Depósito Legal electrónico: ME2016000090 - ISSN Electrónico: 2610797X
http://erevistas.saber.ula.ve/gicos
GICOS - Revista del Grupo de Investigación en Comunidad y Salud
Volumen 11, N 3, 2026

Artículo Original

Fiabilidad diagnóstica de la inteligencia artificial multimodal vs. juicio clínico en la clasificación de fracturas AO/OTA: un estudio exploratorio

Diagnostic reliability of multimodal artificial intelligence vs. clinical judgment in AO/OTA fracture classification: an exploratory study

LEÓN, FRANCISCO[1,2]; PICÓN, KRYSTELL[1,2]

[1]Universidad de Los Andes. Mérida, Venezuela.
[2]Instituto Autónomo Hospital Universitario de Los Andes (IAHULA). Mérida, Venezuela

Autor de correspondencia
franciscojleon95@gmail.com

Fecha de recepción
20/06/2026
Fecha de aceptación
31/07/2026
Fecha de publicación
25/09/2026

Autor
León, Francisco
Médico residente de 4to año de Ortopedia y Traumatología de la Unidad Docente Asistencial de Ortopedia y Traumatología (UDAOT-ULA), Instituto Autónomo Hospital Universitario de Los Andes (IAHULA), Universidad de Los Andes, Mérida, Venezuela.
ORCID: https://orcid.org/0009-0006-6756-9877

Picón, Krystell
Médico Traumatólogo Ortopedista, Fellow en Cirugía de Pierna, Pie y Tobillo, Adjunta de la Unidad Docente Asistencial de Ortopedia y Traumatología (UDAOT-ULA), Instituto Autónomo Hospital Universitario de Los Andes (IAHULA), Universidad de Los Andes, Mérida, Venezuela.
ORCID: https://orcid.org/0009-0008-8686-2774

Citación:
León, F., Picón, K. (2026). Fiabilidad diagnóstica de la inteligencia artificial multimodal vs. juicio clínico en la clasificación de fracturas AO/OTA: un estudio exploratorio. *GICOS, 11*(3), 39-53

RESUMEN

Introducción: La clasificación AO/OTA es el estándar global para la categorización de fracturas, aunque presenta variabilidad interobservador, especialmente en personal en formación. Los modelos de lenguaje de gran escala (*LLM*) multimodales emergen como una alternativa diagnóstica frente a las redes neuronales convolucionales (*CNN*) tradicionales. Materiales y Métodos: Estudio de corte transversal y comparativo. Se seleccionaron 50 radiografías de huesos largos, pelvis y superficies articulares. La muestra fue evaluada por el modelo Gemini 3 Flash y por residentes de la UDAOT-ULA, utilizando como *Gold Standard* el consenso de especialistas. Se empleó el software SPSS v.29 para calcular precisión, sensibilidad, especificidad y concordancia mediante el Índice de Kappa de Cohen ($p < 0,05$). Resultados: La IA alcanzó una precisión global del 88% y una sensibilidad del 95,7%, superando a los residentes de tercer año (R3: 84%). El análisis de concordancia mostró un índice Kappa de 0,79 para la IA (Sustancial), equiparable a los residentes de cuarto año (0,86) y significativamente superior a los R1 (0,08). No obstante, la IA presentó errores cualitativos graves en regiones de alta complejidad anatómica como la pelvis. Conclusiones: La IA multimodal demuestra ser una herramienta de soporte eficaz y un nivelador de competencias para residentes de niveles iniciales. Sin embargo, debido a limitaciones en el razonamiento espacial complejo, su uso debe integrarse como un sistema de doble lectura bajo supervisión del especialista.

Palabras clave: inteligencia artificial multimodal, clasificación AO/OTA, fracturas, residentes, fiabilidad diagnóstica

ABSTRACT

Introduction: The AO/OTA classification is the global standard for fracture categorization, although it presents interobserver variability, especially among personnel in training. Multimodal large language models (*LLM*s) are emerging as a diagnostic alternative to traditional convolutional neural networks (*CNN*s). Materials and Methods: A cross-sectional comparative study. Fifty radiographs of long bones, pelvis, and articular surfaces were selected. The sample was evaluated by the Gemini 3 Flash model and by residents of the UDAOT-ULA, using a consensus of specialists as the *Gold Standard*. SPSS v.29 software was used to calculate accuracy, sensitivity, specificity, and concordance using Cohen's Kappa index ($p < 0.05$). Results: The AI achieved an overall accuracy of 88% and a sensitivity of 95.7%, outperforming third-year residents (R3: 84%). Concordance analysis showed a Kappa index of 0.79 for the AI (Substantial), comparable to fourth-year residents (0.86) and significantly higher than first-year residents (R1: 0.08). However, the AI presented severe qualitative errors in regions of high anatomical complexity such as the pelvis. Conclusions: Multimodal AI proves to be an effective support tool and a competence leveler for junior residents. However, due to limitations in complex spatial reasoning, its use must be integrated as a double-reading system under specialist supervision.

Keywords: multimodal artificial intelligence, AO/OTA classification, fractures, residents, diagnostic reliability

INTRODUCCIÓN

La clasificación de las fracturas es un pilar fundamental en la cirugía ortopédica moderna, ya que permite estandarizar el lenguaje clínico, guiar la toma de decisiones terapéuticas y facilitar la investigación científica. El sistema de clasificación AO/OTA (*Arbeitsgemeinschaft für osteosynthesefragen / Orthopaedic trauma association*) se ha consolidado como el estándar de oro a nivel mundial debido a su estructura jerárquica y lógica anatómica (Meinberg et al., 2018). No obstante, a pesar de su adopción universal, diversos estudios han reportado una variabilidad interobservador significativa, especialmente entre cirujanos en formación, lo que puede comprometer la precisión del plan quirúrgico (Audigé et al., 2005).

Históricamente, la fiabilidad de estas clasificaciones depende de la experiencia del observador y de la calidad de los estudios radiográficos. En centros de trauma de tercer nivel, los médicos residentes enfrentan el desafío de categorizar lesiones de alta complejidad en entornos de alta presión asistencial (Müller et al., 1990). Investigaciones previas sugieren que, aunque el entrenamiento formal mejora la concordancia, incluso los especialistas experimentados presentan discrepancias en fracturas complejas o superficies articulares específicas (Kooistra et al., 2011).

En la última década, la inteligencia artificial, particularmente los modelos basados en visión computacional y aprendizaje profundo (*Deep Learning*), ha demostrado una capacidad creciente para el análisis de imágenes médicas, perfilando una nueva era de medicina de alto rendimiento mediante la convergencia de capacidades humanas y digitales (Topol, 2019). A nivel internacional, se han documentado esfuerzos significativos para validar el desempeño de estas tecnologías frente al ojo humano. Por ejemplo, Olczak et al. (2017) evaluaron el rendimiento de redes de aprendizaje profundo en la interpretación de más de 180,000 radiografías musculoesqueléticas, reportando una precisión diagnóstica equiparable a la de cirujanos ortopédicos experimentados.

Específicamente en el ámbito del trauma ortopédico, Langerhuizen, Janssen et al. (2019) demostraron que, si bien los algoritmos de inteligencia artificial alcanzan niveles de concordancia interobservador similares a los de los médicos residentes en fracturas de huesos largos, el juicio del especialista senior sigue superando las capacidades predictivas de los modelos de primera generación. Asimismo, Chen et al. (2022) evidenciaron en su revisión que la integración de estos modelos automatizados optimiza los flujos de trabajo en urgencias y actúa como una herramienta de soporte que eleva la precisión diagnóstica de los médicos residentes, ayudando a mitigar la brecha de experiencia durante las etapas iniciales de la curva de aprendizaje.

Sin embargo, la implementación de la inteligencia artificial no está exenta de limitaciones. Se ha documentado que estos modelos pueden verse afectados por variables confusoras del entorno sanitario u hospitalario, incurriendo en sesgos cognitivos digitales o “alucinaciones” donde artefactos técnicos son malinterpretados como trazos de fractura, o donde se ignora la visión sistémica de la estabilidad en favor de detalles aislados (Badgeley et al., 2019). Esta dicotomía entre la precisión matemática de la máquina y el juicio clínico integral del cirujano humano plantea la necesidad de evaluar formalmente si el *deep learning* es realmente equiparable

al especialista en la categorización de patrones óseos específicos (Langerhuizen et al., 2020).

La evolución de la inteligencia artificial en el diagnóstico por imagen ha transitado desde las redes neuronales convolucionales (*CNN*), especializadas en tareas cerradas, hacia los modelos de lenguaje extensos (*LLM*) multimodales de alta capacidad para la medicina (Moor et al., 2024). A diferencia de los sistemas convencionales, los *LLM* médicos de última generación integran la visión computacional con capacidades de razonamiento clínico avanzado (Thirunavukarasu et al., 2023). Esta arquitectura permite no solo identificar soluciones de continuidad ósea, sino procesar y codificar el conocimiento clínico y la jerarquía lógica intrínseca a la clasificación AO/OTA (Singhal et al., 2023). Al actuar mediante una comprensión semántica del daño estructural, estos modelos emulan el proceso cognitivo del cirujano, ofreciendo una ventaja potencial en la interpretación de fracturas complejas donde el contexto anatómico y la estabilidad mecánica son determinantes.

No obstante, la validez interna y externa de la evidencia clínica actual sobre estas herramientas se encuentra supeditada a limitaciones críticas. Por un lado, la representatividad de la muestra en contextos locales suele restringirse a casuísticas institucionales acotadas, lo que exige evaluar el comportamiento del algoritmo ante un espectro específico de calidad radiológica y patrones lesionales propios de un único centro. Por otro lado, desde la perspectiva del software, la evaluación de modelos de lenguaje extensos de acceso general plantea un escenario metodológico distinto al de los softwares ortopédicos de nicho; al carecer de un entrenamiento primario y exclusivo para la lectura biomédica, su rendimiento es altamente dependiente de la estructura del *prompt* y es susceptible a sesgos interpretativos bajo entornos de alta presión asistencial.

El presente estudio surge de la necesidad de comparar la eficacia diagnóstica de la inteligencia artificial frente a los distintos niveles de formación del postgrado, utilizando como base las discrepancias observadas en la práctica diaria para fortalecer el proceso de enseñanza-aprendizaje y la seguridad del paciente.

MATERIALES Y MÉTODOS

Diseño del estudio

Se realizó un estudio de corte transversal, observacional y exploratorio en el servicio de ortopedia y traumatología de la Universidad de Los Andes (UDAOT) en Mérida, Venezuela. El protocolo se centró en evaluar la concordancia diagnóstica y la precisión de la inteligencia artificial (IA) frente al juicio clínico humano en diferentes niveles de formación médica.

Muestra

Se seleccionó una muestra no probabilística por conveniencia de 50 estudios radiográficos digitales en proyecciones ortogonales (anteroposterior y lateral) procedentes del archivo del Instituto Autónomo Hospital Universitario de Los Andes (IAHULA).

- *Criterios de inclusión:* radiografías de esqueletos maduros con fracturas de huesos largos, anillo pélvico o superficies articulares, con calidad técnica óptima para la visualización de corticales óseas.

- *Criterios de exclusión:* estudios con artefactos de movimiento, proyecciones inadecuadas o presencia de material de osteosíntesis previo que altere la arquitectura original.

- *Anonimización:* todas las imágenes fueron despojadas de metadatos de identificación personal (*DICOM tags*) para garantizar el cumplimiento de los principios éticos y la confidencialidad del paciente.

Herramienta de inteligencia artificial

Para el procesamiento computacional se empleó el modelo de lenguaje extenso (*LLM*) multimodal Gemini 3 Flash. El procesamiento de las imágenes y la recolección de los diagnósticos emitidos por la inteligencia artificial se llevaron a cabo a través de su interfaz web estándar, con fecha de acceso entre el 17 y el 21 de abril de 2026.

Cabe destacar que no se utilizó acceso mediante API ni se aplicaron ajustes finos (*fine-tuning*) específicos para trauma ortopédico. A diferencia de las redes neuronales convolucionales (*CNN*) convencionales, que operan como sistemas de clasificación cerrada basados en píxeles, este modelo utiliza una arquitectura de *transformer* multimodal. Esta tecnología permite integrar la visión computacional con capacidades de razonamiento clínico y explicabilidad (*Explainable AI*). Mediante un proceso de interrogatorio dinámico (*prompting*), el modelo emula el proceso cognitivo del cirujano al aplicar la lógica jerárquica del sistema AO/OTA, permitiendo no solo la detección del trazo, sino su categorización según estabilidad y morfología.

Grupos de evaluación

La muestra fue evaluada de forma ciega e independiente por cinco niveles de observación:

1. *Grupo inteligencia artificial:* modelo Gemini 3 Flash.

2. *Grupo residentes (R1 a R4):* médicos residentes del postgrado de la UDAOT, subdivididos según su año de formación académica.

3. *Gold standard:* un panel de dos especialistas adjuntos con más de 10 años de experiencia en trauma ortopédico. En caso de discrepancia entre los especialistas, un tercer especialista senior independiente tomó la decisión que actuó como dictamen definitivo.

Procedimiento y protocolo de clasificación

- *Protocolo de la inteligencia artificial:* se presentó cada imagen mediante un *prompt* estandarizado:

"Analice la siguiente radiografía buscando activamente trazos de fractura. Identifique la región anatómica, el segmento y asigne la clasificación AO/OTA detallada (Número-Letra-Número)".

- *Protocolo humano:* los residentes evaluaron las imágenes en monitores de alta resolución, de forma individual, sin límite de tiempo y sin acceso a herramientas de consulta externa.

- *Sistema de clasificación:* se utilizó estrictamente la versión actualizada del compendio de clasificación de fracturas AO/OTA.

Recolección y análisis estadístico

Los datos obtenidos de la evaluación radiográfica de cada grupo (Inteligencia artificial y residentes R1-R4) fueron tabulados en una matriz de recolección de datos diseñada para el estudio. El procesamiento y análisis estadístico se realizó mediante el software IBM SPSS Statistics versión 29.0, empleando los siguientes criterios:

- *Análisis descriptivo:* se utilizaron frecuencias absolutas (n) y porcentajes (%) para describir las características de la muestra radiográfica, incluyendo el hallazgo clínico (presencia o ausencia de fractura), la región anatómica y la clasificación morfológica inicial (articular vs. extraarticular).

- *Análisis de precisión diagnóstica:* se calculó la sensibilidad, especificidad y precisión global de la inteligencia artificial y de cada año de residencia, tomando como referencia (*Gold standard*) el diagnóstico definitivo de dos especialistas en ortopedia y traumatología con más de 10 años de experiencia.

- *Análisis de concordancia:* se determinó el grado de acuerdo interobservador mediante el índice de Kappa de Cohen. Cabe destacar que este análisis se calculó tomando en cuenta la clasificación AO/OTA completa y detallada (incluyendo segmento, grupo y subgrupo morfológico), y no de manera dicotómica (presencia o ausencia de fractura). Para otorgar validez científica a la fuerza de concordancia, cada valor de Kappa fue reportado junto con su error estándar (EE) y su intervalo de confianza al 95% (IC 95%).

- *Significancia estadística:* se calculó el valor p para cada cruce de variables, estableciendo un nivel de significancia de ($p < 0,05$) para determinar que los resultados de concordancia no fueron debidos al azar.

RESULTADOS

La muestra de estudio se analizó mediante una distribución de frecuencias y porcentajes, compuesta por un total de 50 estudios radiográficos. En cuanto al hallazgo clínico inicial, el 92% (n=46) de las imágenes correspondieron a casos patológicos con presencia de fracturas, mientras que el 8% (n=4 restantes) fueron

estudios radiográficamente sanos, utilizados como control de especificidad diagnóstica (tabla 1).

Respecto a la región anatómica, la muestra incluyó una distribución representativa de los diversos segmentos óseos (proximal, diafisario y distal) de distintas regiones anatómicas, permitiendo evaluar la versatilidad de la inteligencia artificial multimodal en distintas zonas de carga y morfología. Finalmente, en la caracterización morfológica de las lesiones, se observó un predominio de fracturas de trazo articular (n=24) sobre las extraarticulares (n=22), lo que constituye una base de datos diversa para la validación de la fiabilidad diagnóstica bajo los criterios internacionales de la clasificación AO/OTA (tabla 1).

Tabla 1.

Caracterización general y distribución anatómico-morfológica de la muestra radiográfica

Variable de estudio	Categoría	Frecuencia (n)	Porcentaje (%)
Hallazgo clínico	Patológico	46	92,0
	Normal (Sano)	4	8,0
Región anatómica	Brazo	4	8.0
	Codo	3	6,0
	Antebrazo	3	6,0
	Muñeca	4	8,0
	Pelvis / Cadera	16	32,0
	Muslo	5	10,0
	Rodilla	5	10,0
	Pierna	5	10,0
	Tobillo	5	10,0
Patología (Fx)	Extraarticulares	22	47,8
	Articulares	24	52,2

Nota: muestra no probabilística de 50 estudios radiográficos evaluados para determinar la presencia, localización anatómica y patrón articular de las lesiones óseas.

En la tabla 2 se presenta el rendimiento diagnóstico detallado de cada grupo de estudio frente al estándar de oro. Se observa una correlación directa entre el nivel de experiencia clínica y el incremento en la precisión global, la cual inicia en un 50% para el grupo R1 y alcanza su punto máximo en el grupo R4 con un 92%.

Respecto al comportamiento de la inteligencia artificial multimodal (Gemini 3 Flash), el modelo alcanzó una precisión global del 88%, derivada de 44 aciertos totales sobre 50 casos, posicionándose por encima del rendimiento del grupo de residentes de tercer año y consolidándose como el evaluador con la mayor sensibilidad del estudio al identificar y clasificar correctamente 44 de las 46 radiografías patológicas (95,7%) (tabla 2).

No obstante, el análisis cualitativo y de especificidad de los 6 errores cometidos por el modelo reveló un contraste metodológico crítico: mientras que los residentes mostraron una capacidad de discriminación de estudios sanos proporcional a su experiencia quirúrgica, la inteligencia artificial presentó una especificidad del 0%, al incurrir en falsos positivos en la totalidad de los controles sanos (4 casos), donde el algoritmo "alucinó" trazos de fractura e inventó una clasificación AO/OTA detallada. Este hallazgo refleja un sesgo de

sobrediagnóstico visual que compromete su capacidad de exclusión de patología (tabla 2).

Tabla 2.

Rendimiento y precisión diagnóstica de la inteligencia artificial multimodal frente a médicos residentes por año de formación

Evaluador	Sensibilidad (%)	Especificidad (%)	Precisión global (%)	Aciertos totales (n/50)
IA	95.7	0.00	88.0	44/50
Primer año (R1)	52.2	25.0	50.0	25/50
Segundo año (R2)	73.9	50.0	72.0	36/50
Tercer año (R3)	84.8	75.0	84.0	42/50
Cuarto año (R4)	91.3	100.0	92.0	46/50

Nota: Los valores de rendimiento diagnóstico son matemáticamente consistentes con la distribución de la muestra (46 estudios patológicos y 4 sanos). El modelo identificó correctamente 44 de 46 fracturas (Sensibilidad: 44/46 = 95,7%) y clasificó erróneamente los 4 casos sanos como fracturas (Especificidad: 0/4 = 0%). En consecuencia, los aciertos totales del modelo fueron 44, lo que resulta en una precisión global del 88% (44 aciertos / 50 casos totales).

En la tabla 3 el análisis de concordancia interobservador respecto al patrón de oro se determinó mediante el coeficiente Kappa de Cohen (κ), cuyos resultados reflejaron una progresión estadística consistente con el nivel de formación quirúrgica y la experiencia clínica. El grupo de residentes de primer año (R1) presentó un índice $\kappa = 0.08$, lo que corresponde a una concordancia insignificante o pobre. La falta de significancia estadística ($p > 0.05$) indica que la tasa de aciertos observada en este grupo (50.0%) no difiere de la esperada por puro azar. A medida que se incrementa el año de residencia, el grado de acuerdo mostró un ascenso sostenido y estadísticamente significativo ($p < 0.001$), alcanzando concordancias moderada (R2), sustancial (R3) y casi perfecto (R4) (tabla 3).

Por su parte, la inteligencia artificial multimodal obtuvo un índice $\kappa=0.79$, situándose en el rango superior de la categoría de concordancia sustancial. La fiabilidad diagnóstica del modelo computacional superó ampliamente a los niveles de residencia R1, R2 y R3, ubicándose en una posición intermedia-alta muy cercana al rendimiento del grupo de residentes de cuarto año (R4). Asimismo, la reducción progresiva del error estándar (EE) desde la inteligencia artificial hacia los R4 confirma una estabilidad diagnóstica en los evaluadores de alto rendimiento (tabla 3).

La figura 1 ilustra la curva de aprendizaje obtenida al contrastar el porcentaje de precisión diagnóstica con el nivel de experiencia de los grupos evaluadores. Se observa una pendiente ascendente pronunciada entre los niveles de residencia iniciales (R1: 50% y R2: 72%), lo que representa la fase de adquisición acelerada de competencias en la clasificación AO/OTA. El punto de inflexión más relevante de la gráfica se sitúa en el rendimiento de la inteligencia artificial multimodal (88%), la cual se posiciona por encima de la media de los residentes de tercer año (R3: 84%). Este hallazgo visual demuestra que la herramienta tecnológica logra insertarse en el segmento de alta competencia de la curva, aproximándose al 'techo' de precisión establecido por el residente de cuarto año (92%).

Esta progresión sugiere que, mientras la curva humana depende de años de exposición clínica y formación quirúrgica, la inteligencia artificial ofrece un estándar de fiabilidad inmediato y comparable al de un residente senior, consolidándose como un recurso de apoyo que podría reducir la brecha de error en los niveles de formación inicial (R1 y R2) (figura 1).

Tabla 3.

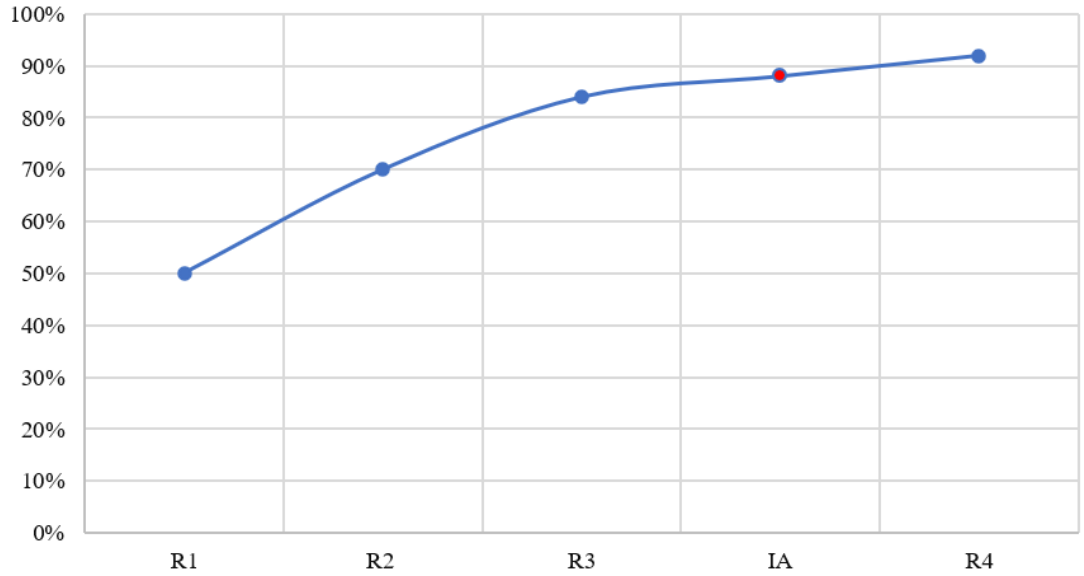
Análisis de concordancia interobservador mediante el índice Kappa de Cohen para la clasificación AO/OTA

Grupo evaluado	Índice kappa (κ)	Error estándar (EE)	Grado de acuerdo	Intervalo de confianza (95%)	Valor p
Primer año (R1)	0.08	0.085	Insignificante / pobre	-0.088 – 0.248	> 0.05
Segundo año (R2)	0.50	0.102	Moderado	0.300 – 0.700	< 0.001
Tercer año (R3)	0.72	0.088	Sustancial	0.548 – 0.892	< 0.001
IA	0.79	0.076	Sustancial	0.641 – 0.939	< 0.001
Cuarto año (R4)	0.86	0.062	Casi perfecto	0.738 – 0.982	< 0.001

Nota: EE: Error estándar; IC 95%: Intervalo de confianza al 95%. Significancia estadística establecida en p < 0,05. Grado de acuerdo según Landis y Koch.

Figura 1.

Comparativa de la precisión diagnóstica global en la curva de aprendizaje médica frente al modelo de inteligencia artificial



Nota: muestra la progresión de la precisión diagnóstica (%) entre los niveles de residencia (R1-R4) y el posicionamiento relativo alcanzado por la IA Gemini 3 Flash (88%). La curva de aprendizaje médica se construyó graficando los valores de precisión global (eje Y) en función del nivel de formación académica (eje X), uniendo los puntos de datos de cada subgrupo mediante una línea de suavizado para ilustrar la tendencia de mejora continua. El rendimiento del modelo de IA se representa como un punto de referencia transversal e independiente de la variable de experiencia temporal.

DISCUSIÓN

El análisis de la fiabilidad diagnóstica entre la inteligencia artificial multimodal (*LLM*) y el juicio clínico de los residentes revela una dinámica de rendimiento que trasciende la precisión estadística convencional. El índice Kappa de 0.79 obtenido por la inteligencia artificial la posiciona en una categoría de concordancia "sustancial" según la escala de Landis y Koch (Landis & Koch, 1977), superando la capacidad diagnóstica de los residentes R1, R2 y R3. Este hallazgo es consistente con la literatura actual, donde se ha demostrado que los algoritmos de aprendizaje profundo pueden igualar o incluso superar al cirujano ortopédico en la clasificación de fracturas óseas (Olczak et al., 2017).

No obstante, la implementación del sistema AO/OTA exige una comprensión morfológica que va más allá de la detección visual, requiriendo una reproducibilidad que históricamente ha sido un desafío analítico, incluso para observadores humanos experimentados en entornos de urgencia (Audigé et al., 2005; Meinberg et al., 2018; Müller et al., 1990).

A pesar de alcanzar una precisión global del 88%, un análisis cualitativo detallado de los 6 errores cometidos por el modelo de inteligencia artificial revela un impacto clínico sustancial. El 66.6% de estas fallas se concentró en la región del anillo pélvico y la cadera, segmento que representaba el 32% de la muestra global. Mientras que los errores cometidos por los residentes senior (R4) correspondieron a discrepancias menores en el subgrupo morfológico, los fallos de la inteligencia artificial en la pelvis implicaron la no detección de trazos de fractura o la "alucinación" en radiografías sanas con la consiguiente confusión catastrófica entre patrones articulares complejos y estructuras anatómicas superpuestas.

Desde una perspectiva anatómica y metodológica, la alta concentración de fallas en el anillo pélvico y la cadera responde a limitaciones intrínsecas de la radiografía convencional en esta región. A diferencia de los huesos largos, la pelvis presenta una arquitectura tridimensional sumamente compleja caracterizada por la superposición de múltiples estructuras óseas densas, tales como las ramas ilio e isquiopúbicas, el reborde acetabular, las alas ilíacas y el sacro. La superposición de estas sombras óseas normales genera patrones visuales que los modelos de lenguaje multimodal (*LLM*), carentes de un procesamiento volumétrico o espacial real, interpretan erróneamente. Al verse impelidos por su diseño taxonómico a forzar una clasificación dentro de la nomenclatura AO/OTA, los modelos transforman las complejidades de la superposición pélvica en falsos positivos estructurales o en la omisión de trazos sutiles.

En la práctica quirúrgica, un error de esta naturaleza por omisión o mala clasificación cambiaría radicalmente la conducta terapéutica, pudiendo pasar de manejos conservadores a intervenciones invasivas no requeridas. Esto demuestra que la gravedad cualitativa de las "alucinaciones" visuales en regiones de anatomía tridimensional superpuesta constituye un riesgo clínico mayor que las imprecisiones conceptuales observadas en los médicos residentes en formación.

Desde una perspectiva clínica, el impacto de estos errores difiere sustancialmente entre el personal en formación

y la herramienta automatizada. Mientras que las imprecisiones de los residentes junior se circunscriben a discrepancias menores en la codificación del subgrupo, los errores de la inteligencia conllevan un riesgo real de iatrogenia diagnóstica. En un entorno de urgencias traumatológicas, aceptar sin verificar un falso positivo pélvico o articular por parte del médico de guardia podría derivar en la indicación innecesaria de estudios complementarios complejos, interconsultas innecesarias o, en escenarios extremos, en planteamientos de intervenciones quirúrgicas invasivas no justificadas. Por lo tanto, la nula especificidad del modelo subraya que, a pesar de su alta sensibilidad como herramienta de triaje, la inteligencia artificial carece del criterio de exclusión necesario para operar con autonomía clínica.

Un hallazgo metodológico fundamental de este estudio radica en el contraste extremo entre la alta sensibilidad del modelo (95.7%) y su nula especificidad (0%). En la práctica clínica y epidemiológica, una prueba diagnóstica con alta sensibilidad es excelente para excluir enfermedad cuando resulta negativa; sin embargo, la incapacidad absoluta del *LLM* para reconocer estudios radiográficos normales, evidencia una limitación cognitiva profunda de los modelos de propósito general.

A diferencia del ojo humano entrenado, que pondera la continuidad cortical y la integridad anatómica global para certificar la normalidad, el algoritmo se ve impelido por su arquitectura a 'encontrar' patrones dentro de la nomenclatura AO/OTA, transformando la superposición de sombras óseas normales en falsos positivos morfológicos. Esto muestra que la optimización de la inteligencia artificial en traumatología no debe enfocarse únicamente en reducir los falsos negativos por omisión, sino de manera prioritaria en desarrollar umbrales de decisión que permitan la validación certera de la normalidad esquelética.

Es imperativo precisar que, a diferencia de las redes neuronales convolucionales (*CNN*) tradicionales, este estudio introduce un paradigma innovador al emplear un modelo de lenguaje de gran escala (*LLM*) con capacidades multimodales (Moor et al., 2024; Thirunavukarasu et al., 2023). Mientras que una *CNN* opera como un clasificador cerrado incapaz de justificar su salida, la arquitectura multimodal utilizada es capaz de articular un razonamiento clínico semántico, integrar la compleja taxonomía jerárquica de la clasificación AO/OTA y procesar jerarquías lógicas complejas (Singhal et al., 2023).

Sin embargo, esta misma naturaleza generativa que simula el pensamiento médico puede introducir "alucinaciones" o errores de razonamiento espacial. En nuestra serie, aunque la inteligencia artificial alcanzó una precisión del 88%, la gravedad de sus errores en regiones de alta complejidad evidenció que el modelo carece aún de la comprensión tridimensional del trazo y la estabilidad mecánica que demostraron los residentes senior (R4), configurando un riesgo de iatrogenia que se debate en la convergencia actual entre la inteligencia humana y la artificial (Topol, 2019).

La transición desde las redes neuronales convolucionales *(CNN*) convencionales hacia los modelos de lenguaje de gran escala (*LLM*) multimodales marca un cambio de paradigma en el diagnóstico por imagen. Las *CNN* operan sobre descriptores locales de textura y bordes a nivel de píxel, destacando en patrones geométricos cerrados, pero careciendo de capacidad interpretativa clínica intrínseca. En contraste, los LLM multimodales

aportan un razonamiento semántico contextual que emula la lógica jerárquica de la clasificación AO/OTA, estructurando diagnósticos explicables. Sin embargo, este razonamiento semántico introduce un tipo de riesgo cualitativamente distinto al de las *CNN*. Mientras que las fallas en las CNN suelen derivarse de ruido en los píxeles o artefactos técnicos, en los *LLM* el sesgo radica en las "alucinaciones" de razonamiento espacial, donde la capacidad generativa del modelo 'inventa' continuidad en sombras anatómicas normales o malinterpreta superposiciones tridimensionales. Por lo tanto, aunque la ventaja semántica del *LLM* enriquece el reporte clínico, su tendencia a generar falsos positivos estructurados representa un riesgo diagnóstico emergente que exige supervisión estricta.

Esta disparidad subraya que la superioridad estadística no debe traducirse en autonomía diagnóstica absoluta en los servicios de traumatología. La alta sensibilidad demostrada por el modelo *LLM* lo convierte en una herramienta de triaje excepcional para la detección inicial de lesiones estructurales en las guardias, actuando como un "nivelador de competencias" que optimiza los flujos de trabajo para el personal en formación (Chen et al., 2022). No obstante, en la clasificación definitiva de fracturas inestables o de morfología atípica, la inteligencia artificial aún carece de la precisión diagnóstica necesaria para sustituir el entrenamiento formal (Langerhuizen et al., 2020).

Como indicador complementario de rendimiento diagnóstico, se evaluó la aplicabilidad del Área Bajo la Curva *ROC* (*AUC-ROC*). No obstante, dado que el diseño del estudio estuvo orientado primariamente a la categorización morfológica alfanumérica de fracturas y no a un escrutinio binario masivo, la representación mediante curvas *ROC* convencionales resulta metodológicamente limitada. Si bien la sensibilidad del 95.7% refleja una elevada capacidad de detección de patología, la especificidad del 0% observada en el grupo de control sano (n=4) colapsa la discriminación del punto de corte óptimo. Por este motivo, y en concordancia con las recomendaciones evaluativas para taxones jerárquicos multinivel como el AO/OTA, el coeficiente Kappa de Cohen ponderado constituye el indicador ordinal de concordancia más robusto para este análisis, reservándose la modelación *ROC* multiclase con remuestreo para subsiguientes cohortes con grupos sanos balanceados.

Por ende, los resultados sugieren que la inteligencia artificial multimodal debe implementarse estrictamente como un sistema de soporte y no como un sustituto en el postgrado de la UDAOT. El modelo complementa el juicio clínico del traumatólogo, advirtiendo que sus errores, aunque numéricamente escasos, pueden ser cualitativamente más graves en patologías que exigen un entendimiento anatómico profundo que los modelos actuales aún no terminan de emular con total fidelidad.

Este estudio exploratorio debe interpretarse a la luz de diversas limitaciones metodológicas y técnicas. En primer lugar, la muestra de 50 estudios radiográficos, aunque diversa en cuanto a segmentos anatómicos, fue seleccionada mediante un muestreo no probabilístico en un único centro asistencial (IAHULA), lo que limita la representatividad estadística y la generalización de los resultados a otros contextos hospitalarios. En segundo lugar, la evaluación se restringió exclusivamente al análisis de radiografías digitales en proyecciones ortogonales. Esto privó tanto a la inteligencia artificial como a los evaluadores humanos de herramientas de

corte tomográfico (TAC) o reconstrucción tridimensional, fundamentales para categorizar fracturas articulares o pélvicas de alta complejidad.

En tercer lugar, el estudio carece de una validación externa con cohortes multicéntricas que permita probar la robustez del algoritmo ante variaciones en calidad técnica o calibración de equipos radiológicos heterogéneos. En cuarto lugar, el diseño no contempló la evaluación de la reproducibilidad intraobservador, por lo que se desconoce la variabilidad diagnóstica de un mismo evaluador (tanto humano como de la inteligencia artificial) al analizar la misma imagen en diferentes momentos. Asimismo, al ser un estudio enfocado en la precisión diagnóstica inicial, no se evaluó el impacto directo de esta herramienta en la toma de decisiones clínicas reales ni en los resultados terapéuticos de los pacientes. Finalmente, desde la perspectiva del modelo informático, el uso de un *LLM* de propósito general (Gemini 3 Flash) plantea sesgos inherentes derivados de la ingeniería de *prompts* y de la falta de un entrenamiento primario exclusivo sobre bases de datos traumatológicas masivas, lo que favorece la aparición de "alucinaciones" visuales cuando el modelo enfrenta superposiciones complejas.

CONCLUSIONES

La inteligencia artificial multimodal demostró una fiabilidad diagnóstica (Kappa de 0,79; precisión global del 88%) superior a la de los residentes de primero a tercer año (R1-R3), logrando una paridad operativa con los residentes de cuarto año (92%). Su mínima tasa de falsos negativos (2%) la posiciona como un "nivelador de competencias" capaz de reducir la brecha de error diagnóstico en el personal en formación inicial. Sin embargo, el modelo presentó una nula capacidad para descartar patología en estudios sanos.

Asimismo, la IA evidenció limitaciones críticas en la interpretación espacial de regiones con alta complejidad anatómica (como la pelvis y superficies articulares). En estos escenarios, sus errores fueron cualitativamente más graves que los observados por los residentes, confirmando que la herramienta aún carece de un análisis lógico-quirúrgico tridimensional profundo.

En consecuencia, el modelo no sustituye el juicio médico y debe emplearse como sistema de soporte de doble lectura bajo supervisión obligatoria. Su integración estratégica en los servicios de traumatología de la región representaría una herramienta clave para optimizar la seguridad clínica y el triaje en las salas de urgencias, respaldando al personal frente a la alta demanda asistencial sin prescindir del criterio humano final.

RECOMENDACIONES

A la luz de los hallazgos obtenidos, se recomienda la integración de modelos de inteligencia artificial multimodal en los servicios de urgencias traumatológicas como un sistema de soporte a la decisión clínica. Su implementación primaria debe orientarse a servir como una herramienta de "doble lectura" para los residentes de niveles iniciales (R1 y R2), con el objetivo de mitigar la brecha de error diagnóstico y reducir la tasa de omisión en fracturas agudas. No obstante, dada la naturaleza de los errores cualitativos observados en regiones de alta complejidad anatómica, es imperativo establecer protocolos de supervisión especializada obligatoria.

Toda clasificación alfanumérica AO/OTA generada por un modelo de lenguaje de gran escala (*LLM*) debe ser validada por un especialista o un residente senior antes de definir una conducta quirúrgica definitiva, garantizando que el juicio humano prevalezca en fracturas articulares o de pelvis.

Asimismo, se propone la actualización de los currículos de postgrado en ortopedia y traumatología para incluir competencias en salud digital e inteligencia artificial aplicada. El médico en formación debe desarrollar un pensamiento crítico que le permita identificar las limitaciones técnicas y las posibles "alucinaciones" visuales de los modelos multimodales, evitando una dependencia tecnológica ciega. Para futuras líneas de investigación, se sugiere ampliar la casuística mediante estudios multicéntricos que incluyan un mayor volumen de fracturas complejas y el uso de tomografía axial computarizada (TAC) como método comparativo. Esto permitiría evaluar si la precisión de la inteligencia artificial mejora en entornos tridimensionales, superando las limitaciones intrínsecas de la radiografía convencional en la clasificación morfológica detallada.

CONFLICTOS DE INTERÉS

Los autores declaran no tener ningún conflicto de interés.

DECLARACIÓN DE USO DE INTELIGENCIA ARTIFICIAL

Los autores declaran que durante la preparación de este trabajo se utilizó el modelo de lenguaje multimodal *Gemini 3 Flash* exclusivamente como herramienta de análisis comparativo y procesamiento experimental dentro del diseño metodológico del estudio, evaluando su rendimiento diagnóstico frente al juicio clínico de los médicos residentes en la clasificación de fracturas AO/OTA. El modelo de inteligencia artificial no participó en la concepción del estudio, el diseño de la investigación, la recolección de los datos clínicos de referencia (*Gold Standard*), ni en la redacción, revisión o edición intelectual del texto final del manuscrito. Los autores asumen la responsabilidad plena y exclusiva de la integridad, precisión y originalidad del contenido publicado en este artículo.

REFERENCIAS

Audigé, L., Bhandari, M., Hanson, B., & Kellam, J. (2005). A concept for the validation of fracture classifications. *Clinical Orthopaedics and Related Research, 430*, 182–190. https://doi.org/10.1097/01.bot.0000155310.04886.37

Badgeley, M. A., Zech, J. R., Oakden-Rayner, L., Glicksberg, B. S., Liu, M., Gale, W., McConnell, M. V., Percha, B., Snyder, T. M., & Dudley, J. T. (2019). Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digital Medicine, 2*, 31, 1-10. https://doi.org/10.1038/s41746-019-0105-1

Chen, K., Stotter, C., Klestil, T., & Nehrer, S. (2022). Artificial intelligence in orthopedic radiography analysis: A narrative review. *Diagnostics, 12*(9), 2235. https://doi.org/10.3390/diagnostics12092235

Kooistra, B. W., Dijkstra, S., Diercks, R. L., & Goslings, J. C. (2011). Reliability of the AO classification for distal radius fractures. *Journal of Orthopaedic Trauma, 25*(7), 409–415. https://doi.org/10.1097/BOT.0b013e3181f21503

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Langerhuizen, D. W. G., Janssen, S. J., Mallee, W. H., van den Bekerom, M. P. J., Doornberg, J. N., Ring, D.,

& Schupte, M. J. (2019). What are the applications and limitations of artificial intelligence for fracture detection and classification in orthopaedic trauma imaging? A systematic review. *Clinical Orthopaedics and Related Research, 477*(11), 2482–2491. https://doi.org/10.1097/CORR.0000000000000848

Langerhuizen, D. W. G., Schupte, M. J., & Doornberg, J. N. (2020). Is deep learning as good as the orthopedic surgeon in classifying distal radius fractures? *Archives of Orthopaedic and Trauma Surgery, 140*(12), 2021–2027. https://doi.org/10.1007/s00402-020-03539-4

Meinberg, E. G., Agel, J., Roberts, C. S., Karam, M. D., & Kellam, J. F. (2018). Fracture and dislocation classification compendium: 2018. *Journal of Orthopaedic Trauma, 32*(Suppl 1), S1–S170. https://doi.org/10.1097/BOT.0000000000001063

Moor, C., Banerjee, O., Abad, Z. S., Kulkarni, N., Reyes, N., Singh, P., Zhang, Y., & Shetty, S. (2024). Med-Gemini: High-capability multimodal models for medicine (arXiv:2404.18416). arXiv. https://doi.org/10.48550/arXiv.2404.18416

Müller, M. E., Nazarian, S., Koch, P., & Schatzker, J. (1990). The comprehensive classification of fractures of long bones. *Springer-Verlag.* https://doi.org/10.1007/978-3-642-75435-7

Olczak, J., Fahlberg, N., Behnam, S., Forghani, B., Kiesel B., Wedin, R., & Gordon, M. (2017). Deep learning, a form of artificial intelligence, open a new window for analysis of musculoskeletal radiographs. *Acta Orthopaedica, 88*(6), 581–586. https://doi.org/10.1080/17453674.2017.1368812

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tan, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamwa, P., Schärli, N., Fisher, A., Bebensee, A., Drucker, J., Schupmann, C., Cole, A., ... Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature, 620*(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Ting, L., & Carin, L. (2023). Large language models in medicine. *Nature Medicine, 29*(8), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7