TESIS DOCTORAL

UCAM
UNIVERSIDAD CATÓLICA
DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

Programa de Doctorado en Tecnologías de la Computación e Ingeniería Ambiental

Caracterización de la malignidad de tumores ováricos mediante imágenes ecográficas y algoritmos de deep learning

Autor:

Igor García Atutxa

Directores:

Dr. D. Andrés Bueno Crespo

Dr. D. José Martínez Más

Murcia, abril de 2026

TESIS DOCTORAL

ESCUELA INTERNACIONAL DE DOCTORADO

Programa de Doctorado en Tecnologías de la Computación e Ingeniería Ambiental

Caracterización de la malignidad de tumores ováricos mediante imágenes ecográficas y algoritmos de deep learning

Autor:

Igor García Atutxa

Directores:

Dr. D. Andrés Bueno Crespo

Dr. D. José Martínez Más

Murcia, abril de 2026

AUTORIZACIÓN DEL DIRECTOR DE LA TESIS
PARA SU PRESENTACIÓN

El Dr. D. Andrés Bueno Crespo y el Dr. D. José Martínez Más como Directores de la Tesis Doctoral titulada "Caracterización de la malignidad de tumores ováricos mediante imágenes ecográficas y algoritmos de deep learning" realizada por D. Igor García Atutxa en el Programa de Doctorado en Tecnologías de la Computación e Ingeniería Ambiental, **autorizan su presentación a trámite** dado que reúne las condiciones necesarias para su defensa.

Lo que firmo, para dar cumplimiento al Real Decreto 99/2011 de 28 de enero, en Murcia a 28 de enero de 2026.

ANDRES|BUENO|CRESPO

Digitally signed by ANDRES|BUENO|CRESPO
Date: 2026.04.27 09:27:47 +02'00'

Firmado por MARTINEZ MAS JOSE - ***3364** el día 27/04/2026 con un certificado emitido por AC FNMT Usuarios

RESUMEN

El cáncer de ovario (CO) representa una de las neoplasias ginecológicas más letales, principalmente debido a su diagnóstico tardío y a la falta de métodos efectivos de detección temprana. Actualmente, el ultrasonido transvaginal (UST) es la modalidad de imagen más utilizada para la evaluación inicial de masas anexiales sospechosas. Sin embargo, presenta limitaciones importantes relacionadas con su especificidad, lo que conduce frecuentemente a incertidumbres diagnósticas y procedimientos invasivos innecesarios.

Esta tesis doctoral tiene como objetivo general desarrollar, implementar y validar modelos híbridos de aprendizaje profundo (DL) basados en una estrategia de fusión temprana adaptativa entre arquitecturas de redes neuronales convolucionales (CNNs) y transformadores de visión (ViTs) para la clasificación multicategoría de tumores ováricos a partir de imágenes ecográficas transvaginales en modo B. El propósito es avanzar hacia un sistema de apoyo a la decisión clínica que, frente a enfoques convencionales de una sola rama, mejore de forma conjunta la precisión diagnóstica, la calibración probabilística, la interpretabilidad y la utilidad clínica.

En primer lugar, se realiza una revisión sistemática que abarca 44 estudios publicados hasta diciembre de 2024, en los que se evalúa la precisión diagnóstica de diferentes modelos de Inteligencia Artificial (IA) aplicados al diagnóstico de CO mediante ultrasonido. El análisis muestra una precisión diagnóstica global elevada (92.3% de exactitud, 91.6% de sensibilidad y 90.1% de especificidad); sin embargo, también revela una heterogeneidad metodológica sustancial y sesgos frecuentes (p. ej., validaciones limitadas y conjuntos de datos pequeños) que restringen la traducción clínica directa de estos modelos y evidencian carencias recurrentes en el reporte de métricas clave para despliegue clínico, como calibración y análisis de utilidad.

En segundo lugar, se propone y valida rigurosamente un modelo híbrido CNN–ViT con una estrategia adaptativa de fusión temprana para la clasificación multiclase de tumores ováricos. El modelo EfficientNetB7–Swin Transformer se entrena y evalúa utilizando un conjunto de datos clínicamente representativo compuesto por 1469 imágenes ecográficas distribuidas en ocho categorías

diagnósticas. El híbrido alcanza un desempeño superior, con 92.13% de exactitud, 92.38% de sensibilidad y un área bajo la curva ROC (AUC) de 0.9904, superando de forma significativa a modelos individuales basados solo en CNNs o solo en ViTs ($p < 0.01$). Además, una estrategia de ensamble suave (combinando los tres mejores híbridos) incrementa el rendimiento hasta 93.3% de exactitud, 93.6% de sensibilidad y 99.0% de especificidad. De manera complementaria, se evalúa el comportamiento probabilístico del sistema mediante calibración (regresión isotónica), se estima su utilidad clínica potencial mediante curvas de decisión clínica (CDC) y se cuantifica la incertidumbre predictiva como criterio operativo para la derivación a revisión experta.

Finalmente, para mejorar la interpretabilidad clínica y la transparencia de las predicciones, se implementa Grad-CAM para generar mapas de activación visual, que muestran correspondencia entre las regiones resaltadas por el modelo y las áreas de relevancia clínica definidas por expertos.

En conclusión, esta investigación doctoral demuestra el potencial clínico de los modelos híbridos CNN–ViT con fusión temprana adaptativa e interpretación visual para apoyar la caracterización de imágenes ecográficas de tumores ováricos. Asimismo, aporta recomendaciones específicas para superar limitaciones metodológicas recurrentes (segmentación, validación, reporte y calibración), incorporando criterios orientados a despliegue (calibración, utilidad e incertidumbre), que pueden facilitar la integración responsable de la IA en la práctica clínica para una detección más precoz y precisa del CO.

PALABRAS CLAVE

Cáncer de ovario, ultrasonido transvaginal, inteligencia artificial, redes neuronales, diagnóstico por imagen, transformadores de visión, modelos híbridos, detección precoz, interpretabilidad.

ABSTRACT

Ovarian cancer (CO) is one of the most lethal gynecologic malignancies, mainly due to late diagnosis and the lack of effective early detection methods. Currently, transvaginal ultrasound (UST) is the most widely used imaging modality for the initial assessment of suspicious adnexal masses. However, it has important limitations related to specificity, frequently leading to diagnostic uncertainty and unnecessary invasive procedures.

The overarching objective of this doctoral thesis is to develop, implement, and validate hybrid deep learning (DL) models based on an adaptive early-fusion strategy that integrates convolutional neural network (CNNs) architectures and vision transformers (ViTs) for multicategory classification of ovarian tumors from B-mode transvaginal ultrasound images. The aim is to advance toward a clinical decision-support system that, compared with conventional single-branch approaches, jointly improves diagnostic accuracy, probabilistic calibration, interpretability, and clinical utility.

First, a systematic review is conducted, encompassing 44 studies published up to December 2024, which evaluate the diagnostic accuracy of different artificial intelligence (AI) models applied to CO diagnosis using ultrasound. The analysis shows high overall diagnostic performance (92.3% accuracy, 91.6% sensitivity, and 90.1% specificity); however, it also reveals substantial methodological heterogeneity and frequent biases (e.g., limited validations and small datasets) that constrain the direct clinical translation of these models and highlight recurring gaps in the reporting of deployment-critical metrics, such as calibration and clinical utility analysis.

Second, a hybrid CNN–ViT model with an adaptive early-fusion strategy is proposed and rigorously validated for multiclass ovarian tumor classification. The EfficientNetB7–Swin Transformer model is trained and evaluated using a clinically representative dataset comprising 1469 ultrasound images distributed across eight diagnostic categories. The hybrid achieves superior performance, with 92.13% accuracy, 92.38% sensitivity, and an area under the ROC curve (AUC) of 0.9904, significantly outperforming single models based only on CNNs or only on ViTs ($p < 0.01$). In addition, a soft-ensemble strategy (combining the three best hybrids) increases performance to 93.3% accuracy, 93.6% sensitivity, and 99.0% specificity.

Complementarily, the system's probabilistic behavior is assessed through calibration (isotonic regression), its potential clinical utility is estimated via decision curve analysis (CDC), and predictive uncertainty is quantified as an operational criterion for referral to expert review.

Finally, to enhance clinical interpretability and prediction transparency, Grad-CAM is implemented to generate visual activation maps, showing correspondence between model-highlighted regions and clinically relevant areas defined by experts.

In conclusion, this doctoral research demonstrates the clinical potential of hybrid CNN–ViT models with adaptive early fusion and visual interpretability to support the characterization of ultrasound images of ovarian tumors. It also provides specific recommendations to overcome recurrent methodological limitations (segmentation, validation, reporting, and calibration) by incorporating deployment-oriented criteria (calibration, clinical utility, and uncertainty), which may facilitate the responsible integration of AI into clinical practice for earlier and more accurate detection of CO.

KEYWORDS

Ovarian cancer, transvaginal ultrasound, artificial intelligence, neural networks, diagnostic imaging, vision transformers, hybrid models, early detection, interpretability.

AGRADECIMIENTOS

Deseo expresar mi más profundo agradecimiento a mi familia, verdadero pilar de este recorrido. Gracias por vuestra comprensión inquebrantable, vuestra paciencia y por acompañarme con afecto sincero en cada etapa de este camino. Sin vuestro apoyo constante, este logro no habría sido posible.

De manera muy especial, agradezco a mi esposa. Tu amor incondicional, tu compañía serena y tu fortaleza en los momentos más difíciles han sido mi refugio y mi impulso. Has sido luz cuando más lo necesitaba, sostén en la incertidumbre y alegría en cada pequeño avance. Esta tesis lleva impresa tu huella en cada página.

A mis padres, José y Arantza, gracias por enseñarme con el ejemplo el valor del esfuerzo, por brindarme siempre libertad para elegir mi rumbo y por sostenerme con generosidad y confianza. Vuestro apoyo ha sido esencial para que pudiera avanzar con firmeza y sin miedo.

A mi hermano Iñigo, por su cariño sincero y por estar siempre dispuesto a tenderme la mano. Tu compañía y apoyo han sido un ancla constante, y tu presencia, una fuente de ánimo en los momentos exigentes.

A mis amigos, gracias por estar cerca aun en la distancia, por las palabras de aliento, las risas compartidas y los silencios cómplices. En especial, a Jon Aitor y Jagoba, por vuestra lealtad, por caminar conmigo incluso cuando el camino se volvía cuesta arriba. Vuestra amistad ha sido un regalo invaluable.

A todos vosotros, gracias de corazón. Esta meta es también vuestra.

"En la vida no hay cosas que temer, solo cosas que comprender". Marie Curie (1867-1934).

ÍNDICE GENERAL

SIGLAS Y ABREVIATURAS

4D: Imagen cuatridimensional (3D + tiempo)

ADF: Algoritmo de detección facial (plural: ADFs)

ADNEX: Assessment of Different NEoplasias in the adneXa (modelo ADNEX)

ANN: Red neuronal artificial (Artificial Neural Network; plural: ANNs)

AUC: Área bajo la curva ROC (Area Under the ROC Curve)

BI-RADS: Breast Imaging Reporting and Data System

CA125: Antígeno de cáncer 125 (Cancer Antigen 125)

CDC: Curvas de decisión clínica

CNN: Red neuronal convolucional (Convolutional Neural Network; plural: CNNs)

CO: Cáncer de ovario

CONSORT-AI: Extensión CONSORT para intervenciones de inteligencia artificial (Consolidated Standards of Reporting Trials – AI extension).

DALY: Año de vida ajustado por discapacidad (Disability-Adjusted Life Year; plural: DALYs)

DE: Desviación estándar

DL: Aprendizaje profundo (Deep Learning)

EMA: Agencia Europea de Medicamentos (European Medicines Agency)

FDA: Administración de Alimentos y Medicamentos de Estados Unidos (Food and Drug Administration)

FIGO: Federación Internacional de Ginecología y Obstetricia (Fédération Internationale de Gynécologie et d'Obstétrique)

GCO: Observatorio Global del Cáncer (Global Cancer Observatory)

Grad-CAM: Mapeo de activación de clase ponderado por gradiente (Gradient-weighted Class Activation Mapping)

HE4: Proteína epididimaria humana 4 (Human Epididymis Protein 4). Se encuentra también en el epitelio del ovario, y por ello es un marcador de lesión ovárica.

IA: Inteligencia artificial

IOTA: International Ovarian Tumor Analysis. Es un grupo internacional de investigación.

MCC: Coeficiente de correlación de Matthews (Matthews Correlation Coefficient)

ML: Aprendizaje automático (Machine Learning)

MMOTU: Ultrasonido de tumores ováricos multimodal (Multi-Modality Ovarian Tumor Ultrasound)

O-RADS: Ovarian-Adnexal Reporting and Data System

OTU-2D: Ultrasonido de tumor ovárico 2D (Ovarian Tumor Ultrasound–2D; subconjunto 2D del conjunto MMOTU)

PACS: Sistema de archivado y comunicación de imágenes (Picture Archiving and Communication System)

PARP: Poli(ADP-ribosa) polimerasa (Poly[ADP-Ribose] Polymerase)

PLCO: Ensayo de cribado de cáncer de próstata, pulmón, colorrectal y ovario (Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial)

PRISMA: Elementos preferidos para informar revisiones sistemáticas y metaanálisis (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)

PROBAST: Herramienta para evaluar el riesgo de sesgo en modelos de predicción (Prediction model Risk Of Bias ASsessment Tool)

PROBAST-AI: Extensión de PROBAST para modelos de predicción basados en IA (PROBAST-AI extension)

QALY: Año de vida ajustado por calidad (Quality-Adjusted Life Year; plural: QALYs)

RGPD: Reglamento General de Protección de Datos

RIS: Sistema de información radiológica (Radiology Information System)

RM: Resonancia magnética

RNN: Red neuronal recurrente (Recurrent Neural Network; plural: RNNs)

ROI: Región de interés (Region of Interest; plural: ROIs)

SaMD: Software como dispositivo médico (Software as a Medical Device)

SHAP: Explicaciones aditivas de Shapley (SHapley Additive exPlanations)

SPIRIT-AI: Extensión SPIRIT para protocolos de ensayos con intervenciones de IA (Standard Protocol Items: Recommendations for Interventional Trials – AI Extension)

STARD-AI: Extensión STARD para estudios de precisión diagnóstica que incorporan IA (Standards for Reporting of Diagnostic Accuracy Studies – AI Extension)

TC: Tomografía computerizada

TRIPOD-AI: Extensión TRIPOD para el reporte transparente de modelos multivariables de predicción que incorporan IA (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis – AI Extension)

UKCTOCS: Ensayo colaborativo del Reino Unido para el cribado de cáncer de ovario (United Kingdom Collaborative Trial of Ovarian Cancer Screening)

UST: Ultrasonido transvaginal

UST-3D: Ultrasonido transvaginal tridimensional

ViT: Transformador de visión (Vision Transformer; plural: ViTs)

ÍNDICE DE FIGURAS, DE TABLAS Y DE ANEXOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ANEXOS

I – INTRODUCCIÓN

I - INTRODUCCIÓN

1.1. PANORAMA CLÍNICO Y CIENTÍFICO DEL CÁNCER DE OVARIO

El cáncer de ovario (CO) constituye uno de los retos más importantes de la oncología contemporánea. No solo es la neoplasia ginecológica más letal, sino también una de las que mayor carga genera en términos de mortalidad, años de vida perdidos y costes asociados a su tratamiento [1–3]. A diferencia de otros cánceres femeninos, como el de mama, en el que la implementación de programas de cribado poblacional ha permitido mejorar significativamente la supervivencia, el CO carece hasta hoy de una estrategia de detección precoz eficaz y que haga un uso eficiente de los recursos disponibles [4–6].

De acuerdo con los datos más recientes del Observatorio Global del Cáncer (GCO) de 2022, se diagnostican anualmente alrededor de 324 000 nuevos casos de CO en el mundo, con más de 206 000 muertes, lo que equivale a una incidencia de 6.7 por cada 100 000 mujeres (Figura 1). Aunque se trata de la octava neoplasia más frecuente en la mujer, ocupa el quinto lugar en mortalidad oncológica femenina [7]. Este desajuste entre incidencia y mortalidad ilustra una característica central: el CO es relativamente menos común que otros tumores, pero su impacto en términos de supervivencia es desproporcionadamente mayor.

La relación entre incidencia y mortalidad en el CO es mucho más estrecha que en otros tumores ginecológicos. En cáncer de mama, por ejemplo, la mortalidad a cinco años es aproximadamente del 15%, mientras que en CO se eleva a cerca del 60%. Ello refleja la incapacidad actual de diagnosticar la enfermedad en fases iniciales, cuando el tratamiento tiene mayor probabilidad de éxito. Más del 70% de las pacientes con CO son diagnosticadas en estadios FIGO III–IV, es decir, cuando el tumor ya se encuentra extendido en la cavidad peritoneal o presenta metástasis a distancia [1, 4]. En estos estadios avanzados, la supervivencia global a cinco años no supera el 30%. Por el contrario, en estadios tempranos (FIGO I–II), la supervivencia puede acercarse al 90%. Este diferencial tan marcado convierte a la

detección temprana en la piedra angular para cambiar la historia natural de la enfermedad [2, 3, 5, 8, 9].

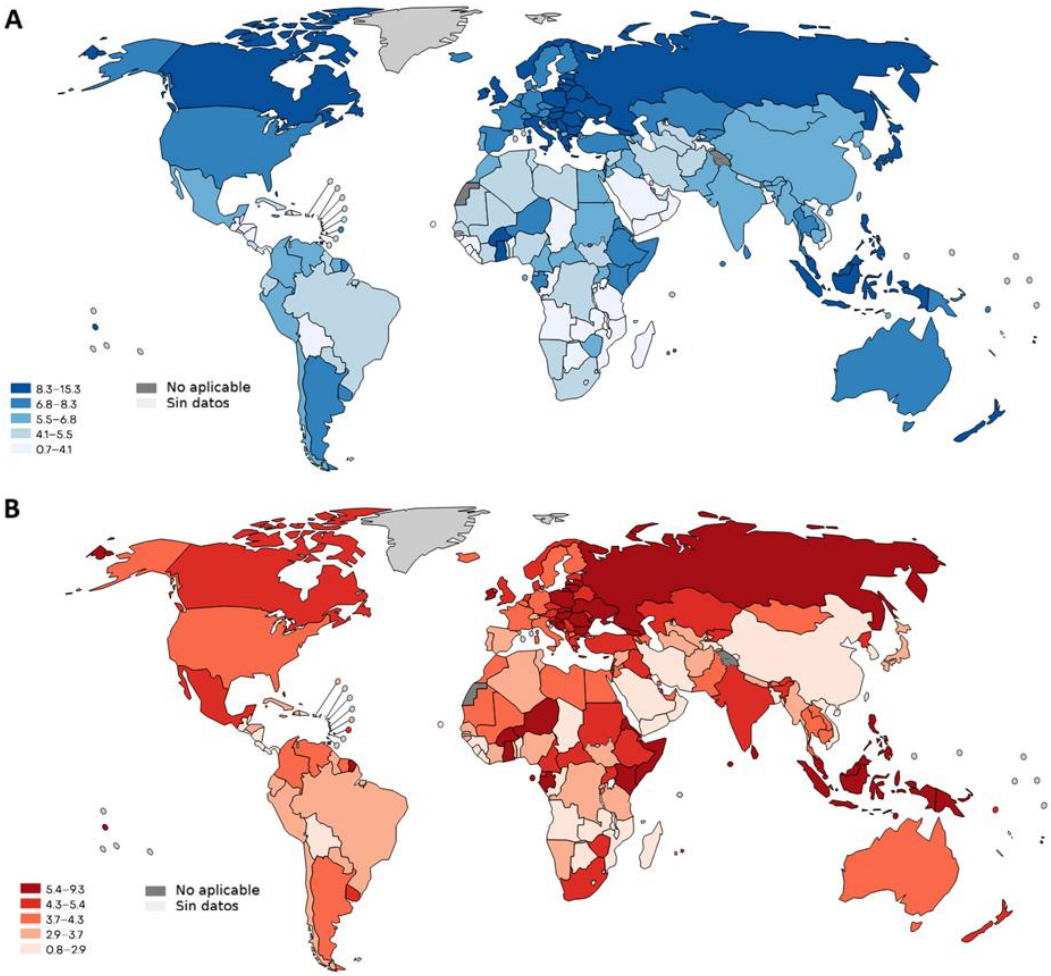

Figura 1. Tasas de incidencia (A) y mortalidad (B) en cáncer de ovario (CO) a nivel mundial (por 100 000 mujeres-año). Fuente: Observatorio Global del Cáncer 2022.

El panorama clínico se ve agravado por la ausencia de síntomas específicos en fases iniciales. La sintomatología del CO suele iniciar con molestias abdominales vagas, distensión o síntomas urinarios y digestivos inespecíficos que retrasan la consulta médica y favorecen el diagnóstico tardío. A diferencia de otros tumores en los que existen biomarcadores fiables para la detección temprana, en el CO el marcador sérico antígeno de cáncer 125 (CA125) tiene una sensibilidad y especificidad insuficientes, con elevaciones también presentes en condiciones benignas como endometriosis o enfermedad inflamatoria pélvica, puesto que es un

antígeno localizado no solo en ovario sino también en zonas serosas (pleura o peritoneo). La proteína 4 del epidídimo humano (HE4) tampoco ha demostrado suficiente eficacia como herramienta de cribado poblacional. De hecho, los dos grandes ensayos clínicos multicéntricos de cribado, el PLCO (Ensayo de cribado de cáncer de próstata, pulmón, colorrectal y ovario) en Estados Unidos [10] y el UKCTOCS (Ensayo colaborativo del Reino Unido para el cribado de cáncer de ovario) en el Reino Unido [11], concluyeron que ni el uso de CA125 ni el ultrasonido transvaginal (UST) disminuyeron de forma significativa la mortalidad por CO en población general. Este fracaso consolidó la idea de que el cribado universal no es factible con las herramientas disponibles actualmente. Revisiones recientes sobre programas de cribado basados en UST y biomarcadores coinciden en que, salvo en subgrupos de muy alto riesgo, el balance riesgo-beneficio del cribado poblacional sigue siendo, en el mejor de los casos, modesto [12, 13].

La epidemiología del CO también muestra importantes desigualdades regionales. En Europa del Este y en países como Letonia, Serbia o Bulgaria, las tasas de incidencia superan los 11 casos por cada 100 000 mujeres, mientras que en África Subsahariana la incidencia es inferior a 3 [14]. Sin embargo, esta menor incidencia no se traduce en mejores resultados: en países de bajos ingresos la mortalidad proporcional es mayor debido a la ausencia de sistemas de salud que garanticen un diagnóstico oportuno y acceso a tratamientos especializados. En Hispanoamérica la incidencia es intermedia, pero la supervivencia global a cinco años es baja (30–35%) en comparación con países de altos ingresos (45–50%). Estas diferencias no responden únicamente a factores biológicos, sino también a determinantes sociales, estructurales y económicos, lo que convierte al CO en un problema de equidad en salud global.

Desde la perspectiva de la carga global de enfermedad, el CO supone más de 6 millones de años de vida ajustados por discapacidad (DALYs) perdidos cada año en el mundo. Esto significa no solo una reducción en la esperanza de vida, sino también un deterioro sustancial en la calidad de vida de las pacientes y sus familias. Los tratamientos de primera línea incluyen cirugía de citorreducción (primaria o intervalar) y quimioterapia basada en platinos, a los que en años recientes se han añadido terapias dirigidas, como los inhibidores de Poli(ADP-ribosa) polimerasa (PARP). Si bien estos avances han mejorado la supervivencia libre de progresión, el impacto en la supervivencia global sigue siendo modesto. La mayoría de las

pacientes sufren recurrencias múltiples que requieren tratamientos repetidos, con efectos secundarios acumulativos, mayor dependencia funcional y un profundo impacto emocional y económico. Diversos análisis recientes destacan que la intensificación terapéutica en estadios avanzados y en líneas sucesivas de tratamiento contribuye de forma sustancial a la carga económica y a la utilización de recursos sanitarios, incluso en sistemas de salud de altos ingresos [4, 6, 9].

En términos de economía de la salud, el CO es uno de los tumores con peor relación coste-efectividad. En Estados Unidos, el gasto anual en su tratamiento supera los 5 mil millones de dólares, mientras que en países de ingresos bajos y medios los costes relativos para las familias son catastróficos, al superar la capacidad de pago de la mayoría de los hogares. En México, se estima que el tratamiento de un caso avanzado de CO puede superar los 10 000 euros anuales, cifra inalcanzable para gran parte de la población si no cuenta con cobertura pública. Este panorama económico evidencia que mejorar la detección temprana no sólo salvaría vidas, sino que también reduciría de manera significativa los costes sanitarios asociados.

En conjunto, el CO representa un problema clínico y científico complejo. Clínicamente, porque se diagnostica tarde, carece de síntomas iniciales específicos y no cuenta con programas de cribado eficaces. Científicamente, porque plantea un reto en el desarrollo de herramientas diagnósticas que superen las limitaciones de los métodos actuales. En este contexto se sitúa la necesidad de explorar enfoques innovadores, como la inteligencia artificial (IA) aplicada al UST, que permitan avanzar hacia un diagnóstico más temprano, preciso y equitativo.

1.2. EL ULTRASONIDO COMO MÓDULO DE DECISIÓN: FORTALEZAS, LÍMITES Y ESTANDARIZACIÓN

El UST ha ocupado, durante más de tres décadas, un lugar privilegiado en la práctica ginecológica y oncológica. Constituye la técnica de primera línea en la evaluación de masas anexiales, en gran medida por su disponibilidad universal, bajo coste, inocuidad, al no implicar radiación ionizante, y capacidad de exploración en tiempo real. Estos atributos lo han consolidado como una herramienta esencial no solo en entornos hospitalarios de alta especialidad, sino también en clínicas de primer contacto, lo que lo convierte en la modalidad

diagnóstica con mayor potencial para impactar en la detección temprana del CO [6, 15]. Diversas guías de práctica clínica y consensos internacionales lo reconocen como la prueba de imagen de primera línea para la caracterización de masas anexiales y la estratificación inicial del riesgo de malignidad en la mayoría de los escenarios asistenciales [16].

1.2.1. Fortalezas clínicas y técnicas del ultrasonido transvaginal

El UST en modo B permite obtener imágenes de alta resolución de los ovarios y estructuras pélvicas adyacentes, facilitando la evaluación de parámetros morfológicos clave: tamaño ovárico, contorno, presencia de septos, papilas, componentes sólidos y vascularización de las masas anexiales. Estos hallazgos constituyen la base de la estratificación diagnóstica inicial y, en manos expertas, permiten orientar la conducta clínica con una sensibilidad superior al 85%. Los estudios de la red *International Ovarian Tumor Analysis* (IOTA) han mostrado que la combinación de evaluación subjetiva por ecografistas expertos y modelos de regresión logística (LR1, LR2) alcanza áreas bajo la curva superiores a 0.90 y supera de forma consistente a índices clásicos como la resonancia magnética (RM) en la discriminación entre tumores benignos y malignos [17, 18].

A diferencia de modalidades de imagen como la RM o la tomografía computarizada (TC), el UST es dinámico: permite explorar la movilidad de las estructuras, la compresibilidad de las masas y la relación con órganos adyacentes, lo que añade un componente funcional a la valoración anatómica. La incorporación del Doppler color y espectral, a partir de los años noventa, supuso un avance en la caracterización vascular de las lesiones, al permitir diferenciar patrones de neovascularización más frecuentes en tumores malignos. Posteriormente, el ultrasonido transvaginal tridimensional (UST-3D) amplió las posibilidades de reconstrucción espacial, y más recientemente la ecografía con contraste ha sido explorada en centros especializados para mejorar la diferenciación entre tumores benignos y malignos. En este contexto, las Reglas Simples de IOTA y modelos multiclase como ADNEX (*Assessment of Different NEoplasias in the adneXa*) proporcionan un marco probabilístico y un vocabulario estructurado que pueden integrarse en la práctica clínica cotidiana manteniendo el bajo coste, la ausencia de radiación ionizante y la portabilidad del UST [19, 20].

Sin embargo, a pesar de estas innovaciones, en la práctica clínica diaria el modo B convencional sigue siendo la técnica más extendida. Esto se debe tanto a su bajo coste como a la simplicidad de su interpretación, aunque estas ventajas vienen acompañadas de limitaciones significativas en términos de reproducibilidad diagnóstica.

1.2.2. Limitaciones y variabilidad interobservador

El principal talón de Aquiles del UST es la dependencia del operador. La adquisición de la imagen, la selección de los planos más representativos y la interpretación final están influidas por la experiencia y pericia del ecografista. Numerosos estudios han documentado una variabilidad inter e intraobservador considerable, que se traduce en diagnósticos dispares ante una misma lesión [21, 22].

En lesiones complejas o *borderline*, esta variabilidad se acentúa. Los tumores serosos de bajo grado o los tumores mucinosos multiloculares pueden presentar características que simulan benignidad en fases iniciales. Del mismo modo, endometriomas o teratomas pueden mostrar patrones ecográficos que imitan malignidad. Estos escenarios se traducen clínicamente en falsos positivos, que conducen a cirugías innecesarias con riesgo de complicaciones, y falsos negativos, que retrasan el diagnóstico de lesiones malignas.

En contextos con especialistas experimentados, la sensibilidad y especificidad del UST para diferenciar masas benignas y malignas pueden superar el 90%. Sin embargo, en manos no expertas o en entornos rurales con recursos limitados, la sensibilidad puede descender hasta el 70% y la especificidad al 60–65% [23, 24]. Estas diferencias reflejan la necesidad de sistemas de apoyo que estandaricen y cuantifiquen la interpretación, reduciendo la subjetividad inherente a la práctica clínica actual. Metaanálisis y estudios de validación externa sugieren, además, que modelos basados en IOTA como ADNEX mantienen un rendimiento elevado cuando se aplican fuera de los centros que los desarrollaron y que las Reglas Simples IOTA pueden emplearse con seguridad por examinadores menos experimentados para la estratificación inicial de masas anexiales [23, 24].

1.2.3. Estandarización: el aporte de IOTA y O-RADS

Conscientes de estas limitaciones, la comunidad científica ha impulsado iniciativas de estandarización del lenguaje y los criterios diagnósticos. El grupo IOTA desarrolló reglas simples y modelos predictivos basados en variables morfológicas y clínicas, los cuales han demostrado mejorar la reproducibilidad interobservador. Posteriormente, el sistema O-RADS (Sistema de informes y datos para ovario-anexos), inspirado en el BI-RADS (Sistema de informe y datos de imagen mamaria) de cáncer de mama, introdujo un marco estructurado de clasificación de lesiones ováricas en categorías de riesgo. Este sistema busca homogeneizar los reportes y facilitar la comunicación entre ecografistas, ginecólogos y oncólogos. O-RADS asigna un valor de riesgo que va desde la categoría 0 (estudio incompleto) hasta la categoría 5 (lesión altamente sospechosa de malignidad) [25–28].

Metaanálisis recientes han mostrado que O-RADS ofrece sensibilidades superiores al 90% en la detección de malignidad. No obstante, la especificidad suele ser más baja (70–80%), lo que implica que un número significativo de lesiones benignas acaban derivándose a cirugía. Además, la implementación de estos sistemas en la práctica diaria requiere capacitación y adherencia estricta a protocolos, lo que no siempre ocurre en entornos de alta carga asistencial [29, 30].

1.2.4. Limitaciones persistentes a pesar de la estandarización

A pesar de la existencia de marcos estandarizados como IOTA y O-RADS, la reproducibilidad sigue siendo insuficiente en la práctica clínica real. En parte, esto se debe a la heterogeneidad fenotípica del CO, que puede presentar patrones morfológicos muy variados. Los tumores *borderline*, por ejemplo, son un área gris en la que ni las reglas simples ni las clasificaciones por riesgo alcanzan sensibilidad y especificidad satisfactorias.

Otro problema es la falta de validación externa amplia de estos sistemas en poblaciones diversas. La mayoría de los estudios de validación externa provienen de centros especializados en Europa y Norteamérica, mientras que en Hispanoamérica, África o Asia las validaciones son escasas. Esto limita la transportabilidad de los resultados y perpetúa las desigualdades en salud global.

1.2.5. Perspectiva crítica

En síntesis, el UST ha sido y seguirá siendo la herramienta diagnóstica de primera línea en la evaluación de masas anexiales. Sus fortalezas (accesibilidad, bajo coste, seguridad y dinamismo) lo convierten en un recurso insustituible. Sin embargo, su dependencia del operador, la variabilidad interobservador y la limitada especificidad en lesiones complejas justifican la necesidad de innovación tecnológica. Diversas revisiones sistemáticas y metaanálisis han subrayado que la mera intensificación de los protocolos ecográficos o el uso aislado de nuevos sistemas de puntuación difícilmente bastan para reducir de forma sustancial la mortalidad por CO si no se acompañan de herramientas de decisión más objetivas, calibradas y reproducibles [31, 32].

La estandarización semántica y los sistemas estructurados como IOTA y O-RADS han representado avances significativos, pero no suficientes para resolver la incertidumbre diagnóstica. La consecuencia es doble: cirugías innecesarias en lesiones benignas y retraso en la detección de tumores malignos incipientes.

De esta manera, el UST se encuentra en un punto de inflexión. Su futuro inmediato pasa por integrarse con herramientas de IA que permitan automatizar, cuantificar y estandarizar la interpretación diagnóstica, reduciendo la subjetividad y aumentando la reproducibilidad. En otras palabras, se trata de aprovechar sus fortalezas mientras se mitigan sus debilidades sistémicas.

1.3. INTELIGENCIA ARTIFICIAL APLICADA AL ULTRASONIDO OVÁRICO: SÍNTESIS CRÍTICA DEL ESTADO DEL ARTE

Durante la última década, la IA ha pasado de ser un concepto teórico para consolidarse como una de las herramientas más disruptivas en medicina. En el ámbito de las imágenes médicas, su impacto ha sido particularmente transformador. La posibilidad de entrenar algoritmos capaces de detectar patrones invisibles al ojo humano, aprender de grandes volúmenes de datos y proporcionar predicciones diagnósticas rápidas y reproducibles ha abierto un horizonte de innovación que afecta directamente a la oncología ginecológica y, en particular, al CO [33, 34].

1.3.1. Evolución histórica de la IA en imágenes médicas

El empleo de algoritmos computacionales en diagnóstico médico no es nuevo: desde la década de los años ochenta se exploraba el potencial de la estadística multivariante y los primeros sistemas expertos. Sin embargo, el verdadero punto de inflexión llegó con la irrupción del aprendizaje profundo (DL). Las redes neuronales convolucionales (CNNs), inspiradas en el procesamiento visual biológico, demostraron a partir de 2012 un rendimiento sin precedentes en competiciones de clasificación de imágenes (ImageNet) con la creación de la red neuronal AlexNet [35]. Su capacidad para extraer automáticamente características jerárquicas complejas, sin necesidad de definir manualmente "descriptores" como ocurría en la visión por computador clásica, cambió por completo el paradigma.

El campo de la radiología adoptó rápidamente estas técnicas. Estudios en mamografía, tomografía de tórax y RM cerebral mostraron que las CNNs podían alcanzar desempeños comparables e incluso superiores a radiólogos expertos en tareas de detección y clasificación de lesiones. Este éxito estimuló su aplicación al ultrasonido.

1.3.2. Primeras aplicaciones en ultrasonido ginecológico

En el caso del CO, las primeras investigaciones se centraron en la clasificación binaria: distinguir entre tumores benignos y malignos a partir de imágenes UST. Estas aproximaciones, aunque simplificadas, marcaron un hito al demostrar que las CNNs eran capaces de superar la variabilidad interobservador y ofrecer métricas de exactitud superiores al 85–90%. En particular, un estudio retrospectivo multicéntrico en China informó un área bajo la curva (AUC) de 0.911, con mejoras significativas en la exactitud diagnóstica de radiólogos cuando contaban con la asistencia del algoritmo [36, 37].

Posteriormente, surgieron modelos más complejos basados en ensamblajes de redes profundas, capaces de alcanzar sensibilidades de hasta 97% con especificidades cercanas al 90%. Estos valores, comparables al desempeño de radiólogos sénior, confirmaron el potencial de la IA como herramienta de apoyo clínico. No obstante, la mayoría de estos estudios eran retrospectivos, con bases de datos limitadas y con validaciones internas, lo que generaba dudas sobre su generalización [38, 39].

1.3.3. Incorporación de modelos multimodales

Una tendencia reciente ha sido la integración de información multimodal: imágenes de ultrasonido junto con datos clínicos (edad, antecedentes familiares, marcadores séricos como CA125 o HE4). Al combinar estas fuentes, los modelos alcanzaron AUC cercanos a 0.98, reduciendo además la tasa de falsos positivos de los lectores humanos. Este enfoque refleja una realidad clínica: el diagnóstico del CO no depende de una sola variable, sino de la integración de múltiples dimensiones.

Otra línea de avance ha sido la incorporación de tareas múltiples en una misma canalización: detección, segmentación y clasificación. Esta aproximación integral ha demostrado ser más robusta, ya que refleja mejor el flujo clínico real. En estudios multicéntricos, estas canalizaciones alcanzaron AUC de 0.941 en validaciones internas y externas, con un rendimiento comparable al de radiólogos sénior incluso en vídeos ecográficos, donde la variabilidad es mayor.

1.3.4. El surgimiento de los Transformers en visión médica

Si bien las CNNs dominaron el campo durante la última década, la aparición del transformador de visión (ViT) cambió nuevamente el panorama. Basados en mecanismos de autoatención, los ViTs procesan la imagen como una secuencia de parches y aprenden relaciones de largo alcance entre regiones distantes. Esta capacidad para modelar el contexto global resulta particularmente útil en ecografía, donde la morfología tumoral debe interpretarse no solo a nivel de detalle local, sino también considerando patrones más amplios [40].

Los ViTs han mostrado un rendimiento comparable al de las CNNs en tareas de clasificación y segmentación, pero con una limitación importante: su menor capacidad para capturar detalles locales finos. En el CO, donde las diferencias entre lesiones benignas y malignas pueden ser extremadamente sutiles, esta debilidad se convierte en un obstáculo relevante.

1.3.5. Modelos híbridos: combinando lo mejor de dos mundos

Para superar estas limitaciones, han emergido modelos híbridos que combinan CNNs y ViTs. La lógica es clara: las CNNs son, por el momento,

insuperables en la extracción de características locales, mientras que los ViTs destacan en el modelado global de relaciones espaciales. Al integrar ambos enfoques, se espera obtener un sistema más robusto y preciso [41].

Hasta ahora, la mayoría de los modelos híbridos han utilizado estrategias de fusión tardía, en las que las características locales y globales se combinan en las capas finales de la red. Sin embargo, este enfoque limita la interacción temprana entre ambas representaciones, lo que podría desaprovechar el potencial de complementariedad. En este punto, la propuesta de esta tesis se inscribe como una innovación conceptual, al permitir que la integración entre detalle local y contexto global ocurra desde las primeras etapas del procesamiento.

1.3.6. Limitaciones metodológicas en la literatura

A pesar de los resultados alentadores, la literatura sobre IA en ultrasonido ovárico presenta limitaciones importantes:

1. **Heterogeneidad de los conjuntos de datos:** los tamaños muestrales varían desde unas decenas hasta unos pocos miles de imágenes, lo que condiciona la potencia estadística. Además, existe un desbalance de clases (más lesiones benignas que malignas), que afecta la estabilidad de los modelos.
2. **Falta de validación externa:** la mayoría de los estudios reportan validaciones internas, con desempeño inflado que no se reproduce en cohortes externas.
3. **Segmentación y anotación subjetiva:** la variabilidad en la delimitación de las lesiones introduce ruido en el entrenamiento y evaluación.
4. **Ausencia de métricas de calibración:** casi ningún estudio reporta curvas de confiabilidad o análisis de incertidumbre, lo que dificulta la implementación clínica.
5. **Sesgo de publicación:** predominan los estudios con resultados positivos, lo que puede distorsionar la percepción real del campo.

1.3.7. Crítica del estado del arte

En conjunto, el estado actual de la IA aplicada al UST es prometedor pero inmaduro. Los algoritmos han demostrado desempeños altos en entornos

controlados, pero su traslado a la práctica clínica real aún enfrenta múltiples barreras. No se trata únicamente de alcanzar una AUC elevada, sino de garantizar reproducibilidad, explicabilidad y calibración en poblaciones diversas.

La literatura sugiere que más que una superioridad consistente de una familia arquitectónica sobre otra (CNN vs. ViT), lo que realmente determina el éxito clínico es el diseño integral del flujo de trabajo, la calidad de las anotaciones, la validación externa y la integración multimodal. La incorporación de técnicas de interpretabilidad Grad-CAM (mapeo de activación de clase ponderado por gradiente), junto con métricas de calibración probabilística y análisis de incertidumbre, se perfila como requisito indispensable para que los modelos ganen aceptación entre clínicos y pacientes.

1.3.8. Perspectiva crítica y lugar de esta tesis

En este contexto, el aporte de la presente investigación se ubica en la frontera entre lo exploratorio y lo traslacional. Al proponer un modelo híbrido CNN–ViT con fusión temprana adaptativa, se busca superar la fragmentación metodológica de la literatura previa y avanzar hacia un sistema que no solo sea preciso, sino también explicable, calibrado y aplicable en la práctica clínica. La validación de este enfoque frente a marcos estandarizados como IOTA y O-RADS constituye un paso crucial para evaluar si la IA puede integrarse como complemento real en la toma de decisiones ginecológicas.

1.4. BRECHAS METODOLÓGICAS QUE BLOQUEAN LA TRASLACIÓN CLÍNICA

La literatura sobre IA aplicada al UST en CO ha generado un entusiasmo comprensible debido a los resultados alentadores en términos de exactitud, sensibilidad, especificidad y AUC. Sin embargo, ese entusiasmo debe ser matizado. La gran mayoría de los modelos propuestos se encuentran todavía en fases exploratorias, con un nivel de madurez tecnológica que dista de lo necesario para su implementación clínica rutinaria. En este sentido, existen brechas metodológicas profundas que obstaculizan la traslación de la IA desde el laboratorio hasta la consulta ginecológica. A continuación, se analizan de manera crítica los principales problemas, sus consecuencias y las posibles vías de solución.

1.4.1. Dependencia de la segmentación y ruido de anotación

Una de las limitaciones más críticas es la fuerte dependencia de los modelos respecto al proceso de segmentación. La ecografía es, por naturaleza, una modalidad de imagen con bajo contraste, artefactos frecuentes y bordes difusos. Bajo estas condiciones, la delimitación precisa de las regiones de interés (ROIs) resulta altamente subjetiva.

Cuando la segmentación se realiza manualmente, la variabilidad interobservador e intraobservador se convierte en una fuente de error sistemática. Diferentes ecografistas, o incluso un mismo ecografista en momentos distintos, pueden marcar contornos diferentes para la misma lesión. Este "ruido de etiqueta" se propaga al entrenamiento de los modelos, generando representaciones menos estables y resultados con menor reproducibilidad.

La evidencia cuantitativa sintetizada en revisiones confirma que el tipo de segmentación es un factor determinante del rendimiento. En una de las metarregresiones más recientes, realizada en este proyecto, la estrategia de segmentación emergió como un predictor independiente de la exactitud de los modelos ($\beta = 0.0656$; $p = 0.007$), mientras que otras covariables, como el tamaño muestral o la familia arquitectónica empleada, no alcanzaron significación estadística [34].

1.4.2. Falta de estandarización en protocolos de adquisición y anotación

Más allá de la segmentación, existe una ausencia generalizada de estandarización en los protocolos de adquisición y anotación de imágenes. Factores aparentemente menores, como la ganancia del ecógrafo (parámetro que controla cuán brillantes u oscuras se ven las imágenes), la presión ejercida con la sonda o la elección de los planos de corte, pueden alterar de manera sustancial las características de la imagen.

En la práctica, los estudios publicados rara vez detallan con precisión estos aspectos técnicos, lo que limita la reproducibilidad. Además, las guías de anotación suelen variar de un grupo a otro, generando criterios dispares sobre qué incluir o excluir en la ROI (p. ej., tabiques finos, sombras acústicas, zonas de hemorragia).

Esta heterogeneidad metodológica explica en parte la dispersión de resultados observada en la literatura y dificulta la comparación entre estudios [42].

1.4.3. Tamaño muestral insuficiente y desbalance de clases

Otra limitación recurrente es el tamaño muestral reducido de muchos estudios. No es infrecuente encontrar trabajos con apenas unas decenas o centenas de imágenes, claramente insuficientes para entrenar modelos de DL con millones de parámetros [43]. Incluso en cohortes más grandes, con varios miles de imágenes, la presencia de clases desbalanceadas constituye un problema mayor [44, 45].

En la práctica clínica, las lesiones benignas son mucho más frecuentes que las malignas. Esta desproporción se refleja en los conjuntos de imágenes, generando modelos sesgados hacia la clase mayoritaria [46, 47]. Como resultado, los algoritmos tienden a clasificar en exceso como benignas las lesiones, logrando una aparente alta exactitud global, pero con una sensibilidad insuficiente para detectar tumores malignos. Las técnicas de sobremuestreo o de aumento de datos pueden mitigar este problema en el entrenamiento, pero en la evaluación la desproporción sigue comprometiendo el rendimiento [48].

1.4.4. Falta de validaciones externas y el problema del dataset shift

La validación externa constituye uno de los pilares fundamentales para la traslación clínica [49]. Sin embargo, la mayoría de los estudios sobre IA en UST ovárico se limitan a validaciones internas (*train-test split* o validación cruzada). Estas estrategias pueden inflar artificialmente las métricas de rendimiento, ya que el algoritmo se evalúa en un entorno muy similar al de su entrenamiento [50].

Cuando los modelos se aplican a datos de otros hospitales, con diferentes equipos ecográficos, operadores y poblaciones, el rendimiento suele caer de manera significativa. Este fenómeno, conocido como desplazamiento de la distribución de los datos (*dataset shift*), refleja la incapacidad de los modelos para generalizar más allá del dominio en el que fueron entrenados. En la práctica clínica, este problema se traduce en un riesgo de errores diagnósticos si los algoritmos no son adaptados y recalibrados a cada nuevo contexto [51].

1.4.5. Ausencia de calibración probabilística y análisis de incertidumbre

Otro aspecto críticamente ausente en la mayoría de los estudios es la calibración probabilística. Los algoritmos de IA generan puntuaciones que se interpretan como probabilidades, pero en la práctica no siempre existe correspondencia entre la probabilidad estimada y la probabilidad real. Un modelo mal calibrado puede asignar una confianza del 95% a un diagnóstico incorrecto, lo que genera un riesgo considerable si se integra en la práctica clínica [52].

De manera complementaria, la falta de análisis de incertidumbre limita la capacidad de los clínicos para decidir cuándo confiar en la predicción de un modelo y cuándo recurrir a su propio juicio. El análisis de entropía de las predicciones o el uso de técnicas bayesianas permitirían establecer umbrales de confianza más seguros, reservando para revisión experta los casos con mayor incertidumbre. Sin este tipo de herramientas, la integración clínica de la IA es poco viable [53, 54].

1.4.6. Problemas de interpretabilidad y la "caja negra"

Uno de los argumentos más repetidos contra la implementación clínica de la IA es su carácter de "caja negra". Muchos modelos de DL ofrecen predicciones sin una explicación clara de cómo se alcanzó el resultado. En un contexto clínico, esta opacidad es problemática tanto desde el punto de vista ético como práctico: los médicos necesitan comprender las razones de un diagnóstico para confiar en él y explicarlo a las pacientes [55].

La ausencia de interpretabilidad limita la aceptación de los modelos, independientemente de su rendimiento cuantitativo. Aunque técnicas como Grad-CAM o SHAP permiten visualizar qué regiones de la imagen contribuyeron a la decisión, su adopción aún no es sistemática en los estudios publicados. Esta falta de explicabilidad constituye una brecha que bloquea la traslación clínica y que debe abordarse de manera prioritaria [56].

1.4.7. Sesgo de publicación y heterogeneidad en los reportes

El campo de la IA en medicina adolece también de un sesgo de publicación favorable a los resultados positivos. Los estudios que reportan desempeños modestos o negativos tienen menos probabilidades de ser publicados, lo que

distorsiona la percepción general del campo. A esto se suma la heterogeneidad en los reportes: métricas diferentes, umbrales distintos, ausencia de intervalos de confianza o de análisis estadísticos adecuados [57].

Para resolver este problema, organismos internacionales han propuesto guías específicas como CONSORT-AI (Extensión CONSORT para intervenciones de inteligencia artificial), SPIRIT-AI (Extensión SPIRIT para protocolos de ensayos con intervenciones de IA) y TRIPOD-AI (Extensión TRIPOD para el reporte transparente de modelos multivariables de predicción que incorporan IA), que buscan estandarizar la forma en que se diseñan, ejecutan y reportan los estudios de IA en salud. Sin embargo, la adopción de estas guías aún es limitada, especialmente en el área de la ecografía ginecológica [58–60].

1.4.8. Consecuencias clínicas y riesgo de implementación prematura

Las brechas metodológicas descritas no son solo problemas académicos: tienen consecuencias directas en la práctica clínica. Implementar algoritmos no validados externamente, mal calibrados o poco interpretables puede generar daños significativos, desde diagnósticos erróneos hasta decisiones terapéuticas inapropiadas. Además, la confianza de los clínicos en la IA podría verse socavada si los primeros sistemas implementados fallan en la práctica, lo que retrasaría la adopción de tecnologías más maduras en el futuro [61, 62].

1.4.9. Perspectiva crítica y vías de solución

Superar estas brechas requiere una estrategia integral:

- **Estandarización de protocolos de adquisición y segmentación**, alineados con guías internacionales como IOTA y O-RADS, pero complementados con métricas de fiabilidad interobservador.
- **Entrenamiento con grandes cohortes multicéntricas**, que permitan capturar la variabilidad de equipos, operadores y poblaciones.
- **Validación externa rigurosa**, incluyendo pruebas prospectivas en tiempo real.
- **Implementación de calibración probabilística y análisis de incertidumbre** como parte del reporte estándar de métricas.

- **Adopción sistemática de técnicas de interpretabilidad**, que hagan transparentes las decisiones de los modelos.
- **Cumplimiento estricto de guías internacionales de reporte (CONSORT-AI, TRIPOD-AI)** para garantizar comparabilidad y transparencia.

Solo con estos pasos será posible avanzar de la prueba de concepto a la traslación clínica, reduciendo la brecha entre innovación tecnológica y beneficio real para las pacientes.

1.5. IMPACTO SOCIOECONÓMICO Y CARGA GLOBAL

El CO no sólo es un problema clínico y científico de enorme magnitud, sino que también representa una carga socioeconómica considerable para los sistemas de salud, las familias y la sociedad en su conjunto. La elevada letalidad del CO y su diagnóstico tardío implican tratamientos más agresivos, múltiples hospitalizaciones, recurrencias frecuentes y cuidados prolongados que generan costes directos e indirectos significativos. El análisis del impacto socioeconómico del CO, por tanto, resulta indispensable para dimensionar el problema y justificar la urgencia de innovaciones diagnósticas que permitan mejorar la detección temprana y reducir tanto la mortalidad como los costes asociados [63–65].

1.5.1. Costes directos del tratamiento

Los costes directos se refieren a los gastos sanitarios asociados de manera inmediata con la atención del CO: consultas especializadas, estudios de imagen, pruebas de laboratorio, intervenciones quirúrgicas, hospitalizaciones, quimioterapia, terapias dirigidas y cuidados de soporte. Estudios de coste de enfermedad en países de altos ingresos muestran que el CO se sitúa entre los tumores ginecológicos con mayor coste anual por paciente, con gastos particularmente elevados en las fases de diagnóstico inicial y de enfermedad metastásica o recurrente, donde los costes totales pueden superar con facilidad los 100 000–200 000 euros anuales por persona cuando se incluyen hospitalizaciones y terapias sistémicas [66, 67].

La cirugía de citorreducción, considerada el pilar del tratamiento, concentra una proporción importante de estos costes directos, dado que requiere centros

especializados, equipos quirúrgicos altamente entrenados, tiempo operatorio prolongado, estancia en unidades de cuidados intensivos y una hospitalización postoperatoria extensa. A ello se añaden los costes de la quimioterapia basada en platinos y taxanos, y, más recientemente, de los inhibidores de PARP y otros fármacos dirigidos empleados como mantenimiento, que han incrementado notablemente el gasto farmacéutico asociado al CO (coste medio por 30 días de 11 000 – 14 000 €) [68, 69].

En países de ingresos bajos y medios, como los de Hispanoamérica, Asia y África, estos tratamientos suelen estar fuera del alcance de los sistemas públicos de salud y resultan inasequibles para la mayoría de las familias. Como consecuencia, muchas pacientes no reciben la atención óptima, lo que contribuye a la elevada mortalidad observada en estas regiones [70].

1.5.2. Costes indirectos y carga para las familias

A los costes directos se suman los costes indirectos, que incluyen la pérdida de productividad laboral de las pacientes y de sus cuidadores, la reducción o abandono del empleo, la disminución de ingresos y la reasignación de tareas domésticas y de cuidados. En muchos contextos, el diagnóstico de CO en una mujer en edad laboral supone la pérdida de una fuente central de ingresos para el hogar, con un impacto que se extiende más allá de la paciente e involucra a toda la unidad familiar [71].

La incapacidad laboral temporal o permanente implica una pérdida de ingresos que afecta no solo a las pacientes, sino también a sus hogares. En contextos donde no existen seguros de salud o sistemas de seguridad social robustos, esta pérdida de ingresos puede tener efectos devastadores. Además, los cuidadores, a menudo familiares cercanos, deben reducir su jornada laboral o abandonar sus empleos para atender a la paciente, lo que amplifica la carga económica indirecta.

El impacto emocional es igualmente relevante: el diagnóstico de un CO avanzado suele ir acompañado de un mal pronóstico, con múltiples recurrencias y un tratamiento prolongado que erosiona el bienestar psicológico tanto de las pacientes como de sus familiares. Los costes intangibles, como el sufrimiento y la pérdida de calidad de vida, son difíciles de cuantificar, pero representan una dimensión crítica de la carga global de la enfermedad.

1.5.3. Carga global de enfermedad en DALYs y QALYs

La carga de enfermedad del CO puede cuantificarse en términos de DALYs, que combinan los años de vida perdidos por muerte prematura y los años vividos con discapacidad, así como en años de vida ajustados por calidad (QALYs) utilizados en evaluaciones de coste-efectividad. Estudios recientes basados en la carga global de enfermedad muestran que el CO genera en torno a 5–6 millones de DALYs anuales, con un incremento superior al 90% en el número absoluto de DALYs entre 1990 y 2019, reflejando tanto el envejecimiento poblacional como la persistencia del diagnóstico en estadios avanzados [72].

En comparación con otros cánceres ginecológicos, el CO tiende a concentrar una mayor proporción de los DALYs por muerte prematura, dado que la supervivencia a cinco años sigue siendo limitada en estadios avanzados y muchas pacientes experimentan múltiples recurrencias que deterioran la calidad de vida. Esta combinación de alta letalidad y discapacidad asociada al tratamiento hace que el impacto del CO sobre los DALYs y QALYs sea especialmente negativo incluso en países con sistemas de salud avanzados [73].

1.5.4. Desigualdades regionales y sociales

El impacto socioeconómico del CO no es homogéneo. Existen marcadas desigualdades regionales y socioeconómicas que agravan la carga de la enfermedad. En países de ingresos altos, el acceso a cirugía especializada y terapias dirigidas permite alcanzar tasas de supervivencia global superiores al 45–50% a cinco años. En contraste, en países de ingresos bajos y medios, la supervivencia rara vez supera el 30%, y en algunos casos desciende al 20%.

Estas brechas se explican en parte por diferencias en el acceso a diagnóstico temprano, cirugía oncológica especializada, quimioterapia basada en platinos y terapias dirigidas. A ello se añaden desigualdades sociales y étnicas: en varios países de altos ingresos, las mujeres pertenecientes a minorías raciales o con menor nivel socioeconómico presentan peor supervivencia, incluso tras ajustar por estadio y tratamiento, lo que pone de manifiesto la influencia de determinantes sociales de la salud [74].

En Hispanoamérica y otras regiones de ingresos medios, factores como la fragmentación de los sistemas de salud, la limitada cobertura financiera, la concentración de servicios oncológicos en grandes ciudades y las barreras geográficas y culturales agravan aún más la inequidad. Además, se ha descrito la contribución de determinantes sociales y ambientales específicos (como la pobreza, la informalidad laboral y la exposición desigual a factores de riesgo) en el perfil de riesgo del CO en Sudamérica [75].

1.5.5. Impacto en los sistemas de salud

El CO representa también un desafío para la sostenibilidad de los sistemas de salud, tanto por el elevado coste de los tratamientos como por la necesidad de infraestructura especializada para cirugía compleja, quimioterapia, terapias dirigidas y cuidados paliativos. A escala global, las neoplasias se asocian a cientos de miles de millones de dólares anuales en gasto sanitario directo, y el CO contribuye de manera desproporcionada a estos costes dentro del grupo de cánceres ginecológicos.

En países de ingresos medios, la combinación de envejecimiento poblacional, transición epidemiológica y recursos limitados genera tensión sobre presupuestos ya restringidos, obligando a tomar decisiones difíciles sobre la asignación de recursos entre el CO y otras enfermedades de alta carga, como las cardiovasculares, la diabetes o las infecciones emergentes [76].

En los países de ingresos altos, el debate es si los nuevos tratamientos (como el mantenimiento con inhibidores de PARP o combinaciones con antiangiogénicos) valen lo que cuestan. En general, ayudan a que el cáncer tarde más en avanzar, pero no siempre está claro que hagan que la paciente viva mucho más tiempo. Por eso, cuando se analizan los costes, se concluye que a veces solo compensa usarlos si el sistema de salud acepta pagar un precio alto por ese beneficio y si se aplican en pacientes con mayor riesgo o con más probabilidad de responder. Esto muestra que hacen falta mejores herramientas de estratificación para usar el tratamiento de forma más adecuada [77].

1.5.6. Necesidad de innovación diagnóstica como estrategia de sostenibilidad

En este contexto, la innovación en diagnóstico precoz no es solo una aspiración clínica, sino una estrategia de sostenibilidad. Mejorar la detección temprana del CO podría reducir la proporción de casos diagnosticados en estadios avanzados, disminuir la necesidad de múltiples líneas de tratamiento de alto coste y mitigar la toxicidad financiera para pacientes, familias y sistemas de salud. Un reciente estudio internacional, de Hutchinson *et al.* en 2025, sobre la carga socioeconómica del CO en 11 países concluye que más del 90% de las pérdidas económicas se atribuyen a mortalidad prematura, lo que resalta el potencial impacto de intervenciones que retrasen o eviten la muerte por CO [65].

En este marco, el desarrollo de modelos de IA aplicados al ultrasonido, capaces de mejorar la clasificación de masas anexiales y apoyar la toma de decisiones en la práctica clínica real, se perfila como una vía prometedora para reducir la carga socioeconómica del CO, al favorecer diagnósticos más tempranos y decisiones terapéuticas mejor informadas.

1.6. HISTORIA Y EVOLUCIÓN DEL ULTRASONIDO EN ONCOLOGÍA GINECOLÓGICA

El ultrasonido es, sin lugar a duda, una de las tecnologías que más ha transformado la práctica médica en el último medio siglo. En ginecología, su introducción marcó un punto de inflexión en la capacidad diagnóstica y en la toma de decisiones clínicas. Para comprender el lugar central que ocupa hoy en la oncología ginecológica, resulta imprescindible recorrer su evolución histórica, desde las primeras experiencias hasta las aplicaciones actuales, y reflexionar críticamente sobre sus logros, limitaciones y perspectivas de futuro [78].

1.6.1. Orígenes del ultrasonido médico

El ultrasonido médico comenzó a explorarse en la década de 1940, inspirado en las tecnologías de sonar utilizadas en la Segunda Guerra Mundial. Los primeros equipos eran rudimentarios y ofrecían imágenes de muy baja resolución, limitadas principalmente a la representación de interfaces tisulares. En los años 50 y 60 se empezaron a realizar las primeras aplicaciones clínicas en obstetricia, permitiendo

visualizar el feto en desarrollo. Este hito revolucionó la atención prenatal y abrió el camino para la incorporación del ultrasonido en otras ramas de la medicina [79].

En ginecología, los primeros usos estuvieron orientados a la evaluación de masas pélvicas y a la identificación de embarazos ectópicos. Sin embargo, la resolución de las imágenes era insuficiente para caracterizar con detalle las estructuras ováricas [80].

1.6.2. La revolución del ultrasonido transvaginal

El verdadero salto se produjo en la década de los ochenta con la introducción de las sondas transvaginales. A diferencia de la ecografía abdominal, el acceso transvaginal permitía situar la sonda más cerca de los órganos pélvicos, aumentando de manera significativa la resolución y la nitidez de las imágenes. Esta innovación supuso un cambio radical en la evaluación de las masas anexiales y permitió, por primera vez, realizar descripciones detalladas de la morfología ovárica.

Con el UST fue posible identificar características como el grosor de la pared quística, la presencia de septos, papilas, componentes sólidos o ascitis, que se convirtieron en criterios fundamentales para diferenciar lesiones benignas de malignas. Desde entonces, el UST se consolidó como la técnica de primera línea en la evaluación de tumores ováricos, desplazando a otras modalidades más costosas y menos accesibles [81].

1.6.3. Intentos de estandarización diagnóstica

A lo largo de esta evolución tecnológica, se hizo evidente la necesidad de estandarizar la interpretación ecográfica. La subjetividad en la descripción de los hallazgos y la variabilidad interobservador generaban inconsistencias que limitaban la utilidad clínica del ultrasonido [82].

El grupo IOTA surgió en respuesta a esta problemática. Sus estudios multicéntricos, iniciados en los años 90, lograron definir reglas simples y modelos de predicción que homogeneizaron el lenguaje y mejoraron la reproducibilidad. Las reglas IOTA, basadas en criterios morfológicos claramente definidos,

alcanzaron sensibilidades y especificidades superiores al 85% en múltiples validaciones internas y externas [83].

El sistema O-RADS, desarrollado posteriormente, buscó ir un paso más allá, inspirándose en el modelo BI-RADS utilizado en mamografía. O-RADS clasificó las lesiones en categorías de riesgo, con recomendaciones claras de conducta clínica para cada una. Aunque su adopción global aún está en proceso, representa un esfuerzo significativo hacia la estandarización universal.

1.6.4. Limitaciones persistentes a pesar de la evolución

A pesar de estos avances, el UST sigue enfrentando limitaciones importantes. La dependencia del operador continúa siendo su principal debilidad. En estudios multicéntricos, la variabilidad interobservador puede llegar al 20–25% en la clasificación de lesiones complejas. Además, incluso con sistemas estandarizados, la especificidad diagnóstica sigue siendo insuficiente en tumores *borderline* o de bajo volumen [84, 85].

Otro problema es la falta de validación externa amplia de los sistemas estandarizados en poblaciones diversas. La mayoría de las validaciones se han realizado en Europa, lo que limita su transportabilidad a contextos de Hispanoamérica, Asia o África, donde las condiciones epidemiológicas y los recursos sanitarios son distintos [86].

1.6.5. El lugar del ultrasonido en la era de la inteligencia artificial

Hoy, el UST se encuentra en un punto de inflexión. Su historia demuestra una progresión continua hacia modalidades más complejas y hacia la búsqueda de estandarización. Sin embargo, la subjetividad persiste y constituye un obstáculo para alcanzar la precisión diagnóstica necesaria en el CO [87, 88].

La irrupción de la inteligencia artificial ofrece la posibilidad de dar el siguiente salto evolutivo: pasar de la interpretación subjetiva y dependiente del operador a un análisis automatizado, cuantitativo y reproducible. En este sentido, la IA puede entenderse como una extensión natural de la trayectoria histórica del ultrasonido en oncología ginecológica: una herramienta destinada a potenciar sus

fortalezas: accesibilidad, bajo coste, inocuidad, mientras se corrigen sus debilidades: variabilidad, falta de reproducibilidad y subjetividad.

1.7. MARCO LEGAL Y BIOÉTICO DE LA INTELIGENCIA ARTIFICIAL EN SALUD

El despliegue de la IA en medicina no se limita a un desafío tecnológico o científico. Su aplicación conlleva profundas implicaciones legales, regulatorias y bioéticas que condicionan su adopción en la práctica clínica. El hecho de que la IA se emplee en un área tan sensible como el diagnóstico del CO obliga a analizar este marco de manera crítica, dado que las decisiones clínicas derivadas de sus predicciones pueden tener consecuencias vitales para las pacientes.

1.7.1. Marco regulatorio internacional: Europa y Estados Unidos

En la Unión Europea, la aprobación del Reglamento de Inteligencia Artificial (2024) constituyó un hito. Este marco normativo clasifica a los sistemas de IA en categorías de riesgo, siendo los algoritmos médicos catalogados como "de alto riesgo". Esto implica requisitos estrictos [89]:

- **Transparencia** en el diseño y funcionamiento del modelo.
- **Trazabilidad** de los datos utilizados para entrenar los algoritmos.
- **Validación externa** en condiciones reales de uso.
- **Supervisión humana obligatoria**, para evitar que las decisiones clínicas se deleguen exclusivamente en sistemas automatizados.

En paralelo, la Agencia Europea de Medicamentos (EMA) y los organismos nacionales de regulación sanitaria han comenzado a explorar cómo integrar estos requisitos con la normativa vigente para dispositivos médicos, ya que la IA en salud suele clasificarse en la categoría de *software* como dispositivo médico (SaMD).

En Estados Unidos, la Administración de Alimentos y Medicamentos de Estados Unidos (FDA) ha publicado guías específicas para el registro de algoritmos médicos. La FDA clasifica a la IA diagnóstica dentro de la categoría SaMD y exige evidencia clínica sólida para su aprobación. Además, en 2021 introdujo el concepto de "aprendizaje adaptativo", reconociendo que los algoritmos de IA pueden actualizarse de forma continua a medida que reciben nuevos datos, lo que plantea

retos regulatorios inéditos: ¿cómo certificar un algoritmo que está en permanente evolución?

Ambas regiones han enfatizado la necesidad de garantizar que los algoritmos sean seguros, eficaces y justos, es decir, que no reproduzcan sesgos de género, etnia o nivel socioeconómico en sus predicciones [90, 91].

1.7.2. Responsabilidad legal y consentimiento informado

Un aspecto crítico es la definición de la responsabilidad legal. En medicina tradicional, la responsabilidad de un error diagnóstico recae en el profesional tratante. Sin embargo, cuando un algoritmo de IA participa en la decisión, la cadena de responsabilidad se complica. Si un modelo clasifica erróneamente un tumor ovárico benigno como maligno, llevando a una cirugía innecesaria, ¿quién es responsable? [92]

Existen tres posiciones principales:

1. **Responsabilidad del médico**: el algoritmo es una herramienta auxiliar y la decisión final sigue siendo humana.
2. **Responsabilidad compartida**: el médico responde parcialmente, pero el fabricante del software también asume responsabilidad si el fallo se debió a deficiencias del algoritmo.
3. **Responsabilidad del desarrollador**: en casos donde el sistema funciona de manera semiautónoma, la empresa que lo creó podría ser la responsable principal.

El consentimiento informado también debe adaptarse a la era digital. Las pacientes tienen derecho a saber si su diagnóstico fue asistido por un sistema de IA, cuáles son sus limitaciones y qué grado de supervisión humana se ejerció. Ocultar esta información podría considerarse una violación al principio de autonomía [93].

1.7.3. Consideraciones bioéticas: principios de la medicina y ética digital

La bioética ofrece un marco fundamental para analizar la incorporación de la IA en salud. Los principios clásicos de Beauchamp y Childress, beneficencia, no

maleficencia, autonomía y justicia, siguen siendo aplicables, pero adquieren nuevas dimensiones en el contexto de la IA [94].

- **Beneficencia**: la IA debe contribuir a mejorar el diagnóstico, reducir errores y beneficiar a las pacientes.
- **No maleficencia**: los riesgos derivados de diagnósticos erróneos, sesgos o falta de interpretabilidad deben minimizarse.
- **Autonomía**: las pacientes deben ser informadas de manera transparente sobre el papel de la IA en su diagnóstico.
- **Justicia**: el acceso a estas tecnologías no debe profundizar las desigualdades entre quienes pueden costear servicios de alta especialidad y quienes dependen de sistemas de salud limitados.

La ética digital añade otros principios relevantes:

- **Transparencia**: los modelos deben ser explicables.
- **Responsabilidad**: los desarrolladores deben rendir cuentas de los sesgos o fallos de los algoritmos.
- **Privacidad**: los datos clínicos utilizados para entrenar la IA deben protegerse mediante normativas de protección de datos [95].

1.7.4. Sesgos algorítmicos y equidad

Un problema cada vez más reconocido es el de los sesgos algorítmicos. Si un modelo se entrena con datos procedentes principalmente de poblaciones europeas, su rendimiento puede ser inferior al aplicarse en mujeres hispanoamericanas o africanas, reproduciendo inequidades existentes. Estos sesgos no solo afectan la precisión, sino que también plantean un dilema ético: la tecnología que debía reducir desigualdades podría terminarlas ampliando [96].

Por ello, la diversidad en los datos de entrenamiento en múltiples contextos no son solo requisitos técnicos, sino también éticos y legales.

1.7.5. Transparencia y explicabilidad

La explicabilidad es uno de los pilares de la regulación internacional. No basta con que un algoritmo ofrezca un diagnóstico con una AUC elevada; debe ser capaz de justificar su decisión de manera comprensible. Herramientas como Grad-

CAM permiten visualizar qué regiones de la imagen fueron determinantes en la clasificación, ofreciendo una transparencia mínima que favorece la aceptación clínica [97, 98].

Desde el punto de vista legal, la ausencia de explicabilidad podría interpretarse como una falta de cumplimiento con el derecho a la información de la paciente. Desde el punto de vista bioético, socava la confianza médico-paciente [99].

1.8. DESAFÍOS DE IMPLEMENTACIÓN CLÍNICA

El paso de la IA desde el laboratorio de investigación hasta la consulta ginecológica constituye uno de los mayores retos de la medicina contemporánea. Aunque los modelos aplicados al ultrasonido en CO han demostrado métricas de rendimiento alentadoras en estudios retrospectivos, su integración en la práctica clínica cotidiana enfrenta múltiples barreras de naturaleza técnica, organizativa, cultural y regulatoria [100].

1.8.1. Integración con sistemas clínicos existentes

Una primera barrera es la integración de los sistemas de IA con la infraestructura digital ya existente en los hospitales, que suele estar conformada por combinaciones heterogéneas de sistemas de archivo y comunicación de imágenes (PACS), sistemas de información radiológica (RIS) y expedientes clínicos electrónicos. Incorporar un modelo de ayuda al diagnóstico basado en ultrasonido requiere compatibilidad con formatos estándar (p. ej., DICOM), interoperabilidad con ecógrafos de distintos fabricantes y capacidad de procesar imágenes en tiempo real sin introducir retrasos en el flujo asistencial [101].

En muchos hospitales, especialmente en países de ingresos bajos y medios, los sistemas PACS son incompletos o inexistentes, lo que hace inviable implementar modelos que dependen de grandes volúmenes de imágenes digitalizadas. Esta limitación técnica amenaza con dejar fuera de la revolución de la IA a los entornos donde más se necesita.

1.8.2. Validación prospectiva en escenarios reales

Un segundo desafío crucial es la validación prospectiva. La mayoría de los estudios publicados son retrospectivos: utilizan bases de datos ya existentes, cuidadosamente seleccionadas, lo que genera resultados optimistas que rara vez se replican en condiciones reales [102, 103].

La validación prospectiva implica aplicar el algoritmo en tiempo real, en pacientes consecutivos, sin exclusión de casos difíciles ni artefactos. Solo así se puede medir su verdadero impacto en la práctica clínica. Además, debe evaluarse no solo la precisión diagnóstica, sino también:

- El tiempo añadido o ahorrado en la consulta.
- La aceptación por parte de los médicos y pacientes.
- El impacto en las decisiones terapéuticas.
- Los desenlaces clínicos a largo plazo.

La falta de estudios prospectivos constituye una de las principales brechas entre la investigación académica y la aplicación práctica.

1.8.3. Formación médica y competencias digitales

Un aspecto frecuentemente subestimado es la necesidad de formación médica en competencias digitales. La mayoría de los programas de residencia en ginecología no incluyen módulos sobre inteligencia artificial, análisis de datos o bioinformática. Esto genera una brecha entre el desarrollo tecnológico y las habilidades de los profesionales que deben utilizarlo.

Si un ginecólogo no comprende cómo funciona un modelo, cuáles son sus limitaciones y cómo interpretar sus salidas, existe un riesgo elevado de uso indebido o de desconfianza. Incorporar la alfabetización digital en los planes de estudio de medicina y en los programas de educación médica continua será imprescindible para una adopción segura y efectiva de la IA [104].

1.8.4. Barreras económicas y desigualdades de acceso

La implementación clínica también está condicionada por factores económicos. Aunque en teoría los algoritmos pueden reducir costes al evitar cirugías innecesarias, en la práctica su adopción inicial exige inversiones en

infraestructura, digitalización de imágenes, servidores de alto rendimiento y licencias de software [105].

En países de ingresos altos, estas inversiones pueden asumirse como parte de la modernización hospitalaria. En países de ingresos bajos y medios, sin embargo, corren el riesgo de ser inaccesibles, lo que profundizaría las desigualdades ya existentes. Paradójicamente, las regiones que más se beneficiarían de un diagnóstico asistido por IA, por tener menos especialistas experimentados, son las que tienen menos posibilidades de implementarlo.

1.8.5. Fiabilidad técnica y robustez

Los algoritmos de IA suelen entrenarse en condiciones controladas. Sin embargo, en la práctica clínica las imágenes pueden tener artefactos, baja resolución, presencia de dispositivos médicos o variaciones en la técnica de adquisición. La robustez frente a estas imperfecciones es esencial para su aplicación. Si un modelo falla cada vez que la imagen es subóptima, se convierte en un riesgo más que en un apoyo. Para ganar robustez, los algoritmos deben entrenarse con conjunto de datos diversos que incluyan variaciones de calidad, diferentes equipos y múltiples operadores. Esta necesidad de diversidad aumenta la complejidad del entrenamiento, pero es indispensable para garantizar fiabilidad clínica [106].

1.9. RELEVANCIA INTERDISCIPLINARIA E IMPACTO SOCIAL

El abordaje del CO y el desarrollo de modelos de IA aplicados al ultrasonido no pueden entenderse desde una única disciplina. Se trata de un problema genuinamente interdisciplinario, en el que convergen la oncología ginecológica, la radiología, la ingeniería biomédica, la informática, la bioética, la salud pública y las ciencias sociales. Cada una de estas áreas aporta preguntas, métodos y marcos interpretativos sin los cuales el proyecto quedaría incompleto [33].

La medicina y la oncología ginecológica proporcionan el contexto clínico, definen los desenlaces relevantes (supervivencia, calidad de vida, reducción de cirugías innecesarias) y garantizan que las preguntas algorítmicas estén alineadas con necesidades reales de las pacientes y de los profesionales. La radiología y la

ecografía ginecológica aportan la experiencia en adquisición de imágenes, estandarización de protocolos y uso de sistemas de clasificación como IOTA u O-RADS [16, 83].

La ingeniería biomédica y la informática médica contribuyen con el diseño de la infraestructura de adquisición, almacenamiento y procesamiento de imágenes, así como con la integración de los modelos en los sistemas hospitalarios (PACS, RIS). Las ciencias de datos y el aprendizaje profundo permiten desarrollar y entrenar modelos híbridos, diseñar estrategias de segmentación automática y aplicar técnicas de validación rigurosas, como la validación cruzada anidada o la evaluación multicéntrica [41, 107].

La bioética y el derecho sanitario analizan las implicaciones éticas, legales y sociales de incorporar sistemas de IA a la toma de decisiones sobre cirugías oncológicas de alto impacto, evaluando cuestiones como la equidad, la explicabilidad, la privacidad de los datos y la distribución de responsabilidades en caso de errores diagnósticos [71].

Desde la perspectiva clínica, el impacto potencial de un modelo de IA bien validado se traduce en al menos tres dimensiones: reducción de cirugías innecesarias, incremento de la detección temprana y mayor consistencia en la toma de decisiones entre hospitales y especialistas. Mejorar la discriminación entre masas benignas y malignas podría evitar cirugías innecesarias en mujeres con tumores benignos, reduciendo morbilidad y costes asociados [23]. Asimismo, favorecer la identificación de tumores en estadios FIGO I–II mediante una mejor interpretación de la ecografía transvaginal contribuiría a desplazar una proporción de casos desde la enfermedad avanzada hacia la enfermedad localizada, con potencial para traducirse en ganancias sustanciales en supervivencia y calidad de vida [63]. Los modelos de IA presentan un potencial singular para homogeneizar el rendimiento diagnóstico entre centros con diferente nivel de experiencia.

En el plano tecnológico, el diseño de modelos híbridos CNN–ViT con fusión temprana adaptativa puede ser extrapolable a otros dominios de la imagen médica: ultrasonido de mama, hígado o tiroides; radiología torácica y abdominal; e incluso aplicaciones veterinarias. De este modo, los avances metodológicos generados en el contexto del CO podrían funcionar como una plataforma de innovación reutilizable en múltiples escenarios [41].

Desde la perspectiva científica, este proyecto contribuye en varios niveles: (i) ofrece una revisión crítica del estado del arte, identificando brechas metodológicas que explican la heterogeneidad de resultados reportados en la literatura; (ii) propone un nuevo enfoque algorítmico de fusión temprana adaptativa que intenta abordar de forma explícita la complementariedad entre detalle local (CNN) y contexto global (ViT) [34, 42].

En términos de impacto social, la relevancia de este proyecto se vincula estrechamente con la equidad y la democratización del acceso a diagnóstico especializado. El CO presenta importantes desigualdades entre países y dentro de los propios sistemas de salud: las mujeres que viven en zonas rurales, con menor nivel socioeconómico o que pertenecen a minorías étnicas suelen ser diagnosticadas más tarde y tienen peores resultados [74].

Si se diseña e implementa de forma responsable, la IA aplicada al ultrasonido puede convertirse en una herramienta democratizadora: modelos entrenados y validados adecuadamente podrían apoyar a ecografistas con menor experiencia en contextos de alta carga asistencial, reduciendo la dependencia de un pequeño número de especialistas altamente cualificados concentrados en grandes ciudades. No obstante, esto solo será posible si se abordan de manera explícita los riesgos de sesgo algorítmico y se garantiza la representación adecuada de poblaciones diversas en los conjuntos de entrenamiento [96].

En síntesis, la relevancia interdisciplinaria y el impacto social de la IA aplicada al ultrasonido ovárico no se limitan a la mejora de métricas diagnósticas, sino que abarcan la posibilidad de transformar circuitos de atención, reducir desigualdades en salud y generar conocimiento metodológico transferible a otros ámbitos de la medicina. La articulación entre clínica, ingeniería, ciencia de datos, bioética y salud pública es, por tanto, condición necesaria para que estos modelos pasen de ser prototipos de laboratorio a herramientas que mejoren de forma tangible la vida de las pacientes.

II – JUSTIFICACIÓN

II - JUSTIFICACIÓN

El CO constituye uno de los retos más significativos en la oncología ginecológica contemporánea debido a su elevada letalidad y a la persistente dificultad de lograr un diagnóstico precoz. A pesar de los avances en cirugía, terapias dirigidas e inmunoterapia, las tasas de supervivencia global se mantienen bajas, principalmente porque la mayoría de los casos son diagnosticados en estadios avanzados. En estadios FIGO III–IV, la supervivencia a cinco años rara vez supera el 30%, mientras que en estadios I–II puede acercarse al 90%, lo que subraya la trascendencia del diagnóstico temprano en términos de pronóstico y años de vida ganados. Este escenario clínico ha generado la necesidad urgente de explorar metodologías diagnósticas innovadoras que complementen y potencien las técnicas actuales de imagen, con el objetivo de identificar de forma más temprana y precisa la malignidad de los tumores ováricos. En esta línea, tanto los análisis de carga global de enfermedad como los informes del *World Ovarian Cancer Coalition Atlas* coinciden en señalar que la ausencia de herramientas eficaces de detección precoz es uno de los principales cuellos de botella para mejorar la supervivencia y reducir la carga socioeconómica asociada al CO [108].

La UST, en particular en su modalidad B, constituye la herramienta de primera línea para la evaluación inicial de masas anexiales. Su accesibilidad, seguridad y bajo coste la hacen insustituible en la práctica clínica. Sin embargo, la interpretación de estas imágenes es altamente dependiente de la experiencia del observador y está sujeta a variabilidad inter e intraobservador. Además, la especificidad es limitada, lo que con frecuencia genera incertidumbre diagnóstica y un número considerable de intervenciones invasivas innecesarias. Estas limitaciones han abierto la puerta a la aplicación de la IA, y en particular del DL, como estrategia de apoyo a la práctica clínica. Diversos estudios multicéntricos y metaanálisis han mostrado que, aun con marcos estandarizados como IOTA y O-RADS, una proporción relevante de lesiones benignas sigue derivándose a cirugía, lo que refuerza la necesidad de herramientas complementarias que aumenten la especificidad sin comprometer la sensibilidad [23, 86].

Dentro del amplio espectro de técnicas de IA, las CNNs han demostrado ser particularmente eficaces para extraer características locales de imágenes médicas, mostrando desempeños notables en dominios como la mamografía, la tomografía de tórax o la resonancia magnética cerebral. No obstante, estas arquitecturas tienden a infraestimar la información contextual global, que en ecografía resulta fundamental para la caracterización de patrones difusos o distribuciones espaciales complejas. Por su parte, los modelos tipo ViT sobresalen en el modelado de relaciones contextuales a gran escala, pero con frecuencia descuidan los detalles finos y locales imprescindibles para una clasificación anatómica precisa. Los ViTs pueden alcanzar rendimientos equiparables o superiores a las CNNs en clasificación de imágenes cuando se dispone de grandes volúmenes de datos, lo que ha motivado su rápida adopción en aplicaciones médicas [40].

Este desajuste metodológico ha propiciado la necesidad de integrar ambos enfoques. La investigación más reciente apunta a que los modelos híbridos CNN–ViT pueden superar las limitaciones de cada arquitectura por separado, siempre que la integración de características se realice en etapas tempranas, permitiendo una interacción eficaz entre los descriptores locales y globales. Sin embargo, la mayoría de los enfoques híbridos desarrollados hasta la fecha han implementado estrategias de fusión tardía, en las que la combinación de características se realiza en fases finales del procesamiento, limitando así el potencial sinérgico de ambas arquitecturas. En el ámbito concreto del ultrasonido ovárico, han empezado a describirse arquitecturas locales–globales para la clasificación de tumores, pero la mayoría se centra en tareas binarias y sigue empleando fusiones tardías o módulos de atención diseñados *ad hoc*, con validaciones metodológicamente heterogéneas [109, 110].

La propuesta de este proyecto doctoral introduce una innovación metodológica decisiva: una estrategia de fusión temprana adaptativa que posibilita que las características extraídas por las CNNs y los ViTs interactúen desde las primeras etapas de procesamiento. Con ello, se pretende optimizar el reconocimiento tanto de microestructuras locales como de patrones de contexto global presentes en las imágenes de UST de ovario. Este planteamiento responde directamente a la brecha existente entre el potencial técnico de la IA y su aplicación rutinaria en la práctica clínica de la ginecología oncológica.

La justificación de este trabajo se articula en distintos niveles: clínico, científico, tecnológico, social y formativo.

2.1. JUSTIFICACIÓN CLÍNICA

Desde una perspectiva clínica, la detección temprana del CO es determinante para mejorar el pronóstico de las pacientes. En este sentido, contar con herramientas de apoyo que incrementen la precisión diagnóstica durante la evaluación de masas anexiales constituye una necesidad crítica. Las guías clínicas y revisiones recientes insisten en que el mayor margen de mejora en CO no reside únicamente en la optimización de las terapias avanzadas, sino en adelantar el diagnóstico a estadios I–II, donde la cirugía citorreductora y los esquemas estándar de platino logran mejores resultados y menores secuelas [1, 111].

La ecografía es insustituible como método inicial de cribado, pero sus limitaciones exigen complementarla con tecnologías que aporten objetividad y reproducibilidad. El desarrollo de algoritmos de IA capaces de discriminar con alta fiabilidad entre tumores benignos y malignos se traduce en beneficios directos para las pacientes: reducción de cirugías innecesarias, menor ansiedad asociada a la incertidumbre diagnóstica y acceso más temprano a tratamientos adecuados. Además, la incorporación de métricas de calibración y análisis de incertidumbre permite a los clínicos contar con información más robusta para la toma de decisiones, reservando la revisión experta para aquellos casos en los que el modelo identifica mayor ambigüedad. La literatura sobre calibración de redes neuronales ha demostrado que modelos bien calibrados pueden transformar probabilidades en riesgos clínicamente interpretables, lo que resulta crucial cuando las salidas del modelo se usan para decidir cirugías o derivaciones a centros especializados [108].

Por tanto, este proyecto se justifica como un aporte a la práctica clínica que busca salvar vidas mediante la detección temprana, mejorar la calidad de vida de las pacientes y optimizar los recursos sanitarios. En particular, la combinación de UST con modelos híbridos calibrados y acompañados de mapas de atención interpretables apunta a reducir la variabilidad interobservador y a ofrecer una segunda opinión objetiva en contextos con menor experiencia ecográfica.

2.2. JUSTIFICACIÓN CIENTÍFICA

El CO se mantiene como un terreno fértil para la investigación biomédica, ya que persisten lagunas significativas en el diagnóstico temprano. Este trabajo doctoral se sitúa en la intersección de dos campos altamente dinámicos: la oncología ginecológica y la IA aplicada a imágenes médicas. La originalidad científica de este proyecto radica en varios aspectos:

- **Fusión temprana adaptativa**: hasta la fecha, la mayoría de las arquitecturas híbridas CNN–ViT han implementado fusiones tardías. La propuesta de fusionar características en etapas iniciales representa una contribución novedosa que permite explotar sinergias entre lo local y lo global desde el inicio del flujo de datos.
- **Caracterización multicategoría**: mientras muchos estudios previos se centran en clasificaciones binarias (benigno vs. maligno), este proyecto aborda la clasificación de múltiples categorías clínicas relevantes, reflejando mejor la diversidad real de los tumores ováricos.
- **Evaluación exhaustiva de incertidumbre y calibración**: más allá de la simple precisión, el proyecto incorpora análisis de confiabilidad y curvas de decisión clínica (CDC), acercando los modelos a escenarios reales de práctica médica.
- **Enfoque en datos reales y desbalanceados**: el uso de un conjunto de datos representativo, con clases minoritarias clínicamente relevantes, añade complejidad y realismo al estudio, aumentando la validez de los hallazgos y su aplicabilidad a contextos clínicos heterogéneos.

Este proyecto no solo busca alcanzar métricas de precisión sobresalientes, sino también aportar conocimiento metodológico sobre cómo diseñar y evaluar algoritmos de IA que puedan traducirse en una implementación clínica real.

2.3. JUSTIFICACIÓN TECNOLÓGICA

El desarrollo de modelos híbridos CNN–ViT con fusión temprana requiere de innovación tecnológica tanto en el diseño e implementación de las arquitecturas. El proyecto plantea una serie de retos tecnológicos que justifican su pertinencia:

- **Generalización y transferibilidad**: se busca diseñar modelos que no solo funcionen en un conjunto de datos local, sino que puedan ser validados y aplicados en contextos multicéntricos, respondiendo a la necesidad de herramientas generalizables.
- **Interpretabilidad**: la integración de técnicas como Grad-CAM responde a la necesidad tecnológica de generar explicaciones visuales que permitan a los médicos comprender y confiar en las decisiones de la IA.

El proyecto se justifica tecnológicamente porque contribuye al desarrollo de una nueva generación de sistemas inteligentes más precisos, confiables e interpretables, alineados con la tendencia global hacia una medicina asistida por IA. Al mismo tiempo, sirve como banco de pruebas para incorporar de forma sistemática análisis de calibración, incertidumbre y explicabilidad en flujos de trabajo de ultrasonido, elementos todavía ausentes en la mayoría de las implementaciones clínicas.

2.4. JUSTIFICACIÓN SOCIAL

La investigación en salud debe trascender el ámbito académico y científico para generar impacto social. El CO no solo afecta la salud física de las mujeres, sino también su bienestar emocional, su vida familiar y su productividad social y laboral. Los estudios de carga global y los informes internacionales estiman que el CO genera millones de DALYs anuales y que su impacto económico es especialmente gravoso en países de ingresos bajos y medios, donde las pacientes afrontan gastos catastróficos y acceso limitado a terapias avanzadas [63, 70].

La aplicación de IA para mejorar la detección del cáncer de ovario puede tener repercusiones sociales directas:

- **Equidad en salud**: el desarrollo de algoritmos que puedan ser aplicados en contextos con escasos recursos humanos especializados democratiza el acceso a diagnósticos de calidad.
- **Reducción de costes sanitarios**: al disminuir intervenciones innecesarias y optimizar recursos, se genera un impacto positivo en la sostenibilidad de los sistemas de salud.

- **Mejora de la calidad de vida**: diagnósticos más tempranos y precisos ofrecen a las pacientes mejores oportunidades terapéuticas, reduciendo la carga emocional y económica sobre las familias.

De este modo, la justificación social del proyecto se fundamenta en su potencial para contribuir a una atención médica más justa, accesible y efectiva.

2.5. JUSTIFICACIÓN ACADÉMICA Y FORMATIVA

Este trabajo doctoral también se justifica como una oportunidad de formación de alto nivel, al integrar investigación aplicada, innovación tecnológica y colaboración internacional. El plan de trabajo contempla estancias de investigación en centros de referencia y la interacción con instituciones especializadas en oncología ginecológica, lo que enriquecerá el perfil del investigador y favorecerá la transferencia de conocimientos. Además, se enmarca en un contexto internacional en el que se reclama de forma creciente la incorporación de competencias en IA y ciencia de datos en la formación de especialistas clínicos, con el fin de garantizar un uso crítico y responsable de estas herramientas [33, 100].

Asimismo, los resultados generados, tanto en forma de publicaciones científicas como de algoritmos y prototipos, aportarán a la comunidad académica nuevos modelos y metodologías que podrán ser replicados, adaptados y mejorados por futuros investigadores. Los artículos ya publicados sobre la revisión sistemática y el modelo híbrido CNN–ViT constituyen los primeros productos científicos de este itinerario y sirven como punto de partida para consolidar una línea de investigación sostenida en IA aplicada a ecografía ginecológica [34, 41].

III – OBJETIVOS

III - OBJETIVOS

El diseño de los objetivos de esta investigación doctoral responde a la necesidad de traducir en metas operativas los retos clínicos, científicos, tecnológicos y sociales expuestos en la Introducción y la Justificación. En concreto, se busca desarrollar y evaluar modelos de IA aplicados al ultrasonido transvaginal (UST) que contribuyan a mejorar la detección y caracterización de tumores ováricos en escenarios clínicos reales.

3.1. OBJETIVO GENERAL

Esta tesis busca desarrollar, implementar y validar modelos híbridos CNN–ViT con fusión temprana adaptativa para clasificar tumores ováricos en imágenes UST aprovechando representaciones locales y globales para mejorar frente a modelos de una sola rama el rendimiento multiclase, la calibración probabilística, la interpretabilidad y la utilidad clínica.

Para ello, (i) se sintetiza cuantitativamente el estado del arte mediante una revisión sistemática, (ii) se realiza una depuración exhaustiva del conjunto de datos para asegurar reproducibilidad y representatividad clínica, (iii) se entrena y compara modelos de referencia (solo CNNs y solo ViTs) frente a (iv) arquitecturas híbridas CNN–ViT, (v) se demuestra la ventaja del enfoque propuesto mediante evaluación robusta y pruebas estadísticas, e (vi) incorpora métodos de explicabilidad para sustentar la transparencia del modelo y su potencial adopción como sistema de apoyo a la decisión clínica.

3.2. OBJETIVOS ESPECÍFICOS

- OE1. Realizar una revisión sistemática de modelos de IA aplicados al diagnóstico y caracterización de tumores ováricos mediante UST, cuantificando desempeño e identificando heterogeneidad, sesgos y brechas metodológicas relevantes para su traslación clínica.

- OE2. Depurar de forma exhaustiva la base de datos empleada, incluyendo control de calidad, armonización de clases y particiones reproducibles para entrenamiento, validación y test.
- OE3. Diseñar, entrenar y evaluar modelos de referencia basados exclusivamente en CNNs y exclusivamente en ViTs para establecer una base robusta y comparable en clasificación multiclase sobre UST en modo B.
- OE4. Proponer, implementar y optimizar modelos híbridos CNN–ViT con fusión temprana adaptativa, integrando características locales y globales, y cuantificar el efecto de la fusión sobre el desempeño multiclase.
- OE5. Validar comparativamente el enfoque híbrido frente a las líneas base mediante evaluación robusta (métricas globales y por clase), pruebas estadísticas, análisis de errores y calibración probabilística. Adicionalmente, explorar estrategias de combinación cuando aporten mejoras consistentes.
- OE6. Incorporar y evaluar métodos de interpretabilidad/explicabilidad para generar mapas de activación y analizar su correspondencia con regiones clínicamente relevantes definidas por expertos, fortaleciendo la transparencia del sistema y su utilidad clínica como apoyo a la decisión de los radiólogos.

IV - MATERIAL Y MÉTODO

IV -MATERIAL Y MÉTODO

4.1. DISEÑO GENERAL DE LA INVESTIGACIÓN DOCTORAL

La presente tesis doctoral se estructura metodológicamente en dos componentes complementarios (Anexo I):

1. **Un estudio de revisión sistemática** sobre el rendimiento diagnóstico, la calidad metodológica y la traslación clínica de los modelos de IA aplicados al UST en modo B en el CO. Este estudio sigue las guías PRISMA 2020 (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) [112] y emplea la herramienta PROBAST (*Prediction model Risk Of Bias ASsessment Tool*) [113] para la evaluación del riesgo de sesgo y la aplicabilidad de los modelos.
2. **Un estudio experimental de desarrollo y evaluación de modelos híbridos CNN–ViT con fusión temprana** para la clasificación multicategoría de tumores ováricos a partir de la base de datos pública OTU-2D, subconjunto del conjunto de imágenes MMOTU (*Multi-Modality Ovarian Tumor Ultrasound*) [114].

4.1.1. Articulación metodológica entre Estudio 1 y Estudio 2

El Estudio 1 tiene un objetivo de síntesis y evaluación crítica: integra la evidencia disponible sobre IA aplicada a UST en modo B, cuantifica el rendimiento diagnóstico reportado y analiza la calidad metodológica. Este enfoque permite identificar patrones globales y factores asociados al rendimiento, pero, por su propia naturaleza, no puede imponer condiciones experimentales homogéneas (misma partición, mismo preprocesado, mismos criterios de evaluación) ni comparar arquitecturas bajo un protocolo único. Además, la revisión pone de manifiesto carencias recurrentes en la literatura, como la escasa estandarización del *pipeline*, el reporte incompleto de métricas (intervalos de confianza, calibración y utilidad clínica) y la falta de análisis explícitos de incertidumbre e interpretabilidad, elementos necesarios para una traslación clínica responsable.

Con base en esas limitaciones, el Estudio 2 se planteó como un estudio experimental, diseñado para cubrir lo que el Estudio 1 no puede resolver: (i) evaluar de manera comparativa modelos CNNs, modelos basados en ViTs y modelos híbridos CNN–ViT bajo un mismo protocolo de entrenamiento; (ii) explorar explícitamente una estrategia de fusión temprana en un escenario multicategoría clínicamente relevante; y (iii) incorporar métricas orientadas a implementación, además de interpretabilidad visual (Grad-CAM). De este modo, el Estudio 2 actúa como validación empírica y extensión traslacional de los hallazgos del Estudio 1, proporcionando un hilo conductor metodológico claro desde la evidencia agregada hacia un *pipeline* experimental verificable.

4.2. ESTUDIO 1: REVISIÓN SISTEMÁTICA SOBRE IA APLICADA A ECOGRAFÍA OVÁRICA

4.2.1. Tipo de estudio y guías de referencia

Se llevó a cabo una revisión sistemática de estudios diagnósticos que evaluaran modelos de IA aplicados a la clasificación de masas ováricas utilizando imágenes de UST en modo B. El estudio se diseñó y condujo conforme a las guías PRISMA 2020, lo que incluye una estrategia de búsqueda predefinida, criterios explícitos de elegibilidad, evaluación de riesgo de sesgo y un plan de análisis estadístico detallado [112].

4.2.2. Fuentes de información y estrategia de búsqueda

En enero de 2025 se realizó una búsqueda sistemática en tres bases de datos de alto impacto: PubMed, IEEE Xplore y Scopus. La estrategia de búsqueda combinó términos relacionados con IA y CO mediante operadores booleanos:

(“machine learning” OR “artificial intelligence” OR “deep learning” OR “neural network”) AND (“ovarian cancer” OR “ovarian tumor”) AND “ultrasound”.

No se aplicaron restricciones de idioma ni de tipo de publicación en la búsqueda inicial, con el objetivo de maximizar la sensibilidad de la estrategia.

4.2.3. Criterios de inclusión y exclusión

Se definieron de antemano criterios de elegibilidad para seleccionar estudios comparables y metodológicamente sólidos:

4.2.3.1. ***Criterios de inclusión***

- Estudios originales que desarrollaran o validaran modelos de IA (incluyendo aprendizaje automático (ML), redes neuronales artificiales (ANNs), CNNs y ViTs) para clasificación de masas ováricas utilizando UST en modo B.
- Uso de un patrón de referencia clínicamente aceptado, preferentemente confirmación histopatológica o seguimiento clínico adecuado.
- Reporte de al menos una métrica de rendimiento: exactitud, sensibilidad, especificidad y AUC.
- Estudios en humanos, independientemente del diseño (retrospectivo, prospectivo, unicéntrico o multicéntrico).

4.2.3.2. ***Criterios de exclusión***

- Estudios que empleasen exclusivamente otras modalidades de imagen (p. ej., TC, RM) sin UST.
- Trabajos centrados solo en segmentación sin evaluación explícita de rendimiento diagnóstico en clasificación.
- Estudios sin patrón de referencia adecuado (p. ej., sin confirmación histopatológica para malignidad).
- Revisiones, editoriales, cartas, resúmenes de congresos sin datos completos o sin acceso al texto íntegro.

4.2.4. Proceso de selección de estudios

La selección de estudios se realizó en dos fases:

- Cribado de títulos y resúmenes: dos revisores independientes evaluaron todos los registros identificados, aplicando los criterios de inclusión/exclusión.

- Revisión de texto completo: los artículos preseleccionados fueron revisados a texto completo por los mismos revisores para confirmar su elegibilidad definitiva.

Las discrepancias se resolvieron mediante consenso y, cuando fue necesario, con la intervención de un tercer revisor. El proceso se documentó mediante un diagrama de flujo PRISMA.

4.2.5. Extracción de datos

De cada estudio incluido se extrajeron de manera sistemática las siguientes variables: autor y año de publicación, tipo de segmentación (manual, automática), arquitectura de IA (CNN, ML, ANN, ViT, híbridos), nombre del modelo o algoritmo, tamaño del conjunto de imágenes, tipo de masas ováricas y número de clases y métricas de rendimiento. Los datos se volcaron en una base estructurada para su análisis estadístico posterior. En estudios con múltiples modelos, se seleccionó el de mejor rendimiento global.

4.2.6. Evaluación del riesgo de sesgo y aplicabilidad

La calidad metodológica y el riesgo de sesgo se evaluaron utilizando la herramienta PROBAST, que consta de 20 ítems agrupados en cuatro dominios: selección de participantes, predictores, resultados y análisis estadístico [113]. Se generaron clasificaciones de riesgo de sesgo (bajo, alto o incierto) por dominio y global, así como tablas detalladas de evaluación disponibles en material suplementario del artículo [34].

4.2.7. Análisis estadístico

El análisis estadístico se realizó en varias etapas:

- **Estadística descriptiva** de las métricas de rendimiento (media, desviación estándar (DE), rango). La normalidad se evaluó con la prueba de Shapiro–Wilk y la homogeneidad de varianzas con la prueba de Levene.
- **Comparación de estrategias de segmentación** (automática vs. manual) mediante la prueba de Mann–Whitney U.

- **Comparación de arquitecturas de IA** mediante pruebas no paramétricas (Kruskal–Wallis) y ANOVA cuando procedía, diferenciando entre CNN, ML, ANN y otros enfoques.
- **Correlación entre tamaño muestral y rendimiento diagnóstico** mediante coeficiente de Pearson, excluyendo estudios con > 5 000 imágenes para evitar la influencia de valores extremos.
- **Metarregresión** con la exactitud como variable dependiente e incluyendo como predictores el tamaño de la base de datos, el tipo de segmentación, el tipo de arquitectura y el riesgo de sesgo.

Adicionalmente, se realizaron análisis de subgrupos según riesgo de sesgo para evaluar la robustez de los resultados. Todos los análisis se llevaron a cabo en Python 3.12.

4.3. Estudio 2: Modelos híbridos CNN–ViT con fusión temprana en OTU-2D

4.3.1. Tipo de estudio

Se realizó un estudio de desarrollo y validación de modelos de predicción diagnóstica, basado en imágenes de UST de tumores ováricos y orientado a la clasificación multicategoría (ocho clases histopatológicas).

Este estudio experimental se concibió como una continuación directa del Estudio 1. Mientras que la revisión sistemática permite sintetizar resultados publicados y detectar patrones generales, no puede (por diseño) imponer un protocolo único de preprocesamiento y entrenamiento para comparar arquitecturas en condiciones idénticas. Además, el Estudio 1 evidenció vacíos recurrentes en la literatura: heterogeneidad en el manejo del desbalance y la segmentación, escasa evaluación de calibración y de utilidad clínica, y poca exploración de modelos híbridos CNN–ViT orientados a clasificación multicategoría. Por ello, el Estudio 2 se diseñó para probar, bajo un *pipeline* uniforme, una estrategia de fusión temprana y contrastarla frente a CNNs y ViTs, incorporando métricas con intervalos de confianza, calibración, análisis de decisión y cuantificación de incertidumbre, con el fin de avanzar desde la evidencia agregada hacia un marco experimental verificable.

4.3.2. Fuente de datos: base OTU-2D

Se utilizó la base de datos pública OTU-2D, subconjunto del conjunto de imágenes MMOTU. OTU-2D contiene 1 469 imágenes ecográficas bidimensionales en modo B, adquiridas bajo condiciones clínicas estandarizadas en el hospital Beijing Shijitan (China). Las imágenes de OTU-2D se proporcionan ya segmentadas y recortadas alrededor de la lesión. Todas las lesiones contaban con confirmación histopatológica por anatomopatólogos expertos [114]. Las lesiones fueron clasificadas por oncólogos ginecológicos en ocho categorías clínicamente relevantes (Figura 2): quiste de chocolate (endometrioma), cistoadenoma seroso, cistoadenoma mucinoso, teratoma, quiste simple, tumor de células de la teca, carcinoma seroso de alto grado y ovario normal.

En esta tesis doctoral todos los experimentos, métricas y análisis se realizaron exclusivamente sobre OTU-2D. Aunque el proyecto cuenta con un protocolo aprobado con el Instituto Estatal de Cancerología "Dr. Arturo Beltrán Ortega" (Anexo II), dicha cohorte local no se empleó en los análisis aquí presentados y queda reservada para futuras validaciones externas.

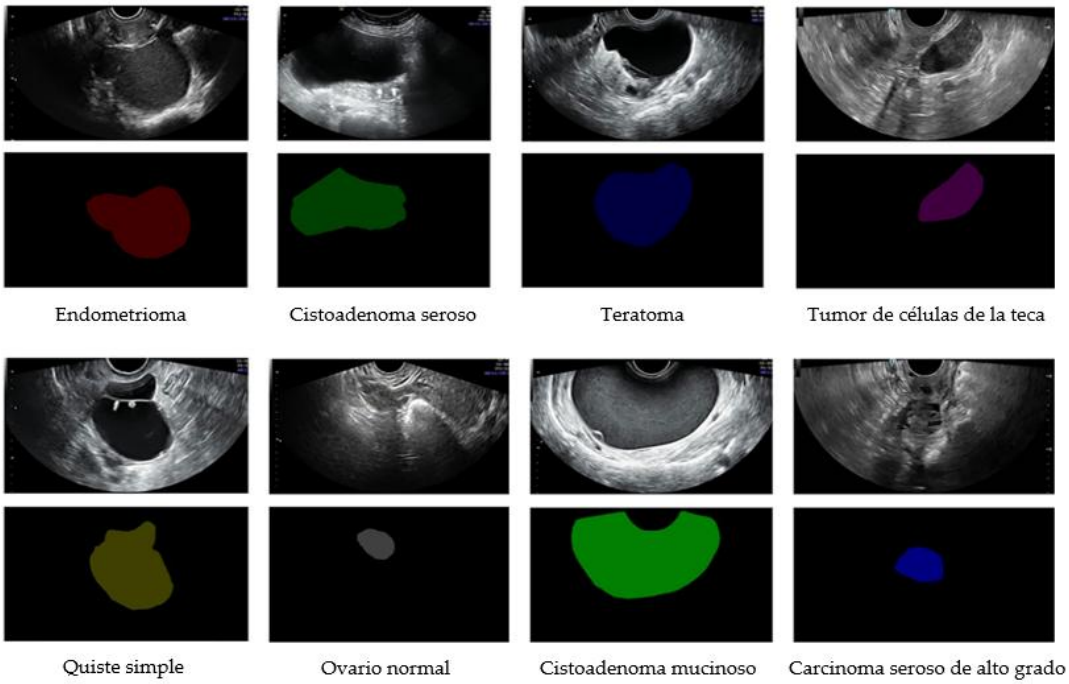

Figura 2. Ejemplos representativos de las ocho categorías clínicas incluidas en el conjunto de datos OTU-2D (imágenes UST y anotaciones). Fuente: Elaboración propia.

4.3.3. Aspectos éticos, privacidad y seguridad

OTU-2D es un recurso público con imágenes y metadatos totalmente anonimizados. Al tratarse de un análisis secundario de datos anonimizados, no fue necesaria una nueva aprobación por parte de un comité de ética ni la obtención de nuevos consentimientos informados.

En paralelo, el protocolo aprobado con el Instituto Estatal de Cancerología "Dr. Arturo Beltrán Ortega" garantiza que, en futuras fases de validación clínica externa con datos locales, la recolección y uso de imágenes se realizará bajo estándares éticos y regulatorios nacionales de México.

La Figura 3 resume el flujo de trabajo completo del Estudio 2, desde la preparación del conjunto OTU-2D hasta la evaluación y análisis de incertidumbre del modelo.

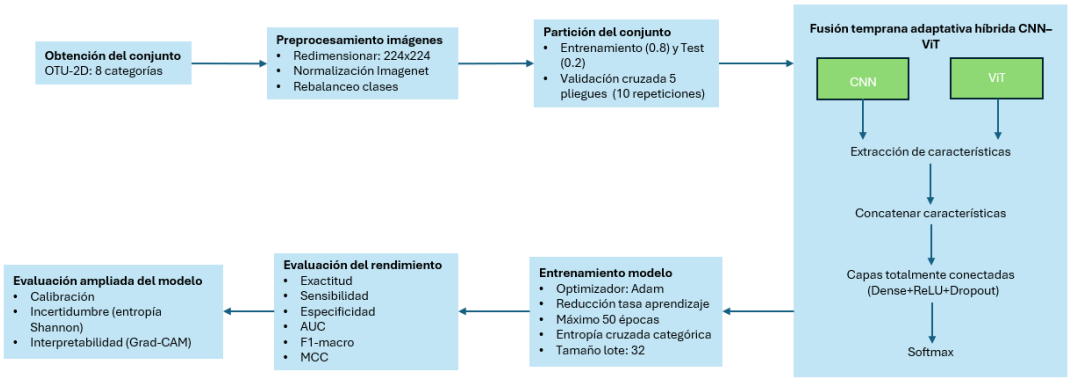

Figura 3. Esquema general del Estudio 2: adquisición y preprocesamiento del conjunto OTU-2D, entrenamiento de modelos híbridos CNN–ViT con fusión temprana, y evaluación mediante métricas de rendimiento. Fuente: Elaboración propia.

4.3.4. Preprocesamiento de imágenes

Todas las imágenes ecográficas se sometieron a un preprocesamiento homogéneo. En primer lugar, se redimensionaron a 224 × 224 píxeles para adecuarse a los requisitos de entrada de las arquitecturas preentrenadas. Posteriormente, se convirtieron a tensores numéricos y se normalizaron utilizando las medias y desviaciones estándar por canal RGB de ImageNet (0.485, 0.456, 0.406 y 0.229, 0.224, 0.225), lo que facilitó el aprendizaje por transferencia. En el conjunto de entrenamiento se aplicaron técnicas de aumento de datos para incrementar la

variabilidad de las imágenes y mejorar la capacidad de generalización del modelo. Estas transformaciones incluyeron rotaciones y traslaciones leves, recortes aleatorios y ajustes de brillo y contraste. Adicionalmente, para reducir el efecto del desbalance entre categorías diagnósticas, se realizó un rebalanceo de clases mediante sobremuestreo aleatorio (*random oversampling*), aplicado exclusivamente al conjunto de entrenamiento, de modo que las clases minoritarias estuvieran más representadas durante el aprendizaje sin alterar los conjuntos de validación y test. La decisión de emplear sobremuestreo aleatorio en lugar de funciones de pérdida específicas (como *focal loss* o *class-balanced loss*) se justificó por su mayor interpretabilidad, su estabilidad durante el entrenamiento y la menor complejidad asociada al ajuste de hiperparámetros.

4.3.5. Arquitecturas de deep learning

Se evaluaron cinco arquitecturas preentrenadas en ImageNet, agrupadas en tres CNN y dos Transformers tipo ViT:

- ResNet152 [115]: redes residuales profundas.
- DenseNet201 [116]: conectividad densa entre capas.
- EfficientNetB7 [117]: escalado compuesto eficiente.
- ViTB16 [40]: ViT basado en parches.
- Swin Transformer [118]: atención jerárquica mediante ventanas.

En las CNN, el último mapa de activación se redujo mediante *adaptive global average pooling* y posteriormente se aplicó *flatten* para obtener un vector 1-D. En los ViT, se utilizó como representación global el *embedding* previo a la capa clasificadora. Todas las arquitecturas se inicializaron con pesos de ImageNet y se ajustaron mediante *fine-tuning* para la tarea de ocho clases.

Para los modelos de rama única (solo CNN o solo ViT) se empleó una cabeza de clasificación homogénea: proyección lineal a 8 *logits*. Los parámetros de entrenamiento se mantuvieron constantes entre modelos para garantizar comparabilidad (Tabla 1).

4.3.6. Bloque de fusión temprana

En los modelos híbridos CNN–ViT se implementó una estrategia de fusión temprana aprendida (*learned early-fusion via joint projection*). Las representaciones locales de la CNN y las globales del ViT se concatenaron en un vector conjunto y se integraron mediante una proyección totalmente conectada con activación ReLU, seguida de una capa final lineal hacia 8 *logits* (y softmax para obtener probabilidades).

La estructura del bloque de fusión se mantuvo fija en todas las combinaciones híbridas; únicamente cambian los *backbones* (y, por tanto, la dimensionalidad de entrada al bloque). La regularización y el esquema de entrenamiento aplicados a los híbridos fueron los mismos que en los modelos exclusivos de CNNs y ViTs (Tabla 1).

Tabla 1. Hiperparámetros y configuración de entrenamiento para todos los modelos.

Categoría	***Parámetro***	***Valor***
Evaluación	*Validación cruzada*	*5*
	Repeticiones	*10*
Entrenamiento	*Tamaño de lote (batch size)*	*32*
	Épocas máximas	*50*
Optimizador	*Tipo*	*Adam*
	Tasa de aprendizaje (learning rate) inicial	1×10^{-4}
Regularización	*L2 / weight decay*	1×10^{-5}
	Dropout	*0.3*
Scheduler	*Tipo*	*ReduceLROnPlateau*
	Reducción (factor)	*0.1*
	Patience	*5*
Parada temprana	*Criterio*	*Pérdida de validación*
	Patience	*5*
Pérdida	*Función de pérdida*	*Entropía cruzada categórica*

4.3.7. Métricas de evaluación y análisis estadístico

Las métricas de desempeño incluyeron exactitud, sensibilidad, especificidad, AUC, F1-macro y coeficiente de correlación de Matthews (MCC). Para comparar los modelos individuales con los modelos híbridos, primero se evaluó la normalidad de las diferencias mediante la prueba de Shapiro–Wilk. Cuando se cumplió el supuesto de normalidad, se emplearon pruebas t de Student pareadas; en caso contrario, se utilizó la prueba de rangos con signo de Wilcoxon.

Dado el carácter multicategoría y el desbalance entre clases, las métricas se computaron tanto de manera global como con agregación macro cuando fue pertinente. La sensibilidad y la especificidad se estimaron a partir de la matriz de confusión por clase y posteriormente se promediaron para evitar que las clases mayoritarias dominen el resultado. Para el AUC se empleó un enfoque *one-vs-rest* y se reportó el promedio macro. Los intervalos de confianza del 95% se obtuvieron mediante *bootstrap* (n = 500) sobre las predicciones en pliegues. En la calibración se aplicó regresión isotónica sobre probabilidades y se evaluó con curvas de calibración. La utilidad clínica se exploró mediante el análisis de CDC en umbrales de riesgo del 5% al 20%. Finalmente, la incertidumbre predictiva se cuantificó mediante la entropía de Shannon de la distribución de probabilidades de clase predicha por el modelo. Esta medida se analizó en relación con la tasa de error diagnóstica y se utilizó para proponer reglas de derivación a revisión experta en los casos de mayor incertidumbre.

4.3.8. Interpretabilidad (Grad-CAM)

Para evaluar la interpretabilidad de los modelos híbridos se utilizó la técnica Grad-CAM, implementada con la librería torchcam. Antes de la capa de clasificación final se generaron mapas de activación, que posteriormente se superpusieron sobre las imágenes originales mediante las librerías OpenCV y PIL, usando un umbral del 50% de la activación máxima para resaltar las regiones más relevantes. La valoración cualitativa se realizó por comparación directa con los ROI señalados por los especialistas en la base de datos OTU-2D, confirmó que los modelos se focalizaban en zonas compatibles con la localización tumoral, lo que refuerza la credibilidad clínica del enfoque híbrido.

4.3.9. Entorno de software y hardware

Los modelos se implementaron en Python 3.12 utilizando PyTorch 2.7.0 con CUDA 11.8. La gestión de datos y análisis adicionales se realizaron con Pandas 2.1.3, NumPy 1.26.4 y Scikit-learn 1.3.2.

Todos los entrenamientos se ejecutaron en un servidor equipado con procesador Intel Xeon Silver 4216 @ 2.10 GHz y una GPU NVIDIA A100 de 40 GB de VRAM.

V – RESULTADOS

V - RESULTADOS

5.1. RESULTADOS DEL ESTUDIO 1: REVISIÓN SISTEMÁTICA

5.1.1. Selección de estudios y características generales

La búsqueda en PubMed, IEEE Xplore y Scopus identificó 823 registros potencialmente relevantes. Tras la eliminación de 58 duplicados, se sometieron a cribado 765 títulos y resúmenes. De ellos, 686 se excluyeron por no cumplir los criterios de inclusión, quedando 79 artículos para revisión a texto completo. Finalmente, 44 estudios cumplieron todos los criterios de elegibilidad y se incorporaron al análisis cuantitativo. Este proceso se representa en el diagrama de flujo PRISMA (Figura 4).

Los estudios incluidos abarcan publicaciones hasta diciembre de 2024 y, en conjunto, analizan más de 650 000 imágenes de UST en modo B utilizadas para la clasificación de masas ováricas. La mayoría de los trabajos, 27 de 44 (61.4%), emplearon segmentación automática, mientras que 17 (38.6%) recurrieron a segmentación manual. Las arquitecturas predominantes fueron las CNN, seguidas de algoritmos clásicos de ML, ANN y, de forma más reciente, modelos basados en ViTs.

5.1.2. Rendimiento global de los modelos de IA

El análisis de los 44 estudios puso de manifiesto un rendimiento diagnóstico globalmente elevado de los modelos de IA aplicados a imágenes de UST. La exactitud media fue del 92.3% con una DE de 5.8%. La sensibilidad media alcanzó el 91.6% ± 7.2 y la especificidad el 90.1% ± 8.1. Solo 23 estudios reportaron AUC, con una media de 0.93 ± 0.04, lo que refleja una capacidad discriminativa global fuerte, aunque el análisis está limitado por el reducido número de trabajos que informan esta métrica.

La comparación de los diez modelos mejor posicionados según exactitud se resume en la Figura 5, donde se representan simultáneamente exactitud, sensibilidad, especificidad y AUC. En dicha figura se observa que varios modelos

alcanzan exactitudes superiores al 95%. Sin embargo, algunos de ellos presentan un desbalance clínicamente relevante entre sensibilidad y especificidad, con modelos que privilegian la detección de casos positivos a costa de aumentar los falsos positivos, o viceversa. Además, muchos de estos mejores modelos se desarrollaron sobre conjuntos de datos pequeños, sin validación externa o con procedimientos de validación interna poco robustos, lo que aumenta la probabilidad de sobreajuste y limita la generalización de los resultados.

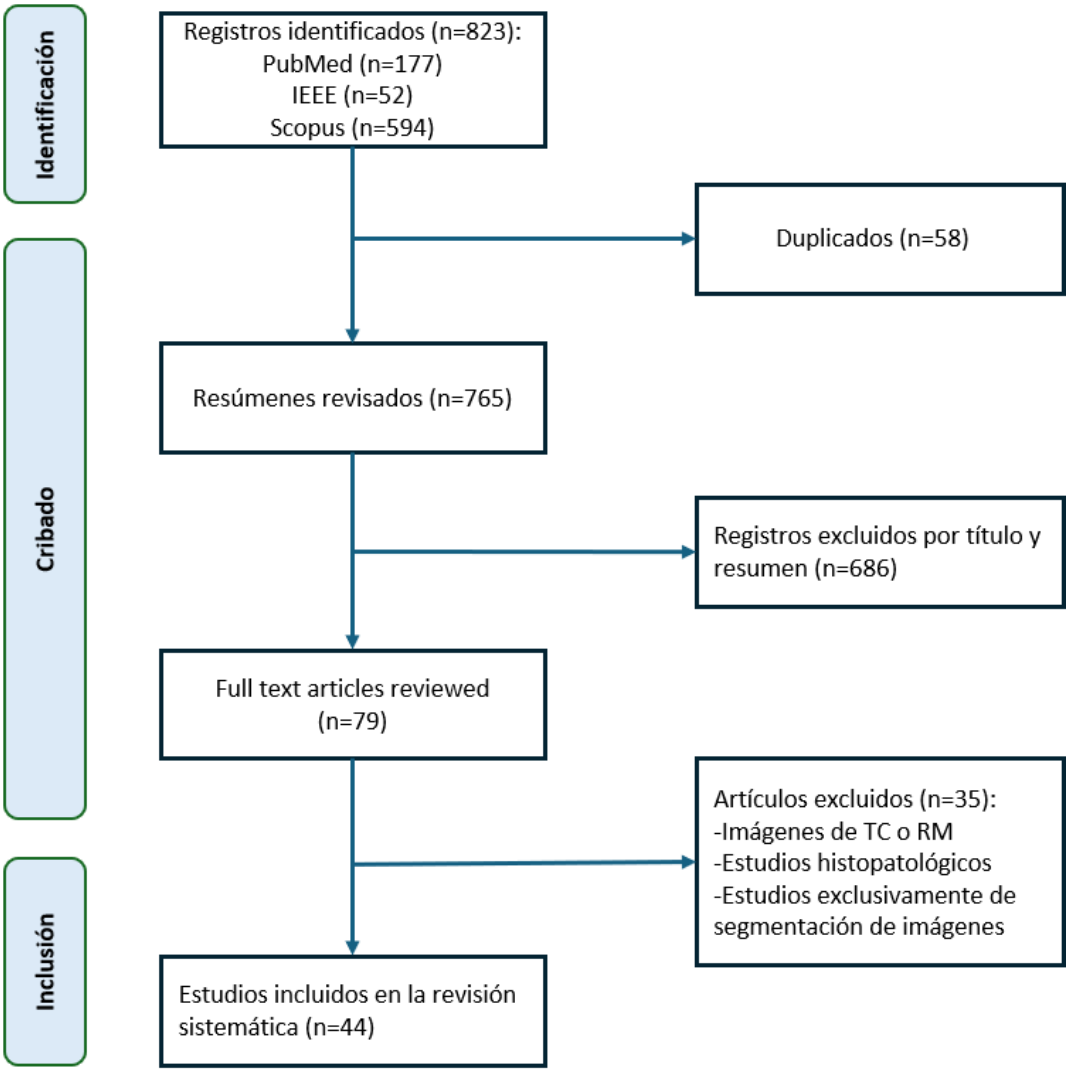

Figura 4. Diagrama de flujo PRISMA del proceso de identificación, cribado e inclusión de estudios; se incluyen 44 trabajos en el análisis cuantitativo. Fuente: Elaboración propia.

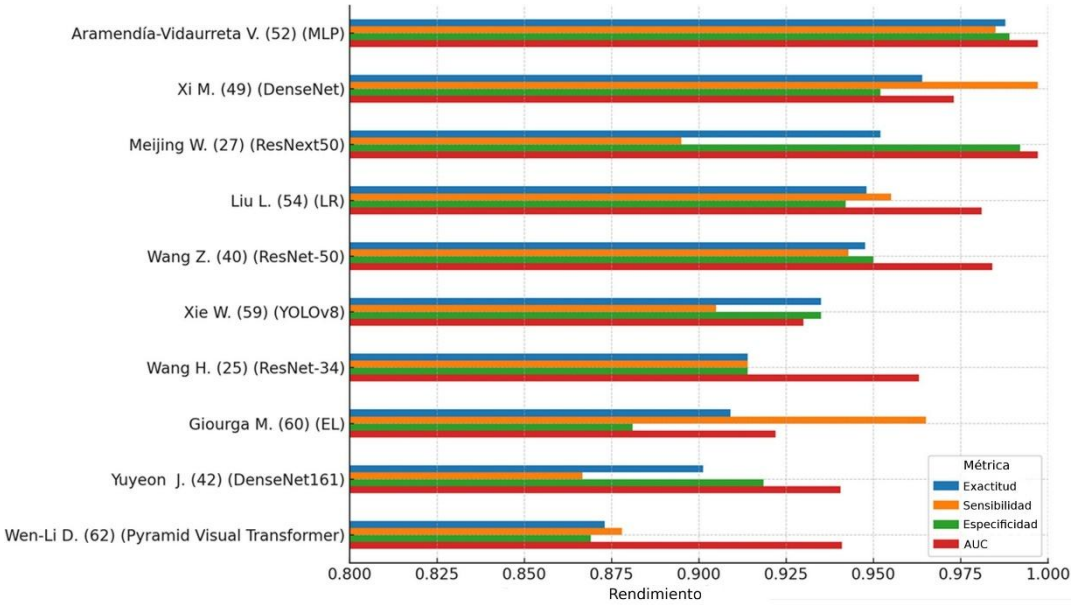

Figura 5. Comparación del desempeño diagnóstico reportado para 10 modelos de IA en ultrasonido transvaginal en modo B. Para cada modelo se muestran: exactitud (azul), sensibilidad (naranja), especificidad (verde) y AUC (rojo). Fuente: Elaboración propia.

5.1.3. Tamaño muestral y rendimiento de los modelos

La relación entre tamaño de la base de datos (limitada a estudios con ≤ 5 000 imágenes) y desempeño se evaluó mediante correlación de Spearman, tras comprobarse con Shapiro–Wilk que tanto el tamaño del conjunto de imágenes como las métricas de rendimiento seguían distribuciones no normales. Para la exactitud, el coeficiente de Spearman fue $\varrho = 0.080$ con $p = 0.653$ y un R^2 de 0.006, es decir, menos del 1% de la variación en exactitud pudo explicarse por el tamaño del conjunto de imágenes en el rango analizado. Para la sensibilidad, $\varrho = 0.246$ ($p = 0.154$; $R^2 = 0.061$) y, para la especificidad, $\varrho = 0.183$ ($p = 0.350$; $R^2 = 0.034$).

Estos resultados se representan en la Figura 6, que muestra tres diagramas de dispersión (exactitud, sensibilidad y especificidad frente al tamaño del conjunto de imágenes, ≤ 5 000), cada uno con un ajuste LOWESS y bandas de confianza del 95%. La figura evidencia una dispersión considerable en todas las métricas y la ausencia de una tendencia clara que sugiera un incremento sistemático del rendimiento con el aumento del tamaño muestral en el rango clínicamente habitual. En consecuencia, dentro de los límites de esta revisión, el tamaño del conjunto de

imágenes no emerge como factor determinante del desempeño, al menos cuando se sitúa por debajo de las decenas de miles de imágenes.

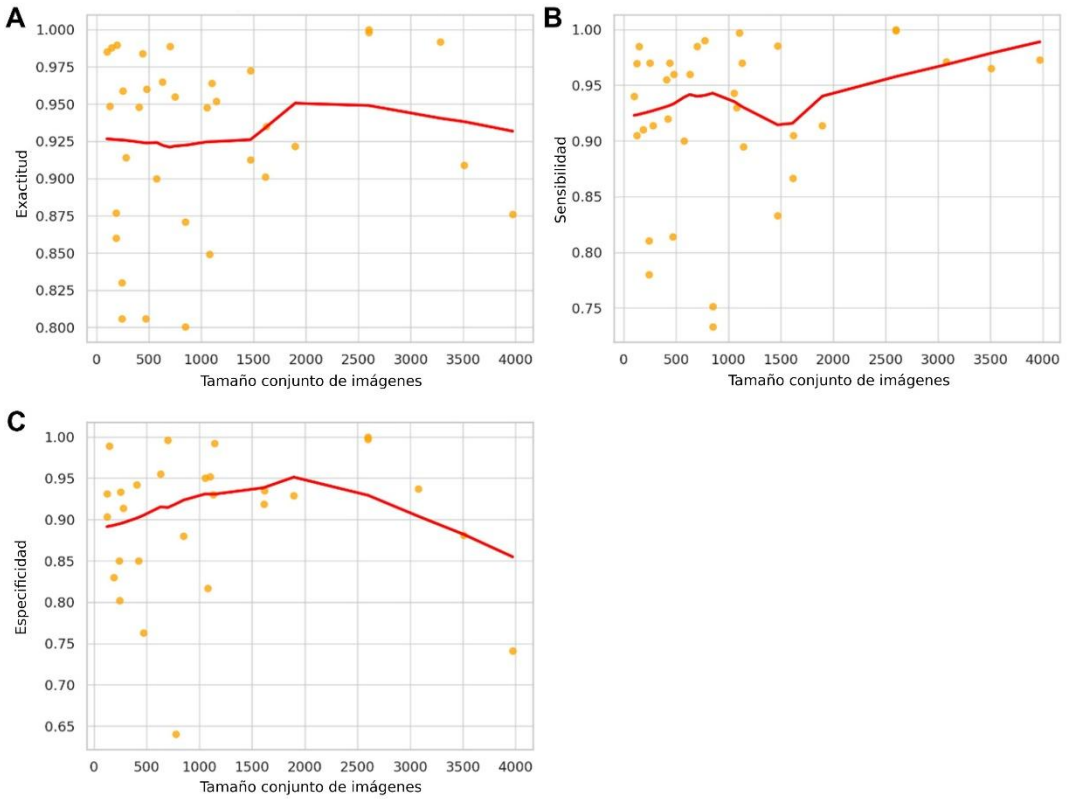

Figura 6. Relación entre el tamaño del conjunto de datos de entrenamiento y el rendimiento en estudios con ≤ 5 000 imágenes. Diagramas de dispersión para (A) exactitud, (B) sensibilidad y (C) especificidad; la línea roja indica el ajuste LOWESS y las bandas del 95% del intervalo de confianza. Fuente: Elaboración propia.

5.1.4. Efecto de la segmentación: automática frente a manual

Uno de los hallazgos más consistentes de la revisión es el impacto de la estrategia de segmentación sobre el rendimiento de los modelos. Al comparar estudios con segmentación automática (n = 27) frente a segmentación manual (n = 17), se observó que la exactitud media de los modelos con segmentación automática fue del 94.2% ± 4.3, mientras que en el grupo con segmentación manual fue del 88.2% ± 6.6. Esta diferencia resultó estadísticamente significativa (Mann–Whitney U = 234.0; p = 0.007). La sensibilidad también fue mayor en el grupo automático (93.7% ± 5.6 frente a 88.6% ± 8.3; p = 0.042). La especificidad media fue de 92.5% ± 6.0 con segmentación automática y 87.3% ± 9.5 en segmentación manual, aunque la

diferencia mostró una tendencia a favor de la segmentación automática, no alcanzó significación estadística (p = 0.084). Las AUC no difirieron de forma significativa entre ambos grupos (p = 0.839).

Estos resultados se representan gráficamente en la Figura 7, que muestra diagramas de cajas de exactitud, sensibilidad, especificidad y AUC estratificados por tipo de segmentación. La figura pone de manifiesto valores medios más altos y menor dispersión en la mayoría de las métricas para los estudios con segmentación automática, así como una mayor presencia de valores atípicos en el grupo de segmentación manual, especialmente para la especificidad, lo que sugiere menor consistencia. Estas diferencias visuales coinciden con la hipótesis de que la segmentación automatizada mejora la reproducibilidad y la estandarización.

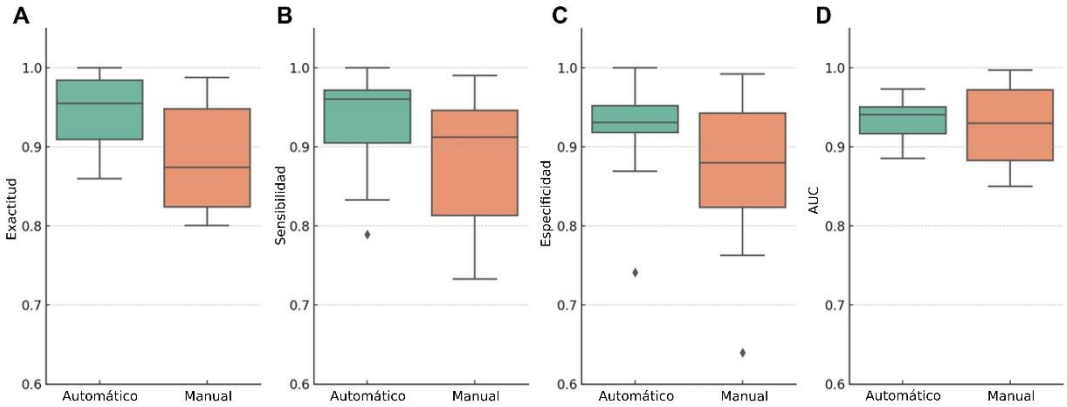

Figura 7. Distribución de métricas de desempeño en estudios que emplean segmentación automática o manual. Diagramas de caja para (A) exactitud, (B) sensibilidad, (C) especificidad y (D) AUC; los puntos representan estudios individuales. Fuente: Elaboración propia.

La Tabla 2 complementa estos hallazgos al mostrar que los estudios con segmentación automática no solo rinden mejor, sino que tienden a utilizar conjuntos de datos notablemente mayores, con desviaciones estándar menores en exactitud y especificidad.

Tabla 2. Tamaño del conjunto de datos y rendimiento según el tipo de segmentación.

Segmentación	*Tamaño conjunto (μ ± σ)*	*Exactitud (μ ± σ)*	*Sensibilidad (μ ± σ)*
Automática	*22941 ± 110530*	*0.94 ± 0.04*	*0.94 ± 0.06*
Manual	*1121 ± 1819*	*0.88 ± 0.07*	*0.89 ± 0.08*

5.1.5. Metarregresión de factores metodológicos

Para analizar de forma conjunta la influencia de distintos factores metodológicos sobre la exactitud, se realizó una metarregresión usando esta métrica como variable dependiente y cuatro predictores: tamaño del conjunto de imágenes, tipo de segmentación, tipo de arquitectura y riesgo de sesgo (alto o bajo). El modelo incluyó 26 estudios con datos completos y resultó globalmente significativo ($F = 3.98$; $p = 0.015$), con un R^2 ajustado de 0.32, lo que indica que los predictores considerados explican aproximadamente el 32% de la variabilidad en la exactitud reportada.

La Figura 8 muestra los coeficientes de la metarregresión con sus intervalos de confianza al 95%. El único predictor estadísticamente significativo fue la segmentación automática, asociada a un incremento de 6.6 puntos porcentuales en exactitud frente a la segmentación manual ($\beta = 0.0656$; $p = 0.007$). El tamaño de la base de datos, el hecho de utilizar o no una CNN y el riesgo de sesgo no alcanzaron significación estadística ($p > 0.05$), si bien el efecto de alto riesgo de sesgo mostró una tendencia a la sobreestimación de la exactitud ($\beta = 0.0427$; $p = 0.096$). Estos resultados refuerzan la noción de que la calidad y estandarización de la segmentación constituye un determinante principal del rendimiento de los modelos de IA en este contexto.

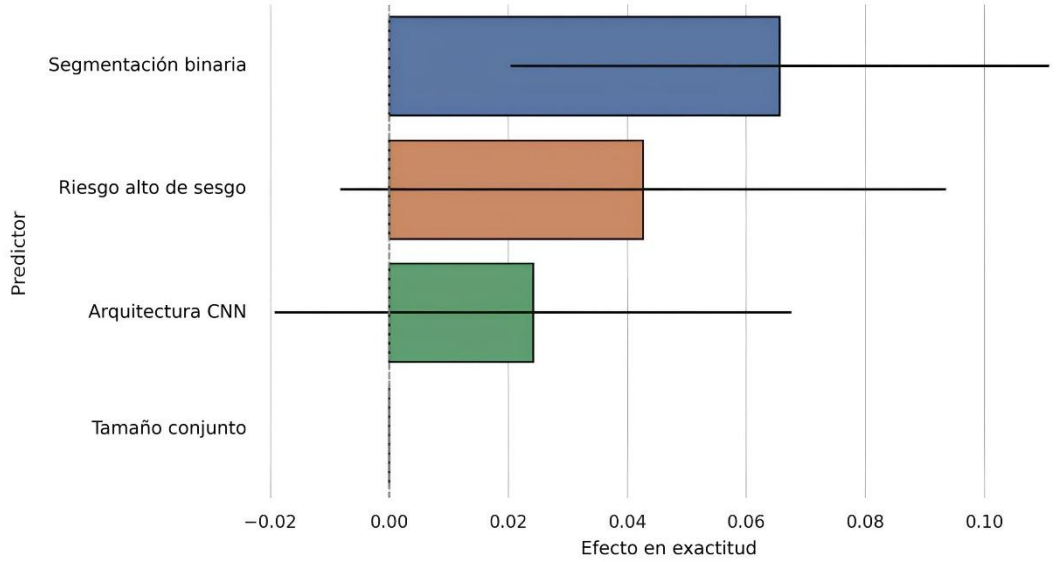

Figura 8. Coeficientes del análisis de metarregresión que evalúa la asociación de factores metodológicos con la exactitud diagnóstica. Se muestran los coeficientes estimados y su significancia estadística para las covariables incluidas. Fuente: Elaboración propia.

5.1.6. Evolución temporal de las arquitecturas de IA

El análisis temporal del tipo de arquitectura empleada en los estudios reveló una transición clara desde técnicas clásicas de ML hacia modelos de DL. Entre 2014 y 2018, aproximadamente el 85% de los trabajos utilizaban algoritmos clásicos como máquinas de soporte vectorial, bosques aleatorios y regresión logística, mientras que las ANNs representaban el 15% restante y no se documentaron modelos basados en CNNs antes de 2020. A partir de 2020, las CNNs comenzaron a aparecer y, en 2022, ya constituían el 65% de las metodologías utilizadas, alcanzando el 78% en 2024.

La Figura 9 presenta un gráfico de barras apiladas que resume esta evolución anual, mostrando el incremento progresivo de modelos de CNNs y la aparición, exclusivamente en 2024, de modelos basados en ViT, que representan alrededor del 10% de las metodologías de ese año. Asimismo, se observa la incorporación puntual de algoritmos de detección facial (ADF) y la presencia limitada de modelos de redes neuronales recurrentes (RNN), ambos con una contribución minoritaria a lo largo del periodo analizado. En conjunto, la figura evidencia una transición metodológica hacia arquitecturas más sofisticadas, coherente con las tendencias observadas en IA aplicada a imagen médica en otros campos.

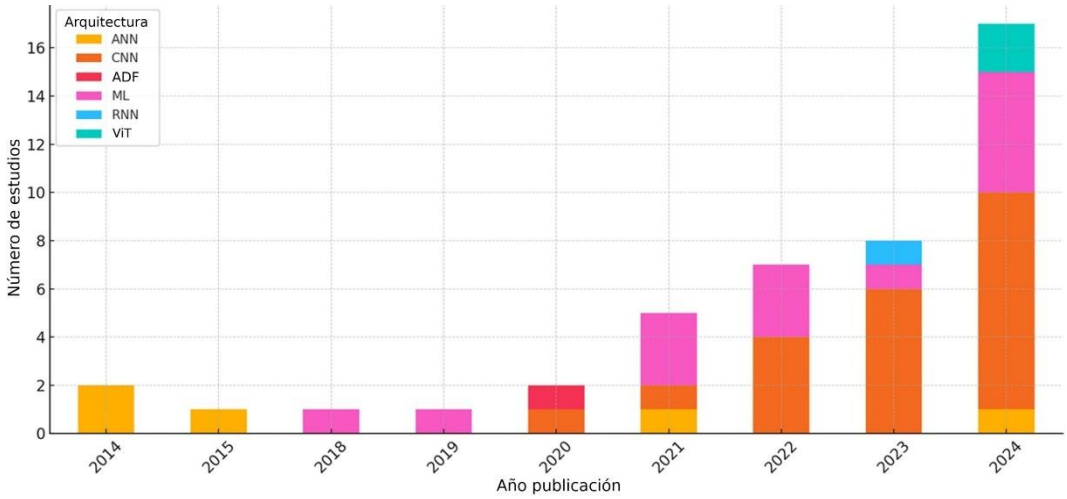

Figura 9. Evolución temporal (2014–2024) de las arquitecturas de IA empleadas en los estudios incluidos, agrupadas por familia de modelos. Fuente: Elaboración propia.

5.1.7. Heterogeneidad y riesgo de sesgo

La evaluación del riesgo de sesgo mediante PROBAST clasificó 26 estudios como de bajo riesgo, 12 como de riesgo incierto y 6 como de alto riesgo. La Figura 10 resume esta distribución por dominios (participantes, predictores, resultados y análisis) y muestra, en un segundo panel, las métricas de rendimiento promedio (exactitud, sensibilidad, especificidad y AUC) estratificadas por nivel de riesgo. Los estudios clasificados como de alto riesgo, aunque reportan exactitudes medias muy elevadas (96.1% ± 7.6), sensibilidades de 95.4% ± 7.9 y especificidades de 93.9% ± 11.7, presentan una elevada dispersión y, de forma notable, ninguno de ellos informa valores de AUC, lo que impide valorar adecuadamente su capacidad discriminativa. En contraste, los estudios de bajo riesgo muestran métricas algo más modestas (exactitud 90.4% ± 5.2; sensibilidad 90.7% ± 7.0; especificidad 89.1% ± 8.2; AUC 0.935 ± 0.039), pero con menor variabilidad y reporte más completo de métricas, lo que sugiere una mayor fiabilidad metodológica y aplicabilidad clínica.

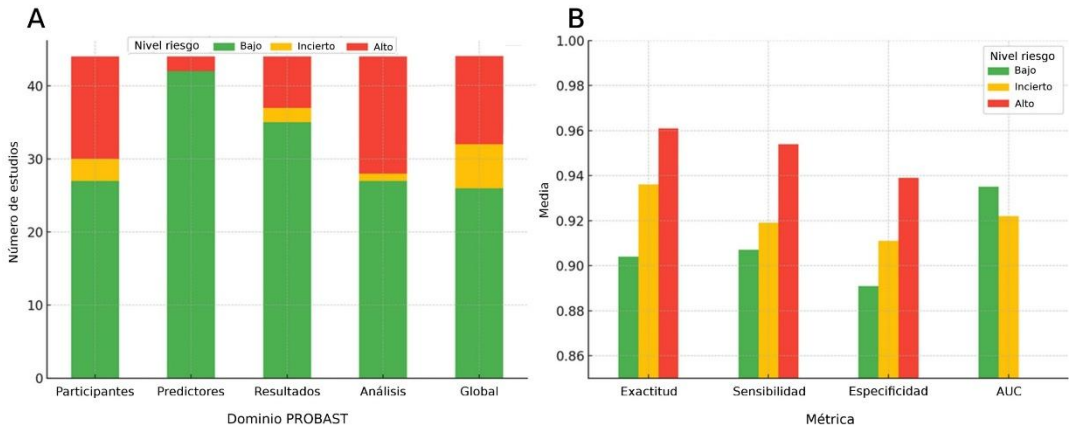

Figura 10. Riesgo de sesgo y rendimiento reportado en los estudios incluidos. (A) Distribución por dominios PROBAST y por riesgo global (bajo, incierto, alto). (B) Medias de exactitud, sensibilidad, especificidad y AUC estratificadas por riesgo global de sesgo. Fuente: Elaboración propia.

En síntesis, el Estudio 1 concluye que los modelos de IA aplicados a UST para la clasificación de masas ováricas alcanzan, en promedio, un desempeño elevado, pero su potencial traslación clínica se ve limitada por la heterogeneidad metodológica, la falta de validaciones externas estandarizadas y, de manera singular, por el papel crítico de la segmentación como factor modulador del rendimiento.

5.2. RESULTADOS DEL ESTUDIO 2: MODELOS HÍBRIDOS CNN–VIT EN OTU-2D

5.2.1. Rendimiento global y comparación entre modelos individuales e híbridos

El Estudio 2 evaluó cinco arquitecturas preentrenadas (tres CNNs y dos ViTs) y sus combinaciones híbridas CNN–ViT con fusión temprana, utilizando exclusivamente las 1469 imágenes UST en modo B de la base OTU-2D. Todos los resultados se obtuvieron mediante validación cruzada de cinco pliegues, repetida diez veces. Por tanto, las métricas reportadas corresponden al promedio de diez ejecuciones independientes.

La Tabla 3 resume las métricas diagnósticas (exactitud, sensibilidad, especificidad, AUC, F1-macro y MCC) con intervalos de confianza al 95% obtenidos mediante *bootstrap* (n = 500) para cada arquitectura individual y para los diferentes modelos híbridos. El modelo híbrido EfficientNetB7–Swin Transformer fue el que alcanzó mejor rendimiento global, con una exactitud de 92.13% [IC95%: 90.7–93.2], una sensibilidad de 92.38% [IC95%: 91.0–93.4], una especificidad de 98.9% [IC95%: 98.7–99.0], un AUC-ROC de 0.9904 [IC95%: 0.987–0.993], un F1-macro de 0.921 [IC95%: 0.908–0.932] y un MCC de 0.910 [IC95%: 0.894–0.923].

Antes de las comparaciones, se comprobó la normalidad de las diferencias mediante Shapiro–Wilk. En función de estos resultados, se aplicaron pruebas t de Student pareadas o, cuando procedía, la prueba no paramétrica de rangos con signo de Wilcoxon. El modelo EfficientNetB7–Swin Transformer mostró diferencias estadísticamente significativas frente a las arquitecturas individuales. En concreto, superó a EfficientNetB7 sola en 17.1 puntos porcentuales de exactitud, 26.6 puntos en sensibilidad, 2.6 puntos en especificidad y 8.0 puntos en AUC ($p < 0.001$ en todas las métricas), y superó a Swin Transformer en 20.1 puntos de exactitud, 28.9 puntos de sensibilidad, 3.1 puntos de especificidad y 7.3 puntos de AUC ($p < 0.001$ en todas las métricas).

La Tabla 4 recoge los p-valores de las comparaciones entre EfficientNetB7–Swin Transformer y el resto de los modelos híbridos, obtenidos mediante prueba de Wilcoxon. Aunque algunos p-valores para AUC no fueron menores de 0.05 (por ejemplo, frente a ResNet152–ViTB16 y DenseNet201–Swin Transformer), las métricas de exactitud, sensibilidad y especificidad mostraron diferencias

significativas en todos los casos, lo que justifica la selección de EfficientNetB7–Swin Transformer como modelo óptimo dentro del conjunto evaluado.

Tabla 3. Métricas de bootstrap (n = 500): exactitud, sensibilidad, especificidad, AUC, F1-macro, MCC para cada algoritmo.

Modelo	Exactitud	Sensibilidad	Especificidad	AUC	F1-macro	MCC
ResNet152	0.735 [0.707, 0.764]	0.632 [0.590, 0.676]	0.959 [0.955, 0.964]	0.926 [0.913, 0.940]	0.661 [0.619, 0.704]	0.676 [0.641, 0.711]
DenseNet201	0.776 [0.749, 0.802]	0.662 [0.622, 0.700]	0.966 [0.962, 0.970]	0.922 [0.905, 0.938]	0.689 [0.649, 0.727]	0.728 [0.697, 0.758]
EfficientNetB7	0.750 [0.721, 0.779]	0.656 [0.618, 0.698]	0.962 [0.958, 0.967]	0.911 [0.893, 0.929]	0.675 [0.635, 0.716]	0.697 [0.663, 0.731]
ViTB16	0.686 [0.653, 0.714]	0.587 [0.547, 0.627]	0.953 [0.948, 0.957]	0.896 [0.877, 0.914]	0.595 [0.553, 0.635]	0.620 [0.580, 0.653]
Swin	0.719 [0.690, 0.747]	0.634 [0.594, 0.673]	0.958 [0.954, 0.962]	0.917 [0.902, 0.933]	0.637 [0.598, 0.672]	0.659 [0.626, 0.692]
ResNet152-ViTB16	0.913 [0.898, 0.925]	0.912 [0.898, 0.925]	0.988 [0.986, 0.989]	0.990 [0.987, 0.992]	0.908 [0.893, 0.920]	0.900 [0.883, 0.915]
DenseNet201-ViTB16	0.908 [0.893, 0.923]	0.907 [0.894, 0.922]	0.987 [0.985, 0.989]	0.988 [0.984, 0.991]	0.907 [0.893, 0.922]	0.895 [0.879, 0.912]
EfficientNetB7-ViTB16	0.905 [0.892, 0.919]	0.906 [0.893, 0.919]	0.986 [0.985, 0.989]	0.985 [0.981, 0.988]	0.904 [0.891, 0.918]	0.892 [0.876, 0.908]
ResNet152-Swin	0.915 [0.901, 0.929]	0.917 [0.903, 0.931]	0.988 [0.986, 0.990]	0.992 [0.989, 0.994]	0.914 [0.900, 0.929]	0.903 [0.887, 0.919]
DenseNet201-Swin	0.913 [0.900, 0.927]	0.915 [0.902, 0.927]	0.988 [0.986, 0.989]	0.990 [0.987, 0.993]	0.913 [0.900, 0.926]	0.901 [0.885, 0.916]
EfficientNetB7-Swin	0.921 [0.907, 0.932]	0.923 [0.910, 0.934]	0.989 [0.987, 0.990]	0.990 [0.987, 0.993]	0.921 [0.908, 0.932]	0.910 [0.894, 0.923]

Los valores se presentan como media ± IC del 95% de validación cruzada de 5 pliegues, repetida 10 veces. Los IC del 95% se calcularon mediante *bootstrap* (n = 500).
AUC, área bajo la curva ROC; F1, puntuación F1 promediada a nivel macro; MCC, coeficiente de correlación de Matthews; ViTB16, Vision Transformer; Swin, Swin Transformer.

Tabla 4. Resultados del valor p del análisis estadístico (pruebas t de Student pareadas o de rangos con signo de Wilcoxon) que compara EfficientNetB7–Swin Transformer con otros modelos híbridos (n = 10 mediciones).

Modelo	Exactitud	Sensibilidad	Especificidad	AUC
ResNet152-ViTB16	<0.01	<0.01	<0.01	0.5930
DenseNet201-ViTB16	<0.01	<0.01	<0.01	<0.01
EfficientNetB7-ViTB16	<0.01	<0.01	<0.01	<0.01
ResNet152-Swin	<0.01	<0.01	<0.01	<0.01
DenseNet201-Swin	0.0488	0.0195	0.0137	0.4922

5.2.2. Rendimiento por categoría clínica y matriz de confusión

La Figura 11 presenta la matriz de confusión normalizada del modelo híbrido EfficientNetB7–Swin Transformer. La matriz muestra una marcada diagonal, indicativa de una alta tasa de aciertos en todas las categorías. La sensibilidad por clase alcanza valores de 1.00 para carcinoma seroso de alto grado, 0.99 para tumor de células de la teca, 0.98 para quiste simple y 0.97 para cistoadenoma mucinoso. En el resto de las categorías, la sensibilidad es menor: cistoadenoma seroso, teratoma, ovario normal y endometrioma se sitúan en 0.89, 0.86, 0.87 y 0.82, respectivamente.

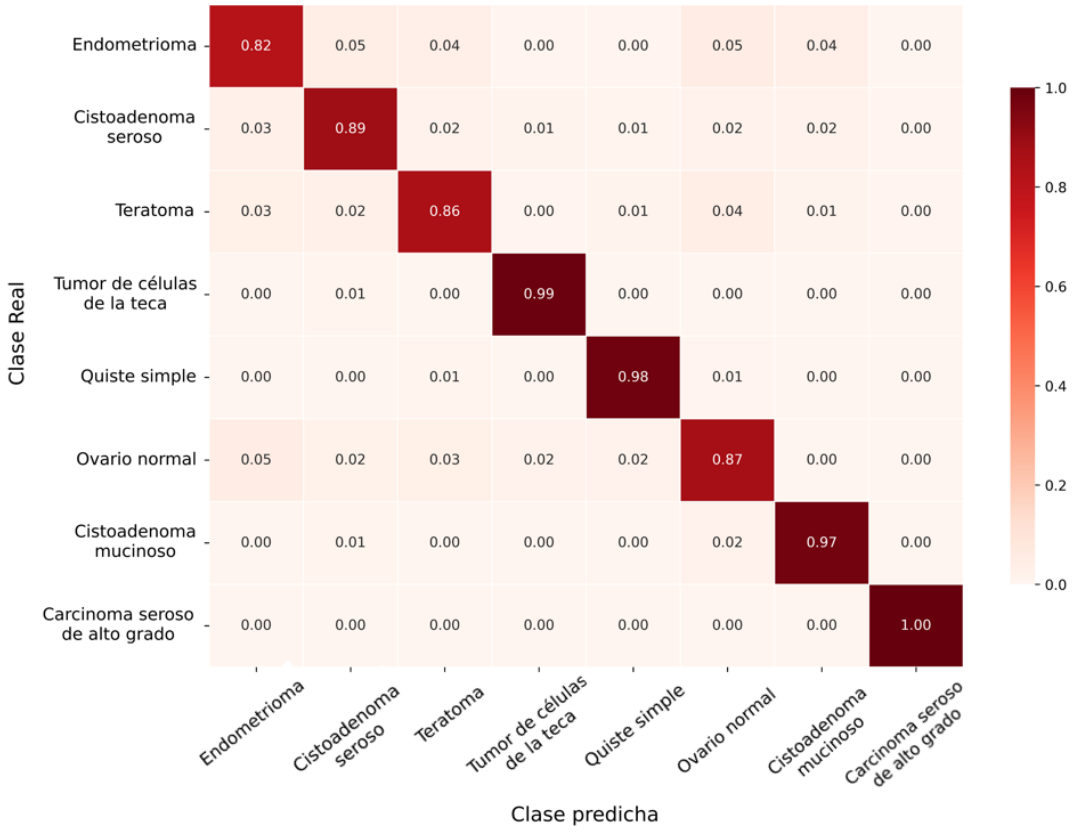

Figura 11. Matriz de confusión normalizada del modelo híbrido EfficientNetB7-Swin Transformer en la tarea de clasificación multiclase (ocho categorías) sobre el conjunto de test. Fuente: Elaboración propia.

La Tabla 5 desglosa las métricas por categoría para el modelo EfficientNetB7–Swin Transformer. En dicha tabla se indica que las categorías carcinoma seroso de alto grado y tumor de células de la teca presentan sensibilidades cercanas al 100%, especificidades superiores al 99% y F1-scores por encima del 97%. Las categorías quiste simple y cistoadenoma mucinoso muestran especificidades superiores al 98%. Incluso en las clases con mayores tasas de error relativo (endometrioma y ovario normal), la sensibilidad se mantiene en al menos 82% y la especificidad en al menos 98%, lo que indica una capacidad de generalización elevada en contextos clínicamente diversos. Los errores se concentran en confusiones entre entidades de morfología ecográfica similar, principalmente entre diferentes tipos de quistes benignos y, en menor medida, con ovario normal, lo que coincide con los patrones de solapamiento sonográfico descritos en la literatura clínica [119–121].

Tabla 5. Métricas por clase (exactitud, sensibilidad, especificidad y F1-score), calculadas como el promedio de las diez repeticiones del modelo híbrido EfficientNetB7–Swin Transformer.

Clase	*Exactitud*	*Sensibilidad*	*Especificidad*	*F1-score*
Endometrioma	*0.8929*	*0.8238*	*0.9842*	*0.8569*
Cistoadenoma seroso	*0.8910*	*0.8859*	*0.9836*	*0.8883*
Teratoma	*0.8969*	*0.8572*	*0.9858*	*0.8766*
Tumor de células de la teca	*0.9617*	*0.9921*	*0.9943*	*0.9765*
Quiste simple	*0.9508*	*0.9835*	*0.9937*	*0.9669*
Ovario normal	*0.8563*	*0.8745*	*0.9802*	*0.8650*
Cistoadenoma mucinoso	*0.9229*	*0.9731*	*0.9895*	*0.9473*
Carcinoma seroso de alto grado	*0.9914*	*1.0000*	*0.9986*	*0.9957*

5.2.3. Estabilidad del entrenamiento y convergencia de las métricas

La Figura 12 ilustra la evolución de las principales métricas de rendimiento y de la pérdida durante 50 épocas de entrenamiento del modelo EfficientNetB7–Swin Transformer, promediadas en diez ejecuciones independientes, junto con la DE asociada. Las curvas muestran una convergencia rápida: la pérdida se estabiliza en

torno a la época 4, la exactitud alrededor de la época 3, la sensibilidad en la época 2, la especificidad en la época 3 y el AUC ya en la época 2. El panel F de la misma figura refleja la disminución progresiva de la DE entre repeticiones a medida que avanza el entrenamiento, hasta converger en valores muy bajos al final de las 50 épocas. Estos resultados indican que el esquema de entrenamiento empleado proporciona una convergencia estable y reproducible.

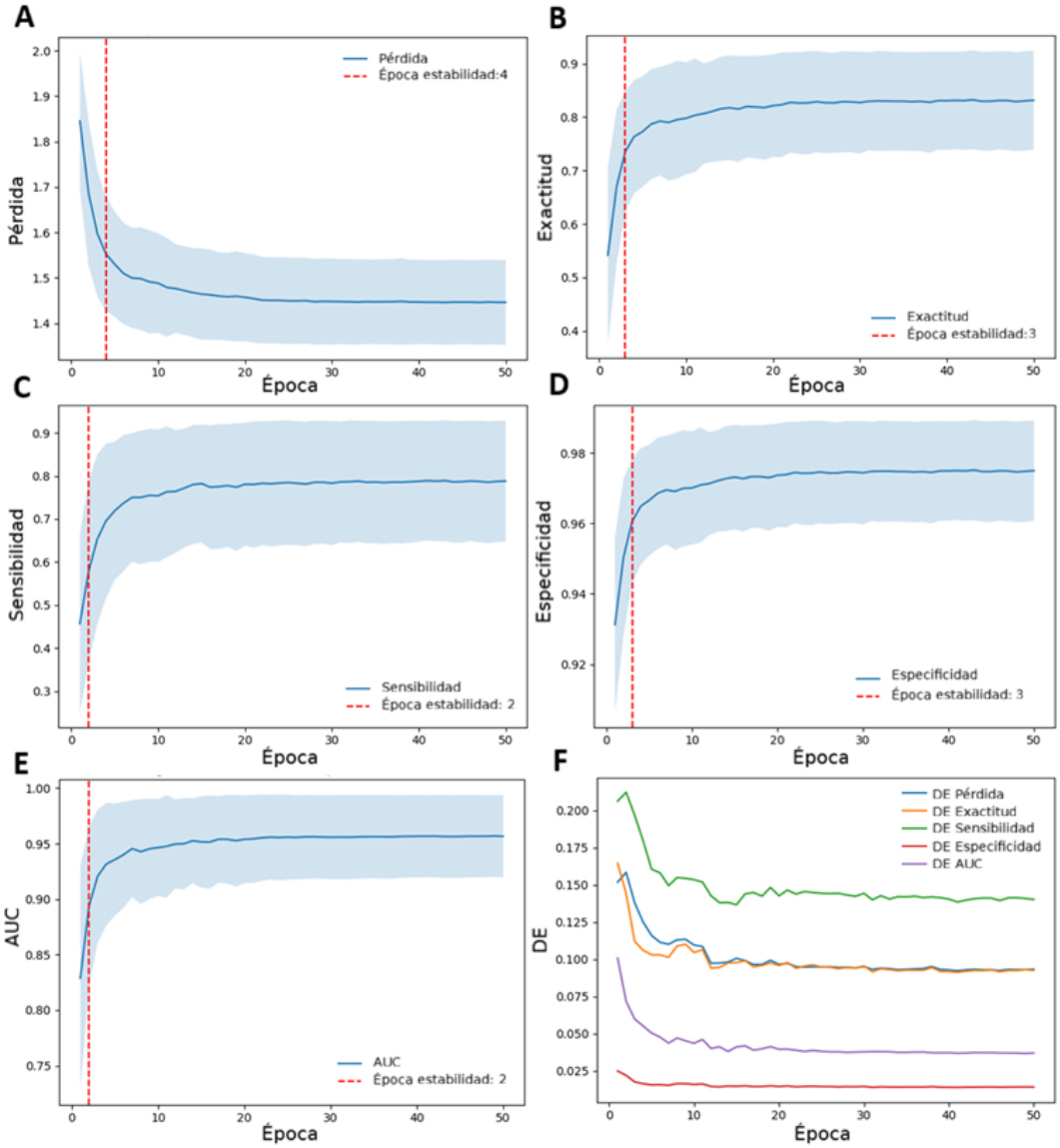

Figura 12. Curvas de entrenamiento del modelo híbrido EfficientNetB7–Swin Transformer a lo largo de 50 épocas en 10 ejecuciones independientes. Media ± desviación estándar (DE) para (A) pérdida, (B) exactitud, (C) sensibilidad, (D) especificidad y (E) AUC; (F) DE entre repeticiones por época para todas las métricas. Fuente: Elaboración propia.

5.2.4. Calibración probabilística y curvas de decisión clínica

La Figura 13 resume dos aspectos complementarios del modelo EfficientNetB7–Swin Transformer: la calidad de las probabilidades que produce (Figura 13A) y su utilidad potencial en un escenario clínico mediante análisis de beneficio neto (Figura 13B). Para este análisis se planteó un objetivo binario potencialmente maligno vs no potencialmente maligno (carcinoma seroso de alto grado, cistoadenoma mucinoso y cistoadenoma seroso). Esta definición no implica equivalencia con malignidad histopatológica, se utiliza para agrupar categorías de mayor relevancia clínica para el triaje.

A partir de la salida multiclase del modelo, se definió una probabilidad de potencial malignidad como la suma de probabilidades de estas tres clases. Sobre esta probabilidad se aplicó regresión isotónica, una técnica de calibración *post-hoc*, que ajusta la confianza para que las probabilidades sean más coherentes con la frecuencia observada. La regresión isotónica se ajusta en validación y posteriormente se aplica y evalúa en test.

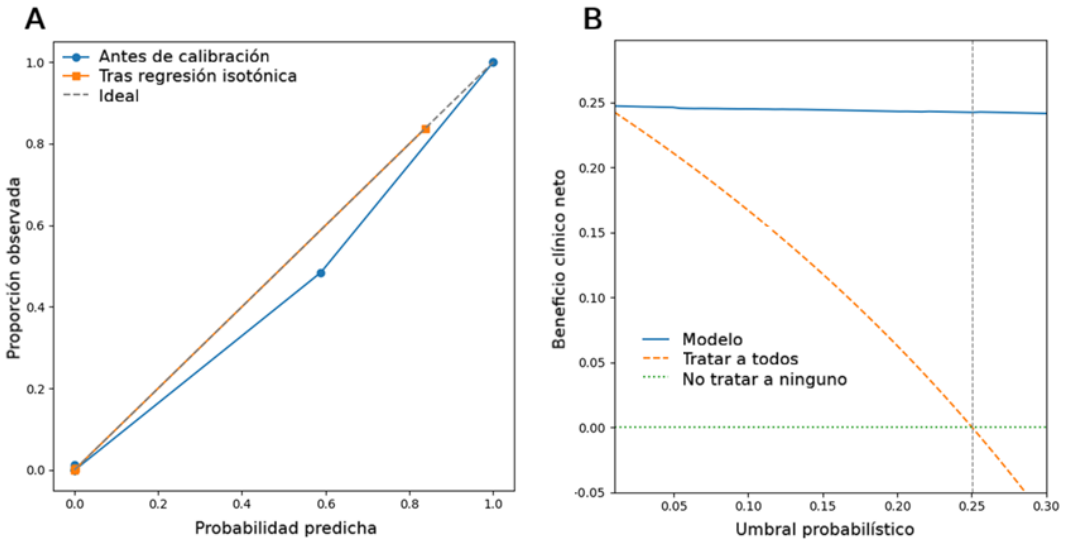

Figura 13. Calibración y utilidad clínica del modelo EfficientNetB7–Swin Transformer para la predicción de potencial malignidad. (A) Curva de calibración antes (azul) y después (naranja) de aplicar regresión isotónica; la línea discontinua indica calibración ideal. (B) Curvas de decisión clínica (CDC): beneficio neto del modelo (azul) frente a "tratar a todos" (naranja) y "no tratar a ninguno" (verde) en función del umbral probabilístico. La línea vertical indica la prevalencia del evento en el conjunto evaluado. Fuente: Elaboración propia.

En la Figura 13A se observa la curva de calibración antes y después de la corrección. Antes de calibrar (curva azul), el modelo muestra una ligera sobreestimación de las probabilidades en rangos intermedios, ya que la curva queda por debajo de la línea ideal. Tras aplicar regresión isotónica (curva naranja), la curva se aproxima a la diagonal, indicando una mejor correspondencia entre probabilidad predicha y proporción observada.

En la Figura 13B se presenta el análisis de CDC, que evalúa el beneficio clínico neto del modelo a distintos umbrales probabilísticos. La curva del modelo (azul) se sitúa por encima de las estrategias "tratar a todos" y "no tratar a ninguno" en un rango de umbrales exploratorio (aprox. 0.05–0.20). Este intervalo se muestra porque es típico de decisiones de triaje orientadas a alta sensibilidad (se aceptan más falsos positivos) y porque queda por debajo de la prevalencia observada (≈0.25), donde suele analizarse la utilidad clínica. La línea vertical indica la prevalencia observada del evento (≈0.25), coherente con el punto donde la estrategia "tratar a todos" cruza beneficio neto cero.

5.2.5. Optimización mediante ensamble suave

Con el objetivo de explorar posibles mejoras adicionales en el rendimiento, se aplicó una estrategia de ensamble suave combinando las probabilidades de las tres arquitecturas híbridas con mejor desempeño: DenseNet201–Swin Transformer, EfficientNetB7–Swin Transformer y ResNet152–Swin Transformer. Las probabilidades de salida se ponderaron de forma proporcional a la AUC obtenida con el conjunto de test de cada modelo. Los resultados de este ensamble se recogen en la Tabla 6Tabla 6, y se discuten adicionalmente en la Tabla 7, que permite la comparación con estudios previos.

Tabla 6. Métricas de rendimiento diagnóstico [exactitud, sensibilidad, especificidad, AUC, F1-score y coeficiente de correlación de Matthews (MCC)] obtenidas mediante optimización por ensamble suave.

Estadístico	*Exactitud*	*Sensibilidad*	*Especificidad*	*AUC*	*F1*	*MCC*
Media ± DE	*0.933 ± 0.005*	*0.936 ± 0.003*	*0.990 ± 0.001*	*0.991 ± 0.004*	*0.933 ± 0.005*	*0.924 ± 0.006*

El ensamble alcanzó una exactitud de 93.3% ± 0.5, una sensibilidad de 93.6% ± 0.3, una especificidad de 99.0% ± 0.1 y una AUC de 0.991 ± 0.004. Estas mejoras

respecto al modelo EfficientNetB7–Swin Transformer individual se acompañan de una mayor estabilidad de las predicciones, lo que refuerza el potencial del uso de ensambles en contextos clínicos donde se requiera maximizar la robustez sin sacrificar eficiencia computacional.

Tabla 7. Comparación de estudios recientes de IA de UST en modo B para la clasificación de tumores ováricos.

Estudio	*Cohorte*	*Nº Clases*	*Modelo*	*Validación externa*	*Métrica principal*	*Explicabilidad / calibración / Análisis CDC*
Nuestro trabajo	*1 469 imágenes; centro único*	*8*	*Híbrido. EfficientNet-B7 (CNN) + Swin Transformer (ViT)*	*No*	*AUC 0.9904; Exac 92.13%; Sens 92.38%; Esp 98.90%*	*Sí / Sí (isotónica) / Sí*
***Gao et al.* [36]**	*~105 000 imágenes multicéntrico*	*2*	*modelo CNN personalizado*	*Sí*	*AUC 0.911; Exac 88.8%; Sens 82.7%; Esp 88.7%*	*No / No / No*
***Dai et al.* [122]**	*8 500 imágenes; multicéntrico*	*2*	*CNN multitarea*	*Sí*	*AUC 0.941; Exac 86.2%; Sens 81.8%; Esp 89.2%*	*Sí / No / No*
***Wu et al.* [123]**	*1 142 imágenes; centro único*	*7*	*ResNeXt50*	*No*	*AUC 0.95; Sens 90%; Esp 99.2%*	*Sí / No / No*
***Giourga et al.* [124]**	*3 510 imágenes; centro único*	*2*	*Ensamble VGG16, ResNet50 y InceptionV3*	*No*	*AUC 0.922; Exac 90.9%; Sens 96.5%; Esp 88.1%*	*No / No / No*
***Xie et al.* [125]**	*519 pacientes; multicéntrico*	*2*	*Ensamble DeepLabV3 y YOLOv8*	*No*	*AUC 0.950; Exac 94.1%; Sens 92.5%; Esp 95.5%*	*Sí (YOLO) / No / No*
***Du et al.* [126]**	*849 pacientes; centro único*	*3*	*Híbrido de radiómica + CNN + características clínicas*	*No*	*AUC 0.84*	*No / No / No*
***Du et al.* [127]**	*849 pacientes; centro único*	*2*	*Híbrido de radiómica + CNN + características clínicas*	*No*	*AUC 0.928*	*Sí / Sí / Sí*
***Barcroft et al.* [128]**	*1 444 imágenes; multicéntrico*	*2*	*Híbrido de CNN + radiómica (segmentación con U-Net)*	*Sí*	*AUC 0.90; Sens 83%*	*No / No / No*

5.2.6. Incertidumbre predictiva y análisis de la entropía

La Figura 14 analiza la relación entre la incertidumbre de las predicciones, cuantificada mediante la entropía de la distribución de probabilidades, y la tasa de error. Las predicciones se agrupan en deciles de entropía, desde el decil 1 (baja entropía, alta confianza) hasta el decil 10 (alta entropía, baja confianza). En los deciles de menor entropía (aprox. 1–7) la tasa de error se mantiene muy baja (cercana a cero), lo que sugiere que, cuando el modelo emite distribuciones de probabilidad concentradas, tiende a acertar con alta consistencia. Sin embargo, conforme aumenta la entropía, el error crece de manera marcada: a partir de los deciles superiores (8–10) la tasa de error aumenta, alcanzando aproximadamente 20% en el decil 9 y alrededor de 43% en el decil 10. Este patrón respalda que la entropía actúa como un indicador útil de incertidumbre, capaz de discriminar entre predicciones típicamente fiables y predicciones con mayor probabilidad de fallo.

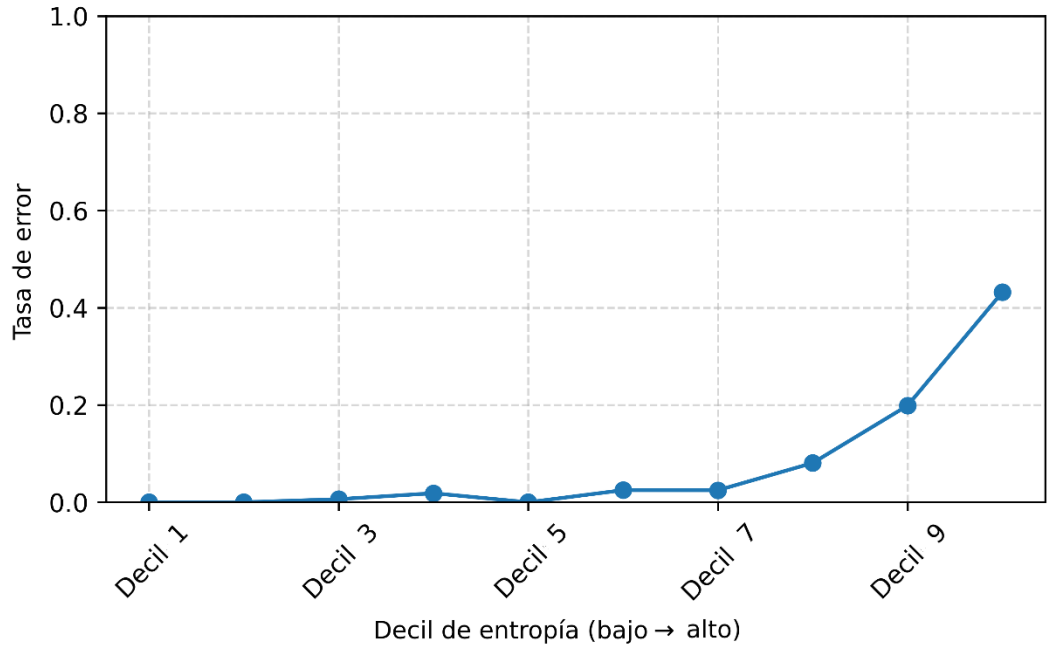

Figura 14. Tasa de error en función de la entropía de la predicción, agrupada por deciles (1 = entropía más baja/mayor confianza; 10 = entropía más alta/menor confianza). Fuente: Elaboración propia.

Estos resultados permiten plantear una estrategia práctica de automatización selectiva. Por ejemplo, si se desea mantener una tasa de error baja en los casos automatizados, podría automatizarse una fracción mayoritaria de predicciones con entropía baja (p. ej., deciles 1–8) y reservar los deciles superiores (9–10) para revisión experta, alternativamente, el umbral puede ajustarse según la tolerancia al riesgo y el contexto clínico (coste relativo de falsos positivos y falsos negativos). En

conjunto, la Figura 14 sugiere una vía concreta para combinar eficiencia operativa y seguridad clínica en la implementación de sistemas de IA, al condicionar la automatización a un criterio explícito de incertidumbre.

5.2.7. Interpretabilidad mediante Grad-CAM

La Figura 15 presenta ejemplos representativos de mapas de activación Grad-CAM generados para el modelo híbrido EfficientNetB7–Swin Transformer. Grad-CAM es un método de interpretabilidad *post-hoc* que estima, para una predicción concreta, qué regiones de la imagen contribuyen más al resultado del modelo. En términos generales, el método utiliza los gradientes del score de la clase objetivo respecto a mapas de activación internos para construir un mapa de relevancia espacial, que se reescala y superpone sobre la imagen original. En el contexto del UST, esta aproximación es útil para comprobar si el modelo apoya sus decisiones en patrones anatómicos (p. ej., morfología de la lesión, contornos) o si, por el contrario, está capturando señales espurias (marcos, ruido periférico, anotaciones, etc.).

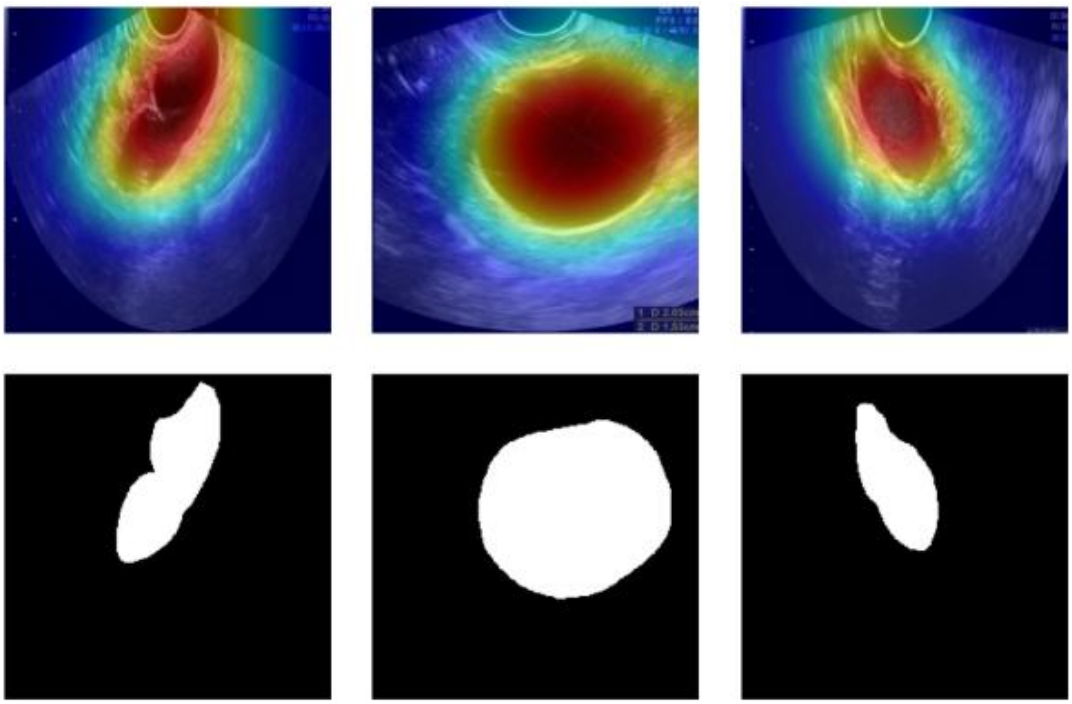

Figura 15. Mapas de activación Grad-CAM representativos en ultrasonido de tumores ováricos. Arriba: mapas de calor superpuestos (rojo/amarillo = mayor activación). Abajo: máscaras binarias de la ROI correspondiente. Fuente: Elaboración propia.

Los mapas de calor mostrados (Figura 15) evidencian que la activación dominante se concentra en torno a la región anatómica donde se ubica la lesión y, de manera cualitativa, guarda concordancia con la zona delimitada por los expertos. Esta convergencia sugiere que el modelo está utilizando información contenida en la lesión y su entorno inmediato para discriminar la clase. Asimismo, en varios casos la activación se distribuye no solo dentro de la máscara, sino también en su periferia cercana, lo cual es esperable en UST: características como el borde lesional, el gradiente de ecogenicidad y la interfaz lesión-tejido adyacente suelen aportar información diagnóstica.

Es importante subrayar que Grad-CAM no prueba causalidad, sino que ofrece una aproximación a la región de influencia. Por ello, su interpretación debe ser prudente y complementarse con verificaciones adicionales. En particular, en aplicaciones biomédicas es recomendable evaluar de manera sistemática: (i) si la activación cae predominantemente dentro de la ROI, (ii) si la activación se desplaza hacia artefactos cuando se altera el contenido de la lesión, y (iii) si los mapas son estables ante pequeñas perturbaciones. En esta línea, una extensión natural de este análisis sería cuantificar la concordancia entre mapas y máscaras mediante métricas de solapamiento y analizar ejemplos de error del modelo para identificar patrones recurrentes de confusión.

VI – DISCUSIÓN

VI -DISCUSIÓN

6.1. SÍNTESIS INTEGRADORA DE LOS DOS ESTUDIOS

Esta tesis doctoral aborda un reto clínico persistente: la caracterización fiable de la malignidad en masas ováricas mediante UST en modo B, en un contexto donde la interpretación sigue siendo dependiente del operador y donde la especificidad limitada puede traducirse en procedimientos invasivos innecesarios. Para responder a este problema, la investigación se estructura en dos componentes complementarios. El Estudio 1 sintetiza cuantitativamente el rendimiento reportado y, sobre todo, evalúa la calidad metodológica de la evidencia publicada. El Estudio 2 traslada esas lecciones a un protocolo experimental controlado y reproducible, en el que se desarrollan y validan modelos híbridos CNN–ViT con fusión temprana sobre la base OTU-2D.

La revisión sistemática mostró que, en promedio, los modelos de IA aplicados a imágenes de UST en modo B alcanzan una exactitud media del 92.3%, con sensibilidades y especificidades generalmente superiores al 90% y AUC alrededor de 0.93, lo que confirma que la IA tiene capacidad para discriminar masas ováricas con un rendimiento global elevado. No obstante, estas cifras deben interpretarse a la luz de una heterogeneidad metodológica sustancial entre estudios, con diferencias en prevalencia, definición de clases, protocolos de adquisición, estrategias de validación y umbrales de decisión. En este contexto, la exactitud funciona como un descriptor útil, pero potencialmente sensible a desbalances y a diferencias de prevalencia; por ello, la comparabilidad real depende de métricas más informativas (p. ej., AUC, F1-score, matrices de confusión e intervalos de confianza), cuyo reporte fue incompleto en una proporción relevante de la literatura incluida.

En este marco, el modelo híbrido EfficientNetB7–Swin Transformer con fusión temprana desarrollada en el Estudio 2 se sitúa en la franja alta de rendimiento observada en la revisión, con una exactitud superior al 92%, especificidad cercana al 99% y AUC en torno a 0.99, obtenidas exclusivamente en la base OTU-2D. La tesis muestra así que es posible diseñar modelos que no solo

igualan, sino que superan rendimientos reportados previamente bajo condiciones controladas, incorporando además componentes metodológicos que la revisión identifica como deficitarios en la literatura: calibración probabilística, evaluación de utilidad clínica mediante CDC y análisis de incertidumbre.

En conjunto, la contribución de la tesis se articula en dos niveles. Primero, proporciona evidencia cuantitativa sobre qué variables metodológicas se asocian de forma más robusta al rendimiento en la literatura, lo que permite priorizar mejoras (segmentación, validación externa…). Segundo, propone un modelo híbrido multicategoría que no se limita a maximizar las métricas, sino que produce salidas más cercanas a la toma de decisiones clínicas: probabilidades interpretables tras calibración, evidencia de utilidad potencial mediante beneficio neto y un criterio operativo basado en incertidumbre para derivación a revisión experta.

6.2. COMPARACIÓN CON REVISIONES PREVIAS Y POSICIONAMIENTO DEL TRABAJO

Los resultados de la revisión sistemática coinciden en gran medida con lo descrito en revisiones previas de IA en CO. El metaanálisis de Xu et al. sobre IA aplicada a distintas modalidades de imagen (UST, TC, RM) describió sensibilidades y especificidades agregadas del 88 y 85%, respectivamente, con AUC global de 0.93, cifras muy similares a las observadas aquí cuando se restringe el análisis a UST en modo B [32]. Mitchell et al. obtuvieron, en una revisión y metaanálisis centrados específicamente en ultrasonido, una sensibilidad global del 81% y especificidad del 92%, destacando también la heterogeneidad entre estudios y la necesidad de mejorar la calidad metodológica y la validación externa [31].

Lo que distingue la presente tesis de estas revisiones es, por un lado, el foco exclusivo en UST en modo B y, por otro, la incorporación de un análisis cuantitativo de factores metodológicos, en el que la estrategia de segmentación aparece como el predictor más robusto del rendimiento, mientras que el tamaño muestral o la familia de arquitectura no alcanzan significancia. Esta aproximación complementa revisiones más amplias sobre IA en CO que mezclan modalidades, algoritmos y escenarios clínicos muy dispares, dificultando extraer conclusiones operativas específicas para el UST.

Por otra parte, el estudio experimental con OTU-2D se inserta en una línea de trabajos que han explorado el uso de DL para la clasificación de masas ováricas a

partir de ecografía. Gao et al. demostraron que un modelo basado en DL podía alcanzar un rendimiento multicéntrico robusto en la clasificación binaria benigno–maligno utilizando ecografía pélvica [36], mientras que Dai et al. desarrollaron un *pipeline* de cribado con validaciones externas múltiples, también limitado a la discriminación binaria [122]. Mitchell et al. y Xu et al. ponen de relieve que la gran mayoría de estudios se circunscribe a tareas binarias y rara vez aborda la clasificación multicategoría de subtipos histológicos, y prácticamente ninguno incorpora calibración probabilística, CDC o análisis explícitos de incertidumbre [31, 32].

En este contexto, el presente trabajo se posiciona como una contribución diferencial al proponer un modelo multicategoría específico para UST 2D, con fusión temprana CNN–ViT y una evaluación que integra discriminación, calibración, utilidad clínica e incertidumbre.

6.3. APORTACIONES METODOLÓGICAS PRINCIPALES

La primera aportación metodológica relevante procede del Estudio 1: la identificación de la segmentación automática como factor clave asociado a una mayor exactitud y sensibilidad, incluso tras ajustar por tamaño de base de datos, arquitectura y riesgo de sesgo en la metarregresión. Este resultado sitúa el problema de la segmentación en el centro de la agenda de investigación: sin una segmentación robusta y estandarizada, el rendimiento aparente de los modelos puede depender más de cómo se recortan las imágenes que de la capacidad real del algoritmo para aprender patrones diagnósticos. Al mismo tiempo, y por tratarse de una metarregresión basada en evidencia agregada, esta asociación debe interpretarse como evidencia explicativa y no como causalidad estricta.

Esta constatación informa directamente el diseño del Estudio 2, donde se trabaja sobre una base (OTU-2D) con regiones de interés ya definidas y se aplica un *pipeline* de preprocesado homogéneo. Aunque en el artículo experimental no se exploran diferentes variantes de segmentación, la coherencia del flujo de procesamiento y el uso de aumentos específicos para UST contribuyen a reducir la variabilidad atribuible a factores ajenos al modelo y permiten una evaluación más limpia de la propuesta híbrida.

La segunda aportación metodológica central es la propia arquitectura de fusión temprana CNN–ViT. Mientras que muchos trabajos recientes en imagen médica utilizan fusiones tardías en las que las salidas de cada rama se combinan solo al final del *pipeline*, el modelo EfficientNetB7–Swin Transformer implementa una proyección conjunta temprana que permite que las representaciones locales y globales se ajusten de manera conjunta desde etapas iniciales del entrenamiento. El hecho de que este modelo supere de forma significativa tanto a las arquitecturas individuales como a otros híbridos evaluados sugiere que la integración precoz de información de contexto y detalle fino puede ser especialmente adecuada para ecografía ovárica, donde la interpretación depende simultáneamente de patrones locales (papilas, componentes sólidos) y de la configuración global de la masa y su entorno.

En tercer lugar, el Estudio 2 incorpora de manera sistemática elementos de evaluación que rara vez aparecen juntos en la literatura de IA en ecografía ovárica: curvas de calibración antes y después de regresión isotónica, CDC para cuantificar beneficio neto, análisis detallado de entropía como marcador de incertidumbre y mapas Grad-CAM para exploración cualitativa de la atención del modelo. Esta combinación responde a recomendaciones de marcos como TRIPOD-AI, PROBAST-AI, STARD-AI, que subrayan que un modelo clínicamente relevante debe ser no solo preciso, sino también bien calibrado, interpretable y acompañado de indicadores claros sobre cuándo sus predicciones son suficientemente fiables como para automatización parcial. En consecuencia, esta tesis contribuye en esa dirección al documentar con detalle configuración y condiciones de entrenamiento para favorecer reproducibilidad y auditoría.

Finalmente, el uso de validación cruzada y las diez muestras por algoritmo ofrece estimaciones más estables y menos dependientes de una única partición, alineándose con las críticas recurrentes a estudios previos que se limitaron a una sola división entrenamiento/validación/test o a validaciones internas poco claras.

6.4. RELEVANCIA CLÍNICA Y POTENCIAL TRASLACIONAL

Desde el punto de vista clínico, los resultados sugieren que un modelo como EfficientNetB7–Swin Transformer podría actuar como módulo de decisión complementario al UST en la estratificación de masas anexiales. La exactitud global

superior al 92%, junto con sensibilidades y especificidades en torno al 90–99% según la categoría, se sitúa al menos al nivel de los rendimientos descritos para ecografistas expertos y modelos basados en reglas como IOTA o O-RADS en estudios de validación bien diseñados [19, 30, 129]. Sin embargo, el valor añadido del enfoque propuesto no reside únicamente en la magnitud de las métricas, sino en la forma en que probabilidades calibradas, beneficio neto y cuantificación de incertidumbre pueden insertarse en decisiones clínicas reales.

Las curvas de calibración muestran que, tras regresión isotónica, las probabilidades del modelo se alinean estrechamente con tasas observadas de acierto, lo que permite interpretar una probabilidad de, por ejemplo, 0.80 como un riesgo clínico razonablemente fiable. Esta interpretabilidad es especialmente relevante cuando el sistema se concibe como apoyo a decisiones de triaje (p. ej., derivación a unidad especializada o intensificación del estudio), donde la salida debe funcionar como estimación de riesgo y no como etiqueta opaca.

En relación con la utilidad clínica, las CDC se interpretan de forma más directa cuando el problema se operacionaliza en términos de una decisión binaria (derivar/intervenir frente a manejo conservador/seguimiento), aun cuando el modelo subyacente sea multicategoría. En ese escenario, el beneficio neto estimado en un rango de umbrales de 5–20% sugiere que el sistema podría aportar valor por encima de estrategias extremas ("operar a todas" o "no operar a ninguna"), con el potencial de reducir intervenciones innecesarias sin pérdida sustantiva de capacidad para detectar enfermedad relevante [130, 131].

El análisis de entropía aporta además una guía explícita para automatización selectiva: las predicciones de baja entropía se asocian a tasas de error prácticamente nulas, mientras que los errores se concentran en los deciles de mayor incertidumbre. Este patrón sugiere un uso prudente del sistema como filtro o "segundo lector": automatizar con alta confianza una fracción sustancial de casos y derivar a revisión experta aquellos en los que el propio modelo indica ambigüedad. En la práctica, esta estrategia reduce el riesgo clínico asociado a errores en casos complejos, y convierte la incertidumbre en una señal operativa para controlar el uso del sistema en producción.

En términos de contexto internacional, trabajos como los de Gao et al. y Dai et al. han demostrado que modelos de IA basados en ecografía pueden insertarse en *pipelines* de cribado multicéntrico binario con resultados robustos, pero sin

abordar la complejidad multicategoría ni la calibración y la incertidumbre de forma tan explícita [36, 122]. La propuesta de esta tesis sugiere un siguiente paso lógico: trasladar estas buenas prácticas metodológicas a escenarios prospectivos y multicéntricos, empleando modelos híbridos que sean capaces de integrarse como "segundo lector".

6.5. AMENAZAS A LA VALIDEZ, LÍMITES DE GENERALIZACIÓN Y LECTURA CONSERVADORA DE LOS RESULTADOS

Aunque el rendimiento obtenido en OTU-2D sitúa al híbrido EfficientNetB7–Swin en el extremo alto de la literatura, estos resultados deben interpretarse con una lectura conservadora de su validez externa. El Estudio 2 se apoya en un único conjunto de datos público, unicéntrico y procedente de un entorno específico (equipos, protocolos de adquisición, selección de cortes y perfil poblacional). Esta dependencia de un dominio concreto implica que el rendimiento observado (en especial la especificidad elevada) puede degradarse ante cambios en la distribución de datos (centros con ecógrafos distintos, variaciones de ganancia/compresión, prevalencias diferentes o casuística más heterogénea). En términos prácticos, la aportación principal del Estudio 2 no es certificar un rendimiento universal, sino demostrar que, bajo condiciones controladas y reproducibles, la fusión temprana CNN–ViT permite capturar señales discriminativas útiles y que estas salidas pueden hacerse clínicamente más accionables mediante calibración, beneficio neto e incertidumbre.

La evidencia experimental se restringe a UST en modo B 2D estático. Esta decisión es metodológicamente defendible, pero delimita el alcance: el modelo aprende a partir de una representación condensada de la exploración y no incorpora información dinámica ni modos complementarios (Doppler, 3D/4D) que contribuyen a la valoración clínica. Por ello, la tesis establece una línea base fuerte sobre la que justificar extensiones, pero no permite inferir directamente el desempeño en escenarios multimodales o temporales.

En el Estudio 1, la síntesis se basa en datos agregados reportados por estudios heterogéneos, lo que limita el control de confusores. La metarregresión aporta evidencia asociativa valiosa, pero no establece causalidad estricta. Además, el riesgo de sesgo alto o incierto observado en una fracción relevante de estudios

sugiere que las estimaciones globales deben interpretarse con incertidumbre. De manera complementaria, en campos con tamaños muestrales pequeños y reportes incompletos no puede descartarse la presencia de efectos de estudios pequeños o sesgo de publicación, lo que refuerza una interpretación prudente del rendimiento agregado.

Finalmente, desde la perspectiva clínica, falta una comparación directa, sobre el mismo conjunto de imágenes, entre el modelo propuesto y expertos humanos con distintos niveles de experiencia. Esta comparación es relevante no solo por el contraste IA vs humano, sino porque permite estimar valor incremental como segundo lector, cuantificar variabilidad interobservador y definir políticas de uso (incluida automatización selectiva basada en incertidumbre) con base empírica.

6.6. HOJA DE RUTA TRASLACIONAL Y AGENDA DE INVESTIGACIÓN DERIVADA

A partir de los resultados y de las amenazas a la validez, la maduración lógica del sistema exige priorizar validación externa y cierre de brechas entre el régimen experimental y el flujo clínico real. Un primer paso es evaluar el modelo en una cohorte prospectiva independiente con confirmación histopatológica y anotación estandarizada, aprovechando el protocolo aprobado con el Instituto Estatal de Cancerología "Dr. Arturo Beltrán Ortega". Esta cohorte permitiría cuantificar degradación por cambio de dominio, recalibrar probabilidades si fuera necesario y verificar si el rango de utilidad observado mediante CDC se mantiene en un entorno real.

De forma complementaria, la robustez debe ponerse a prueba en escenarios multicéntricos que incorporen variabilidad real (equipos, protocolos, prevalencias y operadores). Este paso no solo fortalecería la validez externa, sino que facilitaría definir procedimientos de monitorización y criterios de recalibración o adaptación de dominio, componentes relevantes si se pretende un despliegue sostenido.

Con una base validada en 2D, la extensión hacia información temporal y multimodal (clips ecográficos, Doppler y eventualmente 3D/4D) representa el siguiente salto natural, especialmente para resolver zonas grises clínicas (p. ej., solapamientos morfológicos y casos *borderline*) y reducir falsos positivos sin comprometer sensibilidad. En paralelo, será necesario diseñar estudios prospectivos orientados a impacto, donde se compare la conducta clínica con y sin

apoyo del sistema, midiendo desenlaces como cirugías evitadas, tiempos de derivación, detección temprana, carga de trabajo y aceptabilidad. En este tipo de estudios, la incertidumbre puede operacionalizarse como política de automatización selectiva: automatizar solo cuando el modelo está calibrado y su incertidumbre es baja, y escalar el resto a revisión experta.

Finalmente, la adopción clínica requiere un marco de implementación responsable: definición del uso previsto, supervisión humana, trazabilidad y auditoría, comunicación comprensible de incertidumbre y alineación con requisitos regulatorios y éticos. En esta tesis, la calibración, el análisis de beneficio neto y la incertidumbre ya funcionan como pilares iniciales de esa implementación.

VII – CONCLUSIONES

VII - CONCLUSIONES

A continuación, se detalla de forma explícita cómo se alcanzó cada objetivo específico, indicando el estudio y los capítulos en los que se desarrolló y donde se evidencian los resultados.

7.1. CUMPLIMIENTO DE LOS OBJETIVOS ESPECÍFICOS

OE1. Realización de una revisión sistemática de modelos de IA aplicados al diagnóstico de tumores ováricos mediante UST.

Este objetivo se ha cubierto mediante el Estudio 1 (Cap. IV, 4.2; Cap. V, 5.1), donde se llevó a cabo una revisión sistemática siguiendo PRISMA 2020. Se identificaron y seleccionaron 44 estudios publicados hasta diciembre de 2024, se extrajeron de forma estandarizada sus métricas y características metodológicas, y se evaluó riesgo de sesgo y aplicabilidad con PROBAST. La síntesis cuantitativa incluyó comparaciones por tipo de segmentación, análisis de heterogeneidad, y metarregresión para determinar factores asociados al rendimiento (por ejemplo, el papel crítico de la segmentación automática), delimitando así brechas que condicionan la implementación clínica (validación externa, calibración, reporte y gobernanza).

OE2. Depurar la base de datos empleada.

Este objetivo se ha alcanzado en el Estudio 2 (Cap. IV, 4.3.2–4.3.4), donde se trabajó con el conjunto OTU-2D (1 469 imágenes UST en modo B en ocho categorías con confirmación histopatológica). Se definió un preprocesamiento homogéneo, y se aplicaron aumentos de datos y sobremuestreo de clases únicamente sobre el conjunto de entrenamiento. Finalmente, se estableció una estrategia de partición mediante validación cruzada de 5 pliegues repetida 10 veces con diferentes semillas.

OE3. Diseñar, entrenar y evaluar modelos basados exclusivamente en CNN y exclusivamente en ViT.

Este objetivo se ha cubierto en el Estudio 2 (Cap. IV, 4.3.5–4.3.7; Cap. V, 5.2.1), donde se entrenaron y evaluaron modelos de rama única con pesos preentrenados

(CNN: ResNet152, DenseNet201, EfficientNetB7; ViT: ViTB16 y Swin Transformer). Todos se ajustaron al mismo problema multiclase y bajo un protocolo común, generando así una línea base comparable para cuantificar la ganancia real aportada por la hibridación.

OE4. Proponer, implementar y optimizar modelos híbridos CNN–ViT con fusión temprana adaptativa, y cuantificar su desempeño multiclase.

Este objetivo se ha logrado en el Estudio 2 (Cap. IV, 4.3.6–4.3.7; Cap. V, 5.2.1), mediante la implementación de un bloque de fusión temprana aprendido que concatena la representación local extraída por la CNN y la representación global del ViT. Se evaluaron distintas combinaciones híbridas CNN–ViT bajo el mismo *pipeline*, demostrando que la fusión temprana incrementa de forma consistente el rendimiento frente a modelos de rama única, y seleccionando como configuración de mejor desempeño el híbrido EfficientNetB7–Swin Transformer.

OE5. Validar comparativamente el enfoque híbrido frente a las líneas base mediante pruebas estadísticas, análisis de errores y calibración probabilística.

Este objetivo se ha cubierto en el Estudio 2 (Cap. IV, 4.3.7; Cap. V, 5.2.1–5.2.6). La comparación entre líneas base y modelos híbridos se realizó con validación cruzada, métricas globales y por clase, e intervalos de confianza mediante *bootstrap*. Las diferencias se evaluaron con pruebas estadísticas pareadas y control de significación. Además, se incorporaron evaluaciones clave para traslación clínica: calibración probabilística, análisis de utilidad clínica mediante CDC y análisis de errores (matriz de confusión y patrones de confusión entre clases ecográficamente similares). Como extensión, se exploró un ensamble suave de los mejores híbridos, mostrando mejor rendimiento que el mejor modelo híbrido.

OE6. Incorporar métodos de interpretabilidad y analizar su correspondencia con regiones clínicamente relevantes definidas por expertos.

Este objetivo se ha alcanzado en el Estudio 2 (Cap. IV, 4.3.8; Cap. V, 5.2.7), implementando Grad-CAM para generar mapas de activación que permiten inspeccionar qué regiones de la imagen sustentan la predicción del modelo. La inspección cualitativa evidenció que la atención del modelo se concentra en regiones anatómicas compatibles con la localización tumoral. De forma complementaria, la cuantificación de incertidumbre (Cap. V, 5.2.6) proporciona un criterio operativo para automatización selectiva: predicciones de baja

incertidumbre pueden automatizarse y casos de alta incertidumbre pueden derivarse a revisión experta, reforzando seguridad y aceptación clínica.

7.2. CONCLUSIÓN INTEGRADORA

En conjunto, se demuestra un hilo conductor claro entre ambos estudios: el Estudio 1 define, a partir de evidencia agregada, las limitaciones metodológicas que bloquean la traslación clínica; el Estudio 2 responde a esas limitaciones con un protocolo experimental homogéneo y reproducible, y demuestra que una arquitectura híbrida CNN–ViT con fusión temprana puede mejorar la clasificación multicategoría en UST en modo B, aportando además probabilidades calibradas, estimación de incertidumbre e interpretabilidad visual. Por tanto, los seis objetivos específicos se consideran alcanzados y quedan respaldados por resultados cuantitativos y cualitativos reportados en los capítulos de IV y V.

VIII - LIMITACIONES Y FUTURAS LÍNEAS DE INVESTIGACIÓN

VIII - LIMITACIONES Y FUTURAS LÍNEAS DE INVESTIGACIÓN

8.1. LIMITACIONES

Las limitaciones de esta tesis están determinadas, principalmente, por el alcance de los datos empleados en el Estudio 2 y por las restricciones inherentes al Estudio 1 al apoyarse en evidencia publicada con datos agregados.

En el Estudio 2, la evidencia experimental se deriva exclusivamente de la base pública OTU-2D, de origen unicéntrico y construida a partir de cortes ecográficos 2D representativos seleccionados por especialistas. Esto implica que la distribución de clases, la casuística, los equipos y los parámetros de adquisición responden a un contexto particular; por tanto, la generalización del rendimiento a otros centros, poblaciones, ecógrafos y flujos de trabajo clínicos requiere validación externa. Además, OTU-2D se limita a UST en modo B estático, sin incorporar información dinámica (clips ecográficos) ni modalidades complementarias como Doppler o ecografía 3D/4D. En consecuencia, el modelo se entrena sobre una representación parcial del examen ecográfico y no es posible extrapolar directamente su desempeño a escenarios multimodales o temporales.

En el Estudio 1, la revisión sistemática se basa en resultados agregados reportados por los artículos. Esta circunstancia limita el control completo de la heterogeneidad entre estudios y restringe análisis a nivel paciente, como la estimación uniforme de métricas bajo criterios compartidos. Asimismo, la evaluación del riesgo de sesgo mediante PROBAST mostró una proporción relevante de estudios con riesgo alto o incierto, especialmente en el dominio de análisis, lo que introduce incertidumbre en las estimaciones globales y exige prudencia al interpretar el rendimiento promedio de la literatura.

Finalmente, desde el punto de vista clínico, esta tesis no incluye un estudio de lectura controlada que compare directamente, en OTU-2D, el desempeño del modelo híbrido frente a ecografistas (con distintos niveles de experiencia) bajo marcos clínicos como IOTA u O-RADS. Aunque el posicionamiento se discute con base en valores reportados, la ausencia de comparación empírica en condiciones idénticas limita la cuantificación del valor incremental del sistema como "segundo lector". De forma complementaria, si bien existe un protocolo aprobado con el

Instituto Estatal de Cancerología "Dr. Arturo Beltrán Ortega" para la construcción de una cohorte prospectiva local, dicha cohorte aún no se materializa en datos analizados dentro de esta tesis, lo que acota el peso de la evidencia traslacional directa en población local.

8.2. FUTURAS LÍNEAS DE INVESTIGACIÓN

Las líneas futuras se orientan a reforzar la validez externa, cerrar la brecha entre el entorno experimental y el flujo clínico real. Su priorización y secuenciación se desarrollan en la Sección 6.6.

El paso inmediato más relevante es la validación externa prospectiva del modelo, idealmente activando el protocolo del Instituto Estatal de Cancerología "Dr. Arturo Beltrán Ortega" para recolectar UST en modo B con confirmación histopatológica y anotaciones estandarizadas. Esta cohorte permitiría cuantificar la degradación por cambio de dominio, recalibrar probabilidades cuando sea necesario y verificar la estabilidad de la utilidad clínica sugerida por el análisis de beneficio neto.

En paralelo, la consolidación del sistema requiere validación multicéntrica, incorporando hospitales con equipos, protocolos y prevalencias diferentes. Esto permitiría caracterizar robustez ante variabilidad real, definir esquemas de monitorización y explorar estrategias de adaptación de dominio o reentrenamiento controlado.

Otra extensión natural es ampliar el alcance del modelo hacia información temporal y multimodal; la integración de clips ecográficos y modos complementarios (Doppler, 3D/4D) permitiría evaluar su contribución en subtipos complejos y en escenarios de solapamiento morfológico.

Finalmente, más allá de la validación técnica, será necesario evaluar impacto clínico real mediante estudios prospectivos con diseño de "segundo lector" y comparación directa con clínicos, midiendo desenlaces operativos y clínicos (cirugías evitadas, tiempos de derivación, detección temprana, carga de trabajo y aceptabilidad). De forma complementaria, la implementación responsable exige abordar requisitos de gobernanza: definición del uso previsto, supervisión humana, trazabilidad y auditoría, comunicación de la incertidumbre y alineación con marcos regulatorios y éticos aplicables.

IX - REFERENCIAS BIBLIOGRÁFICAS

IX - REFERENCIAS BIBLIOGRÁFICAS

[1] Lheureux S, Braunstein M, Oza AM. Epithelial ovarian cancer: Evolution of management in the era of precision medicine. CA Cancer J Clin 2019;69:280–304. https://doi.org/10.3322/caac.21559.

[2] Hong MK, Ding DC. Early Diagnosis of Ovarian Cancer: A Comprehensive Review of the Advances, Challenges, and Future Directions. Diagnostics 2025;15:406. https://doi.org/10.3390/diagnostics15040406.

[3] Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer Statistics, 2021. CA Cancer J Clin 2021;71:7–33. https://doi.org/10.3322/caac.21654.

[4] Dexter JM, Brubaker LW, Bitler BG, Goff BA, Menon U, Moore KN, et al. Ovarian cancer think tank: An overview of the current status of ovarian cancer screening and recommendations for future directions. Gynecol Oncol Rep 2024;53:101376. https://doi.org/10.1016/j.gore.2024.101376.

[5] Menon U, Karpinskyj C, Gentry-Maharaj A. Ovarian Cancer Prevention and Screening. Obstet Gynecol 2018;131:909–27. https://doi.org/10.1097/AOG.0000000000002580.

[6] Sideris M, Menon U, Manchanda R. Screening and prevention of ovarian cancer. Med J Aust 2024;220:264–74. https://doi.org/10.5694/mja2.52227.

[7] Ferlay J, Ervik M, Lam F, Laversanne M, Colombet M, Mery L. Global Cancer Observatory: Cancer Today. France: International Agency for Research on Cancer; 2024.

[8] Cortés Morera A, Ibáñez Morera M, Hernández Lara A, García Carranza MA, Cortés Morera A, Ibáñez Morera M, et al. Cáncer de Ovario: Tamizaje y diagnóstico imagenológico. Med Leg Costa Rica 2020;37:54–61.

[9] Christiansen F, Konuk E, Ganeshan AR, Welch R, Palés Huix J, Czekierdowski A, et al. International multicenter validation of AI-driven ultrasound detection of ovarian cancer. Nat Med 2025;31:189–96. https://doi.org/10.1038/s41591-024-03329-4.

[10] Gohagan JK, Prorok PC, Hayes RB, Kramer BS, Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial Project Team. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: history, organization, and status. Control Clin Trials 2000;21:251S-272S. https://doi.org/10.1016/s0197-2456(00)00097-0.

[11] Jacobs IJ. The UK Collaborative Trial Of Ovarian Cancer Screening (Ukctocs): Recruitment, Randomisation, Initial Screen. Int J Gynecol Cancer 2006;16:600–1. https://doi.org/10.1136/ijgc-00009577-200610001-00007.

[12] Bäumler M, Gallant D, Druckmann R, Kuhn W. Ultrasound screening of ovarian cancer. Horm Mol Biol Clin Investig 2019;41. https://doi.org/10.1515/hmbci-2019-0022.

[13] Nebgen DR, Lu KH, Bast RC. Novel Approaches to Ovarian Cancer Screening. Curr Oncol Rep 2019;21:75. https://doi.org/10.1007/s11912-019-0816-0.

[14] Dalmartello M, La Vecchia C, Bertuccio P, Boffetta P, Levi F, Negri E, et al. European cancer mortality predictions for the year 2022 with focus on ovarian cancer. Ann Oncol Off J Eur Soc Med Oncol 2022;33:330–9. https://doi.org/10.1016/j.annonc.2021.12.007.

[15] van Nagell JR, Hoff JT. Transvaginal ultrasonography in ovarian cancer screening: current perspectives. Int J Womens Health 2013;6:25–33. https://doi.org/10.2147/IJWH.S38347.

[16] Andreotti RF, Timmerman D, Strachowski LM, Froyman W, Benacerraf BR, Bennett GL, et al. O-RADS US Risk Stratification and Management System: A Consensus Guideline from the ACR Ovarian-Adnexal Reporting and Data System Committee. Radiology 2020;294:168–85. https://doi.org/10.1148/radiol.2019191150.

[17] Timmerman D, Van Calster B, Testa AC, Guerriero S, Fischerova D, Lissoni AA, et al. Ovarian cancer prediction in adnexal masses using ultrasound-based logistic regression models: a temporal and external validation study by the IOTA group. Ultrasound Obstet Gynecol Off J Int Soc Ultrasound Obstet Gynecol 2010;36:226–34. https://doi.org/10.1002/uog.7636.

[18] Testa A, Kaijser J, Wynants L, Fischerova D, Van Holsbeke C, Franchi D, et al. Strategies to diagnose ovarian cancer: new evidence from phase 3 of the multicentre international IOTA study. Br J Cancer 2014;111:680–8. https://doi.org/10.1038/bjc.2014.333.

[19] Timmerman D, Testa AC, Bourne T, Ameye L, Jurkovic D, Van Holsbeke C, et al. Simple ultrasound-based rules for the diagnosis of ovarian cancer. Ultrasound Obstet Gynecol Off J Int Soc Ultrasound Obstet Gynecol 2008;31:681–90. https://doi.org/10.1002/uog.5365.

[20] Van Calster B, Van Hoorde K, Valentin L, Testa AC, Fischerova D, Van Holsbeke C, et al. Evaluating the risk of ovarian cancer before surgery using the ADNEX model to differentiate between benign, borderline, early and advanced stage invasive, and secondary metastatic tumours: prospective multicentre diagnostic study. BMJ 2014;349:g5920. https://doi.org/10.1136/bmj.g5920.

[21] Yazbek J, Raju SK, Ben-Nagi J, Holland TK, Hillaby K, Jurkovic D. Effect of quality of gynaecological ultrasonography on management of patients with suspected ovarian cancer: a randomised controlled trial. Lancet Oncol 2008;9:124–31. https://doi.org/10.1016/S1470-2045(08)70005-6.

[22] Tavoraitė I, Kronlachner L, Opolskienė G, Bartkevičienė D. Ultrasound Assessment of Adnexal Pathology: Standardized Methods and Different Levels of Experience. Medicina (Mex) 2021;57:708. https://doi.org/10.3390/medicina57070708.

[23] Meys EMJ, Jeelof LS, Achten NMJ, Slangen BFM, Lambrechts S, Kruitwagen RFPM, et al. Estimating risk of malignancy in adnexal masses: external validation of the ADNEX model and comparison with other frequently used ultrasound methods. Ultrasound Obstet Gynecol Off J Int Soc Ultrasound Obstet Gynecol 2017;49:784–92. https://doi.org/10.1002/uog.17225.

[24] Dang Thi Minh N, Nguyen Van T, Duong Duc H, Nguyen Tuan M, Duong Thi Tra G, Do Tuan D, et al. IOTA simple rules: An efficient tool for

evaluation of ovarian tumors by non-experienced but trained examiners - A prospective study. Heliyon 2024;10:e24262. https://doi.org/10.1016/j.heliyon.2024.e24262.

[25] Gareeballah A, Gameraddin M, Alshoabi SA, Alsaedi A, Elzaki M, Alsharif W, et al. The diagnostic performance of International Ovarian Tumor Analysis: Simple Rules for diagnosing ovarian tumors—a systematic review and meta-analysis. Front Oncol 2025;14:1474930. https://doi.org/10.3389/fonc.2024.1474930.

[26] Almeida G, Bort M, Alcázar JL. Comparison of the Diagnostic Performance of Ovarian Adnexal Reporting Data System (O-RADS) With IOTA Simple Rules and ADNEX Model for Classifying Adnexal Masses: A Head-To-Head Meta-Analysis. J Clin Ultrasound 2025. https://doi.org/10.1002/jcu.24048.

[27] Tsili AC, Alexiou G, Tzoumpa M, Siempis T, Argyropoulou MI. Imaging of Peritoneal Metastases in Ovarian Cancer Using MDCT, MRI, and FDG PET/CT: A Systematic Review and Meta-Analysis. Cancers 2024;16:1467. https://doi.org/10.3390/cancers16081467.

[28] Strachowski LM, Jha P, Chawla TP, Davis KM, Dove CK, Glanc P, et al. O-RADS for Ultrasound: A User's Guide, From the AJR Special Series on Radiology Reporting and Data Systems. AJR Am J Roentgenol 2021;216:1150–65. https://doi.org/10.2214/AJR.20.25064.

[29] Vara J, Pagliuca M, Springer S, Gonzalez de Canales J, Brotons I, Yakcich J, et al. O-RADS Classification for Ultrasound Assessment of Adnexal Masses: Agreement between IOTA Lexicon and ADNEX Model for Assigning Risk Group. Diagnostics 2023;13:673. https://doi.org/10.3390/diagnostics13040673.

[30] Zhang Q, Dai X, Li W. Systematic Review and Meta-Analysis of O-RADS Ultrasound and O-RADS MRI for Risk Assessment of Ovarian and Adnexal Lesions. Am J Roentgenol 2023;221:1–13. https://doi.org/10.2214/AJR.22.28396.

[31] Mitchell S, Nikolopoulos M, El-Zarka A, Al-Karawi D, Al-Zaidi S, Ghai A, et al. Artificial Intelligence in Ultrasound Diagnoses of Ovarian Cancer: A Systematic Review and Meta-Analysis. Cancers 2024;16:422. https://doi.org/10.3390/cancers16020422.

[32] Xu HL, Gong TT, Liu FH, Chen HY, Xiao Q, Hou Y, et al. Artificial intelligence performance in image-based ovarian cancer identification: A systematic review and meta-analysis. eClinicalMedicine 2022;53. https://doi.org/10.1016/j.eclinm.2022.101662.

[33] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44–56. https://doi.org/10.1038/s41591-018-0300-7.

[34] Garcia-Atutxa I, Villanueva-Flores F, Barrio ED, Sanchez-Villamil JI, Martínez-Más J, Bueno-Crespo A. Artificial intelligence for ovarian cancer diagnosis via ultrasound: a systematic review and quantitative assessment of model performance. Front Artif Intell 2025;8:1649746. https://doi.org/10.3389/frai.2025.1649746.

[35] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2012;60:84–90. https://doi.org/10.1145/3065386.

[36] Gao Y, Zeng S, Xu X, Li H, Yao S, Song K, et al. Deep learning-enabled pelvic ultrasound images for accurate diagnosis of ovarian cancer in China: a retrospective, multicentre, diagnostic study. Lancet Digit Health 2022;4:e179–87. https://doi.org/10.1016/S2589-7500(21)00278-8.

[37] Hsu ST, Su YJ, Hung CH, Chen MJ, Lu CH, Kuo CE. Automatic ovarian tumors recognition system based on ensemble convolutional neural network with ultrasound imaging. BMC Med Inform Decis Mak 2022;22:298. https://doi.org/10.1186/s12911-022-02047-6.

[38] Christiansen F, Epstein EL, Smedberg E, Åkerlund M, Smith K, Epstein E. Ultrasound image analysis using deep neural networks for discriminating between benign and malignant ovarian tumors: comparison with expert subjective assessment. Ultrasound Obstet Gynecol 2021;57:155–63. https://doi.org/10.1002/uog.23530.

[39] Chen H, Yang BW, Qian L, Meng YS, Bai XH, Hong XW, et al. Deep Learning Prediction of Ovarian Malignancy at US Compared with O-RADS and Expert Assessment. Radiology 2022;304:106–13. https://doi.org/10.1148/radiol.211367.

[40] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale 2021. https://doi.org/10.48550/arXiv.2010.11929.

[41] Garcia-Atutxa I, Martínez-Más J, Bueno-Crespo A, Villanueva-Flores F. Early-fusion hybrid CNN-transformer models for multiclass ovarian tumor ultrasound classification. Front Artif Intell 2025;8. https://doi.org/10.3389/frai.2025.1679310.

[42] Collins GS, Dhiman P, Navarro CLA, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 2021;11:e048008. https://doi.org/10.1136/bmjopen-2020-048008.

[43] Zantvoort K, Nacke B, Görlich D, Hornstein S, Jacobi C, Funk B. Estimation of minimal data sets sizes for machine learning predictions in digital mental health interventions. Npj Digit Med 2024;7:361. https://doi.org/10.1038/s41746-024-01360-w.

[44] Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data 2019;6:27. https://doi.org/10.1186/s40537-019-0192-5.

[45] Hellín CJ, Olmedo AA, Valledor A, Gómez J, López-Benítez M, Tayebi A. Unraveling the Impact of Class Imbalance on Deep-Learning Models for Medical Image Classification. Appl Sci 2024;14:3419. https://doi.org/10.3390/app14083419.

[46] Suresh T, Brijet Z, Subha TD. Imbalanced medical disease dataset classification using enhanced generative adversarial network. Comput Methods Biomech Biomed Engin 2023;26:1702–18. https://doi.org/10.1080/10255842.2022.2134729.

[47] Salmi M, Atif D, Oliva D, Abraham A, Ventura S. Handling imbalanced medical datasets: review of a decade of research. Artif Intell Rev 2024;57:273. https://doi.org/10.1007/s10462-024-10884-2.

[48] Kang GG, So KA, Hwang JY, Kim NR, Yang EJ, Shim SH, et al. Ultrasonographic diagnosis and surgical outcomes of adnexal masses in children and adolescents. Sci Rep 2022;12:3949. https://doi.org/10.1038/s41598-022-08015-4.

[49] Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol 2016;69:245–7. https://doi.org/10.1016/j.jclinepi.2015.04.005.

[50] Geysels A, Garofalo G, Timmerman S, Barreñada L, De Moor B, Timmerman D, et al. Artificial Intelligence Applied to Ultrasound Diagnosis of Pelvic Gynecological Tumors: A Systematic Review and Meta-Analysis. Gynecol Obstet Invest 2025:1–22. https://doi.org/10.1159/000545850.

[51] Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future. Npj Digit Med 2022;5:48. https://doi.org/10.1038/s41746-022-00592-y.

[52] Lindhiem O, Petersen IT, Mentch LK, Youngstrom EA. The Importance of Calibration in Clinical Psychology. Assessment 2020;27:840–54. https://doi.org/10.1177/1073191117752055.

[53] Kurz A, Hauser K, Mehrtens HA, Krieghoff-Henning E, Hekler A, Kather JN, et al. Uncertainty Estimation in Medical Image Classification: Systematic Review. JMIR Med Inform 2022;10:e36427. https://doi.org/10.2196/36427.

[54] Faghani S, Moassefi M, Rouzrokh P, Khosravi B, Baffour FI, Ringler MD, et al. Quantifying Uncertainty in Deep Learning of Radiologic Images. Radiology 2023;308:e222217. https://doi.org/10.1148/radiol.222217.

[55] Abgrall G, Holder AL, Chelly Dagdia Z, Zeitouni K, Monnet X. Should AI models be explainable to clinicians? Crit Care Lond Engl 2024;28:301. https://doi.org/10.1186/s13054-024-05005-y.

[56] Hildt E. What Is the Role of Explainability in Medical Artificial Intelligence? A Case-Based Approach. Bioengineering 2025;12:375. https://doi.org/10.3390/bioengineering12040375.

[57] Koçak B, Köse F, Keleş A, Şendur A, Meşe İ, Karagülle M. Adherence to the Checklist for Artificial Intelligence in Medical Imaging (CLAIM): an umbrella review with a comprehensive two-level analysis. Diagn Interv Radiol 2025;31:440–55. https://doi.org/10.4274/dir.2025.243182.

[58] Lourenço M, Arrufat T, Satorres E, Maderuelo S, Novillo-Del Álamo B, Guerriero S, et al. Ultrasound-Based Deep Learning Models Performance versus Expert Subjective Assessment for Discriminating Adnexal Masses: A Head-to-Head Systematic Review and Meta-Analysis. Appl Sci 2024;14:2998. https://doi.org/10.3390/app14072998.

[59] Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Calster BV, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378.

[60] Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 2020;26:1351–63. https://doi.org/10.1038/s41591-020-1037-7.

[61] Mohajer-Bastami A, Moin S, Ahmad S, Ahmed AR, Pouwels S, Hajibandeh S, et al. Artificial intelligence in healthcare: applications, challenges, and future directions. A narrative review informed by international, multidisciplinary expertise. Front Digit Health 2025;7. https://doi.org/10.3389/fdgth.2025.1644041.

[62] Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clin Kidney J 2021;14:49–58. https://doi.org/10.1093/ckj/sfaa188.

[63] Zhang S, Cheng C, Lin Z, Xiao L, Su X, Zheng L, et al. The global burden and associated factors of ovarian cancer in 1990–2019: findings from the Global Burden of Disease Study 2019. BMC Public Health 2022;22:1455. https://doi.org/10.1186/s12889-022-13861-y.

[64] Tjokroprawiro BA, Novitasari K, Ulhaq RA, Sulistya HA, Martini S. Investigation of the trends and associated factors of ovarian cancer in Indonesia: A systematic analysis of the Global Burden of Disease study 1990–2021. PLOS ONE 2025;20:e0313418. https://doi.org/10.1371/journal.pone.0313418.

[65] Hutchinson B, Euripides M, Reid F, Allman G, Morrell L, Spencer G, et al. Socioeconomic Burden of Ovarian Cancer in 11 Countries. JCO Glob Oncol 2025:e2400313. https://doi.org/10.1200/GO-24-00313.

[66] Chin L, Hansen RN, Carlson JJ. Economic Burden of Metastatic Ovarian Cancer in a Commercially Insured Population: A Retrospective Cohort Analysis. J Manag Care Spec Pharm 2020;26:962–70. https://doi.org/10.18553/jmcp.2020.26.8.962.

[67] Adjei NN, Haas AM, Sun CC, Zhao H, Yeh PG, Giordano SH, et al. Cost of ovarian cancer by the phase of care in the United States. Am J Obstet Gynecol 2025;232:204.e1-204.e13. https://doi.org/10.1016/j.ajog.2024.08.023.

[68] Swiecki-Sikora AL, Craig AD, Chu CS. Financial toxicity in ovarian cancer. Int J Gynecol Cancer 2022;32:1450–4. https://doi.org/10.1136/ijgc-2022-003594.

[69] Liang MI, Chen L, Hershman DL, Hillyer GC, Huh WK, Guyton A, et al. Total and out-of-pocket costs for PARP inhibitors among insured ovarian cancer patients. Gynecol Oncol 2021;160:793–9. https://doi.org/10.1016/j.ygyno.2020.12.015.

[70] Bajwa A, Chidebe RCW, Adams T, Funston G, Soerjomataram I, Cohen R, et al. Challenges and opportunities in ovarian cancer care: A qualitative study of clinician perspectives from 24 low- and middle-income countries. J Cancer Policy 2025;44:100582. https://doi.org/10.1016/j.jcpo.2025.100582.

[71] Chen S, Cao Z, Prettner K, Kuhn M, Yang J, Jiao L, et al. Estimates and Projections of the Global Economic Cost of 29 Cancers in 204 Countries and Territories From 2020 to 2050. JAMA Oncol 2023;9:465–72. https://doi.org/10.1001/jamaoncol.2022.7826.

[72] Azangou-Khyavy M, Ghasemi E, Rezaei N, Khanali J, Kolahi AA, Malekpour MR, et al. Global, regional, and national quality of care index of cervical and ovarian cancer: a systematic analysis for the global burden of disease study 1990-2019. BMC Womens Health 2024;24:69. https://doi.org/10.1186/s12905-024-02884-9.

[73] Che M, Yin R. Analysis of the global burden of ovarian cancer in adolescents. Int J Gynecol Cancer 2025;35. https://doi.org/10.1016/j.ijgc.2024.101620.

[74] Washington CJ, Karanth SD, Wheeler M, Aduse-Poku L, Braithwaite D, Akinyemiju TF. Racial and socioeconomic disparities in survival among women with advanced-stage ovarian cancer who received systemic therapy. Cancer Causes Control 2024;35:487–96. https://doi.org/10.1007/s10552-023-01810-y.

[75] Jara-Rosales S, González-Stegmaier R, Rotarou ES, Villarroel-Espíndola F. Risk Factors for Ovarian Cancer in South America: A Literature Review. J Pers Med 2024;14:992. https://doi.org/10.3390/jpm14090992.

[76] Aggarwal A. The challenge of cancer in middle-income countries with an ageing population: Mexico as a case study. Ecancermedicalscience 2015;9. https://doi.org/10.3332/ecancer.2015.536.

[77] Westwood M, Ramaekers B, Lang S, Grimm S, Deshpande S, de Kock S, et al. Risk scores to guide referral decisions for people with suspected ovarian cancer in secondary care: a systematic review and cost-effectiveness analysis. Health Technol Assess Winch Engl 2018;22:1–264. https://doi.org/10.3310/hta22440.

[78] Dietrich CF, Bolondi L, Duck F, Evans DH, Ewertsen C, Fraser AG, et al. History of Ultrasound in Medicine from its birth to date (2022), on occasion of the 50 Years Anniversary of EFSUMB. A publication of the European Federation of Societies for Ultrasound In Medicine and Biology (EFSUMB), designed to record the historical development of medical ultrasound. Med Ultrason 2022;24:434. https://doi.org/10.11152/mu-3757.

[79] Campbell S. A short history of sonography in obstetrics and gynaecology. Facts Views Vis ObGyn 2013;5:213–29.

[80] Levi S. The history of ultrasound in gynecology 1950–1980. Ultrasound Med Biol 1997;23:481–552. https://doi.org/10.1016/S0301-5629(96)00196-2.

[81] Sayasneh A, Ekechi C, Ferrara L, Kaijser J, Stalder C, Sur S, et al. The characteristic ultrasound features of specific types of ovarian pathology (Review). Int J Oncol 2014;46:445–58. https://doi.org/10.3892/ijo.2014.2764.

[82] Fischerova D, Cibula D. Ultrasound in gynecological cancer: is it time for re-evaluation of its uses? Curr Oncol Rep 2015;17:28. https://doi.org/10.1007/s11912-015-0449-x.

[83] Timmerman D, Calster BV, Testa A, Savelli L, Fischerova D, Froyman W, et al. Predicting the risk of malignancy in adnexal masses based on the Simple Rules from the International Ovarian Tumor Analysis group. Am J Obstet Gynecol 2016;214:424–37. https://doi.org/10.1016/j.ajog.2016.01.007.

[84] Carballo EV, Maturen KE, Li Z, Patel-Lippmann KK, Wasnik AP, Sadowski EA, et al. Surgical outcomes of adnexal masses classified by IOTA ultrasound

simple rules. Sci Rep 2022;12:21848. https://doi.org/10.1038/s41598-022-26441-2.

[85] Chankrachang A, Lattiwongsakorn W, Tantipalakorn C, Tongsong T. Diagnostic Performance of ADNEX Model and IOTA Simple Rules in Differentiating Malignant from Benign Adnexal Masses When Assessed by Non-Expert Examiners. J Clin Med 2025;14:2776. https://doi.org/10.3390/jcm14082776.

[86] Spagnol G, Marchetti M, Tommasi OD, Vitagliano A, Cavallin F, Tozzi R, et al. Simple rules, O-RADS, ADNEX and SRR model: Single oncologic center validation of diagnostic predictive models alone and combined (two-step strategy) to estimate the risk of malignancy in adnexal masses and ovarian tumors. Gynecol Oncol 2023;177:109–16. https://doi.org/10.1016/j.ygyno.2023.08.012.

[87] Masselli G, Bourgioti C. Review of the Imaging Modalities in the Gynecological Neoplasms During Pregnancy. Cancers 2025;17:838. https://doi.org/10.3390/cancers17050838.

[88] Fischerova D, Smet C, Scovazzi U, Sousa DN, Hundarova K, Haldorsen IS. Staging by imaging in gynecologic cancer and the role of ultrasound: an update of European joint consensus statements. Int J Gynecol Cancer 2024;34:363. https://doi.org/10.1136/ijgc-2023-004609.

[89] van Kolfschooten H, van Oirschot J. The EU Artificial Intelligence Act (2024): Implications for healthcare. Health Policy Amst Neth 2024;149:105152. https://doi.org/10.1016/j.healthpol.2024.105152.

[90] Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell 2019;1:389–99. https://doi.org/10.1038/s42256-019-0088-2.

[91] Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: A call for open science. Patterns N Y N 2021;2:100347. https://doi.org/10.1016/j.patter.2021.100347.

[92] Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. Artif Intell Healthc 2020:295–336. https://doi.org/10.1016/B978-0-12-818438-7.00012-5.

[93] Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020;20:310. https://doi.org/10.1186/s12911-020-01332-6.

[94] Beauchamp TL, Childress JF. Principles of biomedical ethics. Eighth edition. New York Oxford: Oxford University Press; 2019.

[95] Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds Mach 2018;28:689–707. https://doi.org/10.1007/s11023-018-9482-5.

[96] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine Learning. ACM Comput Surv 2021;54:1–35. https://doi.org/10.1145/3457607.

[97] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based

Localization. Int J Comput Vis 2020;128:336–59. https://doi.org/10.1007/s11263-019-01228-7.

[98] Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR, editors. Explainable AI: interpreting, explaining and visualizing deep learning. Cham, Switzerland: Springer; 2019.

[99] Wachter S, Mittelstadt B, Floridi L. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. Int Data Priv Law 2017;7:76–99. https://doi.org/10.1093/idpl/ipx005.

[100] Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019;17:195. https://doi.org/10.1186/s12916-019-1426-2.

[101] Shung DL, Sung JJY. Challenges of developing artificial intelligence-assisted tools for clinical medicine. J Gastroenterol Hepatol 2021;36:295–8. https://doi.org/10.1111/jgh.15378.

[102] Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell 2021;3:199–217. https://doi.org/10.1038/s42256-021-00307-0.

[103] Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020;368:m689. https://doi.org/10.1136/bmj.m689.

[104] Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. J Glob Health 2019;9:020318. https://doi.org/10.7189/jogh.09.020318.

[105] Hassan M, Kushniruk A, Borycki E. Barriers to and Facilitators of Artificial Intelligence Adoption in Health Care: Scoping Review. JMIR Hum Factors 2024;11:e48633. https://doi.org/10.2196/48633.

[106] Schwabe D, Becker K, Seyferth M, Klaß A, Schaeffter T. The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. NPJ Digit Med 2024;7:203. https://doi.org/10.1038/s41746-024-01196-4.

[107] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005.

[108] Reid F, Bajwa A. The world ovarian cancer coalition atlas 2023. World Ovarian Cancer Coalition; 2023.

[109] Xu D, He X, Zhang R, Zhang Y, Li M, Zhan Y. Hybrid CNN–Transformer with Fusion Discriminator for Ovarian Tumor Ultrasound Imaging Classification. Electronics 2025;14:4040. https://doi.org/10.3390/electronics14204040.

[110] Bucur C, Balescu I, Petrea S, Gaspar B, Pop L, Varlas V, et al. Artificial Intelligence in Ovarian Cancers—From Diagnosis to Treatment; A Literature Review. J Mind Med Sci 2024;11:277–84. https://doi.org/10.22543/2392-7674.1531.

[111] Bukłaho PA, Kiśluk J, Nikliński J. Diagnostics and treatment of ovarian cancer in the era of precision medicine - opportunities and challenges. Front Oncol 2023;13:1227657. https://doi.org/10.3389/fonc.2023.1227657.

[112] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. https://doi.org/10.1136/bmj.n71.

[113] Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med 2019;170:51–8. https://doi.org/10.7326/M18-1376.

[114] Zhao Q, Lyu S, Bai W, Cai L, Liu B, Cheng G, et al. MMOTU: A Multi-Modality Ovarian Tumor Ultrasound Image Dataset for Unsupervised Cross-Domain Semantic Segmentation 2023. https://doi.org/10.48550/arXiv.2207.06799.

[115] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition 2015. https://doi.org/10.48550/arXiv.1512.03385.

[116] Huang G, Liu Z, Maaten L van der, Weinberger KQ. Densely Connected Convolutional Networks 2018. https://doi.org/10.48550/arXiv.1608.06993.

[117] Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks 2020. https://doi.org/10.48550/arXiv.1905.11946.

[118] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows 2021. https://doi.org/10.48550/arXiv.2103.14030.

[119] Huang CY, Wang HI, Wang PH, Wu YC, Yang MJ, Chen LH, et al. Mean grey value is lower in endometriomas: Differentiating a hypoechogenic adnexal cyst by 3-dimensional power Doppler ultrasound—A preliminary study. J Chin Med Assoc 2011;74:75–80. https://doi.org/10.1016/j.jcma.2011.01.015.

[120] Lupean RA, Ștefan PA, Csutak C, Lebovici A, Măluțan AM, Buiga R, et al. Differentiation of Endometriomas from Ovarian Hemorrhagic Cysts at Magnetic Resonance: The Role of Texture Analysis. Medicina (Mex) 2020;56:487. https://doi.org/10.3390/medicina56100487.

[121] Bašković M, Habek D, Zaninović L, Milas I, Pogorelić Z. The Evaluation, Diagnosis, and Management of Ovarian Cysts, Masses, and Their Complications in Fetuses, Infants, Children, and Adolescents. Healthcare 2025;13:775. https://doi.org/10.3390/healthcare13070775.

[122] Dai WL, Wu YN, Ling YT, Zhao J, Zhang S, Gu ZW, et al. Development and validation of a deep learning pipeline to diagnose ovarian masses using ultrasound screening: a retrospective multicenter study. EClinicalMedicine 2024;78:102923. https://doi.org/10.1016/j.eclinm.2024.102923.

[123] Wu C, Wang Y, Wang F. Deep Learning for Ovarian Tumor Classification with Ultrasound Images. In: Hong R, Cheng W-H, Yamasaki T, Wang M, Ngo C-W, editors. Adv. Multimed. Inf. Process. – PCM 2018, Cham: Springer International Publishing; 2018, p. 395–406. https://doi.org/10.1007/978-3-030-00764-5_36.

[124] Giourga M, Petropoulos I, Stavros S, Potiris A, Gerede A, Sapantzoglou I, et al. Enhancing Ovarian Tumor Diagnosis: Performance of Convolutional Neural Networks in Classifying Ovarian Masses Using Ultrasound Images. J Clin Med 2024;13:4123. https://doi.org/10.3390/jcm13144123.

[125] Xie W, Lin W, Li P, Lai H, Wang Z, Liu P, et al. Developing a deep learning model for predicting ovarian cancer in Ovarian-Adnexal Reporting and Data System Ultrasound (O-RADS US) Category 4 lesions: A multicenter study. J Cancer Res Clin Oncol 2024;150:346. https://doi.org/10.1007/s00432-024-05872-6.

[126] Du Y, Guo W, Xiao Y, Chen H, Yao J, Wu J. Ultrasound-based deep learning radiomics model for differentiating benign, borderline, and malignant ovarian tumours: a multi-class classification exploratory study. BMC Med Imaging 2024;24:89. https://doi.org/10.1186/s12880-024-01251-2.

[127] Du Y, Xiao Y, Guo W, Yao J, Lan T, Li S, et al. Development and validation of an ultrasound-based deep learning radiomics nomogram for predicting the malignant risk of ovarian tumours. Biomed Eng Online 2024;23:41. https://doi.org/10.1186/s12938-024-01234-y.

[128] Barcroft JF, Linton-Reid K, Landolfo C, Al-Memar M, Parker N, Kyriacou C, et al. Machine learning and radiomics for segmentation and classification of adnexal masses on ultrasound. Npj Precis Oncol 2024;8:41. https://doi.org/10.1038/s41698-024-00527-8.

[129] Van Calster B, Valentin L, Froyman W, Landolfo C, Ceusters J, Testa AC, et al. Validation of models to diagnose ovarian cancer in patients managed surgically or conservatively: multicentre cohort study. BMJ 2020;370:m2614. https://doi.org/10.1136/bmj.m2614.

[130] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Mak Int J Soc Med Decis Mak 2006;26:565–74. https://doi.org/10.1177/0272989X06295361.

[131] Leibig C, Allken V, Ayhan MS, Berens P, Wahl S. Leveraging uncertainty information from deep neural networks for disease detection. Sci Rep 2017;7:17816. https://doi.org/10.1038/s41598-017-17876-z.

X – ANEXOS

X - ANEXOS

10.1. ANEXO I. PRODUCTIVIDAD DERIVADA DE LA TESIS DOCTORAL

10.1.1. Publicaciones científicas relacionadas con la tesis

Los resultados derivados de esta tesis doctoral se han publicado en dos artículos científicos en la revista *Frontiers in Artificial Intelligence* (Figura 16):

- **Garcia-Atutxa I**, Villanueva-Flores F, Barrio ED, Sanchez-Villamil JI, Martínez-Más J, Bueno-Crespo A. Artificial intelligence for ovarian cancer diagnosis via ultrasound: a systematic review and quantitative assessment of model performance. Front Artif Intell. 2025;8:1649746. https://doi.org/10.3389/frai.2025.1649746.
- **Garcia-Atutxa I**, Martínez-Más J, Bueno-Crespo A, Villanueva-Flores F. Early-fusion hybrid CNN-transformer models for multiclass ovarian tumor ultrasound classification. Front Artif Intell. 2025;8:1679310. https://doi.org/10.3389/frai.2025.1679310.

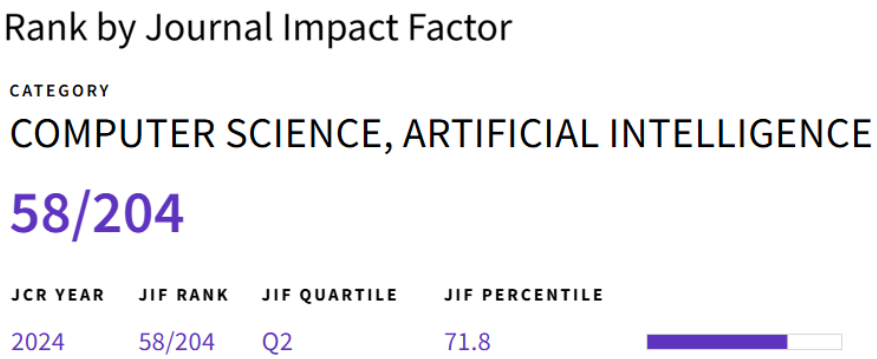

Figura 16. Clasificación en Journal Citation Reports 2024 (JCR 2024) por Journal Impact Factor (JIF) dentro de la categoría Computer Science, Artificial Intelligence (cuartil Q2, percentil 71.8). Fuente: Web of Science.

10.1.2. Participación en congresos relacionada con la tesis

- **Garcia-Atutxa I**, Villanueva-Flores F, Martínez-Más J, Bueno-Crespo A. Detección y clasificación de tumores ováricos en imágenes de ultrasonido

utilizando Inteligencia Artificial. Artificial Intelligence & eHealth, Murcia, 2025. Tipo: oral.

10.1.3. Estancias, proyectos o colaboraciones vinculadas a la tesis

- Estancia internacional en México: Centro de Investigación en Ciencia Aplicada y Tecnología Avanzada (CICATA) del Instituto Politécnico Nacional (IPN), del 1 de octubre de 2024 al 31 de enero de 2025.

10.1.4. Producción científica adicional

Además de la producción científica directamente vinculada a la línea de investigación de esta tesis, durante el periodo doctoral se generó producción adicional que no está relacionada con el objeto de estudio y, por ello, no se desarrolla en este documento. De forma resumida, dicha producción incluye:

- 17 artículos publicados en revistas científicas.
- 4 artículos de divulgación científica.
- 1 congreso internacional.

Como referencia cuantitativa del desempeño académico durante el periodo doctoral, la Figura 17 presenta el perfil bibliométrico en Scopus a fecha 31 de diciembre de 2025, donde se observa un total de 14 documentos indexados (2023–2025), 80 citas acumuladas y un h-index de 4, con una tendencia creciente tanto en producción anual como en citas recibidas.

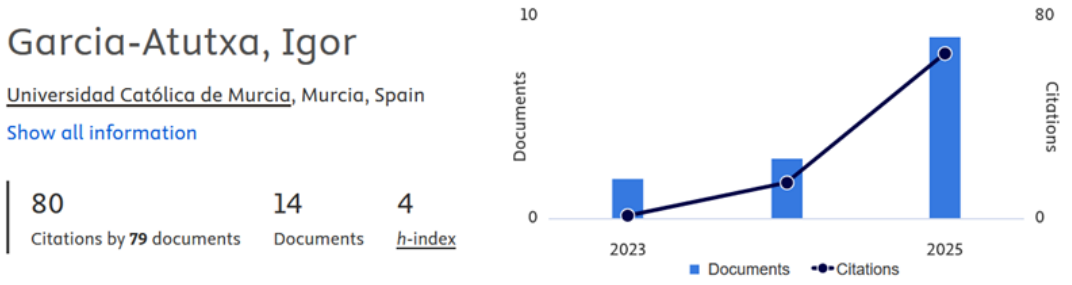

Figura 17. Perfil bibliométrico en Scopus: documentos publicados por año (barras, eje izquierdo) y citas recibidas (línea, eje derecho) durante el periodo doctoral 2023–2025. Fuente: Scopus.

La producción científica completa puede consultarse en el siguiente ORCID: https://orcid.org/0009-0002-1551-2685.

10.1.5. Premios

1. Premio al Proyecto de mayor Impacto Social por la Fundación Ramón Molinas - SpinUOC (2023).
2. Research Competition Award otorgado por The Antibody Society's (2024).
3. Beca Santander Personal Investigador Predoctoral (2025).

10.2. ANEXO II. APROBACIÓN COMITÉ DE ÉTICA EN INVESTIGACIÓN

INSTITUTO ESTATAL DE CANCEROLOGÍA
"DR. ARTURO BELTRÁN ORTEGA"
SUBDIRECCION DE ENSEÑANZA E INVESTIGACION

DICTAMEN DEL COMITÉ DE ÉTICA EN INVESTIGACIÓN

REGISTRO NÚMERO: CONBIOÉTICA-12-CEI-001-20190726

Acapulco Guerrero, a 02 de julio de 2025

NUM. OFICIO: DGIECAN/CCHGC/CEI-088-2025

DRA. AZUCENA OCAMPO BÁRCENAS
PRESENTE:

Título del proyecto:

APLIACIÓN DE LA INTELIGENCIA ARTIFICIAL PARA LA DETECCIÓN TEMPRANA Y CLASIFICACIÓN AUTOMATIZADA DE TUMORES SÓLIDOS EN IMAGENOLOGÍA MÉDICA

Código asignado por el CEI: **DGIECAN/CCHGC/CEI/PRO-054-2025**

Le informamos que su proyecto de referencia ha sido evaluado por el **Comité de Ética en Investigación** y las opiniones acerca de los documentos presentados se encuentran a continuación:

	Nº y/o Fecha Versión	Decisión
SOLICITUD DE EVALUACIÓN	02-07-2025	Aprobado
PROTOCOLO	02-07-2025	Aprobado
HOJA DE PARTICIPANTES EN LA INVESTIGACIÓN	02-07-2025	Aprobado
CARTA DE CONFIDENCIALIDAD	02-07-2025	Aprobado
CONSENTIMIENTO INFORMADO	02-07-2025	Aprobado

Este protocolo queda **Aprobado**, teniendo una vigencia de aprobación del 02 de

En caso de requerir una ampliación, se le solicitará un reporte del progreso de su proyecto con al menos 30 días antes de la fecha de término de su vigencia. Lo anterior forma parte de las obligaciones del investigador las cuales son descritas en la siguiente hoja.

Av. Ruiz Cortines 128-A Col. Alta Progreso, Acapulco, Gro. CP 39570. Tel. (744) 445-66-13

TRANSFORMANDO GUERRERO GOBIERNO DEL ESTADO

INSTITUTO ESTATAL DE CANCEROLOGÍA
"DR. ARTURO BELTRÁN ORTEGA"
SUBDIRECCION DE ENSEÑANZA E INVESTIGACION

LINEAMIENTOS QUE ESTABLECEN LAS OBLIGACIONES DE LOS INVESTIGADORES RESPONSABLES DE PROYECTOS DE INVESTIGACIÓN

1. Contar con la versión actualizada del protocolo, carta de consentimiento, hoja de participantes en la investigación y cualquier otro documento que se haya presentado a revisión y hubiese sido aprobado por el CEI.

   a. Protocolos pendientes de aprobación, resolver observaciones y no exceder los 30 días naturales a partir de la fecha de recepción.

2. Cualquier enmienda o modificación debe ser aprobada por el CEI, de no ser así, el CEI se reserva el derecho de la toma de decisiones en torno a violación o desviación del protocolo (suspensión, cancelación, etc.).

3. El investigador deberá reportar de manera semestral (junio y diciembre) el avance de su proyecto, el cual hará llegar al CEI, describiendo sus avances.

a. Los protocolos que no cuenten con reporte de avance de proyecto, el comité emitirá por escrito marcando una copia a la Dirección General y a la Subdirección de enseñanza e investigación, un aviso de suspensión del protocolo.

b. En caso de no hacer entrega de su informe de avance de proyecto, éste le será cancelado y el investigador no podrá someter a revisión protocolos de investigación por 6 meses contados a partir de la fecha de cancelación del proyecto.

4. Al terminar el proyecto enviar al CEI un reporte final del estudio describiendo los resultados del proyecto.

5. En el caso de protocolos financiados por la Industria Farmacéutica, el investigador responsable notificará a la Secretaría de Salud de la cancelación o suspensión del protocolo de investigación.

Av. Ruiz Cortines 128-A Col. Alta Progreso, Acapulco, Gro. CP 39570. Tel. (744) 445-66-13

TRANSFORMANDO GUERRERO GOBIERNO DEL ESTADO 2021 - 2027

INSTITUTO ESTATAL DE CANCEROLOGÍA
"DR. ARTURO BELTRÁN ORTEGA"
SUBDIRECCION DE ENSEÑANZA E INVESTIGACION

ATENTAMENTE

Nombre	Firma
Dra. Eloísa Ibarra Sierra Presidente del CEI	
Dra. Azucena Ocampo Bárcenas Vocal secretario del CEI	
Dr. Salomón Reyes Navarrete Vocal del CEI	
Dr. Marco Antonio Jiménez López Vocal del CEI	
C. María Elena Serna Hernández Representante ciudadano	