



Aprendizaje profundo como herramienta de apoyo a la decisión en selección embrionaria para FIV: revisión de la literatura

Deep learning as a decision support tool for embryo selection in IVF: A literature review

Areli Monserrat Fuentes Trejo ✉

Universidad Autónoma del Estado Hidalgo, Pachuca, Hidalgo, México

fu466648@uaeh.edu.mx

ORCID: 0009-0002-7795-6257

Silvia Soledad Moreno Gutiérrez

Universidad Autónoma del Estado Hidalgo, Pachuca, Hidalgo, México

silviam@uaeh.edu.mx

ORCID: 0000-0002-8957-3707



<https://doi.org/10.36825/RITI.14.34.003>

Recibido: Mayo 20, 2026

Aceptado: Agosto 10, 2026

Resumen: Esta revisión se sitúa en el campo de las tecnologías de la información aplicadas a la salud, con enfoque en el desarrollo y evaluación de modelos de aprendizaje profundo como sistemas de apoyo a la decisión clínica. El objetivo del trabajo consiste en revisar la evidencia científica disponible de modelos de aprendizaje profundo aplicados a la selección de embriones humanos en la fecundación *in vitro*, para brindar un panorama a los lectores mediante la revisión de artículos publicados en revistas de alto impacto. Se realizó una búsqueda en bases de datos indexadas en Scopus, PubMed y Web of Science, publicados del año 2021 a 2026. La estrategia de búsqueda se llevó a cabo mediante palabras clave como 'fecundación *in vitro*', 'selección embrionaria', 'aprendizaje profundo' e 'inteligencia artificial'. Se identificaron 45 artículos potencialmente relevantes, de los cuales 11 cumplieron los criterios de inclusión por lo que fueron analizados e integrados en la síntesis final priorizando estudios que reportaron métricas de desempeño. La heterogeneidad entre los estudios limitó la comparabilidad de los resultados, no obstante, los modelos inteligentes demuestran potencial para evaluar embriones y predecir implantación, su rendimiento es similar o superior a los métodos convencionales. La literatura expone diversos estudios, sin embargo, la predicción clínica fiable aún requiere profundizar en los modelos integrados y mejor validación.

Palabras clave: *Aprendizaje Profundo, Selección de Embriones, Implantación, Fecundación In Vitro, Inteligencia Artificial.*

Abstract: This review is situated within the field of health information technologies applied to healthcare, with a focus on the development and evaluation of deep learning models as clinical decision support systems. The aim of this work is to review the available scientific evidence on deep learning models applied to the selection of human embryos for *in vitro* fertilization, in order to provide an overview for readers through the analysis of articles published in high-impact journals. A search was conducted in indexed databases, including Scopus, PubMed, and Web of Science, covering publications from 2021 to 2026. The search strategy employed keywords such as '*in*

vitro fertilization', 'embryo selection', 'deep learning', and 'artificial intelligence'. A total of 45 potentially relevant articles were identified, of which 11 met the inclusion criteria and were therefore analyzed and incorporated into the final synthesis, prioritizing studies that reported performance metrics. Heterogeneity among studies limited the comparability of results; nevertheless, intelligent models demonstrate potential for embryo evaluation and implantation prediction, with performance comparable or superior to conventional methods. The literature presents various studies; however, reliable clinical prediction still requires further development of integrated models and improved validation strategies.

Keywords: *Deep Learning, Embryo Selection, Implantation, In Vitro Fertilization, Artificial Intelligence.*

1. Introducción

La integración de modelos de inteligencia artificial (IA) en los sistemas de información en salud es una de las principales tendencias en el desarrollo de tecnologías aplicadas al ámbito clínico [1]. La IA se refiere a la capacidad de los sistemas computacionales para ejecutar tareas que convencionalmente demandan inteligencia humana, tales como el aprendizaje a partir de datos, el razonamiento, el reconocimiento de patrones complejos, la toma de decisiones y la resolución de problemas en entornos clínicos [1]. Dentro de este campo, el aprendizaje profundo (DL) se ha posicionado como una herramienta con alto potencial para el desarrollo de sistemas de apoyo a la decisión clínica, debido a su capacidad para procesar grandes volúmenes de datos e identificar patrones complejos. El DL es una rama del aprendizaje automático fundamentada en redes neuronales artificiales (RNA) con múltiples capas jerárquicas de procesamiento, que permiten a los sistemas computacionales aprender de manera automática representaciones progresivamente más abstractas y complejas a partir de grandes volúmenes de datos [2].

La Organización Mundial de la Salud (OMS) considera la infertilidad un problema de salud pública que afecta al 17.5% de la población a nivel global [3]. La infertilidad se define como la incapacidad de lograr un embarazo después de un año de relaciones sexuales estables sin protección [4], factores médicos y la edad materna son responsables de incrementarla [4]. El creciente número de pacientes con dificultades reproductivas ha incrementado la demanda de tratamientos de reproducción asistida (ART), los cuales comprenden procedimientos en los que ovocitos o embriones son manipulados fuera del organismo femenino con el objetivo de lograr un embarazo exitoso [5].

Entre ellos, la fecundación *in vitro* (FIV) constituye la técnica de mayor aplicación [6]. El término *in vitro* hace referencia a que la fecundación de los ovocitos se realiza fuera del organismo, en condiciones de laboratorio, mientras que la maduración de los ovocitos y el desarrollo posterior del embrión ocurren de forma natural en el ovario y el útero, respectivamente [6]. No obstante, a pesar de su amplio uso, la probabilidad de lograr una gestación continúa siendo limitada [7]. El éxito de la FIV se mide principalmente mediante la tasa de nacimientos vivos por ciclo. Esta tasa se reporta insuficiente, entre el 20 y 30% [7], y puede variar dependiendo de factores como la edad, la calidad de los gametos y las condiciones clínicas del paciente. El proceso de la FIV es largo, invasivo, costoso y con resultados inciertos, por tanto, es causa de estrés, ansiedad, frustración y desgaste psicológico [8].

La selección de embriones se basa en los criterios morfológicos por parte de los embriólogos, este tipo de revisión es subjetiva, existe limitación visual y depende de la experiencia del especialista [9]. La selección precisa es un factor clave para mejorar las tasas de éxito en ART [10]. El reto es conseguir métodos de selección fiables, por ello, la ART ha considerado la aplicación de técnicas de IA en la selección de embriones con potencial de implantación [11]. En el ámbito, la IA adquiere mayor aceptación para fines diagnósticos [1]. En particular, la aplicación de las redes RNA en la FIV está revolucionando la evaluación de viabilidad embrionaria [12].

El uso creciente de incubadores con tecnología *Time-Lapse Imaging* (TLI) ha generado grandes volúmenes de datos embrionarios. En este contexto, las herramientas de IA han comenzado a utilizarse como apoyo para la evaluación objetiva por parte de los especialistas [13]. La combinación de TLI y modelos de DL son una alternativa prometedora para optimizar la selección embrionaria en FIV [14].

Entre los modelos de DL, las redes neuronales convolucionales (CNN) son las más utilizadas en este contexto. Las CNN son modelos de aprendizaje profundo diseñados para aprender de forma automática y adaptativa jerarquías espaciales de características a partir de datos de entrada, en particular imágenes, utilizando capas convolucionales [15].

La literatura sobre IA aplicada a selección embrionaria ha crecido considerablemente en los últimos años, los estudios basados específicamente en modelos de DL presentan diferencias importantes en sus enfoques metodológicos, tipos de datos, métricas de desempeño, estrategias de entrenamiento y objetivos clínicos evaluados. Esta heterogeneidad dificulta obtener una visión integral del estado actual del área e identificar las tendencias, fortalezas y limitaciones de los enfoques propuestos. Por lo que el objetivo de la presente revisión de la literatura es analizar la evidencia científica disponible sobre modelos de DL aplicados a la selección de embriones humanos en FIV, con el fin de proporcionar un panorama integral del estado actual del campo que oriente a futuros investigadores en el desarrollo de nuevos modelos y estudios.

2. Materiales y métodos

2.1. Diseño metodológico

En esta revisión de la literatura se adoptó un enfoque metodológico estructurado para identificar, seleccionar y analizar estudios relacionados con la aplicación de modelos de DL en la selección embrionaria durante procedimientos de FIV. El proceso de búsqueda y selección tomó como referencia los lineamientos de la declaración PRISMA 2020 (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*), los cuales proporcionan un conjunto de recomendaciones para garantizar la transparencia, la completitud y la reproducibilidad de las revisiones de literatura [16]. Se describen las etapas de identificación, cribado, evaluación de elegibilidad e inclusión de los estudios.

2.2. Estrategia de búsqueda

Se realizó una búsqueda en las bases de datos Scopus, PubMed y Web of Science, considerando publicaciones comprendidas entre 2021 y 2026. La estrategia de búsqueda se diseñó mediante la combinación de términos relacionados con el aprendizaje profundo y la selección embrionaria. Las principales palabras clave empleadas se muestran en la Tabla 1 y las cadenas de búsqueda empleadas en cada base de datos se muestran en la Tabla 2.

Tabla 1. Palabras clave.

Concepto	Palabras clave
Inteligencia artificial	<i>Artificial Intelligence, AI</i>
Aprendizaje profundo	<i>Deep Learning, Neural Network, Convolutional, Neural Network</i>
Selección embrionaria	<i>Embryo Selection</i>
Fecundación <i>in vitro</i>	<i>In Vitro Fertilization, IVF</i>
Imagen	<i>Time-Lapse Imaging</i>

Fuente: Elaboración propia.

Tabla 2. Cadenas de búsqueda.

Base de datos	Cadena de búsqueda
Scopus	TITLE-ABS-KEY(("deep learning" OR "machine learning" OR "artificial intelligence" OR "neural network*" OR "convolutional neuronal network*" OR CNN) AND ("embryo selection" OR "embryo evaluation") AND ("in vitro fertilization" OR "IVF" OR "assisted reproductive technolog*") AND ("time-lapse"))
PubMed	("Deep Learning"[Mesh] OR "deep learning" OR "machine learning" OR "artificial intelligence") AND ("embryo selection") AND ("Fertilization in vitro" [Mesh] OR "in vitro fertilization" OR IVF)
Web of Science	TS= (("deep learning" OR "machine learning" OR "artificial intelligence") AND ("embryo selection") AND ("in vitro fertilization" OR IVF))

Fuente: Elaboración propia.

2.3. Criterios de inclusión y exclusión

La selección de los estudios se realizó con base en criterios previamente definidos, con el propósito de garantizar que los artículos incluidos fueran relevantes para el objetivo de la revisión. La Tabla 3 presenta los criterios de inclusión y exclusión aplicados durante el proceso de selección.

Tabla 3. Criterios de inclusión y exclusión

Criterios de inclusión	Criterios de exclusión
Estudios que aplicaran modelos de DL para evaluación, clasificación o selección embrionaria.	Estudios que no utilizaran modelos de DL.
Investigaciones que compararan modelos de DL con métodos convencionales de evaluación embrionaria.	Revisiones de literatura, editoriales y resúmenes de congresos sin resultados primarios cuantificables.
Estudios que reportaran métricas de desempeño (AUC, <i>accuracy</i> , sensibilidad, especificidad, etc.).	Estudios que no presentaran resultados de desempeño.
Estudios realizados con embriones humanos.	Estudios que no abordaran la selección embrionaria en FIV.
Estudios prospectivos y retrospectivos publicados en inglés.	Publicaciones cuyo resumen no describiera claramente objetivos, metodología, resultados y conclusiones.
Publicaciones entre 2021 y 2026.	Publicaciones fuera del periodo establecido.

Fuente: Elaboración propia.

Los estudios retrospectivos son un diseño de investigación que analiza datos o registros existentes recopilados en el pasado para investigar resultados, asociaciones o exposiciones, mientras que los estudios prospectivos son un diseño de investigación en el que se realiza un seguimiento de los participantes a lo largo del tiempo para evaluar la aparición de resultados en relación con exposiciones o intervenciones [17].

2.4. Proceso de selección de estudios

- **Identificación:** Como resultado del proceso de búsqueda y selección se identificaron 45 artículos potencialmente relevantes.
- **Cribado:** En esta etapa se eliminaron los registros duplicados y se llevó a cabo la revisión de los títulos y resúmenes de los estudios recuperados, con el propósito de descartar aquellos que no se alinearan con el objetivo de la revisión.
- **Elegibilidad:** Los artículos potencialmente relevantes fueron sometidos a una revisión del texto para verificar el cumplimiento de los criterios de inclusión y exclusión establecidos.
- **Inclusión:** Finalmente, se incorporaron al análisis únicamente los estudios que cumplieron con todos los criterios de inclusión establecidos. Como resultado del proceso de selección, 11 artículos conformaron el corpus final de la revisión.

2.5. Extracción de datos

De cada estudio incluido se extrajo información como el tipo de estudio, tamaño de la muestra, tipo de modelo de aprendizaje profundo, datos de entrada utilizados, problemática, métricas de desempeño reportadas y principales resultados. La síntesis de la información se realizó considerando dos dimensiones complementarias.

La primera correspondió al desempeño tecnológico de los modelos de aprendizaje profundo, mediante el análisis de métricas como el área bajo la curva (AUC), *accuracy*, sensibilidad y especificidad. La segunda dimensión se enfocó en la validación clínica, considerando estudios que evaluaron desenlaces como la implantación, el embarazo clínico y el nacimiento vivo, así como ensayos clínicos aleatorizados que compararon directamente la selección embrionaria asistida por modelos de DL frente a la evaluación morfológica convencional realizada por embriólogos. Este enfoque permitió integrar tanto el rendimiento computacional de los modelos como la evidencia disponible sobre su utilidad clínica en la selección embrionaria durante procedimientos de la FIV.

2.6. Evaluación de la calidad y pertinencia de los estudios

Con el propósito de fortalecer la selección de los estudios incluidos y garantizar la consistencia del análisis, se elaboró una lista de verificación basada en los objetivos de la presente revisión. La evaluación consideró la pertinencia temática del estudio respecto a la aplicación de modelos de DL en la selección de embriones humanos durante procedimientos de FIV, la claridad del objetivo de investigación, la descripción de los datos utilizados, la especificación del modelo o arquitectura, el reporte de métricas de desempeño, la descripción del procedimiento de entrenamiento y validación. Asimismo, se consideró la indexación de la revista en bases de datos científicas reconocidas y su relevancia dentro del área de investigación como indicadores de calidad de la publicación.

La evaluación permitió determinar la suficiencia metodológica y la pertinencia de los estudios para su incorporación al trabajo final, los trabajos fueron revisados de manera sistemática con el propósito de favorecer la consistencia y transparencia del proceso de selección y análisis.

Tabla 4. Criterios para la evaluación de la calidad de los estudios.

Estudio	Pertinencia	Objetivo	Datos	Modelo	Validación	Métricas	Puntaje	Nivel
1	Si	Si	Si	Si	Si	Si	7/7	Alto
2	Si	Si	Si	Si	Si	Si	7/7	Alto
3	Si	Si	Si	Si	Si	Si	7/7	Alto
4	Si	Si	Si	Si	Si	Si	6/7	Alto
5	Si	Si	Si	Si	Si	Si	7/7	Alto
6	Si	Si	Si	Si	Si	Si	7/7	Alto
7	Si	Si	Si	Si	Si	Si	7/7	Alto
8	Si	Si	Si	Si	Si	Si	7/7	Alto
9	No	Si	Si	Si	Si	Si	6/7	Alto
10	Si	Si	Si	Si	Si	Si	7/7	Alto
11	Si	Si	Si	Si	Si	Si	7/7	Alto

Fuente: Elaboración propia

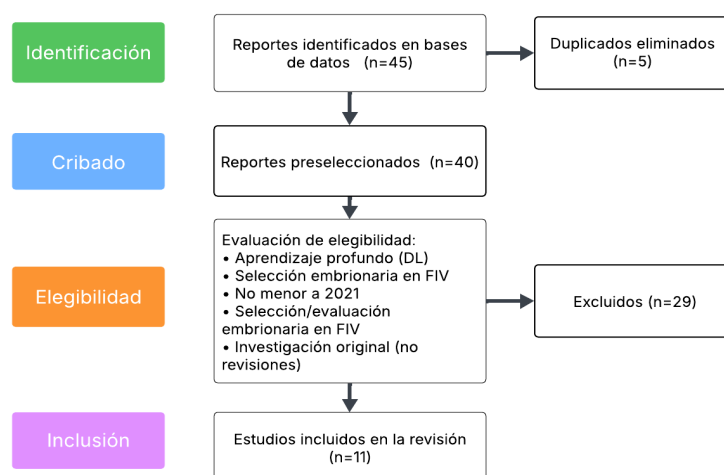


Figura 1. Diagrama Prisma. Fuente: Elaboración propia

3. Resultados

3.1. Selección de estudios

La estrategia de búsqueda permitió identificar 45 registros en las bases de datos seleccionadas. Tras la eliminación de registros duplicados y la aplicación de los criterios de inclusión y exclusión durante las etapas de cribado y elegibilidad, se incluyeron 11 estudios en la síntesis final. El flujo completo del proceso de identificación, cribado, elegibilidad e inclusión se presenta en la Figura 1, mediante el diagrama PRISMA.

3.2. Características de los estudios incluidos

En los últimos años, el campo de la aplicación de DL en la selección de embriones en la FIV ha revolucionado de manera significativa, respecto a la literatura revisada se identificaron artículos con diferente enfoque sea clasificación de embriones, implantación, embarazo clínico y nacido vivo.

Los estudios incluidos en la Tabla 5 presentan heterogeneidad en distintos puntos: los tipos de datos de entrada (imágenes estáticas de microscopio óptico, secuencias de time-lapse, imágenes tomográficas 3D), el enfoque del estudio (clasificación morfológica, implantación, embarazo clínico) y tamaños muestrales (desde 252 hasta más de 181,000 embriones).

Tabla 5. Comparación de artículos de DL en la FIV.

Tipo de estudio	Población	Tipo de datos	Problemática	Técnica de Aprendizaje profundo	Resultados
Retrospectivo unicéntrico [9]	252 embriones	3 imágenes estáticas seleccionadas de videos <i>time-lapse</i> en momentos clave del clivaje	Embarazo clínico	DeepEmbryo: 5 arquitecturas CNN con <i>transfer learning</i> (AlexNet, ResNet18, ResNet34, Inception V3, DenseNet121)	DeepEmbryo 75% de precisión en la predicción de embarazo
Desarrollo y evaluación [12]	1194 imágenes (371 embarazos, 823 no embarazos)	Imágenes estáticas de microscopía de luz (día 5). Edad materna, <i>pseudo-features</i> (<i>stage</i> , ICM, TE continuos)	Mejorar la precisión en la clasificación de embriones en día 5 para optimizar las tasas de éxito	CNN, ResNet-50 y EfficientNet-B0 y GANs	F1: 65.02% Exactitud (<i>Accuracy</i>): 69.04% Sensibilidad (<i>Recall</i> embarazos): 56.76% Especificidad: 74.55% AUC: 66.98%
Retrospectivo observacional unicéntrico [14]	1,580 videos de embriones de 460 pacientes.	TLI (EmbryoScope+). Embriones emparejados del mismo ciclo de estimulación con diferente resultado de implantación	Detectar embriones de buena calidad del mismo ciclo que los métodos convencionales clasificaron como similares	Aprendizaje contrastivo autosupervisado para entrenar CNNs + ajuste fino con red neuronal Siamesa + modelo final XGBoost.	AUC=0.57 y Sin historial: AUC=0.64. AUCs
Retrospectivo, desarrollo y comparación de múltiples CNN con	2,440 embriones para entrenamiento + 742 embriones en	Imágenes estáticas 2D de embriones en día 5 (113 hpi)	Calidad embrionaria (clasificación morfológica en 5 clases según	CNN: Inception v3, ResNet-50, Xception, NASNetLarge, ResNeXt-50, DenseNet-161,	Xception: <i>accuracy</i> 90.97% (blastocisto vs no blastocisto), AUC 0.963

validación externa [18]	<i>test</i> independiente; imágenes capturadas a 113 hpi;		criterios de Gardner)	<i>transfer learning</i> con ImageNet	
RCT doble ciego de no- inferioridad, multicéntrico [19]	1,066 pacientes aleatorizadas; 14 clínicas de FIV. transferencia de embrión único fresco o congelado	Secuencias TLI	Embarazo clínico (latido cardíaco fetal a semana 6- 8)+H8	iDAScore (CNN 3D), sin anotaciones manuales	Tasa de embarazo clínico: iDAScore 46.5% vs. Morfología 48.2%
Multicéntrico, retrospectivo, de gran escala. Entrenamiento y validación interna y externa [20]	181,428 embriones de 22 clínicas FIV en todo el mundo	Secuencias TLI multietapa, embriones incubados por 2, 3 y 5+ días, datos de	Implantación KID	iDAScore v2.0 CNN 3D, identifica patrones espaciales (morfológicos) y temporales (morfocinéticos)	AUC: 0.621 (día 2), 0.649 (día 3), 0.707 (día 5+). Supera KIDScore D5 v3 en día 5+; equivalente a KIDScore D3 en día 3
Retrospectivo, validación independiente [21]	3960 ciclos SVBT (blastocisto único vitricado- desvitricado), Blastocistos días 5–7. Estratificación SART.	Secuencias TLI blastocistos	Embarazo en curso	iDAScore v2.0 (IDA-V2) vs. iDAScore v1.0 (IDA-V1): misma arquitectura CNN 3D pero con 15% más datos de entrenamiento. comparación con KIDScore y criterios de Gardner	iDA-V2 0.736 iDA-V1 0.720 Gardner 0.702
Retrospectivo de cohorte grande, unicéntrico y validación independiente [22]	3,018 ciclos SVBT, edad materna media 39,3 ± 4,0 años.	TLI, secuencias completas, sin anotaciones manuales	Embarazo clínico	iDAScore v1: red neuronal profunda automatizada y sin anotaciones.	AUC iDAScore, KIDScore y Gardner <35 años: 0,72 vs 0,66 y 0,64; 35–37 años: 0,68 vs 0,68 y 0,65; 38–40 años: 0,67 vs 0,65 y 0,64; 41–42 años: 0,66 vs 0,66 y 0,63;

					>42 años: 0,76 vs 0,75 y 0,75
RCT - Prospectivo multicéntrico piloto [23]	Pacientes de 2 clínicas con ≥ 2 blastocistos de buena calidad (≥ 3 BB Gardner) en día 5	Secuencias de imágenes time-lapse, anotaciones morfofocinéticas para KIDScore y automatizado para iDAScore	Embarazo clínico	iDAScore (CNN 3D) vs KIDScore D5 (modelo morfofocinético con anotaciones)	Tasa de embarazo clínico: 50.0% en ambos grupos. hCG+: 64.5% (iDAScore) vs 59.4% (KIDScore). No se demostró superioridad del modelo DL sobre el modelo morfofocinético convencional
Piloto retrospectivo con evaluación clínica interna [24]	977 casos (458 implantación positiva, 519 negativa), 10,747 imágenes tomográficas	Imágenes tomográficas 3D de blastocistos + parámetros clínicos (edad, morfología)	Implantación	Sistema dual de IA: CNN segmentación semántica con <i>transfer learning</i> y CNN 3D para predicción de implantación con parámetros CEE combinados	AUC: 0.774 $\pm$ 0.033 Sensibilidad: 0.727 Especificidad: 0.719 Exactitud: 0.723 VPP 0.727 VPN 0.719
Desarrollo y validación de modelos [25]	700 embriones de FIV-ICSI (443 con nacimiento vivo, 257 sin embarazo),	Imágenes 2D time-lapse (EmbryoScope /EmbryoScope+), etiquetadas por etapa de desarrollo y resultado clínico.	Identificar momentos específicos del desarrollo embrionario que sean predictivos de nacimiento vivo	Pipeline CNN con <i>transfer learning</i> MobileNetV2 como arquitectura	Zygote vs 2-cell: 97.1% <i>accuracy</i> <i>Before vs after</i> NEBD: 94.6% <i>accuracy</i> Clasificación multiclase (5 etapas): 87.7% <i>accuracy</i>

Nota: DL, *deep learning*; CNN, *convolutional neural network*; TLI, *time-lapse imaging*; RCT, *randomized controlled trial*; AUC, *area under the curve*; SVBT, *single vitrified-warmed blastocyst transfer*; D5, día 5; IA, inteligencia artificial; KID, *Known Implantation Data*. Fuente: Elaboración propia a partir de los estudios revisados.

Una diferencia entre los estudios analizados está en el objetivo del modelo. Aquellos enfocados en la predicción de embarazo, implantación o nacimiento vivo presentan de forma consistente valores de AUC moderados, comprendidos entre 0,65 y 0,77, incluso cuando se emplean arquitecturas avanzadas, grandes volúmenes de datos y variables clínicas adicionales. En comparación con estudios de clasificación morfológica, donde las métricas son significativamente superiores, como los reportados por Thirumalaraju *et al.* [18], quienes obtuvieron una exactitud del 90.97% y un AUC de 0.963 mediante arquitecturas CNN para clasificación de calidad embrionaria.

Existe evidencia prospectiva en la literatura revisada, proviene del ensayo clínico aleatorizado (RCT) doble ciego de no inferioridad identificado [19], publicado en *Nature Medicine*. Este estudio multicéntrico, realizado de

14 clínicas de FIV de Australia y Europa con 1,066 pacientes, comparó directamente la selección embrionaria mediante iDAScore (un modelo de DL basado en CNN 3D) frente a la evaluación morfológica estandarizada por embriólogos entrenados. Los resultados mostraron tasas de embarazo clínico similares entre ambos grupos (iDAScore: 46.5% vs. morfología: 48.2%; diferencia de riesgo: -1.7%; IC 95%: -7.7 a 4.3), lo que demuestra la no inferioridad del DL.

Existe evidencia retrospectiva que demuestra la capacidad técnica de los modelos, como, los estudios retrospectivos de iDAScore reportaban AUCs de 0.62 a 0.74 para predicción de implantación [20], [21], con antecedentes desde la validación de su primera versión [22], pero no existe evidencia prospectiva suficiente.

El estudio piloto prospectivo multicéntrico (KiDA Study) que comparó iDAScore frente a KIDScore D5 en dos clínicas suecas reportó tasas de embarazo clínico idénticas del 50.0% en ambos grupos, sin demostrar superioridad del modelo de DL completamente automatizado sobre el modelo morfocinético basado en anotaciones, a pesar de los resultados del modelo los autores destacan la ventaja de iDAScore por su capacidad de ahorrar tiempo y en la reducción de subjetividad [23].

La literatura revisada, iDAScore fue el modelo comercial con mayores evaluaciones. La versión 2.0, descrita por Lassen et al. [20], fue entrenada y validada con 181,428 embriones de 22 clínicas mundiales, el cual representa el dataset más grande identificado en esta revisión, el modelo emplea una arquitectura de CNN 3D (Inflated 3D) que procesa secuencias completas de TLI sin necesidad de anotaciones manuales, identificando simultáneamente patrones espaciales (morfológicos) y temporales (morfocinéticos).

La validación independiente externa realizada por Ueno et al. [21] en 3,960 ciclos de transferencia de blastocisto único vitrificado-desvitrificado confirmó que iDAScore v2.0 (AUC 0.736) superó significativamente a la versión 1.0 (AUC 0.720) y a la clasificación de Gardner (AUC 0.702; $P < 0.0001$), demostrando que el incremento de datos de entrenamiento mejora el rendimiento predictivo.

Miyagi *et al.* [24] propusieron un sistema dual de IA que combina la reconstrucción tridimensional de imágenes tomográficas de blastocistos con parámetros convencionales de evaluación embrionaria (como la edad materna), obteniendo un AUC de 0.774 para predicción de implantación. Este enfoque multimodal que integra datos de imagen con información clínica representa una alternativa prometedora para mejorar la precisión predictiva.

Borna *et al.* [9] desarrollaron DeepEmbryo, un algoritmo que utiliza únicamente tres imágenes estáticas capturadas en momentos estándar del proceso de FIV (19, 43 y 67 horas post-inseminación), logrando una precisión del 75% que superó tanto al modelo de imagen única como a un panel de cinco embriólogos expertos. La relevancia de este estudio es la adaptabilidad a laboratorios de FIV sin equipos de TLI, utilizando imágenes que ya se obtienen en la práctica rutinaria. Una de las principales limitaciones de este estudio es el tamaño muestral reducido (252 embriones), lo que limita la generalización de los hallazgos.

Mapstone *et al.* [25] demostraron que diferentes etapas del desarrollo embrionario (pronúcleo, 2 células, 4 células, 8-16 células, blastocisto) tienen distinto valor predictivo para nacido vivo, y que modelos entrenados en etapas previas al blastocisto pueden añadir información predictiva a embriones ya clasificados como de alta calidad en estadio de blastocisto.

Gilboa *et al.* [26] presentaron una contribución metodológica significativa al proponer un marco metodológico estructurado de cuatro pasos para el desarrollo y validación multicéntrica de IA confiable en la evaluación embrionaria: (I) curación de *datasets* representativos del caso de uso clínico, (II) desarrollo y optimización del modelo, (III) evaluación del rendimiento discriminativo a través de datos variables, y (IV) aseguramiento de interpretabilidad mediante correlación con características morfológicas. Validado con 26,088 embriones de 10 clínicas en 9 países, este marco aborda directamente la falta de estandarización que dificulta la comparación entre estudios y la reproducibilidad de resultados, y puede servir como referencia para futuros investigadores en el campo.

A lo largo de la revisión se identificaron limitaciones que se deben de considerar en futuras investigaciones, primero, la mayoría de los estudios son retrospectivos, y el único ensayo controlado aleatorizado (RCT) disponible no logró demostrar ninguna inferioridad. Segundo, la diversidad de métricas de rendimiento utilizadas (AUC, sensibilidad, especificidad, *accuracy*, F1-score) dificulta la comparación directa entre estudios. Tercero, la representatividad poblacional es limitada, la mayoría de los datos provenientes de países desarrollados, lo que deja en duda sobre la generalizabilidad a otras poblaciones. Los estudios multicéntricos con grandes volúmenes de datos, aunque muestran métricas aceptables, aportan una evidencia sólida sobre la aplicabilidad clínica real de estos sistemas. Estudios mencionan que la variabilidad demográfica entre clínicas suele romper los modelos.

Los resultados analizados indican que el DL es una herramienta prometedora para la evaluación embrionaria en la técnica FIV, especialmente en tareas de clasificación morfológica y como sistema de apoyo a la decisión clínica. Estos modelos analizan imágenes y estiman probabilidades basándose en patrones visuales. Las comparaciones directas entre los modelos que predicen la calidad morfológica y los que predicen los resultados de la implantación pueden ser limitadas, ya que evalúan diferentes criterios de valoración clínicos. La evidencia de estudios prospectivos demuestra que los modelos de DL pueden ser implementados en los tratamientos de FIV, no como sustitución del embriólogo sino como apoyo en la optimización de sus procesos.

4. Discusión

La presente revisión de literatura analizó la evidencia disponible respecto a modelos de DL en la selección de embriones humanos en FIV, desde estudios técnicos hasta ensayos clínicos aleatorizados. El principal hallazgo evidenció diferencia entre el alto rendimiento técnico reportado en diversos estudios retrospectivos y la evidencia clínica prospectiva que no ha demostrado superioridad consistente del DL ante los métodos tradicionales. No obstante, esta diferencia deberá interpretarse considerando la heterogeneidad metodológica de los estudios analizados, debido a que los resultados no se obtuvieron en condiciones equivalentes de diseño, datos, objetivos clínicos o validación.

Al respecto, los estudios retrospectivos también difirieron en su escala y contexto de realización, desde investigaciones unicéntricas con 252 embriones, 977 casos o 1,580 videos de embriones, hasta estudios multicéntricos que analizaron 181,428 embriones procedentes de 22 clínicas de FIV. Los modelos fueron desarrollados considerando fuentes de información diferentes: imágenes estáticas 2D, secuencias completas de TLI, imágenes tomográficas 3D, así como la integración de imágenes con variables clínicas, esta diversidad es relevante en términos metodológicos, resuelven más de una tarea predictiva: clasificar la calidad embrionaria, predecir implantación, embarazo clínico, en curso o nacimiento vivo. Aplicaron validación interna, externa o independiente, no obstante, el desempeño de un modelo podría variar cuando se evalúa con datos diferentes a los utilizados durante su desarrollo o en contextos clínicos distintos, las arquitecturas empleadas fueron diversas: CNN, modelos CNN 3D, *transfer learning*, segmentación semántica y aprendizaje contrastivo autosupervisado. Por lo anterior, no es posible interpretar como equivalentes los resultados de *accuracy*, AUC, sensibilidad o especificidad obtenidos ya que los estudios evalúan resultados clínicos diferentes. Durante la comparación entre modelos, además de identificar las mejores métricas, es fundamental considerar que el rendimiento depende también de tipo de datos, arquitectura y tarea predictiva.

La aplicación de las RNA en la FIV está transformando la evaluación de la viabilidad embrionaria, tanto con TLI como con microscopía de luz convencional [27], no obstante, aunque han mostrado utilidad, la evidencia clínica aún es insuficiente para establecer su impacto sobre las tasas de nacido vivo. En un ensayo clínico piloto el iDAScore y el KIDScore alcanzaron la misma tasa de embarazo clínico (50.0%), mientras que en un ensayo multicéntrico de no inferioridad la tasa de embarazo clínico fue de 46.5% con iDAScore frente a 48.2% con la evaluación morfológica; [26] mencionan que un análisis prospectivo robusto debe considerar posibles factores de confusión clínica, condiciones de cultivo, ambiente uterino, características del ciclo de FIV, estado de ploidía y variabilidad en las anotaciones, estos hallazgos coinciden con la literatura revisada que sugiere que el éxito de la FIV no depende de la selección embrionaria únicamente, sino de diversas variables biológicas y clínicas. Por tanto, la capacidad de un modelo para identificar patrones asociados a la calidad embrionaria o a resultados reproductivos determinados, no implica mejores resultados clínicos.

Asimismo, [28] destacaron la necesidad de mayor estandarización, transparencia y validez comparativa en el entrenamiento, evaluación y comparación de modelos de IA aplicados a la selección embrionaria, la evidencia analizada respalda esta preocupación, considerando que la diversidad de diseños, fuentes de datos, objetivos clínicos y estrategias de validación dificulta establecer comparaciones directas y determinar la superioridad de una arquitectura u otra, por ello, los resultados deben interpretarse dentro del contexto metodológico en el que fueron obtenidos y no sólo a partir de una métrica de desempeño.

Las herramientas de IA poseen alto potencial y podrían implementarse como herramientas de apoyo en la práctica clínica, no obstante, los autores proponen futuras mejoras que incluyan conjunto de modelos o incorporación de datos clínicos adicionales, esto sugiere la necesidad de optimizar las técnicas de extracción de características y entrenar modelos con grandes y diversos conjuntos de datos para mejorar su capacidad de

generalización. Considerando que la mayoría de los estudios se basan en datos retrospectivos, los autores enfatizan la necesidad de realizar estudios prospectivos y validaciones clínicas robustas.

La literatura revisada expone al DL como herramienta prometedora de apoyo a la evaluación embrionaria, no obstante, demostrar su utilidad clínica requiere superar la heterogeneidad metodológica y establecer evaluaciones comparables que permitan determinar si las mejoras observadas en el desempeño técnico se traducen efectivamente en mejores resultados reproductivos.

5. Conclusión

Los modelos de DL para la selección embrionaria en FIV han demostrado, a través de múltiples estudios retrospectivos multicéntricos con muestras que alcanzan cientos de miles de embriones, una alta capacidad técnica para evaluar la morfología y morfocinética embrionaria, predecir la implantación y clasificar la calidad embrionaria, con rendimientos comparables o superiores a los métodos convencionales. Sin embargo, la predicción fiable de resultados clínicos continúa siendo un desafío. Esto requiere el desarrollo de modelos híbridos capaces de integrar información embrionaria, temporal y clínica, así como estrategias de validación robustas orientadas a mejorar la generalización en poblaciones con diferentes características demográficas. En este contexto, el principal valor de estas herramientas podría radicar en su capacidad para estandarizar la evaluación embrionaria y reducir la variabilidad entre embriólogos, favoreciendo una toma de decisiones más consistente y objetiva en la práctica clínica.

6. Referencias

- [1] AlSaad, R., Abusarhan, L., Odeh, N., Abd-Alrazaq, A., Choucair, F., Zegour, R., Ahmed, A., Aziz, S., Sheikh, J. (2025). Deep learning applications for human embryo assessment using time-lapse imaging: Scoping review. *Frontiers in Reproductive Health*, 7, 1-18. <https://doi.org/10.3389/frph.2025.1549642>
- [2] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8 (1), 1-74. <https://doi.org/10.1186/s40537-021-00444-8>
- [3] World Health Organization. (2023). *1 in 6 people globally affected by infertility*. <https://www.who.int/news/item/04-04-2023-1-in-6-people-globally-affected-by-infertility>
- [4] Hamamah, S., Barry, F. (2025). Épidémiologie de la fertilité et facteurs qui l'affectent. *EMC - Gynécologie-Obstétrique*, 61 (1), 1–12. [https://doi.org/10.1016/S1283-081X\(24\)49933-5](https://doi.org/10.1016/S1283-081X(24)49933-5)
- [5] Graham, M. E., Jelin, A., Hoon, A. H., Floet, A. M. W., Levey, E., Graham, E. M. (2023). Assisted reproductive technology: Short- and long-term outcomes. *Developmental Medicine & Child Neurology*, 65 (1), 38–49. <https://doi.org/10.1111/dmcn.15332>
- [6] Mancuso, F., Singh, M., Shanks, A. L. (2026). *In vitro fertilization*. StatPearls Publishing. <https://www.ncbi.nlm.nih.gov/books/NBK562266/>
- [7] Cívico Vallejos, Y., Hernández Dacruz, B., Cívico Vallejos, S. (2022). Selección de embriones en los tratamientos de fecundación in vitro. *Clínica e Investigación en Ginecología y Obstetricia*, 49 (1). <https://doi.org/10.1016/j.gine.2021.100709>
- [8] Grammenou, M., Michou, V., Itziou, A., Tsiotsias, A., Eskitzis, P. (2026). The psychological impact of in vitro fertilization (IVF): A gender systematic review. *Healthcare*, 14 (3), 1-31. <https://doi.org/10.3390/healthcare14030375>
- [9] Borna, M.-R., Sepehri, M. M., Maleki, B. (2024). An artificial intelligence algorithm to select most viable embryos considering current process in IVF labs. *Frontiers in Artificial Intelligence*, 7, 1-10. <https://doi.org/10.3389/frai.2024.1375474>
- [10] Citarella, A. A., Battistoni, P., Coscarelli, C., De Marco, F., Di Biasi, L., Wang, M. (2025). EmbryoVision AI: An explainable deep learning framework for enhanced blastocyst selection in assisted reproductive technologies. *Image and Vision Computing*, 165, 1-11. <https://doi.org/10.1016/j.imavis.2025.105795>
- [11] Dimitriadis, I., Zaninovic, N., Chavez Badiola, A., Bormann, C. L. (2022). Artificial intelligence in the embryology laboratory: A review. *Reproductive BioMedicine Online*, 44 (3), 435–448. <https://doi.org/10.1016/j.rbmo.2021.11.003>

- [12] Onthum, K., Champinyo, N., Suthicharoenpanich, K., Engphaiboon, S., Siricharoen, P., Chaichaowarat, R., Suebthawinkul, C. (2025). Combined input deep learning pipeline for embryo selection for in vitro fertilization using light microscopic images and additional features. *Journal of Imaging*, 11 (1), 1-13. <https://doi.org/10.3390/jimaging11010013>
- [13] Papatheodorou, A., Gilboa, D., Seidman, D., Oraiopoulou, C., Karagianni, M., Papadopoulou, M. I., Tsarfati, M., Kubany, A., Christoforidis, N., Chatziparasidou, A. (2022). P-197 Successful implementation of an end-to-end artificial intelligence (AI) platform in a busy IVF clinic: A prospective observational study. *Human Reproduction*, 37(Suppl. 1), deac107.190. <https://doi.org/10.1093/humrep/deac107.190>
- [14] Boucret, L., Chabrun, F., Boguenet, M., Reynier, P., Bouet, P.-E., May-Panloup, P. (2025). Deep-learning model for embryo selection using time-lapse imaging of matched high-quality embryos. *Scientific Reports*, 15, 1-13. <https://doi.org/10.1038/s41598-025-10531-y>
- [15] Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., Parmar, M. (2024). A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, 57, 1-43. <https://doi.org/10.1007/s10462-024-10721-6>
- [16] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, 1-9. <https://doi.org/10.1136/bmj.n71>
- [17] Lash, T. L., VanderWeele, T. J., Haneuse, S., Rothman, K. J. (2021). *Modern epidemiology* (4th Ed.). Wolters Kluwer.
- [18] Thirumalaraju, P., Kanakasabapathy, M. K., Bormann, C. L., Gupta, R., Pooniwal, R., Kandula, H., Souter, I., Dimitriadis, I., Shafiee, H. (2021). Evaluation of deep convolutional neural networks in classifying human embryo images based on their morphological quality. *Heliyon*, 7 (2), 1-12. <https://doi.org/10.1016/j.heliyon.2021.e06298>
- [19] Illingworth, P. J., Venetis, C., Gardner, D. K., Nelson, S. M., Berntsen, J., Larman, M. G., Agresta, F., Ahitan, S., Ahlström, A., Catrall, F., Cooke, S., Demmers, K., Gabrielsen, A., Hindkjær, J., Kelley, R. L., Knight, C., Lee, L., Lahoud, R., Mangat, M., ... Hardarson, T. (2024). Deep learning versus manual morphology-based embryo selection in IVF: A randomized, double-blind noninferiority trial. *Nature Medicine*, 30 (11), 3114–3120. <https://doi.org/10.1038/s41591-024-03166-5>
- [20] Theilgaard Lassen, J., Fly Kragh, M., Rimestad, J., Nygård Johansen, M., Berntsen, J. (2023). Development and validation of deep learning based embryo selection across multiple days of transfer. *Scientific Reports*, 13, 1-9. <https://doi.org/10.1038/s41598-023-31136-3>
- [21] Ueno, S., Berntsen, J., Okimura, T., Kato, K. (2024). Improved pregnancy prediction performance in an updated deep-learning embryo selection model: A retrospective independent validation study. *Reproductive BioMedicine Online*, 48 (1), 1-10. <https://doi.org/10.1016/j.rbmo.2023.103308>
- [22] Ueno, S., Berntsen, J., Ito, M., Uchiyama, K., Okimura, T., Yabuuchi, A., Kato, K. (2021). Pregnancy prediction performance of an annotation-free embryo scoring system on the basis of deep learning after single vitrified-warmed blastocyst transfer: A single-center large cohort retrospective study. *Fertility and Sterility*, 116 (4), 1172–1180. <https://doi.org/10.1016/j.fertnstert.2021.06.001>
- [23] Bohlin, T., Kurkkio, E., Adolfsson, E., Berntsen, J., Nygård Johansen, M., Apter, S., Caisander, G., Kristensen, K., Schmidt, J. (2024). Comparison of the deep learning tool iDAScore® and the current annotation model, KIDScore™ D5, for finding the embryo with best chance of pregnancy after elective single blastocyst transfer: A randomized multi-centre prospective pilot study. *Reproductive BioMedicine Online*, 48 (Suppl. 1). <https://doi.org/10.1016/j.rbmo.2024.104025>
- [24] Miyagi, Y., Habara, T., Hirata, R., Hayashi, N. (2024). Predicting implantation by using dual AI system incorporating three-dimensional blastocyst image and conventional embryo evaluation parameters: A pilot study. *Reproductive Medicine and Biology*, 23 (1), 1-10. <https://doi.org/10.1002/rmb2.12612>
- [25] Mapstone, C., Hunter, H., Brison, D., Handl, J., Plusa, B. (2024). Deep learning pipeline reveals key moments in human embryonic development predictive of live birth after in vitro fertilization. *Biology Methods and Protocols*, 9 (1), 1-12. <https://doi.org/10.1093/biomethods/bpae052>
- [26] Gilboa, D., Garg, A., Shapiro, M., Meseguer, M., Amar, Y., Lustgarten, N., Desai, N., Shavit, T., Silva, V., Papatheodorou, A., Chatziparasidou, A., Angras, S., Lee, J. H., Thiel, L., Curchoe, C. L., Tauber, Y.,

- Seidman, D. S. (2025). Application of a methodological framework for the development and multicenter validation of reliable artificial intelligence in embryo evaluation. *Reproductive Biology and Endocrinology*, 23 (16), 1-12. <https://doi.org/10.1186/s12958-025-01351-w>
- [27]Salih, M., Austin, C., Warty, R. R., Tiktin, C., Rolnik, D. L., Momeni, M., Rezaatofghi, H., Reddy, S., Smith, V., Vollenhoven, B., Horta, F. (2023). Embryo selection through artificial intelligence versus embryologists: A systematic review. *Human Reproduction Open*, 2023 (3), 1-13. <https://doi.org/10.1093/hropen/hoad031>
- [28]Kragh, M. F., Karstoft, H. (2021). Embryo selection with artificial intelligence: How to evaluate and compare methods? *Journal of Assisted Reproduction and Genetics*, 38 (7), 1675–1689. <https://doi.org/10.1007/s10815-021-02254-6>