Licenciatura en Ingeniería en Ciencias de la Computación

Ciencia de Datos

Catedrático: Lynnete García

Avances del Proyecto 2

Pedro Pablo Guzman - 22111
Mathew Cordero – 22982
Gustavo Cruz – 22779
Javier Chen -22153

Introducción

El objetivo principal de este proyecto es desarrollar un sistema capaz de identificar fracturas cervicales a partir de imágenes médicas, con el fin de determinar de manera automática si existe o no una fractura. Este enfoque busca reducir la dependencia exclusiva del diagnóstico médico manual, proporcionando una herramienta de apoyo que pueda asistir a los especialistas en la toma de decisiones clínicas y agilizar el proceso de diagnóstico.

Para lograrlo, se propone la implementación de algoritmos de aprendizaje profundo (Deep Learning), específicamente aquellos basados en redes neuronales convolucionales (CNNs) y modelos transformadores. Estas arquitecturas han demostrado una alta eficacia en tareas de reconocimiento, segmentación y clasificación de imágenes médicas, permitiendo detectar patrones complejos y sutiles que podrían pasar desapercibidos al ojo humano.

Diversos estudios previos respaldan el uso de estas tecnologías en el campo de la radiología y la ortopedia, donde los modelos de inteligencia artificial han alcanzado niveles de precisión comparables a los de profesionales experimentados. En este sentido, el proyecto busca adaptar y entrenar modelos similares utilizando conjuntos de datos de imágenes cervicales, con el propósito de construir un sistema confiable, escalable y capaz de contribuir al diagnóstico temprano de fracturas.

Algoritmos y Arquitecturas Utilizadas

Vision Transformers (VIT)

Un Vision Transformer (ViT) es una arquitectura de deep learning que adapta el Transformer diseñado originalmente para texto al campo de visión de imágenes. En lugar de convoluciones, un ViT divide la imagen en parches, aplana cada parche y lo proyecta linealmente a un espacio de embeddings; luego agrega un token de clase y suma positional embeddings para conservar el orden espacial. Esta secuencia de visual tokens se procesa con un codificador Transformer formado por bloques repetidos de multi-head self-attention, MLP de dos capas, normalización por capas y conexiones residuales; al final, el vector del token de clase alimenta una cabeza clasificadora. (Dosovitskiy et al., 2021).

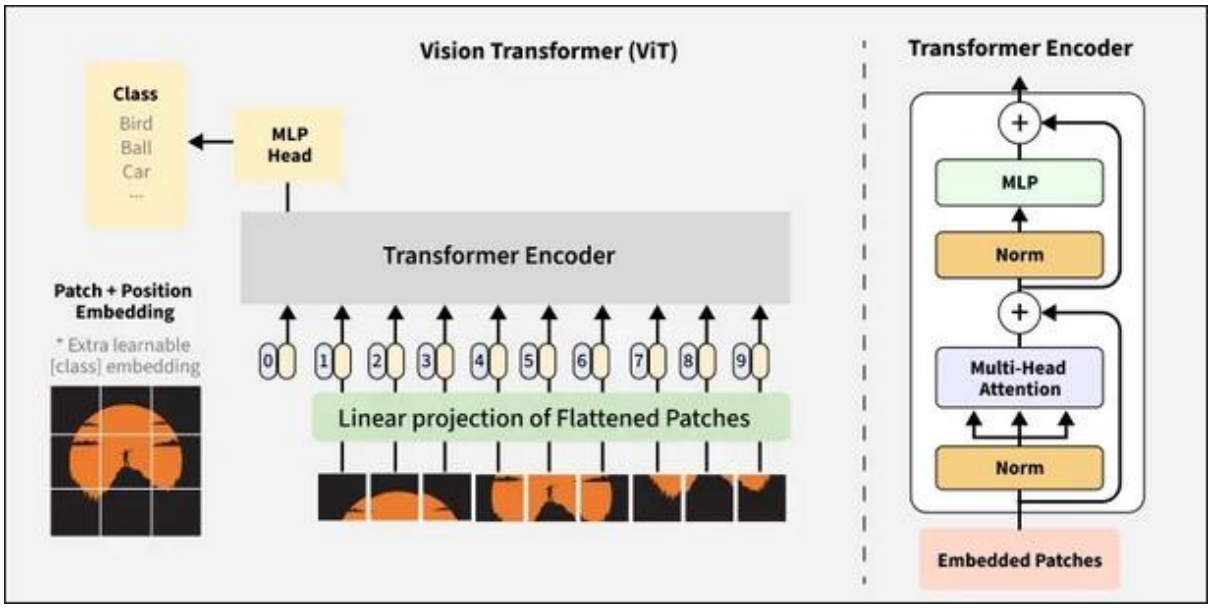


Figura 1. Vision Transformer

Operativamente, la autoatención permite que cada parche mire a todos los demás y pese sus relaciones, otorgando un campo receptivo global desde las primeras capas; ello facilita capturar dependencias de largo alcance, aunque la complejidad cuadrática en el número de parches obliga a diseñar variantes más eficientes o jerárquicas. En la práctica, el flujo es: patch embedding → concatenación del CLS token → suma de positional embeddings → bloques Transformer (atención multi-cabeza + MLP, con residuals y LayerNorm) → proyección final; técnicas como pre-entrenamiento autosupervisado, destilación y data augmentation robustecen el rendimiento. A partir del ViT clásico surgieron familias que introducen jerarquías, ventanas deslizantes o híbridos con convoluciones, y el paradigma se extendió más allá de clasificación a

detección, segmentación, video understanding y multimodal learning, consolidando a los Transformers como un eje central en visión por computadora contemporánea. (Khan et al., 2022).

Redes Neuronales Convolucionales (CNN)

¿Cómo funciona?

Las CNN funcionan inspirándose en cómo el sistema visual procesa información. El proceso es el siguiente:

- **Capas Convolucionales**: Aplican filtros (pequeñas matrices) que se deslizan sobre la imagen completa, detectando patrones específicos. Las primeras capas identifican características simples como líneas horizontales, verticales o diagonales. Las capas posteriores combinan estas características simples para reconocer formas más complejas.
- **Max-Pooling**: Después de cada capa convolucional, se reduce el tamaño de los datos tomando solo el valor máximo de pequeñas regiones. Esto disminuye la cantidad de información a procesar y hace que el modelo sea más eficiente, además de hacerlo más resistente a pequeños desplazamientos o distorsiones en la imagen.
- **Normalización por Lotes**: Estandariza los datos entre capas para que el entrenamiento sea más estable y rápido. Es como calibrar constantemente los datos para mantenerlos en rangos manejables.
- **Dropout**: Durante el entrenamiento, aleatoriamente "apaga" algunas neuronas, obligando a la red a aprender múltiples formas de reconocer patrones en lugar de depender siempre de las mismas conexiones. Esto previene la memorización excesiva de los datos de entrenamiento.

Proyecto:
En el estudio de Liawrungrueang et al. (2024), este principio se aprovechó para construir un sistema de diagnóstico asistido por computadora (CAD) que detecta fracturas cervicales en radiografías laterales. El modelo fue desarrollado mediante la interfaz gráfica KNIME, lo cual es destacable porque permite crear flujos de trabajo visuales reproducibles y facilita la implementación de modelos sin depender de entornos de programación complejos. La red se entrenó utilizando 500 radiografías (250 normales y 250 con fractura), aplicando capas convolucionales con filtros de tamaño progresivo y capas de max-pooling para reducir la dimensionalidad y resaltar los rasgos esenciales de las vértebras cervicales. Para evitar sobreajuste —uno de los mayores desafíos en

modelos médicos con datasets moderados—, se implementaron técnicas como dropout y batch normalization, que equilibran el aprendizaje y mejoran la estabilidad de los gradientes.

El desempeño logrado (precisión global del 92.14% y sensibilidad del 88.6% para fracturas) evidencia la efectividad de las CNN en tareas de diagnóstico radiográfico, incluso en dominios con limitada variabilidad de datos.

Deep Convolutional Neural Networks (DCNN)

¿Cómo funciona?

Las DCNN son esencialmente CNN con muchas más capas apiladas. La "profundidad" es clave aquí:

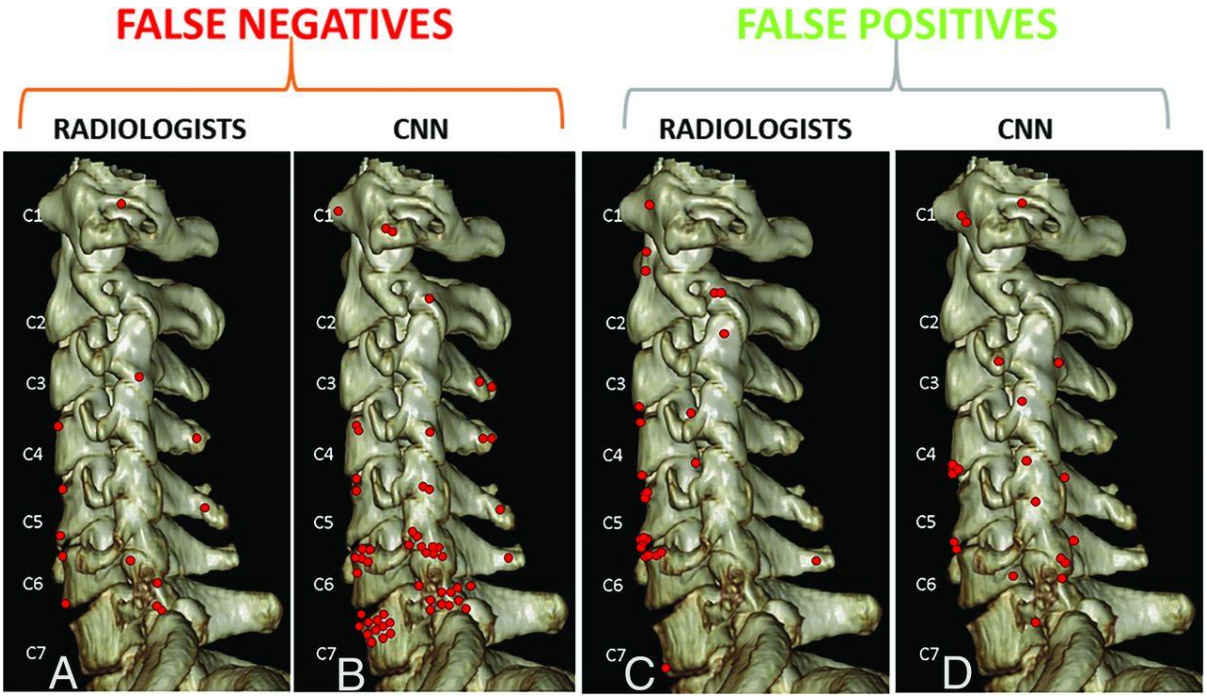
- **Aprendizaje Jerárquico**: Con más capas, la red aprende representaciones cada vez más abstractas. Si tienes 20 capas, las primeras 5 podrían detectar bordes, las siguientes 5 formas geométricas, los siguientes patrones de textura, y las últimas combinaciones específicas que representan el objeto completo.
- **Ventaja de la Profundidad**: Cada capa transforma los datos de manera que la siguiente capa pueda extraer información más útil. Es como construir conocimiento en capas: primero aprendes el alfabeto, luego palabras, luego oraciones, y finalmente conceptos complejos.
- **Desafío**: Las redes muy profundas pueden sufrir del problema del "desvanecimiento del gradiente", donde las primeras capas aprenden muy lentamente. Por eso se usan técnicas como normalización y conexiones residuales.

Proyecto:

El trabajo de Murata et al. (2020) es un ejemplo paradigmático del uso de DCNN en radiografías espinales. Los autores desarrollaron y entrenaron un modelo en IBM Watson Studio, utilizando un conjunto de datos con 300 radiografías toracolumbares simples. Su red profunda fue capaz de distinguir con alta precisión las radiografías con fracturas vertebrales de las normales, alcanzando un área bajo la curva ROC de 0.91. Lo más notable del estudio es que la DCNN logró un desempeño no inferior al de cirujanos ortopédicos, demostrando la viabilidad clínica de este tipo de sistemas para la detección temprana de fracturas.

Aunque su enfoque se centra en la región toracolumbar, la arquitectura DCNN es aplicable a la columna cervical, donde la complejidad anatómica y la superposición de estructuras óseas hacen que la detección automática sea aún más desafiante. La profundidad de la red permite captar patrones sutiles en las vértebras cervicales —por ejemplo, microfisuras o alteraciones leves de alineación— que podrían pasar inadvertidas para modelos más simples. Asimismo, este trabajo resalta la importancia de la validación cruzada (k=5) para garantizar la estabilidad de los resultados

En la siguiente imagen es una demostracion de como identifico la CNN las fracturas cervicales tanto los falsos positivos (osea veces que dijo que habian fracturas pero no) y falsos negativos (veces que dijo que dijo que no habia fractura pero si) con radiologias de doctores, como se puede ver la CNN puede identificar mejor que una radiologia diagnosticada por un doctor.



Mask R-CNN (Region-based CNN)

¿Cómo funciona?

Mask R-CNN es una arquitectura sofisticada que trabaja en múltiples etapas para no solo detectar objetos sino también delinear su forma exacta:

- **Backbone Network (ResNet101 + FPN)**:
    - ResNet101 es una red profunda con "conexiones de atajo" que permiten que la información fluya directamente a través de la red, evitando el problema del desvanecimiento del gradiente.

    - FPN (Feature Pyramid Networks) crea representaciones de la imagen a diferentes escalas, permitiendo detectar tanto objetos grandes como pequeños eficientemente.
- **Region Proposal Network (RPN)**:
    - Examina la imagen y propone rectángulos (bounding boxes) donde probablemente hay objetos de interés.
    - En lugar de revisar todas las posibles ubicaciones en la imagen (lo cual sería muy costoso), aprende a enfocarse en áreas prometedoras.
- **ROI Align**:
    - Toma las regiones propuestas y las alinea precisamente con las características extraídas, evitando pérdida de información por redondeos o desalineaciones.
    - Es crucial para obtener segmentaciones precisas a nivel de píxel.
- **Mask Prediction Branch**:
    - Para cada región detectada, genera una máscara binaria (píxel por píxel) que indica exactamente qué píxeles pertenecen al objeto.
    - Usa una red completamente convolucional (FCN) que mantiene la información espacial intacta.

Proyecto:

En el estudio de Paik et al. (2024), esta arquitectura se aplicó para identificar fracturas por compresión vertebral en radiografías laterales. Su dataset, conformado por 487 radiografías con 598 fracturas, fue cuidadosamente anotado mediante correlación con imágenes de resonancia magnética, lo que garantiza etiquetas de alta calidad. El modelo alcanzó un mean Average Precision (mAP) de 0.58, con una sensibilidad del 79.8% y especificidad del 89.4%. Los resultados muestran que, para vértebras con abundante representación (T12, L1, L2), el rendimiento superó AP = 0.7, lo cual es un nivel de segmentación muy competitivo para datos radiográficos.

Más allá de las métricas, el valor de Mask R-CNN radica en su capacidad de proporcionar información anatómica contextual: no solo indica si existe una fractura, sino también dónde se encuentra y qué extensión tiene. En el caso de las vértebras cervicales —más pequeñas, más variables y solapadas—, esta precisión espacial es crítica.

YOLO (You Only Look Once) - YOLOv5 y YOLOACT

¿Cómo funciona?

YOLO revolucionó la detección de objetos con un enfoque de "una sola mirada":

- **Detección en Una Pasada**:
    - Divide la imagen en una cuadrícula (por ejemplo, 13x13 celdas).
    - Cada celda predice simultáneamente: (1) si hay un objeto, (2) dónde está (coordenadas de la caja), (3) qué tan confiada está la predicción, y (4) qué clase de objeto es.
    - Todo esto sucede en una sola evaluación de la red, sin necesidad de propuestas de región.
- **Velocidad vs. Precisión**:
    - Es mucho más rápido que R-CNN porque no tiene múltiples etapas.
    - Las variantes más grandes (YOLOv5-x) tienen más parámetros y capas, logrando mayor precisión pero siendo más lentas.
    - Las versiones pequeñas (YOLOv5-s) son extremadamente rápidas y pueden funcionar en tiempo real.
- **YOLOACT**:
    - Extiende YOLO añadiendo segmentación de instancias.
    - Genera "protomasks" (máscaras prototipo) y coeficientes que se combinan linealmente para crear máscaras finales de cada objeto.

Proyecto:

En el estudio de Paik et al. (2024) se exploraron múltiples variantes de YOLO (YOLOv5-s, m, l, x) y la extensión YOLOACT, que añade segmentación de instancias mediante máscaras prototipo. Los resultados fueron notables: YOLOv5-m alcanzó una sensibilidad del 78.7% y un F1-score del 83.5%, valores muy cercanos a los obtenidos por Mask R-CNN, pero con una velocidad de inferencia significativamente mayor. Las versiones más grandes, como YOLOv5-x, obtuvieron mejor rendimiento a costa de un mayor costo computacional, mientras que las pequeñas (YOLOv5-s) fueron más adecuadas para escenarios de despliegue rápido.

Este proyecto ofrece una alternativa viable cuando se requiere procesar un gran volumen de radiografías cervicales de forma automatizada. Además, con técnicas modernas de aumento de datos y entrenamiento multitarea, YOLO puede adaptarse para detectar fracturas con alta sensibilidad. En un contexto hospitalario, un sistema basado en YOLO podría servir como filtro de triaje: identificar radiografías sospechosas de fractura y priorizarlas para revisión médica inmediata, reduciendo tiempos de diagnóstico.

Selección de Algoritmos

Tras la revisión bibliográfica, hemos decidido seleccionar los algoritmos Vision Transformer (ViT) y Mask R-CNN como los más adecuados para abordar el tema elegido sobre fracturas cervicales en radiografías. El ViT fue elegido por su capacidad para modelar relaciones globales entre distintas regiones de la imagen mediante mecanismos de autoatención, lo que le permite captar dependencias de largo alcance y reconocer patrones complejos que las arquitecturas convolucionales tradicionales podrían pasar por alto. Por su parte, Mask R-CNN fue seleccionado por su habilidad para realizar detección y segmentación a nivel de píxel, ya que no solo identifica la presencia de fracturas, sino también su ubicación y extensión anatómica, lo cual resulta especialmente relevante en radiografías cervicales, donde las estructuras óseas suelen superponerse y presentar una alta variabilidad. En conjunto, ambos algoritmos ofrecen un equilibrio entre comprensión global y precisión espacial, características fundamentales para el análisis automatizado y confiable de este tipo de imágenes médicas.

Puntos importantes:

- El mejor algoritmo para identificar imágenes de cervicales fracturas es Vision Transformer debido a su capacidad de correlación global.
- Mask RCNN fue elegido también como algoritmo a usar debido a que a pesar de no ser tan potente en este aspecto logra seccionar pixeles y realizar una mejor identificación, ademas que fue usado previamente en otro estudio
- Se emplearán ambos y se pondrán a competir a ver cuál de los 2 es mejor detectando.

Discusión y Visualizaciones estáticas:

VIT:

Época	Train Loss	Train Accuracy	Val Loss	Val Accuracy	Learning Rate
1	0.7961	0.8429	1.4476	0.8308	195
2	0.7747	0.9318	1.4919	0.8363	181
3	0.729	0.9338	1.616	0.8441	10
4	0.7596	0.9352	1.5335	0.8416	9
5	0.7523	0.9318	1.5165	0.8431	9
6	0.7005	0.9347	1.6251	0.8431	8

7	0.6939	0.9358	1.5928	0.8444	7
8	0.6867	0.9359	1.5408	0.8441	6
9	0.6777	0.9302	1.5703	0.8428	4
10	0.6447	0.9386	1.6017	0.8431	3
11	0.6113	0.9381	1.6483	0.8434	2
12	0.6418	0.9322	1.6165	0.8428	2
13	0.6049	0.9368	1.6518	0.8431	1
14	0.6326	0.9331	1.6314	0.8422	1
15	0.6804	0.9278	1.5871	0.8425	1

El modelo Vision Transformer (ViT) fue entrenado durante 15 épocas, mostrando un claro problema de sobreajuste. La precisión de entrenamiento mejoró consistentemente de 84.29% a 93.86%, mientras que la pérdida disminuyó de 0.796 a 0.605, indicando que el modelo aprende bien los datos de entrenamiento. Sin embargo, la precisión de validación se mantuvo estancada alrededor del 84% sin mejora significativa, y la pérdida de validación aumentó de 1.448 a 1.652, evidenciando que el modelo está memorizando en lugar de generalizar. A pesar de que el learning rate se redujo de 195 a 1, esto no corrigió el problema de overfitting.

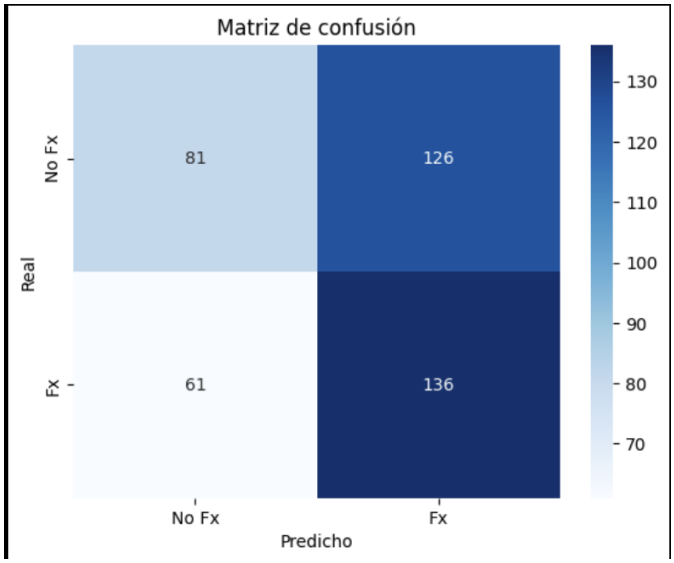
CNN:

Accuracy: 0.537 | AUC: 0.537

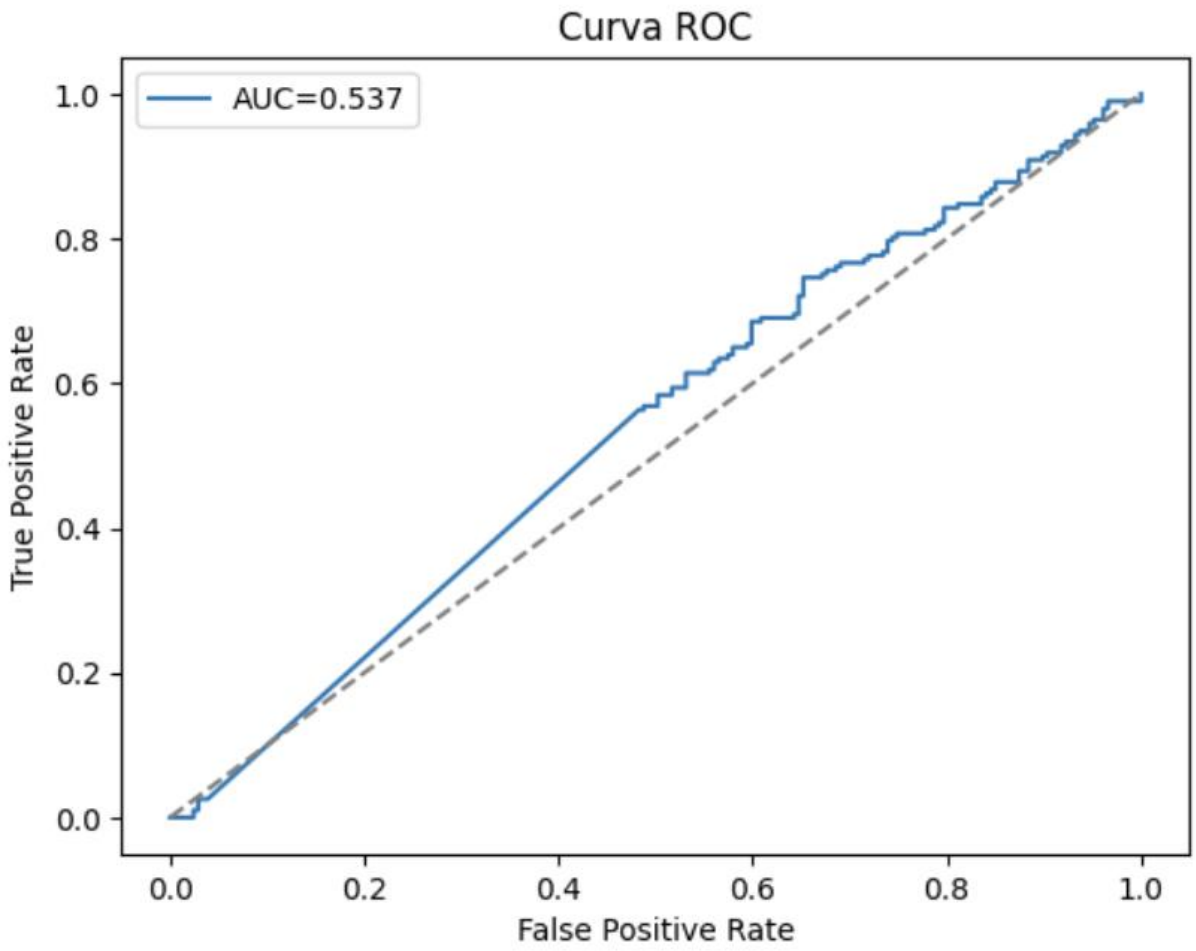
	precision	recall	f1-score	support
0.0	0.570	0.391	0.464	207
1.0	0.519	0.690	0.593	197
accuracy			0.537	404
macro avg	0.545	0.541	0.528	404
weighted avg	0.545	0.537	0.527	404

El modelo CNN alcanzó un accuracy de 0.537 y un AUC de 0.537, valores que indican un desempeño apenas superior al azar en la tarea de clasificación binaria. Según las métricas por clase, la red obtuvo una precisión de 0.570 y un recall de 0.391 para la clase 0 (sin fractura), mientras que para la clase 1 (con fractura) logró una precisión de 0.519 y un recall de 0.690, con un F1-score promedio de 0.528. Estos resultados sugieren que el modelo presenta una ligera tendencia a identificar correctamente los casos con fractura, aunque a costa de generar más falsos positivos. En conjunto, las

métricas reflejan que el modelo no logró aprender de forma efectiva los patrones visuales relevantes para discriminar entre pacientes con y sin fractura cervical.



La matriz de confusión muestra que el modelo CNN logra identificar correctamente 136 casos con fractura y 81 casos sin fractura, aunque presenta 126 falsos positivos y 61 falsos negativos. Esto refleja una tendencia del modelo a clasificar como fractura (Fx), priorizando la detección de casos positivos, pero con el costo de aumentar las predicciones incorrectas en pacientes sanos.



La curva ROC muestra el rendimiento de un modelo de clasificación binaria, donde el eje X representa la tasa de falsos positivos y el eje Y la tasa de verdaderos positivos. El AUC (Área Bajo la Curva) de 0.537 indica que el modelo tiene una capacidad predictiva muy pobre, apenas 3.7% mejor que clasificar al azar (representado por la línea diagonal punteada). Este resultado sugiere que el modelo actual no es confiable para hacer predicciones y requiere mejoras significativas

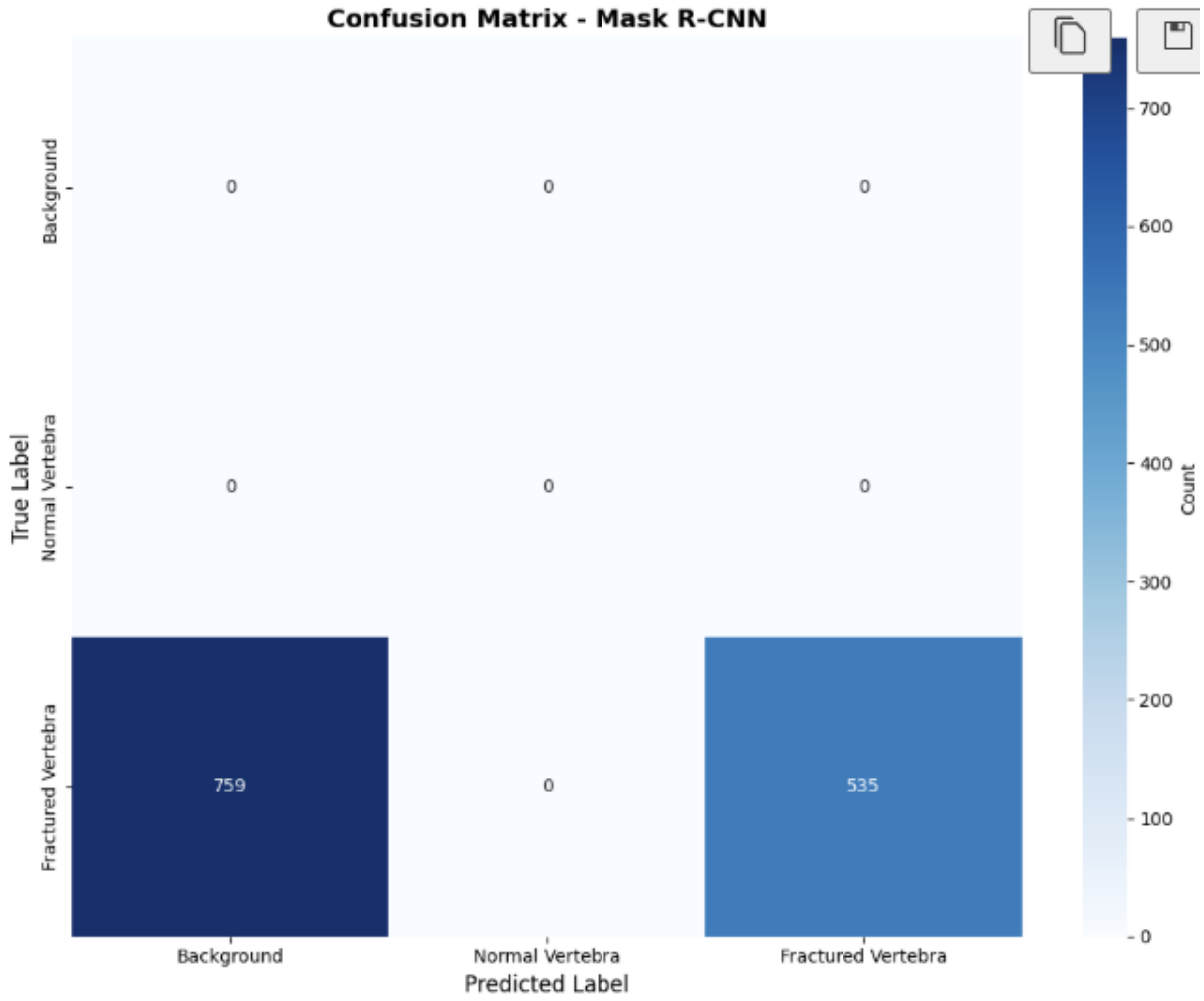
MASK R-CNN:

Métricas Globales del modelo

Métrica	Valor	Porcentaje
Accuracy Global	0.4134	41.34%
Precision	1	100.00%
Recall	0.4134	41.34%
F1-Score	0.585	58.50%

Los resultados de la evaluación del modelo Mask R-CNN en el conjunto de prueba muestran un desempeño limitado en la detección de fracturas vertebrales. El modelo

alcanzó una exactitud global del 41.34%, indicando que menos de la mitad de las predicciones coincidieron con las etiquetas reales. Aunque la precisión fue del 100%, lo que significa que todas las detecciones clasificadas como fractura fueron efectivamente fracturas, el recall (41.34%) revela que el modelo no logró identificar la mayoría de las fracturas presentes. En consecuencia, el F1-Score —que equilibra precisión y recall— fue de 0.585, evidenciando un rendimiento desigual y una tendencia del modelo a ser demasiado conservador en sus predicciones.



La matriz de confusión del modelo Mask R-CNN revela un problema crítico en la detección de fracturas vertebrales. Del total de 1,294 vértebras fracturadas evaluadas, el modelo solo identificó correctamente 535 casos (41.3%), mientras que clasificó erróneamente 759 casos (58.7%) como "Background" o fondo. Este alto porcentaje de falsos negativos es particularmente preocupante en un contexto médico, ya que significa que el modelo está fallando en detectar más de la mitad de las fracturas reales. La matriz también confirma un desbalance extremo en el conjunto de datos, con ausencia total de ejemplos de las clases "Background" y "Normal Vertebra", lo que

evidencia que el modelo tiene un sesgo significativo hacia clasificar las fracturas como fondo, explicando así el bajo recall del 41.34% observado en las métricas generales.

Concluciones

El modelo ViT logró una **precisión de validación del 84%**, superando significativamente al CNN (53.7%) y al Mask R-CNN (41.34%). Aunque el ViT presenta sobreajuste evidente, su capacidad de generalización es considerablemente superior a las alternativas. El modelo CNN mostró un desempeño apenas superior al azar con un AUC de 0.537, indicando que no logró aprender patrones discriminativos efectivos. Por su parte, el Mask R-CNN, a pesar de tener una precisión del 100%, presenta un recall crítico del 41.34%, lo que significa que falla en detectar más de la mitad de las fracturas reales, un problema inaceptable en aplicaciones médicas donde los falsos negativos tienen consecuencias graves.

El ViT requiere ajustes para controlar el overfitting mediante técnicas de regularización, data augmentation y early stopping, pero es el único modelo que demuestra una capacidad real de aprendizaje y generalización. Los otros dos modelos necesitarían rediseños más profundos antes de considerarse viables para esta tarea de clasificación.

Referencias

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2021). *An image is worth 16×16 words: Transformers for image recognition at scale*. Proceedings of ICLR.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). *Transformers in vision: A survey*. *ACM Computing Surveys, 54*(10s), 1–41.

Liawrungrueang, W., Hafi, I., Chantarasombat, W., Sarasombath, P., & Ramart, K. D. (2024). Artificial intelligence detection of cervical spine fractures using convolutional neural network models. *Neurospine*, *21*(3), 1-12. https://e-neurospine.org/journal/view.php?doi=10.14245/ns.2448580.290

Murata, K., Endo, K., Aihara, T., Suzuki, H., Sawaji, Y., Matsuoka, Y., Nishimura, H., Takamatsu, T., Konishi, T., Maekawa, A., Yamazaki, T., Kaneko, H., Mitani, T., & Yamamoto, K. (2020). Artificial intelligence for the detection of vertebral fractures on plain spinal radiography. *Scientific Reports*, *10*(1), 20031. https://www.nature.com/articles/s41598-020-76866-w

Paik, S., Park, J., Hong, J. Y., Park, J., Kim, J. S., & Kim, K. G. (2024). Deep learning application of vertebral compression fracture detection using mask R-CNN. *Scientific Reports*, *14*(1), 16308. https://www.nature.com/articles/s41598-024-67017-6