



DOI: <https://doi.org/10.46296/yc.v10i19.0972>

DEEP LEARNING PARA EL DIAGNÓSTICO AUTOMÁTICO DE LESIONES DE LA MEMBRANA TIMPÁNICA MEDIANTE OTOSCOPIA DIGITAL: PRECISIÓN DIAGNÓSTICA, APLICACIONES CLÍNICAS Y DESAFÍOS DE IMPLEMENTACIÓN

DEEP LEARNING FOR THE AUTOMATED DIAGNOSIS OF TYMPANIC MEMBRANE LESIONS USING DIGITAL OTOSCOPY: DIAGNOSTIC ACCURACY, CLINICAL APPLICATIONS, AND IMPLEMENTATION CHALLENGES

Romero-Ibáñez Heiddy Carolina ¹; Mena-Torres Belén Isabel ²; Pazmiño-Vera Katherine Liliana ³; Rocano-Salinas Bryam Alejandro ⁴; Bermeo-Polo Iván Mauricio ⁵

¹ Médico general, MSc. Salud y seguridad ocupacional. Guayaquil, Ecuador.
Correo: heiddyta100@gmail.com

² Médico general, Centro de Salud de Zumba. Ecuador.
Correo: belenmenat@gmail.com. ORCID ID: <https://orcid.org/0009-0009-1428-8272>

³ Médico general, Hospital General Monte Sinaí. Guayaquil, Ecuador.
Correo: katherineliliana_1999@outlook.es. ORCID ID: <https://orcid.org/0009-0008-2497-886X>

⁴ Médico General, Universidad católica de Cuenca. Ecuador.
Correo: alebars1997@hotmail.com. ORCID: <https://orcid.org/0009-0003-7809-8597>

⁵ Médico General, Ministerio de Salud Pública. Quito, Ecuador.
Correo: ivanb.p@hotmail.com. ORCID: <https://orcid.org/0009-0009-5576-057X>

Resumen

Introducción: La otoscopia constituye el pilar de la evaluación de la membrana timpánica, pero su interpretación depende de la experiencia del examinador, de la calidad de la imagen y de la correcta visualización de los reparos anatómicos. En los últimos años, el deep learning ha permitido automatizar tareas de clasificación, segmentación y cuantificación de lesiones timpánicas, con aplicaciones potenciales en otorrinolaringología, atención primaria y telemedicina. **Objetivo:** Analizar la evidencia reciente sobre deep learning aplicado al diagnóstico automático de lesiones de la membrana timpánica mediante otoscopia digital, describiendo desempeño diagnóstico, utilidad clínica, limitaciones y prioridades para su implementación segura. **Métodos:** Se realizó una revisión narrativa crítica y actualización de evidencia publicada principalmente entre enero de 2020 y agosto de 2026. Se priorizaron estudios originales y revisiones indexadas que evaluaron redes neuronales profundas sobre imágenes otoscópicas u otoendoscópicas para clasificación, detección, segmentación o cuantificación de patología timpánica. Se extrajeron arquitectura, tamaño del conjunto de imágenes, categorías diagnósticas, estrategia de validación y métricas de rendimiento. **Resultados:** Las redes convolucionales alcanzan con frecuencia exactitudes superiores al 90% en validaciones internas para diferenciar membranas normales de patológicas y para clasificar otitis media, perforaciones y colesteatoma. Estudios con 2.272 a 6.066 imágenes reportaron exactitudes aproximadas de 93–97%, mientras que sistemas móviles han mostrado rendimiento elevado en escenarios controlados. Sin embargo, la validación externa evidencia una disminución sustancial del desempeño: un estudio multicohorte mostró una caída del AUROC medio de 0,95 en validación interna a 0,76 en conjuntos externos. La segmentación automática ha alcanzado coeficientes Dice cercanos a 0,93 y permite cuantificar perforaciones y extraer características con valor funcional. **Conclusiones:** El deep learning aplicado a otoscopia digital presenta un alto potencial como herramienta de apoyo diagnóstico, triage y tele-otoscopia, pero la evidencia actual no justifica su uso autónomo. La generalización entre dispositivos, centros y poblaciones, la calidad

Información del manuscrito:

Fecha de recepción: 17 de abril de 2026.

Fecha de aceptación: 22 de junio de 2026.

Fecha de publicación: 10 de julio de 2026.



del estándar de referencia, la calibración, la explicabilidad y la validación prospectiva multicéntrica siguen siendo requisitos esenciales para la transición desde prototipos de investigación hacia sistemas clínicamente confiables.

Palabras claves: deep learning; membrana timpánica; otoscopia; inteligencia artificial; otitis media; perforación timpánica; colesteatoma; imagen médica.

Abstract

Introduction: Otoscopy is central to tympanic membrane assessment, but interpretation depends on examiner expertise, image quality, and adequate visualization of anatomic landmarks. Deep learning has enabled automated classification, segmentation, and quantitative analysis of tympanic membrane abnormalities, creating potential applications in otolaryngology, primary care, and telemedicine. **Objective:** To critically review recent evidence on deep learning for automated diagnosis of tympanic membrane lesions using digital otoscopy, focusing on diagnostic performance, clinical utility, limitations, and priorities for safe implementation. **Methods:** A structured narrative review and evidence update was conducted, focusing mainly on studies published from January 2020 through August 2026. Original studies and indexed reviews evaluating deep neural networks on otoscopic or otoendoscopic images for classification, detection, segmentation, or quantification of tympanic membrane pathology were prioritized. Architecture, dataset size, diagnostic classes, validation strategy, and performance metrics were extracted. **Results:** Convolutional neural networks frequently achieve internal-validation accuracies above 90% for normal-versus-abnormal classification and for otitis media, tympanic membrane perforation, and cholesteatoma. Studies using 2,272–6,066 images reported accuracies of approximately 93–97%, and smartphone-based systems have demonstrated high performance under controlled conditions. Nonetheless, external validation shows a relevant performance drop; a multicohort study reported a decrease in mean AUROC from 0.95 internally to 0.76 on external datasets. Automated segmentation has achieved Dice coefficients around 0.93 and can quantify perforation morphology and derive functionally relevant image features. **Conclusions:** Deep learning applied to digital otoscopy is promising for diagnostic support, triage, and tele-otoscopy, but current evidence does not support autonomous clinical use. Cross-device and cross-population generalizability, robust reference standards, calibration, explainability, and prospective multicenter validation remain essential before routine deployment.

Keywords: deep learning; tympanic membrane; otoscopy; artificial intelligence; otitis media; tympanic membrane perforation; cholesteatoma; medical imaging.

1. Introducción

La inspección de la membrana timpánica mediante otoscopia u otoendoscopia es una de las exploraciones más frecuentes en otorrinolaringología y atención primaria. A partir de características como transparencia, coloración, posición, integridad, vascularización, presencia de líquido, retracciones, placas de timpanoesclerosis, perforaciones o detritos

queratínicos, el clínico establece diagnósticos que abarcan desde variantes normales hasta otitis media aguda, otitis media con efusión, otitis media crónica, perforación timpánica y colesteatoma. Aunque la otoscopia es accesible y rápida, su interpretación presenta variabilidad interobservador y puede ser especialmente difícil cuando la calidad de la imagen es deficiente o el examinador tiene experiencia limitada [1,2].



La digitalización de la otoscopia ha transformado esta limitación en una oportunidad para el análisis computacional. Las imágenes obtenidas con video-otoscopios, endoscopios rígidos y accesorios acoplados a teléfonos inteligentes pueden almacenarse, ampliarse y procesarse mediante visión por computador. A diferencia de los métodos clásicos de aprendizaje automático, el deep learning aprende representaciones jerárquicas directamente desde los píxeles, lo que permite identificar patrones de textura, color, contorno y estructura sin definir manualmente todas las características relevantes [3–8].

Las redes neuronales convolucionales (CNN) —incluyendo ResNet, DenseNet, Inception, MobileNet, EfficientNet y arquitecturas derivadas de U-Net— han sido las más utilizadas en otoscopia automática. Sus aplicaciones se han expandido desde la clasificación binaria normal/anormal hasta la diferenciación multiclase, la segmentación de la membrana y de las perforaciones, la estimación del tamaño lesional y, más recientemente, la predicción de

resultados funcionales auditivos [8–12,20–22]. Este avance sitúa a la otoscopia asistida por inteligencia artificial en la intersección entre otorrinolaringología, imagen médica y salud digital.

El objetivo de esta revisión es sintetizar críticamente la evidencia contemporánea sobre deep learning aplicado al diagnóstico automático de lesiones de la membrana timpánica mediante otoscopia digital, con énfasis en precisión diagnóstica, tareas de imagen, utilidad clínica, validación externa, integración con dispositivos móviles y barreras para su implementación segura.

2. Materiales y métodos

2.1 Diseño de la revisión

Se realizó una revisión narrativa crítica con actualización estructurada de la evidencia. Este diseño fue elegido para integrar estudios de desarrollo de algoritmos, estudios de exactitud diagnóstica, trabajos de segmentación y cuantificación, comparaciones humano–máquina y revisiones sistemáticas, cuya heterogeneidad

metodológica dificulta una síntesis cuantitativa única.

2.2 Estrategia de identificación de evidencia

La actualización bibliográfica se centró principalmente en literatura publicada entre enero de 2020 y agosto de 2026, incorporando algunos trabajos previos considerados fundacionales para contextualizar la evolución del campo. Se priorizaron registros indexados en PubMed/MEDLINE y artículos disponibles en revistas de otorrinolaringología, imagen médica, inteligencia artificial y salud digital. Los términos de búsqueda combinaron conceptos equivalentes a “tympanic membrane”, “otoscopy”, “otoendoscopy”, “otitis media”, “tympanic membrane perforation”, “cholesteatoma”, “artificial intelligence”, “deep learning”, “convolutional neural network”, “computer vision”, “segmentation” y “automated diagnosis”.

2.3 Criterios de selección y extracción

Se priorizaron estudios en humanos que emplearon imágenes otoscópicas u otoendoscópicas y

que informaron resultados cuantitativos de clasificación, detección, segmentación o predicción basada en imagen. Se consideraron variables como tamaño del conjunto de datos, categorías diagnósticas, arquitectura de red, estrategia de partición o validación, sensibilidad, especificidad, exactitud, F1, AUROC, Dice e IoU. También se incluyeron revisiones sistemáticas o metaanálisis relevantes para contextualizar la consistencia de la evidencia. Se excluyeron trabajos puramente técnicos sin relación directa con patología timpánica clínica o sin información suficiente para interpretar el desempeño.

2.4 Enfoque de síntesis

Los hallazgos se organizaron en cinco dominios: clasificación diagnóstica, segmentación y cuantificación, desempeño frente a clínicos, generalización externa e implementación en dispositivos móviles/telemedicina. Dado el carácter narrativo de la revisión, no se realizó metaanálisis de novo ni evaluación formal de riesgo de sesgo; en cambio, se priorizó la interpretación de amenazas recurrentes a la validez, como



sobreajuste, fuga de datos, sesgo de selección, referencia diagnóstica insuficiente y cambio de dominio entre centros o dispositivos.

3. Resultados

3.1 Evolución desde la clasificación básica hacia modelos multiclase

Los primeros sistemas modernos demostraron que las CNN podían reconocer la lateralidad timpánica y detectar perforaciones con rendimiento clínicamente relevante. Lee y colaboradores reportaron una exactitud de 91,0% para identificar perforación timpánica mediante CNN [3]. En 2020, Khan et al. entrenaron modelos sobre 2.484 imágenes otoendoscópicas clasificadas como normal, otitis media crónica con perforación y otitis media con efusión, alcanzando aproximadamente 95% de exactitud y un AUROC medio cercano a 0,99 [5]. Estos trabajos establecieron que la morfología timpánica contiene señales visuales suficientes para una clasificación automatizada robusta en conjuntos controlados.

Posteriormente, Cai et al. evaluaron un sistema de CNN con atención en 6.066 imágenes de 2.022 participantes y cuatro categorías timpánicas. El modelo basado en ResNet50 alcanzó 93,4% de exactitud global en validación cruzada; el F1 fue 94,3% para membranas normales y 96,8% para otitis media con efusión, mientras que la diferenciación entre estados de otitis media crónica fue más exigente [8]. La utilización de mapas de activación permitió además verificar que el algoritmo concentraba la atención en regiones anatómicamente relevantes.

Byun et al. desarrollaron una red ResNet18 + Shuffle con 2.272 imágenes de membrana timpánica, clasificando normalidad, otitis media con efusión, otitis media crónica y colesteatoma. La mejor exactitud alcanzó 97,18%, y el apoyo del modelo mejoró el rendimiento de médicos residentes en hasta 18,4 puntos porcentuales, lo que respalda una función de asistencia más que de sustitución del especialista [10]. En paralelo, Tsutsumi et al. compararon varias arquitecturas sobre 400 imágenes y encontraron que MobileNetV2 alcanzó un

AUROC de 0,902 para la clasificación binaria normal/anormal [11].

3.2 Sistemas móviles y clasificación de múltiples patologías

La integración con teléfonos inteligentes representa una de las líneas más atractivas para trasladar estos modelos fuera de centros especializados. Chen et al. desarrollaron un sistema basado en aprendizaje por transferencia para diez categorías de enfermedad de oído medio a partir de 2.820 imágenes clínicas. Tras optimización, el rendimiento alcanzó 98,0% de exactitud y la aplicación móvil mostró una exactitud de 97,6%; además, superó a médicos generales y residentes en una prueba de interpretación de imágenes [13]. Aunque estos resultados son prometedores, fueron obtenidos dentro de un ecosistema de datos y dispositivos relativamente controlado.

Un avance de mayor escala fue publicado por Dubois et al. en 2024. El algoritmo Inception-v2 se entrenó con 41.664 imágenes y fue evaluado en 3.962 imágenes de validación y

un conjunto de prueba independiente temporal de 326 imágenes. En el conjunto de prueba, la clasificación normal versus anormal alcanzó una sensibilidad de 99,0% y una especificidad de 95,2%. Sin embargo, la sensibilidad para categorías específicas osciló entre 33,3% y 100%, demostrando que un excelente rendimiento binario no implica una identificación multiclase uniforme [20]. Este hallazgo es especialmente relevante para la práctica clínica, donde los errores entre perforación, otitis externa, timpanoesclerosis, injertos o cavidades posquirúrgicas pueden modificar la conducta.

3.3 Segmentación automática y cuantificación de lesiones

La clasificación responde a la pregunta “qué diagnóstico parece mostrar la imagen”, mientras que la segmentación intenta delimitar “dónde está la estructura o lesión”. Esta segunda tarea es crucial cuando se pretende cuantificar una perforación o extraer características morfológicas reproducibles. Pham et al. propusieron EAR-UNet, una arquitectura que integra EfficientNet, puertas de atención y bloques residuales, evaluada en 1.012



imágenes. El coeficiente Dice medio para segmentación de la membrana timpánica fue 0,929, mostrando una delineación anatómica consistente sin preprocesamiento o posprocesamiento manual complejo [9].

Nwosu et al. llevaron esta lógica a la medición de perforaciones. En un estudio de prueba de concepto, un modelo de visión por computador calculó el tamaño relativo de la perforación respecto a la membrana visible. El error promedio del algoritmo fue una sobreestimación de 0,8%, frente a 11,0% en la estimación visual de otorrinolaringólogos encuestados [19]. Aunque el conjunto de perforaciones fue pequeño y relativamente homogéneo, el estudio ilustra el valor de transformar la otoscopia desde una apreciación cualitativa hacia biomarcadores de imagen cuantificables.

En 2026, Liu et al. utilizaron Mask R-CNN para segmentar membrana y perforación en 1.239 imágenes y derivar variables morfológicas con las que se estimó la brecha aéreo-ósea. Los coeficientes Dice fueron 0,875 para la membrana y 0,894

para la perforación, y los modelos de predicción auditiva mostraron rendimiento comparable a mediciones basadas en anotación manual [22]. La misma línea de investigación se ha extendido a modelos que integran características de imagen y variables clínicas para predecir resultados auditivos después de timpanoplastia tipo I [23].

3.4 Colesteatoma: una aplicación de alto valor clínico

El colesteatoma constituye un objetivo especialmente relevante porque puede pasar inadvertido en la exploración inicial y su retraso diagnóstico se asocia con erosión ósea y complicaciones. Tseng et al. analizaron 834 imágenes —197 con colesteatoma, 457 con alteraciones no colesteatomatosas y 180 normales— mediante ocho CNN preentrenadas. DenseNet201 alcanzó 98,5% de exactitud para diferenciar colesteatoma de normalidad, mientras que la discriminación frente a otras membranas anormales fue más difícil, con exactitudes aproximadas de 75,6–90,1% según la arquitectura [17]. Este patrón vuelve a subrayar que el desafío real no es separar una

lesión llamativa de una membrana sana, sino distinguir entre diagnósticos patológicos visualmente parecidos.

3.5 Lo que muestran las revisiones y metaanálisis

La consistencia global de los resultados ha sido analizada en varias síntesis. Habib et al. incluyeron 39 artículos y reportaron una exactitud agrupada de 90,7% para diferenciar imágenes normales de anormales. En estudios con comparación directa, los algoritmos alcanzaron 93,4% de exactitud frente a 73,2% en evaluadores humanos, y las CNN mostraron el mejor rendimiento entre las familias de algoritmos analizadas [14].

Cao et al. incluyeron 16 estudios con 20.254 imágenes timpánicas y estimaron sensibilidad de 93%, especificidad de 85% y AUROC global de 94%. El metaanálisis observó mejor rendimiento con imágenes otoendoscópicas que con otoscopia convencional, lo que sugiere que la calidad de adquisición y el campo visual influyen sustancialmente en la capacidad del algoritmo [15]. Song et al., en otra revisión de 32 estudios, encontraron

una exactitud media de 86% para clasificación basada en imágenes, con un rango amplio de 48,7% a 99,16%, reflejo de la marcada heterogeneidad entre conjuntos de datos, tareas y diseños [18].

3.6 El problema decisivo: generalización externa

La principal barrera para afirmar que estos sistemas están listos para diagnóstico autónomo es la pérdida de rendimiento cuando se aplican a imágenes obtenidas fuera del entorno de entrenamiento. Habib et al. evaluaron 1.842 imágenes procedentes de Turquía, Chile y Estados Unidos. Los modelos alcanzaron un AUROC interno medio de 0,95, pero al probarse en cohortes externas el AUROC descendió a 0,76; la diferencia media fue -0,19 y resultó estadísticamente significativa [16]. Al combinar datos heterogéneos de los tres sitios, el desempeño volvió a mejorar, lo que aporta evidencia directa a favor de conjuntos de entrenamiento multicéntricos y diversos.

Este fenómeno, conocido como domain shift o cambio de dominio, puede originarse por diferencias en iluminación, resolución, tipo de



otoscopio, encuadre, color, limpieza del conducto, prevalencia de diagnósticos, características demográficas o criterios de etiquetado. Por ello, una exactitud interna superior al 95% no debe interpretarse automáticamente como rendimiento clínico generalizable.

3.7 Estudios representativos

Tabla 1. Estudios representativos de deep learning aplicado a otoscopia y membrana timpánica

Estudio	Imágenes / muestra	Tarea principal	Modelo / enfoque	Resultado destacado
Lee et al., 2019 [3]	Imágenes timpánicas clínicas	Perforación y lateralidad	CNN	91,0% para perforación; 97,9% para lateralidad
Khan et al., 2020 [5]	2.484 imágenes	Normal vs COM con perforación vs OME	DenseNet/CNN	≈95% de exactitud; AUROC medio ≈0,99
Cai et al., 2021 [8]	6.066 imágenes; 2.022 participantes	Normal, OME y estados de CSOM	ResNet50 atención/CAM	+ 93,4% exactitud global
Pham et al., 2021 [9]	1.012 imágenes	Segmentación de membrana timpánica	EAR-UNet	Dice medio 0,929
Byun et al., 2021 [10]	2.272 imágenes	Normal, OME, COM, colesteatoma	ResNet18 + Shuffle	97,18% mejor exactitud; mejora de residentes hasta 18,4 pp
Chen et al., 2022 [13]	2.820 imágenes	10 categorías + pass/refer	Ensamble CNN móvil	98,0% tras optimización; app 97,6%
Habib et al., 2023 [16]	1.842 imágenes, 3 países	Normal vs anormal; generalización	Deep learning multicohorte	AUROC 0,95 interno vs 0,76 externo
Tseng et al., 2023 [17]	834 imágenes	Detección de colesteatoma	8 CNN preentrenadas	Hasta 98,5% colesteatoma vs normal
Nwosu et al., 2024 [19]	Prueba de concepto	Segmentación y tamaño de perforación	Visión por computador	Error 0,8% vs 11,0% en estimación clínica Normal/anormal:
Dubois et al., 2024 [20]	41.664 entrenamiento; 326 prueba	11 clases; sistema móvil	Inception-v2	sensibilidad 99,0%, especificidad 95,2% en prueba
Guo et al., 2025 [21]	819 imágenes	Normal, AOM, OME, CSOM	5 CNN; VGGNet-19 mejor	94,51% exactitud; 94,18% sensibilidad; 98,10% especificidad
Liu et al., 2026 [22]	1.239 imágenes	Segmentación + predicción de brecha aéreo-ósea	Mask R-CNN + modelos de regresión	Dice 0,875 TM y 0,894 perforación

Abreviaturas: AOM, otitis media aguda; OME, otitis media con efusión; COM/CSOM, otitis media crónica/supurativa crónica; CAM, class activation map; TM, membrana timpánica; AUROC, área bajo la curva ROC; pp, puntos porcentuales.

4. Discusión

4.1 ¿Qué tan cerca está la IA de un diagnóstico automático real?

La literatura demuestra que la información visual de la membrana timpánica es altamente susceptible de aprendizaje por redes profundas. En tareas bien definidas y con imágenes de calidad, las CNN

pueden alcanzar exactitudes comparables a especialistas y superiores a examinadores menos experimentados [10,13,14]. Sin embargo, el término “diagnóstico automático” debe utilizarse con precisión. La mayoría de los modelos analizan una imagen aislada y no incorporan de forma sistemática otalgia, fiebre, duración de síntomas,

antecedentes de cirugía, audiometría, timpanometría, otorrea, edad o evolución clínica. Por tanto, identifican patrones de imagen compatibles con diagnósticos, pero no reproducen todavía el proceso diagnóstico integral de un otorrinolaringólogo.

El uso más defendible en la actualidad es el de sistema de apoyo a la decisión. En este escenario, el algoritmo actúa como segundo lector, herramienta de triage o guía para usuarios con menor experiencia. La evidencia de Byun et al. es especialmente ilustrativa: la mejora de los residentes cuando recibían la salida del modelo sugiere que la combinación humano-IA puede ser clínicamente más valiosa que una competencia directa entre ambos [10].

4.2 La imagen es el dato: calidad, encuadre y dispositivo

En otoscopia, pequeñas diferencias de adquisición pueden modificar de forma importante la apariencia de la membrana. Cerumen residual, reflejos especulares, desenfoque, baja luminosidad, exposición parcial del tímpano, ángulo oblicuo o variación del balance de blancos

pueden convertirse en fuentes de error. Los estudios de mayor rendimiento suelen aplicar filtros de calidad y excluir imágenes borrosas o insuficientes, una práctica razonable para desarrollo algorítmico pero que puede sobreestimar el rendimiento en condiciones reales [20].

Los sistemas futuros deberían incorporar un módulo previo de control de calidad que determine si una imagen es interpretable antes de emitir una clasificación. Asimismo, la normalización entre dispositivos y la inclusión deliberada de imágenes imperfectas son necesarias para evitar que el algoritmo solo funcione con un otoscopio, un operador o una configuración lumínica específica.

4.3 Sesgo de conjunto de datos y fuga de información

Los conjuntos otoscópicos contienen múltiples imágenes por paciente, oído o sesión. Si fotografías casi idénticas de un mismo paciente aparecen simultáneamente en entrenamiento y prueba, el desempeño puede inflarse por fuga de información. La partición a nivel de paciente —y no únicamente de imagen— debe considerarse un



requisito metodológico. Los conjuntos de prueba temporales o provenientes de otros centros ofrecen una estimación mucho más exigente y clínicamente útil de la generalización [16,20].

Otro problema es el desequilibrio de clases. Patologías frecuentes como OME o tapón de cerumen pueden estar bien representadas, mientras que injertos, colesteatoma o cuerpos extraños son menos numerosos. Un modelo puede alcanzar elevada exactitud global y aun así tener sensibilidad insuficiente para una clase clínicamente relevante. Por ello, deben informarse métricas por clase, intervalos de confianza, matrices de confusión y curvas de calibración, evitando depender exclusivamente de la exactitud total.

4.4 Ground truth: ¿quién define el diagnóstico correcto?

El estándar de referencia es otra fuente crítica de variabilidad. Muchos estudios utilizan la etiqueta asignada por un único otorrinolaringólogo a partir de la imagen o de la consulta clínica. Este enfoque puede ser razonable, pero no elimina la posibilidad de error de etiquetado. Para OME, la incorporación de

timpanometría o hallazgos intraoperatorios puede reforzar la referencia; para colesteatoma, la confirmación quirúrgica representa un estándar particularmente valioso. Los estudios futuros deberían recurrir, cuando sea posible, a consenso de varios especialistas o a estándares complementarios objetivos.

4.5 Explicabilidad y confianza clínica

Los mapas de activación de clase (CAM/Grad-CAM) permiten visualizar las regiones que más influyen en una predicción. En otoscopia son útiles para comprobar que la red presta atención a la membrana y no a artefactos periféricos, texto, marco del endoscopio o características irrelevantes. No obstante, una visualización plausible no demuestra por sí sola causalidad ni seguridad. La explicabilidad debería combinarse con estimaciones de incertidumbre y mecanismos de abstención, de modo que el sistema pueda declarar “imagen no interpretable” o “confianza insuficiente” en vez de forzar un diagnóstico.

4.6 Tele-otoscopia, atención primaria y entornos con recursos limitados

El mayor impacto potencial puede producirse en lugares donde el acceso a otorrinolaringología es limitado. Un otoscopio digital conectado a un teléfono inteligente podría capturar la imagen, evaluar su calidad y generar una recomendación de triage —por ejemplo, normal, manejo inicial/reevaluación o referencia prioritaria— antes de la revisión remota por un especialista [13,20]. Este enfoque puede disminuir derivaciones innecesarias, facilitar programas escolares de tamizaje y ampliar cobertura en zonas rurales.

Sin embargo, el beneficio dependerá de que los modelos se entrenen con poblaciones y dispositivos similares a aquellos donde serán utilizados. Implementar un modelo desarrollado en un centro terciario de alta tecnología en comunidades con cámaras de menor resolución, diferente prevalencia de patología o imágenes tomadas por personal no especializado puede producir errores sistemáticos. Por ello, la validación local y la vigilancia

posterior a la implementación son indispensables.

4.7 De la clasificación al fenotipo cuantitativo

La evolución desde clasificación hacia segmentación abre una ruta particularmente interesante para la imagen otológica. Medir automáticamente área de perforación, cuadrante afectado, distancia al anulus, relación con el manubrio del martillo o cambios de color y textura puede generar biomarcadores objetivos y reproducibles [9,19,22]. Estos datos podrían integrarse con audiometría para predecir pérdida conductiva, con variables perioperatorias para anticipar resultados de timpanoplastia o con series longitudinales para documentar cicatrización.

4.8 Arquitecturas futuras: transformers, video y modelos multimodales

Aunque las CNN siguen dominando la literatura, los Vision Transformers y modelos híbridos ofrecen ventajas potenciales para capturar relaciones globales en la imagen. El video-otoscopio también contiene



información no aprovechada por una fotografía aislada: múltiples ángulos, variación de iluminación y, en algunos sistemas, movilidad timpánica. La selección automática de fotogramas o la creación de imágenes compuestas puede mejorar la robustez.

Una segunda evolución será la multimodalidad. Un modelo que

combine imagen otoscópica con edad, síntomas, timpanometría, audiometría, antecedentes y evolución podría aproximarse más al razonamiento clínico real. No obstante, aumentar la complejidad exige mayor transparencia, calibración y validación prospectiva antes de modificar decisiones terapéuticas.

Tabla 2. Principales barreras para la implementación clínica y estrategias recomendadas

Barrera	Riesgo clínico/metodológico	Estrategia recomendada
Cambio de dominio entre centros/dispositivos	Caída del rendimiento en imágenes externas	Entrenamiento multicéntrico, validación externa y adaptación de dominio
Imágenes de baja calidad	Predicciones erróneas por desenfoque, cerumen o mala exposición	Módulo automático de control de calidad y opción de abstención
Fuga de datos	Sobreestimación del rendimiento	Partición a nivel de paciente/oído/sesión y prueba temporal o externa
Desequilibrio de clases	Alta exactitud global con baja sensibilidad en patologías raras	Métricas por clase, muestreo equilibrado y enriquecimiento de clases críticas
Estándar de referencia débil	Etiquetas incorrectas o inconsistentes	Consenso de expertos y pruebas complementarias cuando sean apropiadas
Falta de calibración	Probabilidades poco confiables	Curvas de calibración, Brier score y umbrales clínicos predefinidos
Automatización excesiva	Dependencia de IA sin contexto clínico	Modelo como apoyo de decisión, no autoridad autónoma
Poca representatividad	Sesgo por edad, dispositivo, región o población	Datasets diversos, auditoría de subgrupos y vigilancia posimplementación

5. Implicaciones para la práctica clínica

Con la evidencia disponible, un marco prudente para el uso clínico del deep learning en otoscopia sería: (1) verificar automáticamente que la imagen sea técnicamente interpretable; (2) realizar una primera clasificación de normalidad versus anormalidad; (3) cuando la imagen

sea anormal, generar un ranking de diagnósticos posibles y nivel de confianza; (4) destacar visualmente la región relevante; y (5) dejar la decisión final al profesional, integrando síntomas, exploración, pruebas audiológicas y antecedentes.

Para atención primaria, el objetivo más realista a corto plazo no es

distinguir todas las enfermedades del oído medio, sino identificar con alta sensibilidad a quienes necesitan referencia, evitar tratamientos innecesarios y mejorar la documentación de la evolución. Para el otorrinolaringólogo, el valor añadido puede estar en la cuantificación objetiva de lesiones, comparación longitudinal y soporte en tareas repetitivas. En investigación, la segmentación automatizada permite construir fenotipos de imagen de gran escala que antes requerían anotación manual intensiva.

6. Conclusiones

El deep learning ha demostrado una capacidad notable para analizar imágenes de la membrana timpánica, con rendimiento elevado en clasificación de otitis media, perforaciones y colesteatoma, así como en segmentación y cuantificación morfológica. Los mejores estudios muestran que la tecnología puede igualar o superar a evaluadores humanos en tareas específicas y puede mejorar el desempeño de clínicos menos experimentados. Además, la combinación de otoscopia digital y

teléfonos inteligentes abre oportunidades concretas para tele-otoscopia y triage en atención primaria.

No obstante, el principal desafío ya no es demostrar que una red neuronal puede aprender una base de imágenes, sino comprobar que mantiene su rendimiento frente a nuevos dispositivos, operadores, centros y poblaciones. La caída observada entre validación interna y externa constituye una señal clara contra la adopción acrítica de cifras de exactitud muy altas. En el estado actual, estos sistemas deben concebirse como herramientas de apoyo diagnóstico con supervisión profesional.

La siguiente etapa de maduración del campo dependerá de datos multicéntricos, estándares de referencia robustos, modelos calibrados y explicables, validación prospectiva y evaluación de resultados clínicos reales. Bajo estas condiciones, la otoscopia asistida por inteligencia artificial podría convertirse en una extensión reproducible de la imagen otológica, capaz no solo de reconocer lesiones, sino también de cuantificarlas, seguirlas longitudinalmente y apoyar



decisiones clínicas de manera segura.

Bibliografía

1. World Health Organization. World report on hearing. Geneva: WHO; 2021.
2. Schilder AGM, Chonmaitree T, Cripps AW, Rosenfeld RM, Casselbrant ML, Haggard MP, et al. Otitis media. Nat Rev Dis Primers. 2016;2:16063. doi:10.1038/nrdp.2016.63.
3. Lee JY, Choi SH, Chung JW. Automated classification of the tympanic membrane using a convolutional neural network. Appl Sci. 2019;9(9):1827. doi:10.3390/app9091827.
4. Cha D, Pae C, Seong SB, Choi JY, Park HJ. Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database. EBioMedicine. 2019;45:606-614.
5. Khan MA, Kwon S, Choo J, Hong SM, Kang SH, Park IH, et al. Automatic detection of tympanic membrane and middle ear infection from otoendoscopic images via convolutional neural networks. Neural Netw. 2020;126:384-394. doi:10.1016/j.neunet.2020.03.023.
6. Habib AR, Wong E, Sacks R, Singh N. Artificial intelligence to detect tympanic membrane perforations. J Laryngol Otol. 2020;134:311-315.
7. Livingstone D, Chau J. Otoscopic diagnosis using computer vision: an automated machine learning approach. Laryngoscope. 2020;130:1408-1413.
8. Cai Y, Yu JG, Chen Y, Liu C, Xiao L, Grais EM, et al. Investigating the use of a two-stage attention-aware convolutional neural network for the automated diagnosis of otitis media from tympanic membrane images: a prediction model development and validation study. BMJ Open. 2021;11:e041139. doi:10.1136/bmjopen-2020-041139.
9. Pham VT, Tran TT, Wang PC, Chen PY, Lo MT. EAR-UNet: a deep learning-based approach for segmentation of tympanic membranes from otoscopic images. Artif Intell Med. 2021;115:102065. doi:10.1016/j.artmed.2021.10.2065.
10. Byun H, Yu S, Oh J, Bae J, Yoon MS, Lee SH, et al. An assistive role of a machine learning network in diagnosis of middle ear diseases. J Clin Med. 2021;10(15):3198. doi:10.3390/jcm10153198.

11. Tsutsumi K, Goshtasbi K, Risbud A, Khosravi P, Pang JC, Lin HW, et al. A web-based deep learning model for automated diagnosis of otoscopic images. *Otol Neurotol*. 2021;42(9):e1382-e1388. doi:10.1097/MAO.00000000000003210.
12. Crowson MG, Hartnick CJ, Diercks GR, et al. Machine learning for accurate intraoperative pediatric middle ear effusion diagnosis. *Pediatrics*. 2021;147:e2020034546.
13. Chen YC, Chu YC, Huang CY, Lee YT, Lee WY, Hsu CY, et al. Smartphone-based artificial intelligence using a transfer learning algorithm for the detection and diagnosis of middle ear diseases: a retrospective deep learning study. *eClinicalMedicine*. 2022;51:101543. doi:10.1016/j.eclinm.2022.101543.
14. Habib AR, Kajbafzadeh M, Hasan Z, Wong E, Gunasekera H, Perry C, et al. Artificial intelligence to classify ear disease from otoscopy: a systematic review and meta-analysis. *Clin Otolaryngol*. 2022;47:401-413. doi:10.1111/coa.13925.
15. Cao Z, Chen F, Grais EM, Yue F, Cai Y, Swanepoel DW, Zhao F. Machine learning in diagnosing middle ear disorders using tympanic membrane images: a meta-analysis. *Laryngoscope*. 2023;133(4):732-741. doi:10.1002/lary.30291.
16. Habib AR, et al. Evaluating the generalizability of deep learning image classification algorithms to detect middle ear disease using otoscopy. *Sci Rep*. 2023;13. doi:10.1038/s41598-023-31921-0.
17. Tseng CC, Lim V, Jyung RW. Use of artificial intelligence for the diagnosis of cholesteatoma. *Laryngoscope Investig Otolaryngol*. 2023;8(1):201-211. doi:10.1002/lto2.1008.
18. Song D, Kim T, Lee Y, Kim J. Image-based artificial intelligence technology for diagnosing middle ear diseases: a systematic review. *J Clin Med*. 2023;12(18):5831. doi:10.3390/jcm12185831.
19. Nwosu O, Suresh K, Knoll R, Lee DJ, Crowson MG. A proof-of-concept computer vision approach for measurement of tympanic membrane perforations. *Laryngoscope*. 2024;134(6):2906-2911. doi:10.1002/lary.31270.
20. Dubois C, Eigen D, Simon F, Couloigner V, Gormish M, Chalumeau M, et al. Development and validation of



- a smartphone-based deep-learning-enabled system to detect middle-ear conditions in otoscopic images. NPJ Digit Med. 2024;7:162. doi:10.1038/s41746-024-01159-9.
- Artif
Intell.
2020;2(2):e200029.
doi:10.1148/ryai.2020200029.
21. Guo Q, Xie L, Zhou L. Deep learning for otitis media classification using otoscopic image. Medicine (Baltimore). 2025;104(48):e46218. doi:10.1097/MD.00000000000046218.
22. Liu TY, Chang HC, Wang PC, Hsu SY, Fang TY, Pham VT, et al. Deep learning-assisted prediction of air-bone gap using tympanic membrane perforation image features. Otolaryngol Head Neck Surg. 2026;174(2):403-412. doi:10.1002/ohn.70074.
23. Liu TY, Chang HC, Wang PC, Hsu SY, Fang TY, Pham VT, et al. Deep learning-assisted prediction of hearing outcomes after anatomically successful type I tympanoplasty. J Otolaryngol Head Neck Surg. 2026;55:19160216261450491. doi:10.1177/19160216261450491.
24. Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. Radiol