Detección de Deepfake en imágenes médicas

Juan Sebastián Ortiz Tangarife
Ingeniería de sistemas
Universidad de Antioquia
Medellín, Colombia
juans.ortiz@udea.edu.co

David Agudelo Ochoa
Ingeniería de sistemas
Universidad de Antioquia
Medellín, Colombia
david.agudeloo@udea.edu.co

Jose Franco Arroyave Cardona
Ingeniería de sistemas
Universidad de Antioquia
Medellín, Colombia
franco.arroyave@udea.edu.co

I. Introducción

A. Contexto del problema

En los últimos años, el desarrollo de redes generativas adversariales (GANs) y otras técnicas avanzadas de inteligencia artificial ha permitido la creación de deepfakes (imágenes, audios o videos manipulados de manera que simulan ser reales). En este contexto, el término "deepfake" se refiere a contenido generado artificialmente, a menudo con el objetivo de engañar o desinformar. Aunque estos avances han permitido mejoras en muchas áreas, también han planteado serios desafíos, especialmente en campos donde la precisión y la veracidad son críticas, como en el ámbito médico.

En el campo de imágenes médicas, las tomografías computarizadas (CT), resonancias magnéticas (MRI) y otras imágenes diagnósticas son fundamentales para el diagnóstico y tratamiento de enfermedades. Las manipulaciones de estas imágenes a través de técnicas de deepfake pueden ser extremadamente peligrosas, ya que pueden alterar los diagnósticos médicos, llevando a decisiones incorrectas y perjudicando el bienestar de los pacientes. Por ejemplo, una imagen manipulada de una radiografía podría hacer que un tumor visible sea invisible, o que un área sana sea interpretada erróneamente como problemática.

El objetivo principal de este proyecto es desarrollar una solución automatizada basada en Machine Learning para la detección de imágenes médicas manipuladas. Esto permitiría a los profesionales de la salud contar con una herramienta que los ayude a identificar si una imagen ha sido alterada de alguna manera, lo que contribuiría a preservar la confiabilidad de los diagnósticos médicos.

La implementación de una solución de este tipo es crucial en un contexto donde la información médica es cada vez más digitalizada y accesible en plataformas online, lo que aumenta la probabilidad de que imágenes manipuladas circulen. La detección eficiente de estos deepfakes en imágenes médicas no solo ayudaría a garantizar diagnósticos correctos, sino también a mejorar la seguridad y la confianza en el uso de imágenes digitales para la toma de decisiones médicas.

B. Composición de la base de datos

La base de datos utilizada se llama Deepfakes: Medical Image Tamper Detection (CT-GAN). Este conjunto de datos fue creado para estudiar la manipulación de imágenes médicas mediante técnicas de aprendizaje profundo (deep learning), específicamente en escaneos 3D de tomografía computarizada (CT) de los pulmones [1]. Se utilizaron redes generativas adversarias (GANs) para alterar los escaneos, ya sea insertando cáncer falso (lesiones falsas) o eliminando cáncer real, simulando un ataque malicioso. El objetivo del dataset es permitir el desarrollo y evaluación de técnicas de detección de imágenes médicas alteradas (deepfakes médicos).

a) Contenido del conjunto de datos:

- 100 estudios de tomografía computarizada (CT scans) de tórax.
- Cada escaneo contiene múltiples cortes (slices) de 512x512 píxeles, formando un volumen 3D.
- Los datos están en formato DICOM, el estándar médico para imágenes.

b) Clasificación de los escaneos: Cada volumen está etiquetado según el tipo de manipulación presente:

- **True-Benign(TB)**: Zona sin cáncer, no manipulada.
- **True-Malignant(TM)**: Zona con cáncer real, no manipulada.
- **False-Benign(FB)**: Cáncer real que fue eliminado con deepfake (engañosamente parece sano).
- **False-Malignant(FM)**: Se añadió un cáncer falso con deepfake (engañosamente parece enfermo).

c) Número de muestras: Cada muestra corresponde a una imagen médica individual en formato `.dcm` (DICOM). Las imágenes están organizadas por paciente, donde cada carpeta representa a un paciente y contiene múltiples cortes (slices) de su estudio médico. El número total de imágenes es el siguiente:

- **Experiment 1 - Blind**: 17,457 imágenes
- **Experiment 2 - Open**: 5,296 imágenes
- **Total**: 22,753 imágenes

d) Explicación experimentos: La base de datos está dividida en dos experimentos: Experiment 1 - Blind y Experiment 2 - Open, estos se diferencian en el nivel de información disponible sobre las manipulaciones.

- **Experiment 1 - Blind:** En este experimento, las anotaciones de las regiones manipuladas no fueron utilizadas durante el entrenamiento. Este enfoque simula un escenario más realista, donde las manipulaciones son desconocidas de antemano, y se espera que el modelo aprenda a detectarlas sin depender de etiquetas precisas sobre la ubicación del deepfake.

- **Experiment 2 - Open:** En este caso, sí se proporcionaron anotaciones explícitas de las regiones manipuladas durante el entrenamiento. El modelo conocía durante el proceso de aprendizaje qué partes de las imágenes estaban alteradas. Este enfoque permite una supervisión más precisa, y se espera que los modelos logren un desempeño más alto al tener acceso directo a la "verdad" respecto a las alteraciones en cada imagen.

e) Anotaciones: Cada experimento incluye un archivo `CSV` con anotaciones que indican regiones alteradas dentro de las imágenes. Cada fila en estos archivos representa una anotación, es decir, una posible manipulación en una ubicación específica de una imagen.

- **Experiment 1 - Blind**: 133 anotaciones
- **Experiment 2 - Open**: 36 anotaciones
- **Total**: 169 anotaciones

Las variables presentes en ambos archivos son:

- **type**: Tipo de manipulación (por ejemplo, `FB`).
- **uuid**: Identificador único del paciente.
- **slice**: Número de corte de la imagen donde se encuentra la anotación.
- **x**, **y**: Coordenadas dentro del corte donde se localiza la alteración. Esto brinda la posibilidad de abordar el problema como una tarea de localización de regiones alteradas, usando técnicas como detección de objetos o segmentación.

f) Datos faltantes: Tras una revisión completa, no se encontraron datos faltantes en los datasets.

g) Codificación de variables: Las variables del conjunto de datos se clasifican como sigue:

- **Variables categóricas**:
  - `type`
- **Variables numéricas**:
  - `uuid`, `slice`, `x`, `y`

C. Paradigma de aprendizaje

El problema de detección de deepfakes en imágenes médicas se aborda como una tarea de clasificación binaria supervisada. Cada imagen se etiqueta como alterada (es decir, manipulada mediante técnicas de deepfake) o no alterada (imagen original sin modificaciones). De esta manera, el objetivo del modelo es aprender a distinguir patrones que permitan identificar si una imagen ha sido manipulada o no. Dado que el objetivo no es identificar el tipo de manipulación ni su ubicación exacta, sino simplemente determinar si existe o no una alteración, este enfoque binario permite simplificar el problema y centrarse en detectar patrones globales de manipulación.

El conjunto de datos está desbalanceado, ya que hay muchas más imágenes manipuladas (FB, FM) que originales (TM, TB). Para reducir un poco ese desbalance, se trabajó principalmente sobre la clase TB, duplicando las muestras y generando cortes adicionales alrededor del corte original. Esto se hizo desplazando el valor del corte (slice) hacia adelante y hacia atrás, como si se tomaran imágenes vecinas dentro del mismo estudio. Así se logra aumentar la cantidad de ejemplos TB sin alterar su clase.

II. Estado del Arte

Los autores Yetiş y Çeçen se centraron en detectar si las imágenes médicas habían sido manipuladas [2], es decir, si correspondían a un *deepfake*, abordando el problema mediante un enfoque de clasificación binaria supervisada. Aunque el conjunto de datos contiene múltiples instancias por tomografía, cada uno de los cortes fue tratado como una muestra completa e independiente. El problema se abordó utilizando distintas variantes de la familia de modelos YOLO (You Only Look Once), especializada en la detección de objetos en imágenes, y cuya aplicación ha demostrado ser relevante en escenarios relacionados con *deepfakes* médicos.

De manera similar, Solaiyappan y Wen también trataron el problema de detección de *deepfakes* médicos como una tarea de clasificación binaria supervisada [3]. En este caso, se exploraron tanto modelos convencionales como modelos de aprendizaje profundo. Entre las técnicas utilizadas se incluyeron clasificadores clásicos como SVM, Random Forest y árboles de decisión, así como redes convolucionales preentrenadas como ResNet, VGG y DenseNet, adaptadas posteriormente mediante *fine-tuning*. Al igual que en el trabajo anterior, cada corte 2D fue tratado de forma independiente, sin modelar la secuencia 3D completa.

Sharafudeen y Chandra [4] propusieron una arquitectura basada en redes neuronales convolucionales tridimensionales (3D-CNN) para abordar el mismo problema, con la ventaja de preservar la información espacial y contextual del volumen completo de la tomografía. El modelo fue entrenado con un 80% de los datos, validado con un 10%, y evaluado con el 10% restante. La métrica principal reportada fue la exactitud (*accuracy*), alcanzando un desempeño cercano al 98%.

Por su parte, Albahli y Nawaz [5] propusieron una arquitectura híbrida llamada MedNet, que combina capas convolucionales con un enfoque de red neuronal multicapa para la clasificación final. A diferencia de los trabajos anteriores, utilizaron validación cruzada como método de evaluación y reportaron múltiples métricas: AUC, F1-score, precisión y exactitud. Sus resultados reflejaron un rendimiento robusto, con valores de exactitud superiores al 99%.

Para la validación de los resultados, los autores Yetiş y Çeçen realizaron una partición del conjunto de datos en 80% para entrenamiento y 20% para prueba. No se menciona el uso de validación cruzada. Por su parte, Solaiyappan y Wen también emplearon una partición de los datos en entrenamiento y prueba, aunque no se especifica la proporción exacta ni el uso de técnicas de validación cruzada. En los otros dos estudios, sí se especifican claramente las estrategias de validación.

El rendimiento de los modelos del primer estudio [2] se evaluó mediante métricas estándar en tareas de detección de objetos: precisión (*precision*), recobrado (*recall*) y media de precisión promedio (*mAP*). Estas métricas se calculan en función de la Intersección sobre la Unión (IoU), que mide el grado de coincidencia entre las cajas predichas y las reales

en la detección. En el segundo y tercer estudio [3], [4], las métricas empleadas se centraron en la exactitud (*accuracy*), con resultados que mostraron valores casi perfectos en los mejores modelos de *deep learning* utilizados. En el último trabajo [5], se utilizaron métricas como la AUC y el F1-score, lo que permitió una evaluación más completa delrendimiento de los modelos.

Según el análisis de resultados, el modelo YOLOv5 su fue el más eficaz para la detección de manipulaciones en imágenes médicas del tipo CT, alcanzando un *recall* de 0,997 y un *mAP* superior a 0,931 [2]. Por otro lado, en el trabajo de Solaiyappan y Wen, los modelos basados en redes profundas como ResNet50 y DenseNet121 lograron desempeños notables, con exactitudes cercanas al 100% [3]. De manera similar, los modelos propuestos por Sharafudeen y Chandra [4] y por Albahli y Nawaz [5] demostraron una alta capacidad de generalización, con exactitudes mayores al 98% y métricas adicionales que respaldan la solidez de sus enfoques.

TABLE I
COMPARACIÓN DE ARTÍCULOS SOBRE DETECCIÓN DE DEEPFAKES MÉDICOS

Artículo	**Modelo usado**	**Tipo de dato**	**Validación**	**Métricas**	**Accuracy / mAP**
Yetiş y Çeçen (2024)	YOLOv5	Cortes 2D independientes	80/20	mAP, Recall, Precision	mAP $>$ 0.931
Solaiyappan y Wen (2022)	ResNet50, SVM	Cortes 2D	No especificada	Accuracy	$\approx$ 100%
Sharafudeen y Chandra (2022)	3D-CNN	Volumen 3D	80/10/10	Accuracy	$\approx$ 98%
Albahli y Nawaz (2023)	MedNet híbrido	CT-GAN	Validación cruzada	AUC, F1, Accuracy	$>$ 99%

REFERENCES

[1] S. Shan, Y. Liu, Y. Ding, J. Li, H. Liu, P. S. Yu, and N. Shah, "Deepfakes: Medical image tamper detection (ct-gan)," Disponible en línea, 2020, accessed: 2025-05-14.

[2] U. Yetiş and H. Şafak Çeçen, "A new approach for effective medical deepfake detection in medical images," *IEEE Access*, vol. 12, pp. 1–11, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10495039/

[3] S. Solaiyappan and Y.-X. Wen, "Machine learning based medical image deepfake detection: A comparative study," *Machine Learning with Applications*, vol. 8, p. 100298, 2022. [Online]. Available: https://doi.org/10.1016/j.mlwa.2022.100298

[4] M. Sharafudeen and S. S. Vinod Chandra, "Medical deepfake detection using 3-dimensional neural learning," in *Artificial Neural Networks in Pattern Recognition*, N. El Gayar, E. Trentin, M. Ravanelli, and H. Abbas, Eds. Cham: Springer International Publishing, 2023, pp. 169–180. [Online]. Available: https://doi.org/10.1007/978-3-031-20650-4_14

[5] S. Albahli and M. Nawaz, "Mednet: Medical deepfakes detection using an improved deep learning approach," *Multimedia Tools and Applications*, vol. 83, no. 16, pp. 48 357–48 375, 2024. [Online]. Available: https://doi.org/10.1007/s11042-023-17562-5