

Modelos Profundos Preentrenados para la Detección Automatizada de 10 Tipos de Fracturas Óseas Un Estudio Comparativo con ResNet50, EfficientNetB3 y MobileNet

Pre-trained Deep Models for Automated Detection of 10 Types of Bone Fractures: A Comparative Study with ResNet50, EfficientNetB3 and MobileNet

Royer Elvis Moreano Condorcuya ^A, Evelyn Naida Luque Ochoa ^B

Resumen— El diagnóstico rápido y preciso de fracturas óseas es fundamental para prevenir complicaciones y optimizar el tratamiento. Las redes neuronales convolucionales (CNN) han demostrado alta eficacia en la clasificación de radiografías, alcanzando resultados comparables a los de especialistas. Este estudio evaluó MobileNet, ResNet50 y EfficientNetB3 para clasificar 10 tipos de fracturas usando 1,129 imágenes balanceadas: 90% para entrenamiento (1,017) y 10% para validación (112). El preprocesamiento incluyó redimensionado a 256x256 píxeles, conversión a RGB, normalización y one-hot encoding. El entrenamiento se realizó en Google Colab con optimizador RMSprop, pérdida `categorical_crossentropy` y `early stopping`. ResNet50 y EfficientNetB3 alcanzaron 95.54% de exactitud y F1-score >0.94, superando a MobileNet 91.51% de exactitud. Las confusiones se presentaron principalmente en fracturas visualmente similares. Las CNN evaluadas son viables para la clasificación automática de fracturas óseas, constituyendo una herramienta de apoyo diagnóstico útil, especialmente en áreas con escasez de especialistas. Se propone como trabajo futuro incluir un conjunto de prueba independiente y aumento de datos para mejorar la generalización.

Palabras clave: Clasificación de imágenes, diagnóstico asistido, fracturas óseas, redes neuronales convolucionales.

Abstract— Rapid and accurate diagnosis of bone fractures is key to avoiding complications. Convolutional neural networks (CNN) have demonstrated high efficiency in the classification of radiographs, achieving results comparable to those of specialists. This study evaluated MobileNet, ResNet50, and EfficientNetB3 to classify 10 fracture types using 1,129 balanced images: 90% for training (1,017) and 10% for validation (112). Preprocessing included resizing to 256x256 pixels, conversion to RGB, normalization and one-hot encoding. The training was carried out in Google Colab with RMSprop optimizer, `categorical_crossentropy` loss and `early stopping`. ResNet50 and EfficientNetB3 achieved 95.54% accuracy and F1-score >0.94, outperforming MobileNet (91.96% accuracy). Confusions occurred mainly in visually similar fractures. The CNNs evaluated are viable for the automatic classification of bone fractures, constituting a useful diagnostic support tool, especially in areas with a shortage of specialists. It is proposed as future work to include an independent test set and data augmentation to improve generalization.

Keywords: Image classification, assisted diagnosis, bone fractures, convolutional neural networks.

1 INTRODUCCIÓN

El diagnóstico oportuno y preciso de fracturas óseas es fundamental para una adecuada intervención médica, ya que estas lesiones, definidas como la pérdida de continuidad del tejido óseo, pueden ocasionar complicaciones severas como discapacidad permanente o pérdida de función. Las fracturas se clasifican según su morfología y mecanismo de lesión en tipos como avulsión, conminuta, dislocación fracturada, en tallo verde (*greenstick*), fisura (*hairline*), impactada, longitudinal, oblicua, patológica y espiral. Cada tipo presenta características radiográficas específicas que requieren de un diagnóstico preciso para orientar el tratamiento.

Tradicionalmente, este diagnóstico depende del análisis

manual de radiografías; sin embargo, el incremento del volumen de estudios y la escasez de personal especializado generan retrasos en la entrega de informes clínicos[1][2]. En los últimos años, las redes neuronales convolucionales (CNN) han mostrado un gran potencial en el diagnóstico médico, alcanzando resultados comparables a los de radiólogos experimentados en la identificación de fracturas óseas. Por ejemplo, en un estudio[3] se demostró que una CNN logro predecir 95.5% de las fracturas superando a los cirujanos que lograron 92.2% de aciertos. Su eficacia es particularmente notable en la clasificación automatizada de imágenes radiológicas", estudios específicos dieron como resultado que las CNN han igualado en desempeño de radiólogos en la identificación de fracturas

Autores:

- A. **ORCID iD** <https://orcid.org/0009-0004-9354-5137>
Royer Elvis Moreano Condorcuya, estudiante de la Escuela Académica Profesional Ingeniería Informática y Sistemas de la Universidad Nacional Micaela Bastidas de Apurímac 1201remec@gmail.com
- B. **ORCID iD** <https://orcid.org/0000-0002-8386-9806>
Evelyn Naida Luque Ochoa, Docente del Departamento Académico de Informática y Sistemas de la Universidad Nacional Micaela Bastidas de Apurímac elujuque@unamba.edu.pe

especialmente en muñecas o extremidades[4]. Diversos estudios han evidenciado que arquitecturas como VGG-16, ResNet50, EfficientNet y modelos basados en YOLO han alcanzado entre 94% y 99% de precisión[5][6]. En este contexto, el presente trabajo tiene como objetivo evaluar el desempeño de MobileNet, ResNet50 y EfficientNetB3 en la clasificación de diez tipos de fracturas óseas a partir de imágenes radiográficas, empleando un conjunto de datos balanceado y un flujo de preprocesamiento optimizado. Los resultados obtenidos podrán servir como base para desarrollar herramientas de diagnóstico asistido que mejoren la precisión y reduzcan los tiempos de respuesta, especialmente en entornos con recursos limitados. Sin embargo, la validación externa con la integración en las clínicas sigue siendo retos importantes para su adopción generalizada [7].

2 TRABAJOS RELACIONADOS

En los últimos años, múltiples investigaciones han evidenciado el potencial de las redes neuronales convolucionales para el diagnóstico automatizado de las fracturas. Por ejemplo, Mohammad et al [8], su modelo fue evaluado sobre múltiples conjuntos de datos, alcanzando precisiones de 95.01%, 94.04% y 96.25%, superando significativamente a varios métodos convencionales en términos de exactitud diagnóstica, el modelo desarrollado con una arquitectura propia, fue evaluado sobre múltiples conjuntos de datos, alcanzando precisiones del 95.01%, 94.04% y 96.25% en la clasificación de fracturas.

Yasenn et al en 2022 [9] en su investigación titulada "Cervical Spine Fracture Detection and Classification Using Two-Stage Deep Learning Methodology", propusieron un enfoque de dos etapas para la detección automática de fracturas en la columna cervical, una condición médica crítica que puede llevar a parálisis permanente o incluso la muerte. En la primera etapa, utilizaron una CNN para determinar la presencia o ausencia de fractura a partir de imágenes de tomografía computarizada (CT). Para mejorar la interpretabilidad, incorporaron técnicas de visualización como Grad-CAM, en la segunda etapa, se enfocaron en la detección específica de fracturas en cada una de las siete vértebras cervicales, entrenando modelos YOLOv5 y YOLOv8 sobre un conjunto de 9,170 imágenes. El desempeño fue evaluado con métricas como precisión 90%.

Por su parte, Medaramatla et al.[10], en su estudio "Detection of Hand Bone Fractures in X-Ray Images Using Hybrid YOLO NAS", desarrollaron un modelo híbrido para la detección de fracturas óseas en las manos, uno de los tipos de lesiones más comunes en la práctica clínica. La propuesta se basa en una combinación de YOLO NAS (You Only Look Once - Neural Architecture Search), EfficientDet y DETR3 (Detection Transformer), algoritmos reconocidos por su precisión en tareas de detección de objetos.

El modelo fue entrenado utilizando un conjunto de 4,736 radiografías de mano, clasificadas en 6 tipos de fracturas. Se realizó una comparación rigurosa con otros modelos existentes para validar el rendimiento del modelo implementado alcanzó el 98% de precisión, Mu et al. [11], en su investigación "Fine-Tuned Deep Convolutional Networks for the Detection of Femoral Neck Fractures on Pelvic Radiographs: A Multi-

center Dataset Validation", utilizaron un total de 1,491 radiografías pélvicas recolectadas en múltiples instituciones para entrenar y validar un modelo denominado Fine-tuned DR-FDS. Este modelo superó a su versión inicial (DR-FDS) y a otras arquitecturas convencionales, alcanzando una AUC de hasta 0.9937 y una precisión del 85.97% en pruebas externas. Además, mostró una alta generalización y rendimiento clínico al reducir el tiempo de lectura e incrementar significativamente la sensibilidad y especificidad diagnóstica.

Así mismo en el estudio realizado por Spoorthy Torne et al. [12], titulado "Enhancing Bone Fracture Classification in X-Ray Using Deep Learning Models: VGG-16, VGG-16 With Random Forest, ResNet50 With SVM, and EfficientNetB0 With XGBoost", se abordó la problemática de la clasificación automatizada de fracturas óseas utilizando modelos híbridos de Deep Learning y Machine Learning clásico. Utilizando un conjunto de datos que incluía 10 tipos diferentes de fracturas óseas en radiografías, los autores compararon el desempeño de varios modelos: Cada modelo fue evaluado según métricas estándar como precisión, recall y F1-score. Entre todos los enfoques, los modelos basados en VGG-16 tanto el modelo original como su versión con Random Forest obtuvieron los mejores resultados, alcanzando una precisión del 95%, superando al resto de modelos en prácticamente todas las clases de fracturas evaluadas.

3 METODOLOGÍA

3.1 Diseño de investigación

El diseño de la investigación es no experimental de corte transversal, lo que implica que se evaluó el rendimiento de diferentes modelos de redes neuronales convolucionales en la detección de fracturas en un momento específico. De acuerdo con Mairer et al. [13] el investigador no interviene ni modifica las variables independientes y los datos se recolectan en un punto único punto temporal. Este enfoque permitió un análisis objetivo de las métricas de rendimiento, como exactitud, recall y F1-score, utilizando un conjunto de datos previamente establecido

3.2 Participantes

La población de este estudio estuvo conformada por un conjunto de 1129 imágenes etiquetadas de fracturas óseas divididas entre 10 tipos: Avulsion fracture, Comminuted fracture, Fracture Dislocation, 'Greenstick fracture, Hairline Fracture, Impacted fracture, Longitudinal fracture, Oblique fracture, Pathological fracture, Spiral Fracture en todo el cuerpo, obtenidas de bases de datos públicas como Kaggle [14], que proporciona un conjunto de imágenes utilizadas en investigaciones científicas para entrenar y evaluar modelos de clasificación.

Dado el tamaño del conjunto de datos se dividió en dos subconjuntos 90% para el entrenamiento, 10% para la validación, siguiendo las prácticas observadas en este estudio Real-time Classification of Bone Fractures Utilizing Different Convolutional Neural Network Approaches [15] el cual empleó una división 90% para el entrenamiento y 10% para la validación. Así mismo se procuró que todas las clases del conjunto

de datos estén balanceadas.

El conjunto de datos fue dividido de la siguiente manera 1017 imágenes par la parte del entrenamiento y 112 imágenes para la parte de la validación, en la tabla 1 se detalla la distribución por clases de las diferentes imágenes dentro de los subconjuntos de datos de entrenamiento y validación.

TABLA 1

Distribución de las imágenes por los subconjuntos entrenamiento y validación

Clase	Entrenamiento (90%)	Validación (10%)
Avulsion fracture	101	12
Comminuted fracture	101	11
Fracture Dislocation	102	11
Greenstick fracture	102	11
Hairline Fracture	102	11
Impacted fracture	102	12
Longitudinal fracture	102	11
Oblique fracture	102	11
Pathological fracture	101	11
Spiral Fracture	102	11

La figura 1 muestra las imágenes de las diferentes clases de fracturas que se encuentran en el conjunto de datos



Fig. 1. Imágenes de fracturas óseas

3.3 Selección de Arquitecturas

Se seleccionaron tres modelos preentrenados para realizar el desarrollo del presente trabajo.

- **MobilNet:** Se escogió porque está diseñado para ser ligero permitiendo inferencia rápida y bajo uso de memoria[16],[17]. Así mismo puede ajustarse a diferentes tipos de fracturas y regiones anatómicas manteniendo el rendimiento estable [18],[19], además logro grandes niveles de predicción entre 95% y 98% a la hora de predecir fracturas[20], [21]. En la figura 2 se muestra la arquitectura del modelo preentrenado MobilNet

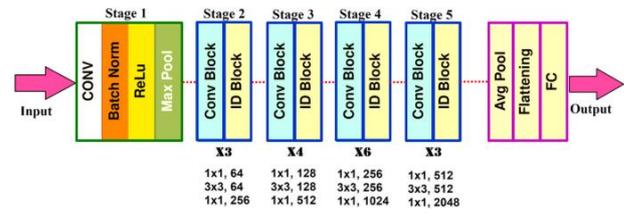


Fig. 2 Arquitectura del modelo ResNet50 extraída de [22]

- **EfficientNetB0:** Se escogió porque se demostró en los siguientes estudios[23],[24], donde se comprobó que EfficientNet es altamente eficiente en la detección de fracturas óseas alcanzados altas tasas de predicción entre un 93.5% y un 99.4% en fracturas de la mano y fracturas cervicales. Así mismo se demostró que EfficientNet destaca por su arquitectura optimizada logrando alto rendimiento con menor parámetros y menor consumo computacional[25]. En la figura 3 se muestra la arquitectura del modelo preentrenado EfficientNetB0



Fig. 3 Arquitectura del modelo EfficientNetB0 extraído de [26]

- **ResNet50:** Se demostró en estos estudios [27],[28] que el uso de este modelo mejoró la capacidad de detección de las fracturas incluso mejor que algunos médicos con poca experiencia. En la figura 4 se muestra la arquitectura de la red convolucional ResNet50

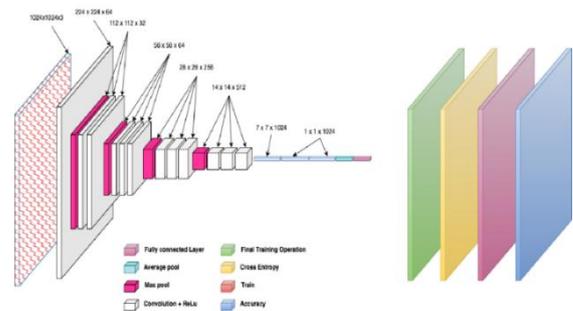


Fig. 4 arquitectura del modelo MobileNet [29]

3.4. Instrumentos y técnicas

Para la implementación de este estudio, se empleó Python como lenguaje de programación, aprovechando su versatilidad y el vasto soporte de bibliotecas especializadas en aprendizaje automático. El entrenamiento de los modelos se llevó a cabo en Google Colab, un entorno de desarrollo que proporcionó acceso a recursos de computación en la nube.

3.5 Preprocesamiento del conjunto de datos

El preprocesamiento se hizo del a siguiente manera:

- Redimensionado y conversión de formato: Todas las imágenes fueron redimensionadas a 256×256 píxeles y convertidas a formato RGB, independientemente de su formato y dimensiones originales, para estandarizar la entrada de datos
- Normalización y preprocesamiento específico del modelo: Se implementaron dos funciones de preprocesamiento, `preprocess_for_model(img)`: convierte a tipo float32 y aplica la función, `preprocess_input()` correspondiente a cada modelo preentrenado, ajustando los valores de píxeles al rango y distribución esperados. `preprocess_for_display(img)`: normaliza los valores de píxeles al rango [0,1] para la visualización en reportes y gráficos. El conjunto de entrenamiento y validación fueron procesados con `preprocess_for_model()`, mientras que `preprocess_for_display()` se utilizó para mostrar ejemplos.
- Agrupamiento en lotes (Batching): Las imágenes se agruparon en lotes de 64 muestras para optimizar el uso de memoria durante el entrenamiento.
- Codificación de etiquetas: Las etiquetas fueron convertidas a formato one-hot encoding, indispensable para usar la función de pérdida `categorical_crossentropy` en la clasificación multiclase.
- Conversión a NumPy para análisis: Los datasets en formato `tf.data.Dataset` se transformaron a arreglos NumPy, lo que permitió realizar análisis estadísticos adicionales y generar matrices de confusión.

Este proceso garantizó que los datos estuvieran en el formato, tamaño y distribución adecuados para maximizar el rendimiento de las arquitecturas de deep learning seleccionadas.

3.6 Análisis de datos

El análisis de datos se llevó a cabo mediante la comparación de las métricas de rendimiento de los modelos entrenados.

Se calcularán las siguientes métricas:

- Exactitud o accuracy: Proporción de predicciones correctas (verdaderos positivos y verdaderos negativos) sobre el total de muestras. La siguiente fórmula muestra accuracy o exactitud

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

- Categorical Crossentropy o Loss: para clasificación multiclase con one-hot, la siguiente fórmula muestra cómo se halla.

$$Loss = -\sum_{i=1}^N \sum_{j=1}^C y_{ij} * \log(\hat{y}_{ij}) \quad (2)$$

- Recall: Qué tan bien el modelo detecta los positivos reales. Es útil cuando te importa no dejar pasar positivos

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

- F1-score: Es el promedio armónico entre precisión y recall. Útil cuando necesitas balancear ambos

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

Para cada modelo utilizando el subconjunto de validación.

La matriz de confusión se empleó para identificar patrones de error específicos, lo que permitió una evaluación más detallada del rendimiento de cada modelo. Los resultados se analizaron estadísticamente para determinar la eficacia de cada red neuronal en la detección de fracturas, garantizando un análisis riguroso y reproducible de los resultados.

4 RESULTADOS EXPERIMENTALES

4.1 Resultados del modelo MobileNet

En la tabla 2 se muestra las métricas de test y validación durante el entrenamiento del modelo MobileNet. El entrenamiento se configuró para 200 épocas, pero se detuvo en la época 150, se usó de optimizador a RMSprop, también se implementó un callback de paciencia en 10 épocas que monitoreaba si la pérdida de la validación no disminuía. El entrenamiento duró un aproximado de 4 horas y 16 minutos.

TABLA 2
Resultados del rendimiento del modelo MobileNet

Métrica	Train	Validación
Exactitud	90.64%	91.51%
Pérdida	0.4804	0.3788
Recall	0.9011	0.9250
F1-score	0.8954	0.9197

En la siguiente tabla se muestra el valor predictivo para cada clase del conjunto

TABLA 3
Resultados predictivos por cada clase del modelo MobileNet

Clase	Precisión
Avulsion fracture	93%
Comminuted fracture	99%
Fracture Dislocation	83%
Greenstick fracture	93%
Hairline Fracture	99%
Impacted fracture	99%

Longitudinal fracture	99%
Oblique fracture	90%
Pathological fracture	92%
Spiral Fracture	91%

La figura 5 muestra el comportamiento de la exactitud durante el entrenamiento del modelo de MobileNet.

- Comportamiento de la exactitud en el entrenamiento: Se observa un comportamiento inconstante pero que aumento gradualmente desde la época 1 con un 09% acabando con un 90.43% en la época 150, indicando que el modelo aprendió acerca de los datos
- Comportamiento de la exactitud en la validación: Se observa un comportamiento inconstante pero que aumento gradualmente desde la época 1 con un 16.96% acabando con un 91.43% en la época 150, indicando que el modelo aprendió acerca de los datos.

Por la brecha y el comportamiento de la exactitud del entrenamiento y validación sugiere un posible underfitting o desaprendizaje a falta de datos

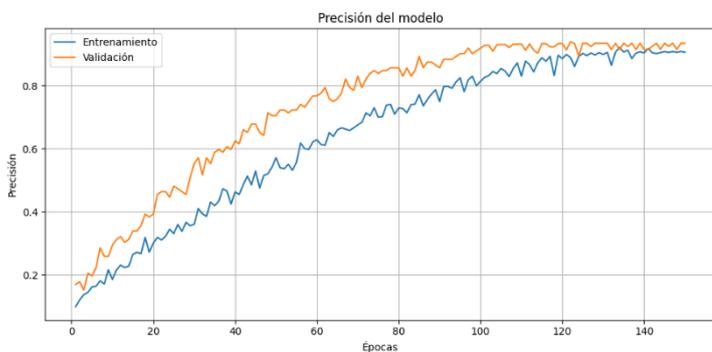


Fig 5. Curva de exactitud durante en entrenamiento del modelo de MobileNet

La figura 6 muestra el comportamiento de la pérdida durante el entrenamiento del modelo de MobileNet

- Comportamiento de la pérdida en el entrenamiento: Se observa un comportamiento inconstante pero que disminuye gradualmente desde la época 1 con un 2.6466 acabando con un 0.4804 en la época 150.
- Comportamiento de la pérdida en la validación: Se observa un comportamiento inconstante pero que disminuye gradualmente desde la época 1 con un 2.2933 acabando con un 0.3788 en la época 150.

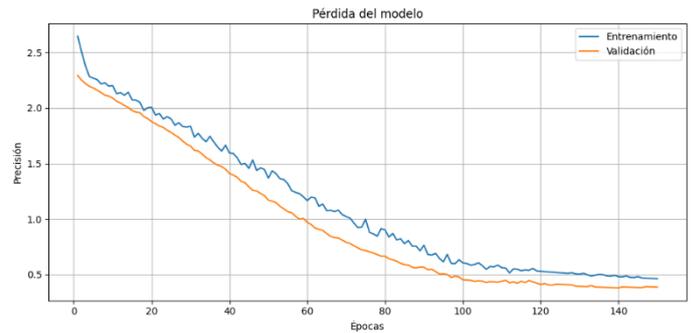


Fig 6 Curva de pérdida (Loss) durante el entrenamiento de MobileNet.

La figura 7 muestra la matriz de confusión del modelo MobileNet hecho durante la variación

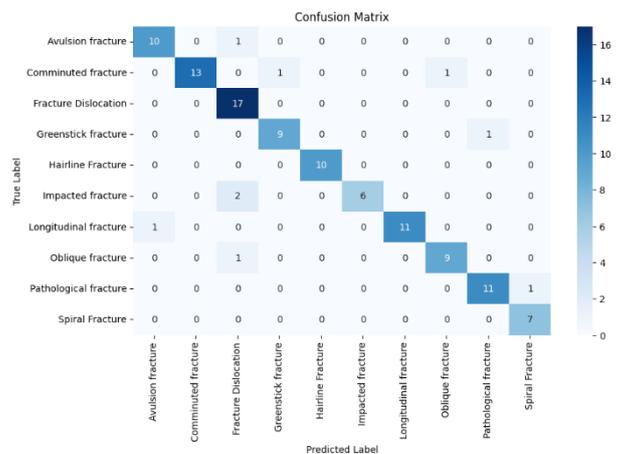


Fig 7 Matriz de confusión de MobileNet en el conjunto de la validación

- Se observa que las clases Fracture Dislocation e Impacted fracture fueron confundidas lo cual indica que hay una similitud visual entre ambas clases, lo mismo para Avulsion fracture fue confundida una vez con Fracture Dislocation y Greenstick fracture con Longitudinal fracture. Esto se debe a que algunas fracturas son similares entre sí.

4.2. Resultados del modelo ResNet50

En la tabla 4 se muestra las métricas de test y validación durante el entrenamiento del modelo ResNet50. El entrenamiento se configuró para 100 épocas, pero se detuvo en la época 62 porque se usó earlystop como callback que monitorea la disminución de la pérdida en la validación con una paciencia de 10 épocas, se usó de optimizador a RMSprop. El entrenamiento duro un aproximado de 1 hora y 30 minutos.

TABLA 4
Resultados del rendimiento del modelo ResNet50

Métrica	Train	Validación
Exactitud	88.48%	95.54%
Pérdida	0.3587	0.1724

Recall	0.9386	0.9564
--------	--------	--------

F1-score	0.9402	0.9466
----------	--------	--------

En la tabla 5 se muestra el valor predictivo para cada clase del conjunto

TABLA 5
Resultados predictivos por cada clase del modelo Res-Net50

Clase	Precisión
Avulsion fracture	99%
Comminuted fracture	88%
Fracture Dislocation	89%
Greenstick fracture	99%
Hairline Fracture	99%
Impacted fracture	99%
Longitudinal fracture	83%
Oblique fracture	99%
Pathological fracture	99%
Spiral Fracture	99%

La figura 8 muestra el comportamiento de la exactitud durante el entrenamiento del modelo de ResNet50.

- Comportamiento de la exactitud en el entrenamiento: Se observa un comportamiento inconstante pero que aumento gradualmente desde la época 1 con un 12% acabando con un 90.30% en la época 62, indicando que el modelo aprendió acerca de los datos.
- Comportamiento de la exactitud en la validación: Se observa un comportamiento inconstante pero que aumento gradualmente desde la época 1 con un 31.25% acabando con un 95.54% en la época 62, indicando que el modelo aprendió acerca de los datos.

Por la brecha y el comportamiento de la exactitud del entrenamiento y validación sugiere un posible underfitting o desaprendizaje a falta de datos

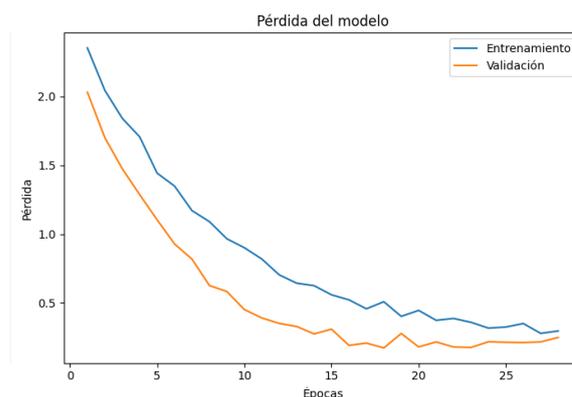


Fig 8. Curva de exactitud durante en entramiento del modelo de ResNet50

La figura 9 muestra el comportamiento de la perdida durante el entramiento del modelo de ResNet50.

- Comportamiento de la perdida en el entrenamiento: Se observa un comportamiento inconstante pero que disminuyo gradualmente desde la época 1 con un 2.8530 acabando con un 0.3201 en la época 62.
- Comportamiento de la perdida en la validación: Se observa un comportamiento inconstante pero que disminuyo gradualmente desde la época 1 con un 2.2933 acabando con un 0.1170 en la época 62

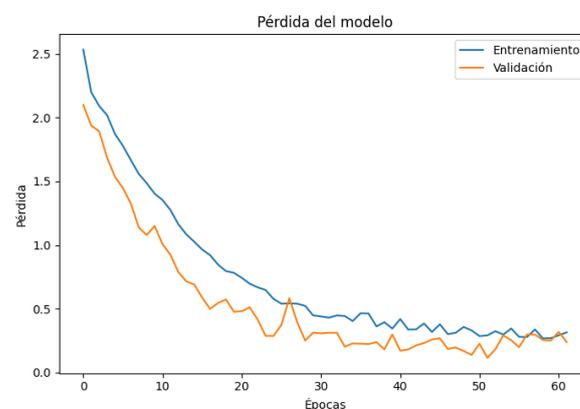


Fig 9 Curva de pérdida durante el entrenamiento de ResNet50.

La figura 10 muestra la matriz de confusión del modelo ResNet50 hecho durante la variación

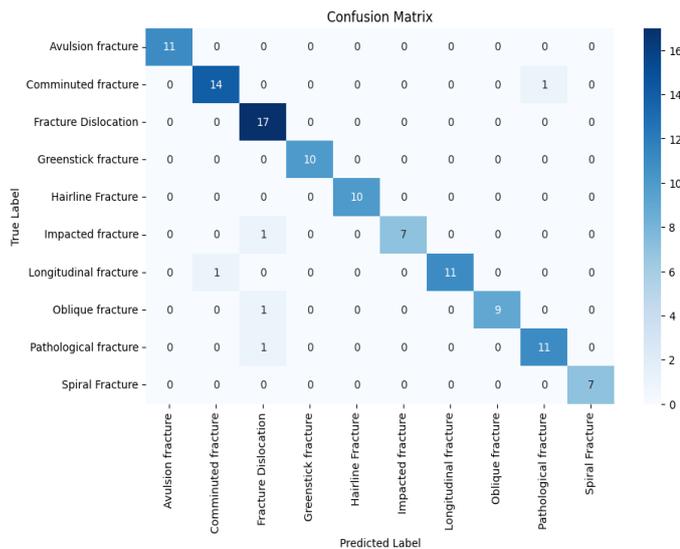


Fig 10 Matriz de confusión de ResNet50 en el conjunto de la validación

- Se observa que Fracture Dislocation tiene 3 falsos positivos, lo que sugiere que otras clases se están confundiendo con esta Impacted fracture, Comminuted, Oblique, Pathological tienen un solo error cada uno. Esto se debe a que algunas fracturas son similares entre sí.

4.3. EfficientNetB3

En la tabla 6 se muestra las métricas de test y validación durante el entrenamiento del modelo EfficientNetB3. El entrenamiento se configuró para 100 épocas, pero se detuvo en la época 28 porque se usó earlystop como callback que monitorea la disminución de la pérdida en la validación con una paciencia de 10 épocas, se usó de optimizador a RMSprop. El entrenamiento duró un aproximado de 1 hora y 30 minutos

TABLA 6
Resultados del rendimiento del modelo EfficientNetB3

Métrica	Train	Validación
Exactitud	88.48%	95.54%
Pérdida	0.3587	0.1724
Recall	0.9304	0.9564
F1-score	0.9300	0.9466

En la siguiente tabla se muestra el valor predictivo para cada clase del conjunto

TABLA 7
Resultados predictivos por cada clase del modelo EfficientNetB3

Clase	Precisión
Avulsion fracture	99%
Comminuted fracture	88%
Fracture Dislocation	89%
Greenstick fracture	99%
Hairline Fracture	99%
Impacted fracture	99%
Longitudinal fracture	83%
Oblique fracture	99%
Pathological fracture	99%
Spiral Fracture	99%

La figura 11 muestra el comportamiento de la exactitud durante el entrenamiento del modelo de EfficientNetB3.

- Comportamiento de la exactitud en el entrenamiento: Se observa un comportamiento inconstante pero que aumento gradualmente desde la época 1 con un 13.83% acabando con un 95.54% en la época 28, indicando que el modelo aprendió acerca de los datos
 - Comportamiento de la exactitud en la validación: Se observa un comportamiento inconstante pero que aumento gradualmente desde la época 1 con un 30.36% acabando con un 95.54% en la época 28, indicando que el modelo aprendió acerca de los datos
- Por la brecha y el comportamiento de la exactitud del entrenamiento y validación sugiere un posible underfitting o desaprendizaje a falta de datos

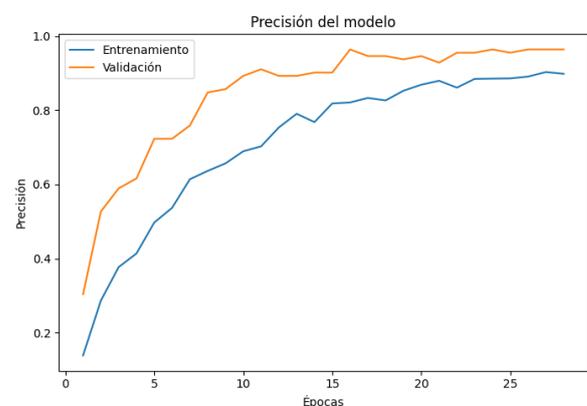


Fig 11. Curva de exactitud durante en entramiento del modelo de EfficientNetB3

La figura 12 muestra el comportamiento de la pérdida durante el entrenamiento del modelo de EfficientNetB3.

- Comportamiento de la pérdida en el entrenamiento: Se observa un comportamiento inconstante pero que disminuye gradualmente desde la época 1 con un 2.3532 acabando con un 0.1724 en la época 28.
- Comportamiento de la pérdida en la validación: Se observa un comportamiento inconstante pero que disminuye gradualmente desde la época 1 con un 2.2933 acabando con un 0.1170 en la época 28

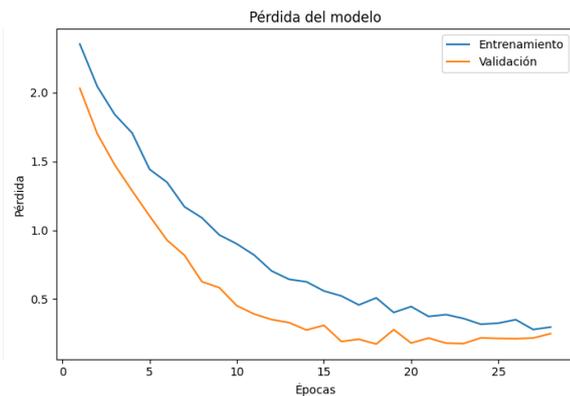


Fig 12 Curva de pérdida durante el entrenamiento de EfficientNetB3.

La figura 13 muestra la matriz de confusión del modelo ResNet50 hecho durante la variación

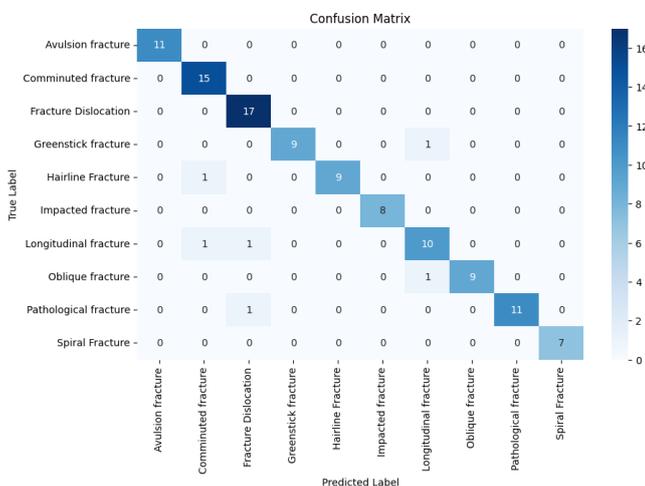


Fig 13 Matriz de confusión de EfficientNetB3 en el conjunto de la validación

- Se observa que Greenstick fracture fue confundida 1 vez como Hairline, Hairline fracture confundida 1 vez como Comminuted, Longitudinal fracture confundida como Comminuted y Fracture Dislocation, Oblique fracture confundida como Longitudinal y Pathological fracture confundida una vez como Fracture Dislocation. Esto se debe a que algunas fracturas son similares entre si.

5 DISCUSIÓN

En los últimos años, múltiples investigaciones han evidenciado el potencial de las redes neuronales convolucionales para el diagnóstico automatizado de fracturas óseas a partir de imágenes médicas. Estos avances han permitido mejorar significativamente la precisión, velocidad y generalización de los modelos, lo cual resulta crucial para su aplicación en entornos clínicos reales. El estudio realizado por Mohammad et al. [1] demostró un enfoque híbrido innovador, integrando técnicas de deep learning con blockchain, obteniendo precisiones de hasta 96.25% en diferentes conjuntos de datos. Este modelo destaca no solo por su alto rendimiento, sino también por su enfoque en la seguridad de los datos médicos, un aspecto crítico en aplicaciones clínicas. Por su parte, Yaseen et al. [2] propusieron un sistema en dos etapas para la detección de fracturas en la columna cervical, utilizando una CNN para el diagnóstico inicial y posteriormente modelos YOLOv5 y YOLOv8 para la localización específica de fracturas en cada vértebra. Su modelo alcanzó una precisión del 90%, lo que demuestra la eficacia del enfoque multi-etapa y la aplicabilidad de las arquitecturas YOLO en contextos clínicos complejos. Asimismo, Medaramatla et al. [3] desarrollaron un modelo híbrido para la detección de fracturas en manos, utilizando YOLO NAS, EfficientDet y DETR3. El modelo fue entrenado con 4,736 imágenes y logró una precisión del 98%, lo que evidencia la capacidad de las arquitecturas modernas para adaptarse a la complejidad morfológica de las fracturas en huesos pequeños. En el trabajo de Mu et al. [4], se utilizó un total de 1,491 radiografías pélvicas para entrenar el modelo Fine-tuned DR-FDS, el cual logró una AUC de hasta 0.9937 y una precisión del 85.97% en pruebas externas, mostrando una alta capacidad de generalización y utilidad clínica en el diagnóstico de fracturas del cuello femoral. De manera similar, Spoorthy Torne et al. [5] exploraron la clasificación multiclase de fracturas en radiografías usando modelos como VGG-16, ResNet50, EfficientNetB0, y sus combinaciones con algoritmos tradicionales como Random Forest, SVM y XGBoost. Entre todos, el modelo VGG-16 con Random Forest alcanzó una precisión del 95%, confirmando la eficacia de los modelos híbridos para problemas de clasificación compleja. En relación con estos trabajos, nuestros resultados muestran un rendimiento competitivo utilizando un conjunto de 1129 imágenes. El modelo MobileNet alcanzó una precisión del 91.96% en validación, con un recall y F1-score de 0.9196 y 0.9197, respectivamente. Por su parte, ResNet50 y EfficientNetB3 mostraron un mejor desempeño, ambos alcanzando una precisión de validación del 95.54%, y un F1-score superior a 0.94. Estos resultados son comparables a los modelos reportados en la literatura, como los de Yasenn et al. y Spoorthy Torne et al., demostrando que los modelos implementados tienen la capacidad de competir con enfoques de vanguardia. Además, nuestros modelos presentan una buena relación entre pérdida (loss) y precisión, lo cual indica un entrenamiento estable y una generalización efectiva en datos no vistos. En particular, el modelo EfficientNetB3 mantiene un equilibrio sólido entre exactitud, recall y F1-score, posicionándose como una de las mejores alternativas en términos de rendimiento global. Estos hallazgos confirman que las arquitecturas modernas, tanto puramente convolucionales como

híbridas, ofrecen soluciones viables y precisas para el diagnóstico automatizado de fracturas óseas. La elección del modelo dependerá del tipo de fractura, la calidad y cantidad de imágenes disponibles, así como de los recursos computacionales en entornos clínicos.

6 FORTALEZAS Y LIMITACIONES

- **Fortaleza:** La principal fortaleza del modelo desarrollado es el uso de EfficientNet, una arquitectura moderna que logra una alta precisión con un bajo costo computacional gracias a su escalado compuesto de profundidad, resolución y ancho. Esta eficiencia estructural permitió obtener buenos resultados incluso con una cantidad limitada de datos y bajo condiciones de entrenamiento restringidas. En el presente trabajo, EfficientNet demostró adaptarse eficazmente al rendimiento requerido, destacando su capacidad para generalizar en clases poco representadas. Como se evidencia en la imagen número 14, el modelo fue capaz de realizar predicciones acertadas.

Real: Longitudinal fracture | Predicción: Longitudinal fracture)



Fig 14 Imagen predicha por el modelo

- **Limitaciones:** En el transcurso de este estudio, surgieron varios obstáculos que quizás hayan afectado el desempeño y la aplicación amplia de los modelos que se prepararon. Uno de los problemas principales fue usar Google Colab sin costo, lo que causó restricciones importantes en cuanto a la potencia de cálculo, el tiempo de uso seguido y el acceso a GPUs de alto nivel. Estas trabas técnicas forzaron a acortar la duración de los entrenamientos y la dificultad del manejo de datos, lo que pudo haber perjudicado la mejora de los hiperparámetros y la cantidad de veces que algunos modelos se repitieron. Otro inconveniente importante fue el tamaño pequeño del grupo de datos, que solo tenía 1129 fotos repartidas en diez tipos. Aunque se intentó que hubiera un equilibrio entre los tipos, esta cantidad es poca para tareas de clasificación de varios tipos con redes neuronales profundas. Esta falta de datos hizo que aumentara el riesgo de un ajuste deficiente, como se vio en las curvas de aprendizaje de los tres modelos evaluados, donde el desempeño en la validación mejoraba, pero sin llegar a un sobreajuste que mostrara una completa asimilación de las características del entrenamiento. Esto indica que los modelos no pudieron aprender patrones complejos de manera óptima debido a la limitada cantidad de ejemplos.

7. CONCLUSIONES Y TRABAJOS FUTUROS

El presente estudio tuvo como objetivo evaluar el desempeño de MobileNet, ResNet50 y EfficientNetB3 en la clasificación automática de diez tipos de fracturas óseas a partir de imágenes radiográficas, utilizando un conjunto de datos balanceado y un flujo de preprocesamiento optimizado. Los resultados evidenciaron que ResNet50 y EfficientNetB3 alcanzaron un rendimiento sobresaliente, con una exactitud del 95,54 % y valores de F1-score superiores a 0,94, superando a MobileNet, que obtuvo un 91,51 % de exactitud. Si bien se identificaron confusiones entre fracturas visualmente similares, los modelos evaluados demuestran ser herramientas viables de apoyo diagnóstico, especialmente en entornos con recursos limitados. Estos hallazgos respaldan la integración de CNN avanzadas en sistemas de asistencia médica para optimizar la precisión y reducir los tiempos de diagnóstico.

Estas métricas son comparables con las mejores propuestas documentadas en la literatura científica reciente, lo que valida la solidez del enfoque adoptado. Además, se ha evidenciado que los modelos presentan buena generalización, bajos niveles de pérdida (loss) durante el entrenamiento y alta capacidad para identificar correctamente múltiples tipos de fracturas. Estas características son fundamentales para su potencial implementación en entornos clínicos reales, donde el diagnóstico rápido y preciso puede significar una diferencia crucial en la atención médica del paciente donde no se cuenta con especialistas. En conjunto, los hallazgos de este trabajo no solo consolidan la viabilidad del uso de CNN en la detección automática de fracturas, sino que también abren la puerta a futuras investigaciones orientadas a mejorar la interpretabilidad, seguridad de los datos y desempeño en condiciones clínicas variables. La combinación de modelos precisos, conjuntos de datos representativos y enfoques centrados en el usuario médico será clave para la adopción masiva de estas tecnologías en el ámbito de la salud.

Para trabajos futuros este estudio implementará un sistema de detección de fracturas con el modelo EfficientNetB3 que desde una página web utilizando el api de Gemini para poder predecir el tipo de fractura que se tiene y sus posibles soluciones diagnósticas ya que en las zonas rurales no existen especialistas, en el contexto peruano, esta problemática se acentúa debido a la alta concentración de equipos de rayos X en la ciudad de Lima, lo que restringe el diagnóstico oportuno en regiones descentralizadas [30]

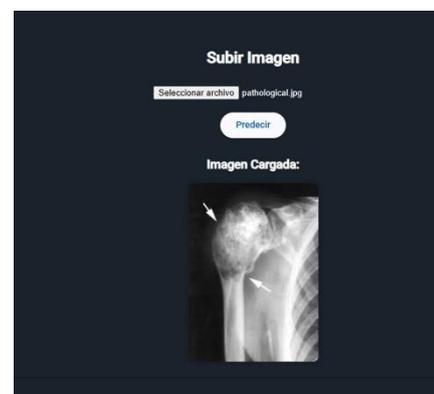


Fig 15 aplicación ingresando la radiografía



Fig 16 aplicación prediciendo que tipo de fractura es y su posible tratamiento.

También se implementará el uso de arquitecturas propias y aumento de datos obteniendo de diversas fuentes como hospitales o clínicas privadas para así mejorar el modelo y descartar el underfitting.

8. REFERENCIAS

- [1] T. S. Omofoye *et al.*, «Backlogs in formal interpretation of radiology examinations: a pilot global survey», *Clin Imaging*, vol. 106, feb. 2024, doi: 10.1016/j.clinimag.2023.110049.
- [2] G. Wahid, Ammara Haroon, Mehreen Samad, y Naila Tamkeen, «Causes of Delay in Radiological Reporting and ways to Reduce them», *Journal of Saidu Medical College, Swat*, vol. 12, n.º 3, pp. 133-137, sep. 2022, doi: 10.52206/jsmc.2022.12.3.697.
- [3] T. Urakawa, Y. Tanaka, S. Goto, H. Matsuzawa, K. Watanabe, y N. Endo, «Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network», *Skeletal Radiol*, vol. 48, pp. 239-244, 2018, doi: 10.1007/s00256-018-3016-3.
- [4] Y. Xie *et al.*, «Artificial intelligence diagnostic model for multi-site fracture X-ray images of extremities based on deep convolutional neural networks», *Quant Imaging Med Surg*, vol. 14, n.º 2, pp. 1930-1943, feb. 2024, doi: 10.21037/qims-23-878.
- [5] N. T. Htun y K. M. M. Tun, «Fuzzy-based Image Enhancement and Ensemble CNN Model for Bone Fracture Detection and Classification System», en *2024 5th International Conference on Advanced Information Technologies (ICAIT)*, 2024, pp. 1-6. doi: 10.1109/ICAIT65209.2024.10754918.
- [6] K. Mittal, K. S. Gill, R. Chauhan, y A. Kapruwan, «Innovative Fracture Diagnosis: MobileNet CNN Approach for Precise Bone Fracture Detection and Classification», en *2024 International Conference on Intelligent Systems for Cybersecurity (ISCS)*, 2024, pp. 1-5. doi: 10.1109/ISCS61804.2024.10581396.
- [7] L. O. Carmo *et al.*, «An increasing number of convolutional neural networks for fracture recognition and classification in orthopaedics», *Bone Jt Open*, vol. 2, pp. 879-885, 2021, doi: 10.1302/2633-1462.210.BJO-2021-0133.
- [8] F. Mohammad, S. Al-Ahmadi, y J. Al-Muhtadi, «Block-Deep: A Hybrid Secure Data Storage and Diagnosis Model for Bone Fracture Identification of Athlete from X-Ray and MRI Images», *IEEE Access*, vol. 11, pp. 142360-142370, 2023, doi: 10.1109/ACCESS.2023.3330914.
- [9] M. Yaseen, M. Ali, S. Ali, A. Hussain, M. Il Joo, y H. C. Kim, «Cervical Spine Fracture Detection and Classification Using Two-Stage Deep Learning Methodology», *IEEE Access*, vol. 12, pp. 72131-72142, 2024, doi: 10.1109/ACCESS.2024.3398061.
- [10] S. C. Medaramatla, C. V. Samhitha, S. D. Pande, y S. R. Vinta, «Detection of Hand Bone Fractures in X-Ray Images Using Hybrid YOLO NAS», *IEEE Access*, vol. 12, pp. 57661-57673, 2024, doi: 10.1109/ACCESS.2024.3379760.
- [11] L. Mu *et al.*, «Fine-Tuned Deep Convolutional Networks for the Detection of Femoral Neck Fractures on Pelvic Radiographs: A Multicenter Dataset Validation», *IEEE Access*, vol. 9, pp. 78495-78503, 2021, doi: 10.1109/ACCESS.2021.3082952.
- [12] S. Torne *et al.*, «VGG-16, VGG-16 With Random Forest, Resnet50 With SVM, and EfficientNetB0 With XGBoost-Enhancing Bone Fracture Classification in X-Ray Using Deep Learning Models», *IEEE Access*, vol. 13, pp. 25568-25577, 2025, doi: 10.1109/ACCESS.2025.3534818.
- [13] C. Maier, J. B. Thatcher, V. Grover, y Y. K. Dwivedi, «Cross-sectional research: A critical perspective, use cases, and recommendations for IS research», *Int J Inf Manage*, vol. 70, p. 102625, jun. 2023, doi: 10.1016/J.IJINFORMGT.2023.102625.
- [14] «Bone Break Classification Image Dataset». Consultado: 14 de julio de 2025. [En línea]. Disponible en: <https://www.kaggle.com/datasets/pkdarabi/bone-break-classification-image-dataset>
- [15] N. Vasker, M. Hasan, M. Nuha, S. Jahan, M. Tahsin, y Md. Y. Emon, *Real-time Classification of Bone Fractures Utilizing Different Convolutional Neural Network Approaches*. 2023. doi: 10.1109/IC-CIT60459.2023.10441387.
- [16] S. Turk, O. Bingol, A. Coşkunçay, y T. Aydin, «The impact of implementing backbone architectures on fracture segmentation in X-ray images», *Engineering Science and Technology, an International Journal*, p., 2024, doi: 10.1016/j.jestch.2024.101883.
- [17] P. M, S. M, R. N, y S. S, «Edge AI-based Bone Fracture Detection using TFLite», *International Journal of Innovative Research in Advanced Engineering*, p., 2025, doi: 10.26562/ijrae.2025.v1204.04.
- [18] A. Khanapure, H. Kashyap, A. Bidargaddi, S. Habib, y A. Anand, «Bone Fracture Detection with X-Ray images using MobileNet V3 Architecture», *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, pp. 1-8, 2024, doi: 10.1109/I2CT61223.2024.10544356.
- [19] R. Bhuria y S. Gupta, «X-Ray Insights: Comprehensive Dataset for Bone Fracture Detection Across Diverse Anatomical Regions», *2024 5th International Conference on Smart Electronics and Communication*

- (ICOSEC), pp. 1242-1247, 2024, doi: 10.1109/ICOSEC61587.2024.10722406.
- [20] M. Goel y G. Singh, «Fracture Detection using MobileNet Model», 2024 *4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS)*, pp. 1574-1579, 2024, doi: 10.1109/ICUIS64676.2024.10866070.
- [21] S. Thota, P. Kandukuru, M. Sundaram, A. Ali, S. Muzamil, y B. Bindu, «Deep Learning based Bone Fracture Detection», 2024 *International Conference on Smart Systems for applications in Electrical Sciences (ICSSES)*, pp. 1-7, 2024, doi: 10.1109/ICSSES62373.2024.10561360.
- [22] S. R. Sannasi Chakravarthy, N. Bharanidharan, C. Vinothini, V. Vinoth Kumar, T. R. Mahesh, y S. Guluwadi, «Adaptive Mish activation and ranger optimizer-based SEA-ResNet50 model with explainable AI for multiclass classification of COVID-19 chest X-ray images», *BMC Med Imaging*, vol. 24, n.º 1, dic. 2024, doi: 10.1186/s12880-024-01394-2.
- [23] A. S. B. Karno *et al.*, «Classification of cervical spine fractures using 8 variants EfficientNet with transfer learning», *International Journal of Electrical and Computer Engineering (IJECE)*, p., 2023, doi: 10.11591/ijece.v13i6.pp7065-7077.
- [24] S.-T. Huang, L.-R. Liu, M.-F. Tsai, M.-Y. Huang, y H.-W. Chiu, «Developing a Deep Learning Model Using Transfer Learning from EfficientNet-b3 to Detect Knee Fracture on X-ray Images», *Proceedings of the 2023 7th International Conference on Medical and Health Informatics*, p., 2023, doi: 10.1145/3608298.3608352.
- [25] K. Teeyapan, «Abnormality Detection in Musculoskeletal Radiographs using EfficientNets», 2020 *24th International Computer Science and Engineering Conference (ICSEC)*, pp. 1-6, 2020, doi: 10.1109/ICSEC51790.2020.9375275.
- [26] H. Amin, A. Darwish, A. E. Hassanien, y M. Soliman, «End-to-End Deep Learning Model for Corn Leaf Disease Classification», *IEEE Access*, vol. 10, pp. 31103-31115, 2022, doi: 10.1109/ACCESS.2022.3159678.
- [27] L. Yeh *et al.*, «A deep learning-based method for the diagnosis of vertebral fractures on spine MRI: retrospective training and validation of ResNet», *European Spine Journal*, vol. 31, pp. 2022-2030, 2022, doi: 10.1007/s00586-022-07121-1.
- [28] Y. Li *et al.*, «Differential diagnosis of benign and malignant vertebral fracture on CT using deep learning», *Eur Radiol*, vol. 31, pp. 9612-9619, 2021, doi: 10.1007/s00330-021-08014-5.
- [29] K. Saini y R. Devi, «A systematic scoping review of the analysis of COVID-19 disease using chest X-ray images with deep learning models», *Journal of Autonomous Intelligence*, vol. 7, dic. 2023, doi: 10.32629/jai.v7i2.928.
- [30] A. P. Bonifaz, C. C. Rodriguez, y R. P. Esparza, «Diagnostic Reference Levels for Common X-ray Procedures in Peru», *Cureus*, vol. 13, p., 2021, doi: 10.7759/cureus.18566.