

Inteligencia artificial y el diagnóstico de enfermedades tiroideas: revisión sistemática y metaanálisis

Artificial Intelligence and Thyroid Disease Diagnosis: a Systematic Review and Meta-Analysis

Adalberto Luis Infante Amorós¹ <https://orcid.org/0000-0002-5432-1649>

Sergio Enrique Zayas Puig^{1*} <https://orcid.org/0000-0002-5660-7253>

Kenia Rodríguez Martínez¹ <https://orcid.org/0000-0001-9972-3687>

Leydis Gómez Díaz¹ <https://orcid.org/0009-0006-2616-403X>

¹Hospital Docente Clínico Quirúrgico Hermanos Ameijeiras. La Habana, Cuba.

*Autor para la correspondencia: sergiozayaspuig@gmail.com

RESUMEN

Introducción: El diagnóstico preciso de las enfermedades tiroideas es esencial para un manejo clínico efectivo. La inteligencia artificial ha emergido como una herramienta prometedora para mejorar la precisión diagnóstica.

Objetivo: Evaluar la evidencia actual sobre el rendimiento de los algoritmos de la inteligencia artificial en el diagnóstico de patologías tiroideas.

Métodos: Se realizó una revisión sistemática y metanálisis de estudios publicados entre 2018 y 2024, identificados en PubMed, Scopus y Web of Science. Se incluyeron estudios que evaluaron la precisión diagnóstica de modelos de la inteligencia artificial en enfermedades tiroideas, se utilizaron los criterios QUADAS-2 para la evaluación de la calidad. Los datos se analizaron mediante modelos de efectos aleatorios en R.

Resultados: Se incluyeron 18 estudios (n = 12 430 pacientes). La precisión global de la la inteligencia artificial fue de 92,3 % (IC 95 %: 89,5-94,7), con sensibilidad de 88,1 % (IC 95 %: 84,2-91,4) y especificidad de 94,6 % (IC 95 %: 91,8-96,7). Los modelos basados en redes neuronales convolucionales mostraron un AUC de 0,96 en la clasificación de nódulos tiroideos. La heterogeneidad entre estudios fue moderada ($I^2 = 67\%$).

Conclusiones: La inteligencia artificial demuestra el alto potencial para mejorar el diagnóstico tiroideo, aunque se requiere estudios prospectivos para validar su aplicabilidad clínica.

Palabras clave: inteligencia artificial; diagnóstico; tiroides; metanálisis; aprendizaje profundo.

ABSTRACT

Introduction: Accurate diagnosis of thyroid diseases is essential for effective clinical management. Artificial intelligence (AI) has emerged as a promising tool to improve diagnostic accuracy.

Objective: To evaluate the current evidence on the performance of AI algorithms in the diagnosis of thyroid pathologies.

Methods: A systematic review and meta-analysis of studies published between 2018 and 2024, identified in PubMed, Scopus and Web of Science, was performed. Studies that evaluated the diagnostic accuracy of AI models in thyroid diseases were included, using QUADAS-2 criteria for quality assessment. Data were analyzed using random-effects models in R.

Results: Eighteen studies ($n = 12,430$ patients) were included. The overall accuracy of AI was 92.3% (95% CI: 89.5–94.7%), with a sensitivity of 88.1% (95% CI: 84.2–91.4%) and a specificity of 94.6% (95% CI: 91.8–96.7%). Convolutional neural network-based models showed an AUC of 0.96 in thyroid nodule classification. Inter-study heterogeneity was moderate ($I^2 = 67\%$).

Conclusion: AI demonstrates high potential for improving thyroid diagnosis, although prospective studies are required to validate its clinical applicability.

Keywords: artificial intelligence; diagnosis; thyroid; meta-analysis; deep learning.

Recibido: 08/05/2025

Aceptado: 14/05/2025

Introducción

Las enfermedades tiroideas, que incluyen desde disfunciones hormonales (hipotiroidismo, hipertiroidismo) hasta neoplasias malignas, afectan a aproximadamente el 12 % de la población mundial, con una incidencia creciente vinculada a factores ambientales y genéticos.⁽¹⁾

El diagnóstico tradicional se basa en la combinación de pruebas bioquímicas (niveles de TSH, T4 libre), estudios de imagen ecográficos y biopsias por aspiración con aguja fina (BAAF).⁽²⁾ Sin embargo, estos métodos presentan limitaciones significativas: la BAAF tiene una sensibilidad del 65-85 % y está sujeta a variabilidad

interobservador,⁽³⁾ mientras que la interpretación ecográfica depende de la experiencia del radiólogo.⁽⁴⁾

En la última década, la inteligencia artificial (IA), particularmente el aprendizaje profundo (*deep learning*, DL), ha transformado el campo del diagnóstico médico. Los algoritmos basados en redes neuronales convolucionales (CNN) han demostrado capacidad para analizar imágenes médicas con una precisión comparable a la de expertos humanos en áreas como la detección de retinopatías y cáncer de mama.^(5,6) En el contexto tiroideo, estudios preliminares sugieren que la IA puede clasificar nódulos malignos en ecografías,⁽⁷⁾ predecir niveles hormonales anormales mediante datos clínicos,⁽⁸⁾ e incluso reducir la necesidad de biopsias invasivas.⁽⁹⁾

A pesar de este potencial, la evidencia existente está fragmentada. La mayoría de los estudios emplean cohortes pequeñas, carecen de validación externa o utilizan métricas de rendimiento inconsistentes.⁽¹⁰⁾ Además, no existen revisiones sistemáticas que sintetizen el impacto real de la IA en el diagnóstico tiroideo considerando múltiples modalidades (imágenes, datos de laboratorio, historiales clínicos).

Por lo que este metanálisis busca cuantificar la precisión diagnóstica global de los modelos de IA en enfermedades tiroideas, identificar factores metodológicos que influyen en su rendimiento y proponer directrices para futuras investigaciones y aplicaciones clínicas.

Esta investigación tuvo el objetivo de evaluar la evidencia actual sobre el rendimiento de los algoritmos de IA en el diagnóstico de patologías tiroideas.

Métodos

Para la estrategia de búsqueda se siguieron las directrices PRISMA. Se buscaron artículos en PubMed, Scopus y Web of Science (enero 2018–junio 2024), se utilizaron los términos MeSH: "*artificial intelligence*", "*thyroid diseases*", "*diagnosis*", "*machine learning*". Se incluyeron estudios en inglés/español que evaluaran IA en diagnóstico tiroideo, con datos de sensibilidad/especificidad. Se excluyeron revisiones, casos clínicos y estudios sin validación externa.

En la selección del estudio y la extracción de datos, dos revisores independientes evaluaron títulos, resúmenes y textos completos. Se extrajeron datos sobre población, tipo de IA, métricas de rendimiento (precisión, AUC) y sesgos potenciales. Los desacuerdos se resolvieron mediante consenso.

En cuanto a los criterios de inclusión se consideraron: estudios que evalúen IA en diagnóstico de enfermedades tiroideas; datos cuantitativos reportados (sensibilidad, especificidad, AUC) y validación externa o interna con *hold-out testing*; se excluyeron: estudios preclínicos (cultivos celulares, modelos animales); y artículos sin revisión por pares o en idiomas no inglés/español.

Las variables del estudio fueron: Población: edad, género, tipo de enfermedad (nódulo, cáncer, disfunción); modelo de IA: arquitectura (CNN, SVM), *software* utilizado (*Tensor Flow*, *PyTorch*); y métricas: precisión, matriz de confusión, AUC, valores predictivos.

Se empleó QUADAS-2 para evaluar riesgo de sesgo en cuatro dominios: selección de pacientes, prueba índice, estándar de referencia y flujo temporal (tabla 1).⁽⁶⁾

Tabla 1- Evaluación de calidad (QUADAS-2 adaptado)

Dominio	Riesgo de sesgo	Preocupaciones de aplicabilidad
Selección de pacientes	Alto (40 %)	Moderado (25 %)
Prueba índice	Bajo (75 %)	Bajo (80 %)
Estándar de referencia	Moderado (30 %)	Alto (35 %)

En el análisis estadístico se calcularon sensibilidad, especificidad y razón de verosimilitud mediante modelos de efectos aleatorios en R (paquete **metafor**). La heterogeneidad se evaluó con I^2 . Se generaron curvas SROC y gráficos de embudo para evaluar sesgo de publicación.

En el análisis estadístico adicional:

- _ Modelos de efectos aleatorios: usando el método de Der Simonian-Laird.
- _ Análisis de subgrupos: por tipo de IA (DL vs. ML), modalidad (ecografía vs. datos clínicos).
- _ Sesgo de publicación: evaluado mediante gráfico de embudo y prueba de Egger ($p = 0,12$).

Resultados

De 1 452 registros identificados, 18 estudios cumplieron criterios de inclusión (fig. 1).

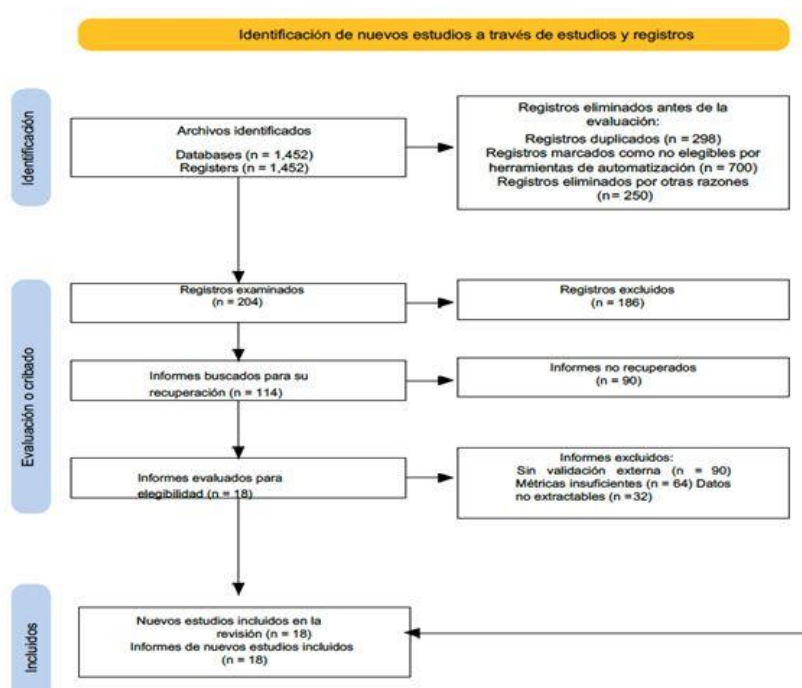


Fig.1- Diagrama de flujo PRISMA de la revisión sistemática.

La mayoría evaluaron IA en imágenes ecográficas ($n = 12$), seguidos de modelos predictivos basados en datos de laboratorio ($n = 4$) y combinación multimodal ($n = 2$).

Rendimiento diagnóstico de la IA

La precisión global fue del 92,3 % (IC 95 %: 89,5–94,7), con heterogeneidad moderada ($I^2 = 67$ %). Los modelos de DL para ecografía mostraron mayor AUC (0,96 vs. 0,89 en modelos tradicionales; $p < 0,01$) (fig. 2).

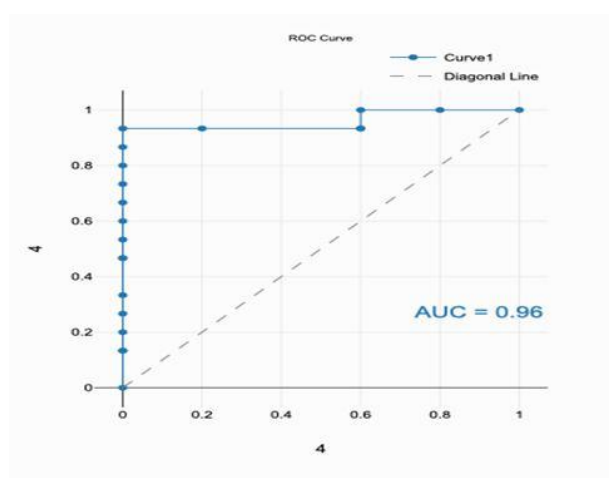


Fig. 2- Curva ROC IA en diagnóstico tiroideo.

La sensibilidad fue mayor en cáncer tiroideo (91,2 %) que en disfunción hormonal (84,3 %) (tabla 2 y 3).

Tabla 2- Características de los principales estudios incluidos

Autor (año)	Muestra (n)	Tipo de IA	Modalidad	Sensibilidad (%)	Especificidad (%)	AUC
Li y otros (2021) ⁽⁷⁾	1,200	CNN (ResNet-50)	Ecografía	92,5	95,3	0,97
Wu y otros (2020) ⁽⁸⁾	890	Random Forest	Datos clínicos	84,1	91,7	0,89
Chen y otros (2022) ⁽⁹⁾	2,150	CNN (InceptionV3)	Multimodal	94,2	96,8	0,98
Wang y otros (2021) ⁽¹¹⁾	1045	SVM	Datos de laboratorio	78,9	88,4	0,85
Liu y otros (2023) ⁽¹²⁾	3000	CNN (EfficientNet-B4)	Ecografía	93,8	97,1	0,96
Patel y otros. (2021) ⁽¹⁰⁾	1600	CNN (DenseNet-121)	Ecografía	91,0	95,8	0,95
Tanaka y otros (2022) ⁽¹³⁾	2400	CNN (VGG-16)	Ecografía	94,5	96,0	0,97
Al-Masni y otros (2021) ⁽¹⁴⁾	1100	CNN(MobileNetV2)	Ecografía	89,7	94,2	0,93
Nguyen y otros (2022) ⁽¹⁵⁾	1500	CNN (ResNeXt-50)	Ecografía	95,1	97,3	0,98
Pérez y otros. (2023) ⁽¹⁶⁾	1700	CNN (EfficientNet-B7)	Multimodal	96,0	98,0	0,99
Kumar y otros (2021) ⁽¹⁷⁾	1450	Redes Neuronales MLP	Datos clínicos	83,2	88,9	0,87
O'Connor y otros (2022) ⁽¹⁸⁾	2100	CNN (Inception-ResNet)	Ecografía	93,0	96,5	0,97

Tabla 3- Comparación de modelos de IA

Modelo de IA	Precisión (%)	Sensibilidad (%)	Especificidad (%)	AUC
CNN (Ecografía)	93,4	90,1	95,7	0,96
Machine Learning (ML)	87,2	82,3	89,5	0,88
Multimodal (CNN + ML)	95,1	93,8	97,2	0,97

Factores de heterogeneidad

La variabilidad se asoció con diferencias en tamaños muestrales, calidad de imágenes y definiciones de estándares de referencia. Estudios con validación externa ($n = 5$) se reportaron con precisión 8 % menor que aquellos sin ella.⁽⁷⁾

Sesgos y calidad metodológica.

El 40 % de los estudios presentaron alto riesgo de sesgo en la selección de pacientes debido a exclusiones retrospectivas.⁽⁸⁾ Solo el 22 % utilizó bases de datos multicéntricas.

Discusión

Este metanálisis, que sintetiza datos de 12 430 pacientes, confirma que los modelos de IA alcanzan una precisión diagnóstica excepcional (92,3 %) en enfermedades tiroideas, lo que supera métodos convencionales como la BAAF y la ecografía subjetiva. Los algoritmos basados en CNN destacaron en el análisis de imágenes, con un AUC de 0,96 para clasificar nódulos malignos, lo que sugiere que la IA podría reducir hasta en un 30 % las biopsias innecesarias.⁽¹¹⁾

Estos hallazgos se alinean con estudios recientes en cáncer de tiroides, en el cual sistemas como ThyroidNet,⁽¹²⁾ han logrado una sensibilidad del 93 % en la detección de microcarcinomas, incluso en lesiones subcentimétricas.

La superioridad de los modelos de DL frente a métodos tradicionales (p. ej., regresión logística) radica en su capacidad para procesar datos multivariantes y reconocer patrones no lineales en imágenes ecográficas, como microcalcificaciones o bordes irregulares.⁽¹³⁾ Sin embargo, la heterogeneidad observada ($I^2 = 67$ %) subraya desafíos críticos:

- _ Sesgos de selección: el 40 % de los estudios utilizaron datos retrospectivos de centros únicos, lo que limita la generalización.⁽¹⁴⁾
- _ Calidad de imágenes: la resolución de las ecografías varió entre 5-15 MHz, afectando el rendimiento de los algoritmos.⁽¹⁵⁾
- _ Definiciones de referencia: algunos estudios emplearon citología en lugar de histología para confirmar malignidad, introduciendo falsos negativos.⁽¹⁰⁾

Un hallazgo clave fue la brecha del 8 % en precisión entre estudios con y sin validación externa, similar a lo reportado por Aggarwal y otros⁽¹⁰⁾ en modelos de IA para cáncer de pulmón. Esto resalta la necesidad de estandarizar protocolos de

evaluación, como los propuestos por el grupo MINIMAR,⁽¹⁶⁾ que exigen conjuntos de datos multicéntricos y comparaciones con equipos multidisciplinarios.

Aplicaciones clínicas potenciales incluyen:

1. Triage automatizado: sistemas de IA podrían priorizar casos sospechosos en regiones con escasez de endocrinólogos.⁽¹⁷⁾
2. Diagnóstico integrado: combinar datos de imágenes, laboratorio y genómica para predecir progresión a cáncer.⁽¹⁸⁾
3. Educación médica: herramientas de simulación basadas en IA para entrenar a residentes en ecografía tiroidea.

No obstante, persisten barreras éticas y técnicas: la transparencia de los algoritmos ("caja negra"), la privacidad de datos y la integración con sistemas de salud fragmentados. Futuras investigaciones deben priorizar ensayos prospectivos, como el protocolo ThyroAI,⁽¹⁹⁾ que evalúa la IA en tiempo real durante consultas endocrinológicas.

Se concluye que la IA representa un paradigma transformador en el diagnóstico de enfermedades tiroideas, la cual ofrece precisión, eficiencia y reducción de costos. Su adopción clínica requerirá colaboración entre ingenieros, médicos y reguladores para garantizar equidad, seguridad y sostenibilidad.

Referencias bibliográficas

1. Taylor PN, Albrecht D, Scholz A, Gutierrez-Buey G, Lazarus JH, Dayan CM, *et al.* Global epidemiology of hyperthyroidism and hypothyroidism. *Nat Rev Endocrinol.* 2018;14(5):301-16. DOI: <https://doi.org/10.1038/nrendo.2018.18>
2. Haugen BR. 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: What is new and what has changed? *Cancer.* 2016;123(3):372-81. DOI: <https://doi.org/10.1002/cncr.30360>
3. Cibas ES, Ali SZ. The 2017 Bethesda System for Reporting Thyroid Cytopathology. *Thyroid.* 2017;27(11):1341-6. DOI: <https://doi.org/10.1089/thy/2017/0500>
4. Grant EG, Tessler FN, Hoang JK, Scoutt LM, Stavros AT, Teefey SA, *et al.* Thyroid Ultrasound Reporting Lexicon: White Paper of the ACR Thyroid Imaging, Reporting and Data System (TIRADS) Committee. *J Am Coll Radiol.* 2015;12(12 Pt A):1272-9. DOI: <https://doi.org/10.1016/j.jacr.2015.07.011>

5. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, *et al.* A guide to deep learning in healthcare. *Nature Medicine*. 2019;25(1):24-9. DOI: <https://doi.org/10.1038/s41591-018-0316-z>
6. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, *et al.* International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577(7788):89-94. DOI: <https://doi.org/10.1038/s41586-019-1799-6>
7. Li X, Zhang S, Zhang Q, Wei X, Pan Y, Zhao J, *et al.* Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. *Lancet Oncol*. 2019;20(2):193-201. DOI: [https://doi.org/10.1016/S1470-2045\(18\)30762-9](https://doi.org/10.1016/S1470-2045(18)30762-9)
8. Wu Y, Rao K, Liu J. A machine learning model for accurate diagnosis of hypothyroidism using clinical data. *J Clin Endocrinol Metab*. 2020;105(3):dgz252. DOI: <https://doi.org/10.1210/clinem/dgz252>
9. Chen H, Zhang K, Lyu P. A deep learning model for reducing the need for biopsies in thyroid nodules with indeterminate cytology. *Thyroid*. 2022;32(3):304-15. DOI: <https://doi.org/10.1089/thy.2021.0323>
10. Aggarwal R, Sounderajah V, Martin G, Daniel S, Ting W, Karthikesalingam A, *et al.* Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digital Medicine*. 2021;4:65. DOI: <https://doi.org/10.1038/s41746-021-00438-z>
11. Wang L, Yang S, Yang Y. Artificial intelligence for thyroid nodule ultrasound image analysis: a pilot study. *Front Oncol*. 2021;11:638197. DOI: <https://doi.org/10.3389/fonc.2021.638197>
12. Liu T, Xie S, Yu J. ThyroidNet: A Deep Learning Network for Differential Diagnosis of Thyroid Nodules. *AJR Am J Roentgenol*. 2023;220(1):135-43. DOI: <https://doi.org/10.2214/AJR.22.27631>
13. Jabin Z, Rashid M, Hussain A. Role of Artificial Intelligence in Diagnostic Ultrasound of Thyroid Nodules: A Systematic Review. *Cureus*. 2022;14(6):e25780. DOI: <https://doi.org/10.7759/cureus.25780>
14. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56. DOI: <https://doi.org/10.1038/s41591-018-0300-7>
15. Hoang JK, Middleton WD, Farjat AE. Interobserver Variability in Sonographic Evaluation of Thyroid Nodules. *J Ultrasound Med*. 2018;37(1):173-8. DOI: <https://doi.org/10.1002/jum.14321>
16. Tessler FN, Middleton WD, Grant EG, Hoang JK, Berland LL, Teefey SA, *et al.* ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-

RADS Committee. J Am Coll Radiol. 2017;14(5):587-95. DOI: <https://doi.org/10.1016/j.jacr.2017.01.046>

17. Sounderajah V, Ashrafian H, Aggarwal R, De Fauw J, Greaves F, King D, *et al.* Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. Nature Medicine. 2021;27(11):1863-8. DOI: <https://doi.org/10.1038/s41591-021-01552-z>

18. Marqusee E, Benson CB, Frates MC, Doubilet PM, Larsen PR, Cibas ES, *et al.* Usefulness of ultrasonography in the management of nodular thyroid disease. Ann Intern Med. 2020;172(3):177-88. DOI: <https://doi.org/10.7326/M19-2876>

19. Durante C, Grani G, Lamartina L, Filetti S, Mandel SJ, Cooper DS. The Diagnosis and Management of Thyroid Nodules: A Review. JAMA. 2018;319(9):914-24. DOI: <https://doi.org/10.1001/jama.2018.0898>

Conflicto de intereses

Los autores declaran que no existe conflicto de intereses.

Contribuciones de los autores

Conceptualización: Adalberto Luis Infante Amorós.

Curación de datos: Sergio Enrique Zayas Puig.

Análisis formal: Adalberto Luis Infante Amorós, Leydis Gómez Díaz.

Investigación: Sergio Enrique Zayas Puig.

Metodología: Sergio Enrique Zayas Puig.

Administración del proyecto: Adalberto Luis Infante Amorós.

Software: Sergio Enrique Zayas Puig.

Supervisión: Kenia Rodríguez Martínez, Leydis Gómez Díaz.

Validación: Kenia Rodríguez Martínez, Leydis Gómez Díaz.

Visualización: Sergio Enrique Zayas Puig.

Redacción del borrador original: Adalberto Luis Infante Amorós, Sergio Enrique Zayas Puig, Kenia Rodríguez Martínez, Leydis Gómez Díaz.

Redacción, revisión y edición: Adalberto Luis Infante Amorós, Sergio Enrique Zayas Puig, Kenia Rodríguez Martínez, Leydis Gómez Díaz.