Desarrollo y evaluación de modelos de Inteligencia Artificial para la estimación de la edad ósea en pacientes pediátricos

Germán Casabó Vallés

Resumen — Tradicionalmente, los métodos más comunes para la estimación de la edad ósea son el método de Greulich y Pyle y el método de Tanner y Whitehouse; sin embargo, ambos métodos presentan limitaciones como la variabilidad interobservador debido a la subjetividad de estos. En los últimos años, han surgido diferentes aproximaciones para el cálculo automático de la edad ósea mediante el uso de algoritmos de Inteligencia Artificial (IA). En el presente trabajo, hemos desarrollado y evaluado diferentes modelos de IA basados en redes neuronales convolucionales (CNNs) para el cálculo automático de la edad ósea y hemos comparado sus rendimientos entre sí y respecto a los métodos tradicionales y automáticos publicados. Para ello, se ha utilizado un conjunto de datos formado por 12.611 radiografías de la mano izquierda de pacientes pediátricos anotadas con su edad ósea y el sexo del paciente y se han explorado diferentes arquitecturas y técnicas de optimización. Las CNNs que mejores resultados han obtenido están basadas en la arquitectura ResNet50 y presentan errores medios absolutos (MAEs) de 12,15 y de 12,49 meses para imágenes de varones y hembras, respectivamente. Finalmente, se ha entrenado una CNN con imágenes únicamente de varones de entre 10 y 15 años, obteniendo un MAE de 9,09 meses. Estos resultados están en línea con la variabilidad descrita en la práctica clínica (entre 5,4 y 9,96 meses) y en otros modelos de IA (entre 4,2 y 9,96 meses).

Palabras clave —edad ósea, inteligencia artificial, redes neuronales convolucionales, MAE

Abstract— Traditionally, the most common methods for bone age estimation are the Greulich and Pyle method and the Tanner and Whitehouse method; however, both methods have limitations such as interobserver variability due to the subjectivity of them. In recent years, different approaches have emerged for the automatic calculation of bone age using Artificial Intelligence (AI) algorithms. In the present work, we have developed and evaluated different AI models based on convolutional neural networks (CNNs) for the automatic calculation of bone age and have compared their performances with each other and with respect to traditional and automatic published methods. For this purpose, a dataset consisting of 12,611 left hand radiographs of pediatric patients annotated with their bone age and patient sex has been used and different architectures and optimization techniques have been explored. The best performing CNNs are based on the ResNet50 architecture and present mean absolute errors (MAEs) of 12.15 and 12.49 months for male and female images, respectively. Finally, a CNN has been trained with images only of males between 10 and 15 years old, obtaining a MAE of 9.09 months. These results are in line with the variability described in clinical practice (between 5.4 and 9.96 months) and in other AI models (between 4.2 and 9.96 months).

Index Terms— bone age, artificial intelligence, convolutional neural networks, MAE

1 Introducción

1.1 Edad ósea

La edad ósea es un indicador clínico que permite evaluar el estado de maduración esquelética de un individuo, comúnmente a partir de los cambios de los centros de osifi cación a lo largo del tiempo. [1].

La determinación de la edad ósea ha sido útil en una variedad de contextos clínicos durante más de 75 años, destacando en el ámbito pediátrico, en el que la estimación de la edad ósea ayuda a detectar y tratar desórdenes de crecimiento, predecir la potencial altura futura, problemas endocrinológicos, etc. [1].

En este sentido, la edad ósea es el único indicador de madurez biológica, independiente del tamaño, que se usa de forma rutinaria desde el nacimiento hasta la adultez [2].

Además de en el entorno clínico, la determinación de la edad ósea es útil en el campo del deporte de élite para la

- *E-mail de contacto: gcasabovalles@gmail.com*
- *Trabajo tutorizado por: Antonio Lozano Bagen*
- *Curso 2023/24*

selección de atletas, en contextos forenses, e incluso en programas de inmigración internacionales para estimar la edad de menores solicitantes de asilo [1].

1.2 Métodos tradicionales para la estimación de la edad ósea

Tradicionalmente, los métodos más comunes para la estimación de la edad ósea son el método de Greulich y Pile [3] y el método de Tanner y Whitehouse [4], ambos basados en el análisis de radiografías de la mano izquierda de pacien tes pediátricos.

El método de Greulich y Pyle se basa en la comparación visual de las radiografías a analizar frente a un atlas con radiografías de referencia, mientras que el método de Tanner y Whitehouse se basa en asignar una puntuación en función de los diferentes estados que pueden presentar los centros de osificación y combinar dichas puntuaciones para obtener una estimación de la edad ósea.

Un estudio de 2016 [5] identificó que el método de Greulich y Pyle [3] era el más utilizado por los especialistas pediátricos americanos para el cálculo de la edad ósea, llegando a alcanzar un 97,4% de uso en niños mayores de 3 años. Otro estudió identificó, en cambio, que el método de Tanner y Whitehouse es el preferido entre especialistas endocrinólogos europeos [6].

No obstante, a pesar de su amplio uso entre los profesionales clínicos, ambos métodos presentan varias limitaciones, como por ejemplo una inherente variación inter e intraobser vador debida a la subjetividad de los métodos [6]. Diversos estudios [7], [8], [9], [10] se han centrado en abordar esta variabilidad y han demostrado que la desviación estándar sobre una determinación en estudios interobservador varía entre 0,45 y 0,83 años, es decir, entre 5,4 y 9,96 meses, aproxi madamente [6].

1.3 Desafío de la Sociedad Radiológica de América del Norte

En los últimos años, el aumento de la capacidad computacional y los avances en los algoritmos de inteligencia artificial (IA) han revolucionado el campo de las imágenes médicas con la aparición de un tipo específico de aprendizaje profundo conocido como redes neuronales convolucionales (CNNs, por sus siglas en inglés) [11].

Las CNNs son especialmente efectivas en la detección de patrones complejos dentro de imágenes, y su capacidad para aprender de grandes volúmenes de datos las convierte en una herramienta prometedora para superar las limitaciones de los métodos tradicionales de análisis de las imágenes médicas [12].

En 2017, como parte de sus esfuerzos para impulsar el uso de herramientas basadas en IA para radiología, la Sociedad Radiológica de América del Norte (RSNA, por sus siglas en inglés) organizó un desafío para evaluar el rendimiento que presentaban los algoritmos de IA ejecutando una actividad de lo más común para muchos radiólogos pediátricos: estimar la edad ósea de pacientes pediátricos a partir de radiografías de sus manos [13], [14].

Los resultados del desafío mostraron el enorme potencial de las CNNs para el cálculo de la edad ósea, reduciendo la variabilidad entre observadores y mejorando la precisión y consistencia de las mediciones [13], [14].

2 Objetivos

El objetivo principal de este trabajo es desarrollar modelos basados en redes neuronales convolucionales (CNNs) para calcular la edad ósea de manera automática a partir de radiografías de mano de pacientes pediátricos.

Los objetivos específicos incluyen:

- Revisar el estado del arte relativo a métodos basados en IA para la estimación de la edad ósea.
- Explorar y aplicar diferentes arquitecturas de CNNs para el análisis de las radiografías.
- Comparar el rendimiento de los modelos desarrollados entre sí y respecto a los métodos tradicionales y automáticos para la evaluación de la edad ósea.

3 Estado del arte

En los últimos años, se han publicado diferentes trabajos

originales [15], [16], [17], [18] y revisiones bibliográficas [19], [20] sobre la aplicación de CNNs para la estimación automática de la edad ósea.En uno de estos trabajos, publicado en el año 2.019, Dallora y colaboradores realizan un análisis comparativo de diferentes modelos de IA para la estimación de la edad ósea y obtuvieron un error promedio absoluto (MAE, por sus siglas en inglés) de 9,96 meses [20].

Sin embargo, muy pocas de estas herramientas basadas en IA han sido comercializadas [19]. En la actualidad, BoneXpert (Visiana, Dinamarca) es el único sistema con marcado CE, es decir, es la única herramienta basada en IA que se puede utilizar en un entorno clínico real en la Unión Europea para la estimación automatizada de la edad ósea. Su uso está indicado para niños de 2,5 a 17 años y niñas de 2 a 15 años independientemente de su etnia y presenta una desviación estándar de 0,63 años (7,56 meses) cuando se compara con el método de Greulich y Pyle [10], [21].

Por último, tal y como hemos comentado en la introducción, el presente trabajo se basa en el reto que propuso la RSNA en el año 2.017 [13], [14] que, a su vez, se basa en los datos publicados por Larson y colaboradores [12]. El equipo de Larson consiguió un MAE de 0,52 años (6,24 meses) [12], mientras que los 5 primeros equipos clasificados en el reto de la RSNA consiguieron MAEs de entre 4,2 y 4,5 meses [14].

4 Material y métodos

4.1 Conjunto de datos

El conjunto de datos utilizado consta de 12.611 radiografías de la mano izquierda de pacientes pediátricos (10,8 ± 3,5 años de edad cronológica y 10,6 ± 3,4 años de edad ósea estimada) procedentes de dos hospitales americanos: el Lucile Packard Children's Hospital de Stanford (California, Estados Unidos) y el Children's Hospital Colorado de Aurora (Colorado, Estados Unidos). En la Figura 1 se muestra una radiografía incluida en el conjunto de datos como ejemplo.

Fig. 1. Ejemplo de radiografía de mano.

Cada una de las imágenes está etiquetada con la edad ósea (en meses) y el sexo del paciente (el conjunto de datos consta de 6.833 varones y 5.778 hembras).

Las imágenes fueron inicialmente utilizadas, previa aprobación por los Comités de Ética (*Institutional Review Boards*) de ambas instituciones, por Larson y colaboradores [12] para comparar el rendimiento en la estimación de la edad ósea de un modelo de aprendizaje profundo respecto a las estimaciones de radiólogos expertos y de los modelos automatizados existentes.

Posteriormente, los Comités de Ética aprobaron la curación y el uso de las radiografías por parte de la Sociedad Norteamericana de Radiología para su competición de inteligencia artificial [14] en la plataforma Kaggle (California, Estados Unidos), siempre y cuando se utilicen para propósitos académicos o educacionales y se atribuya el origen de los datos adecuadamente.

En este sentido, debemos destacar que las imágenes se encuentran accesibles en la dirección https://www.rsna.org/rsnai/ai-image-challenge/rsna-pediatric-bone-age-challenge-2017 y también como *dataset*

de Kaggle en la dirección https://www.kaggle.com/datasets/kmader/rsna-bone-age.

4.2 Selección de un subconjunto de datos

Por motivos computacionales, durante la fase inicial del desarrollo de modelos de inteligencia artificial, se seleccionaron 2.018 imágenes (un 16% del total de las 12.611) para evaluar el mayor número posible de CNNs diferentes y explorar cuál era la arquitectura con un mayor potencial para entregar un buen rendimiento.

Para la selección de las 2.018 imágenes, se categorizaron las imágenes a partir de su edad esquelética en 15 grupos etarios (el número de *bins* se obtuvo mediante el método de Sturges [22]) y se realizó una partición estratificada del conjunto de datos para asegurar que la distribución de los valores de edad **ósea** de las 2.018 imágenes seleccionadas era representativa de la distribución de los valores de edad ósea del conjunto de datos.

4.3 *Software* y *hardware*

Para la realización de este trabajo, se ha utilizado la plataforma de Kaggle, ya que proporciona acceso a entornos de alto rendimiento con unidades de procesamiento gráfico (GPU, por sus siglas en inglés) y unidades de procesamiento central (CPU, por sus siglas en inglés) optimizadas para el entrenamiento de modelos de redes neuronales.

El desarrollo y entrenamiento de los modelos se ha realizado utilizando *notebooks* de Kaggle y el lenguaje de programación Python (versión 3.10.14).

Respecto al *software*, se han utilizado librerías especializadas de aprendizaje automático y aprendizaje profundo como scikit-learn [23], Keras [24], y TensorFlow [25]. Además, también se han utilizado las librerías Pandas [26], Numpy [27], Matplotlib [28], os, scipy [29], pickle, tqdm [30] y PIL [31] para funciones relacionadas con el manejo, la transformación, y la visualización de las imágenes y los datos.

Respecto al *hardware*, se ha utilizado la GPU P100 (incluida en las opciones de Kaggle) para acelerar los tiempos de entrenamiento de las CNNs y un ordenador personal con conexión a internet para acceder a los recursos de Kaggle.

4.4 Desarrollo de los modelos de CNNs

Durante el desarrollo del presente trabajo, se desarrollaron diferentes modelos de redes neuronales para el cálculo de la edad ósea, partiendo de redes neuronales sencillas y poco profundas hasta la utilización de redes neuronales complejas, como VGG16 [32], ResNet50 [33], InceptionV3 [34], EfficientNetB0 [35], Xception [36], MobileNetV2 [37], DenseNet121 [38] y NASNetMobile [39].

Se han desarrollado modelos con un subconjunto de los datos formado por 2.018 imágenes escogidas de manera que fuesen representativas del total del conjunto de datos, con el conjunto total de los datos (separando pacientes masculinos y femeninas), y con un subconjunto de pacientes masculinos de entre 10 y 15 años de edad ósea. En todos los casos, se ha realizado una partición estratificada 80/20 del conjunto de datos entre el subconjunto de entrenamien to y el subconjunto de validación.

Se han implementado diferentes técnicas para la optimización de los modelos, como la inclusión de pasos de Dropout [40] y Batch Normalization [41], la reducción del Learning Rate, la configuración de Early Stopping, y el uso técnicas avanzadas como aprendizaje por transferencia o *transfer learning*, ajuste fino o *fine-tuning*, y aumentación de datos o *data augmentation*.

Todos los modelos generados han utilizado el algoritmo Adam [42] y el error cuadrático medio (MSE, por sus siglas en inglés) como función de pérdida para la optimización de los modelos. La métrica escogida para la evaluación del rendimiento de los modelos ha sido el error absoluto medio (MAE, por sus siglas en inglés).

4.5 Manejo de la variable sexo

Las imágenes venían etiquetadas únicamente con la edad ósea (variable objetivo) y el sexo (variable predictora). Se han utilizado 3 aproximaciones diferentes para tratar la variable sexo:

- No inclusión en los modelos.

- Concatenación de la entrada de la variable sexo a la salida de la rama convolucional de la red neuronal.
- Creación de redes neuronales independientes para varones y para hembras.

5 Resultados y discusión

5.1 Desarrollo de CNNs preliminares con un subconjunto de los datos

En primer lugar, se desarrollaron diferentes CNNs utilizando un subconjunto representativo de las imágenes disponibles con el objetivo de observar el comportamiento de diferentes arquitecturas y técnicas de optimización y de escoger la mejor aproximación con la que analizar todas las imágenes.

Este subconjunto, formado por 2.018 de las 12.611 imágenes, fue seleccionado de manera que se garantizase una distribución equilibrada y representativa de la población completa.

Para validar esta selección, se realizó una prueba de la t para comparar la distribución de los valores de edad ósea del conjunto de datos original y del conjunto de datos selec cionado, y se obtuvo un valor p de 0,98, lo que indicó la alta similaridad entre ambas distribuciones de datos.

Todos los modelos se entrenaron con la misma partición 80/20 entre el subconjunto de entrenamiento y el subconjunto de validación de las 2.018 imágenes, durante un máximo de 300 épocas, un tamaño de lote de 16 imágenes por iteración, y con *early stopping* para detener el entrenamiento si no se observaba una disminución del MSE en el conjunto de validación durante 10 épocas (paciencia).

5.1.1 Redes neuronales poco profundas

En primer lugar, se utilizó una red neuronal poco profunda consistente en 2 capas de convolución con 32 y 64 filtros, activación ReLU, y tamaño de kernel de 3x3, seguidas de capas de MaxPooling, un aplanamiento de la salida de la parte convolucional, y una capa densa de 128 neuronas antes de la capa de salida para la predicción final de la edad ósea (Modelo 1). Además, se desarrolló otro modelo igual a este, pero concatenando la entrada de la variable sexo a la salida de la parte convolucional (Modelo 2).

En segundo lugar, se añadieron sobre estas redes técnicas de regularización mediante la inclusión de capas de Batch Normalization y de Dropout después de cada capa convolucional y densa (Modelos 3 y 4).

En la Tabla 1 se presentan los resultados de la época con menor MSE en el conjunto de validación de cada uno de los modelos desarrollados en este apartado. Tal y como se puede observar, la adición de las técnicas de regularización mejoró los resultados; sin embargo, las redes siguieron sobreajustándose a los datos de entrenamiento (se aprecia un MAE mucho menor en el conjunto de entrenamiento (*train*) que en el conjunto de validación (*val*)).

TABLA 1

Resultados de las CNNs poco profundas

Modelo	*Época*	*MAE (train)*	*MAE (val)*
1	4	27,46	34,93
2	14	8,59	32,96
3	14	17,08	29,63
4	19	15,46	27,65

Modelo 1: CNN poco profunda sin regularización ni información sobre sexo; Modelo 2: CNN poco profunda sin regularización, pero con la variable sexo; Modelo 3: CNN poco profunda con regularización, sin la variable sexo; Modelo 4: CNN poco profunda con regularización y variable sexo.

En la Figura 2 se muestra el gráfico de entrenamiento de la CNN que mejor rendimiento entregó en este apartado (Modelo 4), en el que puede observarse como el modelo sufrió sobreajuste desde antes de la época 10.

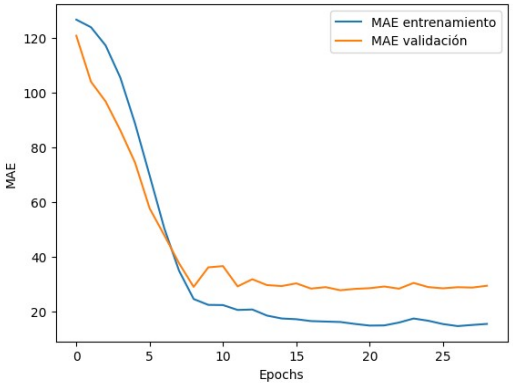

Fig. 2. Gráfico del entrenamiento de una CNN poco profunda con técnicas de regularización e información sobre la variable sexo.

5.1.2 Redes neuronales con arquitectura VGG16

Los resultados del apartado anterior evidenciaron que las redes neuronales simples y poco profundas no eran capaces de conseguir extraer los patrones necesarios de las imágenes para el cálculo de la edad ósea, así como de gene ralizar a datos no vistos y presentar buenos resultados en el conjunto de validación, por lo que se optó por evaluar arquitecturas más complejas como la VGG16, que consta de 16 capas de profundidad [32].

En primer lugar, se implementó una red neuronal con arquitectura VGG16 desde 0 (Modelo 5). Además, se desarrolló otro modelo igual a este, pero concatenando la entrada de la variable sexo a la salida de la parte convolucional (Modelo 6). Para el entrenamiento de ambos modelos se redujo el *learning rate* a 0,00001.

A continuación, se optó por incorporar técnicas avanzadas de aprendizaje profundo como *transfer learning*, *fine-tuning*, y *data augmentation*.

Respecto al *transfer learning*, se cargó el modelo VGG16 preentrenado con los pesos de ImageNet [11], [43] disponible en Keras con las capas convolucionales congeladas y sin añadir las capas superiores, ya que se añadió una capa densa de 512 neuronas seguida de una capa de Dropout (Modelo 7) después de aplanar la parte convolucional de la red y antes de la capa de salida. Además, se desarrolló otro modelo igual a este, pero concatenando la entrada de la variable sexo a la salida de la parte convolucional (Modelo 8). Para el entrenamiento de ambos modelos se redujo el *learning rate* a 0,00001.

Posteriormente, se implementó la técnica de *fine-tuning* cargando la red VGG16 con los pesos de ImageNet, como en el Modelo 7, pero descongelando las últimas 4 capas para ajustar mejor la CNN al conjunto de datos (Modelo 9). Se redujo el *learning rate* a 0,000001 para garantizar la estabilidad del entrenamiento.

Finalmente, se desarrolló otro modelo consiste en añadir sobre el Modelo 9 un paso de *Data Augmentation* para aumentar la diversidad de las imágenes de entrenamiento y, por tanto, reducir el riesgo de sobreajuste de la CNN (Modelo 10). El *learning rate* se mantuvo en 0,000001.

En la Tabla 2 se muestran los resultados de la época con menor MSE en el conjunto de validación obtenidos con las diferentes aproximaciones de VGG16.

TABLA 2

Resultados de las CNNs basadas en VGG16

Modelo	*Época*	*MAE (train)*	*MAE (val)*
5	69	10,57	21,91
6	65	11,50	23,40
7	42	10,75	18,33
8	53	10,48	18,12
9	73	11,19	17,50
10	29	24,01	24,77

Modelo 5: VGG16 desde 0; Modelo 6: VGG16 desde 0 + variable de sexo; Modelo 7: VGG16 preentrenada; Modelo 8: VGG16 preentrenada + variable de sexo; Modelo 9: VGG16 preentrenada + *fine tuning*; Modelo 10: VGG16 preentrenada + *fine tuning* + *data augmentation*

En primer lugar, se observa claramente como los dos modelos basados en VGG16 que peor funcionaron son aquellos en los que la red neuronal se entrenó desde 0 (Mo delos 5 y 6). Esto podría ser debido a que el número de imágenes (2.018) era demasiado bajo como para ajustar adecuadamente todos los parámetros de la red.

En este sentido, los modelos que utilizaron *transfer learning* se manifestaron como aquellos con los mejores rendimientos, con MAEs en el conjunto de validación de alrededor de 18 meses, aunque continuaron presentando un sobreajuste a los datos de entrenamiento.

El modelo que mejores resultados presentó, el Modelo 9, que incluyó *tranfer learning* y *fine-tuning* en las últimas 4 capas convolucionales, alcanzó un MAE de 17,50 meses en el conjunto de validación. Estos resultados indicaron que la técnica de *transfer learning* utilizando los pesos de ImageNet era útil para mejorar los resultados en nuestro conjunto de datos.

Por último, el modelo que incorporó la técnica de *data augmentation* consiguió evitar el sobreajuste, ya que el MAE tanto en el conjunto de entrenamiento como en el de valida ción fue muy similar; sin embargo, presentó un rendimiento notablemente en el conjunto de validación inferior a su equivalente sin *data augmentation* (24,77 vs. 17,50).

En la Figura 3 se muestra el gráfico del entrenamiento del Modelo 9, ya que fue el que mejores resultados presentó. Se observa cómo, aunque el rendimiento es el mejor hasta el momento, se produjo un sobreajuste a partir de la época 20.

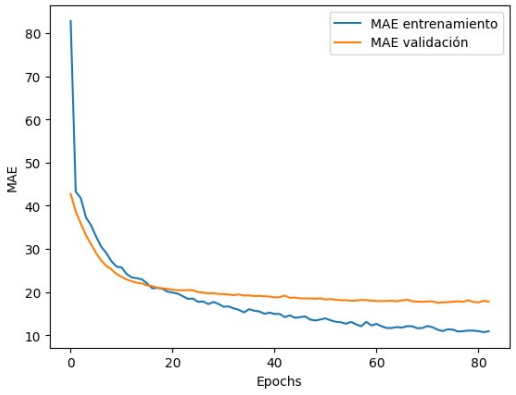

Fig. 3. Gráfico de entrenamiento de una CNN con arquitectura VGG16 con pesos preentrenados y *fine-tuning* en las últimas 4 capas.

5.1.3 Redes neuronales con otras arquitecturas

Los resultados del apartado anterior indicaron que el uso de la técnica de *transfer learning*, es decir, cargar las redes con pesos preentrenados con el conjunto de imágenes de ImageNet, era la estrategia que mejores resultados presentaba, por lo que a continuación se optó por evaluar dife rentes arquitecturas preentrenadas para evaluar cuál era la arquitectura con mayor potencial de rendimiento.

En particular, se evaluaron las arquitecturas ResNet50 [33], InceptionV3 [34], EfficientNetB0 [35], Xception [36], MobileNetV2 [37], DenseNet121 [38] y NASNetMobile [39].

La estrategia adoptada fue la misma en todos los casos (Modelos 11-17): se concatenó la entrada de la variable sexo a la salida de la base preentrenada de cada una de las redes (todas las capas convolucionales se mantuvieron congeladas), y se añadió una capa densa de 128 neuronas y una capa de Dropout antes de generar la salida de la red. Se utilizó un *learning rate* de 0,00001.

En la Tabla 3 se presentan los resultados obtenidos en la época con menor MSE en el conjunto de validación de cada uno de los modelos.

TABLA 3

Resultados de las CNNs basadas en otras arquitecturas

Modelo	*Época*	*MAE (train)*	*MAE (val)*
11	32	15,18	16,00
12	48	21,20	27,41
13	77	18,31	16,76
14	62	19,36	27,17
15	111	19,76	24,26
16	63	19,43	22,26
17	103	25,09	30,07

Modelo 11: ResNet50; Modelo 12: Inception V3; Modelo 13: EfficientNetB0; Modelo 14: Xception; Modelo 15: MobileNetV2; Modelo 16: DenseNet121; Modelo 17: NASNetMobile

Tal y como puede observarse, el rendimiento de los modelos fue muy desigual en función de la arquitectura utilizada. De hecho, únicamente las CNNs con arquitectura ResNet50 (MAE en conjunto de validación de 16,00 meses) y EfficientNetB0 (MAE en conjunto de validación de 16,76 meses) mejoraron el rendimiento obtenido con la arquitectura VGG16.

En la Figura 4 se muestra el gráfico del entrenamiento del Modelo 13, es decir, de la CNN con arquitectura ResNet50 que mejores resultados presentó en este apartado. Puede observarse como durante las 30 primeras épocas no se aprecia sobreajuste del modelo.

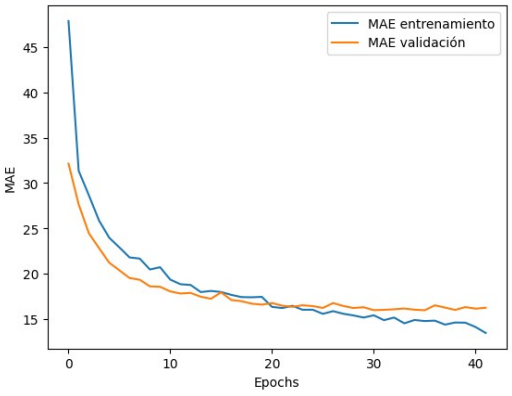

Fig. 4. Gráfico de entrenamiento de una CNN con arquitectura ResNet50 con pesos preentrenados.

5.2 Desarrollo de CNNs con todos los datos y específicas de sexo

Llegados a este punto, se concluyó que la arquitectura que más potencial tenía para entregar un mejor rendimiento en nuestro conjunto de datos era ResNet50, arquitectura que cuenta con 50 capas de profundidad [33], por lo que se decidió utilizar todos los datos disponibles para desarrollar los modelos definitivos.

Respecto a la inclusión o no de la variable sexo, se observó que, aunque si bien era cierto que los modelos que incluían esta información presentaban un rendimiento ligeramente mejor que sus homólogos que no tenían en cuenta esta información, la mejora en el rendimiento tampoco era sustancial. Por este motivo, y también para aligerar la carga computacional de entrenar las CNNs con todos los datos, se optó por desarrollar CNNs independientes y específicas para cada sexo.

Por tanto, para cada conjunto de datos (6.833 radiografías procedentes de pacientes masculinos, y 5.788 radiografías procedentes de pacientes femeninas) se desarrollaron CNNs utilizando 4 aproximaciones distintas, todas ellas basadas en *transfer learning* con los pesos preentrenados en ImageNet:

- *Transfer learning* (Modelos 18 y 22)

- *Transfer learning + data augmentation* (Modelos 19 y 23)
- *Transfer learning + fine-tuning* (Modelos 20 y 24)
- *Transfer learning + fine-tuning + data augmentation* (Modelos 21 y 25)

De forma parecida a la dispuesta en el apartado anterior, todos los modelos se entrenaron con la misma partición 80/20 entre el subconjunto de entrenamiento y el subconjunto de validación, con un tamaño de lote de 32 imágenes por iteración, un máximo de 300 épocas con parada prema tura a las 10 épocas de no mejorar el MSE en el conjunto de validación, y con *learning rates* variables entre 0,00001 y 0,000001 en función de las necesidades de estabilizar los resultados del entrenamiento.

En la Tablas 4 y 5 se presentan los resultados obtenidos en las CNNs desarrolladas con las imágenes procedentes de pacientes masculinos y femeninas, respectivamente.

TABLA 4

Resultados de las CNNs basadas en ResNet50 con datos de pacientes masculinos

Modelo	*Época*	*MAE (train)*	*MAE (val)*
18	39	9,12	12,15
19	30	16,69	13,71
20	17	9,72	12,32
21	42	16,68	14,53
22	31	9,23	13,00
23	50	15,31	13,48
24	86	6,71	12,49
25	97	14,52	13,00

Modelo 18: ResNet50 con *transfer learning;* Modelo 19: ResNet50 con *transfer learning* y *data augmentation*; Modelo 20: ResNet50 con *transfer learning* y *fine-tuning*; Modelo 21: ResNet50 con *transfer learning*, *fine-tuning*, y *data-augmentation*.

TABLA 5

Resultados de las CNNs basadas en ResNet50 con datos de pacientes femeninos

Modelo	*Época*	*MAE (train)*	*MAE (val)*

Modelo 22: ResNet50 con *transfer learning;* Modelo 23: ResNet50 con *transfer learning* y *data augmentation*; Modelo 24: ResNet50 con *transfer learning* y *fine-tuning*; Modelo 25: ResNet50 con *transfer learning*, *fine-tuning*, y *data-augmentation*.

Tal y como puede observarse, esta estrategia produjo una mejora significativa del MAE en el conjunto de validación, situándose alrededor de 12 meses en las CNNs entrenadas con imágenes de cada uno de los sexos.

Interesantemente, se observó un comportamiento similar al observado durante el entrenamiento de las redes neuronales basadas en la arquitectura VGG16 con un subconjunto de los datos (Tabla 2). Tanto en el caso de la CNN entrenada con radiografías de pacientes masculinos como en la CNN entrenada con radiografías de pacientes femeni nas, los mejores resultados en el conjunto de validación se obtuvieron en los modelos que no incluyen un paso de *data augmentation*.

Si bien es cierto que la inclusión de esta técnica provocó que desapareciese el ligero sobreajuste que presentaban las CNN sin *data augmentation* (véase resultados de los modelos 19 y 21 en la Tabla 4 y de los modelos 23 y 25 en la Tabla 5, en los que el MAE en el conjunto de validación es incluso menor que el MAE en el conjunto de entrenamiento), esta inclusión provocó también un aumento absoluto del MAE en el conjunto de validación respecto a los modelos sin *data augmentation*.

Así pues, la CNN con mejores resultados (MAE de 12,15 meses en el conjunto de validación) en las imágenes de pacientes masculinos es aquella que incluyó únicamente *transfer learning*, sin descongelación de ninguna capa convolucional. En la Figura 5 se muestra el gráfico de entrenamiento de esta CNN.

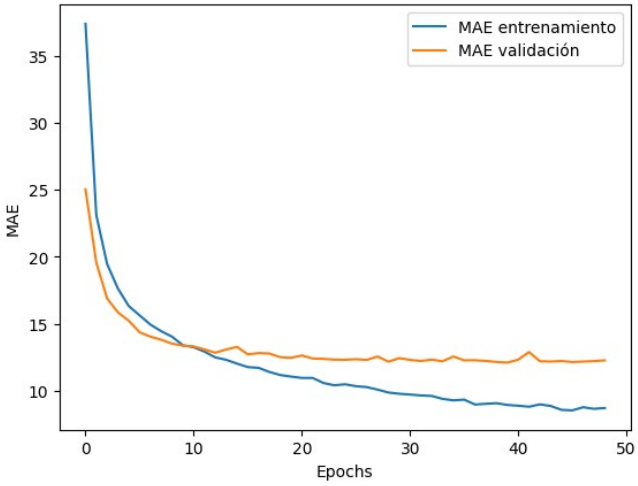

Fig. 5. Gráfico de entrenamiento de la CNN con arquitectura ResNet50 con pesos preentrenados (*transfer learning*) y datos de pacientes masculinos.

Por su parte, en el caso de las pacientes femeninas, el modelo con mejores resultados es aquel que utilizó tanto el *transfer learning* como el *fine-tuning* mediante la descongelación de las últimas capas convolucionales. En la Figura 6 se muestra el gráfico de entrenamiento de esta CNN.

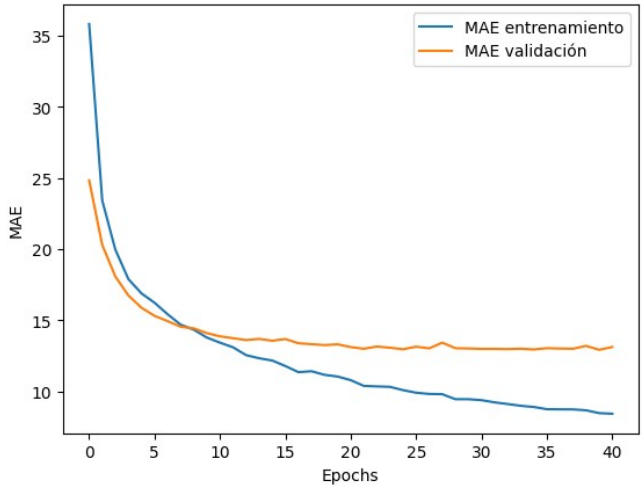

Fig. 6. Gráfico de entrenamiento de la CNN con arquitectura ResNet50 con pesos preentrenados (*transfer learning*), *fine-tuning* y *data augmentation* y datos de pacientes femeninos.

5.3 Exploración de estratificación por grupo etario

Por último, se realizó un experimento consistente en seleccionar únicamente las radiografías procedentes de pacientes masculinos de entre 120 y 180 meses de edad ósea estimada para evaluar si una posible estratificación por grupo etario ayudaba a conseguir mejores resultados. Se seleccionaron 4.424 imágenes, lo que suponían un 64,74 % del total de imágenes correspondientes a hombres.

Se utilizó la misma estrategia descrita anteriormente, es decir, sobre la base de una CNN con arquitectura ResNet50 preentrenada con los pesos de ImageNet, se aplicó:

- *Transfer learning* (Modelo 26)
- *Transfer learning + data augmentation* (Modelo 27)
- *Transfer learning + fine-tuning* (Modelo 28)
- *Transfer learning + fine-tuning + data augmentation* (Modelo 29)

En la Tabla 6 se presentan los resultados obtenidos. De igual modo que en los apartados anteriores, se realizó una partición 80/20 entre el subconjunto de entrenamiento y el de validación, y se entrenaron los modelos durante un máximo de 300 épocas (parada prematura a las 10 épocas sin mejorar el MSE en el conjunto de validación) con un tamaño de lote de 32 imágenes por iteración y *learning rates* variables entre 0,00001 y 0,000001.

TABLA 6

Resultados de las CNNs basadas en ResNet50 con datos de hombres de entre 120 y 180 meses de edad ósea

Modelo	*Época*	*MAE (train)*	*MAE (val)*
26	20	8,47	9,09
27	33	12,77	9,46
28	12	9,92	9,15
29	57	11,79	10,62

Modelo 26: ResNet50 con *transfer learning;* Modelo 27: ResNet50 con *transfer learning* y *data augmentation*; Modelo 28: ResNet50 con *transfer learning* y *fine-tuning*; Modelo 29: ResNet50 con *transfer learning*, *fine-tuning*, y *data-augmentation*.

Interesantemente, el mejor resultado lo obtuvo la CNN que utilizó únicamente la técnica de *transfer learning* con un MAE en el conjunto de validación de 9,09 meses. En la Figura 7 se muestra el gráfico del entrenamiento de esta red.

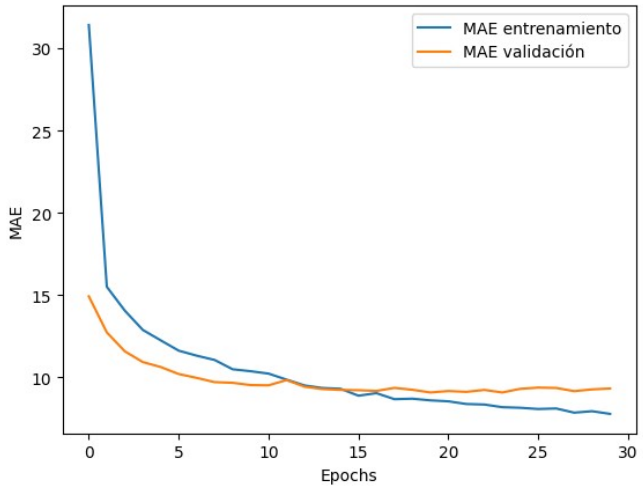

Fig. 7. Gráfico de entrenamiento de la CNN con arquitectura ResNet50 con pesos preentrenados (*transfer learning)* y datos de pacientes masculinos de entre 120 y 180 meses de edad ósea.

Como se puede observar, este resultado mejoró sustancialmente (de 12,15 a 9,09 meses de MAE) cualquier resultado obtenido previamente, lo que indicó que restringir el modelo a un grupo de sexo y edad específico mejoró la capacidad de los modelos para predecir la edad ósea.

5.4 Comparación de los resultados de los modelos de IA desarrollados frente al estado del arte y a la práctica clínica

Durante los apartados anteriores, hemos ido observando como el desarrollo sucesivo de CNNs fue mejorando el rendimiento de estas, entendiendo como mejora del rendimiento una disminución del MAE en el conjunto de validación.

Así pues, las primeras CNNs sencillas y poco profundas entrenadas sobre un subconjunto de los datos consiguieron MAEs de entre 27,65 y 34,93 meses. Posteriormente, las CNNs basadas en la arquitectura VGG16 mejoraron estas cifras a MAEs de entre 17,50 y 24,77 meses. En ese momento, se probaron diferentes arquitecturas de CNNs, con rendimientos variables que fueron desde los 16,00 meses de MAE en el caso de la ResNet50 hasta 30,07 en el caso de NASNet.Mobile.

Estos resultados han mejorado notablemente cuando se han entrenado CNNs basadas en ResNet50 específicas de sexo con todos los datos disponibles para ello, pasando de un MAE de 16,00 meses a MAEs de 12,15 y 12,49 meses para pacientes masculinos y femeninos, respectivamente. Finalmente, la CNN entrenada únicamente con radiografías de pacientes masculinos de entre 10 y 15 años ha conseguido un MAE de 9,09 meses, lo que supone una mejora sustancial del rendimiento de los modelos previos.

Tal y como hemos comentado en los apartados de Introducción y de Estado del arte, los resultados obtenidos están en línea con lo descrito en la bibliografía respecto a la variabilidad existente en los métodos tradicionales y en algunos métodos basados en IA.

Por un lado, algunos autores han cifrado la variabilidad interobservador existente en la práctica clínica entre 5,4 y 9,96 meses [6]. Por otro lado, los modelos basados en IA para el cálculo automático de la edad ósea presentan rendimientos de entre 4,2 [14] y 9,96 meses [20].

Si bien es cierto que algunos modelos de IA, como los 5 primeros clasificados del reto de la RSNA, obtienen rendimientos muy buenos y muy por encima de los encontrados en la práctica clínica habitual, podemos concluir que nuestro modelo de CNN para niños de entre 10 y 15 años obtiene resultados equivalentes a los encontrados en la práctica clínica.

6 Conclusiones y líneas abiertas

En el presente trabajo hemos desarrollado y evaluado diferentes modelos de IA basados en CNNs para el cálculo automático de la edad ósea en pacientes pediátricos y hemos comparado sus rendimientos entre sí, y frente a la variabilidad asociada tanto a los métodos tradicionales utilizados en la práctica clínica como a otros modelos de IA.

Se han evaluado innumerables CNNs (los 29 modelos presentados en el trabajo son solo una muestra representa-

tiva de ellas), desde redes neuronales sencillas y poco profundas que no eran capaces de aprender la complejidad de los datos, hasta redes neuronales muy profundas y comple jas que han permitido mejorar sustancialmente el rendimiento de los modelos hasta alcanzar un MAE de 9,09 meses en el conjunto de validación, rendimiento equiparable al encontrado en la práctica clínica [6] y a otros modelos de IA publicados [20].

El aprendizaje por transferencia, o *transfer learning*, se ha revelado como una técnica necesaria para obtener el máximo aprovechamiento de los datos. También se ha observado, en línea con lo publicado por Larson y colaboradores [12], que, aunque desarrollar CNNs con un subconjunto de las imágenes no permite obtener los mejores resultados, sí que puede ser una herramienta útil en contextos dónde la capacidad computacional es limitada para explorar diferentes arquitecturas antes de desarrollar los modelos finales con todos los datos.

Por último, respecto a futuras líneas de investigación, creemos que es especialmente interesante la sustancial mejora de rendimiento observada al acotar el rango de edad ósea de los pacientes utilizados en el entrenamiento de la red neuronal.

En nuestra opinión, este resultado abriría las puertas al desarrollo de un *pipeline* basada en IA que conste de 2 pasos para la estimación de la edad ósea. En primer lugar, se utilizaría una CNN (específica de sexo) para clasificar la radiografía en un grupo etario concreto. En segundo lugar, se utilizaría una CNN previamente entrenada con imágenes del sexo y grupo etario correspondiente para afinar en la estimación de la edad ósea. Creemos que la aplicación de estos dos modelos de IA permitiría mejorar los resultados en la estimación automática de la edad ósea, tal y como hemos visto en la prueba piloto realizada con una CNN entrenada con imágenes de pacientes masculinos de entre 10 y 15 años de edad ósea.

Bibliografía

[1] A. L. Creo and W. F. Schwenk, "Bone Age: A Handy Tool for Pediatric Providers," *Pediatrics*, vol. 140, no. 6, Dec. 2017, doi: 10.1542/peds.2017-1486.

[2] D. D. Martin *et al.*, "The use of bone age in clinical practice - Part 1," Jul. 2011. doi: 10.1159/000329372.

[3] W. Walter. Greulich and S. Idell. Pyle, *Radiographic atlas of skeletal development of the hand and wrist*. California: Stanford University Press, 1959.

J. Tanner, R. Whitehouse, N. Cameron, W. Marshall, M. Healy, and H. Goldstein, *Assessment of skeletal maturity and prediction of adult height (TW2 method)*. London: Academic Press, 1983.

M. A. Breen, A. Tsai, A. Stamm, and P. K. Kleinman, "Bone age assessment practices in infants and older children among Society for Pediatric Radiology members," *Pediatr Radiol*, vol. 46, no. 9, pp. 1269–1274, Aug. 2016, doi: 10.1007/s00247-016-3618-7.

R. R. Van Rijn and H. H. Thodberg, "Bone age assessment: Automated techniques coming of age?," Nov. 2013. doi: 10.1258/ar.2012.120443.

R. K. Bull, P. D. Edwards, P. M. Kemp, S. Fry, and I. A. Hughes, "Bone age assessment: A large scale comparison of the Greulich and Pyle, and Tanner and Whitehouse (TW2) methods," *Arch Dis Child*, vol. 81, no. 2, pp. 172–173, 1999, doi: 10.1136/adc.81.2.172.

G. Frank Johnson, J. P. Dorst, J. P. Kuhn, A. F. Roche, and G. H. Davila, "Reliability os skeletal age assesments," Jun. 1973. [Online]. Available: www.ajronline.org

A. F. Roche, C. G. Rohmann, N. Y. French, and G. H. Davila, "Effect of training on replicability of assessments of skeletal maturity (Greulich-Pyle)." [Online]. Available: www.ajronline.org

H. H. Thodberg and L. Sävendahl, "Validation and reference values of automated bone age determination for four ethnicities," *Acad Radiol*, vol. 17, no. 11, pp. 1425–1432, Nov. 2010, doi: 10.1016/j.acra.2010.06.007.

A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", Accessed: Oct. 28, 2024. [Online]. Available: http://code.google.com/p/cuda-convnet/

D. B. Larson, M. C. Chen, M. P. Lungren, S. S. Halabi, N. V. Stence, and C. P. Langlotz, "Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs," *Radiology*, vol. 287, no. 1, pp. 313–322, Apr. 2018, doi: 10.1148/radiol.2017170236.

E. L. Siegel, "What can we learn from the RSNA pediatric bone age machine learning challenge?," Jan. 01, 2019, *Radiological Society of North America Inc.* doi: 10.1148/radiol.2018182657.

S. S. Halabi *et al.*, "The rSNA pediatric bone age machine learning challenge," *Radiology*, vol. 290, no. 3, pp. 498–503, Mar. 2019, doi: 10.1148/radiol.2018180736.

C. Spampinato, S. Palazzo, D. Giordano, M. Aldinucci, and R. Leonardi, "Deep learning for automated skeletal bone age assessment in X-ray images," *Med Image Anal*, vol. 36, pp. 41–51, Feb. 2017, doi: 10.1016/j.media.2016.10.010.

H. Lee *et al.*, "Fully Automated Deep Learning System for Bone Age Assessment," *J Digit Imaging*, vol. 30, no. 4, pp. 427–441, Aug. 2017, doi: 10.1007/s10278-017-9955-8.

[17] S. H. Tajmir *et al.*, "Artificial intelligence-assisted interpretation of bone age radiographs improves accuracy and decreases variability," *Skeletal Radiol*, vol. 48, no. 2, pp. 275–283, Feb. 2019, doi: 10.1007/s00256-018-3033-2.

[18] S. J. Son *et al.*, "TW3-Based Fully Automated Bone Age Assessment System Using Deep Neural Networks," *IEEE Access*, vol. 7, pp. 33346–33358, 2019, doi: 10.1109/ACCESS.2019.2903131.

[19] B.-D. Lee and M. S. Lee, "Automated Bone Age Assessment Using Artificial Intelligence: The Future of Bone Age Assessment," *Korean J Radiol*, vol. 22, no. 5, p. 792, 2021, doi: 10.3348/kjr.2020.0941.

[20] A. L. Dallora, P. Anderberg, O. Kvist, E. Mendes, S. Diaz Ruiz, and J. Sanmartin Berglund, "Bone age assessment with various machine learning techniques: A systematic literature review and meta-analysis," *PLoS One*, vol. 14, no. 7, p. e0220242, Jul. 2019, doi: 10.1371/journal.pone.0220242.

[21] H. H. Thodberg, S. Kreiborg, A. Juul, and K. D. Pedersen, "The BoneXpert method for automated determination of skeletal maturity," *IEEE Trans Med Imaging*, vol. 28, no. 1, pp. 52–66, Jan. 2009, doi: 10.1109/TMI.2008.926067.

[22] H. A. Sturges, "The Choice of a Class Interval," *J Am Stat Assoc*, vol. 21, no. 153, pp. 65–66, 1926, doi: 10.1080/01621459.1926.10502161.

[23] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011, Accessed: Oct. 28, 2024. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html

[24] F. Chollet and others, "Keras," 2015.

[25] Martín~Abadi *et al.*, " TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," 2015. [Online]. Available: https://www.tensorflow.org/

[26] W. McKinney, "Data Structures for Statistical Computing in Python," *Proceedings of the 9th Python in Science Conference*, pp. 56–61, 2010, doi: 10.25080/MAJORA-92BF1922-00A.

[27] C. R. Harris *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020, doi: 10.1038/s41586-020-2649-2.

[28] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Comput Sci Eng*, vol. 9, no. 3, pp. 90–95, 2007, doi: 10.1109/MCSE.2007.55.

[29] P. Virtanen *et al.*, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nature Methods 2020 17:3*, vol. 17, no. 3, pp. 261–272, Feb. 2020, doi: 10.1038/s41592-019-0686-2.

[30] C. O. da Costa-Luis, "`tqdm`: A Fast, Extensible Progress Meter for Python and CLI," *J Open Source Softw*, vol. 4, no. 37, p. 1277, May 2019, doi: 10.21105/JOSS.01277.

[31] A. Clark, "Pillow (PIL Fork) Documentation," 2015, *readthedocs*. [Online]. Available: https://buildmedia.readthedocs.org/media/pdf/pillow/latest/pillow.pdf

K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, Sep. 2014, Accessed: Oct. 28, 2024. [Online]. Available: https://arxiv.org/abs/1409.1556v6

K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 770–778, Dec. 2015, doi: 10.1109/CVPR.2016.90.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 2818–2826, Dec. 2015, doi: 10.1109/CVPR.2016.308.

M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *36th International Conference on Machine Learning, ICML 2019*, vol. 2019-June, pp. 10691–10700, May 2019, Accessed: Oct. 28, 2024. [Online]. Available: https://arxiv.org/abs/1905.11946v5

F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 1800–1807, Oct. 2016, doi: 10.1109/CVPR.2017.195.

M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, Jan. 2018, doi: 10.1109/CVPR.2018.00474.

G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 2261–2269, Aug. 2016, doi: 10.1109/CVPR.2017.243.

B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning Transferable Architectures for Scalable Image Recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 8697–8710, Jul. 2017, doi: 10.1109/CVPR.2018.00907.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014, Accessed: Oct. 28, 2024. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *32nd International Conference on Machine Learning, ICML 2015*, vol. 1, pp. 448–456, Feb. 2015, Accessed: Oct. 28, 2024. [Online]. Available: https://arxiv.org/abs/1502.03167v3

D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization," *3rd International Conference on Learning Representations, ICLR*

2015 - Conference Track Proceedings, Dec. 2014, Accessed: Oct. 28, 2024. [Online]. Available: https://arxiv.org/abs/1412.6980v9

[43] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," *2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*, pp. 248–255, 2009, doi: 10.1109/CVPR.2009.5206848.