Aprendizagem automática por inteligência artificial na avaliação otoscópica da membrana timpânica

Artigo Original

Autores

Rafael Pires
Hospital Garcia de Orta, Unidade Local de Saúde Almada Seixal, Portugal

Rodrigo Pires
Portsmouth Hospitals University NHSTrust, UK

Sammy Gasana
Carnegie Mellon University, USA

Sofia Sousa Teles
Hospital Garcia de Orta, Unidade Local de Saúde Almada Seixal, Portugal

Ana Miguel Couto
Hospital Garcia de Orta, Unidade Local de Saúde Almada Seixal, Portugal

Leonor Oliveira
Hospital Garcia de Orta, Unidade Local de Saúde Almada Seixal, Portugal

Jorge Dentinho
Hospital Garcia de Orta, Unidade Local de Saúde Almada Seixal, Portugal

Luís Antunes
Hospital Garcia de Orta, Unidade Local de Saúde Almada Seixal, Portugal

Correspondência:
Rafael Pires
rafael.pires@ulsas.min-saude.pt

Artigo recebido a 8 de Maio de 2025.
Aceite para publicação a 4 de Novembro de 2025.

Resumo

Objetivos: Desenvolver e avaliar um modelo de aprendizagem para a classificação automatizada de patologias da membrana timpânica e do ouvido médio, utilizando imagens de otoscopia.
Desenho do Estudo: Retrospectivo.
Materiais e Métodos: Rede neural convolucional (CNN) baseada no EfficientNet com 618 imagens classificadas em cinco categorias: membrana timpânica normal, otite média aguda, otite serosa, atelectasia do ouvido médio e perfuração timpânica. Os dados foram divididos em treino (70%), validação (20%) e teste (10%). O desempenho do modelo foi analisado através da precisão, sensibilidade, especificidade, F1 score e área sob a curva ROC (AUC).
Resultados: O modelo demonstrou uma precisão global de 88.98%, sensibilidade de 91.85%, especificidade de 96.37%, F1 score de 0.89 e AUC de 0.975. As visualizações Grad-CAM confirmaram o foco do modelo em áreas clinicamente relevantes.
Conclusões: Este estudo demonstrou que um modelo baseado em CNN é capaz de classificar com precisão patologias comuns do ouvido médio, apresentando potencial como ferramenta de apoio no diagnóstico em contexto de telemedicina e cuidados de saúde primários.
Palavras-chave: Otoscopia; avaliação da membrana timpânica por inteligência artificial; otite média aguda; otite serosa; atelectasia do ouvido médio; perfuração timpânica.

Introduction

Otoscopic examination remains a cornerstone in diagnosing middle ear diseases (MED) such as acute otitis media (AOM), otitis media with effusion (OME), middle ear atelectasis (MEA) and tympanic membrane (TM) perforations. These conditions are prevalent, especially in paediatric populations, and their timely diagnosis is critical for preventing complications and guiding management. Nevertheless, diagnostic variability among practitioners-especially in non-specialist settings—poses challenges to consistent patient care as it depends significantly on the examiner's training and experience.[1]

A notable study reported that the average diagnostic accuracy of video-presented otoendoscopies presenting AOM, OME and normal findings was lower for general practitioners (46%) and paediatricians (51%) compared to otolaryngologists (74%) in the United States.[2]

An automated otoscopic image-based diagnostic technology can serve as a diagnostic aid for clinicians conducting video otoscopies across various healthcare settings, and assist in the decision to refer to a specialist.[1]

Recent advances in machine learning, particularly in deep learning in the domain of computer vision, have enabled the development of automated tools capable of analysing medical images with high precision. Convolutional neural networks (CNNs) are a specific subtype of deep learning models and use spatial and structural information from images to identify patterns, classify objects, and make predictions in various tasks such as image recognition, and object detection.[3] EfficientNet, a family of CNN models developed by Google Research™, is recognised for achieving high performance with relatively low computational cost through compound scaling of depth, width, and resolution.[4]

This study aims to develop, validate and assess the performance of a CNN-based machine learning model (MLM) in classifying video otoscopic and otoendoscopic images from a clinical database. We hypothesise that such a model could reliably distinguish among common TM and middle ear pathologies and support clinical decision-making in telemedicine and primary care settings.

Methods

Data collection

We retrospectively reviewed medical records for patients who had undergone flexible otoendoscopy, rigid otoendoscopy and video otoscopy at Hospital Garcia de Orta between January 2022 and January 2025. Clinical data and associated otoscopic and otoendoscopic images were collected from a storage software system DiVAS (XION GmbH, Berlin, Germany) and a shared cloud-based folder. Images were included if they were from patients with a normal TM or with a diagnosis of AOM, OME, MEA or TM perforation, and provided by otolaryngology specialists and residents. An online publicly available database of otoscopy images from Department of Pediatrics of University of Winsconsin School of Medicine and Public Health, USA,[5] was also consulted to retrieve otoscopic images of patients with diagnosis of AOM and OME. All images were then de-identified and re-categorized by one otolaryngologist to enable cross-checking. Images of a wide range of resolutions, clarity, lighting, and distance from surface were included in model training, validation and testing. Images were excluded from patients with altered anatomy from prior otologic surgery. Images were excluded if the lens was foggy, blurred or obscured by cerumen or otorrhea.

Classification Model

We adopted the EfficientNet-B model, a CNN model proven to have a for high accuracy and efficiency through an effective compound scaling method that uses fewer parameters and computational resource requirements.[4] The model was initialised with weights pretrained on ImageNet and fine-tuned using our labelled otoscopic and otoendoscopic images. Model implementation and training were conducted in Python (version 3.1) using PyTorch (v1.X), on a NVIDIA GeForce RTX 3060 GPU with 16 GB of RAM. The model consisted of 5,288,548 trainable parameters with a model size of approximately 20.17 MB. Stratified random sampling was used during training to ensure balanced class representation in each batch. Training was conducted over 20 epochs with performance monitored on validation and test sets. The dataset was divided into 70% training (432 images), 20% validation (123 images), and 10% test (62 images) splits.

To tailor the EfficientNet model for our tympanic membrane classification task, we replaced its final output layer with a fully connected dense layer, followed by a Rectified Linear Unit (ReLU) activation and a Dropout layer to reduce overfitting. The final classification layer was modified to output five categories: normal TM, MEA, OME, AOM, and TM perforation. Unlike segmentation-based approaches, our model was trained directly on unlabelled otoscopic and otoendoscopic images without pixel-level annotations, to better reflect the type of input typically encountered in primary care settings.

To address class imbalance—particularly for underrepresented categories such as atelectasis and TM perforation—we applied rotational data augmentation and image flipping (both horizontal and vertical), Gaussian blur and sharpness adjustment, enhancing image diversity and model robustness. During training, we ensured equal class representation in each epoch through stratified random sampling. All input images were center-cropped and resized to 224 × 224 pixels, and normalized using ImageNet for RGB channel statistics (mean: [0.485, 0.456, 0.406]; std: [0.229, 0.224, 0.225]). The categorical cross-entropy loss function was used.

Advanced training strategies included the use of a ReduceLROnPlateau scheduler to decrease the learning rate by a factor of 0.15 when validation loss plateaued (patience = 3 epochs), early stopping to prevent overfitting, and gradient clipping (max norm = 1.0) for training stability. The Adam optimizer was used with a learning rate of 0.0001. Cross-validation and repeated runs were performed to ensure model robustness and reproducibility.

Performance Evaluation

In this study, we assessed the performance of our tympanic membrane classification model using widely accepted machine learning metrics to evaluate diagnostic accuracy. The metrics included accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score. These were calculated separately for each of the five diagnostic categories: normal TM, MEA, OME, AOM, and TM perforation.

To ensure reliability, model evaluation was performed over 20 independent epochs on the validation or test set, and the results were aggregated. Sensitivity, specificity, PPV, and NPV were averaged across all diagnostic classes and across the 20 epochs. Overall accuracy, however, was computed directly from the confusion matrix for each run to reflect the true proportion of correct predictions (both true positives and true negatives) relative to the full dataset. The final reported overall accuracy was the average across all 20 epochs for test set, providing a robust measure of the model's general diagnostic performance.

For further evaluation of model discrimination capability, receiver operating characteristic (ROC) curves were plotted by computing the true positive rate (TPR) and false positive rate (FPR) using the following formulas:

- TPR = TP / (TP + FN)
- FPR = FP / (FP + TN)

Where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

The area under the ROC curve (AUC) was calculated for each class using a one-vs-rest binarization approach, suitable for multiclass classification problems. All ROC curves and AUC values were computed in Python.

To gain insight into the model's decision-making process during image classification, we employed the Gradient-weighted Class Activation Mapping (Grad-CAM) technique using Python. Grad-CAM generates heatmaps that highlight the specific regions of each otoscopic image that most influenced the model's prediction. This visualization method allowed us to assess whether the model was focusing on clinically relevant areas of the tympanic membrane, such as regions of perforation, signs of effusion, retraction, or inflammation. These attention maps were qualitatively reviewed during training to confirm that the model's predictive focus aligned with anatomical and pathological features typically used by otolaryngologists in diagnostic assessment. To gain insight into the model's decision-making process, we applied Grad-CAM (gradient-weighted class activation mapping). This technique produces attention heatmaps that highlight the image regions most influential to the model's classification decision. By reviewing these visualizations, we confirmed that the model's focus corresponded to medically relevant features of the TM, such as perforation margins, retraction patterns, effusion opacities, and inflammatory signs, thereby supporting the clinical validity of its predictions.

Results

A total of 618 otoscopic images were obtained. The dataset included 235 images of normal TM, 141 TM perforation, 82 MEA, 95 OME and 62 OMA. Over the course of 20 training epochs, the EfficientNet-based classification model demonstrated steady improvements in accuracy, F1 score, and ROC-AUC across the training, validation, and test sets (figure 1). Training loss decreased from 1.21 (epoch 1) to 0.11 (epoch 19), while training accuracy increased from 57.9% to 96.0%, and F1 score from 0.52 to 0.96, indicating strong learning capacity. Similarly, the training ROC-AUC rose from 0.82 to 0.998, reflecting excellent class separability. The model also performed well on unseen data. Validation accuracy peaked at 81.8% (epoch 11), with a corresponding F1 score of 0.81 and ROC-AUC of 0.94. On the test set, the model achieved a final accuracy of 88.98%, F1 score of 0.89, and ROC-AUC of 0.977 at epoch 20.

Despite this, mild overfitting was observed after epoch 14, as training accuracy and loss continued to improve while validation accuracy plateaued or slightly declined (e.g., from 81.8% at epoch 11 to 72.7% at epoch 20). Validation loss also fluctuated, suggesting a growing gap between training and validation performance. Nonetheless, ROC-AUC values remained high across all sets, indicating that classification confidence remained stable.

Table 1 reveals a class-wise diagnostic metrics analysis on the test set. The model performed well in detecting AOM (100.00% sensitivity)

Figure 1
Area under the receiver-operating characteristic curve (ROC AUC) for correct otoscopic diagnosis classification over 20 epochs in training, validation and test sets.

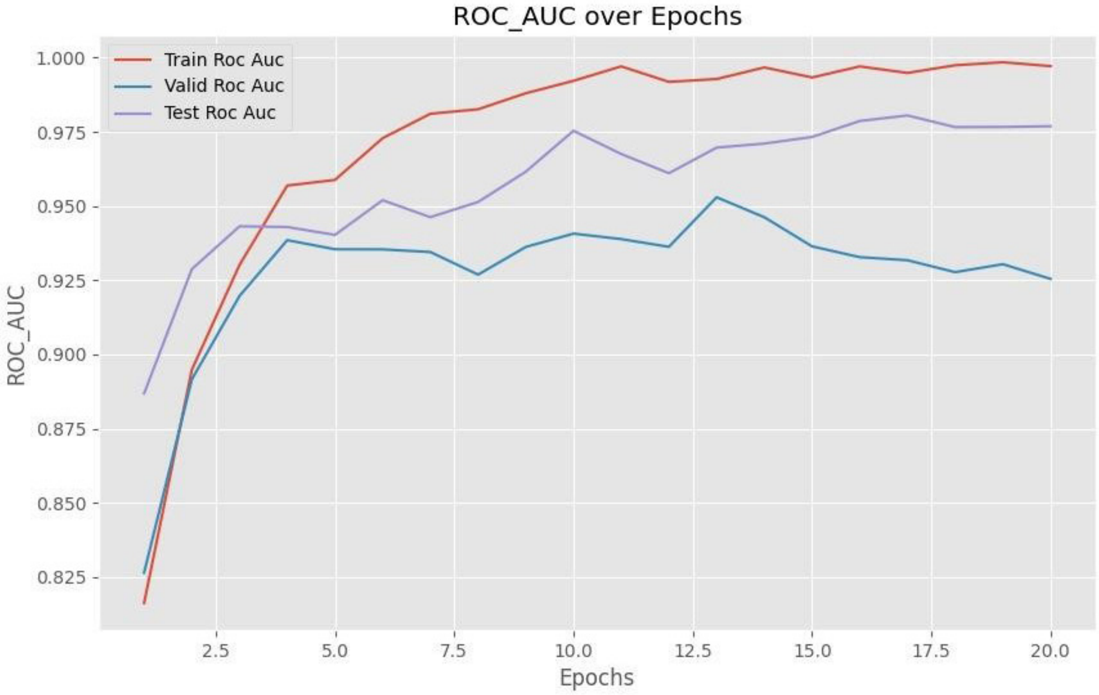

Table 1
Average of performance metrics - sensitivity, specificity, PPV and NPV - for test set.

Average	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
Normal	95.43	97.56	95.65	94.11
AOM	100.00	98.30	87.50	100.00
OME	78.97	92.70	82.89	92.98
MEA	90.88	96.64	91.65	97.78
TM Perforation	92.58	98.00	91.78	92.25

Table 2
Overall average performance metrics for test set over 20 epochs

	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F1-score	Accuracy (%)	ROC AUC
Overall	91.65	96.37	89.21	95.02	0.89	88.98	0.975

and normal TM (95.43% sensitivity). Diagnostic performance was also strong for TM perforation. However, sensitivity was lower for OME (78.97%), a more visually subtle condition, which aligns with its higher confusion rate in the classification results.

Table 2 shows overall average performance metrics for test set over 20 epochs, with and overall average sensitivity of 91.65%, specificity of 96.37% and accuracy of 88.98%.

Discussion

In this study, we demonstrate that machine learning can be used to develop models for the detection and classification of MED, based on the average overall accuracy and sensitivity of our model (91.65% and 88.98%, respectively). A meta-analysis by Cao *et al* (2023)[1] with 16 studies encompassing 20,254 tympanic membrane images, demonstrated a combined sensitivity and specificity for applying ML approaches to diagnose MED in validation or test datasets of 93% and 85%, respectively, an AUC of 94% and an accuracy that ranged from 76% to 98%. In our investigation, we were able to obtain higher overall specificity and AUC ROC values for testing than the majority of these studies.

While several studies have explored AI-based classification of tympanic membrane conditions, our study offers additional value in several key areas. First, we demonstrate that deep learning can be effectively applied to classify otoscopic images into five clinically MED using a compact and efficient CNN architecture. Our EfficientNet-based model achieved high training, validation and test accuracy, indicating strong potential for integration into clinical workflows, particularly in primary care settings.

Importantly, our model was trained on a dataset that deliberately included a wide range of image quality, simulating the variability encountered in real-world practice—such as differences in lighting, focus, framing, and TM visibility. This approach contrasts with prior studies that often relied on highly curated, high-quality image sets, and may therefore limit generalisability. By embracing this variability, our model may perform more robustly when deployed in everyday clinical environments where image quality cannot be guaranteed.

We opted to use a classification model rather than a segmentation-based approach, as this reflects the current practical needs of clinicians in primary care, where a fast, point-of-care decision tool is preferable to more computationally intensive image analysis. That said, segmentation models—capable of highlighting specific regions of interest like perforation margins or effusion shadows—may further enhance interpretability and diagnostic precision, and represent a natural next step for future work.

Our findings also highlighted the specific diagnostic challenges faced by the model. For instance, atelectasis was occasionally misclassified as OME, and this latter was sometimes confused with normal TM—likely due to subtle and overlapping visual features. These misclassifications underscore the importance of standardised image acquisition protocols, including consistent positioning, framing, and lighting during otoendoscopy, which could enhance model accuracy.

Although interactive detection systems were developed as a real-time diagnostic supporting tool for classifying MED in two studies[6,7], to the best of our knowledge, no published study has yet tested such an AI model in a real-world clinical setting. This is an essential future step we aim to pursue: implementing and evaluating our model within primary care or urgent care environments to assess usability, diagnostic impact, and potential to reduce unnecessary referrals to ENT specialists.

Several limitations of this investigation should be acknowledged. Our dataset was of relatively small size and derived from a single institution, which may introduce selection bias and limit generalisability. While the inclusion of varied image quality improves validity, it may have modestly impacted model performance. Only five categories of TM conditions were included, limiting validity of comparison to other studies that included more and different diagnoses. Furthermore, our dataset excluded images from patients with prior ear surgery, so model performance in such population remains unknown.

Mild overfitting was also noted when comparing training and validation accuracy and loss curves, suggesting the model began to memorise features specific to the training data rather than learning generalised patterns applicable to unseen cases. These findings highlight the importance of employing overfitting mitigation strategies.

Nonetheless, our study reinforces the feasibility of using AI to automate the classification of middle ear conditions from otoscopic images and offers practical insights into real-world model training and deployment. As further work continues in refining models, expanding datasets, and validating performance across diverse clinical contexts, AI-assisted otoscopy holds strong potential to enhance diagnostic accuracy, triage efficiency, and early intervention in otology.

Conclusion

This study demonstrates the feasibility of using CNN-based AI models to classify otoendoscopic images with high accuracy and promising results in distinguishing common MED. Future work should aim to expand the dataset in both size and diversity, validate model performance in prospective, multi-institutional clinical settings, and explore integration into user-friendly applications that support frontline healthcare providers. Such developments could significantly enhance early diagnosis, streamline referrals, and ultimately improve outcomes in otologic care.

Conflicts of interest

The authors declare that they have no conflict of interest regarding this article.

Data confidentiality

The authors declare that they followed the protocols of their work in publishing patient data.

Human and animal protection

The authors declare that the procedures followed are in accordance with the regulations established by the directors of the Commission for Clinical Research and Ethics and in accordance with the Declaration of Helsinki of the World Medical Association. Privacy policy, informed consent and Ethics committee authorisation. The authors declare that they have obtained signed consent from the participants and that they have local ethical approval to carry out this work.

Financial support

This work did not receive any grant contribution, funding or scholarship.

Scientific data availability

There are no publicly available datasets related to this work.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this manuscript, the author(s) used ChatGPT (OpenAI) for assistance with organizing sections of the manuscript, particularly the Methods and Discussion sections, and to improve language and readibility of the text. After using this tool, the authors thoroughly reviewed and edited the content as needed and assume full responsibility for the final version of the manuscript.

References

1. Cao Z, Chen F, Grais EM, Yue F, Cai Y, Swanepoel W. et al. Machine learning in diagnosing middle ear disorders using tympanic membrane images: a meta-analysis. Laryngoscope. 2023 Apr;133(4):732-741. doi: 10.1002/lary.30291.

2. Pichichero ME, Poole MD. Comparison of performance by otolaryngologists, pediatricians, and general practitioners on an otoendoscopic diagnostic video examination. Int J Pediatr Otorhinolaryngol. 2005 Mar;69(3):361-6. doi: 10.1016/j.ijporl.2004.10.013.

3. Zhao X, Wang L, Zhang Y, Han X, Deveci M, Parmar M. A review of convolutional neural networks in computer vision. Artif Intell Rev. 2024;57(99): 1-43. doi:10.1007/s10462-024-10721-6

4. Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning. 2019;97:6105–14. doi:10.48550/arXiv.1905.11946

5. University of Wisconsin–Madison Department of Pediatrics. Images – Acute Otitis Media Exercises [Internet]. Madison (WI): University of Wisconsin School of Medicine and Public Health; [cited 2025 May 4]. Available from: https://www.pediatrics.wisc.edu/education/acute-otitis-media/exercises/images/

6 - Yokananth R, Gosula V. OtoVision: bridging machine learning and edge computing for effective and affordable ear disease diagnosis. Neural Comput Appl. 2025;37(3):1565–73. doi:10.1007/s00521-024-10426-5.

7. Zeng X, Jiang Z, Luo W, Li H, Li H, Li G. et al. Efficient and accurate identification of ear diseases using an ensemble deep learning model. Sci Rep. 2021 May 25;11(1):10839. doi: 10.1038/s41598-021-90345-w.

8. Habib AR, Wong E, Sacks R, Singh N. Artificial intelligence to detect tympanic membrane perforations. J Laryngol Otol. 2020 Apr;134(4):311-315. doi: 10.1017/S0022215120000717.

9. 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 Sep 7;12(18):5831. doi: 10.3390/jcm12185831.

10. Levi L, Ye K, Fieux M, Renteria A, Lin S, Xing L. et al. Machine learning of endoscopy images to identify, classify, and segment sinonasal masses. Int Forum Allergy Rhinol. 2025 May;15(5):524-535. doi: 10.1002/alr.23525.