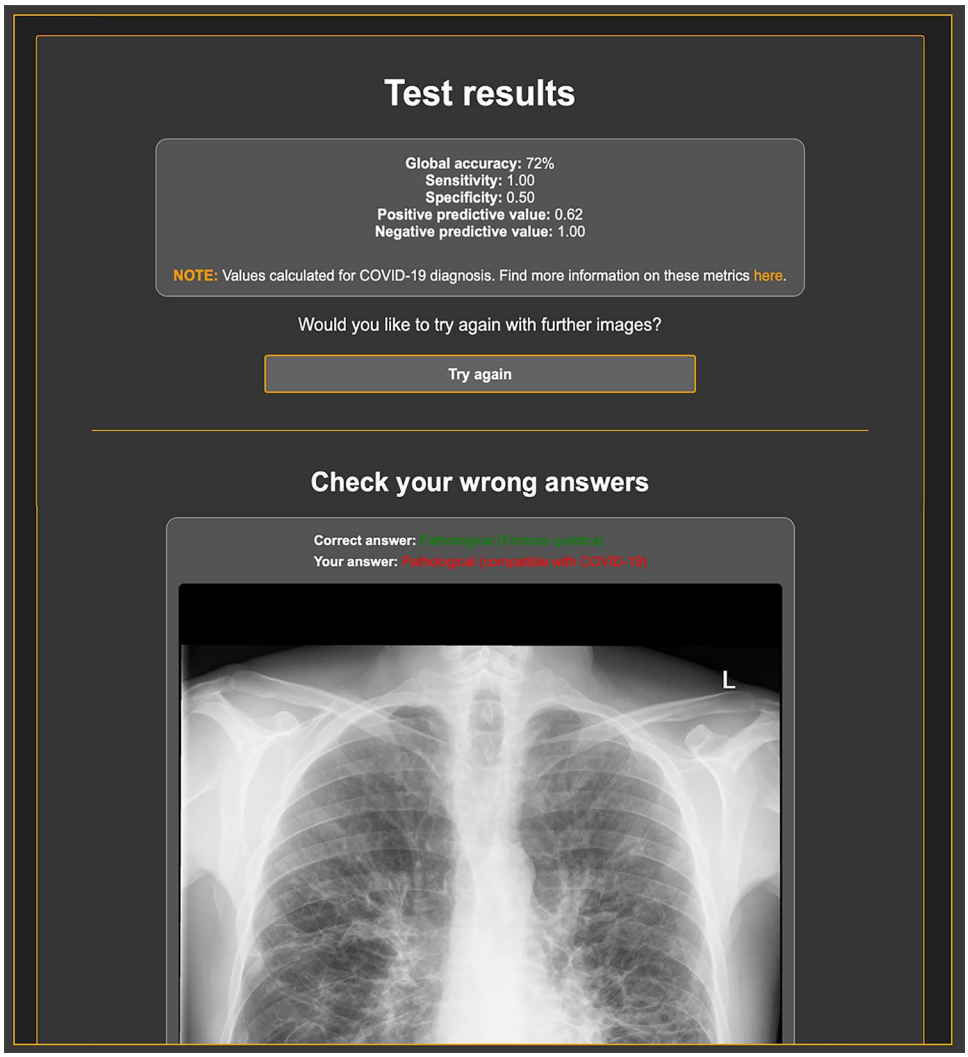


Fig. 6 Screen of results, as viewed from a desktop device

Overall, this first version collected the following answers: 9026 (39.02%) TP; 7334 (31.71%) TN; 3545 (15.33%); FP; and 3225 (13.94%) FN. The mean of TP, TN, FP, and FN per user were 12.82 (SD 15.91), 10.42 (SD 12.88), 4.58 (SD 6.31), and 5.04 (SD 7.65), respectively (Table 3).

The global population diagnostic accuracy values registered were the following: sensitivity 0.74, 95% CI [0.73, 0.74]; specificity 0.67, 95% CI [0.67, 0.68]; PPV 0.72, 95% CI [0.71, 0.72]: NPV 0.69, 95% CI [0.69, 0.70]; and accuracy 0.71, 95% CI [0.70, 0.71] (Table 4). Differently, the means for the diagnostic values obtained by the users were as follows: sensitivity 0.72, 95% CI [0.71, 0.74]; specificity 0.64, 95% CI [0.62, 0.66]; PPV 0.70, 95% CI [0.69, 0.72]; NPV 0.68, 95% CI [0.66, 0.69]; and accuracy 0.68, 95% CI [0.67, 0.69] (Table 5).

The hardest CXR obtained an odds of answering correctly of 0.32 (Table 6). On the other side, the easiest CXR odds of answering correctly was 18.70 (Table 7). Both CXR belonged to the category *pathological compatible with COVID-19* (Figs. 11a and 12).

As would be expected, users who achieved a sensitivity equal or higher to the average sensitivity (0.72), also registered less specificity ($p < 0.0001$). Similarly, the user's with higher specificity, those who achieved the average specificity (0.64) or more, decrease their sensitivity ($p < 0.0001$). In addition, the users who sent more answers than the 50th percentile (68

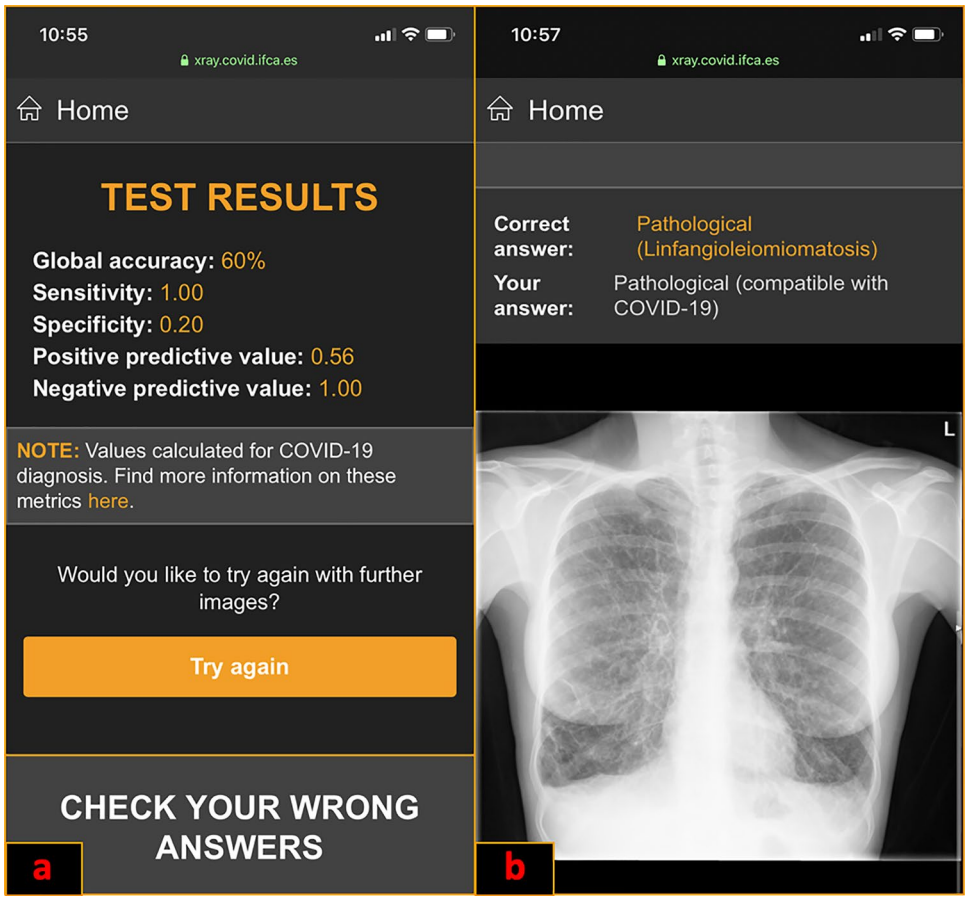


Fig. 7 Screen of results, as viewed from a mobile device

answers) achieved better final average global accuracy ($p < 0.0001$); excluding users who sent less than 10 answers from the analysis, results were also statistically significant ($p = 0.0002$) (Table 8).

Beta version did not ask for the user's category or specialty, so no data about it was recorded before the final version that was launched on the 24th of April.

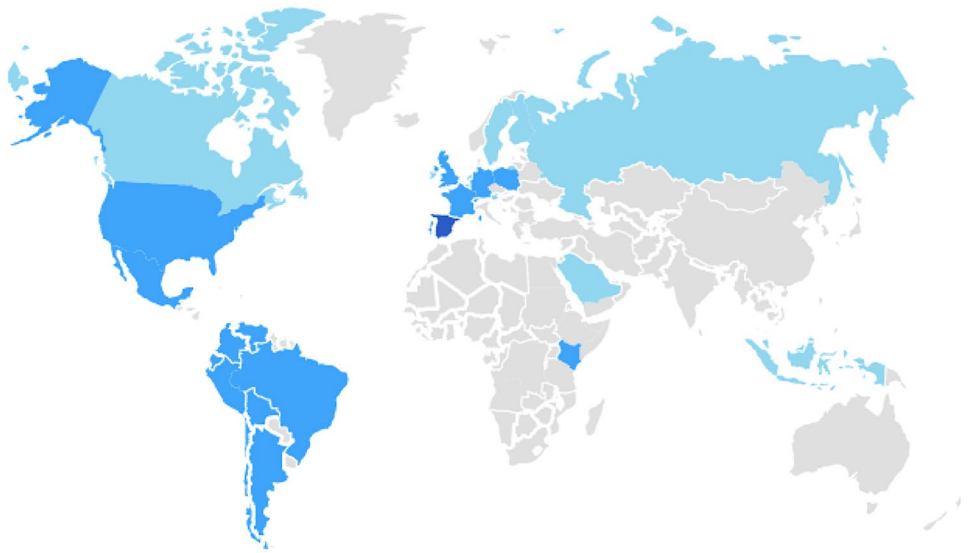


Fig. 8 Countries of the application users. The first version of the application extended to more than 20 countries as the map shows in blue color: a darker color means more visits

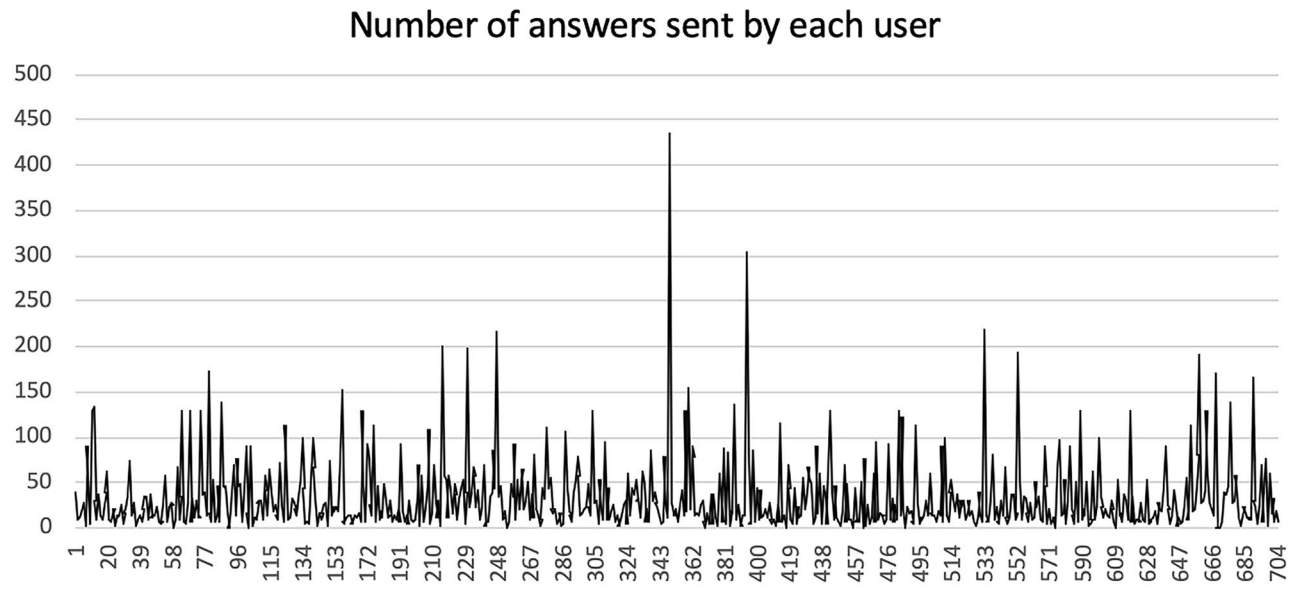


Fig. 9 A frequency distribution graph showing the number of answers sent by each user. The *x*-axis represents the users and the *y*-axis represents the number of responses

Discussion

Several medical educational applications have been developed [33–35], but only a few of them are focused on the diagnosis training, and even less on COVID-19 diagnosis. Currently, the society is facing a health situation which has no precedents [36]. In this context, the role of the medical community is vital [37, 38] and the medical education of those who are taking part of the solution is essential [18, 39].

The application *COVID-19 TRAINING* was developed to help professionals to acquire the required competencies to diagnose COVID-19 on CXR. This application brings a new manner to ease the learning of the necessary skills to successfully evaluate COVID-19 on CXR. According to us, this is a different educational technique that could set a precedent for developing further applications on the training on the diagnosis of other pathologies.

Furthermore, we have introduced an original evaluation method based on the estimation of the users' diagnostic accuracy values for COVID-19. In our opinion, this mechanism of assessment could prove useful for medical professionals since it provides them with information about their sensitivity, specificity, predictive values, and global accuracy. Armed with that knowledge, they will recognize and improve their weaknesses.

The results obtained by the first version of the application showed that those who achieved better sensitivity also decrease their specificity. Similarly, those who registered better specificity also had lower sensitivity. These results

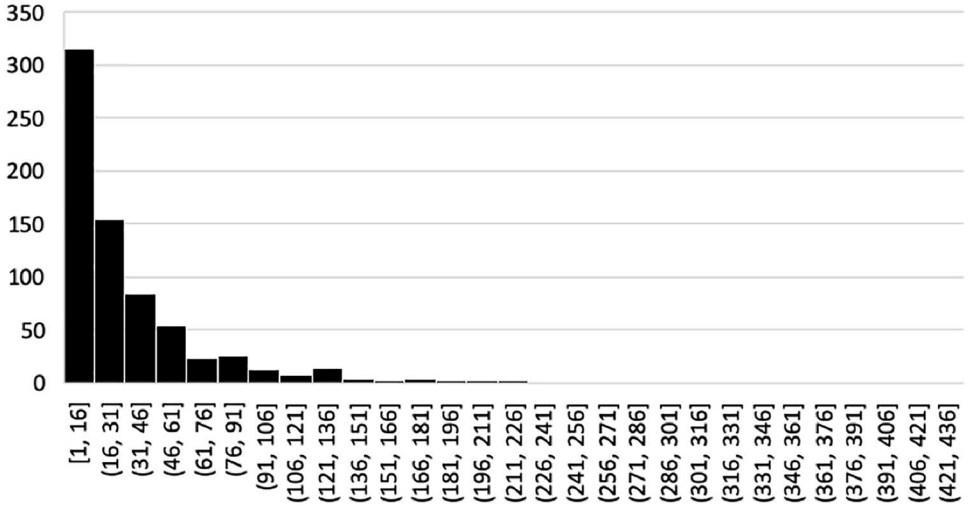


Fig. 10 An histogram depicting the number of answers collected by the first version of the application. The *x*-axis shows intervals of 16 responses width and the *y*-axis shows the frequency of the intervals

Table 3 Basic statistics for the user's answers

Answers	*n* (%)	Mean ± SD (per user)	Range (per user)
TP	9026 (39.02)	12.82 ± 15.91	0–136
FN	3225 (13.94)	4.58 ± 6.31	0–91
TN	7334 (31.71)	10.42 ± 12.88	0–88
FP	3545 (15.33)	5.04 ± 7.65	0–123
Correct responses	16,360 (70.73)	23.24 ± 28.09	0–220
Incorrect responses	6770 (29.27)	9.62 ± 12.66	0–214
Total responses	23,130 (100)	32.86 ± 39.21	1–434

TP true positive, *FN* false negative, *TN* true negative, *FP* false positive

evidence that usually, increasing sensitivity entails decreasing specificity, and vice versa. Our application can help the users to find an optimal balance between their sensitivity and specificity and consequently, to achieve their best possible global accuracy.

In addition, the analysis and display of the failed answers is also a key feature of this tool. At the end of the test, users can review their incorrect responses so they will become aware about their mistakes and they will recognize their weak points. This feature can be also useful in research to identify the cases missed by numerous users. Later on, the examination of these problematic images may give some clues to find solutions to improvable areas in the medical practice not detected before.

Following this approach, we have analyzed the hardest and the easiest CXR trying to understand why users found difficult or obvious the respective cases.

The most difficult CXR was a COVID-19 case that had subtle and ill-defined ground-glass opacities in the right lung. The most evident findings on this image were in the right upper lobe, nearly overlapped with the first costochondral junction (Fig. 11). This first costochondral junction is a well-recognized pitfall on CXR and it sometimes mimics a rounded opacity [40]. Perhaps, in the situations where a COVID-19 is suspected and a prominence on either the first costochondral or first costosternal or sternoclavicular junctions is seen on CXR, further evaluation should be performed.

On the other side, the easiest CXR was also a COVID-19 patient, but in this case, the image illustrated evident peripheral bilateral opacities, which are one of the most typical radiological finding of this disease [7, 13] (Fig. 12).

Overall, this application seems to be helpful. From the beginning, it had a dizzyingly fast diffusion with hundreds of users registered from more than 20 different countries within the first days and more than 20.000 answers sent. Additionally, many medical societies spread and shared this first beta version on their webpages.

Results showed that the users who sent more answers achieved better global accuracy. In addition, only 6 users (0.9%) exceeded the number of cases contained in our dataset, so most of them did not repeat answers. Users who sent less than 10 answers were excluded from the analysis to avoid potential biases. These facts could indicate that the application improves the users' diagnostic skills.

Finally, we also found that medical educational applications may be used in research as a new method to collect relevant information. Since the application was officially launched on the 24th of April, data about the users' specialty or professional category have been recorded. Our intention is to use this data to analyze the difference in the diagnostic values between the users belonging to different specialties or categories and to try to estimate the real utility of the CXR on the evaluation of COVID-19.

Table 4 Population diagnostic accuracy values

Statistic	Value	95% CI
Sensitivity	.74	[.73, .74]
Specificity	.67	[.67, .68]
Positive predictive value	.72	[.71, .72]
Negative predictive value	.69	[.69, .70]
Global accuracy	.71	[.70, .71]

Values computed from the total answers collected

CI confidence interval

Table 5 Average of user's diagnostic accuracy values

Statistic	Mean ± SD	95% CI of the mean
Sensitivity	.72 ± .21	[.71, .74]
Specificity	.64 ± .25	[.62, .66]
Positive predictive value	.70 ± .21	[.69, .72]
Negative predictive value	.68 ± .22	[.66, .69]
Global accuracy	.68 ± .16	[.67, .69]

Values estimated from the diagnostic values achieved by each user

CI confidence interval

Table 6 The top 10 hardest chest X-rays

Position	Image	TP	FN	TN	FP	Odds of answering correctly
1	F06	47	149	0	0	.32
2	F020_2	0	0	48	117	.41
3	M053	53	128	0	0	.41
4	M034	0	0	57	130	.44
5	M06	0	0	56	125	.45
6	M030_2	52	116	0	0	.45
7	M09	66	121	0	0	.55
8	F09	66	118	0	0	.56
9	M032_2	0	0	62	98	.63
10	M031	70	104	0	0	.67

TP true positive, *FN* false negative, *TN* true negative, *FP* false positive

Table 7 The top 10 easiest chest X-rays

Position	Image	TP	FN	TN	FP	Odds of answering correctly
1	M027	187	10	0	0	18.70
2	F024	0	0	185	10	18.50
3	M027_2	182	10	0	0	18.20
4	F022	0	0	163	10	16.30
5	F014	175	11	0	0	15.91
6	M019	158	11	0	0	14.30
7	M018	181	13	0	0	13.92
8	F06_3	173	13	0	0	13.31
9	M02_2	151	12	0	0	12.58
10	M016	163	14	0	0	11.64

TP true positive, *FN* false negative, *TN* true negative, *FP* false positive

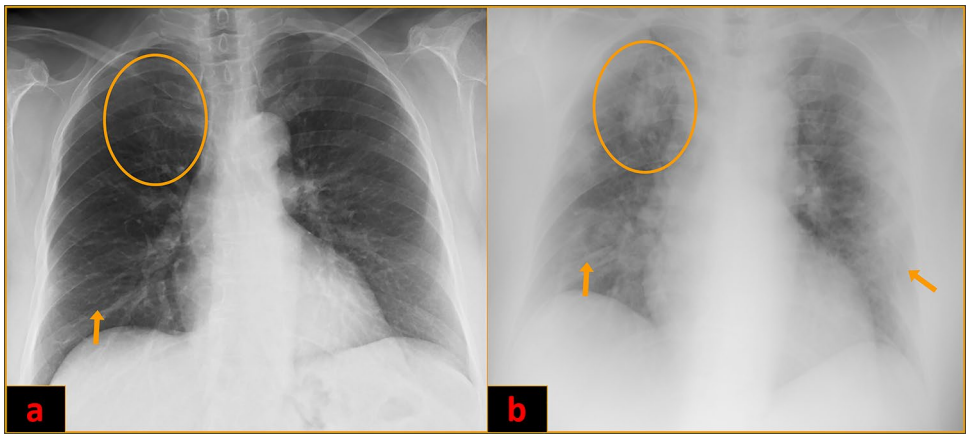


Fig. 11 **a** The most difficult chest X-ray. It is a 66-year-old female with a chest X-ray belonging to the category *pathological compatible with COVID-19*. On this image, subtle ground-glass opacities on the right lung are visible (arrow): the most evident opacity is in the right upper lobe (circle). **b** On a subsequent chest X-ray performed to the same patient after three days, this ground-glass opacity in the upper lobe became even clearer (circle). In addition, new opacities appeared in the lower lobes on this chest X-ray (arrows)

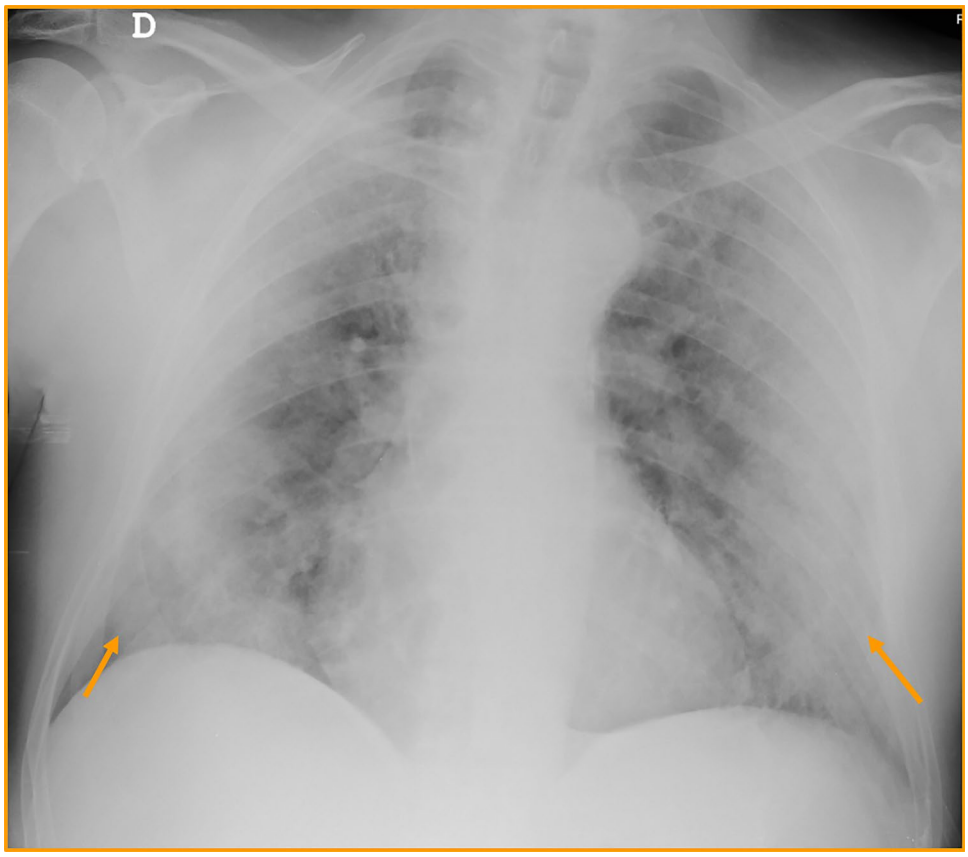


Fig. 12 The easiest chest X-ray. It is a 69-year-old male diagnosed of COVID-19. The image shows extensive, multiple and bilateral opacities (arrows) indicating a severe form of COVID-19 pneumonia

Table 8 Comparison of the results between groups and *p* values for chi-squared tests

Statistic	Group 1	Group 2	Proportion difference	95% CI of difference	χ^2 value	*p* value
Sensitivity [TP/(TP + FN)][a]	.76 [3504/4604]	.72 [5479/7591]	− .04	[− .06, − .02]	22.814	< .0001
Specificity [TN/(TN + FP)][b]	.72 [3192/4425]	.64 [4130/6431]	− .08	[− .10, − .06]	74.795	< .0001
Global accuracy [(TP + TN)/(TP + TN + FP + FN)][c]	.70 [10430/14956]	.73 [5930/8174]	.03	[.02, .04]	20.147	< .0001
Global accuracy (TP + TN)/(TP + TN + FP + FN)[d]	.70 [9815/13988]	.73 [5930/8174]	.02	[.01, .04]	14.203	.0002

CI confidence interval, *TP* true positive, *FN* false negative, *TN* true negative, *FP* false positive

[a]Group 1: users who achieved a specificity lesser than the mean (.64); group 2: users who achieved a specificity equal to or higher than the mean (.64)

[b]Group 1: users who registered a lower sensitivity than the mean (.72); group 2: users who registered a sensitivity equal to or higher than the mean (.72)

[c]Group 1: users who sent a number of answers equal to the 50th percentile (68 answers) or less; group 2: users who sent more than 68 answers

[d]Group 1: users who sent a number of answers between 10 and the 50th percentile (68 answers); group 2: users who sent more than 68 answers

Conclusion

In present COVID-19 pandemic, the medical education of the professionals involved in patients care is vital and *COVID-19 TRAINING* brings a different solution to help them in this purpose. Applications focused on the training on diagnosis could provide a new original manner of evaluation for medical professionals. The assessment of users by estimating their diagnostic accuracy values make them aware of their weak points. In addition, this kind of application also collects valuable information that can be used for research purposes.

Appendix 1

A link to a video navigating the application: https://api.cloud.ifca.es:8080/swift/v1/covid19/VIDEO%20APP.mov.

Acknowledgments We would like to acknowledge all the professionals who have battled in an exemplary manner to safeguard the health and life of all citizens worldwide since the beginning of this pandemic.

Funding The authors received no financial support for this work.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al: A Novel Coronavirus from patients with pneumonia in China, 2019. N Engl J Med 382(8): 727-733, 2020. https://doi.org/10.1056/NEJMoa2001017
2. Raoult D, Zumla A, Locatelli F, Ippolito G, Kroemer G: Coronavirus infections: Epidemiological, clinical and immunological features and hypotheses. Cell Stress 4(4): 66-75, 2020. https://doi.org/10.15698/cst2020.04.216
3. WHO. WHO's Coronavirus disease (COVID-19) outbreak situation dashboard. (n.d.) Available at https://www.who.int. Accessed 6 May 2020.
4. Cucinotta D, Vanelli M: WHO Declares COVID-19 a Pandemic. Acta Biomed 91(1): 157-160, 2020. https://doi.org/10.23750/abm.v91i1.9397
5. Wu Z, McGoogan JM: Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72314 Cases From the Chinese Center for Disease Control and Prevention. JAMA 323(13): 1239-1242, 2020. https://doi.org/10.1001/jama.2020.2648
6. Xiea M, Chen Q: Insight into 2019 novel coronavirus — an updated interim review and lessons from SARS-CoV and MERS-CoV. Int J Infect Dis 94: 119-124, 2020. https://doi.org/10.1016/j.ijid.2020.03.071
7. Cao Y, Liu X, Xiong L, Cai K: Imaging and clinical features of patients with 2019 novel coronavirus SARS-CoV-2: A systematic review and meta-analysis. J Med Virol, https://doi.org/10.1002/jmv.25822, April 3, 2020.
8. Lechien JR, Chiesa-Estomba CM, De Siati DR, Horoi M, Le Bon SD, Rodríguez A, et al: Olfactory and gustatory dysfunctions as a clinical presentation of mild-to-moderate forms of the coronavirus disease (COVID-19): a multicenter European study. Eur Arch Otorhinolaryngol, https://doi.org/10.1007/s00405-020-05965-1, April 6, 2020.
9. Udugama B, Kadhiresan P, Kozlowski HN, Malekjahani A, Osborne M, Li VYC, et al: Diagnosing COVID-19: The Disease and Tools for Detection. ACS Nano 14(4): 3822-3835, 2020. https://doi.org/10.1021/acsnano.0c02624
10. Fang Y, Zhang H, Xie J, Lin M, Ying L, Pang P, et. al: Sensitivity of Chest CT for COVID-19: Comparison to RT-PCR. Radiology, https://doi.org/10.1148/radiol.2020200432, February 19, 2020.
11. He JL, Luo L, Luo ZD, Lyu JX, Ng MY, Shen XP, et al: Diagnostic performance between CT and initial real-time RT-PCR for clinically suspected 2019 coronavirus disease (COVID-19) patients outside Wuhan, China. Resp Med 168:105980, 2020. https://doi.org/10.1016/j.rmed.2020.105980
12. Li Y, Yao L, Li J, Chen L, Song Y, Cai Z, et. al: Stability issues of RT-PCR testing of SARS-CoV-2 for hospitalized patients clinically diagnosed with COVID-19. J Med Virol, https://doi.org/10.1002/jmv.25786, March 26, 2020.
13. Wong HYF, Lam HYS, Fong AHT, Leung ST, Chin TWY, Lo CSY, et al: Frequency and distribution of chest radiographic findings in COVID-19 positive patients. Radiology, https://doi.org/10.1148/radiol.2020201160, March 27, 2020.
14. Jacobi A, Chung M, Bernheim A, Eber C: Portable chest X-ray in coronavirus disease-19 (COVID-19): A pictorial review. Clin Imaging, 64:35-42, 2020. https://doi.org/10.1016/j.clinimag.2020.04.001
15. Borghesi A, Maroldi R: COVID-19 outbreak in Italy: Experimental chest X-ray scoring system for quantifying and monitoring disease progression. Radiol Med 125(5):509–513, 2020. https://doi.org/10.1007/s11547-020-01200-3, 2020.
16. Kooraki S, Hosseiny M, Myers L, Gholamrezanezhad A: Coronavirus (COVID-19) Outbreak: What the Department of Radiology Should Know. J Am Coll Radiol 17(4): 447-451, 2020. https://doi.org/10.1016/j.jacr.2020.02.008
17. Hwang EJ, Nam JG, Lim WH, Park SJ, Jeong YS, Kang JH, et al: Deep learning for chest radiograph diagnosis in the Emergency Department. Radiology 293(3):573-580, 2019. https://doi.org/10.1148/radiol.2019191225
18. van der Gijp A, Ravesloot CJ, Jarodzka H, van der Schaaf MF, van der Schaaf IC, van Schaik JPJ, et al: How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology. Adv Health Sci Educ Theory Pract 22: 765–787, 2017. https://doi.org/10.1007/s10459-016-9698-1
19. Yoon JS, Boutis K, Pecaric MR, Fefferman NR, Ericsson KA, Pusic MV: A think-aloud study to inform the design of radiograph interpretation practice. Adv Health Sci Educ Theory Pract, https://doi.org/10.1007/s10459-020-09963-0, March 5, 2020.
20. American College of Radiology. ACR practice parameter for radiologist coverage of imaging performed in hospital emergency departments. Available at https://www.acr.org/-/media/ACR/Files/Practice-Parameters/HospER.pdf?la=en. Accessed 19 May 2020.
21. Eng J, Mysko WK, Weller GE, et al: Interpretation of Emergency Department radiographs: a comparison of emergency medicine physicians with radiologists, residents with faculty, and film with digital display. AJR Am J Roentgenol 175(5):1233-1238, 2000. https://doi.org/10.2214/ajr.175.5.1751233
22. Gatt ME, Spectre G, Paltiel O, Hiller N, Stalnikowicz R: Chest radiographs in the emergency department: is the radiologist really

necessary?. Postgrad Med J 79(930):214-217, 2003. https://doi.org/10.1136/pmj.79.930.214
23. Petinaux B, Bhat R, Boniface K, Aristizabal J: Accuracy of radiographic readings in the emergency department. Am J Emerg Med 29(1):18-25, 2011. https://doi.org/10.1016/j.ajem.2009.07.011
24. Chung JH, Duszak R Jr, Hemingway J, Hughes DR, Rosenkrantz AB: Increasing Utilization of Chest Imaging in US Emergency Departments From 1994 to 2015. J Am Coll Radiol 16(5):674-682, 2019. https://doi.org/10.1016/j.jacr.2018.11.011
25. Sellers A, Hillman BJ, Wintermark M: Survey of after-hours coverage of emergency department imaging studies by US academic radiology departments. J Am Coll Radiol 11(7):725-730, 2014. https://doi.org/10.1016/j.jacr.2013.11.015
26. Yi Y, Lagniton PNP, Ye S, Li E, Xu RH: COVID-19: What has been learned and to be learned about the novel coronavirus disease. Int J Biol Sci 16(10):1753-1766, 2020. https://doi.org/10.7150/ijbs.45134
27. Python 3 Programming Language, RRID:SCR_008394.
28. Anonymized
29. Baratloo A, Hosseini M, Negida A, El Ashal G: Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity. Emerg (Tehran) 3(2):48-49, 2015.
30. Šimundić AM: Measures of Diagnostic Accuracy: Basic Definitions. EJIFCC. 2009;19(4):203-211, 2009.
31. Google Analytics: http://google.com/analytics.
32. McGuckin C, Crowley N: Using Google Analytics to evaluate the impact of the CyberTraining project. Cyberpsychol Behav Soc Netw 15(11): 625-9, 2012. https://doi.org/10.1089/cyber.2011.0460
33. SPSS, RRID:SCR_002865; IBM, New York, USA.
34. Wood LE, Picard MM, Kovacs MD: App Review: The Radiology Assistant 2.0. J Digit Imaging 31: 383–386, 2018. https://doi.org/10.1007/s10278-018-0070-2.
35. Mosa AS, Yoo I, Sheets L: A systematic review of healthcare applications for smartphones. BMC Med Inform Decis Mak 12:67, 2012. https://doi.org/10.1186/1472-6947-12-67
36. WHO. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. Available at https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020. Accessed 5 April 2020.
37. Adams JG, Walls RM: Supporting the Health Care Workforce During the COVID-19 Global Epidemic. JAMA 323(15):1439–1440, 2020. https://doi.org/10.1001/jama.2020.3972
38. DeWitt DE: Fighting COVID-19: Enabling Graduating Students to Start Internship Early at Their Own Medical School. Ann Intern Med, https://doi.org/10.7326/M20-1262, April 7, 2020.
39. Lei P, Huang Z, Liu G, Wang P, Song W, Mao J, et al: Clinical and computed tomographic (CT) images characteristics in the patients with COVID-19 infection: What should radiologists need to know? J X-ray Sci Technol, https://doi.org/10.3233/XST-200670, April 7, 2020.
40. Haramati LB, Haramati N: Pulmonary pseudonodules on computed tomography: a common pitfall caused by degenerative arthritis. J Thorac Imaging 11(4):283-285, 1996. https://doi.org/10.1097/00005382-199623000-00007

9.5.2. Artículo científico en la revista *Insights into Imaging*: *A primer on deep learning and convolutional neural networks for clinicians*

Iglesias *et al. Insights Imaging* (2021) 12:117
https://doi.org/10.1186/s13244-021-01052-z

Insights into Imaging

EDUCATIONAL REVIEW

Open Access

A primer on deep learning and convolutional neural networks for clinicians

Lara Lloret Iglesias[1*], Pablo Sanz Bellón[2,3], Amaia Pérez del Barrio[2,3], Pablo Menéndez Fernández-Miranda[2,3], David Rodríguez González[1], José A. Vega[4,5], Andrés A. González Mandly[2,3] and José A. Parra Blanco[2,3]

Abstract

Deep learning is nowadays at the forefront of artificial intelligence. More precisely, the use of convolutional neural networks has drastically improved the learning capabilities of computer vision applications, being able to directly consider raw data without any prior feature extraction. Advanced methods in the machine learning field, such as adaptive momentum algorithms or dropout regularization, have dramatically improved the convolutional neural networks predicting ability, outperforming that of conventional fully connected neural networks. This work summarizes, in an intended didactic way, the main aspects of these cutting-edge techniques from a medical imaging perspective.

Keywords: Deep learning, Image processing, Medical imaging, Educational

Introduction

Artificial intelligence (AI) is defined as the intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans. Despite the hype that is currently encountering, it is not something as new as one may imagine. Most of the people consider that the historical article written by Turing in 1950 [1] established the beginning of this new field by asking a simple question: *Can machines think?* AI includes what was later called Machine Learning, but the first AI algorithms did not actually learn. The first AI computer programs were based on the so-called symbolic AI. This first approach consisted on programming a set of rules large enough to manipulate knowledge and reached its height of popularity during the boom of expert systems in the eighties. Symbolic AI was probed to work fine for logic problems where the rules were clear, such as chess playing, but were useless for more diffuse and perceptual problems such as the recognition and manipulation of images, voice, language. This is where the *learning* approach comes into play.

The learning approach

The concept of machine learning arose from the need of answering certain questions that were not covered by the symbolic AI, where all the rules to solve certain problem should be coded by some expert. Some of these open issues were:

- Can a computer program go beyond what we know how to code?
- Can a computer program learn just by looking at the data?
- Can a computer program even surprise us?

The main idea behind Machine Learning is to let the computer learn directly by exposing it to a large number of examples of a given situation or class. The Machine Learning algorithm will then automatically develop a model that can deduce and generalize the examples it was exposed to and make predictions from it for totally new cases. This allows to develop systems capable of tasks so diverse as predicting house prices by *looking* at

*Correspondence: lloret@ifca.unican.es
[1] Advanced Computation and e-Science, Instituto de Fsica de Cantabria - CSIC, Santander, Spain
Full list of author information is available at the end of the article

the historical behavior of the real-estate market (regression problem) or a system that is able to learn how to distinguish two different varieties of glioblastoma that it has never seen before just by *looking* at many different samples of magnetic resonance images (MRI) of both categories (classification problem). At the end of the day, this is similar to the way in which we humans learn: by exposition to many examples allowing us to generalize a certain concept. In the context of medical imaging, there are currently two main different types of learning: supervised learning and unsupervised learning. In the supervised learning approach, the learning algorithm receives as input a series of data tagged with the correct answer or label. This means that, in the case of the glioblastoma classification, the machine learning system would have access to several MRI of each of the two glioblastomas types with a label for each of them indicating to which category it belongs. One can easily see the improvement with respect to the symbolic AI: in the symbolic AI approach, some expert should have found and coded the rules allowing to distinguish the two different types of tumors. This often implies having a well-established metric on how to distinguish the categories to be predicted, something that is usually not the case in computer vision problems. Actually, most of the mechanisms used by humans to perform daily actions, such as recognizing faces or even speaking or understanding the context of a sentence, would be very difficult to wrap up into some coded rules. The paradigm shift between classical programming or symbolic AI and the machine learning approach can be easily understood by looking at Fig. 1.

In the unsupervised learning approach, the learning algorithm does not receive the labels. Instead, it only receives some input data and the algorithm alone will work on its own to extract the information needed to solve the problem under study. One of the most common tasks to be solved with an unsupervised approach is the clustering. It consists in grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar, according to a certain metric, to each other than to those in other groups. For the glioblastoma case previously mentioned, this would imply to just give the algorithm as input a set of brain MRI without any further indication and let the system learn whether there are some meaningful features allowing to separate the dataset into categories. This can be an easy task if we want to separate very obvious categories such as images of green circles from images of red circles, but can become really complicated for more multidimensional tasks. Some of the advantages of the unsupervised learning approach is that it could allow us to really learn new things about the problem. For instance, it may find three different tumor categories instead of two, potentially allowing us to discover nuances in brain cancer diagnosis that escaped the human experience. Besides, it is much easier to find unlabelled data than finding some expert willing to tag a big dataset for training a supervised algorithm. On the other hand, the main disadvantage of the unsupervised learning is that, nowadays, the algorithms are still less accurate and trustworthy than the supervised methods.

It is important to note that, even if we have focused on supervised and unsupervised learning, there are other learning approaches less used for now in the medical context, but still worth mentioning:

- The semi-supervised learning approach combines both supervised and unsupervised learning. In this case, the algorithm can learn from a mixture of labelled and unlabelled data. This approach includes a range of possible techniques that are outside the scope of this paper.

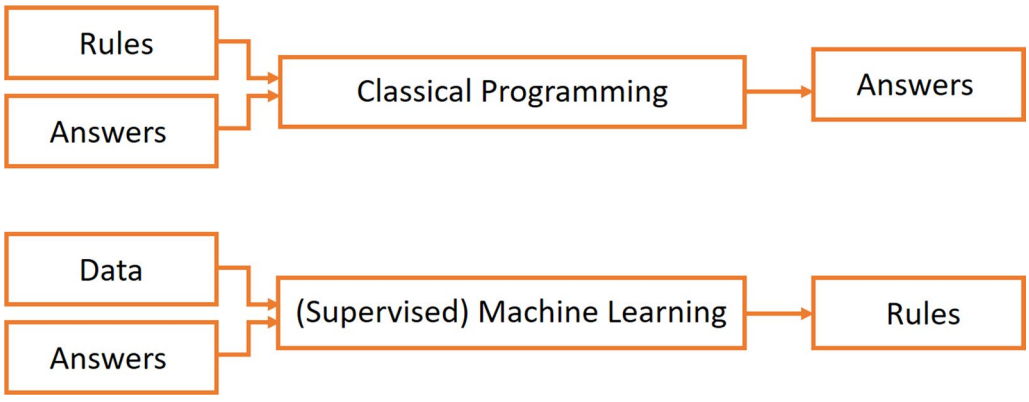


Fig. 1 Classical programming vs supervised learning approach

- The reinforcement learning approach is concerned with how software agents should take actions in an environment by maximizing some portion of a certain cumulative reward. These are goal-oriented algorithms, which learn how to attain a complex objective (goal) or how to maximize along a particular dimension over many steps. For example, they can maximize the points won in a game over many moves. This approach is widely used in robotics nowadays.

In this section, the different types of learning available for machine learning algorithms have been summarized. For the examples in this article, we will be focused on classification problems using the supervised learning approach since it is the most widely used nowadays in medicine as it has proved its success for many different applications.

How computers see images

In order to understand how machine learning algorithms and, more specifically Convolutional Neural networks, work with images, one needs to understand how computers actually *see* these images. Figure 2 shows the matrix representation of a black and white image [2]. Since this is a 8 bits image, the pixels can take values from 0 to 255 ($2^8 = 256$) depending on the tonality of gray. For the images in color, the matrix representation would be the same, but instead of a single matrix, there will be three matrices corresponding to the red, green and blue color whose values will also range from 0 to 255. The mix of these three color channels will form the final image as can be seen in Fig. 3.

Now that we know how computers represent images, let us do a small exercise. Let us take for instance the left image in Fig. 4. Imagine you need to separate the green points from the blue points. This would be what we have previously defined as a classification problem. Just by looking at the image any person would imagine the straight line separating both categories. But, what happens if you are told to give the equation of this straight line? Since there are no graduated axes and probably you have not done this exercise in quite a longtime, probably it will take you a couple of minutes to figure out which is the answer. Let us now imagine that we rotate and translate the axes as in the central image in Fig. 4. For more clarity, the right image in Fig. 4 shows the new x axis parallel to the text. The problem has not changed. Just the representation of the data has changed by performing a linear transformation. If we are asked now to separate green points and blue points, we would not even need 10 s: points with $x > 0$ are blue, and points with $x < 0$ are green. What has happened here? By changing the representation of the data, we have transformed a slightly difficult problem into a much easier problem to solve. This concept of representation change is paramount when understanding how Machine Learning and Deep Learning algorithms work. Since we have already seen that, for a computer, an image is just a set of values forming a matrix, we can easily understand that there is no much difference between the two-dimensional example in Fig. 4 and a multidimensional image matrix. Thus, our goal in an image-based classification problem would be to find alternative representations of the raw image values until we find a new way of looking at the data where the problem becomes easier. The two-dimensional problem in Fig. 4 was very easy to solve by eye, but in real live

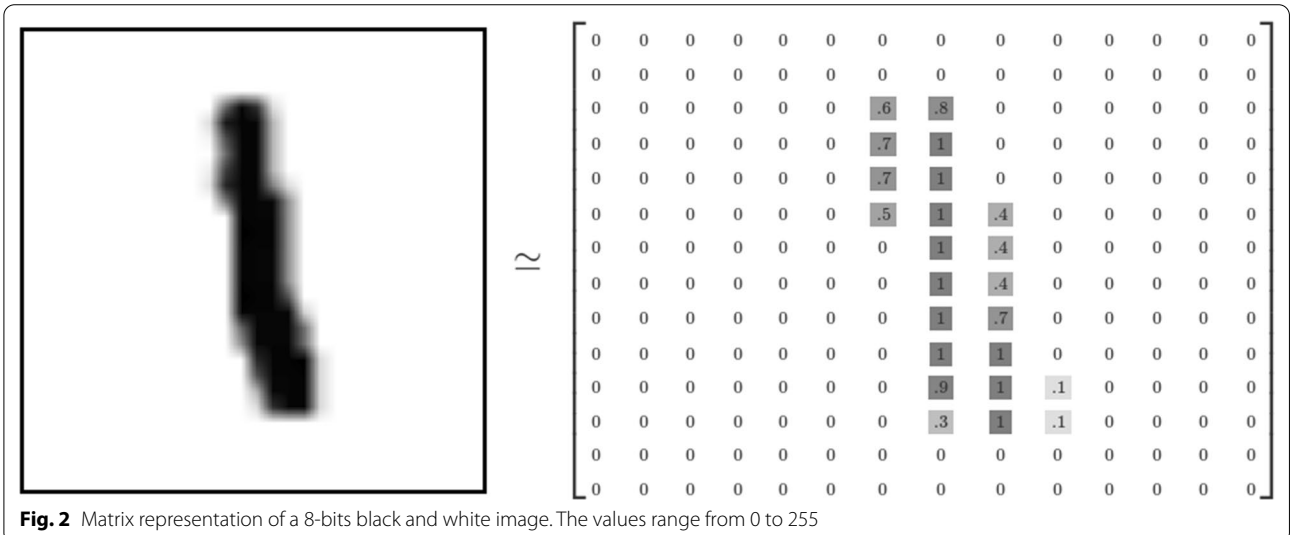


Fig. 2 Matrix representation of a 8-bits black and white image. The values range from 0 to 255

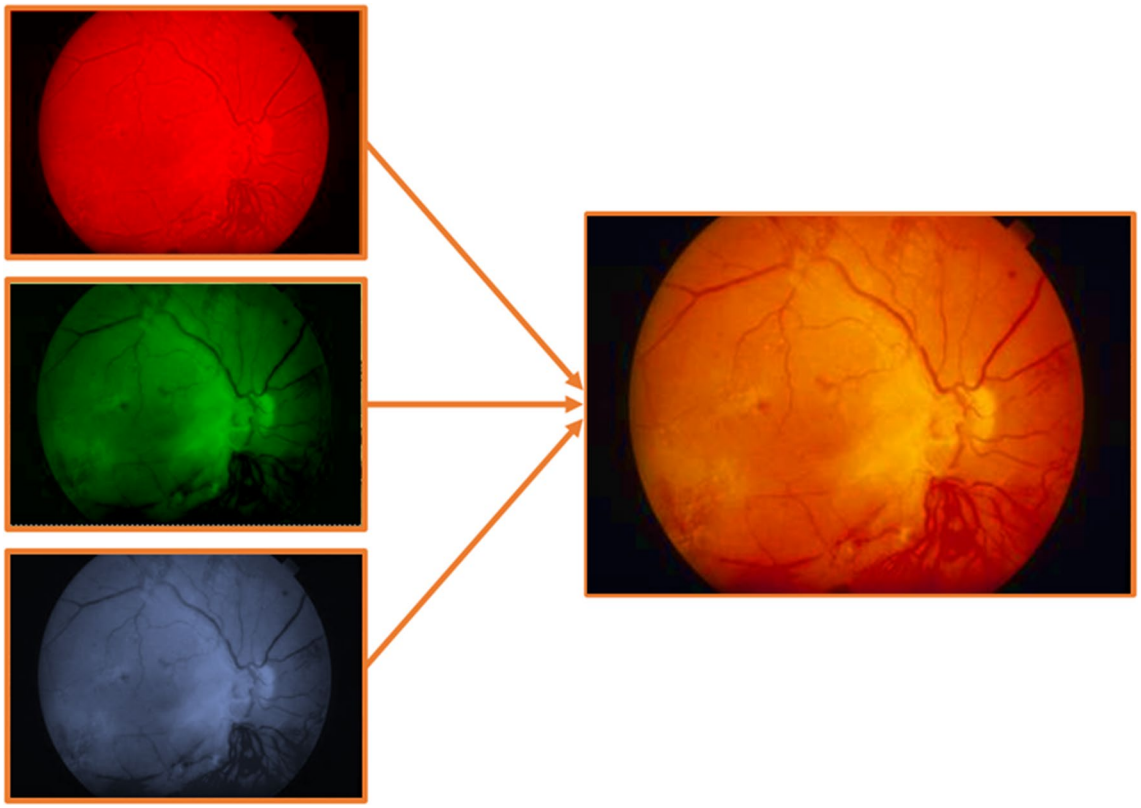


Fig. 3 The three color channels (red, green, blue) of a retina image for diabetic retinopathy detection

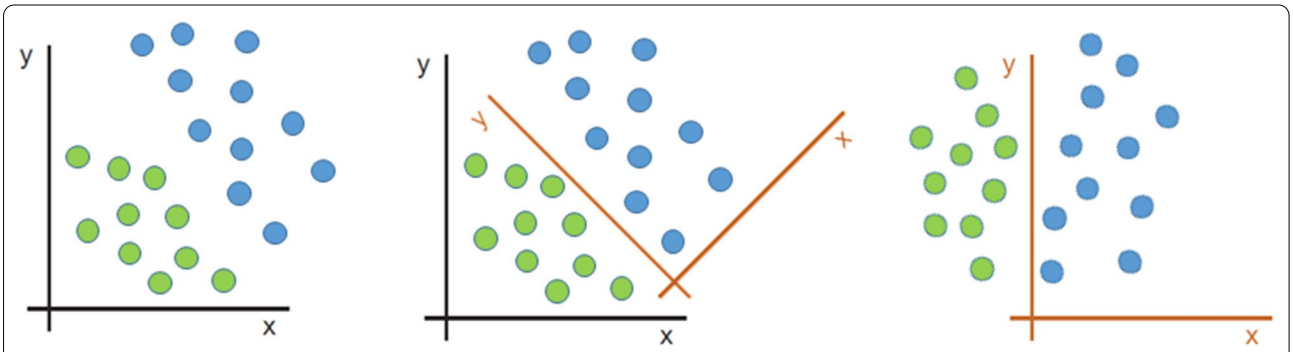


Fig. 4 Solving a two-dimensional classification problem by eye by changing the data representation

problems are much more complicated. This means that we would like to have some way for efficiently and systematically finding and testing new representations until we find the optimal one. This process of automatically searching for the optimal solution (or representation) for our problem is exactly what the Machine Learning algorithms do.

Neural networks

Artificial neural networks offer a way of achieving a systematic search for the optimal solution to the problem to be solved. Their elementary units, the neurons, are slightly inspired by the biological neurons: an artificial neuron receives one or more inputs (corresponding to postsynaptic potentials of the biological neurons dendrites in the biological analogy) and sums them to produce an output (or activation, representing a neuron's

action potential which is transmitted along its axon). Each input is separately weighted, and their sum is passed through a nonlinear function known as activation function. Without the activation function, the neural network could only solve linear problems. These functions work similarly to the threshold potentials needed to regulate and propagate signaling in the nervous system. The choice of these activation functions is one of the *hyperparameters* in a learning algorithm, which means that it is up to us to decide which one to use. Figure 5 shows the structure of an artificial neuron.

The W matrix and the b or bias factor are first applied to the input data x in order to perform a linear transformation (rotation + translation, respectively) similarly as in the 2D example in Sect. 3. The σ in this case is a particular and widely used activation function called sigmoid. The shape of the sigmoid function can be seen in Fig. 6. The sigmoid gives smoother output values than a simple step function, is differentiable and presents a very nice property: its derivative depends on the function itself.

$$\sigma'(x) = \sigma(x) * (1 - \sigma(x))$$

We will later see that this is a desirable feature for an activation function.

Learning process

Let us now see how the learning process works. For most of its range, the sigmoid function will return values close to 1 or close to 0. This makes this function very appropriate for binary problems, like in the case of the classification problem in Fig. 4. The output of the activation function can then be taken as the answer to the problem: an output value of 0 can correspond to blue and a value of 1 to green, or the other way around: it does not matter as long as we keep the criteria consistent. We can take each of the data points in Fig. 4 and pass it through the neuron. In this case, the input is bidimensional corresponding to the x and y coordinates of each of the data points, i.e., the input value is $X = (x_{\text{input}}, y_{\text{input}})$. Taking one set of values from our input dataset through one single neuron (also called forward propagation), the full expression is shown in Fig. 7. where the W matrix and the bias b correspond, in this first iteration, to some random values. The L is the so-called loss function. This is a function accounting for how well we are performing. It depends on a (the prediction) and y (the true label of the data points, i.e.: 0 in the case of blue and 1 in the case of green). This is where the labels in the supervised learning approach come into play. In order to learn with a supervised algorithm, we need to know beforehand the category of our data points. The L should be small when we are performing well (i.e., predicting properly the category of each data point) and should be big when we are not doing a good job. This means that, for an optimal solution, the loss function should be as small as possible. This transforms our learning problem into a minimization problem. We need to minimize the loss function.

This minimization process is what we call *learning*. One can decide which is the loss function depending on the problem to be solved, for instance, as an example, the loss function can just be the mean squared error between the prediction a and the real label y. The mean squared error is calculated as the average of the squared

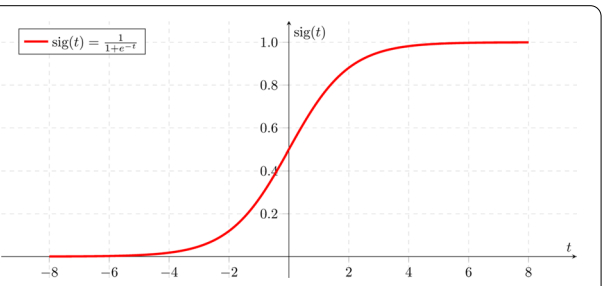


Fig. 6 Sigmoid function

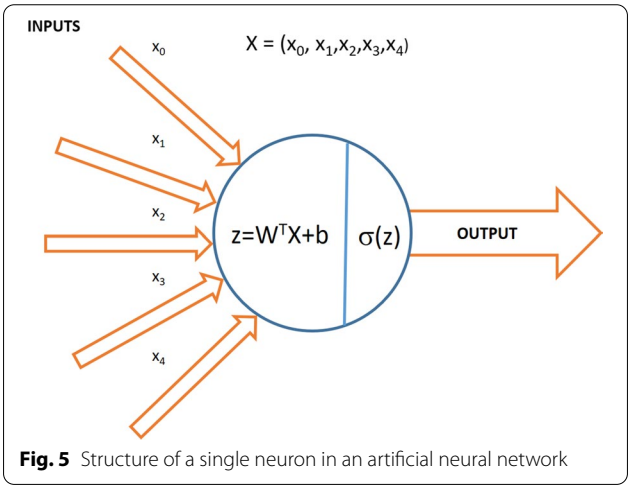


Fig. 5 Structure of a single neuron in an artificial neural network

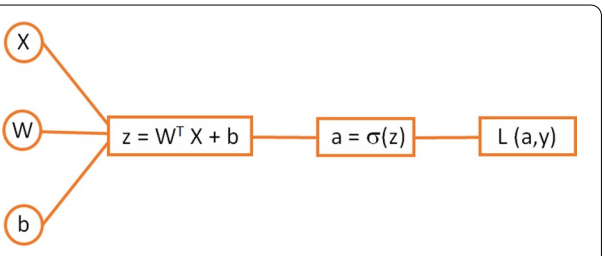


Fig. 7 Forward propagation of a single neuron

differences between the predicted and actual values or labels, i.e., $\Sigma(a_i - y_i)^2$. The result is always positive, and a perfect algorithm will give a value of 0. The squaring means that larger mistakes result in more error than smaller mistakes, i.e., the model is punished for making larger mistakes. Another possible choice for loss function is the cross-entropy function, widely used for classification problems. For the educational purpose of this document, understanding the mean squared error would be enough.

After the first pass through the neuron with random values for W and b, the second iteration will not be random at all. We want to learn which are the values of W and b so that L is as small as possible. For this, we will calculate the derivatives and update the values of W and b according to the direction of the minimum. This can be better understood taking a look at Fig. 8 where a hypothetical curve of L versus W is plotted. After the first iteration of our neuron, the W value can be either greater or smaller than the W minimizing the L function ($W_{\min}$). For this particular problem, if the W value is smaller than $W_{\min}$ (point 1 in Fig. 8), the slope at that point, and hence, the derivative dL/dW, would be negative. In the same way, if the W value is greater than $W_{\min}$ (point 1 in Fig. 8), the derivative at that point would be positive. This means that, in both cases, if we subtract the derivative from the initial random value of W, we would be going in the direction of the $W_{\min}$ as indicated by the red arrows in Fig. 8.

The same is valid also for the value of b. After the first iteration, the values of W and b can be updated by the formula in Fig. 9. The computation of the gradient of the loss function with respect to the weights of the network (W and b) is called backpropagation. The term α in Fig. 9 is known as the learning rate. The derivative gives the direction of the step in the direction of the minimum, and the learning rate gives the magnitude of the step. α is another of the model *hyperparameters.* In the case of α, we must take into account that if it is too large, we may take steps that are too big and we may miss the minimum. If the learning rate is too small, the learning process will be too slow since we will be approaching the minimum with tiny steps.

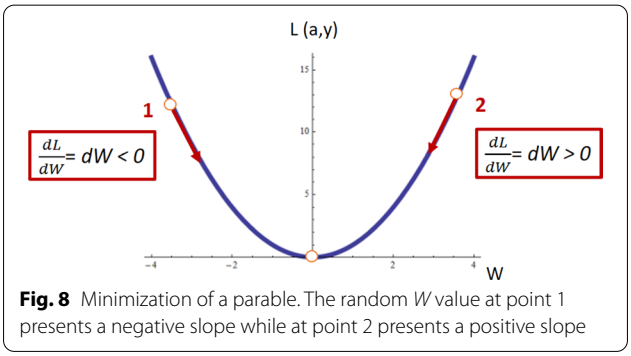


Fig. 8 Minimization of a parable. The random W value at point 1 presents a negative slope while at point 2 presents a positive slope

$$\frac{dL}{dw} = dW = \frac{dL}{da}\frac{da}{dz}\frac{dz}{dw} \qquad \frac{dL}{db} = db = \frac{dL}{da}\frac{da}{dz}\frac{dz}{db}$$

$$W = W - \alpha\, dW \qquad b = b - \alpha\, db$$

Fig. 9 Derivative calculation using the chain rule and update of the W and b parameters in the direction of the minimum

The process will be then repeated with all the points in Fig. 4, updating for each iteration the values of W and b so that they come always closer to the ones that minimize the loss function, optimizing thus the performance of the algorithm.

Multilayer neural networks

For very simple problems, one single neuron can be enough. Usually in a real-life problem, one needs more than one data representation change in order to solve it. This is where the term *network* comes into play. For complicated problems, one would need many different representations of the data, that will be combined among them to create further representations, that will in turn combine, etc., in order to reach the optimal representation allowing to solve the image classification problem or any other problem one may want Machine Learning to solve. This stack of combined representations can be nicely visualized with the shape of a network. Figure 10 shows one of these neural networks, where each of the connected nodes represents one neuron as the one showed in "Convolutional Neural Networks" section. We call this type of networks fully connected neural networks since all the neurons in one layer are connected with all the neurons in the following layer. The *layer* of a neural network is a collection of neurons operating together at a specific depth within a neural network. The input layer contains the raw data. The hidden layers are the ones between input layers and output layers. The number of hidden layers is another model hyperparameter to be chosen during

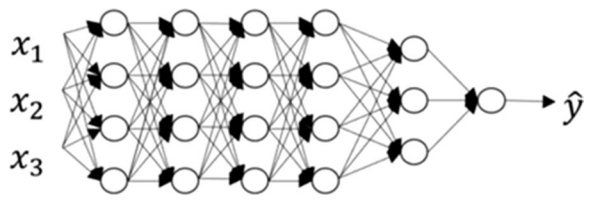


Fig. 10 Neural network with 5 hidden layers

optimization. Figure 10 represents a neural network with 5 hidden layers.

The optimization process described in the previous section, where a function is minimized by iteratively moving in the direction of steepest descent as defined by the negative of the gradient, is called gradient descent.

The learning process described for a single neuron would be the same in a neural network with several layers. The expressions are similar to the ones in Fig. 9, the fact of having many layers will be reflected in the amount of terms in the different expressions when applying the chain rule to calculate the gradient, but the procedure will be similar as in the case of just one neuron.

We have based our example in a 2D problem for clarity purposes, but everything explained here can be trivially generalized to a *N*-dimensional dataset.

Some additional comments

It can be observed that all the expressions in this section are written in matrix form. This is one of the advantages of the gradient descent method: it allows to compress the equations with a very simple notation. Besides, many programming languages work optimally with matrices, speeding dramatically the calculations with respect to looping over all the variables at every iteration. Also, as it was previously mentioned, choosing activation functions whose derivative depends on the function itself allows to have the value of the derivative calculated when performing the backpropagation (explained in "Learning process" section), since it was already obtained in the forward propagation step. This greatly improves the computational performance of the learning algorithm.

On the other hand, an appropriate choice of the model hyperparameters plays a key role in the success of a neural network model. The learning rate, the type of activation functions and the number of hidden layers have already been mentioned as some of the main ones. Another hyperparameter that is worth mentioning is the number of epochs which indicates the number of times that the learning algorithm will work through the entire training dataset. Datasets are usually grouped into batches, especially when the amount of data is very large. During the learning process, the value of the model weights is updated every time the algorithm works through each of the data batches. The amount of data from the original dataset in each of the batches is also a hyperparameter to be set by the machine learning practitioner.

Why the *deep* in deep learning?

The depth of a neural network is its number of neuron layers. Deep Learning refers to the fact of having many more layers than in the so-called Machine Learning algorithms together with all the issues arose from it. The philosophy is the same for both cases but, having more layers usually implies further problems. We summarize here three of the most common issues in Deep Learning:

- Vanishing gradient: the sigmoid activation function presents derivatives very close to zero for most of its range. This is not important when having a few layers, but for very deep neural networks, the product of many values too close to zero can result in a neural network that is unable to learn. This is easy to understand taking a look at the expressions in Fig. 9 and imagining what happens if the *da/dz* factor is very small. For Deep Learning algorithms, specific activation functions are used, such as ReLU [3] or leaky ReLU (see Fig. 11), where the derivatives values (and hence the gradients) are greater than for a sigmoid. Each problem is different, but a rule of thumb can be to use ReLU (or leaky ReLU) for the hidden layers and use the sigmoid only for the last layer when working on a binary classification problem.
- Overfitting problems: Overfitting occurs when a good fit is achieved on the training data, while the model does not generalize well on new, unseen data. This means that the model has learned patterns specific to the training data, which are irrelevant to other new or different data. This can also happen in Machine Learning algorithms, but are more of an issue for Deep Learning models due to the larger amount of neurons. It can be easily understood thinking about the number of neurons as if it were the degree of a polynomial: the greater the degree, the easier it is to find a curve passing by all the N points. But if we add a new point to the dataset, most probably the perfect N-points fit would fail on the $N+1$ point, meaning that the model is not general enough for this specific problem. Several regularization methods have been developed to avoid the overfitting, such as dropout, data augmentation or *L1*/*L2* regularization. The dropout technique consists of randomly dropping out nodes during the training to avoid the over specialization of some neurons. The

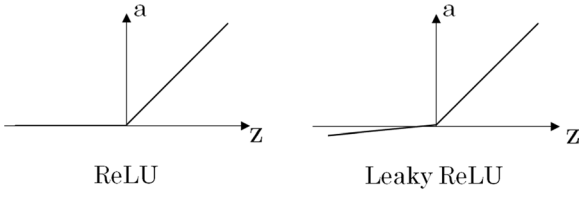


Fig. 11 ReLU and leaky ReLU activation functions

data augmentation technique is used to increase the amount of data by adding slightly modified copies of already existing data or newly created synthetic data from existing data, so that the algorithm does not *see* the same data twice. When working with images, the modifications can be geometric transformations, flipping, color modification, cropping, rotation, noise injection, random erasing, etc.

The $L1/L2$ regularization techniques consist of adding a penalty term to the loss function: the absolute value of the magnitude of the network weights for L1 and the squared magnitude for $L2$.

When minimizing this modified loss function, the penalty term (that is always positive since it is either a magnitude or its squared value) will advantage network weights as small as possible while still trying to optimize the network performance. This will help the neural network to regularize itself, since it will favor simpler models. This can be understood again in terms of a fit to a polynomial: trying to have smaller polynomial coefficients will lead to a simpler model that will not try to perfectly fit every single outlier in the training dataset, but to give a more general result that will potentially have a better performance when exposed to new data.

In any case, the generalization to new data is still one of the major problems of AI, especially in the medical context where datasets are sometimes not as large, varied and balanced as desirable.

- Convergence problems: a fast convergence of the gradient descent algorithm to the minimum is not guaranteed. The optimization algorithm to be used is another hyperparameter and a correct choice can mean the difference between good results in minutes, hours, or days. Several optimization algorithms have been developed to improve the convergence problems, such as the Adam optimization [4]. This algorithm uses the squared gradients to scale the learning rate and it takes advantage of adaptive momentum by using the moving average of the gradient instead of gradient itself.

Convolutional neural networks

Until now, we have described how the learning algorithms work on general N-dimensional data points, but all what has been explained previously can be applied to images. The computational structure of images was described in Sect. 3, and it was already explained that, at the end of the day, images are just matrices and can be treated in the same way as we have shown in the previous sections. Nevertheless, to deal with images we must introduce Convolutional Neural Networks. This type of Neural Networks has established the state of the art in computer vision since 2012, when they beat all their competitors at the ImageNet Large Scale Visual Recognition Challenge in 2012 [5].

The convolutional neural networks employ a specialized kind of linear operation called convolution. Convolutional networks are simply neural networks that use convolution in place of general matrix multiplication in at least one of their layers. Figure 12 shows the general architecture of a ConvNet where multiple filters are taken to slice through the image and map them one by one learning different portions of an input image. One can imagine a small filter sliding left to right across the image from top to bottom and that moving filter is looking for, say, some vertical edge. Each time a match is found, it is mapped out onto an output image called feature map.

An example of a filter can be seen in Fig. 13. The image on the left shows the pixel representation of a vertical line filter and the image on the right shows its visualization.

During the sliding through the input images in the neural network, the filter pixel values are multiplied by

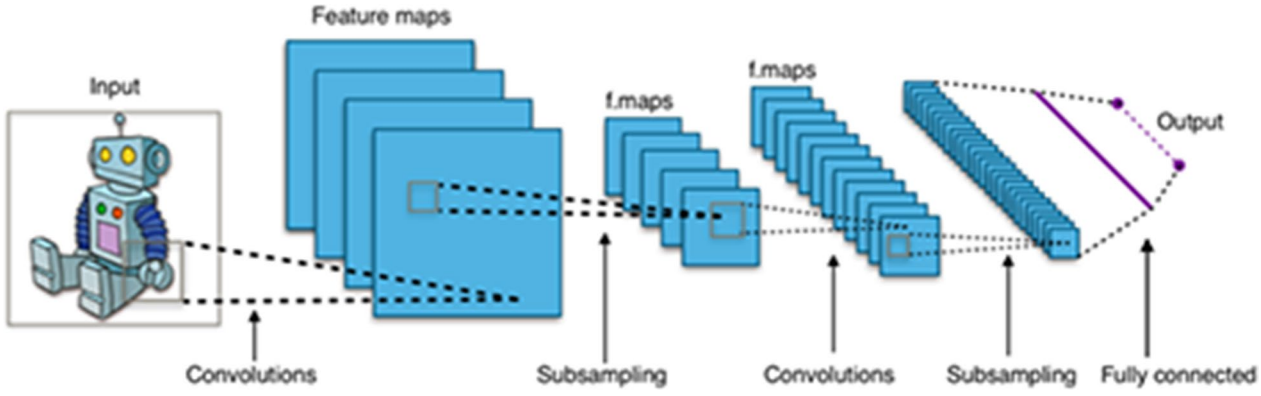


Fig. 12 General convolutional neural network architecture

0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	0	0	0	0

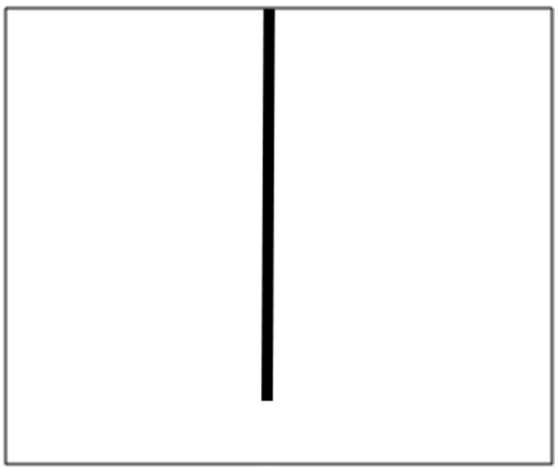


Fig. 13 Pixel representation of a vertical line filter (left) and its visualization (right)

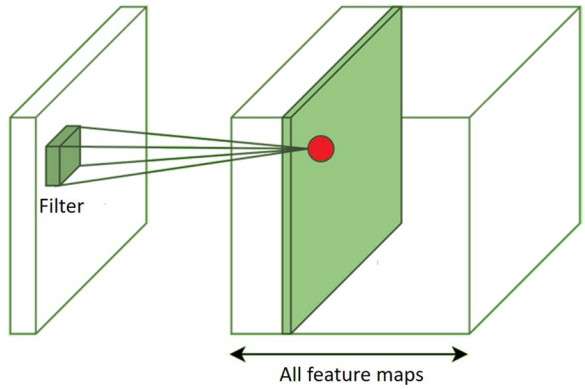


Fig. 14 Extraction of features from the original image by using sliding filters

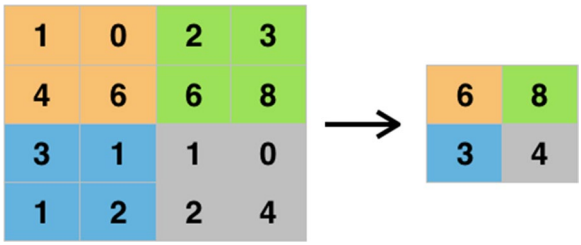


Fig. 15 Max-pooling applied to a feature map, reducing its dimensionality

the pixel values of the image section. If the image presents some feature similar to the vertical line at that particular position, the result of the multiplication will be a high value scalar while, if the shape at that point is completely different from a vertical line, the resulting scalar will be of smaller value. The feature map is then a matrix formed by all these scalars, giving information on the presence and on the location of the particular feature: if the original input image had a vertical straight line on the top left corner, the corresponding feature map matrix will present greater values on that same corner. This means that, each feature map is the mapping of a certain feature in the original image (see Fig. 14). Hence, the convolutional part of a neural network creates a new representation of the original image by extracting and separating the main relevant features of it. By *main relevant features*, we mean here the optimal features for solving a certain problem such as how to distinguish the two different varieties of glioblastoma mentioned in Sect. 2 or the classification of the objects that a self-driven car has in front of it.

Since usually various dozens of filters are needed to solve an image classification problem and each filter generates a corresponding feature map that is given as input to the following layer in the neural network, the problem's size can scale very quickly.

To alleviate this effect, there are several methods to reduce the dimensionality of the feature maps. One of the most widely used is the so-called *max-pooling*.

It consists on going through the feature map, usually with a sliding window of 2×2, taking the maximal pixel value on that square. A graphical illustration of this can be seen in Fig. 15.

Another well-known type of pooling is the *average pooling* which, instead of taking the maximal value, takes the average of all the values in the feature map.

For the convolutional neural networks, the filter values are the ones to be learnt (the W and b following our terminology). As for the fully connected neural networks, the first iteration will have random filters, and all the

optimization process will take place by applying the gradient descent method until it reaches the optimal set of filters.

After the convolutional part where the features are extracted, these features are given as input to a fully connected network that will perform the classification as explained in previous sections. The fact that the features of interest are learnt during the optimization represents another great step forward with respect to more classical algorithms in computer vision, where some expert should first extract by hand the features under consideration for solving the problem (feature engineering) and then give them as input to the classification algorithm.

An interesting fact that makes the convolutional neural networks so competitive when dealing with images is that they are translation invariant. This means that if one shifts an image a bit, it would have a similar activation map as the image before shifting. This is because the convolution is a feature detector, independently of the position of the feature. The translation invariance of the convolutional neural networks allows to learn using only very few parameters with respect to classical methods.

Transfer learning

The traditional supervised learning approach breaks down when we do not have sufficient labelled data for the task we care about to train a reliable model. This is often the case when dealing with medical images. Transfer learning is the capacity of learning a new task through the transfer of knowledge from a related task that has already been learned. This can be achieved thanks to the hierarchical way in which the neuronal networks learn. The first layers in a model working to classify images will learn very general visual features such as intersections, straight lines, simple curves, dots. These basic image components are common to any type of picture we will be working with. As one goes deeper into the layers of the model, the algorithm will extract more complex features that are also more specific of the particular problem being studied.

This means that we can train a model to perform image classification using a big dataset (typically ImageNet [6]) and then re-use it to perform another task where we do not have that many data. For doing so, the first layers of the neural network can be left frozen during the new training, i.e., the filter weights will not change. The small dataset will then be only used to learn the filter weights on the last layers who are in charge of the problem specific features extraction. This allows to make a better exploitation of small datasets, fully using the relevant information to extract the most meaningful features and *transferring* the most general parts of the model from a different problem where more data are available. This technique is widely used nowadays in the medical field. As an example, an extensive review of its use for the diabetic retinopathy case can be found at [7].

Limitations

There is no doubt that deep learning has managed to achieve a very important scientific milestone in many different areas, including medicine. But these new techniques still present some limitations to be widely used in the clinical practice. One of the problems come from the variability of the data itself (e.g., contrast, resolution, signal to noise) which make the Deep Learning models suffer from a poor generalization when the training data come from different machines (different vendor, model, etc.) with different acquisition parametrization or any underlying component that can cause the data distribution to shift. These over-parametrized models have a high tendency to rely on superfluous correlations and are very sensitive to any shift caused by external factors such as different scanner or acquisition protocols.

This generalization gap can be partially mitigated through some techniques. Some of them have already been introduced in this article, and some of them are still being widely explored. The easiest one relies on the fact that all deep learning methods perform better when there is more data to train the model. This is also a problem in medicine, since it is not easy for single sites to generate a large amount of data and manual labels. Working in multi-center initiatives and crowd sourcing can be an efficient approach to achieve such useful resources. Other approaches that will have a great impact in this kind of problem rely on generative models. One of the disadvantages is that the most prominent generative models such as generative adversarial networks (GANs) [8] usually require a large amount of data, and it can take non-desirable shortcuts to model the underlying distribution. Recent likelihood-based models [9] showed some improvements; however, it is still very difficult to model such high dimensional distributions. Another approach, also concerning unlabelled data, is the use of semi-supervised learning methods, that can yield improvements even when working with small datasets. A possible way to minimize the problem of creating a great amount of manual annotations is to use active learning, where the most uncertain predictions are selected for manual correction before re-training the model.

Another big issue is the lack of model interpretability and explainability of the deep learning models. This is common to all areas, but in some of them, such as medicine and health care, not addressing such challenge might

seriously limit the chances of adoption, in real practice, of computer-based systems that rely on these complex nonlinear models. Currently, many techniques tackling this very important issue are being explored and great advances are being achieved in this direction [10].

Finally, it is necessary to address one of the main problems that is transversal to all the other mentioned issues: the feedback loop between the deep learning practitioner and the health professional is paramount to be able to make real advances and to build robust models both from a mathematical and from a medical point of view. Projects with a multidisciplinary approach, containing people from different domains, are thus essential to make the most of these very powerful techniques and to be able to use them confidently in a clinical routine.

Conclusions

This document summarizes for non-deep learning experts and clinicians in particular, the main aspects to understand how neural networks work, placing emphasis on the convolutional neural networks that represent the state-of-the-art algorithms for image analysis. The concepts of automatic learning and data representation have been reviewed together with the functioning of the neurons in a neural network and the main advantages and problems, emphasizing that deep learning is not just a multilayered machine learning approach, but it also has to do with the consequent improvement and optimization of the learning algorithms to make neural networks more robust. The intention of this work was to introduce the basics and set a strong foundation for clinicians on the topic so that they can continue building on top of it with more advanced concepts.

Abbreviations

AI: Artificial intelligence; ConvNet: Convolutional neural network; MRI: Magnetic resonance images.

Acknowledgements

We acknowledge support of the publication fee by the CSIC Open Access Publication Support Initiative through its Unit of Information Resources for Research (URICI). We acknowledge the support from the Advanced Computing and e-Science group at the Institute of Physics of Cantabria (IFCA-CSIC-UC) and the Radiology Department at the Hospital Universitario Marques de Valdecilla (Santander, Spain).

Authors' contributions

All authors have contributed to the general writing and revision of the manuscript. All authors read and approved the final manuscript.

Funding

Consejo Superior de Investigaciones Cientificas (JS-CSIC-BMCSO-0920) DeepHybrid DataCloud (H2020—Grant agreement No 777435) Servicio Cantabro de Salud.

Availability of data and materials

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Advanced Computation and e-Science, Instituto de Fsica de Cantabria - CSIC, Santander, Spain. [2]Servicio de Radiodiagnostico, Hospital Universitario Marques de Valdecilla, Santander, Spain. [3]Instituto de Investigación Sanitaria Valdecilla (IDIVAL), Santander, Spain. [4]Departamento de Morfologa y Biologa Celular, Universidad de Oviedo, Oviedo, Spain. [5]Facultad de Ciencias de la Salud, Universidad Autonoma de Chile, Santiago de Chile, Chile.

Received: 16 February 2021 Accepted: 1 July 2021
Published online: 12 August 2021

References

1. Alan MT (1950) Computing machinery and intelligence. Mind 59:433–460
2. LeCun Y, Cortes C (2010) MNIST handwritten digit database
3. Hara K, Saito D, Shouno H (2015) Analysis of function of rectified linear unit used in deep learning. In: 2015 International joint conference on neural networks (IJCNN). IEEE, pp 1–8
4. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980
5. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
6. http://www.image-net.org/
7. Kandel I, Castelli M (2020) Transfer learning with convolutional neural networks for diabetic retinopathy image classification: a review. Appl Sci 10(6):2021
8. Menick J, Nal K (2018) Generating high fidelity images with subscale pixel networks and multidimensional upscaling. arXiv preprint arXiv:1812.01608
9. Liu Q, Yu L, Luo L, Dou Q, Heng PA (2011) Semi-supervised medical image classification with relation-driven self-ensembling model. IEEE Trans Med Imaging 39(11):3429–3440
10. Fan F, Xiong J, Li M, Wang G (2021) On interpretability of artificial neural networks: a survey. arXiv:2001.02522

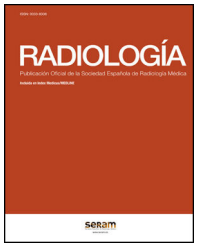
Publisher's Note

9.5.3. Artículo científico en la revista *Radiología*: Inteligencia artificial en Radiología - introducción a los conceptos más importantes

Radiología xxx (xxxx) xxx–xxx

RADIOLOGÍA

www.elsevier.es/rx



ACTUALIZACIÓN

Inteligencia artificial en Radiología: introducción a los conceptos más importantes

A. Pérez del Barrio[a,*], P. Menéndez Fernández-Miranda[a], P. Sanz Bellón[a], L. Lloret Iglesias[b] y D. Rodríguez González[b]

[a] *Servicio de Radiodiagnóstico, Hospital Universitario Marqués de Valdecilla, Santander, Cantabria, España*
[b] *Instituto de Física de Cantabria, Consejo Superior de Investigaciones Científicas (CSIC), Universidad de Cantabria, Santander, Cantabria, España*

Recibido el 7 de noviembre de 2021; aceptado el 15 de marzo de 2022

PALABRAS CLAVE
Inteligencia artificial;
Radiología;
Aprendizaje automático;
Aprendizaje profundo;
Redes neuronales artificiales;
Imagen médica

Resumen La interpretación de la imagen médica es una de las principales tareas que realiza el radiólogo. Conseguir que los ordenadores sean capaces de realizar este tipo de tareas cognitivas ha sido, durante años, un reto y a la vez un objetivo en el campo de la visión artificial. Gracias a los avances tecnológicos estamos ahora más cerca que nunca de conseguirlo y los radiólogos debemos involucrarnos en ello para garantizar que el paciente siga siendo el centro de la práctica médica.

Este artículo explica de forma clara los conceptos teóricos más importantes de esta área y los principales problemas o retos actuales; además, aporta información práctica en relación con el desarrollo de un proyecto de inteligencia artificial en un servicio de Radiología.

KEYWORDS
Artificial intelligence;
Radiology;
Machine learning;
Deep learning;
Artificial neural networks;
Medical imaging

Artificial Intelligence in Radiology: an introduction to the most important concepts

Abstract The interpretation of medical imaging tests is one of the main tasks that radiologists do. For years, it has been a challenge to teach computers to do this kind of cognitive task; the main objective of the field of computer vision is to overcome this challenge. Thanks to technological advances, we are now closer than ever to achieving this goal, and radiologists need to become involved in this effort to guarantee that the patient remains at the center of medical practice.

* Autor para correspondencia.
Correo electrónico: amaia.pdb@gmail.com (A. Pérez del Barrio).

https://doi.org/10.1016/j.rx.2022.03.003
0033-8338/

Cómo citar este artículo: A. Pérez del Barrio, P. Menéndez Fernández-Miranda, P. Sanz Bellón et al., Inteligencia artificial en Radiología: introducción a los conceptos más importantes, Radiología, https://doi.org/10.1016/j.rx.2022.03.003

This article clearly explains the most important theoretical concepts in this area and the main problems or challenges at the present time; moreover, it provides practical information about the development of an artificial intelligence project in a radiology department.

Inteligencia artificial, aprendizaje automático y aprendizaje profundo

Conceptos básicos

La inteligencia artificial (IA) se define como *la capacidad de las máquinas de realizar tareas intelectuales habitualmente realizadas por humanos*[1]. Este término se utiliza como concepto general que engloba tanto el aprendizaje automático (AA) como el aprendizaje profundo (AP). Ambos conceptos pertenecen a un subcampo de la IA que se caracteriza por crear sistemas que son capaces de *aprender*, es decir, capaces de generar sus propias reglas sirviéndose únicamente de los datos *(Data-driven Artificial Intelligence)*[2].

Algunos autores diferencian entre el AA y el AP basándose en que, en el primero, existe intervención humana en el entrenamiento del algoritmo a través de la manipulación de los datos (extracción y selección de las características más importantes), mientras que en el AP la intervención humana es mínima dado que no existe este paso previo[2]. Sin embargo, un enfoque más correcto es considerar el AP como una evolución del AA: los sistemas de AP son sistemas de AA, pero más profundos (y de ahí viene su nombre), es decir, constan de muchas más capas y, precisamente, son estas capas extra las que les confieren la capacidad de extraer las características más relevantes de los datos por sí solos (fig. 1)[2]. En este campo, el concepto de *característica* se refiere a aquellas variables o propiedades de los datos que son mensurables como, por ejemplo, el valor del píxel o la edad del paciente. Las características más relevantes serán aquellas que ayuden en la resolución del problema que se plantea. Otro concepto importante es el modelo de aprendizaje del sistema de AA o AP, ya que se distinguen 3 tipos (fig. 2)[3].

La IA y, especialmente el AP, han protagonizado innumerables artículos en los últimos años, muchos de ellos relacionados con la Radiología. Sin embargo, estos conceptos no son tan novedosos como se cree. En realidad, la IA surgió a mediados de los años 50[4] y, a lo largo de su historia, ha presentado momentos de estancamiento (los llamados *inviernos de la IA*) y momentos de repunte. Actualmente, estamos viviendo uno de esos momentos de repunte sin precedentes, principalmente gracias al desarrollo de la tecnología necesaria para su funcionamiento óptimo como, por ejemplo, las unidades de procesamiento gráfico. En el ámbito médico, la Radiología es una de las especialidades que más se está viendo revolucionada por estos nuevos sistemas[5].

Principales hitos en el desarrollo de la inteligencia artificial

Desde finales de los años 50, una serie de sucesos relacionados con la IA tuvieron gran impacto mediático.

En 1970-1976, el *teorema de los 4 colores*, problema matemático no resuelto, consigue probarse gracias a la ayuda de un ordenador, convirtiéndose así en el primer ejemplo de la inclusión de los ordenadores en la resolución de problemas humanos[6]. Es en esta misma década cuando nace el término *inteligencia aumentada* para expresar el uso de los ordenadores dedicados a ensalzar la cognición humana[7].

Posteriormente, en 1997, el ordenador *Deep Blue* de IBM consigue vencer en un torneo de ajedrez al campeón del mundo, Garry Kasparov. *Deep Blue* poseía información de miles de partidas previas y era capaz de analizar todas las posibles situaciones de los siguientes[6–8] movimientos: se trataba de IA *simbólica (knowledge-based artificial intelligence)* (fig. 1), puesto que estos sistemas no *aprendían* nada, sino que simplemente aplicaban las reglas del juego programadas por humanos, con la ventaja que tienen los ordenadores de poder procesar muchos datos en poco tiempo[8].

En 2015, el sistema *AlphaGo*, desarrollado por Google DeepMind, se convierte en el primer sistema en vencer a uno de los mejores jugadores de go, un juego con reglas sencillas, pero más complejo que el ajedrez en el aspecto estratégico[9]. Este sistema se basaba ya en técnicas de AA implementadas a través de redes neuronales de AP, puesto que el algoritmo era capaz de inferir o *aprender* él solo las reglas del juego con base en los datos de partidas previas.

Acercándonos al momento actual, en 2017, se presenta *AlphaZero* que, a diferencia de *AlphaGo*, es capaz de aprender enfrentándose a sí mismo, es decir, se basa en aprendizaje por refuerzo y no se le proporcionan previamente datos de partidas anteriores, evitando, por tanto, cualquier intervención humana. Con tan solo unas pocas horas de entrenamiento autónomo, este algoritmo fue capaz de ganar al go a otros programas y versiones previas[10].

En la actualidad, son las redes neuronales profundas, generalmente entrenadas mediante aprendizaje supervisado, los sistemas con mayor éxito en el ámbito médico y científico[11]. Estas redes se engloban dentro del AP dado que aprenden directamente de los datos y sin necesidad de estar estos previamente seleccionados por los humanos. El desarrollo de estas técnicas ha supuesto un cambio de paradigma en este campo y, sobre todo, en el análisis

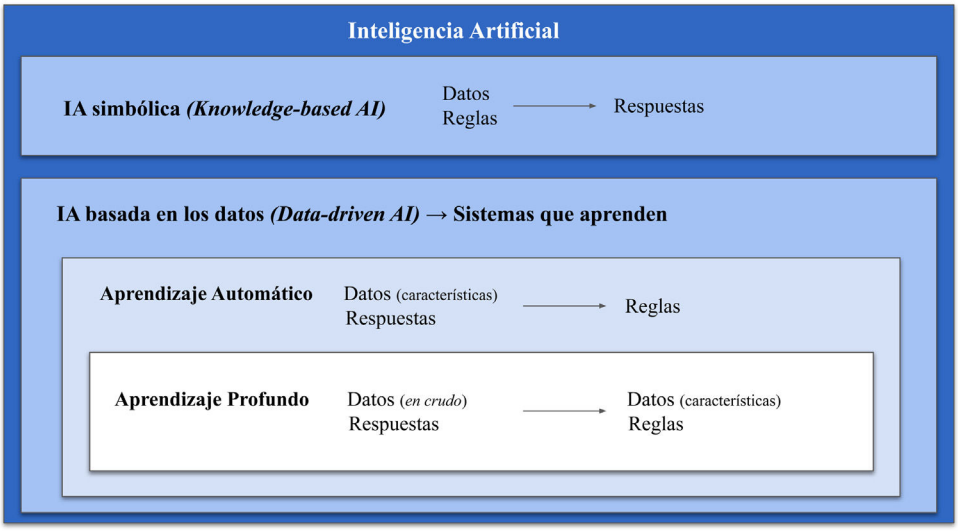


Figura 1 Esquema que engloba los diferentes tipos de inteligencia artificial (IA). Se distinguen 2 campos principales: la IA simbólica, formada por sistemas que necesitan la programación previa de unas reglas, y la IA basada en datos, formada por sistemas que *aprenden*, dentro de la cual se encuentra el aprendizaje automático, con algoritmos a los que hay que ofrecerles los datos ya *depurados* y, dentro de este, el aprendizaje profundo con sistemas que, gracias a la mayor cantidad de capas, son capaces de hacer esto ellos mismos.
AI: artificial intelligence.

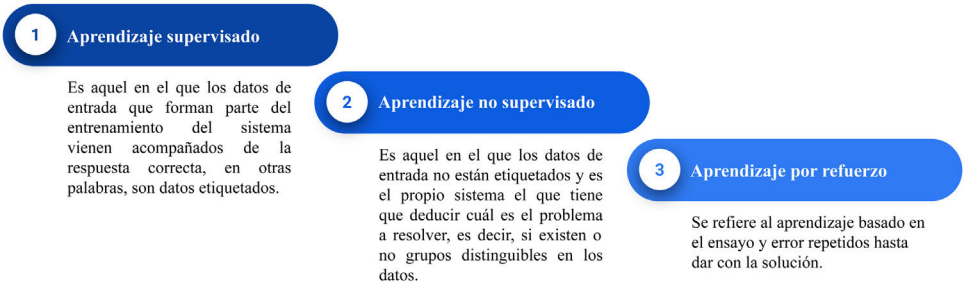


Figura 2 Se distinguen 3 tipos de aprendizaje: supervisado, no supervisado y por refuerzo.

de imágenes y es por ello que son el foco de este artículo.

Las redes neuronales

Las redes neuronales son modelos de predicción, es decir, dados unos datos previos, son capaces de producir una predicción al enfrentarse a datos nuevos. Otros modelos de predicción más conocidos son la regresión lineal simple, la múltiple o la regresión logística. Las redes neuronales obtienen mejores resultados que los anteriores ante problemas más complejos. En general, podemos dividir los modelos de predicción en modelos de clasificación y modelos de regresión. Los modelos de clasificación se basan en encontrar una predicción discreta para la variable de entrada como, por ejemplo, predecir la presencia o no de una enfermedad concreta a partir de una imagen. Mientras que los modelos de regresión se utilizan para encontrar predicciones continuas para la variable de entrada como, por ejemplo, predecir el valor del dímero D a partir de unas variables de entrada (a saber, la edad, la presencia de enfermedad oncológica, etc.). A continuación, se explican las bases de las redes neuronales, comenzando con la neurona artificial y terminando con las redes neuronales convolucionales, el tipo de red neuronal que más éxito ha demostrado en visión artificial.

La neurona artificial

Las redes neuronales artificiales están compuestas por múltiples neuronas artificiales interconectadas, también llamadas perceptrones simples, que se pueden comparar con las neuronas biológicas (fig. 3). La neurona artificial o perceptrón simple consta de varias vías de entrada, que se asemejan a las dendritas de las neuronas biológicas y que transmiten la información hacia el soma. El soma de la neurona artificial es una función que integra toda la información de las entradas y que, tras aplicar una función de activación, genera una salida[3].

La función de activación se podría asemejar al proceso biológico de despolarización de las membranas, que no sigue una función lineal, sino que responde a la *ley del todo o nada*. Las neuronas biológicas reciben muchos impulsos

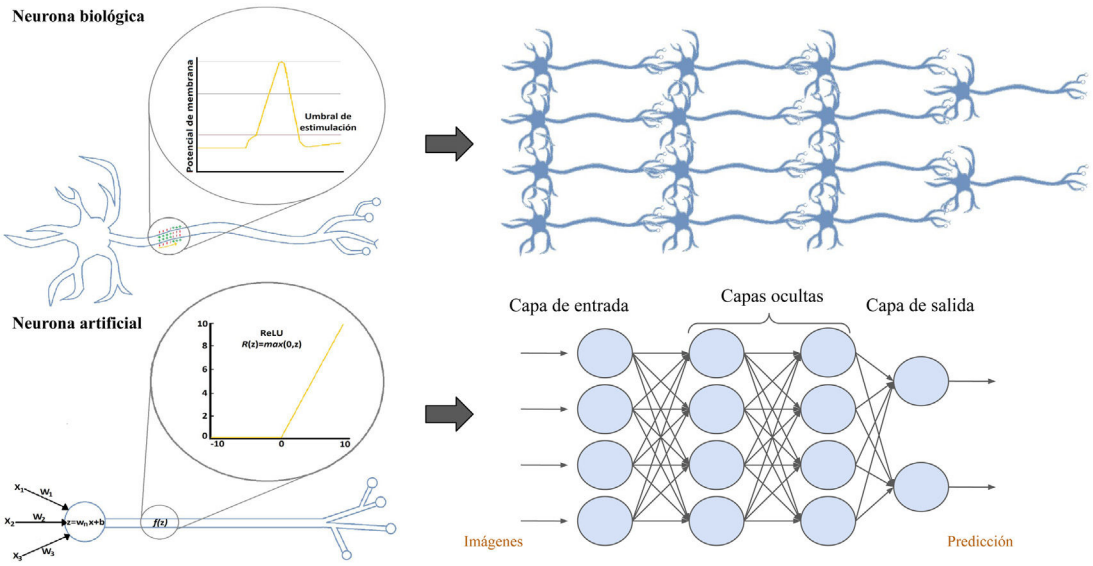


Figura 3 Comparativa entre la neurona biológica y la artificial y la red neuronal biológica y la artificial. La red neuronal artificial se divide en 3 partes principales: la *capa de entrada* es una capa de perceptrones especializados en recibir la información; las *capas ocultas* son aquellas capas capaces de extraer las características de los datos e ir transformándolos en busca de la mejor representación del problema a resolver, y la *capa de salida* es una capa preparada para ofrecer la información de salida, como, por ejemplo, en los problemas de clasificación, la clase a la que corresponde la imagen de entrada según la predicción de la red.

que no consiguen *activarlas* o despolarizarlas, hasta que llega uno con suficiente potencia para despolarizarla consiguiendo generar una salida o potencial de acción que viaja a través del axón, el cual transmitirá el impulso a las neuronas contiguas[12]. A las neuronas artificiales también les llegan una serie de estímulos y, si alguno de ellos logra *activar* la función de activación, esta dará lugar a una salida. La razón matemática que explica que estas funciones sean indispensables es que son las encargadas de introducir la *no linealidad* en la neurona, lo que permite poder aproximarse a funciones mucho más complejas y así resolver, por ejemplo, problemas de clasificación que no sean separables por una recta[3].

La red neuronal artificial clásica

Al igual que las neuronas biológicas se organizan en capas para formar redes neuronales biológicas, las neuronas artificiales hacen lo mismo formando redes artificiales, por lo tanto, la asociación de perceptrones en capas y la concatenación de sucesivas capas es lo que da lugar a una red neuronal (fig. 3)[13].

La arquitectura de las redes neuronales profundas se puede asemejar al modelo biológico de la corteza visual primaria propuesto por D. H. Hubel y T. Wiesel, ambos premios Nobel, en 1959. Según este modelo biológico, la corteza visual primaria está compuesta por 2 tipos de células: las células simples y las células complejas. Las células simples, también llamadas *detectoras de bordes*, responden positivamente al detectar el borde de un objeto en una determinada orientación, mientras que las células complejas utilizan la contribución de las anteriores para encontrar todos los bordes del objeto. La organización de estas células en capas, de modo jerárquico, hace que los objetos se vayan reconociendo de forma secuencial, empezando por las características más simples para acabar por las más complejas. Las redes neuronales mantienen una arquitectura similar: la primera capa de la red se encarga de extraer características groseras de la imagen, como los bordes, el contraste de color, etc. Para, posteriormente, pasar la información por sucesivas capas que van extrayendo detalles más finos[14].

El proceso de aprendizaje o entrenamiento

Antes de comenzar el entrenamiento de la red, se deben seleccionar unas variables llamadas *hiperparámetros*. Los *hiperparámetros* son variables que determinan la estructura de la red y de cómo se entrena, por lo tanto, se definen antes de comenzar el entrenamiento y se van ajustando en función de los resultados del mismo. El tipo de función de activación utilizada y el número de capas ocultas del algoritmo son ejemplos de *hiperparámetros*[2].

El proceso de aprendizaje o entrenamiento de una red neuronal consiste en ajustar unos parámetros llamados pesos. Los pesos se entienden como la intensidad de las conexiones existentes entre las neuronas artificiales. En la neurona biológica, los pesos se podrían asemejar a la intensidad de las sinapsis entre neuronas. Así, ajustando las *sinapsis* conseguiremos un resultado final óptimo[15].

Cada vez que una imagen entra en la red, se activan secuencialmente todas las neuronas de todas las capas de la red, es decir, se van generando unos pesos para cada conexión neuronal, a lo que se le llama *propagación hacia delante*. Finalmente, al llegar a la última capa, gracias a este proceso, se genera una predicción para esa imagen. Pero, dado que todavía estamos entrenando la red, ¿cómo podemos saber si esa predicción es correcta o errónea?

¿Cómo podemos hacer que la red vaya mejorando con cada imagen de entrenamiento?

En un modelo de aprendizaje supervisado, la red neuronal comienza utilizando unos pesos aleatorios y *aprende* cuando esos pesos se van ajustando al comparar los resultados de la red para un ejemplo con el resultado de referencia o etiqueta[11]. Para ello necesita, por un lado, las referencias o etiquetas; por otro, una función que mida el error generado *(función de pérdida);* un algoritmo de optimización que calcule la magnitud y la dirección en la que deben modificarse los pesos con el objetivo de minimizar ese error *(descenso de gradiente)*, y, por último, otra función capaz de trasladar este *ajuste* de forma retrógrada a través de la red y que modifique los pesos de cada neurona en función de cuánto de responsable sea esa neurona del resultado final *(propagación hacia atrás)* (fig. 4)[15].

Este proceso de ajuste de pesos es lo que se llama *aprendizaje* de la red y se da principalmente en las capas ocultas. A lo largo de la red, y mediante el ajuste de los pesos, las capas ocultas van formando representaciones de los datos cada vez más complejas, pero que se ajustan cada vez más al problema (fig. 5)[2]. Así, la base tanto del AA como del AP es transformar los datos sucesivamente hasta encontrar la mejor representación que permita resolver el problema. El término *profundo* no hace referencia a un entendimiento más profundo de los datos, sino a aprender capas sucesivas de representaciones cada vez más significativas y el número de estas capas es lo que se conoce como la profundidad del modelo[3].

Para entrenar una red neuronal, es necesario disponer de al menos 2 subconjuntos de datos: un grupo para el entrenamiento propiamente dicho, con el que el modelo ajustará sus pesos de acuerdo con un mínimo en la *función de pérdida*; y otro grupo de datos con el que evaluar el rendimiento de este, denominado el conjunto de validación. Así, se van realizando iteraciones (llamadas *épocas)* sobre estos grupos de datos y el modelo irá obteniendo cada vez mejores resultados que se irán observando en la evaluación del rendimiento del modelo con el conjunto de validación en cada iteración. Si el rendimiento del modelo no es bueno, el experto puede realizar cambios en los *hiperparámetros*.

Finalmente, una vez concluido el entrenamiento, es decir, una vez ajustados tantos los pesos como los *hiperparámetros*, se prueba el modelo con datos nuevos (conjunto de datos denominado test) para evaluar su rendimiento real. Es decir, se expone al modelo a datos nuevos, por ejemplo, imágenes nuevas y no etiquetadas, y se obtiene una predicción; por ejemplo, la clase a la que pertenece esa imagen en un problema de clasificación. Los datos del test no deben nunca utilizarse para modificar pesos o *hiperparámetros* del modelo.

Las redes neuronales convolucionales

Con la extensión del uso de las redes neuronales clásicas, empezaron a surgir problemas que impulsaron el desarrollo de formas más complejas de redes neuronales. En el caso de la imagen y el reconocimiento de objetos, el principal problema era que, generalmente, el mismo objeto podía tener formas y posiciones diferentes, lo que reducía el rendimiento de las redes. Así, surgieron las redes neuronales convolucionales (RNC), las más usadas para imagen médica[15].

Las redes neuronales clásicas mencionadas anteriormente están compuestas por capas totalmente conectadas. Esto significa que todas las neuronas de una capa están conectadas con las de la siguiente capa y, por lo tanto, la imagen se interpreta en su totalidad, tomando como entrada el valor de todos los píxeles y realizando operaciones que incluyen toda la información de la imagen. Así, si por ejemplo la red tiene como objetivo aprender a identificar coches y en una de las imágenes aparece un coche en la esquina superior izquierda y en otra en la esquina inferior derecha, la red tendrá que aprender unos pesos diferentes para cada una de esas imágenes, dado que la diferente localización del mismo objeto hace que sean interpretados como objetos diferentes, cada uno con sus pesos y representaciones específicas. Esto hace que estas redes no funcionen bien ni sean eficientes en tareas como la interpretación de la imagen o la identificación de objetos. Por el contrario, las RNC disponen de unas matrices denominadas *filtros* capaces de analizar la composición de la imagen y que conceden a la red la capacidad de identificar el coche independientemente de su localización, lo que las hace mucho más eficientes que las redes neuronales clásicas para la interpretación de la imagen[15].

Cada capa convolucional de una RNC puede constar de varios *filtros*. Estos *filtros* son matrices numéricas que van recorriendo la imagen realizando operaciones de convolución sobre grupos de píxeles, dando lugar a mapas de características. Cada *filtro* representa una característica (fig. 6). Así, capa tras capa, se van extrayendo características cada vez más complejas y se van formando representaciones cada vez más *groseras* de los datos de entrada. Es habitual que las capas convolucionales se sigan de capas de *pooling*, capas que reducen la dimensionalidad de los mapas y así el coste computacional (fig. 7). Las representaciones de la última capa de la parte convolucional son transformadas a un vector final a través de una o más capas completamente conectadas y, finalmente, a una predicción. A esta segunda parte de la red se le denomina comúnmente el clasificador[15].

En conclusión, las RNC, gracias a los *filtros*, aprenden patrones locales y, por lo tanto, son capaces de reconocer dicho patrón independientemente de que se realice una traslación, mientras que las redes neuronales clásicas aprenden patrones globales y no son capaces de abstraerse de la localización, orientación o forma del objeto en la imagen. Así pues, las RNC han demostrado ser las más adecuadas para trabajar con imagen médica, siendo capaces de realizar tareas complejas como la clasificación de imágenes[2].

Principales problemas de las redes neuronales y algunas soluciones

Entre los obstáculos a los que nos podemos enfrentar durante el entrenamiento de una red neuronal y que somos capaces de detectar gracias al conjunto de validación destacan el sobreajuste y el subajuste. El sobreajuste ocurre cuando el modelo se especializa tanto en los datos de entrenamiento que no es capaz de generalizar y, por lo tanto, al enfrentarse a datos nuevos no obtiene buenos resultados. El

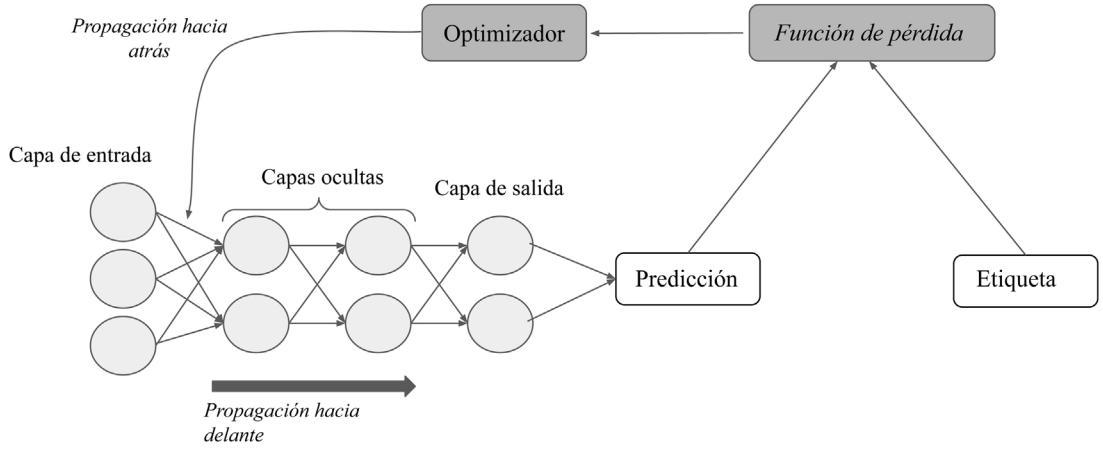


Figura 4 El proceso de aprendizaje supervisado (el entrenamiento). Mediante la *función de pérdida* se cuantifica la diferencia entre la predicción de la red y la etiqueta de cada entrada; a continuación, a través de la *propagación hacia atrás* de este error y del algoritmo de optimización, se ajustan los pesos de las distintas neuronas hasta que correspondan con un mínimo en *la función de pérdida*.

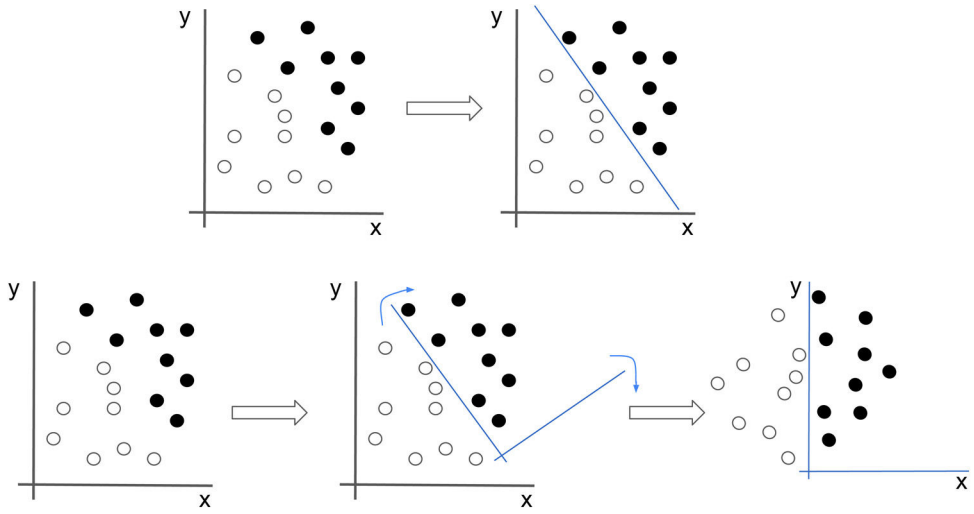


Figura 5 Las redes neuronales buscan la mejor representación de los datos que les permita resolver el problema. En este ejemplo, si tratamos de clasificar los puntos en negros y blancos tendríamos que trazar una recta que correspondería con una ecuación no intuitiva. Sin embargo, si aplicamos una transformación en los datos que hagan que la imagen rote, el problema, de repente, se vuelve mucho más sencillo ($x = 0$).

subajuste, por su parte, se refiere a un modelo que, debido a su insuficiente especialización o excesiva simplicidad, no es capaz de obtener buenos resultados ni siquiera con los datos de entrenamiento. En ninguno de los casos el modelo habrá encontrado aquellas características importantes que le permitirían resolver el problema de forma general o con datos nuevos, es decir, en ninguno de los casos el modelo habrá *aprendido* a generalizar. En conclusión, las redes neuronales deben aprender transformaciones y no ejemplos concretos[3].

El sobreajuste está directamente relacionado con uno de los obstáculos más importantes con los que nos encontramos a la hora de desarrollar sistemas relacionados con la imagen médica: la escasez de datos etiquetados. Una de las razones de este problema es que la creación de amplias bases de datos de imágenes debidamente etiquetadas requiere mucho tiempo y esfuerzo por parte del experto, en nuestro caso, del radiólogo. Además, dado que no siempre concuerdan el diagnóstico por imagen con el histológico, se debe tener muy en cuenta qué prueba es la que se debe considerar diagnóstica de la enfermedad a estudio para así crear la etiqueta. La imagen se considera diagnóstica en algunas entidades como las fracturas. Sin embargo, la mayoría de las enfermedades necesitan de otras pruebas para realizar el diagnóstico definitivo, ya sea la histología o los hallazgos clínico-analíticos, como, por ejemplo, el cáncer, en el que en la mayoría de los casos es necesario el resultado histológico para realizar su diagnóstico[16]. A su vez, esta falta de imágenes etiquetadas muchas veces se ve acentuada debido al complicado marco ético y legal en la transferencia de datos de carácter médico[17].

No obstante, existen varios proyectos en marcha con el objetivo de crear amplias bases de datos con imágenes

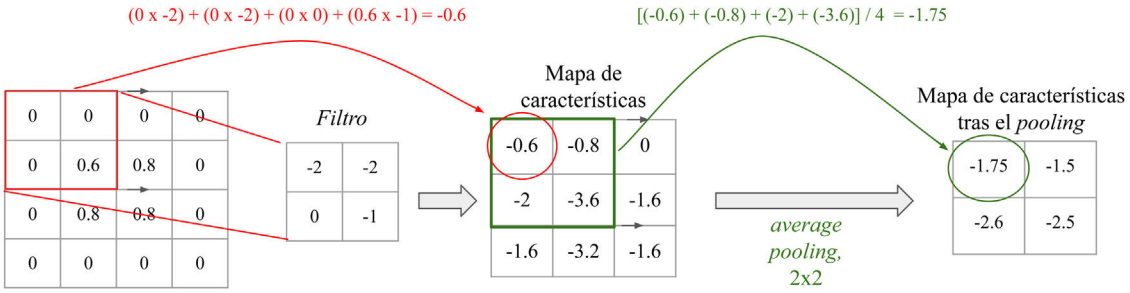


Figura 6 *Filtros* y *pooling*: El *filtro* es una matriz que va recorriendo la imagen y va realizando una operación de multiplicación elemento a elemento y suma para obtener un valor (convolución). Obtenemos valores altos cuando aplicamos el *filtro* sobre una característica similar al mismo y valores bajos cuando lo aplicamos sobre una distinta. En un *filtro* que detecte bordes verticales, si obtenemos valores altos significa que se ha detectado un borde vertical. ¿Recuerda esto a las células de la corteza visual primaria? Por otro lado, la capa de *pooling*, de tipo *average* y de tamaño 2 × 2 en este ejemplo, realiza la media de los valores del rango, reduciendo la dimensionalidad del mapa de características.

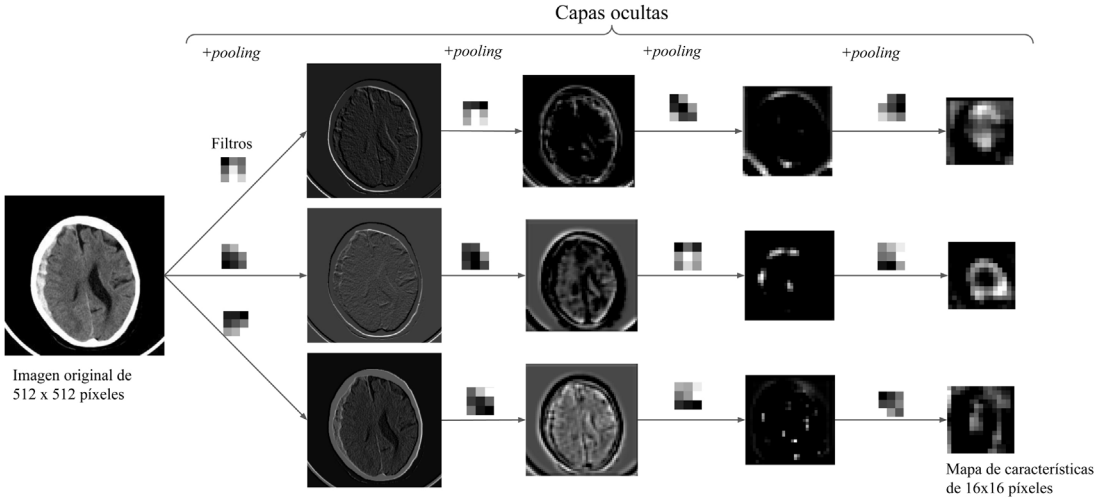


Figura 7 Ejemplo con una imagen: se muestra un ejemplo en el que la imagen de una TC de cráneo va pasando por sucesivas capas que contienen 3 *filtros* de 3 × 3 y la función de *pooling*, lo que da lugar a diferentes mapas de características cada vez de menor resolución espacial. A mayor profundidad en la red, los mapas son cada vez más groseros y de menor tamaño, transmitiéndose hacia delante solo la información más relevante.

médicas etiquetadas, como el Cancer Imaging Archive[18], o empresas como Savana[19], que ofrecen soluciones de IA para la explotación de los datos médicos en formato de texto libre. También se plantean estrategias como el informe interactivo, en el que el radiólogo puede crear vínculos (*hipertexto*) a otros textos o etiquetas en el propio informe[16]; el informe estructurado; o incluso proyectos de colaboración internacionales que involucran a muchos radiólogos, como la preparación del conjunto de datos para el *RSNA 2019 Brain CT Hemorrahge Challenge*[20]; plataformas como OpenNeuro[21], que facilita el acceso a bases de datos tanto de imágenes cerebrales como de electroencefalogramas; la red europea de imagen de tumores cerebrales ENBIT[22], o consorcios como ENIGMA[23], que une a investigadores en genómica e imagen cerebral.

Otra de las soluciones que más éxito está demostrando, principalmente en el ámbito médico, es la transferencia de aprendizaje *(transfer learning)*. Esta técnica consiste en poder trasladar a nuestro modelo, desde una red ya entrenada, tanto la arquitectura como los pesos de las primeras capas. Se elige trasladar los pesos de las primeras capas debido al aprendizaje jerárquico de las redes neuronales, ya comentado anteriormente, según el cual son las primeras capas las que se encargan de extraer características más simples, es decir, menos específicas del problema a resolver y que se asume que son comunes para ambos conjuntos de imágenes. En este sentido, se puede o bien entrenar únicamente la última parte de la red, el clasificador, y mantener la parte convolucional *congelada*, o bien entrenar también un número variable de capas de la parte convolucional, a lo que se denomina *descongelar* capas. En cualquier caso, este nuevo modelo inicia su proceso de aprendizaje *con ventaja*, al tener que ajustar los pesos desde una posición favorable en lugar de partir de valores aleatorios. Es por ello que estos modelos pueden obtener buenos resultados con menos datos que aquellos modelos completamente nuevos[24].

Y, por último, uno de los problemas más importante de estos sistemas y, al mismo tiempo, más difícil de solucionar es su escasa transparencia. Ya que, aunque se pueda explicar el proceso matemático mediante el cual se construyen

los algoritmos, no se conoce claramente cómo llegan a sus conclusiones. Es por ello que, todavía a día de hoy, las RNC son consideradas cajas negras[25] y mejorar su explicabilidad es motivo de estudio. Una de las soluciones que está siendo muy utilizada son las *Grad-CAM*[26], sistemas de localización mediante gradiente de las áreas de la imagen en las que el algoritmo se fija para tomar la decisión final. Al mismo tiempo, esta escasa explicabilidad y transparencia dificultan el desarrollo de un marco ético-legal para la regulación de la implementación de estos sistemas en la práctica médica habitual[27].

Desarrollo de sistemas de inteligencia artificial en los servicios de Radiología

Dentro de un servicio de Radiodiagnóstico, los sistemas de IA pueden aplicarse en múltiples áreas, como en tareas relacionadas con la citación de los pacientes[28], la selección del mejor protocolo de imagen y dosis de radiación[29], la colocación del paciente en el equipo[30], el posprocesado de la imagen (reconstrucciones, mejora de la calidad de la imagen, etc.)[31] y, por supuesto, y como ya hemos explicado, en la interpretación de la imagen[32]. En este último campo, no solo se están desarrollando sistemas de IA que realicen un diagnóstico, sino también sistemas capaces de segmentar órganos y detectar lesiones, así como monitorizarlas[33]. Y, yendo un poco más allá, se están estudiando sistemas que predigan, por ejemplo, la supervivencia estimada o la gravedad de la enfermedad en función del tipo de lesión u otros datos clínico-analíticos del paciente[32–34]. La suma de datos clínicos del paciente a los datos propios de la imagen puede aportar mejoras sustanciales en los resultados de estos modelos de IA, lo que ha llevado a crear redes que combinan métodos de AP y de AA, las redes híbridas[34].

Cuando se plantea el desarrollo de un sistema de IA relacionado con la interpretación de imagen médica, lo primero es obtener la aprobación del comité de ética del hospital. Generalmente, para estudios retrospectivos en los que la obtención del consentimiento informado no es factible y los riesgos de fuga de datos médicos son mínimos, el consentimiento informado del paciente suele ser prescindible[35].

Posteriormente, se procede a la selección de aquellos pacientes a incluir en el estudio y a la recolección de sus imágenes. Este es uno de los pasos más importantes en el desarrollo de estos sistemas que, como ya hemos mencionado, dependen, en gran medida, de la cantidad y la calidad de los datos *(data-driven systems)*. A día de hoy, debido a que el etiquetado de las imágenes radiológicas no está extendido y que los sistemas de información radiológicos no están preparados para este tipo de búsquedas, la obtención de imágenes de una enfermedad concreta no es una tarea fácil[35]. Aquí, el procesamiento del lenguaje natural proporciona técnicas capaces de obtener datos estructurados de los informes radiológicos, con resultados muy prometedores[36]. Asimismo, es de gran relevancia la desidentificación de las imágenes que, en el caso del formato DICOM, puede ser compleja. Una vez obtenidos los datos, estos deben ser preprocesados en función del tipo de red neuronal que se vaya a entrenar[16].

Cuando la base de datos ya está preparada, esta se divide en los 3 subconjuntos mencionados anteriormente: el de entrenamiento, el de validación y el de test, con un porcentaje aproximado del 80, el 10 y el 10%, respectivamente[16].

Ofrecemos un repositorio en GitHub donde ponemos a su disposición un ejemplo de una neurona que resuelve un problema de regresión y otro ejemplo de una red neuronal que resuelve un problema de clasificación, en el siguiente enlace: https://github.com/deepMedicalImaging/RedNeuronalArtificial.git.

Autoría

1. Responsable de la integridad del estudio: AP, PM, PS, LL y DR.
2. Concepción del estudio: AP, PM, PS, LL y DR.
3. Diseño del estudio: AP.
4. Obtención de los datos: no procede.
5. Análisis e interpretación de los datos: no procede.
6. Tratamiento estadístico: no procede.
7. Búsqueda bibliográfica: AP.
8. Redacción del trabajo: AP.
9. Revisión crítica del manuscrito con aportaciones intelectualmente relevantes: AP, PM, PS, LL y DR.
10. Aprobación de la versión final: AP, PM, PS, LL y DR.

Conflicto de intereses

Los autores declaran no tener ningún conflicto de interés.

Bibliografía

1. European Society of Radiology. What the radiologist should know about artificial intelligence —an ESR white paper. Insights Imaging. 2019;10(1):44.
2. Chollet F. Deep learning with python. 2ª. ed. Shelter Island (NY): Manning Publications Co;. 2021.
3. Iglesias L, Bellón P, del Barrio A, Fernández-Miranda P, González D, Vega J, et al. A primer on deep learning and convolutional neural networks for clinicians. Insights Imaging. 2021;12:117.
4. Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychol Rev. 1958;65:386–408.
5. Hosny A, Parmar C, Quackenbh J, Schwartz L, Aerts H. Artificial intelligence in radiology. Nat Rev Cancer. 2018;18:500–10.
6. Detlefsen M, Luker M. The four-color theorem and mathematical proof. J Philos. 1980;65:803–20.
7. Von Ahn L. Augmented intelligence: The Web and human intelligence. Philos Trans R Soc A Math Phys Eng Sci. 2013;371, 20120383.
8. Murray C, Hoane AJ. Feng Hsiung Hsu F-H. Deep blue. Artif Intell. 2002;134(1–2):57–83.
9. Chen J. The evolution of computing: AlphaGo. Comput Sci Eng. 2016;18:4–7.
10. Bratko I. AlphaZero —what's missing? Informatica. 2018;42:7–11.
11. Erickson B, Korfiatis P, Akkus Z, Kline T. Machine learning for medical imaging. Radiographics. 2017;37:505–15.
12. Silbernagl S, Despopoulos A. Fisiologia: texto y atlas. 7th ed. Médica Panamericana;. 2021.
13. Soffer S, Ben-Cohen A, Shimon O, Amitai MM, Greenspan H, Klang E. Convolutional neural networks for radiologic images: A radiologist's guide. Radiology. 2019;290:590–606.
14. Hubel D, Wiesel T. Receptive fields of single neurones in the cat's striate cortex. J Physiol. 1959;148:574–91.

15. Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, et al. Deep learning: A primer for radiologists. RadioGraphics. 2017;37:2113–31.
16. Willemink M, Koszek W, Hardell C, Wu J, Fleischmann D, Harvey H, et al. Preparing medical imaging data for machine learning. Radiology. 2020;295:4–15.
17. Schönberger D. Artificial intelligence in healthcare: A critical analysis of the legal and ethical implications. Int J Law Inf Technol. 2019;27:171–203.
18. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository. J Digit Imaging. 2013;26:1045–57.
19. Espinosa L, Tello J, Pardo A, Medrano I, Ureña A, Salcedo I, et al. Savana: A global information extraction and terminology expansion framework in the medical domain. Procesamiento del Lenguaje Natural. 2016;57:23–30.
20. Flanders AE, Prevedello LM, Shih G, Halabi SS, Kalpathy-Cramer J, Ball R, et al., Construction of a machine learning dataset through collaboration: The RSNA 2019 Brain CT Hemorrhage Challenge. Radiol Artif Intell. 2020;2:e190211.
21. Markiewicz CJ, Gorgolewski KJ, Feingold F, Blair R, Halchenko YO, Miller E, et al. The OpenNeuro resource for sharing of neuroscience data. Elife. 2021;10:e71774.
22. INCF. European Network for Brain Imaging of Tumours (ENBIT) [consultado 6 Nov 2021]. Disponible en: https://www.enbit.ac.uk/.
23. Thompson PM, Stein JL, Medland SE, Hibar DP, Vasquez AA, Renteria ME, et al. The ENIGMA Consortium: Large-scale collaborative analyses of neuroimaging and genetic data [consultado 6 Nov 2021]. Brain Imaging Behav. 2014;8:153–82. Disponible en: http://enigma.ini.usc.edu/.
24. Long M, Cao Y, Wang J, Jordan M. Learning transferable features with deep adaptation networks. Proc 32nd Int Conf Mach Learn. 2015;37:97–105.
25. Wadden J. Defining the undefinable: The black box problem in healthcare artificial intelligence. J Med Ethics. 2021: 107529.
26. Selvaraju R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. IEEE International Conference on Computer Vision (ICCV). 2017:618–26.
27. Geis JR, Brady AP, Wu CC, Spencer J, Ranschaert E, Jaremko JL, et al. Ethics of artificial intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement. J Am Coll Radiol. 2019;16:1516–21.
28. Chong LR, Tsai KT, Lee LL, Foo SG, Chang PC. Artificial intelligence predictive analytics in the management of outpatient MRI appointment no-shows. AJR Am J Roentgenol. 2020;215:1155–62.
29. McCollough CH, Leng S. Use of artificial intelligence in computed tomography dose optimisation. Ann ICRP. 2020;49 1_Suppl:113–25.
30. Gang Y, Chen X, Li H, Wang H, Li J, Guo Y, et al. A comparison between manual and artificial intelligence-based automatic positioning in CT imaging for COVID-19 patients. Eur Radiol. 2021;31:6049–58.
31. Higaki T, Nakamura Y, Tatsugami F, Nakaura T, Awai K. Improvement of image quality at CT and MRI using deep learning. Jpn J Radiol. 2019;37:73–80.
32. Nasrullah N, Sang J, Alam MS, Mateen M, Cai B, Hu H. Automated lung nodule detection and classification using deep learning combined with multiple strategies. Sensors (Basel). 2019;19:3722.
33. Kim D, Lee S, Kwon S, Nam W, Cha I, Kim H. Deep learning-based survival prediction of oral cancer patients. Sci Rep. 2019;9:6994.
34. Chieregato M, Frangiamore F, Morassi M, Baresi C, Nici S, Bassetti C. A hybrid machine learning/deep learning COVID-19 severity predictive model from CT images and clinical data. arxiv.org. 2021 (2105.06141).
35. Montagnon E, Cerny M, Cadrin-Chênevert A, Hamilton V, Derennes T, Ilinca A, et al. Deep learning workflow in radiology: a primer. Insights Imaging. 2020;11:22.
36. Pons E, Braun L, Hunink M, Kors JA. Natural language processing in Radiology: A systematic review. Radiology. 2016;279:329–43.

9.5.4. Artículo en proceso de revisión: *Generalization of Deep Learning Algorithms for X-rays: the Influence of the Radiography Device and Other Potential Factors*

Generalization of Deep Learning Algorithms for X-rays: the Influence of the Radiography Device and Other Potential Factors

Pablo Menéndez Fernández-Miranda[a,d,e,*], Enrique Marqués Fraguela[b,d], David Rodríguez González[c,*], Nicolás Ferreiros Vázquez[b,d], Lara Lloret Iglesias[c], José A. Vega[e,f]

[a]*Departamento de Radiología, Hospital Universitario Marqués de Valdecilla, Av. Valdecilla, Santander 39008, Spain*
[b]*Departamento de Radiofísica y Protección Radiológica, Hospital Universitario Marqués de Valdecilla, Av. Valdecilla, Santander 39008, Spain*
[c]*Grupo de Computación Avanzada y e-Ciencia, Instituto de Física de Cantabria, (IFCA), Consejo Superior de Investigaciones Científicas (CSIC), Av. de los Castros, Santander 39005, Spain*
[d]*Instituto de Investigación Sanitaria de Valdecilla, c/ Cardenal Herrera Oria, Santander 3901, Spain*
[e]*Departamento de Morfología y Biología Celular, Universidad de Oviedo, Av. Julián Clavería, 6, Oviedo 33006, Spain*
[f]*Facultad de Ciencias de La Salud, Universidad Autónoma de Chile, Av. Pedro de Valdivia 425, Providencia, Región Metropolitana, Santiago, Chile*

Abstract

Generalization of Deep Learning (DL) algorithms is still a controversial issue whose causes and solutions remain unknown. This issue has been highlighted in the context of the COVID-19 pandemic, which brought about a raise of papers reporting new DL algorithms that aid in the assessment of X-rays. Although these algorithms seem to perform successfully in the training institutions, they often show a lack of generalization to external sites. We present a study of several factors which may hinder DL algorithms' generalization ability, namely: the institution where the data are acquired, the acquisition protocol, the image processing applied by the X-ray machine, and the type of response function of the X-ray machine. To study this, we compared the performance of three CNNs trained for the classification of COVID-19 and Control chest radiographs, using different combinations of data acquired in two institutions by three different models of X-ray machine. Our results found that the type of response function is the main handicap for generalization, while other factors are surmountable. Our results also found that CNN-feature values depend on the X-machine that acquires the image, making differences among images from different X-ray machines one of the potential primary causes behind DL algorithms' generalization deficiency.

Keywords:
2000 MSC: 68T07, 68T20, 68T37, 68T05, 92-08 Deep Learning, Generalization, X-ray machine, Chest Radiograph, COVID-19, causal relationship.

1. Introduction

In December 2019, an initial cluster of a previously unrecognized viral entity termed coronavirus disease-19 (COVID-19) was described in Wuhan, China (Zhu et al. (2020)). Since then, the infection has rapidly spread globally (Srivastava et al. (2020)), and the efforts to improve the quality of the detection and diagnosis methods have been continuously growing. In this context, chest radiographs (CXR) have been widely used both for primary disease evaluation and for the management and follow-up of patients with pneumonia (Borghesi and Roberto (2020)).

However, chest radiography is a complex imaging

*Corresponding authors: pablomenendezfernandezmiranda@gmail.com (P.M. Fernández-Miranda); drodrig@ifca.unican.es (D.R. González)

modality to interpret (Al aseri (2009)), and their evaluation could be a challenging task requiring experience and expertise (Hwang et al. (2019)). Given this background, Deep Learning (DL) classification algorithms may help improve the quality of radiographic interpretation and lead to more accurate diagnoses (Hwang et al. (2019)).

Consequently, different DL algorithms for COVID-19 diagnosis on CXRs have been published in the last year. These algorithms often perform successfully on the training, validation, and test subsets, however, fail catastrophically in deployment when the distribution of data suddenly shifts as a result of a generalization deficiency (Maguolo and Nanni (2021); Pan et al. (2019); Pooch et al. (2020); Sathitratanacheewin et al. (2020); Subbaswamy and Saria (2018); Zech et al. (2018)).

The aforementioned deficiency in DL algorithms' capacity to generalize is often misidentified during the algorithm evaluation. This occurs since algorithms are evaluated on test subsets that are, in reality, holdout sets originating from the same population sample from which training and validation subsets were obtained. Evaluating algorithms through this process allows developers to estimate the algorithm's internal validation performance but not the algorithm's external validation performance, also known as generalization. Therefore, DL algorithms' performance should also be assessed on test subsets coming from a different source than those from which training and validation subsets were drawn (Maguolo and Nanni (2021); Pan et al. (2019); Pooch et al. (2020); Sathitratanacheewin et al. (2020); Zech et al. (2018)).

Additionally, the source from where training, validation, and usually test subsets are obtained, may significantly differ from the real-world environment in which the algorithm is going to be deployed (Pooch et al. (2020); Sathitratanacheewin et al. (2020); Zech et al. (2018)). This issue is especially relevant in medical data, where it has been referred to as distribution or domain shift (Pooch et al. (2020); Sathitratanacheewin et al. (2020)). Distribution shifts can lead to a reduction in the algorithms' performance when they are deployed on the target population, which may result in a lack of generalization. Overcoming this issue is still a big challenge in Machine Learning (Pooch et al. (2020); Sathitratanacheewin et al. (2020)).

Surprisingly, DL algorithms' generalization deficiency for image classification has only been argued by a handful of authors within the medical field. While the causes behind it remain unclear, most of the authors who have addressed this issue have only studied the generalization through testing different DL algorithms on external datasets from external institutions (Pan et al. (2019); Rajpurkar et al. (2020); Sathitratanacheewin et al. (2020); Zech et al. (2018)). Contrary to previous works, this research separately looks at the influence of the institution and the X-ray device on both DL algorithms' internal validation and generalization performances. Another distinct aspect of this paper is the assessment of how the image processing applied by the X-ray device and its type of response function impact DL algorithms' internal validation and generalization.

2. Theory

As mentioned in the Introduction, the lack of generalization of DL algorithms is secondary to distribution shifts, which previous literature attributes to inter-institutional differences (Zech et al. (2018)). This paper analyzes the potential sources of distribution shifts, known to us, affecting DL algorithms for medical image classification. We divided these sources into two categories: X-ray device related factors, and institutional related factors.

Device related factors are those which may change the distribution of pixel values, including the image protocol, the equipment's image processing, and the type of response function applied by the radiography device (Lanca and Silva (2013)). On the other hand, institutional related factors are those which do not modify the image pixel values, and include differences among hospitals, such as the image labeling criteria, population demographics, disease epidemiology, or the workflow in radiology departments.

This investigation hypothesizes that differences in medical images introduced by device related factors are the principal sources of distribution shifts which lead DL algorithms for radiography classification to suffer a lack of robustness and generalization deficiency. In our opinion, radiological images could have different textures depending on the model of X-ray machine that acquires the image, given that each model of X-ray device applies a different image processing and response function. These textural variations among images from different X-ray device

models induce a significant distribution shift that might hinder the algorithm's generalization.

According to our hypothesis, a high level of generalization across institutions could be achieved for the same make and model of X-ray device, thus inter-institutional differences might be surmountable. Additionally, training a different algorithm for each model of radiography device could achieve high performance algorithms using smaller training samples.

With the purpose to test our assumptions, we studied the influence of a variety of factors on both DL algorithms' internal validation and generalization. These factors are: the institution where datasets are obtained; the image acquisition protocol; the image processing applied by the X-ray machine; and the type of response function employed by the X-ray machine.

3. Materials and Methods

We designed three experiments to estimate both the influence of institutional related (labeling, hospital, demographics...) and X-ray device related (acquisition protocol, image processing applied by the X-ray machine, and type of response function of the X-ray machine) factors on DL algorithms' performance (Fig. 1, Fig. 2). Particularly, we assessed the influence of these factors both on internal validation and generalization performances. To accomplish this, three convolutional neural networks (CNNs) were trained to classify COVID-19 and non-COVID-19 CXRs. It is important to highlight that all the DL models trained for this work were CNNs pretrained on ImageNet. Specifically, all the models had the same architecture, which consisted in a VGG16 with a classifier, and were all trained with the same hyperparameters.

3.1. Experiment designs

3.1.1. Experiment 1

The first experiment analyzed the influence of institutional and device related factors on DL algorithm's internal validation performance (Fig. 2). To do so, we compared the performance of three DL models, each trained with 300 images to classify COVID-19 and non-COVID-19 CXRs. The first model, Model-F1F2, was trained with images acquired by a Fujifilm FDR Smart FGX device from Institution-1. The second model, Model-F1F3, was trained with images half acquired by the Fujifilm FDR Smart FGX device from Institution-1, and half by another Fujifilm FDR Smart FGX device from Institution-2. The third model, Model-F1'F3'S2, was trained with 101 images from each of the two Fujifilm FDR Smart FGX devices (Institution-1 and Institution-2) and 98 images from a General Electric Revolution XRD device from Institution-2.

After training, the internal validation performances of Model-F1F2 and Model-F1F3 were compared to estimate the influence of institutional related factors on the models' internal validation performance. This comparison also allowed investigators to evaluate the influence of the acquisition protocol, as the two Fujifilm devices had a different acquisition protocol. Later, the internal validation performance of Model-F1'F3'S2 was compared with that of Model-F1F2 and Model-F1F3. Given that both the image processing and the type of response function were different between the Fujifilm and the General Electric devices, the comparison among these models allowed us to evaluate the influence of X-ray machine related factors on the algorithms' internal validation performance.

Finally, Gradient-weighted Class Activation Mapping (Grad-CAM) heatmaps were used to highlight the important regions in the image for predicting the concept (Selvaraju et al. (2017)).

3.1.2. Experiment 2

The second experiment studied the influence of institutional and device related factors on DL algorithms' generalization (Fig. 2).

This experiment used Model-F1F2, which was trained in Experiment 1 with 300 images acquired by the Fujifilm FDR Smart FGX from Institution-1 to classify COVID-19 and non-COVID-19 CXRs. This model's performance was compared on four test subsets, which included images from two different institutions and three different models of X-ray devices. Specifically, subset T1 contained 94 images acquired by the Fujifilm FDR Smart FGX from Institution-1, subset T2 contained 94 images acquired by the Fujifilm FDR Smart FGX from Institution-2, subset T3 contained 94 images acquired by the General Electric device from Institution-2, and subset T4 contained 44 images acquired by a Carestream device from Institution-2.

It is important to highlight that Fujifilm and Carestream devices had the same type of response function, however

they had a different image processing. On the other hand, the General Electric device had a different image processing and type of response function than the Fujifilm and Carestream devices'. Therefore, the influence of institutional and device related factors on DL algorithm's generalization was estimated by comparing the performance of Model-F1F2 on the four subsets. Similar to Experiment 1, Grad-CAM heatmaps were used to increase the understanding of the DL model's performance.

3.1.3. Experiment 3

The third experiment studied the influence of institutional and device related factors on feature values extracted by an ImageNet CNN (Fig. 2). We hypothesized that radiological images have different textures depending on the model of X-ray machine that acquires the images. Thus, ImageNet CNN-feature values might depend on the model of X-ray machine that acquires the images. As such, algorithms' generalization could be hindered by image differences produced by X-ray devices and ultimately, by device related factors.

To do so, we used a hierarchical clustering algorithm (MJ and Wilson (2002)) to group the test images used in Experiment 2 (images from subsets T1, T2, T3 and T4) into clusters based on their CNN-feature values similarities. Afterwards, we examined if the clusters found by the algorithm corresponded to images from the target classes (COVID-19 or Control), or from other hidden classes, such as the institution or the model of X-ray device.

Through this experiment we evaluated if target classes were the most evident in the dataset based on CNN-feature values, or if other hidden classes, such as the X-ray machine, were even more evident. We termed categories which were more evident than the target classes as high-level hidden classes.

This experiment was conducted four times. First, using images from subsets T1, T2, and T3, and features extracted by Model-F1F2. Later, using a cropped version of the same images to remove metallic tokens from CXRs and ensure that tokens did not influence the clusterization. Then, we used images from subsets T1, T2, T3, and T4, and features extracted by Model-F1F2. Lastly, images from subsets T1, T2, and T3, and features extracted by both a pre-trained and a trained version of Model-F1'F3'S2 were used to evaluate if fine-tuning technique could change the level of importance of the hidden classes.

3.2. Ethical Approval

This research involved patients from two different medical institutions: Hospital Universitario Marqués de Valdecilla, located in Santander, Spain – referred to in the text as Institution-1; and Hospital de Sierrallana in Torrelavega, Spain – referred to as Institution-2. The Ethics Committees of both institutions approved this research.

3.3. Patients Recruitment

Patients were randomly recruited from four databases. The first database contained all patients who had chest radiographs acquired by a Fujifilm FDR Smart FGX from Institution-1 between September 15th, 2019 and November 25th, 2020. The second database incorporated all patients with CXRs acquired by a Fujifilm FDR Smart FGX from Institution-2, during the same period. The third database included all patients with CXRs acquired by a General Electric Revolution XRD from Institution-2 during the period of January 1st, 2020 and November 25th, 2020. Lastly, the fourth database contained all patients with CXR acquired by a Carestream DRX Evolution Plus from Institution-2, from January 1st, 2018 to November 25th, 2020.

Only the first frontal view radiograph from each patient was included to avoid potential biases that could be derived from including multiple images per patient. Afterwards, images were manually labeled by expert radiologists in two classes, COVID-19 and Control, according to the inclusion criteria summarized in Table 1. These classes are referred to as target classes in this research.

3.4. Dataset and Subsets

As previously mentioned, frontal view radiographs from four different radiography devices, including three different models of X-ray machine from two different institutions, were manually labeled as COVID-19 or Control. Among the device models chosen for this research, two of them had a logarithmic response function (Fujifilm and Carestream), whereas the other device model had a linear response function (General Electric) (KCARE (2005)). After image gathering, radiographs were randomly sampled with stratification to build the main dataset. The dataset contained 394 images acquired

Table 1: Inclusion criteria for target classes

Target class	Clinical history	Imaging Findings
COVID-19	Positive RT-PCR testing	Three radiologists, where at least one was a thoracic radiologist, reported COVID-19 compatible findings on the chest radiograph
Control	No symptoms of COVID-19	Three radiologists, where at least one was a thoracic radiologist, reported no pathological findings on the chest radiograph

RT-PCR = reverse-transcriptase polymerase chain reaction.
Meeting all the inclusion criteria was required for an image to be included in a class.

by the Fujifilm FDR Smart FGX device from Institution-1, 244 acquired by the Fujifilm FDR Smart FGX device from Institution-2, 192 acquired by the General Electric Revolution XRD device from Institution-2, and 44 acquired by the Carestream DRX Evolution Plus device from Institution-2 (Fig. 1). These sample sizes were designed based on the number of images available from each digital radiography device after the cleaning and filtering process.

Later, the dataset was split into eight subsets based on the radiography device which acquired the images and on the institution where they came from (Fig. 1). The first and second subsets, F1 and F2, were each comprised of 150 images acquired by the Fujifilm device from Institution-1. The third subset, F3, included 150 images acquired by the Fujifilm device from Institution-2. The fourth subset, S2, contained 98 images acquired by the General Electric device from Institution-2. The fifth, sixth and seventh subsets, T1, T2 and T3, included 94 images each. While images from subset T1 were acquired by the Fujifilm device from Institution-1, images from subset T2 were acquired by the Fujifilm device from Institution-2, and images from subset T3 by the General Electric device from Institution-2. The eighth subset, T4, contained 44 images acquired by the Carestream device from Institution-2. Subsets F1, F2, F3, and S2 were used to train the DL models, while subsets T1, T2, T3, and T4 were used to test the DL models (Fig. 2).

It is important to highlight that all the subsets were stratified by target class, so that each of them contained an equal number of COVID-19 and Control images.

3.5. Image Preprocessing

Images were collected as 16-bit unsigned integer monochrome pixels stored in a DICOM format. After data collection, an image preprocessing was applied. The first step of the preprocessing was to assign an appropriate window. Then, images' pixel values were inverted when required and resized to 512 x 512 pixels using a cubic spline interpolation. Finally, pixel values were rescaled to [0, 1] and images were stacked into three channels, since ImageNet pre-trained models require a three channel image input.

3.6. Image Acquisition Protocols

The most relevant image acquisition protocol parameters for each radiography device are summarized in Table 2.

The acquisition protocols of both Fujifilm equipment (Institution-1 and Institution-2) were compared using the devices' exposure index. In doing so, it was feasible to explore whether a statistically significant difference existed between the acquisition protocol of the Fujifilm device from Institution-1 and the acquisition protocol of the Fujifilm device from Institution-2. The exposure index was selected as the metric to compare the protocols since it measures the air kerma at the detector surface (Butler et al. (2010)). However, each manufacturer often has its own exposure index computed by a different method (Seibert and Morin (2011)), thus exposure indexes were only compared among devices of the same manufacturer.

3.7. Models' Training

Three CNNs with weights pre-trained on ImageNet were fine-tuned. The chosen architecture was a VGG16, as it had satisfactory results when classifying COVID-19 images in other previous works (Das et al. (2021); Shelke et al. (2021); Sitaula and Hossain (2020)). On top the VGG16, a classifier including a Global Max Pooling layer

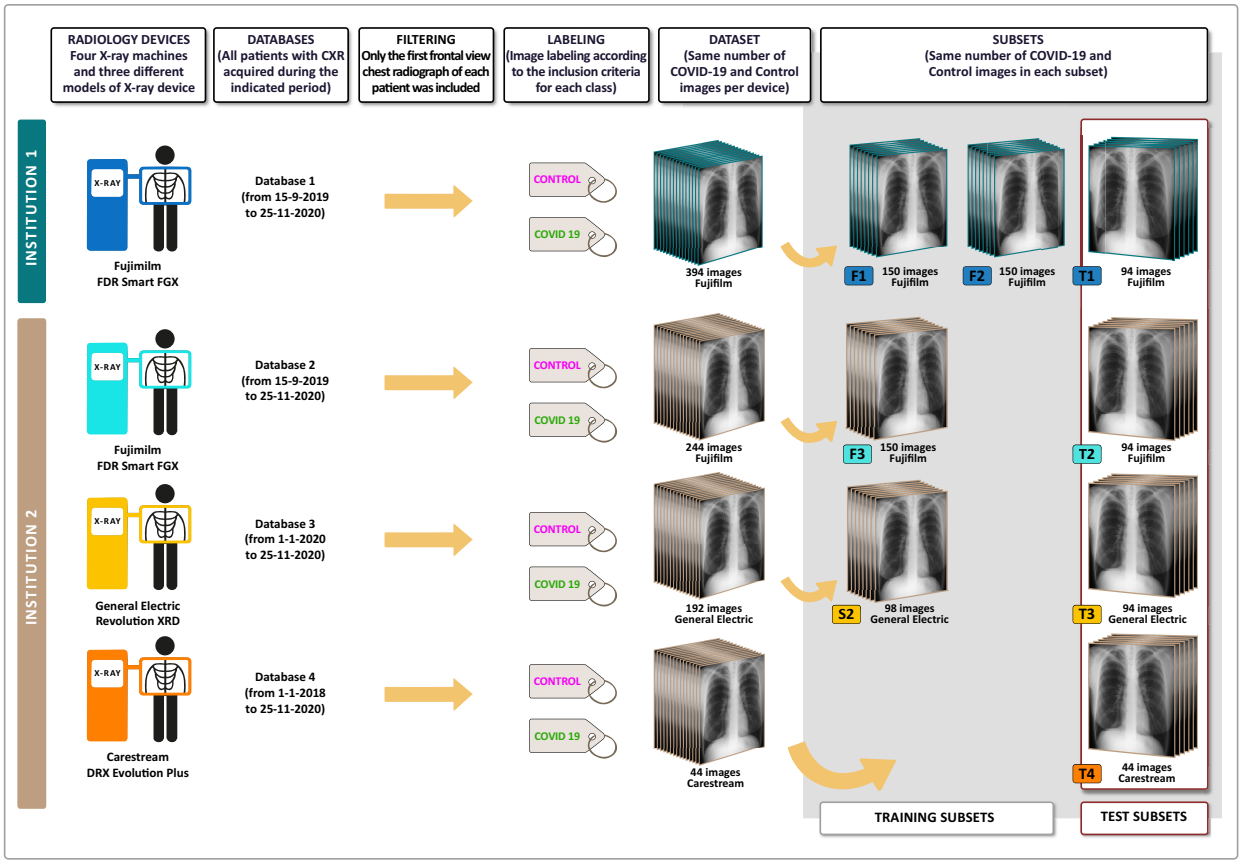


Figure 1: Process by which dataset was collected and split into training and tests subsets.

and an output single neuron with a sigmoid activation function, was added. No layers were frozen during the training. Models were trained following a stratified 5-fold cross-validation with the same hyperparameters (Fig. 2).

The first model, Model-F1F2, was trained with subsets F1 and F2 (300 images Fujifilm - Institution-1). The second model, Model-F1F3, was trained with the subsets F1 and F3 (150 images Fujifilm - Institution-1 and 150 images Fujifilm - Institution-2). Finally, the third model, Model-F1'F3'S2, was trained with 101 images from subset F1 (subset F1'), 101 images from subset F3 (subset F3'), and all images from subset S2 (98 images General Electric - Institution-2) (Fig. 2).

3.8. Models' Evaluation

Models' performance was evaluated through the area under the receiver operating characteristic curve (AUC) on the four test subsets: T1 (94 images Fujifilm - Institution-1), T2 (94 images Fujifilm - Institution-2), T3 (94 images General Electric - Institution-2), and T4 (44 images Carestream - Institution-2).

The first experiment analyzed the influence of institutional and device related factors on models' internal validation performance. To evaluate the effect of institutional

Table 2: Descriptive Statistics of the Most Relevant Acquisition Protocol Parameters

Radiography device	Exposure index /Relative X-ray exposure*	Peak kilovoltage (kvp)	X-ray tube current (mA)	Exposure expressed (μAs)	Distance source-detector (mm)	Duration of X-ray exposure (ns)
Fujifilm[1]						
Median	130.00	120.00	200.00	2500.00	1000.00	12.00
Mean	132.84	119.99	200.66	2655.01	1000.00	13.29
SD	36.00	0.09	4.39	1160.39	0.00	5.86
Min	37.00	119.00	192.00	600.00	1000.00	3.00
Max	330.00	120.00	212.00	11700.00	1000.00	59.00
Fujifilm[2]						
Median	114.00	119.00	192.00	2500.00	1000.00	13.00
Mean	120.52	119.99	192.60	2680.74	1000.00	14.00
SD	41.17	1.11	2.09	1252.42	0.00	6.37
Min	69.00	104.00	188.00	900.00	1000.00	5.00
Max	547.00	124.00	210.00	12900.00	1000.00	63.00
General Electric[2]						
Median	64.50	120.00	200.00	1360.00	1800.00	6.00
Mean	70.88	119.90	202.08	1485.94	1780.10	6.89
SD	32.88	1.79	20.36	688.48	116.27	3.44
Min	26.00	108.00	200.00	540.00	1000.00	2.00
Max	309.00	130.00	400.00	6360.00	1930.00	31.00
Carestream[2]						
Median	148.57	125.00	125.00	1650.00	1795.50	13.50
Mean	146.81	124.89	125.80	1734.09	1796.89	13.84
SD	36.00	0.75	5.28	674.78	5.94	5.19
Min	7.65	120.00	125.00	100v	1792.00	1.00
Max	244.57	125.00	169.00	3500.00	1820.00	25.00

*Exposure index for Fujifilm and Carestream equipment; Relative X-ray exposure for General Electric equipment.
[1]X-ray machine from Institution-1; [2]X-ray machine from Institution-2.

related factors and the influence of the image acquisition protocol, the AUC of Model-F1F2 on subset T1 was compared with the AUC of Model-F1F3 on the same subset. Then, to analyze the influence of device related factors on the models' internal validation, the AUCs of Model-F1F2 and Model-F1F3 on subset T1 were compared with the AUC of Model-F1'F3'S2 on the same subset (Fig. 2).

The second experiment aim was to estimate the influence of institutional and device related factors on DL models' generalization capacity. With this purpose, the performance of Model-F1F2, was compared between the four test subsets (T1, T2, T3 and T4) (Fig. 2). This model was chosen for this experiment as it was the only model trained with images from only one institution and one X-ray device model.

Finally, attempting to understand the models' perfor-

mance, we implemented the Grad-CAM algorithm, which "uses the gradients of any target concept, flowing into the final convolutional layer to produce a coarse localization heatmap highlighting important regions in the image for predicting the concept" (Selvaraju et al. (2017)). Through Grad-CAM heatmaps we sought to distinguish whether the algorithm was focusing and predicting based on clinical features that defined the target classes (COVID-19 and Control), such as lung opacities, or based on non-clinical features related to the radiography device, such as textural characteristics. In other words, we examined if the models' predictions were based on spurious relationships or on causal relationships, the latter is essential to achieve generalization. We analyzed how institutional and device related factors affect the quality of Grad-CAM heatmaps (Fig. 2).

3.9. Hierarchical Clustering

Hierarchical clustering algorithms group data into clusters based on data similarities through an unsupervised approach (MJ and Wilson (2002)). Therefore, we considered this kind of analysis a valuable method to demonstrate if the institution where images came from, or the X-ray device model used to acquire the images, could be considered more relevant classes (confounding factor) than the intended target class (COVID-19 or Control). Thus, the institution and the radiography equipment could behave as hidden classes, which in the eyes of the algorithm could have more discriminative power than the intended target class. Consequently, the institution and the X-ray machine, including the device's image processing and its type of response function, were considered as potential hidden classes.

For this purpose, feature maps were extracted from the DL models' last convolutional layer. Subsequently, a Global Max Pooling was applied to the feature maps to obtain the feature values, which were later standardized. Images were then clustered by an agglomerative hierarchical clustering algorithm based on feature values similarities. Finally, we examined if the clusters of images found by the algorithm corresponded with the target class or with any of the proposed hidden classes.

Clusters were visually recognizable through heatmaps in which labels identifying the target class and the potential hidden classes were plotted. Hierarchical clustering was run four times. Firstly, the clustering algorithm was run with features extracted by Model-F1F2 from subsets T1, T2 and T3. This resulted in the grouping of images into different clusters. Subsets T1, T2, and T3 contained 94 images each including the same number of COVID-19 and Control images.

The second time the algorithm was run, images from the same subsets (T1, T2 and T3) were cropped preserving only the central areas of the image, which were completely free of metallic tokens (Fig. 3). Features extracted from these cropped images by Model-F1F2 were used to run a new clustering. The purpose of this process was to ensure that metallic tokens indicating laterality did not have an influence on the clusterization.

Thirdly, the clustering was run with features extracted by Model-F1F2 from 44 randomly sampled images from T1, T2 and T3 and from all the images in T4 (44 images). This clustering analysis allowed investigators to evaluate the clusterization of features extracted from 44 images of each radiography device with the same number of COVID-19 and Control images.

The last two clustering algorithms were run with features extracted by Model-F1'F3'S2 before and after being trained from subsets T1, T2 and T3. The purpose of this analysis was to assess whether fine-tuning could reduce differences among radiography devices graphically showing the differences in the feature values introduced by the fine-tuning process.

3.10. Statistical Analysis

The statistically significant difference analysis among the exposure indexes comprised both the verification of the normality assumption with the Lillifors correct Kolmogorov-Smirnov (KS) test (Dallal and Wilkinson (1986); Lilliefors (1967); Dodge (2008)) and the comparison of the metric distributions with the U-Mann-Whitney-Wilcoxon (MWU) (Mann and Whitney (1947); Wilcoxon (1945)) and KS for two independent samples (Dodge (2008)) tests. These statistical tests were conducted using the Python libraries statsmodels (Seabold and Perktold (2010)) and SciPy (Virtanen et al. (2020)).

The Cross-Validated AUCs 95% confidence intervals (CIs) were computed with the R package cvAUC (LeDell et al. (2014)). The AUC differences with their 95% CIs were calculated using the bootstrap method (Efron and Tibshirani (1986)). Any difference where the CI excluded the 0 value was considered to be statistically significant

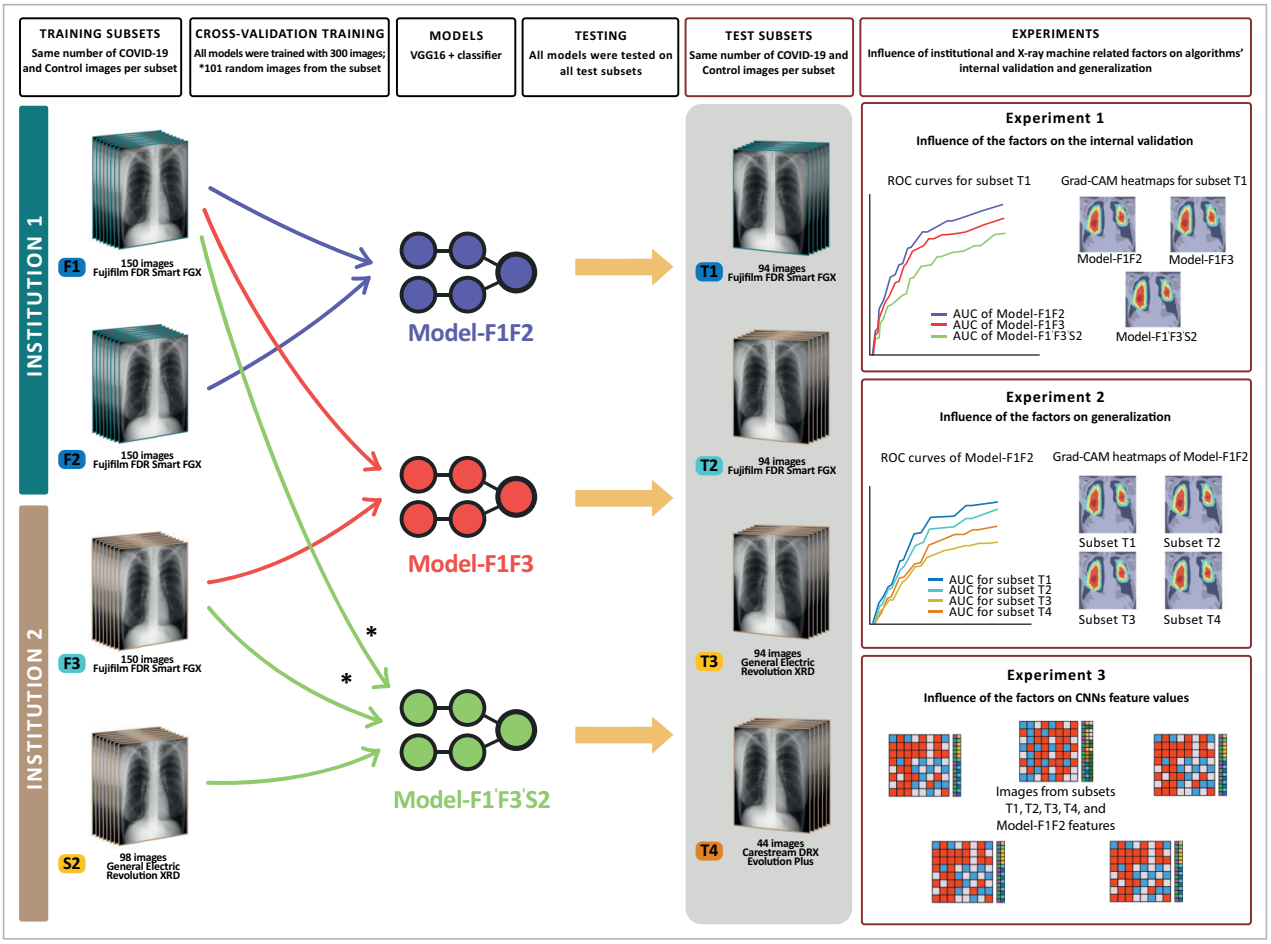


Figure 2: Experimental design. Three CNNs were trained with 300 images. The first model (Model-F1F2) was trained with images from only one model of X-ray device from one institution. The second model (Model-F1F3) was trained with images from only one model of X-ray device from two different institutions. The third model (Model-F1'F3'S2) was trained with images from two models of X-ray device from two different institutions. Subsequently, the algorithms' performance was evaluated on images obtained from different models of X-ray machines and institutions to assess the influence of institutional and X-ray machine related factors on both DL algorithms' internal validation and generalization performance.

with a $p - value < 0.05$. In addition, the DeLong test for paired samples (DeLong et al. (1988)) was also conducted as a second comparing method using the R package pROC (Robin et al. (2011)). Considering that the DeLong test is sometimes exceptionally conservative (Vickers et al. (2011)), bootstrapping was chosen as the reference technique when discrepancies occurred between the results of the bootstrapping and the DeLong test. A two-tailed $p - value$ of 0.05 was considered to indicate statistical significance in all the tests.

3.11. Programming Resources

For this investigation, Python 3.6.8 (Van Rossum and Drake Jr (1995)) and R 4.1.0 (R Core Team (2020)) were utilized. Python libraries used included pydicom 2.1.1 (Mason (2011)), TensorFlow 2.0.0 (Abadi et al. (2015)), Keras 2.2.4 (Chollet et al. (2015)), Seaborn 0.11.1 (Waskom et al. (2017)), statsmodels 0.12.2 (Seabold and

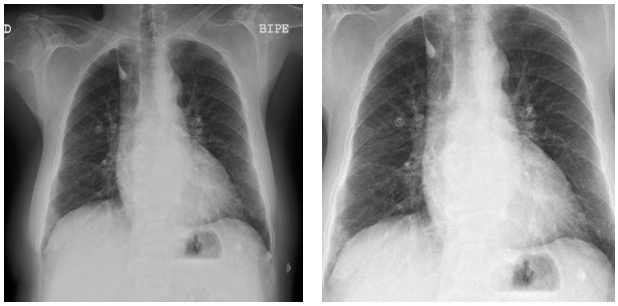


Figure 3: Example of how images were cropped to remove metallic tokens. The image on the left is the original image, while the image on the right is the cropped version of the same image.

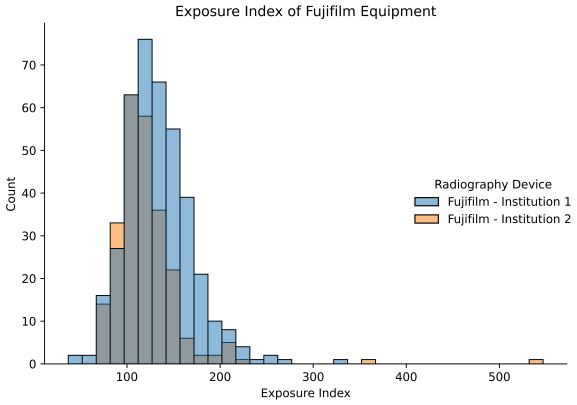


Figure 4: Exposure index of Fujifilm equipment from both institutions.

Perktold (2010)), and SciPy 1.5.4 (Virtanen et al. (2020)). R packages used included cvAUC 1.1.0 (LeDell et al. (2014)) and pROC 3.6 (Robin et al. (2011)).

4. Results

4.1. Patients

This research included 874 patients, 45.08% (394) from Institution-1 and the remaining from Institution-2. The study sample comprised of 42.56% (372) females and 57.44% (502) males. The median age was 62 years, while the average age was 60.21 ± 17.14 years (5-96). Population descriptive statistics for all the subsets of the dataset are summarized in Table 3.

4.2. Image Acquisition Protocols

The exposure indexes of the Fujifilm devices from both institutions were compared to determine whether a statistically significant difference existed between the acquisition protocols of both institutions. The distribution of the exposure indexes did not follow a normal distribution ($p < 0.001$), and a statistically significant difference was detected among the exposure indexes of the two Fujifilm devices (MWU $p = 0.001$, KS $p = 0.002$) (Fig. 4) (Table 4). Therefore, the acquisition protocol of the Fujifilm from Institution-1 and the acquisition protocol of the Fujifilm from Institution-2 were different.

4.3. Models' Evaluation

4.3.1. Experiment 1

There were no significant differences between the internal validation performances of Model-F1F2 and Model-F1F3 (Fig. 5) (Table 5 and Table 6). Thus, the addition of images to the training sample which were acquired in different institutions with different image protocols, did not have a significant impact on the algorithm's internal validation.

Grad-CAM heatmaps based on the predictions of Model-F1F2 and Model-F1F3 for images from subset T1 showed similar activation patterns among each other. The heatmaps for COVID-19 images depicted activations exclusively inside the lungs, while Control images lacked activations within any region of the image (Fig. 6). Seemingly, both Model-F1F2 and Model-F1F3 were able to learn the radiological findings of COVID-19. Therefore, both models made predictions based on causal relationships instead of spurious relationships.

By contrast, the internal validation performances of Model-F1F2 and Model-F1F3 on subset T1 were, respectively, 8% ($p < 0.01$) and 5.2% ($p < 0.05$) higher than the internal validation performance of Model-F1'F3'S2 on the same subset (Fig. 5) (Table 6). As it was previously described, Model-F1F2 and Model-F1F3 were both trained exclusively with 300 images from Fujifilm devices. While Model-F1'F3'S2 was also trained with 300 images, rather than only including images from the Fujifilm devices, Model-F1'F3'S2 incorporated images from

Table 3: Descriptive Statistics of Population Age, and Gender Distribution

	Age					Gender		
Subsets of the Dataset	Median	Mean	SD	Min	Max	Female % (n)	Male % (n)	N
Fujifilm[1]	59.00	58.01	16.92	16	93	44.16 (174)	55.84 (220)	394
Subset F1	59.50	58.53	17.32	16	93	40.67 (61)	59.33 (89)	150
Subset F2	56.50	56.34	17.10	17	89	54.00 (81)	46.00 (69)	150
Subset T1	61.00	59.83	15.87	24	92	34.04 (32)	65.96 (62)	94
Fujifilm[2]	66.00	63.42	16.50	5	96	38.52 (94)	61.48 (150)	244
Subset F3	61.00	61.89	16.29	20	92	38.00 (57)	62.00 (93)	150
Subset T2	68.00	65.85	16.63	5	96	39.36 (37)	60.64 (57)	94
General Electric[2]	63.00	60.09	17.78	18	91	43.75 (84)	56.25 (108)	192
Subset S2	63.00	59.79	17.59	18	87	42.86 (42)	57.14 (56)	98
Subset T3	63.00	60.41	18.06	19	91	44.68 (42)	55.31 (52)	94
Carestream[2]	65.00	62.57	17.33	20	86	45.45 (20)	54.55 (24)	44
Subset T4	65.00	62.57	17.33	20	86	45.45 (20)	54.55 (24)	44
TOTAL	62.00	60.21	17.14	5	96	42.56 (372)	57.44 (502)	874

Subsets F1, F2, F3 and S2 were used to train the models, while subsets T1, T2, T3, and T4 were used to test the models.
[1]X-ray machine from Institution-1; [2]X-ray machine from Institution-2.

Table 4: P-values for the Statistical Analysis of both Fujifilm EI*

Radiography device	KSL	MWU	KS
Fujifilm[1]	< 0.001	< 0.001	< 0.002
Fujifilm[2]	< 0.001		

EI = Exposure Index; KSL = Kolmogorov-Smirnov-Lilliefors; MWU = U-Mann-Whitney-Wilcoxon; KS = Kolmogorov-Smirnov for two samples. A two-tailed $p - value < 0.5$ was considered statistically significant.
*Descriptive statistics of both Fujifilm EI are in Table 2.
[1]X-ray machine from Institution-1; [2]X-ray machine from Institution-2.

the General Electric device in the training as well.

In addition, Grad-CAM heatmaps obtained from the predictions of Model-F1'F3'S2 showed more activation areas on both COVID-19 and Control images than those observed for Model-F1F2 and Model-F1F3, many of which were located outside the lungs and did not have clinical or radiological meaning (Fig. 6). The addition of images acquired by multiple models of X-ray device to the training sample decreased the algorithm's internal validation performance and led the algorithm to learn spurious relationships (confounding factors) instead of causal relationships.

4.3.2. Experiment 2

Experiment 2 results showed that Model-F1F2, which was trained with images from only one institution and acquired by only one X-ray device model (Fujifilm - Institution 1), managed to generalize across hospitals. This being said, Model-F1F2 suffered a variable decrease in performance when having to perform on images from external institutions or acquired by external X-ray device models (Fig. 7). Particularly, Model-F1F2 generalized to Fujifilm images from Institution-2 (subset T2) with a loss in the AUC of 9.8% ($p = 0.06$), and to Carestream images from Institution-2 (subset T4) with a loss in the AUC of 18.9% ($p < 0.05$) (Table 7). Conversely, Model-F1F2 did not generalize to General Electric images from Institution-2 (subset T3), as it showed a loss in the AUC

Table 5: Models' AUCs with 95%CI

Tests	Model-F1F2	Model-F1F3	Model-F1'F3'S2
Subset T1	0.878 (0.847, 0.909)	0.836 (0.799, 0.873)	0.785 (0.745, 0.826)
Subset T2	0.780 (0.738, 0.822)	0.728 (0.682, 0.773)	0.679 (0.631, 0.727)
Subset T3	0.544 (0.491, 0.596)	0.555 (0.502, 0.607)	0.751 (0.708, 0.795)
Subset T4	0.689 (0.618, 0.760)	0.622 (0.548, 0.700)	0.590 (0.514, 0.664)

95% Confidence Intervals (CI) are reported in parentheses.

Table 6: Bootstrapping Differences between the Models' AUCs with 95%CI

Tests	Model-F1F2 & Model-F1F3	Model-F1F2 & Model-F1'F3'S2	Model-F1F3 & Model-F1'F3'S2
Subset T1	0.029 (−0.008, 0.068)	0.080 (0.024, 0.141)**	0.052 (0.008, 0.101)*
Subset T2	0.040 (−0.021, 0.107)	0.096 (0.010, 0.184)*	0.055 (−0.024, 0.134)
Subset T3	−0.008 (−0.049, 0.032)	−0.234 (−0.361, −0.095)***	−0.226 (−0.352, −0.092)***
Subset T4	0.039 (−0.068, 0.161)	0.092 (−0.052, 0.234)	0.053 (−0.054, 0.160)

95% Confidence Intervals (CI) are reported in parentheses.
$P - values$ for two-tailed DeLong tests: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

of 33.5% ($p < 0.001$) which led the model to perform randomly (Table 5 and Table 7). Ultimately, Model-F1F2 generalized across institutions and across different makes and models of X-ray device with the same type of response function, however, not across X-ray devices with a different type of response function.

Additionally, Grad-CAM heatmaps obtained from Model-F1F2 predictions for test images acquired by the Fujifilm device from Institution-1 (subset T1) and by the Fujifilm device from Institution-2 (subset T2) were similar to each other. This being said, Grad-CAM heatmaps based on predictions for test images from the Carestream device (subset T4) showed a significant decrease in the model's sensitivity to detect the radiological findings of COVID-19. Considering that the Carestream and the Fujifilm devices had the same type of response function, the decrease in the model's sensitivity to detect COVID-19 could be secondary to differences in the image processing applied by the X-ray machine. Finally, Grad-CAM heatmaps obtained from predictions for test images from the General Electric device (subset T3) did not show any activation areas, probably due to the fact that this device had a different type of response function than the Fujifilm and Carestream devices'.

4.4. Hierarchical Clustering

4.4.1. Experiment 3

The hierarchical clustering algorithm grouped images from subsets T1 (Fujifilm - Institution-1), T2 (Fujifilm - Institution-2), and T3 (General Electric - Institution-2) into two well-defined clusters based on the features extracted by Model-F1F2. These two clusters corresponded to the Fujifilm and General Electric device models which acquired the images from the subsets (Fig. A1). The algorithm did not separate the images from the two Fujifilm devices, despite these having a different acquisition protocol and belonging to different institutions. In other words, the clustering algorithm successfully separated images from different makes and models of X-ray device. Additionally, images belonging to the different target classes (COVID-19 and Control) were not separated.

The same result was also observed when clustering was run with test images excluding metallic tokens (Fig. A2). Moreover, the addition of images acquired by a third model of X-ray machine, subset T4 (Carestream - Institution 2), resulted in the grouping of images from the three X-ray device models into separated clusters, mixing images from the two Fujifilm devices (Fig. 8).

Table 7: The Influence of Institutional and X-ray Device Related Factors on the Generalization Performance of Model-F1F2

Tests	Performance difference	Factors that might hinder generalization
Subset T1-Subset T2	0.098 (0.045, 0.154)$^{+}$	Institution
Subset T1-Subset T3	0.335 (0.291, 0.368)***	Institution + DIP + DRF
Subset T1-Subset T4	0.189 (0.141, 0.236) *	Institution + DIP
Subset T2-Subset T3	0.237 (0.192, 0.266)***	DIP + DRF
Subset T2-Subset T4	0.091 (0.042, 0.135)	DIP
Subset T3-Subset T4	−0.146 (−0.171, −0.111)*	DRF

DIP = X-ray device's image processing, DRF = X-ray device's type of response function.
95% Confidence Intervals (CI) are reported in parentheses.
$P - values$ for two-tailed DeLong tests: $^{+}p = 0.06$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

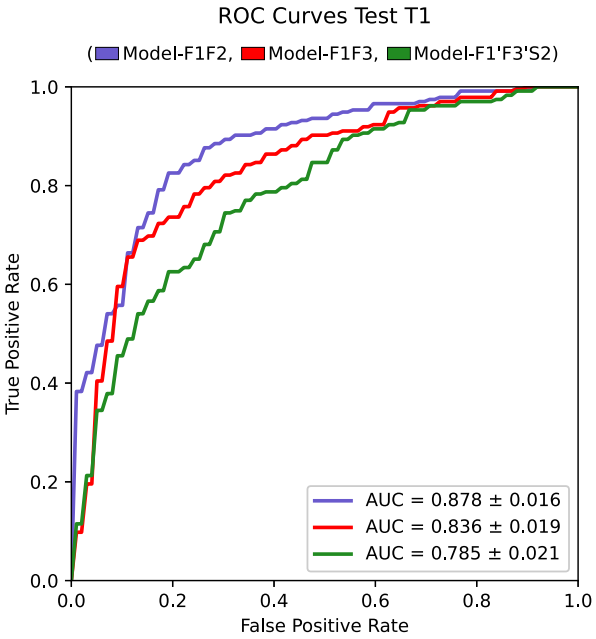


Figure 5: Models' ROC curves for subset T1. Model-F1F2 and Model-F1F3 were trained with images acquired by only one model of X-ray machine, while Model-F1'F3'S2 was trained with images acquired by two different X-ray device models. Model-F1F2 and Model-F1F3 achieved a better internal validation performance than Model-F1'F3'S2. This performance was evaluated using subset T1, which contained 94 images acquired by the Fujifilm device from Institution-1.

The clustering algorithm also managed to group images from subsets T1, T2, T3, and T4 into two clusters. These clusters coincided with the different types of response functions used by the X-ray devices which acquired the subset images (Fig. 8). Therefore, images acquired by X-ray devices with the same type of response function had feature values which were more similar among each other, than feature values from images acquired by X-ray devices with different type of response function.

Finally, the clustering algorithm was also run with features extracted from subsets T1, T2 and T3 by Model-F1'F3'S2 before and after being trained. The algorithm found evident clusters of images that corresponded with images acquired by each model of X-ray machine. This result was observed both when image features were extracted by the pre-trained version of Model-F1'F3'S2 and when features were extracted by the trained version of Model-F1'F3'S2 (Fig. A3). Thus, fine-tuning did not overcome feature values differences found among images acquired by different X-ray device models.

5. Discussion

5.1. Experiment 1 –Factors that Affect DL Algorithms' Internal Validation

Through this investigation, we observed that device related factors significantly affect DL algorithms' internal validation performance. Specifically, algorithms' internal validation performance decreases as more X-ray machine models acquire the training images. According to this, Model-F1F2 and Model-F1F3, which were trained with images from only one make and model of X-ray machine, achieved a better performance in internal validation than

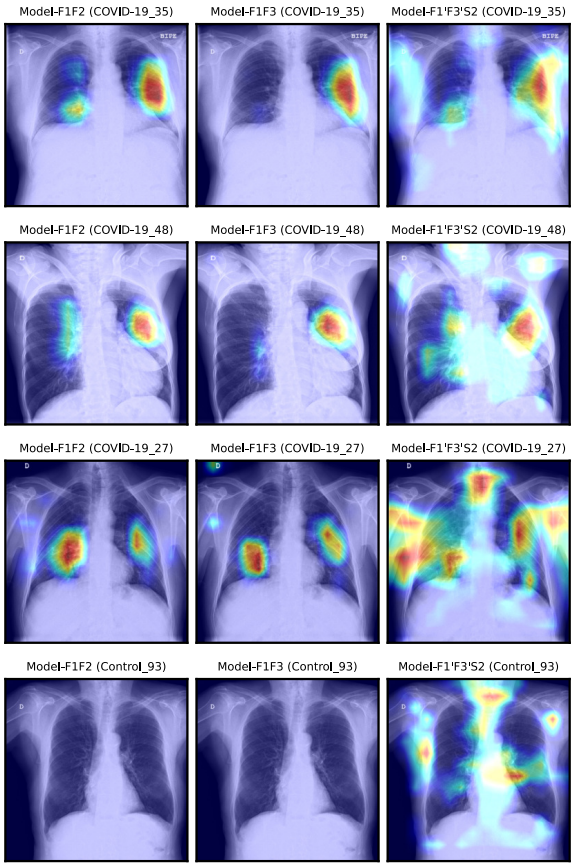


Figure 6: Example of five radiographs with their predicted Grad-CAM heatmaps generated by the three DL models. Model-F1F2 was trained with 300 images acquired by the Fujifilm device from Institution-1, Model-F1F3 with 300 images, half acquired by the Fujifilm from Institution-1 and half by the Fujifilm from Institution-2, and Model-F1'F3'S2 with 202 images from the Fujifilm devices (101 from each Institution) and 98 images from the General Electric device from Institution-2. The addition of images acquired in different institutions with different image protocols, did not have a significant impact on the algorithm's internal validation. On the other hand, the addition of images acquired by multiple X-ray device models to the training sample led the algorithm to learn spurious relationships instead of causal relationships.

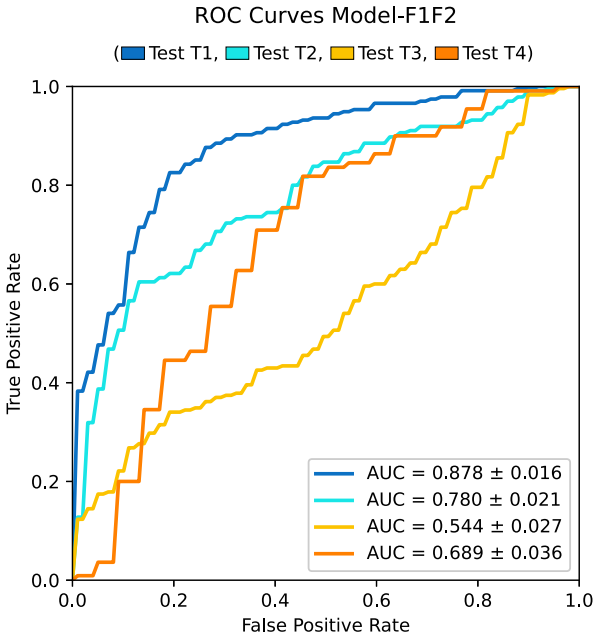


Figure 7: Model-F1F2 was trained with images acquired by the Fujifilm device from Institution-1. This model generalized to images acquired by the Fujifilm device from Institution-2 (subtest T2), and also to images acquired by the Carestream device from Institution-2 (subtest T4). By contrast, Model-F1F2 did not generalize to images acquired by the General Electric device from Institution-2 (subtest T3). Subset T1 contained images from the Fujifilm device from Institution-1.

Model-F1'F3'S2, which was trained with images from two different make and models of X-ray device.

This being said, the decrease in the algorithms' performance is not the only consequence encountered when training with images from multiple device models. Grad-CAM heatmaps showed that as more models of X-ray machine acquired the training images, more textural and non-radiological activation areas appeared in the heatmaps. Particularly, Model-F1F2 and Model-F1F3 showed similar activation patterns in the Grad-CAM heatmaps, which included activation areas inside the lungs exclusively for COVID-19 patients, and absence of activation areas for Control patients. By contrast, Model-F1'F3'S2 showed several activation areas without radiological meaning, including activation areas outside the lungs in COVID-19 patients, and activation areas inside the lungs in Control patients. In other words, Model-F1F2 and Model-F1F3 seemed to predict based on the detection of the radiological findings of COVID-19 rather than based on other

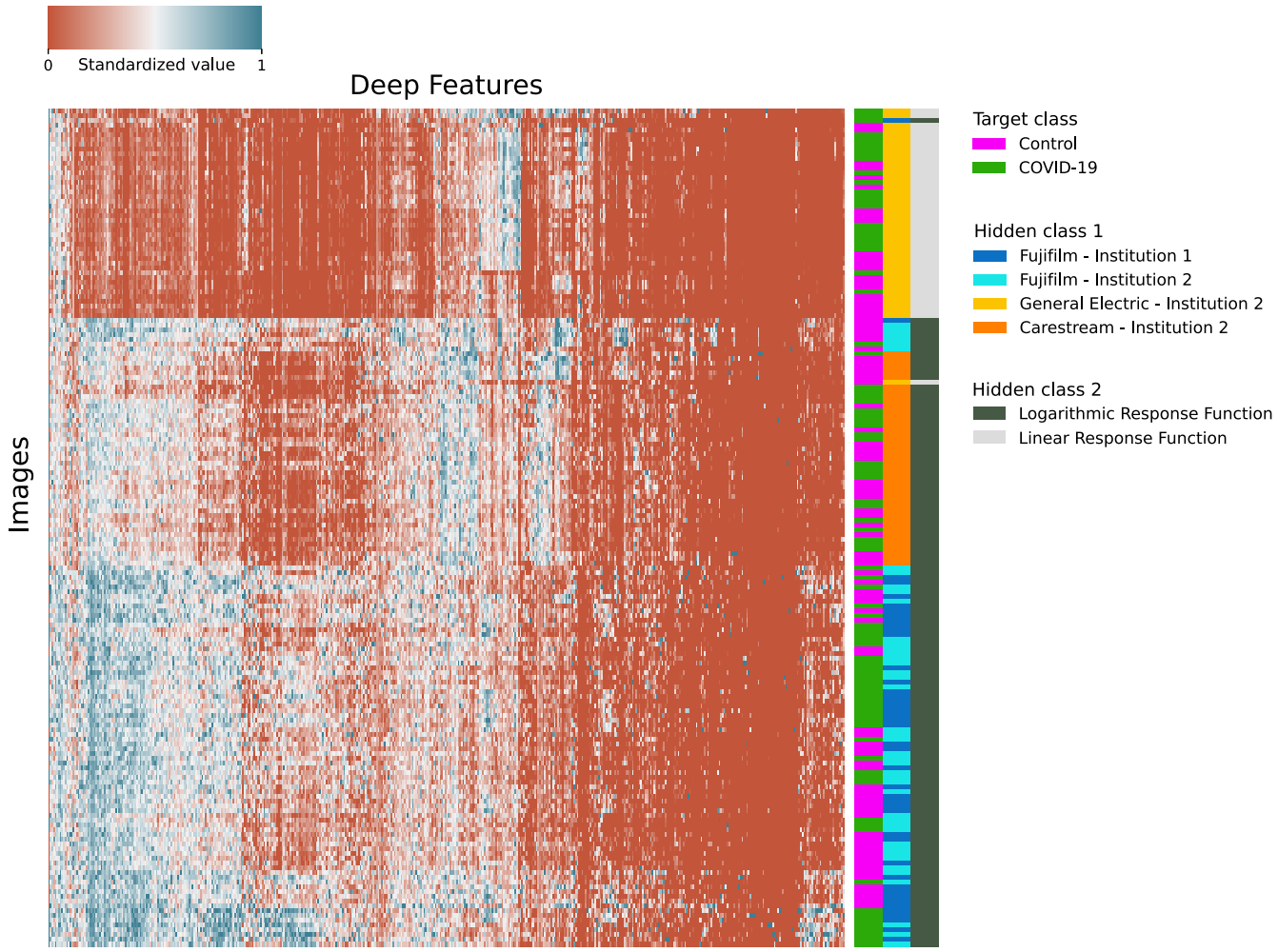


Figure 8: Clusterization of images from subsets T1, T2, T3, and T4 based on Model-F1F2 features. Images clusterization from subsets T1, T2, T3, and T4 using a hierarchical clustering algorithm based on Model-F1F2 features resulted in three clusters. These three clusters corresponded with images from each of the three models of X-ray machine that acquired the images. The radiographs from the two Fujifilm machines were grouped together, despite being from different institutions. Additionally, images were also clustered into two groups based on the type of response function that the X-ray machines which acquired the images had. Finally, images belonging to the different target classes (COVID-19 and Control) were not separated.

image findings without radiological meaning, as Model-F1'F3'S2 did. This issue is of vital importance when discussing the generalization power of DL algorithms. To build robust DL models that can generalize, a key requirement is to be sure that the algorithm predicts based on causal relationships (Luo et al. (2020)), such as pathological radiographic findings, rather than based on spurious relationships, unrelated to the disease.

Unlike device related factors, institutional related factors and the image acquisition protocols do not signif-

icantly affect algorithms' internal validation. Model-F1F2 and Model-F1F3 did not show differences in performance, despite Model-F1F2 being trained exclusively with images from Fujifilm Institution-1 and Model-F1F3 being trained with images from both Fujifilm devices (Institution-1 and Institution-2). Grad-CAM heatmaps were very similar for both DL models.

To our knowledge, this paper is the only one of its kind which studies the effects of training a DL algorithm for medical image classification with CXRs from multiple institutions and acquired by different device models on the algorithm's internal validation performance.

In light of the results accomplished in Experiment 1, we propose an alternative training technique to develop DL algorithms for medical image classification. This technique consists in training different algorithms using, for each one, images acquired by only one X-ray device model. It is important to highlight that within this technique, it does not matter if images are acquired by different devices or in different institutions, as long as all devices are the same model of X-ray machine. Through this technique, it is possible to achieve high performance algorithms that are able to learn causal relationships with a smaller training sample.

5.2. Experiment 2 - Factors that Affect DL Algorithms' Generalization

This research attempts to quantified the influence that institutional and device related factors might have on DL algorithms' generalization for radiography classification. Our results found that DL algorithms can generalize across institutions and X-ray devices with the same type of response function. However, this being said, algorithms may suffer a variable decrease in their performance when deployed on external datasets. On the other hand, generalization across X-ray devices with a different type of response function was not observed.

When it comes to institutional related factors, Model-F1F2 was able to generalize across institutions with a decrease in its performance of 9.8%. This measure was computed as the difference between the model's performance on subset T1 and subset T2. Both subsets contained 94 images acquired by the same make and model of X-ray device (a Fujifilm FDR Smart FGX), however, while images from subset T1 were acquired in Institution-1, images from T2 were acquired in Institution-2, which used a different image acquisition protocol. Therefore, the 9.8% reduction in the generalization performance was attributed to institutional related factors. Experiment 2 results suggest that institutional related factors do not impede the generalization of DL algorithms, however, they can decrease the algorithm's performance.

The acquisition protocol probably had a non-significant influence on the model's generalization performance, as clustering algorithms in Experiment 3 did not separate images from both Fujifilm devices into different clusters, despite having different acquisition protocols.

Model-F1F2 also generalized across X-ray devices with the same type of response function. Specifically, the performance of Model-F1F2 on subset T2 (images from Fujifilm - Institution-2) was 9.1% higher than the performance of Model-F1F2 on subset T4 (images from Carestream - Institution-2). Images from subsets T2 and T4 were all acquired in Institution-2 by two devices which both had a logarithm response function. Thus, the main difference between the images from subsets T2 and T4 was the image processing applied by the X-ray device. Therefore, we assumed that differences in the image processing applied by the X-ray machine reduced the algorithm's generalization performance by 9.1%. In other words, Model-F1F2 generalized across X-ray devices with the same type of response function with a loss in its performance of 9.1%. By contrast, Model-F1F2 performed randomly on subset T3 (images from General Electric – Institution-2), which contained images acquired in Institution-2 by a device which had a linear response function. This result indicates that the algorithm was not able to generalize across devices with a different type of response function.

The previously mentioned results were also supported by Grad-CAM heatmaps. Heatmaps based on the predictions of Model-F1F2 for subset T1 images showed activation areas in locations where radiological findings of COVID-19 could be found. Although heatmaps based on predictions for subset T2 were similar, the model slightly reduced its sensitivity to detect lung opacities in the images from this subset. A higher reduction in the model's sensitivity was observed for heatmaps based on predictions for subset T4, however, these heatmaps still showed activation areas on several lung opacities. By contrast, heatmaps based on predictions for subset T3 images did not show any activation areas, as the algorithm did not

generalize to this subset.

The generalization of DL algorithms for radiography classification to external datasets has been argued by a handful authors. Pooch et al. (Pooch et al. (2020)) conclude that state-of-the-art DL algorithms do not generalize to external data which have differences with the training data. Similar to this, Zech et al. (Zech et al. (2018)) and Sathitratanacheewin et al (Sathitratanacheewin et al. (2020)) defend that CNNs do not generalize to external sites. Additionally, Zech et al. (Zech et al. (2018)) and Maguolo et al. (Maguolo and Nanni (2021)) warn that neural networks can often distinguish the dataset or the hospital where the images come from. For Maguolo et al. (Maguolo and Nanni (2021)), this issue is very important since most papers obtain images of each class to predict from different datasets. Trying to understand how CNNs distinguish the source of the dataset, Cohen et al. (Cohen et al. (2020)) propose discrepancies in image labeling criteria among medical centers to be the potentially cause. On the other hand, for Rajpurkar et al. (Rajpurkar et al. (2020)) and Pan et al. (Pan et al. (2019)) DL algorithms for CXR classification can generalize to datasets from external institutions with a decrease in their performance.

Our research can shed light on the controversy surrounding DL algorithms' generalization, as this work separately analyzed the influence of multiple factors on this issue. We observed that the X-ray device's response function is probably the most important factor for DL algorithms' generalization, as it can impede it. The second most relevant factor might be the device's image processing. Although this factor does not impede the algorithm to generalize, it may significantly decrease its performance. Finally, institutional related factors and the image acquisition protocol can also reduce the algorithm's performance, however, probably less than the device's image processing. Taking all this into account, we propose a hierarchy of factors that might affect the generalization of DL algorithms for medical image classification. This hierarchy is shown in Fig. 9.

In summary, although institutional and device related factors may reduce algorithm's generalization, DL algorithms can generalize across institutions and X-ray devices with the same type of response function. By contrast, these algorithms might not generalize across devices with a different type of response function.

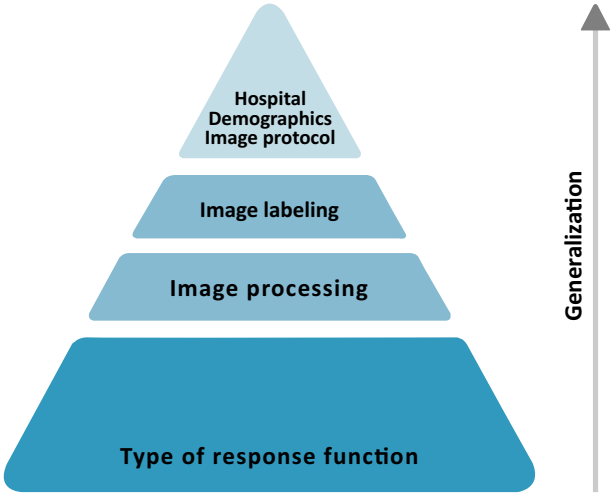


Figure 9: Diagram shows the hierarchy of factors affecting the generalization of Deep Learning algorithms for medical image classification. The type of response function of the radiography device is the most relevant factor since it is the only one that impedes generalization. The second most important factor is the image processing applied by the X-ray machine. The image processing hinders generalization, however, it does not impede it. Finally, institutional related factors, including the image labeling, the institution, and population demographics, have a surmountable influence on algorithm's generalization.

5.3. Experiment 3 – Factors that Affect CNN-Feature Values

The unsupervised clustering algorithm demonstrated that feature values extracted by CNNs from images might be highly dependent on the model of X-ray device that acquires the images. This happens since each model of X-ray device applies a characteristic image processing and has a specific response function. Therefore, radiological images have different textures depending on the model of X-ray device that acquires the images. This leads to disparities in CNN-feature values among images from different X-ray devices that could hinder generalization.

The clustering algorithm in Experiment 3 separated images from each model of X-ray device into different clusters. Images acquired by devices with a logarithmic response function and images acquired by devices with a linear response function were also separated from each other into two different clusters. Additionally, the algorithm did not separate images from the two Fujifilm de-

vices into separate clusters, despite them belonging to different institutions and having different acquisition protocols. Therefore, the model of X-ray device and ultimately, its image processing and its type of response function, had an important influence on CNN-feature values. In contrast, the institution and the acquisition protocol did not significantly affect CNN-feature values.

Similar results were observed when metallic tokens were removed, thus confirming their lack of effect on the clusterization process. Clusterization was also not affected after fine-tunning was applied.

On the other hand, the algorithm did not clearly separate the target classes (COVID-19 and Control) into different clusters. Hence, following our terminology, the equipment, its image processing, and its type of response function were considered high-level hidden classes. We use this term to refer to imaging categories in the dataset which are even more evident for CNNs than the target class. In other words, in the eyes of the VGG16, images from different models of X-ray device were more different from each other than images from COVID-19 and Control classes were from each other.

Experiment 3 results suggest that the model of X-ray device has a strong influence on CNN-feature values, which can be even higher than the influence of target classes when faced with complex problems, such as the classification of COVID-19 CXRs. The feature values susceptibility to the X-ray machine, indicates that features extracted by VGG16 and probably by other ImageNet CNNs, are mainly based on textures instead of shapes. This issue is extremely relevant for DL algorithms' generalization, as features based on shapes are potentially more robust and invariant than features based on textures. Consequently, neural networks able to predict based on shape features rather than on textural features could potentially solve the generalization issue.

Outside of the medical field, Geirhos et al. (Geirhos et al. (2019)) have proposed that ImageNet-trained CNNs are biased towards recognizing textures rather than shapes. These authors also suggest that shape biased networks are inherently more robust than texture biased networks (Geirhos et al. (2019)), something that our research also puts forward in the medical field.

According to us, our investigation is the first which explores the effect of X-ray devices on the feature values extracted by ImageNet-trained CNNs. We conclude that features extracted by ImageNet-trained CNNs from CXRs are mainly based on textures that depend on the X-ray device which acquired the images. Therefore, Transfer Learning from ImageNet-trained CNNs might not be the best solution to achieve robust DL algorithms for medical image classification, when training with small or medium datasets. However, an exception to this rule might be tasks where geometric differences among classes are particularly evident, like the classification of anatomical areas.

Moreover, we introduce the use of hierarchical clustering as a useful technique to graphically visualize the fine-tunning process of CNNs, and to detect high-level hidden classes in datasets. In our opinion, finding these hidden classes could be advantageous, since rather than training only one algorithm using all training images, it is possible to train different algorithms using for each one, images from only one hidden class. This method could achieve high performing algorithms with a smaller training sample. Additionally, this technique could also facilitate the algorithm to learn causal relationships instead of spurious relationships, thus leading to more reliable Grad-CAM heatmaps, as Experiment 1 showed.

6. Conclusion

Factors that affect DL algorithms' internal validation and generalization can be classified into two groups: institutional and device related factors. Institutional related factors are those which do not modify the image pixel values, while device related factors refer to those which may change the distribution of pixel values.

In our research, institutional related factors did not affect the algorithm's internal validation, however, they decreased the algorithm's generalization performance when deployed to an external institution. On the other hand, device related factors significantly reduced both algorithm's internal validation and generalization performances. In particular, the type of response function of the radiography device impeded the algorithm to generalize, while other device related factors hindered, but did not impede, the algorithm's generalization. Thus, the model of X-ray machine which acquires the training and test images is a potentially crucial factor in the performance of DL algorithms.

Furthermore, clustering analysis results indicated that image textures are different for each model of X-machine.

These results also suggest that ImageNet-trained CNNs are probably biased towards identifying textures rather than shapes. Therefore, feature values extracted by ImageNet-trained CNNs depend on the X-ray machine which acquires the image. We found this influence of the X-ray machine on CNN-feature values to be one of the primary causes behind DL algorithms' generalization deficiency for radiography classification.

As a result of this research, we propose two approaches to manage the lack of DL algorithms' generalization. The first approach we propose is to focus on training a different algorithm for each high-level hidden class. This strategy is based on the idea that a good level of generalization is achievable inside the same hidden class. The second approach tries to find a solution through the development of neural networks able to predict based on shape features rather than on textural features. Resulting in the algorithm's to focus on shapes instead of textures.

Acknowledgments

The authors would like to thank the institutions that participated in this study, the patients who made this research possible by providing their images, and the radiologists who generously labeled the images. Additionally, the authors would like to show special gratitude to the Spanish National Research Council (CSIC) which supported this work (grant number 202050E107, 2020).

Conflict of interests

Authors declare that they have no conflict of interest.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems.

Al aseri, Z., 2009. Accuracy of chest radiograph interpretation by emergency physicians. Emergency radiology 16, 111–114.

Borghesi, A., Roberto, M., 2020. Covid-19 outbreak in italy: experimental chest x-ray scoring system for quantifying and monitoring disease progression. La radiologia medica 125, 509–513.

Butler, M.L., Rainford, L., Last, J., Brennan, P.C., 2010. Are exposure index values consistent in clinical practice? a multi-manufacturer investigation. Radiation Protection Dosimetry 139, 371–374.

Chollet, F., et al., 2015. Keras. URL: https://github.com/fchollet/keras.

Cohen, J.P., Hashir, M., Brooks, R., Bertrand, H., 2020. On the limits of cross-domain generalization in automated x-ray prediction, in: Medical Imaging with Deep Learning. URL: https://openreview.net/forum?id=VB2MOuOKyq.

Dallal, G.E., Wilkinson, L., 1986. An analytic approximation to the distribution of lilliefors's test statistic for normality. The American Statistician 40, 294–296.

Das, A.K., Kalam, S., Kumar, C., Sinha, D., 2021. Tlcovan automated covid-19 screening model using transfer learning from chest x-ray images. Chaos Solitons Fractals 144, 110713.

DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics 44, 837–845.

Dodge, Y., 2008. The Concise Encyclopedia of Statistics. Springer New York, New York, NY. chapter Kolmogorov-Smirnov Test. pp. 283–287.

Efron, B., Tibshirani, R., 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science 1, 54–75.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W., 2019. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.,

in: International Conference on Learning Representations. URL: https://openreview.net/forum?id=Bygh9j09KX.

Hwang, E.J., Nam, J.G., Lim, W.H., Park, S.J., Jeong, Y.S., Kang, J.H., Hong, E.K., Kim, T.M., Goo, J.M., Park, S., Kim, K.H., Park, C.M., 2019. Deep learning for chest radiograph diagnosis in the emergency department. Radiology 293, 573–580.

KCARE, 2005. Quantitative evaluation of digital detectors for general radiography. Technical Report 05078. KCARE.

Lanca, L., Silva, A., 2013. Digital Imaging Systems for Plain Radiography. Springer-Verlag New York.

LeDell, E., Petersen, M., van der Laan, M., 2014. cvAUC: Cross-Validated Area Under the ROC Curve Confidence Intervals.

Lilliefors, H.W., 1967. On the kolmogorov-smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association 62, 399–402.

Luo, Y., Peng, J., Ma, J., 2020. When causal inference meets deep learning. Nature Machine Intelligence 2, 426–427.

Maguolo, G., Nanni, L., 2021. A critic evaluation of methods for covid-19 automatic detection from x-ray images. Information Fusion 76, 1–7.

Mann, H., Whitney, D., 1947. On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics 18, 50–60.

Mason, D., 2011. Su-e-t-33: pydicom: an open source dicom library. Medical Physics 38, 3493–3493.

MJ, G., Wilson, S., 2002. Introduction to hierarchical clustering. Journal of clinical neurophysiology: official publication of the American Electroencephalographic Society 19, 144–151.

Pan, Ian AU Agarwal, S.P.I., Agarwal, S., Merck, D., 2019. Generalizable inter-institutional classification of abnormal chest radiographs using efficient convolutional neural networks. Journal of Digital Imaging 32, 888–896.

Pooch, E.H.P., Ballester, P., Barro, R.C., 2020. Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification, in: Thoracic Image Analysis, Springer International Publishing. pp. 74–83.

R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Rajpurkar, P., Joshi, A., Pareek, A., Chen, P., Kiani, A., Irvin, J., Ng, A.Y., Lungren, M.P., 2020. Chexpedition: Investigating generalization challenges for translation of chest x-ray algorithms to the clinical setting. arXiv:2002.11379.

Robin, X., TurcK, N., HainarD, A., TibertI, N., Lisacek, F., Sanchez, J.C., Müller, M., 2011. proc: an open-source package for r and s+ to analyze and compare roc curves. BMC Bioinformatics 12, 77.

Sathitratanacheewin, S., Sunanta, P., Pongpirul, K., 2020. Deep learning for automated classification of tuberculosis-related chest x-ray: dataset distribution shift limits diagnostic performance generalizability. Heliyon 6, e04614.

Seabold, S., Perktold, J., 2010. statsmodels: Econometric and statistical modeling with python, in: 9th Python in Science Conference.

Seibert, J.A., Morin, R.L., 2011. The standardized exposure index for digital radiography: an opportunity for optimization of radiation dose to the pediatric population. Pediatric Radiology 41, 573–581.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV).

Shelke, A., Inamdar, M., Shah, V., Tiwari, A., Hussain, A., Chafekar, T., Mehendale, N., 2021. Chest x-ray classification using deep learning for automated covid-19 screening. SN computer science 2, 300.

Sitaula, C., Hossain, M.B., 2020. Attention-based vgg-16 model for covid-19 chest x-ray image classification. Applied Intelligence 17, 1–14.

Srivastava, N., Baxi, P., Ratho, R.K., Saxena, S.K., 2020. Coronavirus Disease 2019 (COVID-19): Epidemiology, Pathogenesis, Diagnosis, and Therapeutics. Springer Singapore, Singapore. chapter Global Trends in Epidemiology of Coronavirus Disease 2019 (COVID-19). pp. 9–21.

Subbaswamy, A., Saria, S., 2018. Counterfactual normalization: Proactively addressing dataset shift and improving reliability using causal mechanisms, in: Silva, R., Globerson, A., Globerson, A. (Eds.), 34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018, Association For Uncertainty in Artificial Intelligence (AUAI). pp. 947–957.

Van Rossum, G., Drake Jr, F.L., 1995. Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam.

Vickers, A.J., Cronin, A.M., Begg, C.B., 2011. One statistical test is sufficient for assessing new predictive markers. BMC Medical Research Methodology 11.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors, 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods 17, 261–272.

Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lukauskas, S., Gemperline, D.C., Augspurger, T., Halchenko, Y., Cole, J.B., Warmenhoven, J., de Ruiter, J., Pye, C., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Bachant, P., Martin, M., Meyer, K., Miles, A., Ram, Y., Yarkoni, T., Williams, M.L., Evans, C., Fitzgerald, C., Brian, Fonnesbeck, C., Lee, A., Qalieh, A., 2017. mwaskom/seaborn: v0.8.1 (september 2017). URL: https://doi.org/10.5281/zenodo.883859, doi:10.5281/zenodo.883859.

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics Bulletin 1, 80–83.

Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K., 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine 15, 1–17.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G.F., Tan, W., 2020. A novel coronavirus from patients with pneumonia in china, 2019. New England Journal of Medicine 382, 727–733.

Appendix A

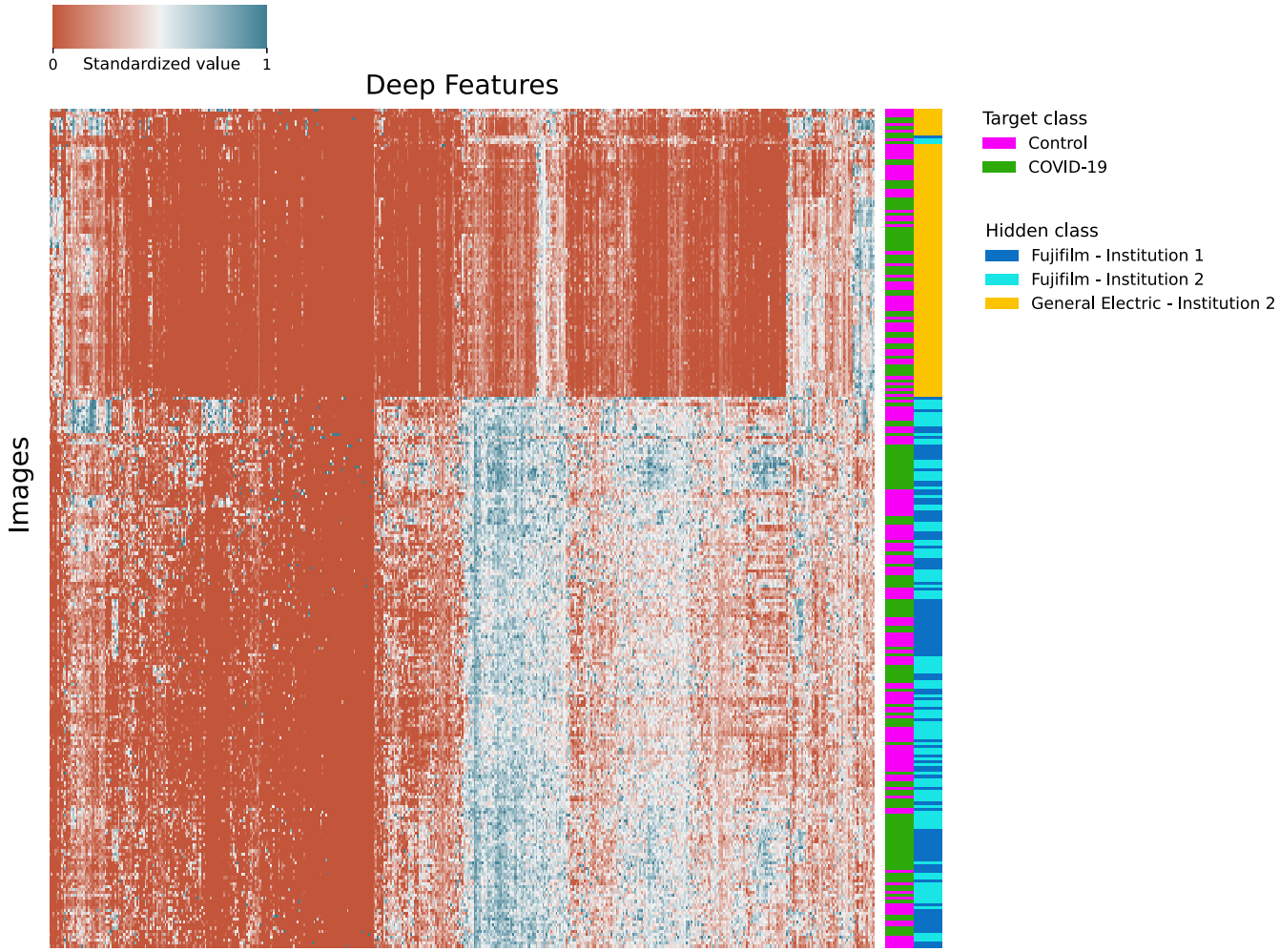


Figure A1: Clusterization of images from subsets T1, T2, and T3 based on Model-F1F2 features. The clusterization of images using hierarchical clustering resulted in two clusters. One cluster contained the images acquired by the Carestream device and the other cluster contained the images acquired by the Fujifilm devices. Therefore, images were separated based on the device which acquired them. Additionally, images acquired by the two Fujifilm devices were not separated, despite being from different institutions and having different acquisition protocols.

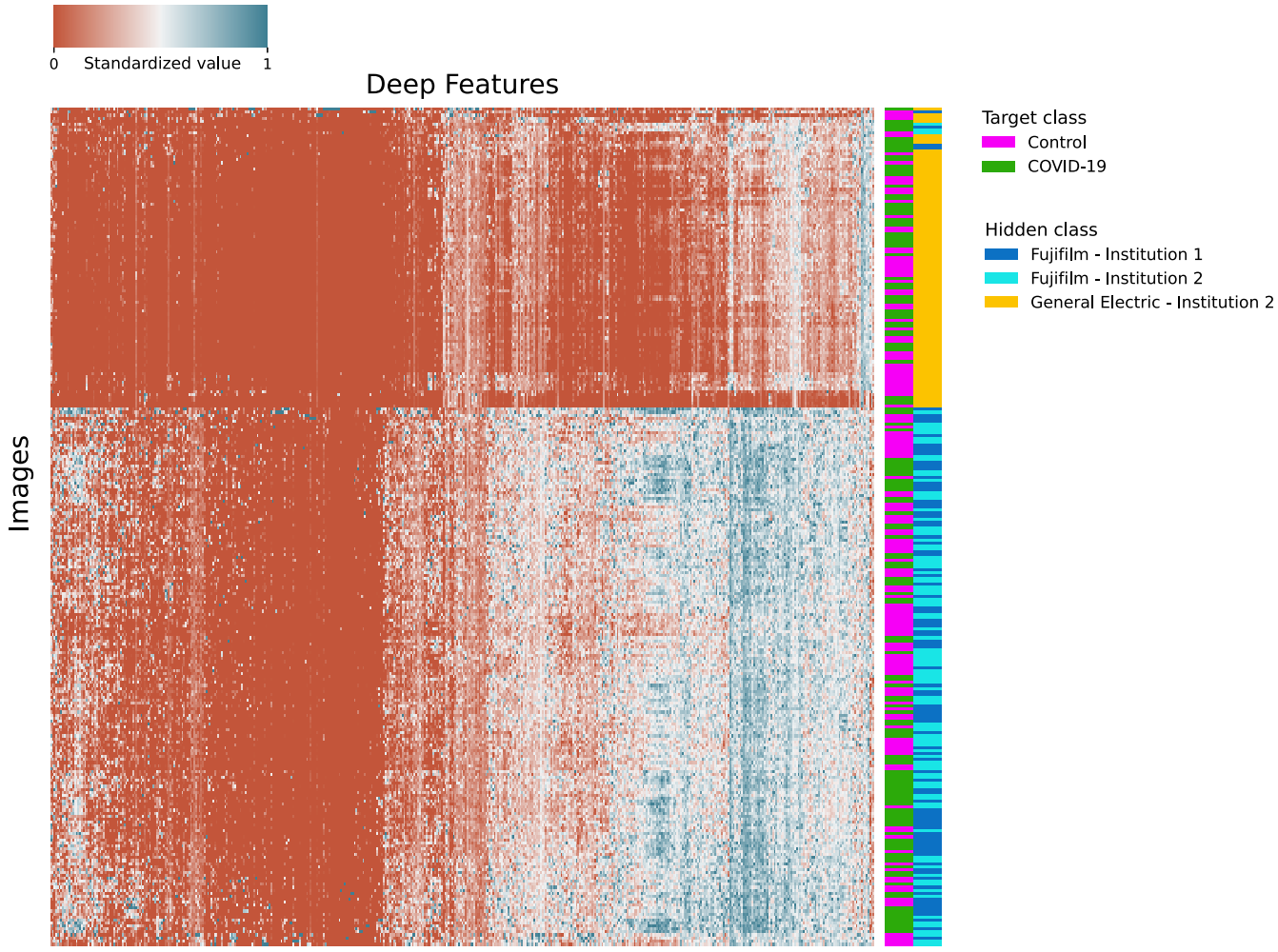


Figure A2: Clusterization of cropped images from subsets T1, T2, and T3 based on Model-F1F2 features. When images were cropped to eliminate metallic tokens, no relevant changes were observed in regards to the results shown in Fig. A1.

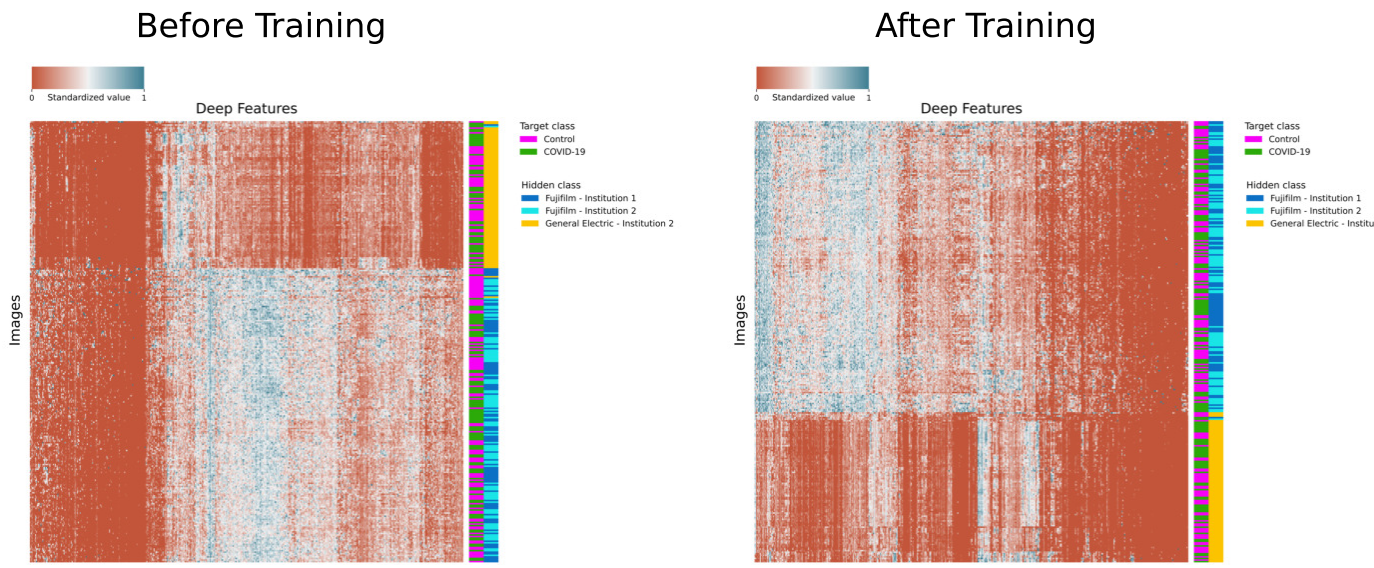
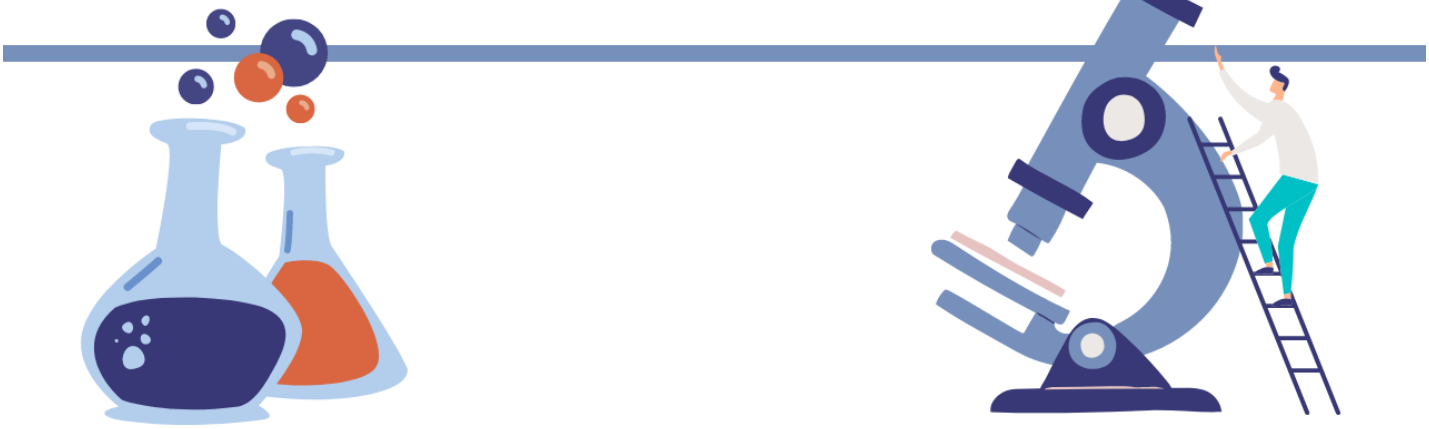


Figure A3: Clusterization of images from subsets T1, T2, and T3 based on features extracted by Model-F1'F3'S2 before and after being trained. Images were clustered based on features extracted by the pre-trained and trained versions of Model-F1'F3'S2. In both cases, images were successfully grouped into different device clusters. Thus, fine-tunning did not reduce differences in feature values between images from different X-ray machines.

9.5.5. Capítulo en el libro *"Curso: Tendencias en Investigación clínica 2021"*: Inteligencia artificial en medicina

CURSO: TENDENCIAS EN INVESTIGACIÓN CLÍNICA 2021



Capítulos Curso Tendencias en Investigación Clínica

AUTORES

- *Mª del Mar García Sáiz*. Jefe del Servicio de Farmacología Clínica. Hospital Universitario Marqués de Valdecilla. Responsable de la UIC de IDIVAL, miembro de la Plataforma SCReN (ISCIII).
- *Lucía Lavín Alconero*. Doctora en Biología. IDIVAL- Hospital Universitario Marqués de Valdecilla.
- *Dr. Ignacio Duran Martínez.* Servicio de Oncología Médica. Hospital Universitario Marques de Valdecilla. IDIVAL. Santander.
- *Paula Iruzubieta.* Facultativo especialista del Servicio Digestivo. Hospital Universitario Marqués de Valdecilla. IDIVAL. Santander. Cantabria.
- *María Teresa Arias.* Facultativo especialista del Servicio Digestivo. Hospital Universitario Marqués de Valdecilla. IDIVAL. Santander. Cantabria.
- *Javier Crespo.* Jefe de Servicio de Servicio Digestivo. Hospital Universitario Marqués de Valdecilla. IDIVAL. Santander. Cantabria.
- *Javier Loricera García*. Adjunto del Servicio de Reumatología. Hospital Universitario Marqués de Valdecilla, IDIVAL. Santander
- *Miguel Ángel González-Gay Mantecón.* Jefe del Servicio de Reumatología. Hospital Universitario Marqués de Valdecilla, IDIVAL. Santander
- *Ricardo Blanco Alonso.* Jefe de sección del Servicio de Reumatología. Hospital Universitario Marqués de Valdecilla, IDIVAL. Santander
- *Pascual Sánchez-Juan.* Unidad de Deterioro Cognitivo. Hospital Universitario Marqués de Valdecilla.
- *Carmen Lage-Martínez.* Unidad de Deterioro Cognitivo. Hospital Universitario Marqués de Valdecilla.
- *Sara López-García.* Unidad de Deterioro Cognitivo. Hospital Universitario Marqués de Valdecilla.
- *Luis Alberto Vázquez Salví*. Doctor en Medicina. Médico del Servicio de Endocrinología, Diabetes y Nutrición – Hospital Universitario Marqués de Valdecilla. Profesor Asociado de Ciencias de la Salud – Facultad de Medicina – Universidad de Cantabria. Santander. España.
- *Laura Ramos Ramos*. Doctora en Medicina. Médica del Servicio de Endocrinología, Diabetes y Nutrición – Hospital Universitario Marqués de Valdecilla. Santander. España.
- *María Piedra León*. Doctora en Medicina. Médica del Servicio de Endocrinología, Diabetes y Nutrición – Hospital Universitario Marqués de Valdecilla. Santander. España.
- *Francisco Arnaiz de las Revillas*. Adjunto del Servicio de Enfermedades Infecciosas. Hospital Universitario Marqués de Valdecilla, IDIVAL. Universidad de Cantabria
- *Claudia González-Rico.* Técnico de apoyo a la investigación / Study Coordinator del Servicio de Enfermedades Infecciosas. Hospital Universitario Marqués de Valdecilla, IDIVAL. Universidad de Cantabria
- *M. Carmen Fariñas.* Jefe del Servicio de Enfermedades Infecciosas. Hospital Universitario Marqués de Valdecilla, IDIVAL. Universidad de Cantabria
- *Pilar Alonso Lecue*. Doctora en Biología Molecular. Técnico Superior de apoyo a la Investigación. IDIVAL, Hospital Universitario Marqués de Valdecilla.
- *José Manuel Cifrián Martínez.* Jefe de Servicio de Neumología. Hospital Universitario Marqués de Valdecilla, IDIVAL.

- *Francisco González Vílchez*. Profesor Asociado de Medicina. Departamento de Medicina y Psiquiatría. Universidad de Cantabria. Facultativo Especialista de Área de Cardiología. Hospital Universitario Marqués de Valdecilla. Santander. Investigador del Grupo de Investigación Cardiovascular. Instituto de Investigación Valdecilla (IDIVAL). Cantabria.
- *José Antonio Vázquez de Prada Tiffé*. Profesor Titular de Medicina. Departamento de Medicina y Psiquiatría. Universidad de Cantabria. Facultativo Especialista de Área de Cardiología. Hospital Universitario Marqués de Valdecilla. Santander. Director del Grupo de Investigación Cardiovascular. Instituto de Investigación Valdecilla (IDIVAL). Cantabria.
- *José M de la Torre Hernández*. Jefe de Sección en Cardiología Intervencionista-Servicio de Cardiología. Hospital Universitario Marqués de Valdecilla, Santander
- *Jose Luis Gutiérrez Baños*. Jefe De Servicio De Urología. Hospital Universitario Valdecilla. Profesor Asociado De Urología. Universidad De Cantabria. Instituto De Investigación Sanitaria Valdecilla (IDIVAL)
- *Pedro José Prada Gómez.* Jefe de Servicio de Oncología Radioterápica. Hospital Universitario Marqués de Valdecilla
- *Felix Campos Juanatey*. F.E.A. Urología. Hospital Universitario Valdecilla. Instituto De Investigación Sanitaria Valdecilla (IDIVAL)
- *Mario Domínguez Esteban*. F.E.A. Urología. Hospital Universitario Valdecilla. Instituto De Investigación Sanitaria Valdecilla (IDIVAL)
- *Javier Vázquez Bourgon.* Profesor Asociado de Psiquiatría, Departamento de Medicina y Psiquiatría, Facultad de Medicina, Universidad de Cantabria. Facultativo Especialista de Área de Psiquiatría, Servicio de Psiquiatría, Hospital Universitario Marqués de Valdecilla. Grupo de Psiquiatría, Instituto de Investigación Sanitaria Valdecilla (IDIVAL) Grupo G26, CIBERSAM
- *Mª Henar Rebollo Rodrigo*. Jefe de Servicio de Medicina Preventiva y Seguridad del Paciente del Hospital Universitario Marqués de Valdecilla.
- *Javier Canelas Fernández*. Residente de Medicina Preventiva y Salud Pública de la Unidad Docente de Medicina Preventiva y Salud Pública de Cantabria.
- *Borja Suberviola Cañas* (Ph.D). Servicio de Medicina Intensiva. Hospital Universitario Marqués de Valdecilla. Santander
- *María Ángeles Ballesteros Sanz* (Ph.D). Servicio de Medicina Intensiva. Hospital Universitario Marqués de Valdecilla. Santander
- *Elena Cuenca Fito* (MD). Servicio de Medicina Intensiva. Hospital Universitario Marqués de Valdecilla. Santander
- *Gina Lladó Jordan*. Doctora en biomedicina y ciencias de la salud. Profesora de metodología de investigación, evidencia científica y epidemiología. Monitora del estudio observacional MIRCAST en IDIVAL
- *Daniel García López.* Facultativo Especialista de Área. Servicio de Anestesiología, Reanimación y Unidad del Dolor. Hospital Universitario Marqués de Valdecilla
- *Lucía Lavín Alconero*. Doctora en Biología. IDIVAL- Hospital Universitario Marqués de Valdecilla.
- *Marcos Gómez Ruiz*. CBC (Hi) FEBS. Adjunto, Unidad de Cirugía Colorrectal, Servicio de Cirugía General. Director de Programas de Cirugía Robótica, Hospital Universitario Marqués de Valdecilla. Director del Grupo de Innovación Quirúrgica, Instituto de Investigación Sanitaria Valdecilla, IDIVAL
- *Marcos López Hoyos.* Jefe de Servicio Inmunología. Profesor Titular Inmunología,

Hospital Universitario Marqués de Valdecilla. Universidad de Cantabria. Director Científico de IDIVAL Santander

- *Teresa Gimenez Poderós*. Servicio de Farmacia. Hospital Universitario Marqués de Valdecilla
- *Marta Valero Domínguez*. Servicio de Farmacia. Hospital Universitario Marqués de Valdecilla
- *Pablo Menéndez Fernández-Miranda***.** Médico residente de radiodiagnóstico en el Servicio de Radiodiagnóstico del Hospital Universitario Marqués de Valdecilla.
- *Enrique Marqués Fraguela.* Facultativo especialista en el área de radiofísica en el Servicio de Radiofísica y Protección Radiológica del Hospital Universitario Marqués de Valdecilla.
- *Pablo Sanz Bellón.* Médico residente de radiodiagnóstico en el Servicio de Radiodiagnóstico del Hospital Universitario Marqués de Valdecilla.
- *Marta Drake Pérez.* Médico facultativa especialista en el área de radiodiagnóstico en el Servicio de Radiodiagnóstico del Hospital Universitario Marqués de Valdecilla.
- *Andrés González Mandly.* Médico facultativo especialista en el área de radiodiagnóstico y jefe de servicio del Servicio de Radiodiagnóstico del Hospital Universitario Marqués de Valdecilla.
- *José Luis Arroyo, MD, PhD.* Director Banco de Sangre y Tejidos de Cantabria – Fundación Marqués de Valdecilla.
- *Oscar M Pello, PhD.* Responsable del procesamiento de progenitores hematopoyéticos y la terapia celular; Banco de Sangre y Tejidos de Cantabria – Fundación Marqués de Valdecilla.

CAPÍTULO 20. INTELIGENCIA ARTIFICIAL EN MEDICINA

*Pablo Menéndez Fernández-Miranda***.** Médico residente de radiodiagnóstico en el Servicio de Radiodiagnóstico del Hospital Universitario Marqués de Valdecilla.

Enrique Marqués Fraguela. Facultativo especialista en el área de radiofísica en el Servicio de Radiofísica y Protección Radiológica del Hospital Universitario Marqués de Valdecilla.

Pablo Sanz Bellón. Médico residente de radiodiagnóstico en el Servicio de Radiodiagnóstico del Hospital Universitario Marqués de Valdecilla.

Marta Drake Pérez. Médico facultativa especialista en el área de radiodiagnóstico en el Servicio de Radiodiagnóstico del Hospital Universitario Marqués de Valdecilla.

Andrés González Mandly. Médico facultativo especialista en el área de radiodiagnóstico y jefe de servicio del Servicio de Radiodiagnóstico del Hospital Universitario Marqués de Valdecilla.

1. INTRODUCCIÓN

Las publicaciones en medicina referentes a sistemas de inteligencia artificial (IA) se han multiplicado de forma exponencial en las últimas dos décadas y, especialmente, en el último lustro. Sin embargo, la IA es un campo tan antiguo o tan nuevo, como los ordenadores personales. Sus orígenes se remontan al verano de 1956, cuando John McCarthy acuñó por primera vez el término IA en una conferencia en Dartmouth, EEUU, tan solo 5 años más tarde de la comercialización de la primera computadora personal. La explicación que subyace al hecho de que esta disciplina no hubiera protagonizado incursiones de relevancia en el campo de la medicina hasta los últimos años, se encuentra en que no fue hasta la primera década del siglo XXI, cuando la industria de los videojuegos desarrolló las unidades de procesamiento gráfico o GPUs, que incrementaron de forma sustancial la capacidad de computo y abrieron la posibilidad a entrenar redes profundas. Fue este hecho junto al nacimiento del *Big Data* – obtención y almacenamiento de grandes cantidades de datos – los puntos clave que situaron a la IA y, en concreto al *Machine Learning* (ML), a la vanguardia de la investigación en medicina.

2. INTELIGENCIA ARTIFICIAL

La IA se define como la capacidad de las máquinas de simular funciones cognitivas humanas. El primer abordaje para lograrlo, dio lugar a lo que se denominó IA simbólica o IA guiada por el conocimiento, que experimentó su mayor auge en la década de los años 1980 con la aparición de los sistemas expertos (Fig.1).

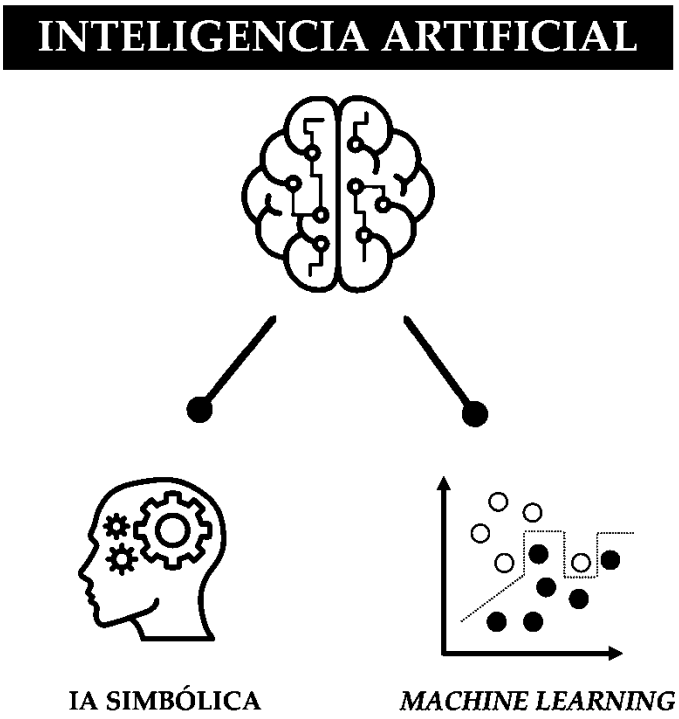


Figura 1: Los algoritmos de Inteligencia Artificial se pueden dividir en dos grandes grupos: los sistemas expertos o IA simbólica y el *Machine Learning* o IA guiada por datos.

Esta aproximación se basaba en la idea de que una inteligencia artificial que alcanzara el nivel de la inteligencia humana, podría lograrse a través de la programación de algoritmos que simularan la toma de decisiones de un humano experto. Un ejemplo conocido de este tipo de IA, fue Mycin, un popular sistema experto orientado al diagnóstico y tratamiento de enfermedades comunes desarrollado por la Escuela de Medicina de Stanford. Mycin fue programado con cerca de 500 reglas que trataban de emular la toma de decisiones que llevaría a cabo un humano experto.

Los sistemas expertos aportaron soluciones eficaces a problemas lógicos, sin embargo, problemas no resolubles mediante reglas explícitas, como el reconocimiento de imágenes o él del propio lenguaje, resultaban aún inabordables.

El segundo abordaje se denominó *Machine Learning* o IA guiada por datos, y surge de la pregunta que 1950 se formuló Alexander Turing acerca de si un ordenador podría aprender por sí mismo y adquirir conocimiento original. El ML supuso un cambio de paradigma. Ahora, los algoritmos no se programarían con reglas explícitas, como en el caso de la IA simbólica, si no que los algoritmos se programarían para que fueran capaces de aprender esas reglas a partir de la exposición a ejemplos resueltos, es decir, se entrenarían con datos. Una vez extraídas las reglas, estas se aplicarían a la resolución de problemas nuevos, lo que se denomina generalizar. Desde los años 1980, el campo del *Machine Learning* se ha postulado como el principal en la IA.

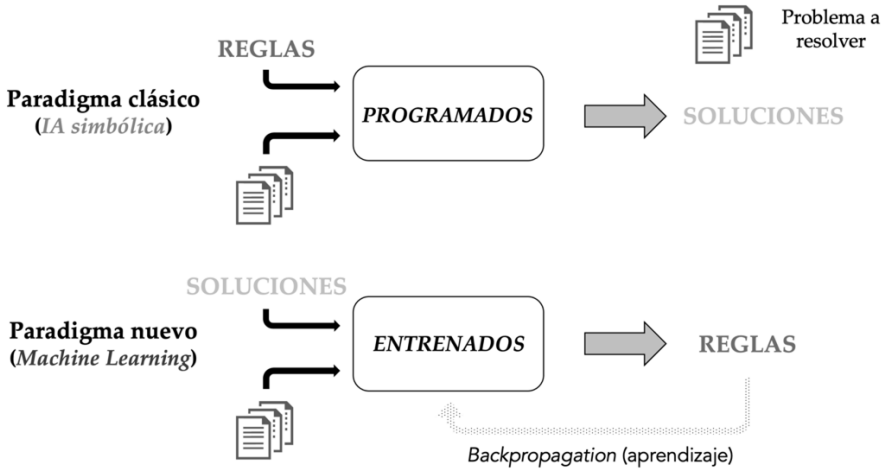


Figura 2: Cambio de paradigma que supuso el *Machine Learning*, ya no se programan los sistemas con reglas explícitas, sino que se entrenan algoritmos que son capaces de extraer esas reglas.

3. MACHINE LEARNING

El *Machine Learning*, traducido como aprendizaje automático, es el campo de la IA que como se indicó anteriormente, se basa en la búsqueda de métodos para dotar a las máquinas de capacidad de aprendizaje.

Pongamos un ejemplo aplicado a la medicina: distinguir entre radiografías de COVID-19 y radiografías normales. En la fase de aprendizaje, al igual que haría un radiólogo, el sistema de IA trataría de aprender la llamada frontera de decisión, es decir, los hallazgos que propiciarían la clasificación de las radiografías en una de las categorías propuestas (COVID-19 o control). Posteriormente, se aplicaría esa frontera de decisión a nuevas imágenes, siendo así capaz de realizar la tarea de diagnóstico. En otras palabras, al igual que los seres humanos, los sistemas de ML aprenden extrayendo reglas a partir de ejemplos conocidos.

En este ejemplo se puede apreciar una de las características más importantes que aporta el ML con respecto a la IA simbólica. Si quisiéramos realizar la misma tarea con un sistema de IA simbólica, un experto debería encontrar y codificar un conjunto de reglas que, bajo su criterio, permitieran distinguir las dos clases de radiografías. Sin embargo, debemos tener en cuenta que, a menudo, en los problemas médicos no existen reglas definidas y que el conocimiento humano no es completo y es, a veces, incorrecto. Además, el conocimiento humano es en ocasiones intuitivo y difícilmente formulable en una serie de reglas lógicas. Frente a la IA simbólica, el ML no requiere de la definición de reglas por un ser humano y por ello, no presenta los problemas descritos.

3.1. Algoritmos de Machine Learning en función del grado de intervención del humano

Los algoritmos de *Machine Learning* se clasifican en función del grado de intervención del humano en el aprendizaje. De esta forma se distinguen los siguientes tipos:

3.1.1. Supervisado

Estos algoritmos se aplican principalmente en tareas de clasificación o segmentación. En este abordaje, se utilizan datos para el entrenamiento del algoritmo que han sido etiquetados o segmentados por un ser humano, de modo que el aprendizaje se encontrará condicionado por dichas elecciones.

En un ejemplo de clasificación de radiografías de tórax pertenecientes a paciente con COVID-19 o sanos, las radiografías utilizadas durante el entrenamiento deberían ir acompañadas por la etiqueta COVID-19 o control. El algoritmo aprenderá a identificar las características distintivas del COVID-19 a partir de la exposición a esos datos etiquetados.

3.1.2. No Supervisado

En este tipo de algoritmos no es necesario disponer de datos de entrenamiento etiquetados. Estos sistemas son capaces de aprender patrones de similitud en los datos que permiten su agrupación en diferentes categorías abstractas (*clustering*), para posteriormente intentar darle un sentido a las agrupaciones.

Continuando con el ejemplo propuesto de clasificación de radiografías COVID-19, un algoritmo de ML no supervisado partiría de las imágenes sin etiquetar y trataría de agruparlas en función de patrones extraídos automáticamente. Posteriormente un humano analizaría las categorías de clasificación propuestas por el algoritmo y trataría de encontrar su sentido.

Una de las grandes ventajas de los sistemas no supervisados es que permiten aprender aspectos nuevos sobre el problema. Quizás en el ejemplo anterior, un análisis a posteriori podría revelar diferentes pronósticos entre las clases propuestas, habiendo encontrado de forma automática el algoritmo un patrón que permite predecir la evolución.

3.1.3. Por refuerzo

Basado en el aprendizaje conductista, estos algoritmos aprenden obteniendo un refuerzo o recompensa cuando toman decisiones correctas. En este caso, no es necesario preparar un conjunto de datos de entrenamiento, sino que solo se requiere de la definición de una métrica que otorgue a cada estado obtenido por el algoritmo una puntuación, de modo que el algoritmo tratará de encontrar los estados que optimicen dicha puntuación.

Un ejemplo sencillo, se podría encontrar en los sistemas de IA que aprenden a jugar a juegos de mesa. Los algoritmos entrenados para jugar al ajedrez son, en muchos casos, algoritmos de ML que aprenden por refuerzo, tratando de maximizar una métrica que tiene tres posibles resultados: -1 (partida perdida), 0 (partida empatada) y 1 (partida ganada).

Por último, es importante señalar que existe un cuarto tipo de aprendizaje que combina aprendizaje supervisado y aprendizaje no supervisado, ya que utiliza para su entrenamiento datos etiquetados y no etiquetados. Esta técnica recibe el nombre de aprendizaje semi-supervisado, y da lugar en ocasiones a confusión con el aprendizaje por refuerzo, sin embargo son conceptos completamente diferentes.

3.2. Algoritmos de Machine Learning en función de la forma de representación interna

Como se ha indicado previamente, el objetivo principal de los algoritmos de ML es el de extraer un conjunto de reglas que permitan resolver una tarea concreta a partir de ejemplos conocidos, cumpliendo la premisa de que esas reglas sean aplicables a nuevos casos, es decir, que permitan generalizar el conocimiento. Existen dos técnicas para lograrlo:

3.2.1. Basados en instancias

Se establece una métrica que permite medir la similitud entre los datos de entrenamiento y datos nuevos a los que tengamos que aplicar el algoritmo. De esta forma, un dato nuevo se clasificará dentro de la categoría a la que pertenezcan los datos más parecidos. Un ejemplo es la técnica de los K-vecinos más cercanos. Para decidir si una radiografía sin diagnóstico pertenece a la clase COVID o control, se proyecta en el espacio de características y se examina su distancia a aquellas radiografías que fueron almacenadas por el sistema inicialmente, deduciendo que pertenece a la clase de las radiografías a las que más se parezca, con las que más características comparta.

3.2.2. Basados en modelos

Se extrae un modelo a partir de los datos de entrenamiento. El sistema no almacena los datos de entrenamiento, sino ciertas transformaciones que permiten resolver la tarea para la que esta diseñado el algoritmo. En el ejemplo de clasificación de radiografías, una red neuronal artificial sería capaz de generar fronteras de decisión y determinar cuáles son las regiones que ocupan las distintas clases de radiografías.

3.3. Algoritmos de Machine Learning en función de la forma de delimitar las fronteras de decisión

3.3.1. Modelos de regresión

Basados en la aplicación de los principios de la estadística al análisis de datos, están considerados como la puerta de entrada al ML. Uno de los más famosos es la regresión logística binaria. En ella, se puede distinguir entre dos clases encontrando los coeficientes óptimos por los que multiplicar las diferentes variables independientes para obtener la variable dependiente, es decir, se trata de ajustar una curva.

3.3.2. Métodos de Kernel

Consisten en un grupo de algoritmos de clasificación. El más conocido es el *Support Vector Machine* (SVM). Tienen como objetivo encontrar un límite de decisión en la forma de un hiperplano que permite separar los datos de entrenamiento en dos subespacios correspondientes a las dos categorías de entrenamiento. La clasificación de nuevos datos consistiría en determinar el subespacio al que pertenecen.

3.3.3. Modelos basados en árboles de decisión

Los árboles de decisión son estructuras en forma de diagrama de flujo que permiten clasificar en función de una serie de condiciones que ocurren de forma sucesiva. Existen técnicas que

tratan de ensamblar varios árboles de decisión entrenados con diferentes muestras de los datos de enteramiento, que reciben el nombre de bosques aleatorios.

Otro ejemplo de un algoritmo de este tipo, son las máquinas de aumento de gradiente, que de manera similar a los bosques aleatorios, ensamblan varios modelos de predicción débiles, generalmente árboles de decisión, para obtener un modelo de mayor potencia predictiva. Utilizan el impulso de gradiente: una forma de optimizar cualquier modelo de aprendizaje automático mediante el entrenamiento iterativo, desarrollando nuevos modelos que se especializan en abordar los puntos débiles de los modelos anteriores. Este algoritmo es hoy en día uno de los mejores para lidiar con datos no perceptuales. Junto con el aprendizaje profundo, es una de las técnicas más utilizadas en competiciones de *Kaggle*. Habitualmente, el aumento de gradiente es el más utilizado para problemas donde los datos están estructurados y para problemas de aprendizaje superficial, mientras que el aprendizaje profundo se utiliza para problemas de percepción, como la clasificación de imágenes.

3.3.4. Redes neuronales primitivas

Frank Rosenblatt creó en 1958 el perceptrón o neurona artificial o unidad básica de inferencia, en forma de discriminador lineal.

Esta unidad elemental está inspirada en las neuronas biológicas. Al igual que en estas, una neurona artificial recibe una o más entradas y las pondera con unos pesos, para producir una salida. Cada entrada se pondera por separado y su suma se pasa a través de una función no lineal conocida como función de activación. Sin la función de activación, la red neuronal solo podría resolver problemas lineales. Estas funciones son equivalentes a las que determinan los potenciales umbral de activación de las neuronas biológicas.

3.3.5. Deep Learning o aprendizaje profundo

Para problemas muy simples, una sola neurona puede ser suficiente, sin embargo los problemas de la vida real suelen requerir de representaciones más complejas de los datos y sus relaciones. Para ello se agrupan varios perceptrones simples en lo que se denomina red neuronal o perceptrón multicapa.

Una red neuronal artificial esta formada por múltiples capas que permiten resolver problemas no lineales, definiéndose como un aproximador universal, lo cual es la principal limitación del perceptrón simple. Estas capas se organizan en una capa de entrada, un conjunto de capas ocultas y una capa de salida. Un ejemplo sería las puertas lógicas: *AND* y *OR* sólo requieren de definir una frontera de decisión lineal y se pueden implementar con un perceptrón simple, mientras que la puerta *XOR* requiere de un perceptrón multicapa dado que se necesita más de una frontera de decisión.

Con el incremento de la capacidad de cómputo, se abrió la posibilidad de entrenar redes con muchas capas, dando lugar a un subconjunto de algoritmos de ML que se denominó *Deep Learning* (DL) o aprendizaje profundo.

Dentro del DL, se distinguen varios tipos de redes neuronales: redes convolucionales, redes recurrentes, redes adversarias, etc. Dentro de estas, cabe destacar las redes densas, en la que cada neurona de la red se encuentra conectada con todas las neuronas de la capas siguiente, y las redes convolucionales, que constituyen una de las arquitecturas más utilizadas en medicina.

3.3.6. Redes Neuronales Convolucionales

Las redes neuronales convolucionales se consideran el estado del arte en visión artificial desde 2012, cuando vencieron a sus competidores en la competición de reconocimiento de imágenes *ImageNet*.

Las redes neuronales convolucionales se basan en la extracción de mapas de características de las imágenes utilizando una serie de filtros aprendidos por la propia red durante el entrenamiento. Para extraer las características la red aplica estos filtros aprendidos a las imágenes a través de la operación matemática de convolución, de ahí el nombre que reciben estas redes.

Con el fin de reducir la dimensionalidad del modelo y la complejidad del mismo, este tipo de redes utilizan además de la operación de convolución otras operaciones de reducción espacial de los mapas de características llamadas *pooling*.

Finalmente, las características extraídas son dadas como entrada a una red neuronal densa que actúa como clasificador.

El hecho de que estas redes realicen una extracción automática de las características representa un gran paso con respecto a los algoritmos clásicos de visión artificial, en los que un experto debía extraer manualmente las características para resolver el problema y luego utilizarlas como entrada para el clasificador.

4. ENTRENAMIENTO DE ALGORITMOS DE *MACHINE LEARNING*

El proceso de entrenamiento consiste en el autoajuste de los parámetros de la red a través de un proceso repetitivo que comienza con el cálculo de las predicciones del algoritmo para el conjunto de datos de entrenamiento, continua con el cálculo del error de dichas predicciones comparándolas con la etiqueta de cada dato, y prosigue con el ajuste de los parámetros en la manera en que permita reducir el error. Esto se consigue gracias a la técnica matemática del descenso de gradiente y al algoritmo de retropropagación, y cada uno de esos ciclos recibe el nombre de época.

Es importante distinguir entre parámetros, que son los pesos y los sesgos que se fijan automáticamente durante el entrenamiento, y los hiperparámetros, que son aquellos ajustes que define el desarrollador antes de comenzar el entrenamiento.

Uno de los problemas que presentan estos algoritmos, es el del sobre-entrenamiento u *overfitting*. Esta situación ocurre cuando el sistema memoriza los datos de entrenamiento no siendo capaz de generalizar su conocimiento a datos nuevos. Para evitarlo, se utiliza en el entrenamiento, además del conjunto de datos de entrenamiento, un conjunto de datos que recibe el nombre de validación. Cuando las métricas de entrenamiento y validación dejan de evolucionar de forma paralela, sabremos que el algoritmo esta cayendo en sobre-entrenamiento.

Por último, se utilizará un tercer conjunto de datos llamado de test, que permitirá evaluar el resultado final del modelo.

Bibliografía

1. Chollet F. (2018). *Deep Learning with Python*. Manning Publications Co.
2. Collen M. F. (1994). The origins of informatics. Journal of the American Medical Informatics Association : JAMIA, 1(2), 91–107.
3. Dias, R., & Torkamani, A. (2019). Artificial intelligence in clinical and genomic diagnostics. Genome medicine, 11(1), 70.
4. European Society of Radiology (ESR) (2019). What the radiologist should know about artificial intelligence - an ESR white paper. Insights into imaging, 10(1), 44.
5. Jones, L. D., Golan, D., Hanna, S. A., & Ramachandran, M. (2018). Artificial intelligence, machine learning and the evolution of healthcare: A bright future or cause for concern?. Bone & joint research, 7(3), 223–225.
6. Choudhury, A., & Asan, O. (2020). Role of Artificial Intelligence in Patient Safety Outcomes: Systematic Literature Review. JMIR medical informatics, 8(7), e18599.
7. Montani, S., & Striani, M. (2019). Artificial Intelligence in Clinical Decision Support: a Focused Literature Survey. Yearbook of medical informatics, 28(1), 120–127.
8. Shortliffe E. H. (1977). Mycin: A Knowledge-Based Computer Program Applied to Infectious Diseases. Proceedings of the Annual Symposium on Computer Application in Medical Care, 66–69.
9. Turing A.M. (1950). Computing Machinery and Intelligence. Mind, 59(236), 433-460.
10. West, E., Mutasa, S., Zhu, Z., & Ha, R. (2019). Global Trend in Artificial Intelligence-Based Publications in Radiology From 2000 to 2018. AJR. American journal of roentgenology, 213(6), 1204–1206.

.