

INTELIGENCIA ARTIFICIAL PARA LA CARACTERIZACIÓN DE MASAS MAMARIAS BENIGNAS Y MALIGNAS MEDIANTE MAMOGRAFÍA Y ECOGRAFÍA

ARTIFICIAL INTELLIGENCE FOR THE CHARACTERIZATION OF BENIGN AND MALIGNANT BREAST MASSES USING MAMMOGRAPHY AND ULTRASOUND

Juliana del Rocio Carrasco Lascano¹, Elida Maria Hidalgo Gualan², Raúl Eduardo Chávez Robalino³, Abraham Oliver Balderas⁴

{rocio19914.4@gmail.com¹, correo elyta66@yahoo.es², rolyraul22@gmail.com³, droliiver.radiologiacyc@gmail.com⁴}

Fecha de recepción: 13/09/2026 / Fecha de aceptación: 08/10/2026 / Fecha de publicación: 14/10/2026

RESUMEN: El cáncer de mama es la neoplasia maligna más frecuente en mujeres a nivel mundial, con aproximadamente 2,3 millones de casos nuevos anuales. La capacidad para diferenciar con precisión entre las masas benignas y las malignas en la mamografía y la ecografía es una cuestión de diagnóstico clínico que presenta una considerable variabilidad entre observadores. En este artículo se analizan los estudios publicados entre 2023 y 2026 que evalúan el rendimiento de los sistemas de inteligencia artificial (IA), con su complejo de «aprendizaje profundo», para ayudar a diferenciar entre masa benigna y masa maligna. Se realizó una búsqueda estructurada en PubMed/MEDLINE, Scopus, Web of Science, ScienceDirect e IEEE Xplore. La evidencia disponible indica que los sistemas de deep learning alcanzan sensibilidades combinadas cercanas a 0,92–0,93 y especificidades de 0,88–0,91 en ecografía mamaria, mientras que en mamografía los valores de AUC en estudios de lectura se sitúan entre 0,87 y 0,90. Las técnicas multimodales que combinan información procedente de diferentes fuentes parecen tener mejor capacidad discriminativa en términos técnicos, aunque esto depende por completo del diseño del estudio y del grupo de pacientes estudiado. La mayoría de los estudios son retrospectivos, de un único centro y carecen de una validación externa prospectiva. Por lo tanto, parece que la inteligencia artificial podría resultar útil como apoyo en la práctica radiológica, pero aún se necesitan estudios prospectivos y multicéntricos para demostrar su precisión, así como un análisis detallado de sus implicaciones éticas y legales.

¹Hospital General Latacunga, Ecuador, Pelileo, <https://orcid.org/0000-0003-3635-8955>.

²Universidad Central del Ecuador, Ecuador, Quito, <https://orcid.org/0000-0002-7978-6864>

³Universidad Central del Ecuador, Ecuador, Quito, <https://orcid.org/0009-0006-2075-5699>

⁴IMSS, Star médica, México, Mérida, Yucatán, <https://0000-0001-9312-2346>

Palabras clave: *Inteligencia artificial; Mamografía; Ecografía mamaria; Neoplasias de la mama; Diagnóstico asistido por computadora*

ABSTRACT: Breast cancer is the most frequent malignant neoplasm in women worldwide, with approximately 2.3 million new cases annually. The ability to accurately differentiate between benign and malignant masses on mammography and ultrasound is a clinical diagnostic challenge that shows considerable interobserver variability. This article reviews studies published between 2023 and 2026 that evaluate the performance of artificial intelligence (AI) systems, particularly those based on deep learning, to assist in distinguishing between benign and malignant masses. A structured search was conducted in PubMed/MEDLINE, Scopus, Web of Science, ScienceDirect, and IEEE Xplore. Available evidence indicates that deep-learning systems achieve combined sensitivities of approximately 0.92–0.93 and specificities of 0.88–0.91 in breast ultrasound, while in mammography the AUC values in reader studies range from 0.87 to 0.90. Multimodal techniques that combine information from different sources appear to have better technical discriminatory capacity, although this depends entirely on study design and the patient population examined. Most studies are retrospective, single-center, and lack prospective external validation. Therefore, artificial intelligence may prove useful as a support tool in radiological practice, but prospective multicenter studies are still needed to demonstrate its accuracy, along with a detailed analysis of its ethical and legal implications.

Keywords: *Artificial intelligence; Mammography; Breast ultrasound; Breast neoplasms; Computer-assisted diagnosis*

INTRODUCCIÓN

El cáncer de mama es la neoplasia maligna más frecuente en mujeres, con 2,3 millones de nuevos casos diagnosticados en todo el mundo en 2020 (1). Es una de las principales causas de mortalidad en la población femenina, lo que hace que el desarrollo de métodos de evaluación y detección de lesiones mamarias sea una prioridad absoluta.

La mamografía digital y la tomografía digital de mama (DBT) se consideran los métodos «gold standard» para el cribado mamario; sin embargo, tienen una sensibilidad limitada cuando las mamas presentan tejido denso. Dado que esta es la principal limitación de la mamografía convencional, la ecografía se ha sugerido como complemento de la mamografía para detectar y caracterizar lesiones, así como para evaluar aquellas detectadas mediante mamografía. La interpretación subjetiva de los exámenes mamarios ecográficos es un importante «bucle de retroalimentación» de la ecografía, que se practica habitualmente en todo el mundo, lo que plantea la necesidad de encontrar una solución innovadora y rentable.

Todas las soluciones anteriores requieren que el usuario identifique el tipo de tarea: detección, caracterización, triaje o cribado. Las tareas pueden diferenciarse en función de que la imagen de la mama se utilice para detectar una lesión (detección), caracterizar si una lesión detectada es benigna o maligna (caracterización), seleccionar la lesión de mayor probabilidad para la biopsia (triaje) o determinar si una mujer presenta una neoplasia mamaria (cribado). Este artículo aborda

únicamente la segunda tarea, el problema de clasificación de distinguir entre masa benigna y masa maligna.

Las primeras soluciones informáticas se basaban generalmente en métodos de gran escala, en los que los autores extrajeron ellos mismos las características radiómicas pertinentes para detectar un tumor mamario maligno y construyeron modelos de CAD. Sin embargo, estos métodos no lograron hacer frente al «dilema de los falsos positivos», lo que dio lugar a elevados costes de evaluación. En comparación con la ingeniería de características, el enfoque de la CNN (red neuronal convolucional) obtiene características de las imágenes de forma automática, sencilla y eficiente.

Metaanálisis recientes han reportado valores de área bajo la curva (AUC) de aproximadamente 0,87 para sistemas de IA autónoma en estudios de lectura de mamografía, comparables o superiores a los de radiólogos, así como sensibilidades combinadas cercanas a 0,93 y especificidades de 0,90 en ecografía mamaria (2,3). Existe una serie de cuestiones en este sentido, entre ellas, en primer lugar, la variedad de estudios, los diseños, la prevalencia de estudios retrospectivos unicéntricos, la falta, en general, de estudios de validación externa prospectiva con resultados clínicamente relevantes y la aplicabilidad de un algoritmo a una población diferente.

La pregunta de investigación es la siguiente: ¿cuál es el rendimiento de los sistemas de información artificial en la caracterización de lesiones mamarias benignas y malignas en mamografías y ecografías, según la evidencia publicada entre 2023 y 2026. Se trata de una investigación destinada a comprender y explicar el valor diagnóstico de los algoritmos de inteligencia artificial para la detección de lesiones mamarias en imágenes de ecografía y mamografía. El objetivo es tener una idea clara del valor de dichos algoritmos para distinguir entre lesiones benignas y malignas. Por lo tanto, se plantean los siguientes objetivos: determinar qué algoritmos se utilizan, en resumen, comparar su rendimiento en términos de precisión, especificidad y precisión, y comprender qué limitaciones tienen, así como cómo pueden aplicarse en la práctica.

MATERIALES Y MÉTODOS

Se llevó a cabo una revisión narrativa estructurada con síntesis temática de la bibliografía científica. Este tipo de estudio permite la incorporación de información procedente de diversas fuentes, ya que integra un conjunto de literatura heterogéneo que abarca diferentes modalidades de imágenes y algoritmos de inteligencia artificial (IA). La revisión siguió las recomendaciones generales de la escala SANRA para revisiones narrativas, que no requieren un proceso sistemático de detección ni el registro en la base de datos de registros de revisiones, como PROSPERO (6).

La búsqueda se realizó en las bases de datos PubMed/MEDLINE, Scopus, Web of Science, ScienceDirect y IEEE Xplore, y se limitó al período comprendido entre enero de 2023 y marzo de 2026. La búsqueda se realizó utilizando conceptos relacionados con la inteligencia artificial (inteligencia artificial, aprendizaje profundo, aprendizaje automático, redes neuronales convolucionales, diagnóstico asistido por computadora) y las modalidades de imagen

(mamografía, mamografía digital, ultrasonido mamario, tomosíntesis mamaria, imágenes mamarias por ultrasonido), junto con la entidad principal (neoplasmas mamarios, masa mamaria, benigno, maligno, BI-RADS). El operador lógico «Y/OR» se aplicó en los diferentes aspectos de la búsqueda. Para acceder a la lista completa de términos de búsqueda, consulte los autores.

Se incluyeron estudios que evalúan sistemas de IA en mamografía y/o ultrasonido mamario para la caracterización de masas (adjudicación de riesgo), publicados entre enero de 2023 y marzo de 2026, y que informen de al menos una medida de rendimiento (sensibilidad, especificidad, área bajo la curva (AUC) o precisión), junto con diagnósticos de referencia (histología o seguimiento clínico), escritos en inglés o español y que aporten datos suficientes. Se excluyeron los estudios de investigación que utilizan únicamente imágenes de detección, los estudios de predicción y de riesgo, las actas de congresos que carecen de datos suficientes, los artículos que no aportan datos sobre ninguna de las medidas de rendimiento requeridas y los estudios que no informan de la validación interna o externa.

La selección de estudios se realizó de forma iterativa mediante tamizaje de títulos y resúmenes, seguido de la revisión de texto completo de los artículos potencialmente elegibles. No se empleó doble revisor independiente ni se elaboró un diagrama PRISMA formal, coherente con el carácter narrativo de la revisión. La extracción de datos incluyó información bibliográfica, diseño del estudio (prospectivo o retrospectivo), características de la población, modalidad de imagen, tipo de algoritmo de IA, referencia estándar empleada y métricas de rendimiento diagnóstico.

La calidad metodológica se evaluó cualitativamente siguiendo los dominios de QUADAS-2 y los principios de QUADAS-AI (selección de pacientes, prueba índice, referencia estándar, flujo y tiempos, independencia de conjuntos de entrenamiento y prueba, validación externa y reproducibilidad) (5). Se diferenciaron explícitamente las herramientas de evaluación del riesgo de sesgo de las guías de reporte como CLAIM y TRIPOD+AI. Los resultados de esta evaluación se presentan de forma narrativa en la sección de Resultados, sin una tabla numérica exhaustiva por estudio debido a la heterogeneidad de los diseños y al carácter narrativo de la revisión.

RESULTADOS

De las bibliografías revisadas, se desprende que la mayor parte de los estudios utilizan redes neuronales profundas (CNN) para la clasificación y la segmentación, en comparación con los demás enfoques. Además, se han implementado arquitecturas de tipo U-Net para la segmentación de lesiones, modelos híbridos que combinan convoluciones con características radiómicas manuales, máquinas de vector soporte (SVM), árboles extremos y XGBoost, así como vision transformers o atención.

La terminología se mantiene en inglés en cursiva cuando no existe una traducción consolidada en la literatura de habla hispana (deep learning, machine learning, Vision Transformers).

En el ámbito de la ecografía mamaria, el metaanálisis de Li et al., que incluyó 20 estudios y un total de 14.955 casos (de los cuales 4.197 se utilizaron para prueba de modelos), reportó una

sensibilidad combinada de 0,93, una especificidad de 0,90 y un AUC de 0,732. En el análisis de subgrupos, los modelos multimodales (que integran B-mode con Doppler color, elastografía o ecografía volumétrica automatizada) alcanzaron un AUC de 0,787, superior al de los modelos basados únicamente en B-mode (AUC 0,642) (3). Una revisión sistemática posterior de Dan et al., que analizó 16 estudios con 9.238 participantes, concluyó que la evidencia disponible es insuficiente para afirmar que el deep learning supera de forma consistente a los lectores humanos o mejora de manera significativa su rendimiento en entornos clínicos reales (4).

Estudios primarios multicéntricos han mostrado resultados más favorables. Gu et al. desarrollaron un modelo de atención dual basado en CNN que, mediante la combinación de las imágenes en modo B y de Doppler color, logró un AUC de entre 0,91 y 0,96 en los conjuntos de prueba externos procedentes de diferentes hospitales y equipos de fabricantes distintos. El uso de la IA como asistencia mejoró el rendimiento de radiólogos noveles y redujo la tasa de falsos positivos (7,8). Otros trabajos recientes que integran radiómica intratumoral y peritumoral, deep learning y variables clínicas han reportado AUC superiores a 0,86–0,93 en la caracterización de lesiones BI-RADS 4A, con potencial para reducir el número de biopsias innecesarias manteniendo una alta sensibilidad (12-15).

En mamografía, el metaanálisis de Yoon et al., que analizó 16 estudios con más de 1,1 millones de exámenes en casi 500.000 mujeres, encontró un AUC de 0,87 para la IA autónoma frente a 0,81 de los radiólogos en estudios de lectura de mamografía digital (diferencia estadísticamente significativa). En tomosíntesis digital de mama el AUC de la IA fue 0,90 frente a 0,79 de los radiólogos. La IA tendió a mostrar mayor sensibilidad y menor especificidad, lo que se traduce en tasas de recuerdo más elevadas cuando se utiliza como lector único (2,9).

La integración de múltiples fuentes de información (características radiómicas de distintas regiones, datos clínicos y modalidades de imagen complementarias) muestra de forma consistente una mejor discriminación técnica en entornos controlados de evaluación. Sin embargo, esta superioridad no es universal, ya que depende del diseño del estudio, del tipo de pacientes incluidos y de las condiciones de la validación (interna o externa). Existe también un sesgo relacionado con el abanico de pacientes que podría propiciar una sobreestimación de la precisión de los modelos (16-20).

Las limitaciones más comunes que determinaron la calidad de los modelos y de las predicciones fueron el diseño de estudios de tipo retrospectivo y unicéntrico; la falta de análisis de modelos de validación externa (en cuyo caso el rendimiento se veía reducido); las diferencias en los procedimientos de imagen y los umbrales de clasificación; las distribuciones desiguales de casos; y las características demográficas de determinados grupos de pacientes. Estas características suponen una limitación a la hora de aplicar los modelos en otras situaciones clínicas y explotar su verdadero potencial.

Tabla 1. Resumen de rendimiento diagnóstico reportado en metaanálisis principales.

Modalidad	Tipo de evidencia	Sensibilidad	Especificidad	AUC	Referencia
Ecografía	Metaanálisis (20 estudios)	0,93	0,90	0,732	(3)
Ecografía multimodal	Subgrupo del metaanálisis	0,93	0,88	0,787	(3)
Mamografía	Estudios de lectura	—	—	0,87	(2)
DBT	Estudios de lectura	—	—	0,90	(2)

DISCUSIÓN

Los hallazgos de esta revisión indican que la inteligencia artificial basada en deep learning ha alcanzado un nivel de desarrollo técnico que permite considerarla una herramienta complementaria en la caracterización de masas mamarias. Los valores de sensibilidad y especificidad reportados en metaanálisis de ecografía, así como los AUC de estudios de lectura en mamografía, sugieren un potencial clínico significativo, particularmente en escenarios donde la disponibilidad de radiólogos especializados es limitada o donde se busca reducir la variabilidad interobservador (2,3).

Sin embargo, hay que tener en cuenta que se ha constatado una discrepancia entre los resultados experimentales y el rendimiento de estas herramientas en el ámbito clínico real. El éxito técnico de las soluciones multimodales debe evaluarse en contextos específicos, teniendo en cuenta las características de los pacientes, las lesiones analizadas y el diseño del estudio. En mi opinión, basada en los resultados experimentales presentados y en las declaraciones de las sociedades europeas de referencia, la mejor opción es considerar a la IA como una herramienta de apoyo o una segunda opinión, más que como un sustituto de la opinión humana profesional real (7,10,11).

Los estudios mencionados anteriormente presentan ciertas limitaciones que deben tenerse en cuenta a la hora de aplicar las mismas tecnologías en la práctica clínica. Entre ellas, cabe destacar que son retrospectivos, de carácter monocéntrico, carecen de validación externa y están formados por un grupo de población limitado. Es evidente que las aplicaciones prácticas de estos sistemas requieren una investigación adicional, en la que se estudien sus efectos reales sobre parámetros clínicos importantes, como la tasa de detección de tumores, las tasas de cáncer entre los que fueron pasados por alto, la tasa de detección y la necesidad de biopsia.

También es necesario considerar la implementación de decisiones similares en entornos clínicos reales, con sus múltiples dilemas éticos, legales y políticos, algunos de los cuales pueden tener repercusiones catastróficas para todos los implicados.

CONCLUSIONES

Los sistemas de inteligencia artificial basados en deep learning demuestran un rendimiento diagnóstico prometedor para la caracterización de masas mamarias benignas y malignas en mamografía y ecografía, con sensibilidades combinadas cercanas a 0,92–0,93 y especificidades de 0,88–0,91 en ecografía, y AUC de aproximadamente 0,87–0,90 en estudios de lectura de mamografía. Por otro lado, hay que reconocer que los estudios son principalmente retrospectivos y de un solo centro, y que se necesitan investigaciones prospectivas y multicéntricas para confirmar estos resultados.

Los modelos que combinan diversas fuentes de información suelen ofrecer mejor discriminación que aquellos que solo utilizan una, pero dicha discriminación no está exenta de reactividad. Habría que llevar a cabo estudios prospectivos y multicéntricos para evaluar las aplicaciones clínicas de la inteligencia artificial (IA) en materia de imagenología mamaria, con el fin de establecer si la tecnología permite un diagnóstico más preciso; además, habría que abordar las preocupaciones normativas y éticas expresadas. En la actualidad, la posición más defendible es la de la IA como herramienta complementaria al radiólogo.

REFERENCIAS BIBLIOGRÁFICAS

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2024;74(3):229-263. doi:10.3322/caac.21834
2. Yoon JH, Strand F, Baltzer PAT, Conant EF, Gilbert FJ, Lehman CD, Morris EA, Mullen LA, Nishikawa RM, Sharma N, Vejborg I, Moy L, Mann RM. Standalone AI for breast cancer detection at screening digital mammography and digital breast tomosynthesis: a systematic review and meta-analysis. *Radiology.* 2023;307(5):e222639. doi:10.1148/radiol.222639
3. Li H, Zhao J, Jiang Z. Deep learning-based computer-aided detection of ultrasound in breast cancer diagnosis: a systematic review and meta-analysis. *Clin Radiol.* 2024;79(11):e1403-e1413. doi:10.1016/j.crad.2024.08.002
4. Dan Q, Xu Z, Burrows H, Bissram J, Stringer JSA, Li Y. Diagnostic performance of deep learning in ultrasound diagnosis of breast cancer: a systematic review. *NPJ Precis Oncol.* 2024;8(1):21. doi:10.1038/s41698-024-00514-z
5. Sounderajah V, Ashrafian H, Rose S, Shah NH, Ghassemi M, Golub R, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. *Nat Med.* 2021;27(10):1663-1665. doi:10.1038/s41591-021-01517-0
6. Baethge C, Goldbeck-Wood S, Mertens S. SANRA—a scale for the quality assessment of narrative review articles. *Res Integr Peer Rev.* 2019;4:5. doi:10.1186/s41073-019-0064-8
7. Gu Y, Xu W, Liu T, An X, Tian J, Ran H, et al. Deep Learning-assisted Diagnosis of Breast Lesions on US Images: A Multivendor, Multicenter Study. *Radiol Artif Intell.* 2023;5(5):e220185. doi:10.1148/ryai.220185

8. Gu Y, Xu W, Liu T, et al. Ultrasound-based deep learning in the establishment of a breast lesion risk stratification system: a multicenter study. *Eur Radiol.* 2023;33(4):2954-2964. doi:10.1007/s00330-022-09263-8
9. Scaranelo AM. Standalone AI in Breast Cancer Screening: Where We Are and What Is to Be Achieved. *Radiology.* 2023;307(5):e230935. doi:10.1148/radiol.230935
10. Athanasiou A, Appelman L, Pijnappel RM, Gilbert FJ, Pediconi F, Mann R. ESR Essentials: diagnostic work-up in patients with symptomatic breast disease—practice recommendations by the European Society of Breast Imaging. *Eur Radiol.* 2025;35(2):723-732. doi:10.1007/s00330-024-10980-5
11. Klontzas ME, Groot Lipman KBW, Akinci D'Antonoli T, Andreychenko A, Cuocolo R, Dietzel M, et al. ESR Essentials: common performance metrics in AI—practice recommendations by the European Society of Medical Imaging Informatics. *Eur Radiol.* 2026;36(2):1528-1540. doi:10.1007/s00330-025-11890-w
12. Singla V, Garg D, Negi S, Mehta N, Pallavi T, Choudhary S, Dhiman A. Deep learning powered breast ultrasound to improve characterization of breast masses: a prospective study. *Acta Radiol.* 2026;67(1):13-23. doi:10.1177/02841851251377927
13. Vallez N, Mateos-Aparicio-Ruiz I, Rienda MA, Deniz O, Bueno G. Comparative analysis of deep learning methods for breast ultrasound lesion detection and classification. *Phys Med.* 2025;134:104993. doi:10.1016/j.ejmp.2025.104993
14. Lin X, Liao T, Yang Y, Ouyang R, Zhou Y, Lai X, Ma J. Value of deep learning model for predicting Breast Imaging Reporting and Data System 3 and 4A lesions on mammography. *Quant Imaging Med Surg.* 2025;15(5):4047-4058. doi:10.21037/qims-24-1523
15. Breast Cancer Diagnosis Using a Dual-Modality Complementary Deep Learning Network With Integrated Attention Mechanism Fusion of B-Mode Ultrasound and Shear Wave Elastography. *Ultrasound Med Biol.* 2025;51(11):2135-2143. doi:10.1016/j.ultrasmedbio.2025.08.007
16. Tian R, Lu G, Tang S, Sang L, Ma H, Qian W, Yang W. Benign and malignant classification of breast tumor ultrasound images using conventional radiomics and transfer learning features: A multicenter retrospective study. *Med Eng Phys.* 2024;125:104117. doi:10.1016/j.medengphy.2024.104117
17. Bruno P, Macrì M, Dodaro C, et al. A Dual-stage Deep Learning Framework for Breast Ultrasound Image Segmentation and Classification. *J Med Syst.* 2025;49(1):162. doi:10.1007/s10916-025-02298-6
18. Bunnell A, Valdez D, Wolfgruber TK, Quon BK, Hung K, Hernandez BY, Seto TB, Killeen J. Prediction of mammographic breast density based on clinical breast ultrasound images using deep learning: a retrospective analysis. *Lancet Reg Health Am.* 2025;46:101096. doi:10.1016/j.lana.2025.101096
19. Acosta-Jiménez S, Mendoza-Mendoza MM, Galván-Tejada CE, Celaya-Padilla JM, Galván-Tejada JI, Soto-Murillo MA. Explainable Deep Learning for Breast Lesion Classification in Digital and Contrast-Enhanced Mammography. *Diagnostics (Basel).* 2025;15(24):3143. doi:10.3390/diagnostics15243143

20. Shia WC, et al. Deep learning-based classification of benign and malignant breast microcalcifications in mammography. Sci Rep. 2025;15(1):42768. doi:10.1038/s41598-025-27149-9