Universidad de Lima

Facultad de Ingeniería y Arquitectura

Carrera de Ingeniería de Sistemas

PREDICCIÓN DE EDAD ÓSEA CON RED BASADA EN VGG-16 Y TRANSFER LEARNING

Tesis para optar el Título Profesional de Ingeniero de Sistemas

Alvaro Daniel Quispe Carbonel

CÓDIGO 20151099

ASESOR

José Antonio Taquía Gutiérrez

Lima – Perú

Julio de 2020

Predicción de edad ósea con red basada en VGG-16 y transfer learning

Alvaro Daniel Quispe Carbonel
20151099@aloe.ulima.edu.pe
Universidad de Lima

Resumen: Para el diagnóstico de enfermedades hereditarias y desórdenes endocrinos en niños y jóvenes suele utilizarse el cálculo de la edad ósea a través de una evaluación visual de una radiografía de la muñeca y de la mano. La presente investigación se ha centrado en el desarrollo de un modelo de deep learning para la predicción de la edad ósea utilizando 12611 imágenes radiológicas de niños y jóvenes de 0 a 18 años de la Sociedad de Radiología de Norte América (RSNA). Para lograrlo se realizaron tres procesos: preprocesamiento mediante una red neuronal convolucional U-Net para la generación de máscaras que permitan eliminar el fondo de las imágenes y su posterior ecualización; el desarrollo de una de red neuronal convolucional basada en VGG-16, transfer learning, cuyos pesos fueron obtenidos de ImageNet y un mecanismo de atención con la que se entrenaron 6 modelos con imágenes preprocesadas y originales; y una comparación entre la predicción de 230 imágenes locales, con la edad ósea determinada por médicos para dicho subconjunto. Con la validación del dataset de la RSNA, se obtuvo un MAE en meses promedio de 9.4 para el modelo propuesto de hombres. Por el lado de la validación local, se alcanzó un MAE en meses promedio de 13.7 en hombres contando todos los grupos de edad y se alcanzó un MAE en meses de 12.4 en mujeres de 2 a 13 años.

Palabras Clave: Deep Learning, Determinación de edad ósea, U-Net-, VGG-16, Visión computacional

Abstract: Bone age assessment through a visual evaluation of an X-ray of the wrist and hand is often used for the diagnosis of hereditary diseases and endocrine disorders in children and young people. This research is focused on the development of a deep learning model for bone age prediction using 12611 radiological images of children and young people aged 0-18 from the Radiological Society of North America (RSNA). To achieve this, three processes were carried out: preprocessing through a convolutional U-Net neuronal network for the generation of masks that eliminate the background of the images and their subsequent equalization; the development of a convolutional neuronal network based on VGG-16, transfer learning with weights obtained from ImageNet and an Attention Mechanism, which trained 6 models with preprocessed and original images; and a comparison between the prediction of 230 local images, with the bone age determined by doctors for this subset. With the validation of the RSNA dataset, an average MAE of 9.4 months was obtained for the proposed male model. The validation using the local images reached an average MAE of 13.7 months in men counting all age groups and a MAE of 12.4 months was obtained in women from 2 to 13 years old.

Keywords: Bone Age Assessment, Computer Vision, Deep Learning, U-Net, VGG-16

1. INTRODUCCIÓN

El ritmo de crecimiento depende de cada persona, de tal manera que la edad cronológica, basada en años, puede ser diferente a la maduración del esqueleto, que está controlada por hormonas endocrinas (Pose Lepe, Villacrés, Silva Fuente-Alba, y Guiloff, 2018). Este, es un indicador de desarrollo fisiológico utilizado en el diagnóstico de enfermedades hereditarias y desórdenes endocrinos (Giordano, Spampinato, Scarciofalo, y Leonardi, 2010). Si no se identifican las razones de una maduración esquelética tardía en un niño, es probable que también tenga una pubertad tardía (Satoh, 2015). De acuerdo con la Organización Mundial de la Salud (2014), el retraso del crecimiento afecta a unos 162 millones de niños menores de 5 años y por lo tanto es uno de los obstáculos más significativos para el desarrollo humano como establecen en sus metas mundiales al 2025.

Durante décadas, la determinación de la edad ósea se ha basado en la evaluación visual de una radiografía de la mano y la muñeca, mediante la aplicación de los métodos como Tanner-Whitehouse (TW) o Greulich Pyle (GP). Sin embargo, las estimaciones manuales por los métodos mencionados anteriormente tienen cierto grado de variabilidad entre los evaluadores (Manzoor, Hassan, y Ahmed, 2014). De hecho, en medicina humana entre el 2% al 30% de los informes radiológicos pueden tener errores (Elizalde, 2012).

Con la introducción del uso de imágenes digitales, se han realizado múltiples intentos para desarrollar técnicas de procesamiento de imágenes que extraigan automáticamente las características clave de la osificación en los huesos para proporcionar un enfoque más eficaz y objetivo a las evaluaciones de la madurez esquelética (Gilsanz y Raitb, 2005). Entre estas, la más relevante debido a su uso comercial es el sistema BoneXpert, desarrollado por Thodberg (2009). Sin embargo, su elevado precio ha restringido el uso e impedido la validación en el entorno local. Debido a esto, el presente trabajo propone un método propio basado en *Deep Learning* para determinar la edad ósea a partir de una imagen radiológica de la mano de niños y jóvenes de 0 a 18 años que será validado en el Perú. La investigación consistió en tres procesos: preprocesamiento, entrenamiento de 6 modelos a partir de una red neuronal convolucional y la validación. Las imágenes utilizadas en el preprocesamiento y entrenamiento pertenecen al reto de edad ósea de la Sociedad de Radiología de Norte América (RSNA Pediatric Bone Age Challenge, 2017). Para el primer proceso se entrenó una red neuronal convolucional U-Net para la generación de máscaras que permitan eliminar el fondo de las imágenes. A partir de estas, se aplicó una ecualización por histogramas para realzar los huesos y se descartaron manualmente aquellas imágenes que no mostraban los huesos completos de las manos. El segundo proceso fue desarrollado a partir de una red VGG-16 con los pesos obtenidos de ImageNet. Una vez desarrollada la arquitectura se entrenaron 6 modelos, 3 con las imágenes preprocesadas y 3 con las imágenes originales. Asimismo, estos fueron divididos en 3 categorías: hombres, mujeres y mixto (ambos géneros). A partir de los modelos con mejores resultados, el último proceso consistió en la validación de estos con 230 imágenes radiológicas anonimizadas de la mano de niños y jóvenes de 0 a 18 años, provistas por un médico radiólogo. Estas imágenes locales, de acceso restringido, fueron conseguidas mediante un consentimiento informado. Para lo mencionado, esta investigación se basó en las 6 categorías de Gilsanz y Raitb (2005) sobre el crecimiento y se formaron 3 grupos.

A continuación, se hará una revisión del estado del arte. Luego, se desarrollará el marco teórico, que explicará conceptos sobre la edad ósea, así como las redes neuronales utilizadas para realizar las predicciones. Después, se desarrollará la experimentación a partir de la metodología propuesta. Finalmente, se discutirán los resultados obtenidos y finalmente se establecerán las conclusiones de la presente investigación.

2. ESTADO DEL ARTE

La automatización de la edad ósea ha sido una de las áreas de mejora en la radiología de los últimos años. Desde comparaciones a partir de histogramas, tal como hizo Mansourvar et al (2012), hasta la aplicación de sistemas complejos utilizando *Deep Learning* en más de una etapa del desarrollo como Iglovikov, Rakhlin, Kalinin, y Shvets (2018). A continuación se hará una revisión de diferentes propuestas a través de los años.

2.1 Enfoque en regiones de interés de epífisis y metáfisis (EMROI) y carpo (CROI)

Rucci et al (1995) plantearon un sistema basado en redes neuronales artificiales para determinar la edad ósea con el método Tanner-Whitehouse 2 (TW2) utilizando solo los huesos del carpo (CROI). Para ubicar cada hueso utilizaron un módulo de atención y después un clasificador para identificar cada hueso. De esta manera, demostraron que una red neuronal puede ser una herramienta relevante para determinar la edad ósea con el método Tanner-Whitehouse 2 (TW2). Por otro lado, Sato et al (1999) se concentraron en la epífisis y metáfisis (EMROI) al aplicar un algoritmo para extraer la tercera falange distal, media y proximal. Esto con el fin de obtener las distancias en las EMROI y realizar una regresión múltiple para obtener la edad ósea.

Por otro lado, Mahmoodi, Sharif, Graeme Chester, Owen, y Lee (2000) propusieron un sistema que aplicó una búsqueda jerárquica para focalizar los huesos y luego un modelo de forma activa mediante un contorno óseo. El sistema extrajo tres características en donde existía una relación razonable entre la epífisis-metafisis y la edad cronológica. El sistema fue evaluado con una técnica de *leave one out*. Los investigadores presentaron la precisión del sistema con 82% para los pacientes masculinos y 84% para los femeninos.

En el estudio de Pietka et al (2001), propusieron una metodología para poder realizar la extracción de las EMROI para la determinación de la edad ósea. Para esto, se realizaron tres etapas: detección de la punta de la falange, extracción de las EMROI, y localización de sus diámetros y borde inferior. La investigación demostró que a partir de las características extraidas se podia determinar la edad ósea de manera más objetiva que observando toda la imagen.

Por otra parte, Bocchi, Ferrara, Nicoletti, y Valli (2003) se enfocaron tanto en EMROI como en CROI. Se aplicó un filtro de Gabor y diferencia Gaussiana respectivamente para resaltar los huesos. Después, se entrenaron 20 redes neuronales con backpropagation para clasificar cada hueso y finalmente se calculó la edad ósea a partir del método TW2.

En el mismo año, Niemeijer, van Ginneken, Maas, Beek, y Viergever (2003) desarrollaron un sistema automatizado basado en el método TW2 se enfocaron en la tercera falange utilizando un modelo de forma activa. El modelo empleó la descripción de los modos de variación y una matriz de covarianza para la medición estadística. La función de

emparejamiento fue ejecutada por la relación más alta entre la escala de píxeles para cada ROI del hueso y la escala de píxeles de las imágenes de la muestra, la precisión del sistema investigado para categorizar la edad ósea fue de 73% a 80% comparada con la de un observador.

Siete años más tarde, Giordano et al (2010) presentaron un sistema automático según el método Tanner-Whitehouse (TW2), basado en la integración de dos sistemas, el EMROI y el CROI. El sistema asegura una evaluación precisa de la edad ósea para rangos de edad de 0 a 10 años para los hombres y de 0 a 7 años para las mujeres. Para ambos enfoques se propusieron nuevas técnicas de segmentación. La extracción de las regiones de interés se llevó a cabo combinando una extensión de la técnica de contornos activos y la diferencia derivada de un filtro gaussiano. Se obtuvo una media de error de 0.46 con una desviación estándar de 0.37 años.

En el mismo sentido, Somkantha et al (2011) probaron la eficiencia de la extracción de límites y la evaluación de la edad ósea diseñando una técnica de seguimiento de bordes del carpo. A partir de los huesos delimitados, se determinó la edad ósea en niños pequeños de 0 a 6 años con una regresión con el método de Máquinas de Soporte de Vectores y una regresión de la red neuronal que aplicaba el algoritmo *backpropagation*. Los resultados mostraron que la regresión del vector de soporte tenía mejor eficiencia que la regresión de la red neuronal. Spampinato, Palazzo, Giordano, Aldinucci y Leonardi (2017) obtuvieron resultados similares, y produjo resultados cercanos a los de los radiólogos expertos.

2.2 Enfoque en RUS

Tristán-Vega y Arribas (2008) propusieron un sistema para determinar la edad ósea con el método Tanner-WhiteHouse 3 (TW3) basado en el radio, cúbito y huesos cortos (RUS). Este sistema fue asistido con contornos manuales y luego se aplicó una técnica de agrupamiento adaptativo para la segmentación. Asimismo, se aplicó una red neuronal perceptrón en el estado de decisión para hacer probabilidades a posteriori que predigan la tasa de error. Si bien la diferencia de la media entre el sistema y las observaciones por radiólogos es amplia, los investigadores afirmaron que su método podría ampliarse mejorando la segmentación ósea y que sería valioso seguir ahondando en las redes neuronales.

Por otro lado, Liu, Qi, Liu, Ning, y Luo (2008) entrenaron 2 redes neuronales con el algoritmo backpropagation. Una basada en RUS y otra en el carpo (CROI). Para segmentar los huesos de imágenes utilizaron algoritmos de enjambre de partículas. Los investigadores determinaron que no hubo una diferencia significativa entre los modelos ni la evaluación manual.

Once años después, aparte de la localización del RUS, Son et al (2019) introdujeron el concepto de bordes de región de interés (bROIs) para reducir el área que una red neuronal convolucional debe explorar para extraer un ROI. De esta manera, aplicaron diferentes técnicas para cada tipo de bROI como binarización, rotación o cascos convexos. Después, entrenaron una red Faster R-CNN para extraer los ROIs. Finalmente se entrenó una red VGG para clasificar cada hueso con el fin de compararlo con la tabla de madurez del método TW3 y obtener la edad ósea.

2.3 Sistemas comerciales

Según lo citado por Thodberg (2009), BoneXpert, que es un software comercial, se compone de tres capas. La capa A determina los bordes de los huesos utilizando un modelo de apariencia activa, que permite coincidir los bordes predeterminados con los ingresados. La capa B determina la edad ósea basada de cada hueso. Finalmente, la capa C transforma las edades a fases de TW o calcula el promedio de los huesos requeridos para el método GP. Pose Lepe et al (2018) indicaron que los niveles de correlación entre la determinación manual y con este sistema se encontró en un rango de 0.91 y 0.93. Sin embargo, este no puede aceptar imágenes de mala calidad, así como solo se basa en 15 huesos en donde se deja de lado distintos huesos cortos y del carpo. Caso contrario a Kim et al (2017), quienes a partir de un sistema utilizando *Deep Learning* desarrollaron VUNO Med-Bone Age que también es un sistema de predicción de edad ósea, esta vez solo orientado al método GP. A diferencia de otras propuestas, el resultado final son tres imágenes de referencia más probables marcadas por la edad según la radiografía del paciente, es decir, las tres edades óseas más probables. Se obtuvo una precisión de 96.3% con respecto a la edad ósea referencial. De igual manera, se determinó que el tiempo de lectura de los radiólogos se redujo en un 29% tras usar este sistema como apoyo.

2.4 Uso de modelos preentrenados

Lee et al (2017) propusieron un sistema compuesto de dos componentes principales: un motor de preprocesamiento y un clasificador. El motor de preprocesamiento divide toda la imagen de rayos X en muestras y la clase de cada una es determinada por una red neuronal convolucional. Utilizando los resultados de la clasificación de las muestras de imagen, se creó una máscara para la mano y la muñeca. Para cada imagen de rayos X preprocesada, el clasificador estima la edad ósea utilizando una red neuronal convolucional basada en un modelo GoogLeNet preentrenado.

Un enfoque similar es el de Chen (2017), que empleó un modelo preentrenado, como VGGNet, para la inicialización del peso y la sintonización de todas las capas para predecir la edad ósea. De acuerdo con su análisis, se propuso una función de pérdida con regularización L2 para medir la diferencia entre la edad ósea determinada automáticamente y la determinada por un especialista. Aunque la precisión de la predicción general de la solución propuesta no es notable, se comentó que este enfoque puede lograr resultados competentes que se aproximen al desempeño de un radiólogo. Iglovikov, Rakhlin, Kalinin, y Shvets (2018), tras eliminar el fondo de las imágenes con una U-Net, plantearon 18 modelos para evaluar diferentes regiones de la mano y su impacto en la predicción de la edad ósea. Estos consistieron en una combinación entre el tipo (regresión y clasificación), el género (masculino, femenino o mixto) y la región de la mano (mano completa, carpo, metacarpo y falanges). Lograron determinar que, si bien la evaluación ósea puede ser realizada con ciertas regiones de la mano, estas presentan un aumento de error de un 10% al 15% contra la evaluación de la mano completa.

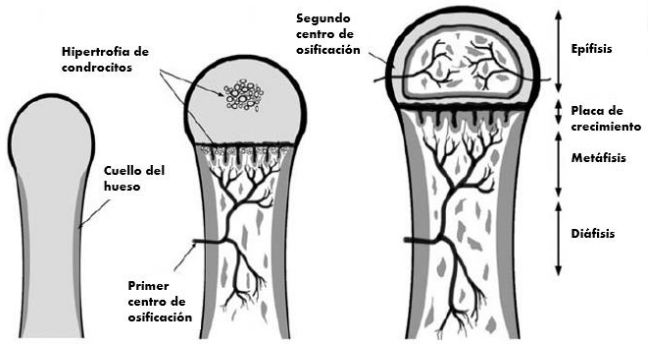
Adicionalmente, Ren et al (2018), utilizaron un módulo de atención y una regresión a partir de una red Inception V3. Para el módulo de atención se aplicó una atención “gruesa” y “fina”. La primera sirvió para localizar la mano y eliminar el fondo innecesario y la segunda para realzar los huesos de la mano. De esta manera, se entrenó la red mencionada con el conjunto de datos de la RSNA y a partir de parámetros obtenido, se aplicó *transfer learning* para determinar la edad ósea otro en un segundo conjunto de datos privado. Se obtuvo un MAE de 5.2 con las imágenes de la RSNA y un 5.3 con las del conjunto de datos privado. Asimismo, se evaluó que el impacto en la determinación de la edad ósea fue consistentemente mayor en casos en los que se utilizaron pesos transferidos a pesos aleatorios para inicializar el entrenamiento de la red propuesta.

3. MARCO TEÓRICO

3.1 Desarrollo de los huesos

La madurez del esqueleto es una medida de desarrollo, que toma en cuenta el tamaño, la forma, y el grado de osificación del hueso. Esta se denota con la calcificación inicial, que comienza en el centro de los huesos largos en una región llamada centro de osificación primaria(diáfisis) y que a pesar de que existen algunos huesos, como los carpianos, que se osifican a partir de este, el resto desarrolla centros de osificación secundaria (epífisis) y se encuentran en el cartílago de las extremidades del hueso (Gilsanz y Raitb, 2005). En la figura 3.1 se pueden ver 3 etapas de la maduración de los huesos de izquierda a derecha: sin un centro de osificación, con diáfisis y, con diáfisis y epífisis. Conforme avanza el proceso de osificación, el hueso reemplaza el cartílago dejando al final una delgada capa de este.

Figura 3.1
Proceso de osificación de los huesos



Nota. Adaptado de Hand Bone Age por Gilsanz, V. y Raitb, O, 2005, Springer.
(http://www.chospab.es/biblioteca/DOCUMENTOS/Atlas_of_Hand_Bone_Age.pdf)

3.2 Métodos para la determinación de la edad ósea

Son et al (2019) indica que no existe un método estándar para la determinación de la edad ósea y que los métodos más utilizados en la práctica clínica son el el Greulich-Pyle (GP) y Tanner-Whitehouse (TW).

Greulich Pyle (GP)

Consiste en un atlas, que contiene una serie de imágenes radiológicas base que indican las etapas de crecimiento de los niños con diferentes niveles de madurez ósea. La imagen de rayos X de la mano no dominante de un paciente, por lo general la izquierda, es comparada con las imágenes del atlas y se elige aquella en la que exista una mayor similitud para determinar la edad ósea del paciente. El mayor inconveniente de este método es que si bien los huesos de la mano pueden considerarse como una sola entidad, pueden haber huesos con diferentes estados de madurez como describe Rucci et al (1995). Sin embargo, de acuerdo con Giordano et al (2010) este método es el más utilizado debido a su facilidad de uso.

Tanner-Whitehouse (TW)

Este método reemplaza la búsqueda de similitud entre imágenes como un todo. (Tristán-Vega y Arribas, 2008) indican que a partir de la primera, tercera y quinta falange además del radio, la ulna y huesos cortos (RUS), se determina la madurez ósea comparando cada uno con la imagen correspondiente en la tabla de conversión, que indica la forma del hueso junto con un puntaje. Una vez que se tengan los puntajes por hueso, estos son sumados para obtener la edad ósea. Navarro, Tejedor, y López Siguero (2014) comentan que este método tiene hasta 3 versiones, siendo TW3 la última. Rucci et al (1995) adicionan que aparte de la versión clásica, previamente explicada, se pueden usar diferentes tablas de conversión para estimar la edad ósea ya sea a partir solamente del carpo (CROI) o RUS.

3.3 Categorías para determinar edad ósea

Gilsanz y Raitb (2005) dividieron el desarrollo del esqueleto en 6 categorías por género: infancia, niñez, pre-puertad, pubertad temprana y tardía, pubertad tardía y post pubertad. La principal diferencia entre géneros es que se muestra una madurez del esqueleto más temprana para las mujeres que para los hombres, que se muestra en la tabla 3.1. Esta clasificación permite agrupar las fases de desarrollo por edad y género.

Tabla 3.1

Categorías de desarrollo del esqueleto

Categorías	Infancia	Niñez	Pre-pubertad	Pubertad temprana y media	Pubertad tardía	Post- pubertad
Mujeres	Hasta 10 meses	Hasta 2 años	Hasta 7 años	Hasta 13 años	Hasta 15 años	Hasta 17 años
Hombres	Hasta 14 meses	Hasta 3 años	Hasta 9 años	Hasta 14 años	Hasta 16 años	Hasta 19 años

Nota. Obtenido de Hand Bone Age por Gilsanz, V. y Raitb, O, 2005, Springer. (http://www.chospab.es/biblioteca/DOCUMENTOS/Atlas_of_Hand_Bone_Age.pdf)

3.4 Redes neuronales convolucionales

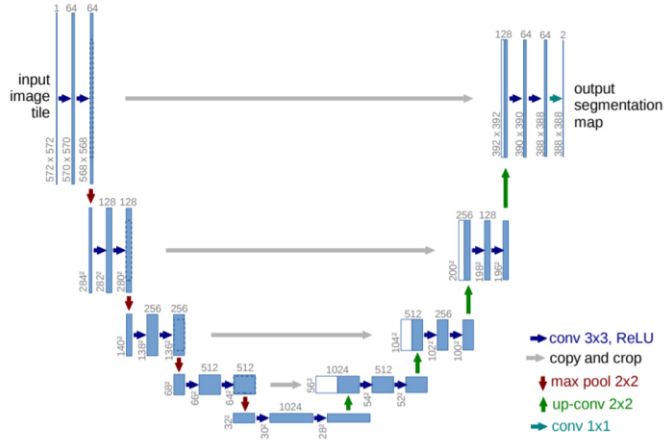
U-Net

Ronneberger, Fischer, y Brox (2015) propusieron esta red neuronal convolucional para casos de segmentación y ha probado tener éxito en el sector biomédico. La arquitectura de esta red, de 23 capas convolucionales, mostrada en la figura 3.2, consiste de dos etapas: de contracción (izquierda) y de expansión (derecha). La primera aplica convoluciones a la imagen de entrada, comprimiendo la información y detectando las diferentes características en la imagen. Esta compuesta de 2 convoluciones 3x3 en varias repeticiones, cada una seguida de una función de activación *ReLU* y una operación de *max pooling* de 2x2. En cada repetición se duplica el número de características. La segunda se encarga de recuperar el tamaño original con el uso de lo aprendido en la etapa anterior. Esta compuesta de un muestreo ascendente del mapa de características seguido de una convolución 2x2 ("convolución ascendente") que reduce a la mitad el número de características. Asi como su contraparte, contiene 2 convoluciones 3x3 en varias

repeticiones, cada una seguida de una función de activación *ReLU*. En la capa final se utiliza una convolución 1x1 para ubicar cada vector de característica al número deseado de clases.

Asimismo, esta red se caracteriza por no necesitar una gran cantidad de imágenes para conseguir buenos resultados. Esto se debe a la aplicación de *data augmentation* en la cuál se aplican diferentes rotaciones a las imágenes originales.

Figura 3.2
Arquitectura de U-NET

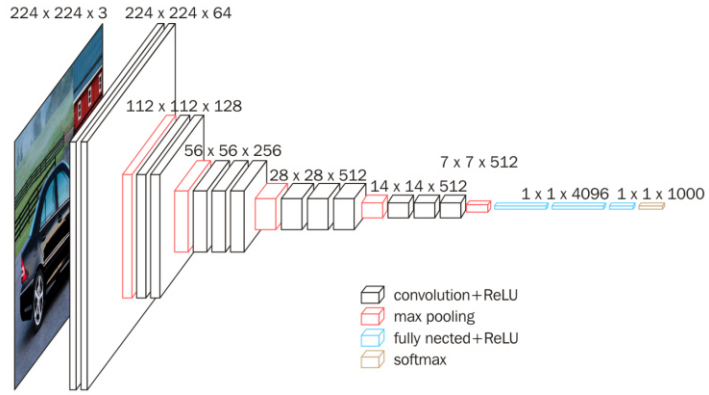


Nota. Obtenido de U-net: Convolutional networks for biomedical image segmentation, por Ronneberger, O., Fischer, P. y Brox, T, 2015, Lecture Notes in Computer Science (https://doi.org/10.1007/978-3-319-24574-4_28)

VGG-16

Simonyan y Zisserman (2014), desarrollaron la red neuronal convolucional, Oxford Visual Geometry Group (VGG-16). Como se puede ver en la figura 3.3, esta arquitectura esta constituida de 13 capas convolucionales. Cada grupo es seguido de una capa de *max pooling*. A estas le siguen 3 capas totalmente conectadas, de ahí su nombre incluye 16, y finalmente un clasificador *softmax*. Esta red se caracteriza por contar con varios parámetros para entrenar, lo que puede hacer que su entrenamiento dure más que el resto de las redes. Asimismo, tiene la ventaja de contar con pesos preentrenados disponibles del dataset de Imagenet, que podrían ser utilizados para inicializar el entrenamiento de la red propuesta como comenta.

Figura 3.3
Arquitectura de VGG-16



Nota. Obtenido de A review of deep learning in the study of materials degradation, por Nash, W., Drummond, T. y Birbilis, N, 2018, Materials Degradation (https://doi.org/10.1038/s41529-018-0058-x)

3.5 Transfer Learning

Ren et al (2018) comentan que esta técnica busca inicializar los parámetros de una red a partir de otra con la misma arquitectura. Esto con el fin de otorgar una mejor capacidad de reconocimiento y reducir el tiempo de entrenamiento. Un aspecto a tener en cuenta es la disponibilidad de pesos preentrenados para una arquitectura de red determinada.

Castillo, Tong, Zhao, y Zhu (2017) comentan que la aplicación de transfer learning suele ser de estas 3 maneras:

1. Al reusar la arquitectura de un modelo y usar pesos preentrenados como un punto de inicio y después entrenar la red entera.
2. Utilizar la misma arquitectura del modelo. Después aplicar pesos preentrenados en las primeras capas y congelar estas capas durante el entrenamiento, de tal manera que solo las últimas capas sean entrenadas. Esto tiene como fundamento que se considera que las primeras capas buscan en su mayoría extraer características, mientras que las últimas estan asociadas a una tarea específica.
3. Finalmente, también se aplica esta técnica cuando se cargan los pesos, se congelan las primeras capas de una red y reemplazan las últimas capas para armar una nueva arquitectura para un problema diferente al planteado por el modelo del cual se extrayeron los pesos.

3.6 Atención en redes neuronales convolucionales

Jetley, Lord, Lee, y Torr (2018) comentaron que las redes neuronales convolucionales han demostrado resultados impresionantes en una variedad de tareas de reconocimiento de imagen. Sin embargo, muchas veces no se llega a entender el por qué en algunos casos se falla y en algunos casos no. Por lo tanto, es necesario tener en cuenta 2 esquemas sobre la atención en características en redes neuronales convolucionales.

Atención entrenada

Consiste en ayudar al modelo a tener en cuenta características importantes en la imagen. Sea x un vector de características, a un vector de atención(máscara), $f(x)$ un mecanismo de atención y en donde $\odot$ es multiplicación por elemento, la atención podría verse de la siguiente manera

$$\begin{aligned} a &= f(x) \\ g &= a \odot x \end{aligned} \tag{1}$$

Este tipo de atención a su vez puede ser dividida en dos categorías, suave y dura. La atención suave se refiere cuando las características son multiplicadas por una máscara de valores a entre 0 y 1. Mientras, la atención dura corresponde a una multiplicación de 0 o 1.

Atención post hoc

Consiste en un grupo de técnicas que ayudan a visualizar qué partes son consideradas importantes por un modelo. Estas no tienen incidencia alguna en la forma en que el modelo aprende. Una de las técnicas utilizadas es el mapa de atención, mostrado en la figura 3.4, que según Simonyan, Vedaldi, y Zisserman (2014) es una matriz escalar que representa la importancia relativa de las activaciones de capas en diferentes ubicaciones espaciales con respecto a la tarea objetivo.

Figura 3.4
Imagen original con su mapa de atención.

Nota. Obtenido de Learn to pay attention, por Jetley, Lord, Lee, y Torr, 2018 (https://arxiv.org/pdf/1804.02391.pdf)

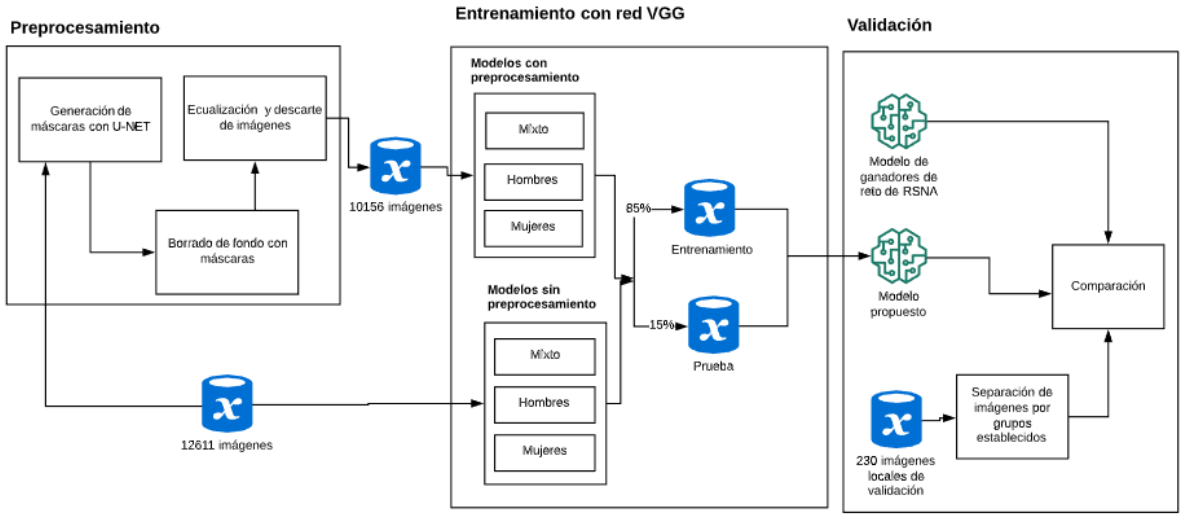
3.7 Capa localmente conectada

Keras (2015) comenta que una capa localmente conectada es como una capa convolucional, a excepción de no tener pesos compartidos. Es así como cada píxel tiene un filtro propio. A pesar de aumentar el número de parámetros, también permite que una red aprenda de diferentes tipos de características en diferentes regiones. Tanto Taigman, Ranzato, Aviv, y Park (2014) como Chen et al (2015) obtuvieron buenos resultados al aplicar este tipo de capa.

4. METODOLOGÍA

La investigación consistió en tres procesos: preprocesamiento, entrenamiento de 6 modelos a partir de una red neuronal convolucional y la validación, que se muestran en la figura 4.1.

Figura 4.1
Metodología propuesta



Para el primer proceso se entrenó una red neuronal convolucional U-Net para la generación de máscaras que permitan eliminar el fondo de las imágenes. A partir de estas, se aplicó una ecualización por histogramas y se descartaron manualmente, a partir de una evaluación visual, aquellas imágenes que no mostraban los huesos completos de las manos. La segunda etapa fue desarrollada a partir de una red VGG-16 con los pesos obtenidos de ImageNet. Una vez desarrollada la arquitectura se entrenaron 6 modelos, como se muestra en la tabla 4.1, 3 con las imágenes preprocesadas y 3 con las imágenes originales. Asimismo, estos fueron divididos en 3 categorías: hombres, mujeres y mixto (ambos géneros). Esto significa que cada tipo de modelo tuvo un entrenamiento con las imágenes con y sin preprocesamiento.

Tabla 4.1
Detalle de 6 modelos entrenados

Tipo de modelo	Tipo de imagen	
Mixto(Ambos géneros)	Con preprocesamiento	Sin preprocesamiento
Hombres	Con preprocesamiento	Sin preprocesamiento
Mujeres	Con preprocesamiento	Sin preprocesamiento

A partir de los modelos con mejores resultados, el último proceso consistió en la validación de estos con 230 imágenes anonimizadas proporcionadas por un médico radiólogo que pertenecían a niños y jóvenes de 0 a 18 años. Para lo mencionado, esta investigación se basó en las 6 categorías de Gilsanz y Raitb (2005) sobre el crecimiento y se formaron 3 grupos. Cada uno incluye 2 categorías. De esta manera, como se puede ver en la tabla 4.2, el primer grupo incluye a niños de hasta 3 años y el de niñas de hasta 2 años. El segundo hasta 14 años para hombres y 13 años para mujeres y el tercer grupo incluye a hombres hasta 19 años y mujeres de hasta 17 años.

Tabla 4.2
Grupos formados para validación

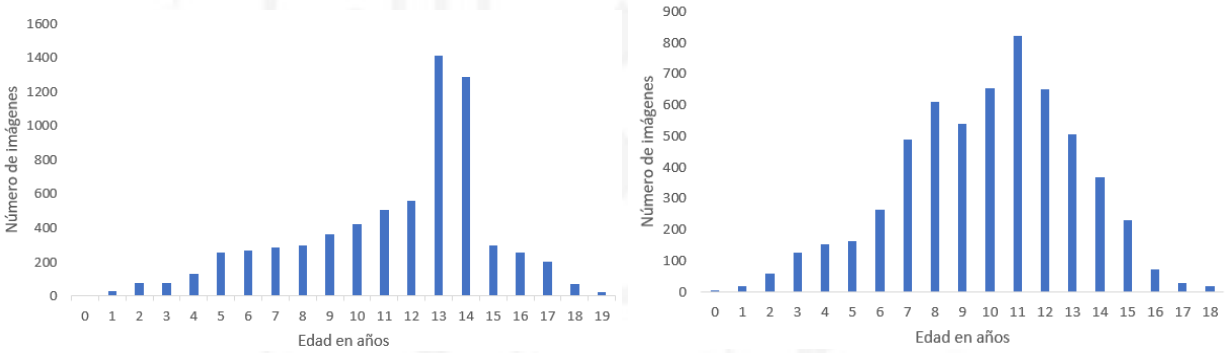
Grupo	Edad límite
Hombres Grupo 1	Hasta 3 años
Hombres Grupo 2	Hasta 14 años
Hombres Grupo 3	Hasta 19 años
Mujeres Grupo 1	Hasta 2 años
Mujeres Grupo 2	Hasta 13 años
Mujeres Grupo 3	Hasta 17 años

4.1 Experimentación

Datos

Se utilizaron 12611 imágenes radiológicas (en formato .png, tipo RGB y tamaño promedio de 1,694 × 1,284) de la mano de jóvenes entre 0 a 18 años. La distribución de géneros corresponde a un 54.2% de hombres y 45.8% de mujeres. En la figura 4.2 se muestra la distribución por edad y género. Las imágenes fueron publicadas en el reto de edad ósea de la Sociedad de Radiología de Norte América (RSNA Pediatric Bone Age Challenge, 2017) y fueron provistas por 2 hospitales en Estados Unidos. Todas las imágenes están asociadas con una edad ósea que fue resultado de un reporte radiológico de la institución de donde se obtuvo la imagen, la revisión de 2 radiólogos de cada institución, y una segunda revisión despues de un año por parte de uno de estos. El método utilizado para la determinación de la edad ósea fue el de Greulich Pyle.

Figura 4.2
Distribución por edad y género del conjunto de datos de la RSNA



(a) (b)

Nota. (a) Hombres (b) Mujeres

4.1.1 Preprocesamiento de imágenes

Al revisar las imágenes se detectó una gran variabilidad en cuanto a tonos y presencia de marcadores ajenos a la mano del paciente. Se asumió que estos podrían tener una incidencia importante para encontrar las características más relevantes para la predicción de la edad ósea. Debido a esto, el preprocesamiento constó de 2 pasos: la eliminación del fondo de las imágenes y la aplicación de una ecualización por histograma. Ambos tomados como referencia de lo hecho por Iglovikov, Rakhlin, Kalinin y Shvets (2018), que también identificaron el mismo problema con respecto al conjunto de imágenes del reto de la RSNA.

Para el primer paso se usó una red convolucional U-Net con el fin de eliminar los marcadores. Iglovikov, Rakhlin, Kalinin y Shvets (2018) entrenaron la red con 100 máscaras, que ellos generaron de manera semiautomática con la web supervisely (delineado manual con correcciones automáticas del software) y obtuvieron resultados positivos. Debido a que para esta investigación no se pudo acceder a las funcionalidad de corrección automática para la generación de máscaras, se generaron manualmente 200 máscaras.

Debido a que el entrenamiento de la U-Net fue realizado en la plataforma *google colab,* se estandarizó el tamaño de las imágenes a 256 x 256 con el fin de reducir la carga computacional. Como la finalidad de esta red es determinar si los píxeles forman parte de una región de interés o del fondo, se utilizó la función de pérdida *binary cross entropy*. Asimismo, el modelo fue entrenado con 60 *epochs* y utilizando el optimizador de Adam tal como Ibtehaz y Rahman (2019) lo hicieron para aplicar esta red en un caso de segmentación biomédica.

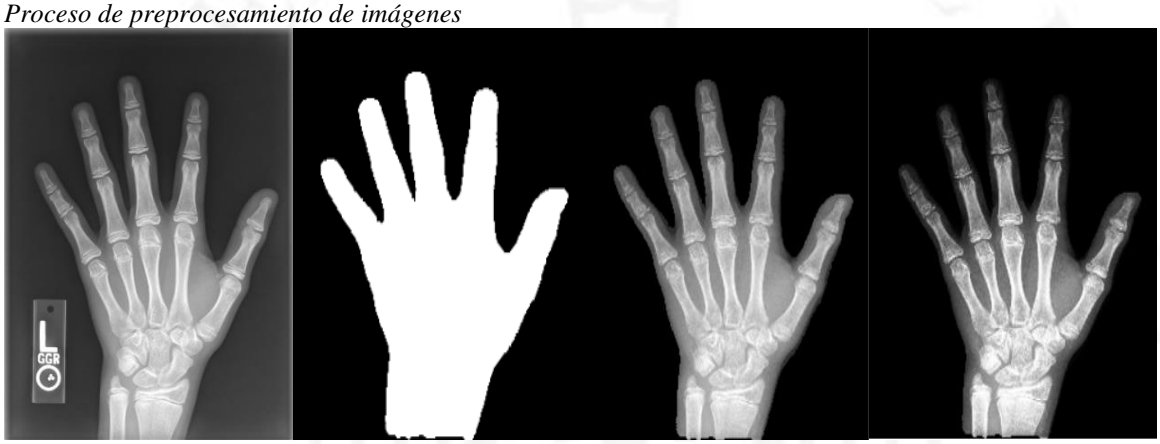
Para determinar si el entrenamiento de la U-Net fue satisfactorio, se utilizó el índice de Jaccard para la comparación de dos imágenes (Kalaiselvi, Vasanthi, y Sriramakrishnan, 2017). Este indicador mide el grado de similitud entre 2 imágenes (A y B) mediante la división de su intersección sobre la su unión. El modelo entrenado obtuvo un índice de 0.92.

$$\text{Índice de Jaccard} = \frac{|\text{A} \cap \text{B}|}{|A \cup B|} \tag{2}$$

A partir de este resultado, se realizó la predicción de 12411 imágenes. Una vez que se obtuvieron las máscaras predichas se aplicaron a las imágenes originales. El segundo paso del preprocesamiento consistió en la aplicación de una ecualización por histograma a cada imagen utilizando la función respectiva en la librería *numpy*. Esto, con la finalidad de realzar los huesos. El resultado del preprocesamiento es mostrado en la figura 4.3.

Figura 4.3

Proceso de preprocesamiento de imágenes



(a) (b) (c) (d)

Nota. (a) imagen original, (b) máscara predicha por U-Net, (c) imagen sin marcadores, (d) aplicación de ecualización

Asimismo, se hizo una evaluación visual, de forma manual, de todas la imágenes para corroborar que no se haya perdido alguna parte de la región de interés. Se encontró que en algunos casos la máscara predicha no cubria toda la mano, por lo que se descartaron. Al hacerlo, el conjunto de imágenes se redujo a 10156. Para asegurar que el descarte no haya afectado a la distribución propuesta por la RSNA se hizo una revisión del porcentaje de la cantidad de las imágenes sin descartar tras el preprocesamiento con respecto a las imágenes originales para cada grupo que será usado en el proceso de validación. En la tabla 4.3 se observa que el porcentaje es de por lo menos 68 %, e incluso, si no fuera por el grupo 1 de los hombres el mínimo sería de un 78%. De esta manera, se consideró innecesario el esfuerzo de modificar la U-Net para reducir el impacto mencionado.

Tabla 4.3

Cantidad de imágenes originales y tras preprocesamiento con su respectivo porcentaje

Tipo de modelo	Cantidad de imágenes originales	Cantidad de imágenes con preprocesamiento sin descartar	Porcentaje
Hombres Grupo 1	185	125	68%
Hombres Grupo 2	5797	4685	81%
Hombres Grupo 3	851	662	78%
Mujeres Grupo 1	80	63	79%
Mujeres Grupo 2	4980	4089	82%
Mujeres Grupo 3	718	554	78%

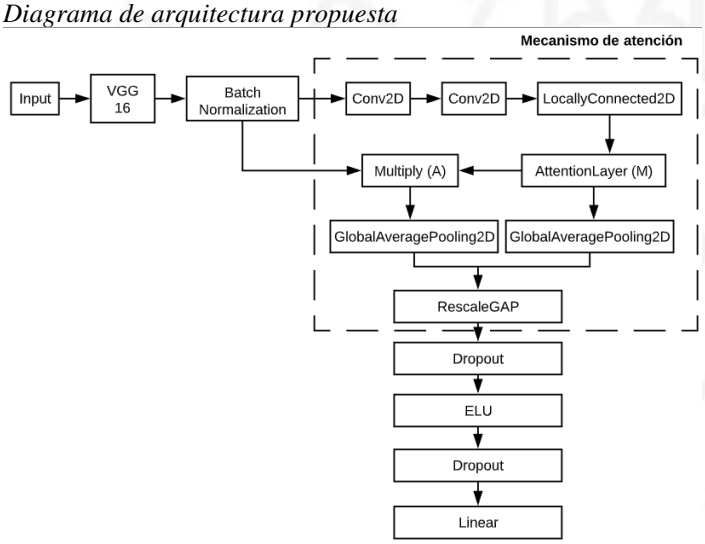
4.1.2 Desarrollo de la red convolucional y entrenamiento

Desarrollo de la red convolucional

Para poder realizar la predicción de la edad ósea, esta investigación se basó en una red VGG-16. Esta supone que la cantidad de parámetros a ser entrenados es mayor que en otras redes como GoogleNet o AlexNet. Sin embargo, la investigación de Canziani, Paszke, y Culurciello (2016) indicó que presentaba una mayor precisión. Debido a que la exactitud de la predicción de la edad ósea es uno de los indicadores más importantes, si no el más importante, se decidió priorizar este aspecto frente al costo computacional que pueda significar el entrenamiento de esta red neuronal convolucional.

Adicionalmente, se aplicó transfer learning para la predicción de la edad ósea. Lee et al., (2017) determinó que no hacerlo le generó muchos casos en donde la función de pérdida no logró converger. Debido a esto, asumió que la inicialización aleatoria de pesos no era un método estable dada la cantidad de imágenes, que fue similar a la que se dispuso para la presente investigación. Los pesos que se utilizaron fueron obtenidos del rendimiento de VGG-16 en ImageNet. Russakovsky et al. (2014) comenta que esta base de datos es un punto de referencia para la clasificación de objetos con alrededor de 15 millones de imágenes en alta resolución clasificadas en 22000 categorias. Si bien los pesos transferidos no están basado en imágenes médicas, Lee et al (2017) y Chen (2017) no presentaron problemas para aplicarlo a la predicción de la edad ósea. Para cargar estos pesos se congelaron las primeras capas, es decir antes de las capas completamente conectada*s* de la VGG-16, con el fin de armar una nueva arquitectura ya que la predicción de la edad ósea es un problema de regresión. La arquitectura propuesta se muestra en la figura 4.4.

Figura 4.4
Diagrama de arquitectura propuesta



Partiendo de la estructura mencionada, se añadió una capa de batch normalization pues Lathuilière, Mesejo, Alameda-Pineda, y Horaud, (2018) determinaron que añadirla antes de la función de activación puede ser significante y beneficioso en casos de redes neuronales convolucionales con una salida de regresión.

Después, se aplicó un mecanismo de atención inspirado en el propuesto por Mader (2018) y aplicado por Górriz, Antony, McGuinness, Giró-i-Nieto, y O'Connor (2019) e Islam et al (2019). Este consiste que a partir de una capa convolucional de entrada con características, se pueden apilar 2 capas convolucionales de 1x1 para extraer caracteristicas espaciales. El resultado despues se pasa a una capa 1x1 localmente conectada (convolución con pesos no compartidos) con activación sigmoidal para obtener una capa de atención (M). Esta máscara será multiplicada por elemento por los mapas de características obtenidos de la capa de batch normalization con el fin de acentuar las áreas informativas y obtener una capa con atención (A). Para terminar se aplicó *Global Average Pooling* (GAP) para la capa de atención (M) y la capa con atención (A). Esto con la finalidad de utilizar una capa Lambda para reescalar las características de A con la de M (*RescaleGAP*).

De esta manera, se pudieron resaltar aprovechar los pesos de cada característica para reforzar los puntos de interés en la imagen. Para visualizar la atención, se pudo extraer el mapa de la figura 4.5, que resaltó las áreas que más se tomaron en cuenta.

Figura 4.5

Imagen original y correspondiente mapa de atención

(a) (b)

Nota.(a) Imagen radiológica (b) Mapa de atención generado por modelo

Tras el mecanismo de atención, se aplicó una capa de *dropout* con el fin de mejorar la capacidad de la red y reducir el riesgo de overfitting (Ren et al. 2018). Posteriormente se aplicó una capa *ELU* (*Exponential Linear Unit*) y después otra capa *dropout.* Finalmente se utilizó una capa con una función de activación linear para realizar la regresión.

Entrenamiento

Debido a la suposición que la presencia de los marcadores y diferencia de tonos en las imágenes radiológicas influirían en la predicción de la edad ósea, se decidió entrenar un modelo con las imágenes originales y preprocesadas, con ambos géneros y con cada uno llegando a un total de 6 modelos. Para realizarlo, se utilizó Keras con Tensorflow y Python 3.6 en una máquina GPU NVIDIA Quadro M4000(8 GB GDDR5 de RAM).

El tamaño de las imágenes fue de 500 x 500, ya que Castillo, Tong, Zhao, y Zhu (2017) determinaron que este era el tamaño óptimo para este dataset. Además, se realizaron transformaciones a las imágenes de entrenamiento que representen las variaciones en imágenes radiológicas mediante *data augmentation*. Cicero y Bilbily (2017) demostraron que su aplicación mejora notablemente el entrenamiento. Los parámetros utilizados se muestran en la tabla 4.4.

Tabla 4.4

Parámetros de data augmentation

Parámetros	Valor
Rotation Range	20
Width Shift Range	0.2
Height Shift Range	0.2
Zoom Range	0.2
Horizontal flip	True

Las imágenes utilizadas fueron separadas en 85% y 15% de entrenamiento y prueba respectivamente. Además, se entrenó el modelo con 50 epochs. Tanto la distribución de imágenes como la cantidad de epochs se deben a que con estos parámetros se obtuvo el mejor resultado. El ratio de aprendizaje utilizado fue de 0.0001. La razón de que sea una cantidad baja se debe a que se buscó que no se sobrescriban los pesos del modelo previamente entrenado (Lee et al 2017).

Al igual que Iglovikov et al (2018) y Chen (2017) se utilizó el optimizador Adam, que buscó minimizar el error absoluto medio (MAE):

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \quad (3)$$

Este indicador muestra la diferencia entre la predicción (x_i), el valor real (y_i) en meses entre la cantidad de imágenes (n)

4.1.3 Validación en entorno local

Se realizó una comparación entre la predicción de la edad ósea con los modelos entrenados con las imágenes originales y la determinación manual de un experto mediante el método de Greulich-Pyle mediante el indicador de MAE en meses. Esto a partir de 230 imágenes radiológicas anonimizadas de la mano de niños y jóvenes entre 0 y 18 años, provistas por un médico radiólogo. Estas imágenes locales, de acceso restringido, fueron conseguidas mediante un consentimiento informado. Para contar con un punto de comparación adicional se usó la web provista por Cicero y Bilbily (2017), cuyo modelo fue el ganador del reto de la RSNA, para predecir las edad óseas de las imágenes mencionadas.

A partir de la revisión de diferentes automatizaciones de edad ósea, se consideró un MAE en meses de 9 para considerar la propuesta como válida.

Cómo se comentó en la sección de metodología, para validar el modelo en el contexto local, se formaron 3 grupos. Cada uno incluye 2 categorías. De esta manera, el primer grupo incluye a niños de hasta 3 años y el de niñas de hasta 2 años. El segundo hasta 14 años para hombres y 13 años para mujeres y el tercer grupo incluye a hombres hasta 19 años y mujeres de hasta 17 años.

5. RESULTADOS Y DISCUSIÓN

En la tabla 5.1 se observa que, con imágenes originales se obtuvo un mejor rendimiento (11.71 de MAE en meses) que con las imágenes preprocesadas (13.1 de MAE en meses). Sorprende el resultado, pues Iglovikov et al (2018) indicó un rendimiento contrario. Puede deberse a que en el preprocesamiento hubo una reducción en la cantidad de imágenes. Si bien se tuvo en cuenta que la distribución de las edades no fuera afectada (se mantuvo en promedio el 78% de las imágenes para cada grupo de edad), y se implementó *data augmentation* para realizar variaciones en las imágenes, aquellas faltantes pueden haber impactado en la generalización.

Tabla 5.1
Resultados de prueba del proceso de desarrollo del modelo

Técnica de validación	MAE en meses con modelo con preprocesamiento	MAE en meses con modelo sin preprocesamiento
Ambos géneros	13.1	11.7
Hombres	9.8	9.4
Mujeres	10.6	9.5

Por otro lado, también se ha obtenido un mejor MAE en meses al separar las imágenes por género con imágenes originales o preprocesadas respectivamente. Tanto para hombres (9.4 y 9.8) como para mujeres (9.5 y 10.6), se obtuvo un mejor rendimiento que al entrenar con ambos (11.7 y 13.1). De acuerdo con Gilsanz (2010), las mujeres siempre tienen una edad ósea más avanzada que los hombres. Por lo tanto, pueden existir ciertas características que diferencien el nivel de madurez esquelética e influencien el entrenamiento.

Como se muestra en la tabla 5.2, el tiempo de entrenamiento representó un obstáculo para la experimentación de diferentes hipótesis y a la hora de encontrar los parámetros más adecuados. Si bien la cantidad de epochs fue la misma para cada modelo (50 epochs), el tiempo varió desde 3 horas, en el caso del modelo de mujeres con imágenes preprocesadas, a 12 horas, en el caso del modelo mixto con las imágenes originales. La razón de la variación entre los tiempos de entrenamiento de los modelos se deben a que la separación por género representó menos imágenes y por lo tanto, un tamaño del *batch* más pequeño. Sin embargo, otra razón por la que puede haber disminuido el tiempo de entrenamiento en casi un 50% puede ser que el preprocesamiento haya disminuido características en la imagen por procesar.

Tabla 5.2
Tiempo de entrenamiento de los modelos

Técnica de validación	Horas con modelo con preprocesamiento	Horas con modelo sin preprocesamiento
Ambos géneros	7	12
Hombres	4	7
Mujeres	3	5

De acuerdo a la tabla 5.3, el MAE en meses conseguido con el modelo de hombres basado en imágenes originales (9.4) se aproxima a los resultados de otras automatizaciones mencionadas en el estado del arte por Iglovikov et al (2018) con 8.1de MAE en meses Sin embargo, aun se encuentra lejos de lo conseguido por Cicero y Bilbily (2017) y Ren et al (2018) con 6 de MAE en meses . Si bien, se probó con el mismo conjunto de imágenes, la cantidad de epochs con los que entrenaron su modelo (Cicero y Bilbily (2017) aplicaron 500) y la diferente arquitectura de red neuronal convolucional pueden haber marcado la diferencia. Cabe recalcar que los resultados mostrados son los comentados por sus respectivos autores como parte de su investigación.

Tabla 5.3
Resultados de automatizaciones de edad ósea

Autores	MAE en meses
Mejor modelo propuesto(hombres)	9.4
Iglovikov et al (2018)	8.1
Cicero y Bilbily (2017)	6.0
Ren et al (2018)	6.0

Con respecto a las 230 imágenes locales, en la tabla 5.4 se muestra que el MAE en meses obtenido por los modelos propuestos de hombres y mujeres fue de un 13.7 y 13.9 respectivamente, mientras que el modelo de Cicero y Bilbily (2017) de 13.1 y 10.1 respectivamente. En ningún caso el modelo propuesto obtuvo un MAE en meses menor al suyo. Sin embargo, solo el modelo de mujeres de los ganadores de la RSNA (10.2) se encuentra cerca a lo establecido para ser considerado como válido (9 de MAE en meses).

Tabla 5.4
Resultados promedio de predicción de edad ósea

Tipo de modelo	MAE en meses con modelo propuesto	MAE en meses con plataforma de ganadores de la RSNA
Hombres	13.7	13.1
Mujeres	13.9	10.2

Al desagregar las predicciones por 3 grupos de edad para cada género, la tabla 5.5 muestra que por el lado de las mujeres el modelo ganador de la RSNA consiguió un resultado dentro del rango objetivo, de 8.8 para el grupo 1 y en los demás grupos obtuvo resultados cercanos, de 10.3 y 10.5 para el grupo 2 y 3 respectivamente, mientras que el modelo de mujeres propuesto obtuvo 17.7, 12.8 y 17.4, aún lejos del MAE objetivo.

Por el lado de los hombres, el modelo de Cicero y Bilbily (2017) tambien obtuvo un mejor resultado pero esta vez solo con los grupos 1 y 3 (10.6 y 13 MAE en meses), pues en el grupo 2 se obtuvo 14, mientras que en el modelo propuesto de hombres se obtuvo un 12.8 de MAE en meses. Lo que es resaltante es que para el grupo 3, con el modelo propuesto de hombres, se obtuvo un 29.1 de MAE en meses, mucho más alto que los demás resultados.

Tabla 5.5
Resultados promedio por grupos de desarrollo

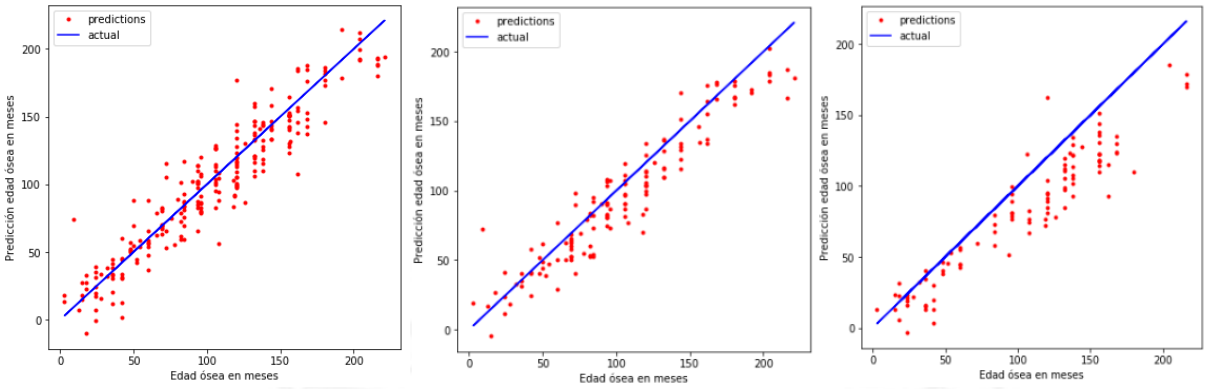
Tipo de modelo	MAE en meses con modelo propuesto	MAE en meses con plataforma de ganadores de la RSNA
Hombres Grupo 1 (hasta 3 años)	12.9	10.6
Hombres Grupo 2 (hasta 14 años)	12.8	14
Hombres Grupo 3 (hasta 19 años)	29.1	13
Mujeres Grupo 1 (hasta 2 años)	17.8	8.8
Mujeres Grupo 2 (hasta 13 años)	12.8	10.3
Mujeres Grupo 3 (hasta 17 años)	17.4	10.5

Al revisar cada predicción y edad ósea determinada por los radiólogos, con apoyo de un gráfico de dispersión, se identificó la presencia de residuales altos de hasta 63 meses, que se muestran en la figura 5.1. Esto conllevó a solicitar a un radiólogo experto que vuelva a determinar la edad ósea de 14 imágenes radiológicas con esta característica con

el fin de validar si la diferencia entre la predicción y la edad ósea determinada era la correcta. Cabe resaltar que solo se le solicitó al radiólogo experto una reevaluación de las imágenes, sin indicar los resultados presentados.

Figura 5.1

Resultados de predicción con valores determinados:



(a) (b) (c)

Nota. (a) modelo mixto, (b) modelo de mujeres, (c) modelo de hombres

En la tabla 5.6 se muestra que tras la reevaluación para los modelos propuestos, solo no hubo una mejora de MAE en meses en el grupo 3 de las mujeres pero en ningún caso hubo un acercamiento al MAE objetivo. A pesar de no haber obtenido resultados favorables, esta prueba confirmó la variabilidad en la determinación de la edad ósea.

Tabla 5.6

Resultados promedio por grupos de desarrollo tras recálculo de residuales altos

Tipo de modelo	MAE en meses con modelo propuesto	MAE en meses con plataforma de ganadores de la RSNA
Hombres Grupo 1	12.6	10.9
Hombres Grupo 2	12.7	14
Hombres Grupo 3	19.5	7.4
Mujeres Grupo 1	12.6	10.9
Mujeres Grupo 2	12.4	10.1
Mujeres Grupo 3	17.4	10.5

Tambien se debe tener en cuenta que a pesar de que la RSNA determinó una muestra de 200 imágenes para determinar a un ganador, la cantidad de imágenes utilizadas en la validación local (230) pueden no ser suficientes para generalizar la aplicabilidad de la propuesta. Además, si bien las edades óseas con la que se han comparado las predicciones han sido utilizadas por un médico, no se ha tenido en consideración un consenso entre radiólogos en su determinación, como se realizó en el conjunto de imágenes de la RSNA.

6. CONCLUSIONES

En esta investigación se ha implementado un proceso de Deep Learning para la predicción de la edad ósea. Debido a la presencia de marcadores y tonos de color diferentes en las imágenes radiológicas, se entrenó una U-Net para la generación de máscaras con las que se eliminó el fondo. Además, para realzar los huesos de las manos, se ecualizaron las imágenes. Además, se implementó una red basada en VGG-16 con transfer learning y un mecanismo de atención. Con esta se entrenaron 6 modelos tomando en cuenta las imágenes preprocesadas y las originales, así como con ambos géneros, y por cada uno.

Se comprobó que el entrenamiento con imágenes originales obtuvo un mejor rendimiento que con imágenes en las que se eliminó el fondo de la mano y se ecualizaron. Asimismo, la separación por género obtuvo un MAE en meses de 9.4 para hombres y 9.5 para mujeres con las imágenes de validación del dataset del reto de la RSNA. Estos resultados se encuentran dentro del rango de otras automatizaciones revisadas, pero aun no esta cerca de las mejores.

Al realizar una validación con 230 imágenes de Perú, se obtuvieron resultados lejanos al MAE en meses objetivo. Además, al separar las imágenes por grupos de edad, se mostraron tanto mejores como peores resultados. Tras esto,

un radiólogo experto volvió a determinar la edad ósea de 14 imágenes radiológicas, que presentaban residuales altos. Lo mencionado mejoró el MAE a excepción del grupo 3 de las mujeres, con lo cual se comprobó la variabilidad de la determinación de la edad ósea entre radiólogos mencionada en otros estudios, que ratifica la necesidad de contar con una automatización como lo pretende esta propuesta.

Por primera vez se ha aplicado una automatización de cálculo edad ósea en un contexto local que nos indica que este tipo de acercamiento puede ser válido. Sin embargo, es necesario realizar más esfuerzos para analizar los distintos casos que existen.

7. TRABAJOS FUTUROS

Los pesos preentrenados obtenidos de ImageNet se encuentran en formato RGB, mientras que la mayoría de las imágenes radiológicas estan bajo el formato estándar DICOM, que está compuesto de imágenes en blanco y negro. Si bien las imágenes y pesos utilizados en esta investigación estaban en RGB, se recomienda transformar las imágenes utilizadas asi como los pesos preentrenados a blanco y negro como lo hizo Chen (2017), con la finalidad de basar el aprendizaje en características propias de una imagen radiológica.

Otro aspecto que se podria tener en cuenta es la homogenización de la rotación de las imágenes para el entrenamiento. Iglovikov et al (2018) incluyó en su fase de preprocesamiento un modelo para identificar puntos estrategicos de la mano y a partir de estos realizar transformaciones.

Además, sería valioso orientar el aprendizaje de características a ciertas partes de la mano. En diferentes investigaciones, como las de Tristán-Vega y Arribas (2008) y Liu et al (2008), la predicción de la edad ósea se ha basado en zonas como el carpo, el radio o la ulna.

De igual manera, se puede reducir el rango de edad de entrenamiento como Kim et al (2017). Lo mencionado podria aumentar la precisión para los diferentes grupos de edad. Tambien seria importante contar con una mayor cantidad de imágenes radiológicas de la mano (800) y que las edades ósea que le correspondan sean consensuadas con por lo menos 2 radiologos.

8. AGRADECIMIENTO

Se reconoce al doctor Carlos Ugas por su asesoría en temas relacionados a la edad ósea y los procesos para determinarla, así como a los profesores José Antonio Taquía y Daniel Cárdenas, de la Carrera de Ingeniería de Sistemas, por el apoyo y consejos dados para el desarrollo de la presente investigación. También se agradece a Fabricio Paredes, docente de la Carrera de Ingeniería Industrial, quien me facilitó el acceso y uso del hardware necesario para el entrenamiento de los modelos propuestos.

REFERENCIAS

Bocchi, L., Ferrara, F., Nicoletti, I. y Valli, G. (2003). An artificial neural network architecture for skeletal age assessment. *IEEE International Conference on Image Processing*, *1*, 1077–1080.

Canziani, A., Paszke, A. y Culurciello, E. (2016). An Analysis of Deep Neural Network Models for Practical Applications, 1–7. Recuperado a partir de http://arxiv.org/abs/1605.07678

Castillo, J. C., Tong, Y., Zhao, J. y Zhu, F. (2017). RSNA Bone-age Detection using Transfer Learning and Attention Mapping.

Chen, M. (2017). Automated Bone Age Classification with Deep Neural Networks, 1–7.

Chen, Y. H., Lopez-Moreno, I., Sainath, T. N., Visontai, M., Alvarez, R. y Parada, C. (2015). Locally-connected and convolutional neural networks for small footprint speaker recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2015-Janua*, 1136–1140.

Cicero, M. y Bilbily, A. (2017). Machine Learning and the Future of Radiology: How we won the 2017 RSNA ML Challenge. Recuperado a partir de https://www.16bit.ai/blog/ml-and-future-of-radiology

Gilsanz, V. y Raitb, O. (2005). *Hand Bone Age*. *Springer*. Recuperado a partir de

http://www.chospab.es/biblioteca/DOCUMENTOS/Atlas_of_Hand_Bone_Age.pdf

Giordano, D., Spampinato, C., Scarciofalo, G. y Leonardi, R. (2010). An automatic system for skeletal bone age measurement by robust processing of carpal and epiphysial/metaphysial bones. *IEEE Transactions on Instrumentation and Measurement*, *59*(10), 2539–2553. https://doi.org/10.1109/TIM.2010.2058210

Górriz, M., Antony, J., McGuinness, K., Giró-i-Nieto, X. y O'Connor, N. E. (2019). Assessing Knee OA Severity with CNN attention-based end-to-end architectures, (April). Recuperado a partir de http://arxiv.org/abs/1908.08856

Ibtehaz, N. y Rahman, M. S. (2019). MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Networks*, *121*, 74–87. https://doi.org/10.1016/j.neunet.2019.08.025

Iglovikov, V. I., Rakhlin, A., Kalinin, A. A. y Shvets, A. A. (2018). Pediatric bone age assessment using deep convolutional neural networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11045 LNCS*, 300–308. https://doi.org/10.1007/978-3-030-00889-5_34

Islam, N., Saeed, U., Naz, R., Tanveer, J., Kumar, K. y Shaikh, A. A. (2019). DeepDR: An image guideddiabetic retinopathy detection technique using attention-based deep learning scheme. *2019 2nd International Conference on New Trends in Computing Sciences, ICTCS 2019 - Proceedings*, 1–6. https://doi.org/10.1109/ICTCS.2019.8923097

Jetley, S., Lord, N. A., Lee, N. y Torr, P. H. S. (2018). Learn to pay attention. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 1–14.

Keras. (2015). LocallyConnected2D. Recuperado el 5 de noviembre de 2020, a partir de https://faroit.com/keras-docs/1.2.2/layers/local/#locallyconnected2d

Kim, J. R., Shim, W. H., Yoon, H. M., Hong, S. H., Lee, J. S., Cho, Y. A. y Kim, S. (2017). Computerized bone age estimation using deep learning-based program: Evaluation of the accuracy and efficiency. *American Journal of Roentgenology*, *209*(6), 1374–1380. https://doi.org/10.2214/AJR.17.18224

Lathuilière, S., Mesejo, P., Alameda-Pineda, X. y Horaud, R. (2018). A Comprehensive Analysis of Deep Regression, 1–17. Recuperado a partir de http://arxiv.org/abs/1803.08450

Lee, H., Tajmir, S., Lee, J., Zissen, M., Yeshiwas, B. A., Alkasab, T. K., … Do, S. (2017). Fully Automated Deep Learning System for Bone Age Assessment. *Journal of Digital Imaging*, *30*(4), 427–441. https://doi.org/10.1007/s10278-017-9955-8

Liu, J., Qi, J., Liu, Z., Ning, Q. y Luo, X. P. (2008). Automatic bone age assessment based on intelligent algorithms and comparison with TW3 method. *Chinese Journal of Medical Imaging Technology*, *24*(10), 1661–1664. https://doi.org/10.1016/j.compmedimag.2008.08.005

Mader, K. (2018). Attention on pretrained-vgg16 for bone age. Recuperado a partir de https://www.kaggle.com/%0Akmader/attention-on-pretrained-vgg16-for-bone-age,

Mahmoodi, S., Sharif, B. S., Graeme Chester, E., Owen, J. P. y Lee, R. (2000). Skeletal growth estimation using radiographie image processing and analysis. *IEEE Transactions on Information Technology in Biomedicine*, *4*(4), 292–297. https://doi.org/10.1109/4233.897061

Mansourvar, M., Raj, R. G., Ismail, M. A., Kareem, S. A., Shanmugam, S., Wahid, S., … Idris, N. (2012). Automated web based system for bone age assessment using historam technique. *Malaysian Journal of Computer Science*, *25*(3), 107–121.

Navarro, M. M., Tejedor, B. M. y López Siguero, J. P. (2014). El uso de la edad ósea en la práctica clínica. *Anales de Pediatria Continuada*, *12*(6), 275–283.

Niemeijer, M., van Ginneken, B., Maas, C. A., Beek, F. J. A. y Viergever, M. A. (2003). Assessing the skeletal age from a hand radiograph: automating the Tanner-Whitehouse method. *Medical Imaging 2003: Image Processing*, *5032*, 1197. https://doi.org/10.1117/12.480163

Pose Lepe, G., Villacrés, F., Silva Fuente-Alba, C. y Guiloff, S. (2018). Correlación en la determinación de la edad ósea radiológica mediante el método de Greulich y Pyle versus la evaluación automatizada utilizando el software BoneXpert. *Revista chilena de pediatría*, (ahead), 0–0. https://doi.org/10.4067/s0370-

41062018005000705

Ren, X., Li, T., Yang, X., Wang, S., Ahmad, S., Xiang, L., … Wang, Q. (2018). Regression Convolutional Neural Network for Automated Pediatric Bone Age Assessment from Hand Radiograph. *IEEE Journal of Biomedical and Health Informatics*, *PP*(c), 1. https://doi.org/10.1109/JBHI.2018.2876916

Ronneberger, O., Fischer, P. y Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9351*, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

RSNA Pediatric Bone Age Challenge. (2017). Recuperado el 1 de agosto de 2019, a partir de rsnachallenges.cloudapp.net/competitions/4%0D

Rucci, M., Coppini, G., Nicoletti, I., Cheli, D. y Valli, G. (1995). Automatic analysis of hand radiographs for the assessment of skeletal age: A subsymbolic approach. *Computers and Biomedical Research*. https://doi.org/10.1006/cbmr.1995.1016

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Fei-Fei, L. (2014). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Sato, K., Ashizawa, K., Anzo, M., Otsuki, F., Kaneko, S., Tanaka, T., … Murata, M. (1999). Setting up an automated system for evaluation of bone age. *Endocrine Journal*, *46*(SUPPL.).

Simonyan, K., Vedaldi, A. y Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. *2nd International Conference on Learning Representations, ICLR 2014 - Workshop Track Proceedings*, 1–8.

Simonyan, K. y Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition, 1–14. Recuperado a partir de http://arxiv.org/abs/1409.1556

Somkantha, K., Theera-Umpon, N. y Auephanwiriyakul, S. (2011). Bone age assessment in young children using automatic carpal bone feature extraction and support vector regression. *Journal of Digital Imaging*, *24*(6), 1044–1058. https://doi.org/10.1007/s10278-011-9372-3

Son, S. J., Song, Y., Kim, N., Do, Y., Kwak, N., Lee, M. S. y Lee, B.-D. (2019). TW3-based Fully Automated Bone Age Assessment System using Deep Neural Networks. *IEEE Access*, *7*, 1–1. https://doi.org/10.1109/access.2019.2903131

Spampinato, C., Palazzo, S., Giordano, D., Aldinucci, M. y Leonardi, R. (2017). Deep learning for automated skeletal bone age assessment in X-ray images. *Medical Image Analysis*, *36*, 41–51. https://doi.org/10.1016/j.media.2016.10.010

Taigman, Y., Ranzato, M. A., Aviv, T. y Park, M. (2014). Taigman_DeepFace_Closing_the_2014_CVPR_paper. https://doi.org/10.1109/CVPR.2014.220

Thodberg, H. H. (2009). An automated method for determination of bone age. *Journal of Clinical Endocrinology and Metabolism*. Endocrine Society. https://doi.org/10.1210/jc.2008-2474

Tristán-Vega, A. y Arribas, J. I. (2008). A radius and ulna TW3 bone age assessment system. *IEEE Transactions on Biomedical Engineering*, *55*(5), 1463–1476. https://doi.org/10.1109/TBME.2008.918554