The Laryngoscope
© 2022 The Authors. *The Laryngoscope*
published by Wiley Periodicals LLC on
behalf of The American Laryngological,
Rhinological and Otological Society, Inc.

Machine Learning in Diagnosing Middle Ear Disorders Using Tympanic Membrane Images: A Meta-Analysis

Zuwei Cao, MD, PhD; Feifan Chen, MSc; Emad M. Grais, PhD; Fengjuan Yue, MSc; Yuexin Cai, MD, PhD; De Wet Swanepoel, PhD; Fei Zhao, MD, PhD

Objective: To systematically evaluate the development of Machine Learning (ML) models and compare their diagnostic accuracy for the classification of Middle Ear Disorders (MED) using Tympanic Membrane (TM) images.

Methods: PubMed, EMBASE, CINAHL, and CENTRAL were searched up until November 30, 2021. Studies on the development of ML approaches for diagnosing MED using TM images were selected according to the inclusion criteria. PRISMA guidelines were followed with study design, analysis method, and outcomes extracted. Sensitivity, specificity, and area under the curve (AUC) were used to summarize the performance metrics of the meta-analysis. Risk of Bias was assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 tool in combination with the Prediction Model Risk of Bias Assessment Tool.

Results: Sixteen studies were included, encompassing 20254 TM images (7025 normal TM and 13229 MED). The sample size ranged from 45 to 6066 per study. The accuracy of the 25 included ML approaches ranged from 76.00% to 98.26%. Eleven studies (68.8%) were rated as having a low risk of bias, with the reference standard as the major domain of high risk of bias (37.5%). Sensitivity and specificity were 93% (95% CI, 90%–95%) and 85% (95% CI, 82%–88%), respectively. The AUC of total TM images was 94% (95% CI, 91%–96%). The greater AUC was found using otoendoscopic images than otoscopic images.

Conclusions: ML approaches perform robustly in distinguishing between normal ears and MED, however, it is proposed that a standardized TM image acquisition and annotation protocol should be developed.

Key Words: otoscopy, tympanic membrane, artificial intelligence, machine learning, deep learning, middle ear disorders, otitis media, hearing healthcare.

Level of Evidence: Not Applicable

***Laryngoscope*, 00:1–10, 2022**

INTRODUCTION

Middle Ear Disorder (MED) refers to conditions that disturb the normal function of the middle ear and are mainly caused by inflammation or trauma.[1] It is the most common acquired otologic disease, and a significant contributor to the global burden of disease. Otitis media alone affects close to 750 million people annually.[2] MED has a number of forms, for example, tympanic membrane (TM) perforation, acute otitis media (AOM), otitis media with effusion (OME), and chronic otitis media (COM).[3] Without timely diagnosis and treatment, severe and persistent MED can lead to permanent hearing loss, developmental delay in children, and even life-threatening complications.[4]

Otoscopy or otoendoscopy is a clinical examination routinely used by healthcare pro.fessionals, including otologists, audiologists, pediatricians, family practitioners, and those who work in urgent and emergency care services. The image allows visualization of the condition of the ear canal, TM, and middle ear, which facilitates diagnosis of MED.[5,6] However, evidence shows that the rate of correct diagnosis of otitis media by non-

From the Center for Rehabilitative Auditory Research (Z.C.), Guizhou Provincial People's Hospital, Guiyang City, China; Centre for Speech and Language Therapy and Hearing Science, Cardiff School of Sport and Health Sciences (F.C., E.M.G., F.Z.), Cardiff Metropolitan University, Cardiff, UK; Medical Examination Center (F.Y.), Guizhou Provincial People's Hospital, Guiyang City, China; Department of Otolaryngology, Sun Yat-sen Memorial Hospital (Y.C.), Sun Yat-sen University, Guangzhou City, China; and the Department of Speech-Language Pathology and Audiology (D.W.S.), University of Pretoria, Pretoria, South Africa.

Additional supporting information may be found in the online version of this article.

Editor's Note: This Manuscript was accepted for publication on June 21, 2022.

This research was funded by NIHR (AI Award, 02305), Sêr Cymru III Enhancing Competitiveness Infrastructure Award (MA/KW/5554/19), Great Britain Sasakawa Foundation (5826), Cardiff Metropolitan University Research Innovation Award, and The Global Academies Research and Innovation Development Fund, National Natural Science Foundation of China (81860871), Guizhou Provincial Science and Technology Projects (QKHPTRC [2020] 5011). Dr Fei Zhao won Global Academies and Santandar 2021 Fellowship Award.

The authors have no other funding, financial relationships, or conflicts of interest to disclose.

Send correspondence to Fei Zhao, Centre for Speech and Language Therapy and Hearing Science, Cardiff School of Sport and Health Sciences, Cardiff Metropolitan University, Cardiff CF5 2YB, UK.
Email: fzhao@cardiffmet.ac.uk
De Wet Swanepoel, Department of Speech-Language Pathology and Audiology, University of Pretoria, Pretoria, South Africa.
Email: dewet.swanepoel@up.ac.za

DOI: 10.1002/lary.30291

specialist healthcare professionals is lower in comparison to otolaryngologists due to their lack of skills and experience.[7,8] Severely limited resources in terms of ear and hearing specialists in low- and middle-income countries (LMICs) do not help the situation in terms of early diagnosis and intervention for people with MED.[9] Therefore, poor diagnostic accuracy leads to misdiagnosis and subsequent delay in treatment, which may cause preventable complications.[10]

Artificial Intelligence (AI) is a rapidly evolving discipline and has already been used to support and improve health services in many areas.[11] For example, AI applications have successfully provided automatic diagnostic tools for different diseases, being particularly useful in the interpretation of medical images, for example, an AI system for breast cancer screening,[12] using X-ray and computed tomography (CT) scans.[13] AI application in the field of ENT and Audiology has provided automated diagnostic tools for diagnosing MED by analyzing clinical data. This presents its possible implementation in hearing healthcare services as a clinical decision support system. These approaches have significant potential to improve the timely identification and treatment. The majority of these studies have developed machine learning (ML) or deep learning (DL) models for the automated diagnosis of external and middle ear diseases using otoscopic images. For example, Myburgh et al.[9] built a neural network using 389 images to classify 5 categories of video-otoscopic images; normal tympanic membrane, obstructing wax or foreign body in the external ear canal, acute otitis media, OME, and CSOM, achieving a classification accuracy of 86.84%. Additionally, a recent study by Cai et al.[3] used a two-stage convolutional neural network (CNN) method using attention mechanisms for endoscopic image classification. The results showed a classification performance in identifying normal, OME, and COM in active and static stages, equivalent to the diagnostic level of an associate professor in otolaryngology. However, several studies have indicated that the reliability and interpretability of the models need to be further validated and tested in non-specialist hearing healthcare settings. As a result, the proposed research question is; "*What is the diagnostic ability and applicability of ML algorithms in the diagnosis of MED in the real world environment?*" The review with meta-analysis aimed to systematically evaluate the performance of ML models in terms of the diagnostic accuracy in classifying MED from TM images. We also critically appraised model development and the challenges they present.

METHODS

The review was conducted in accordance with the Preferred Reporting Items for a Systematic Review and Meta-Analysis of Diagnostic Test Accuracy Studies (PRISMA-DTA) guidelines.[14,15] According to the list of essential items, a modified PICOS was adopted to raise and formulate the research questions, that is, **P**articipants (adults and children with MED), **I**ndex test and setting (ML models for MED classification based on otoscopic images), **C**omparison (reference standards and target conditions), **O**utcome (diagnosis accuracy of the ML models), **S**tudy design (both prospective and retrospective study designs). The review protocol was registered on the International Perspective Register of Systematic Review (PROSPERO; CRD42021254036).

Search Strategy

A comprehensive search was performed in the bibliographic databases; PubMed, EMBASE, CINAHL and CENTRAL up until the November 30, 2021. We used the combined terms of Medical Subject Headings (MeSH), keywords or free text words to model search strategies. The search strategies were adapted to the requirements of each database. Details for PubMed were: ("artificial intelligence" [Mesh] OR "machine learning" [Mesh] OR "deep learning" [Mesh] OR (artificial [tiab] AND intelligence [tiab]) OR (machine [tiab] AND learning [tiab]) OR (deep [tiab] AND learning [tiab]) OR AI [tiab] OR ML [tiab] OR DL [tiab] OR "computational intelligence" [tiab] OR "computer assisted" [tiab] OR "machine intelligence" [tiab] OR "computer reasoning" [tiab] OR "computer vision system*" [tiab] OR "knowledge acquisition" [tiab] OR "knowledge representation*" [tiab] OR "neural network" [tiab]) AND (diagnos* [tiab] OR detect* [tiab] OR identif* [tiab]) AND ("Tympanic Membrane" [Mesh] OR ("ear drum*" [tiab] or eardrum*[tiab] or tympanic[tiab]) OR "Otitis Media" [Mesh] OR "Otitis Media" [tiab] OR "Glue ear" [tiab] OR AOM [tiab] OR OME [tiab] OR ("Middle Ear" [tiab] AND (Infect* [tiab] OR Inflam* [tiab] OR effusion* [tiab] OR disease* [tiab])) OR ((nonsuppurative[tiab] OR "non suppurative" [tiab] OR "secretory" [tiab] OR muco*[tiab]) AND otitis [tiab])). In addition, manual searches were performed from identified publications to avoid any potential risk.

Eligibility Criteria

Studies on the development of ML approaches for the automatic diagnosis of MED using TM images were retrieved. Prospective and retrospective study designs were eligible. Types of ML approaches used for MED classification were not specifically defined. Sources of TM images used for analysis could be proprietary or open-access. Attempts were made to contact the corresponding authors for additional information if the studies had no extractable numbers of true positive (TP), true negative (TN), false positive (FP), and false negatives (FN). These studies were finally excluded if there was no response from the authors. Studies were excluded if they were comments, letters, case reports, conference abstracts, or animal studies.

Three authors (Z.C., F.C., F.Y.) independently examined the titles and abstracts of studies identified to check relevance. After reviewing the full-text studies meeting the eligibility criteria were identified and included. Any disagreements that could not be solved after discussion led to arbitration by one of the other authors (F.Z., E.M.G.).

Data Extraction and Data Processing

Two authors (Z.C., F.C.) independently used an agreed data collection form to extract data from the included studies. Study design, data source, dataset size, ML approaches, sensitivity, specificity, TP, TN, FP, and FN were extracted from the validated or test datasets for each ML model. Sensitivity, specificity and AUC were then calculated. A large heterogeneity was found in the included literature in terms of types of MED. The ML approaches also varied considerably in methods of data processing and statistical analysis, and consequently, the accuracy in diagnosing or classifying MED by the different ML approaches was critically analyzed. Moreover, sensitivity and specificity together with AUC were used as the main performance metrics of the meta-analysis. The extracted data were double checked for accuracy and any disagreements after discussion were arbitrated by one of the other authors (F.Z., E. M.G.).

Quality Assessment

Risk of Bias was assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2)[16] tool combined with the Prediction Model Risk of Bias Assessment Tool (PROBAST)[17] (Supplement S1). Three authors (Z.C., F.C., F.Y.) independently used the tool to assess the risk of bias in the included studies. One of the other authors (F.Z., C.X., D.W.S.) arbitrated any disagreements after discussion.

Data Synthesis and Statistical Analysis

A bivariate random-effects regression model was constructed and performed using the MIDAS module for Stata 16.0.[18] The meta-analysis of diagnostic test performance was undertaken to obtain combined sensitivity and specificity. A summary receiver operating characteristic (SROC) analysis was carried out to assess the diagnostic accuracy based on sensitivity and specificity obtained from the meta-analysis. Heterogeneity across studies was assessed using the Q and I^2 statistic.[19] $I^2 > 50\%$ was considered sustainable heterogeneity and then further analyzed to investigate the possible source of heterogeneity using meta-regression analysis. In addition, the assessment for publication bias was performed using Deeks' funnel-plot asymmetry test.[20]

RESULTS

A total of 725 studies were retrieved with 249 remaining after the removal of duplicates. A further 156 were removed after screening of titles and abstracts. Ninety-three records were read full text and reference checked for extra articles. Twenty-five articles met the inclusion criteria for further examination. Of these, 9 were excluded because of missing data for analysis, even after we contacted the authors for additional information. As a result, 16 studies were included in the quality assessment and meta-analysis. Figure 1 shows the study selection process in a PRISMA flow diagram. Sixteen studies[3,6,9,21–33] were eventually included for the quality assessment and meta-analysis with the first reported in 2016.[22] The key characteristics of the included studies are summarized in Table I.

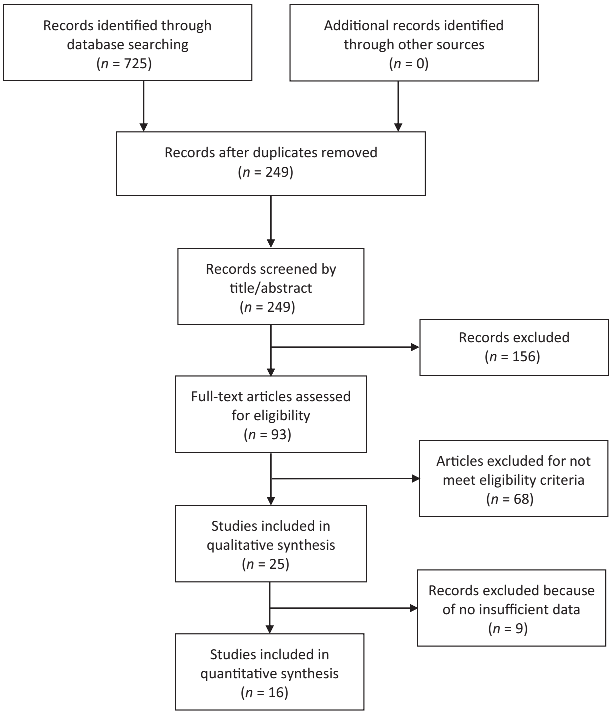
Records identified through database searching (n = 725)
Additional records identified through other sources (n = 0)
Records after duplicates removed (n = 249)
Records screened by title/abstract (n = 249)
Records excluded (n = 156)
Full-text articles assessed for eligibility (n = 93)
Articles excluded for not meet eligibility criteria (n = 68)
Studies included in qualitative synthesis (n = 25)
Records excluded because of no insufficient data (n = 9)
Studies included in quantitative synthesis (n = 16)

Fig. 1. Flow diagram of study selection.

A total of 20,254 TM images were used for ML development and data analysis in the 16 studies. Of these, 7025 were normal TM images and 13,229 images of MED. Data size ranged from 45 to 6066 with a median of 507 per study. It should be noted that the studies did not always distinguish children and adults in their datasets, except for two[25,32] which mention that the data source is based on pediatric patients. Several studies had a very small sample size for example Livingstone et al.[28] with 45 cases (normal TM = 16 and perforations = 29). Indeed, imbalanced datasets across the different types of MED were found in most of the studies that intended to classify the various types of MED. This imbalance between the different classes may lead the classifiers to perform better with the majority class than the minority class.[34]

As shown in Table I, acquisition of the TM images varied from either proprietary or open-access sources (e.g., Google library). In addition, the TM images were taken using different equipment including; smartphone camera, otoscope, and professional otoendoscope. As a result, the original image resolution of the otoscopic images and otoendoscopic images varied from 500 × 500 to 1920 × 1080 pixels. To prepare and augment image data, different methods of image pre-processing were applied, including; cropping, blur detection, and other image augmentation techniques. The different quality of TM images is likely to affect the identification of anatomic structures and their pathological characteristics. Therefore, a standardized data acquisition process should be considered in future studies.

ML Model Development

As shown in Table I, a total of 25 ML models were identified in the 16 included studies. These could be categorized into; general ML approaches (e.g., decision tree, SVM, k-NN), and ML approaches based on DL models. Several recent studies have developed the DL models in combination with attention mechanisms. To resemble the procedure used to identify the significant lesion areas in an otoscopic image, the attention mechanism used in CBAM[21] and Attention Unet[22] guided the network to devote more attention to the important parts of the otoscopic image data by learning which parts of the data are essential in driving the classification decision for the network. These models tend to perform better than general ML/DL models. It should be noted that the reliability of the combined classification method appears

TABLE I.
A Summary of Included Studies with the Model Development and the Performance Evaluation.

Authors	Data Classification and Sample Size	Data Source	Image Pre-processing	Algorithms	Performance Evaluation				
					TP	FP	FN	TN	Accuracy
Alhudhaif 2021, Saudi Arabia	Normal (535) * AOM (119) * CSOM (63) * Earwax (140)	Open access (Otoscopic images)	Cropping Image augmentation (rotate, width shift, height shift, shear, zoom)	CBAM	63	2	1	106	98.26%
Basaran, 2020 Turkey	Normal (154) * AOM (69) * CSOM (14) * Earwax (21) * Myringosclerosis (4) * Tympanostomy tubes (2) * Otitis externa (18)	Publicly available (Otoscopic images)	Image augmentation (rotate and flip)	AlexNet	182	14	48	263	87.77%
				VGG16	199	27	31	250	88.56%
				VGG19	187	28	43	249	86.00%
				GoogLeNet	190	35	40	242	85.21%
				ResNet50	179	33	51	244	83.43%
				ResNet101	179	33	51	244	82.64%
Byun, 2021 Korea	Normal (19) * OME (17) * COM (17) * Cholesteatoma (18)	Proprietary (Otoscopic images)	Cropped and resized to 270 × 270 resolution, Image augmentation (flip, flop, and rotation), Randomly cropping 256 × 256 patches	ResNet18	50	2	0	19	97.18%
Cai, 2020 China	Normal (1040) * OME (2613) * CSOM (2413)	Proprietary (Otoendoscopic images)	Image augmentation (random shift, shear, zoom and flip) Size standardization	ResNet50	4719	307	96	944	93.36%
Cha, 2019 Korea	Normal (4342) * Abnormal (6202) (OME * Perforation * Otitis externa)	Proprietary (Otoendoscopic images)	Image augmentation (random rotation, random shift, random scales, flip)	InceptionV3	1127	113	28	840	93.31%
				ResNet101	1115	123	48	820	91.88%
				Ensemble classifier1	1139	101	22	846	94.17%
Crowson, 2021 USA	Normal (126) * OME (212)	Proprietary (Otoendoscopic images)	Random crops of original images to a minimum scale of 0.15	ResNet34	28	9	2	30	84.06%
Habib, 2020 Australia	Normal (105) * Perforation (128)	Google (Otoscopic images)	Cropping	InceptionV3	19	6	6	19	76.00%
Livingstone, 2019 Canada	Normal (346) * Earwax (63) * Tympanostomy tubes (120)	Proprietary and Google (Otoscopic images)	Image augmentation (rotate and flip)	CNN	23	6	1	15	84.44%
Livingstone, 2020 Canada	Normal (538) * AOM (26) * OME (87) * Earwax (273) * Myringitis (29) * Myringosclerosis (173) * Tympanostomy tubes (260) * Perforation (86) * Cholesteatoma (21) * Otomycosis (13) * Otitis externa (97)	Proprietary and Google (Otoscopic images)	Text or annotations were removal.	Multilabel classifier architecture	79	7	2	11	90.91%
Myburgh, 2016 South Africa	Normal (123) * AOM (80) * OME (80) * CSOM (86) * Earwax (120)	Proprietary (Otoscopic images)	Cropping size standardization	Decision tree	27	9	14	58	78.70%
Myburgh, 2018 South Africa	Normal (123) * AOM (51) OME (69) * CSOM(86) * w/o (60) (Earwax)	Proprietary (Otoscopic images)	Cropping Blur detection	Decision tree	43	10	4	21	82.05%
				Neural network	45	7	3	21	86.84%
Sundgaard, 2021 Denmark	Normal (658) * AOM (145) * OME (533)	Proprietary (Otoscopic images)	Cropping Image augmentation (flip with random erasing)	Deep metric learning	548	130	57	601	86.00%
Uçar, 2021 Turkey	Normal (220) * COM (220) *Earwax (220) * Myringosclerosis(220)	Publicly available (Otoscopic images)	Cropping Blur detection	Bi-LSTM	114	6	6	34	92.50%
Viscaino, 2020 Chile	Normal (220) * CSOM (220) * Earwax (220) * Myringosclerosis (220)	Proprietary (Otoscopic images)	Cropping Blur detection	SVM	107	13	6	34	88.13%
Wu, 2020 China	Normal (3235) * AOM (3355) * OME (4113)	Proprietary (Otoscopic images)	Image augmentation (rotate, width shift, height shift, shear, zoom and flip)	Xception	51	10	4	37	86.27%
				MobileNetV2	50	11	6	35	83.33%

(Continues)

TABLE I.
Continued

Authors	Data Classification and Sample Size	Data Source	Image Pre-processing	Algorithms	Performance Evaluation				
					TP	FP	FN	TN	Accuracy
Zeng, 2021 China	Normal (468) * CME (45) * CSOM (402) * EACB (38) * IC(605) * OE(251) * SOM (272) * TMC (115)	Proprietary (Otoendoscopic images)	Cropping size standardization	Ensemble classifier2	1646	74	33	443	95.13%

FN = false negative; FP = false positive; TN = true negative; TP = true positive.

questionable because the simple aggregation of two independent approaches to illustrate reliability and interpretability is hard to justify with no effective underlying rationale.[3]

Methodological Quality Assessment of Included Studies

Risk of bias was assessed using the combined tool based on QUADAS-2 and PROBAST, as shown in Supplement S1. The methodological quality assessment mainly includes; patient selection, number of participants, index test, internal validation techniques, reference standard, outcome, and applicability domains (Table II). Eleven out of the sixteen studies (68.8%) were marked as low risk of bias, whereas five studies were scored as high risk due to identification of bias in study design, internal validation, reference standard, outcome, and/or applicability domains. As indicated in Supplement S1, the reference standard is considered the method to correctly classify participants as having or not having a target condition. As there was no agreed reference standard to diagnose MED, the reference standard was the major domain of high risk of bias (37.5%, 6/16). Moreover, limited sample size in four studies could result in overfitted approaches, leading to high risk of bias in the analysis domain.[17]

Publication Bias Assessment of Included Studies

As described in the section of "Data synthesis and statistical analysis," the publication bias assessment of the included studies was evaluated using Deeks' funnel-plot asymmetry test. In Figure 2, the DOR is presented in a natural logarithm for the x-axis, and a reciprocal of the square root of the effective sample size (1/√ESS) is displayed on the y- axis. The regression test of asymmetry was conducted using the proposed ML models from the individual studies. The result was not statistically significant ($p = 0.29$) indicating that there is no evidence for publication bias.

Approach Performance Evaluation and Meta-Analysis

In this review, accuracy was reported in all included studies as one of the performance metrics. The accuracy of included approaches ranged from 76.00% to 98.26% with a median of 87.11%. Over 98% accuracy was reached using the CBAM algorithm in the study by Alhudhaif et al.[21] Although the risk of overfitting should be considered, unlike other DL architectures that required a fixed feature size, the approaches detected key points from the TM images, followed by extracting hypercolumn deep features from the ResNet18 approaches.

In the study by Cai et al.,[24] the dataset was significantly larger than the other studies (6066 TM images), and it included major categories of EAC pathologies and MED. The DL approach achieved an accuracy of 94.17% using two of the best-performing approaches (Inception-V3 and ResNet101). In contrast, the low accuracy of 76.00% and 78.70% found in the studies by Habi et al,[26] and Myburgh et al.,[6] were mainly due to small sample size. Therefore, it is useful to have a large dataset when developing a deep network model to identify the features that identify the various MEDs. Cai et al.[3] suggested that the achievement of high accuracy with a relatively small database may be attributed to the combined use of the main classifier and focal classifier when using attention mechanisms.

To evaluate the performance of ML in diagnosing MED, apart from accuracy, different performance evaluation metrics were used. These included; sensitivity,[21,22] specificity,[21,22] *F*-score,[23,25] AUC-ROC[23,25] and in some studies PPV.[22,27,30] As there is no guideline on reporting these diagnostic test accuracy studies using ML approaches, not all studies reported the other evaluation parameters, such as sensitivity, specificity, F-score, or AUC. Therefore, in this review, sensitivity, specificity, and AUC for individual studies without these evaluation metrics were re-calculated on the basis of TP, FP, FN, and TN.

As shown in Figure 3, the forest plot shows an overview of diagnostic test accuracy by different types of ML approaches for the detection of MED. A total of 25 algorithms used in the included studies are summarized. The combined sensitivity and specificity for applying ML approaches to diagnose MED in validation or test datasets were 93% (95% CI, 90%–95%) and 85% (95% CI, 82%–88%), respectively and the AUC was 94% (95% CI, 91%–96%). These results indicate excellent performance of ML approaches in diagnosing MED from TM images.

However, a significant heterogeneity was identified among included algorithms as shown in Figure 3 (sensitivity: $Q = 199.83$, $I^2 = 97.00$, $p = 0.00$; specificity: $Q = 210.78$, $I^2 = 88.61$, $p = 0.00$). Further meta-regression analyses were conducted to explore the potential sources of heterogeneity. The significant sources of heterogeneity in sensitivity and specificity included data size, data quality, data source, classification numbers for

TABLE II.
Quality Assessment Results of the Included Studies According to the Combined Tool Based on QUADAS-2 and PROBAST.

	Participants		Analysis					
	Selection	Number	Index test	Internal validation techniques	Reference Standard	Outcome	Applicability	Overall
Alhudhaif, 2021 Saudi Arabia	–	–	–	–	–	–	–	–
Basaran, 2020 Turkey	–	–	–	–	+	–	–	–
Byun, 2021 Korea	–	+	–	–	–	–	–	–
Cai, 2020 China	–	–	–	–	+	–	–	–
Cha, 2019 Korea	–	–	–	–	–	+	+	+
Crowson, 2021 USA	–	+	–	–	+	–	–	+
Habib, 2020 Australia	+	+	–	+	–	+	+	+
Livingstone, 2019 Canada	–	–	–	–	–	+	+	+
Livingstone, 2020 Canada	+	–	–	+	–	+	+	+
Myburgh, 2016 South Africa	–	+	–	–	–	–	–	–
Myburgh, 2018 South Africa	+	–	–	–	–	–	–	–
Sundgaard, 2021 Denmark	–	–	–	–	+	–	–	–
Uçar, 2021 Turkey	–	–	–	–	+	–	–	–
Viscaino, 2020 Chile	–	–	–	–	+	–	–	–
Wu, 2020 China	–	–	–	–	–	–	–	–
Zeng, 2021 China	–	–	–	–	–	–	–	–

"–" is equal to low risk, "+" is equal to high risk. The overall risk of bias was considered high if at least two categories were at high risk.

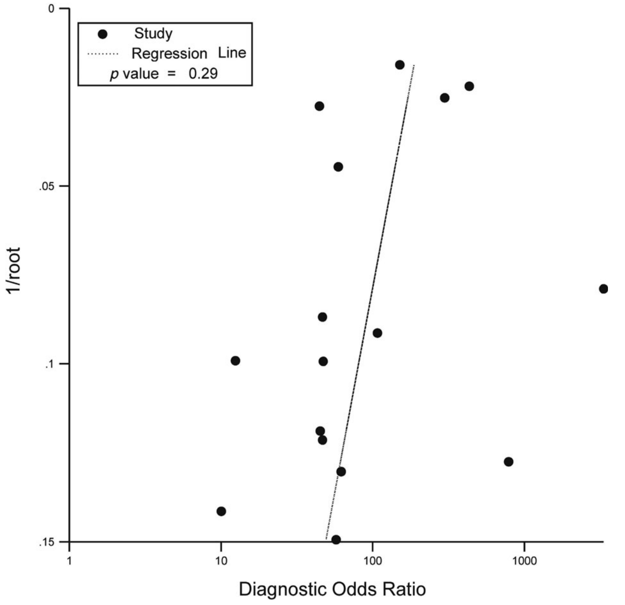

Fig. 2. Deeks' funnel plot to assess the likelihood of publication bias.

MED, reference standard, outcome, and applicability except for patient selection and internal validation techniques (Fig. 4).

Further analysis of the summary operating receiver operation characteristic (SROC) was undertaken to assess performance based on data from a meta-analysis, that is, sensitivity and specificity on the curves and a 95% confidence contour around these points. As shown in Figure 5, the AUC of total TM images was 0.94 (95% CI, 0.94–0.96). However, further analysis revealed that the AUC of the otoendoscopic images (i.e., 0.98, 95% CI: 0.97–0.99) was higher than the AUC of the otoscopic images (i.e., 0.93, 95% CI: 0.91–0.95). This result implies a negative impact on the ML performance when using low-quality TM images obtained from the otoscope.

DISCUSSION

The present systematic review with meta-analysis is a distinctive approach to evaluating the quality and feasibility of ML approaches for the automated diagnosis of MED from TM images. Collectively, ML/DL classification approaches demonstrate excellent accuracy for correctly classifying various types of MED (i.e., 76.00%–98.26%). Moreover, several other metrics of diagnostic performance, for example sensitivity, specificity, and AUC, also suggest good diagnostic performance superior to the accuracy rate of healthcare practitioners.[3,6] For example, ML approaches show better accuracy in diagnosing MED than junior Otolaryngologists in the study by Cai et al. (i.e., 93.36% vs. 79.1–86.6%).[3] In addition, a recent study by Crowson et al.[25] found that diagnostic accuracy for AOM has yet to consistently surpass 70% for primary care providers, pediatricians, and physicians in different disciplines. According to a review from *Nature Reviews Disease Primers*,[5] AOM tends to be overdiagnosed, particularly in the primary care setting, due to difficulties in confirming middle ear effusion.

Although an interactive detection system was developed as a real-time diagnostic supporting tool for classifying MED in the study by Zeng et al.,[33] it should be noted that it has not been used in a real clinical scenario.

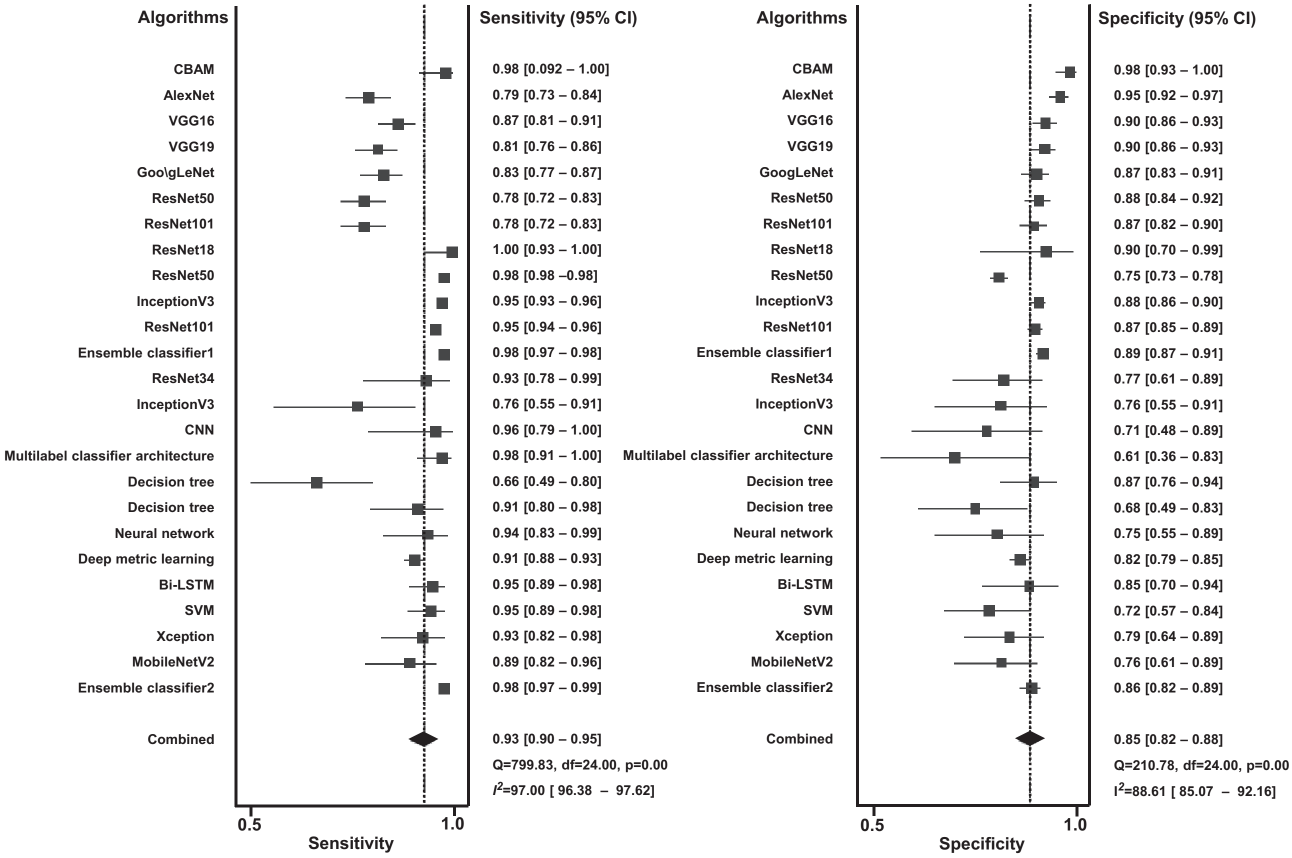

Fig. 3. Forest plot of the pooled sensitivity and specificity for applying ML tools to diagnose MED. The algorithms are in the same order listed in Table I.

Therefore, prospective clinical trials are needed to provide high-quality evidence for applying AI tools as an effective decision-making device used in hearing healthcare.[35,36]

The heterogeneity assessment is a crucial component in meta-analysis, because the presence or absence of true heterogeneity can affect the statistical models to be applied. The current meta-analyses indicate several important sources for the heterogeneity found in sensitivity and specificity, such as data-related issues (i.e., quality, size, and source). There was also considerable heterogeneity caused by the lack of reference standard for MED definitions among the studies, which further limits pooling to assess the classification performance.

Data quality is one of the biggest challenges to the successful development and implementation of AI systems in healthcare. As indicated in this review, although the recent research outcomes show the achievement of a high level of accuracy, the included studies lacked standard protocols in terms of data collection and performance evaluation metrics. Studies that do not meet strict methodological standards usually over-or under-estimate the indicators of test performance as well as limiting the applicability of the results. Therefore, it is important to develop a recommendation criterion for collecting, storing and managing datasets to avoid the influence of data quality.

Several approaches for improving the transparency and the quality in the AI studies of diagnostic accuracy have been suggested, such as Standards for Reporting Diagnostic Accuracy (STARD). The STARD initiative is generally accepted as an important step toward consistency in reporting essential information.[6,24,27] Such efforts provide the best possible evidence for the best patient care. According to the newly published recommendations on the collection and annotation of otoscopy images for intelligent medicine,[37] a standardized data acquisition process is crucial to guarantee the high quality of otoscopic images for developing reliable and comparable ML approaches. It is evidenced by better ML performance using the high-resolution otoendoscopic images in comparison to the use of otoscopic images as in the present study. Moreover, apart from the image resolution, better clarity and scope of the TM is also important to help clearly identify important features of the TM structure and the pathological characteristics. This is often affected by using different otoscopy systems, such as a standard otoscope[26,27,29] and endo-otoscope.[3–5]

A further challenge is that the diagnostic accuracy of the ML approaches is affected by the degree and type of pathological changes in the middle ear. For example, Cai et al.[3] indicated that it is more difficult to distinguish between normal and OME, whereas there was difficulty in classifying the cases of cerumen and tympanostomy tube in the study by Livingstone et al.[24] Therefore, to minimize this challenge, multi-label classifiers trained

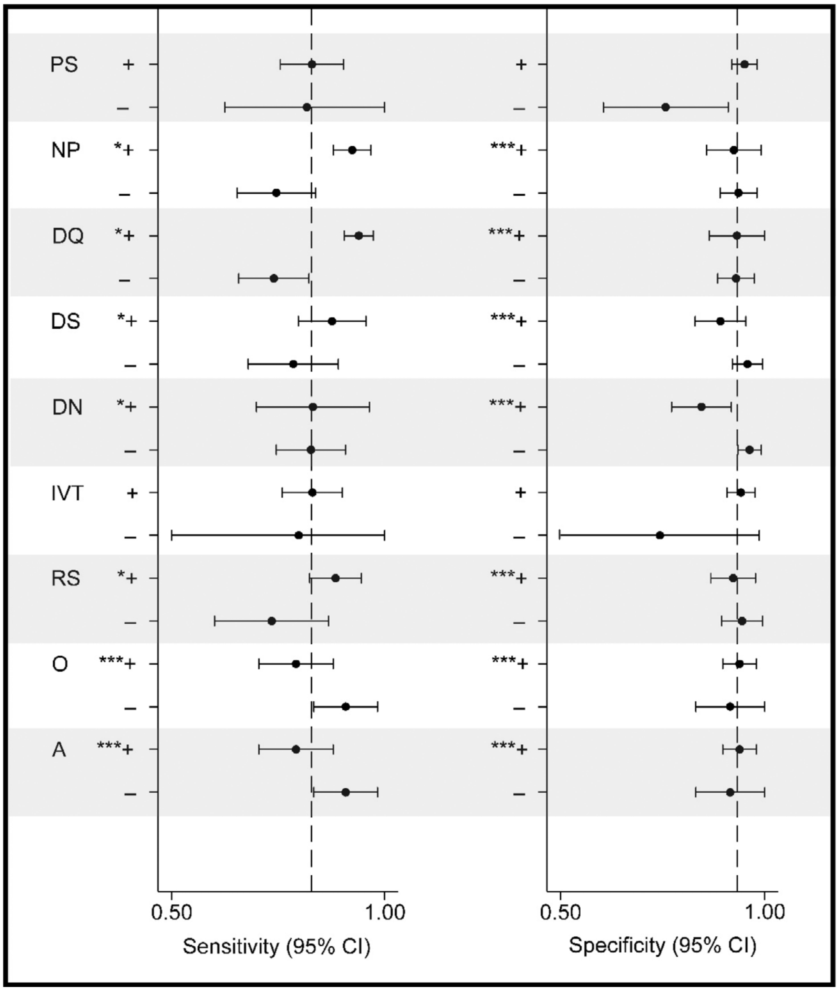

Fig. 4. The results of meta-regression analyses to explore the potential sources of heterogeneity.
"+" is high risk, "–" is low risk. A = applicability; DN = disease number; DQ = data quality; DS = data source; IVT = internal validation techniques; NP = number of participants; O = outcome; PS = patient selection; RS = reference standard. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

using larger training datasets may lead to better classification results, and thus improve accuracy in diagnosing MED using ML approaches.

Not having access to big data is a common challenge when analyzing medical images.[38] As a result, some studies have used augmentation to increase data but clear and rigorous inclusion and exclusion criteria were not reported.[39] To overcome the difficulties of identifying big data, apart from using more data, accuracy and reliability can be further improved by using advanced ML techniques, such as Transfer learning (TF), Data augmentation, and Few-shot learning. The building of powerful CNN models will facilitate the best performance in automated diagnosis using small datasets.[39] Nie et al.[40] used transfer learning tools to analyze wideband absorbance immittance data obtained from a small group of patients with otosclerosis ($n = 135$), and achieved excellent performance in terms of accuracy (94%). Therefore, transfer learning offers an important method for DL applications to medical imaging using various pre-trained CNN models.[41,42]

It should be noted that the global burden of hearing loss is higher than ever and is growing persistently. Ensuring access to appropriate hearing health services presents a significant challenge and requires key barriers to be overcome, specifically the limited healthcare infrastructure, availability of routine ENT and Audiological equipment, and the critical shortage of healthcare professionals such as ENT specialists and audiologists in LMICs.[43,44] AI approaches show great potential for improving the delivery of hearing health services in resource-poor settings.[45] However, the endo-otoscope devices used in middle-income and high-income countries are too expensive to be widely available to primary healthcare workers in many low-income countries.

Recent development of a screening device in combination with a smartphone using AI technology for non-specialist healthcare settings is suggested as an effective diagnostic tool in LMICs. For example, AI offers significant potential for maternal and child health, which is one of the major public health issues in LMICs, such as pregnancy monitoring, prediction of birth asphyxia, mother and/or child malnutrition.[46] Therefore, these successful examples of AI applications endorse the feasibility of improvement in the situation of limited professionals/specialists in hearing healthcare. The future development of a low-cost device

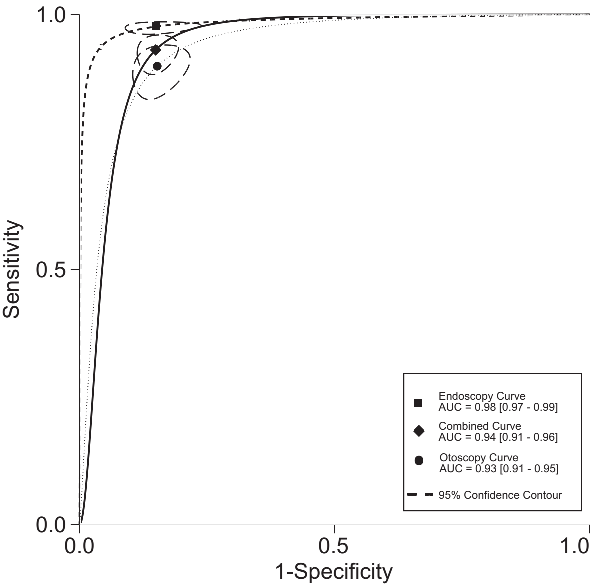

Fig. 5. The plot of summary operating receiver operation characteristic (SROC).

based on a smartphone as an effective and accurate diagnostic tool for MEDs appears urgently needed to provide solutions to the challenges of shortage of specialist training, unaffordable equipment, and un-sustainable hearing care services in LMICs.[46] However, AI diagnostic support systems will nevertheless need to be validated in terms of the accuracy and reliability in house as well as primary care services where otolaryngology referral may be challenging, such as in rural areas, before they are implemented into low-resourced environments.

Limitations and Future Studies

As there is no standardized protocol for the development of AI diagnostic tools using TM images at present, this review finds a large heterogeneity in included studies, in terms of data source and data quality. As a result, data is only classified as to whether there was middle ear disease. It was not possible to further analyze the subsets of individual MED. Therefore, future studies should include (1) development of a guideline on minimum standards for TM images data collection, for example specifications of equipment, and TM image resolution; (2) extracted TM features for individual middle ear pathologies; (3) professional qualification, and (4) assessment criteria on how to classify poor quality data.

CONCLUSION

ML approaches analyzing TM images can diagnose MED with high levels of sensitivity and specificity. The ML approaches demonstrate significant potential for improved access to early diagnosis and timely treatment of MED. However, a standardized TM images acquisition and annotation protocol should be developed, which will further enhance applications of DL approaches using big dataset and high-quality otoscopic images. In the meantime, to minimize the influence of the degree and type of pathological changes on the diagnostic accuracy of MED using the ML approaches, the use of an advanced neural network for multi-label classifiers trained using larger training datasets may lead to robust classification results.

BIBLIOGRAPHY

1. Lee JY, Choi S-H, Chung JW. Automated classification of the tympanic membrane using a convolutional neural network. *Appl Sci*. 2019;9:1827.
2. Graydon K, Waterworth C, Miller H, Gunasekera H. Global burden of hearing impairment and ear disease. *J Laryngol Otol*. 2019;133:18-25.
3. Cai Y, Yu JG, Chen Y, 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.
4. World Health Organization. Childhood Hearing Loss: Strategies for Prevention and Care. 2016.
5. Schilder AG, Chonmaitree T, Cripps AW, et al. Otitis media. *Nat Rev Dis Primers*. 2016;2:1-18.
6. Myburgh HC, van Zijl WH, Swanepoel D, Hellstrom S, Laurent C. Otitis media diagnosis for developing countries using tympanic membrane image-analysis. *EBioMedicine*. 2016;5:156-160.
7. Guan Q, Huang Y, Zhong Z, Zheng Z, Zheng L, Yang Y. Diagnose like a radiologist: attention guided convolutional neural network for thorax disease classification. arXiv preprint arXiv:180109927. 2018.
8. Liskowski P, Krawiec K. Segmenting retinal blood vessels with deep neural networks. *IEEE Trans Med Imaging*. 2016;35:2369-2380.
9. Myburgh HC, Jose S, Swanepoel DW, Laurent C. Towards low cost automated smartphone- and cloud-based otitis media diagnosis. *Biomed Signal Process Control*. 2018;39:34-52.
10. Schilder AG, Marom T, Bhutta MF, et al. Panel 7: otitis media: treatment and complications. *Otolaryngol Head Neck Surg*. 2017;156:S88-S105.
11. Shapshay SM. Artificial intelligence: the future of medicine? *JAMA Otolaryngol Head Neck Surg*. 2014;140:191.
12. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577:89-94.
13. Hayden JA, Ogilvie R, Stewart SA, et al. Development of a clinical decision support tool for diagnostic imaging use in patients with low back pain: a study protocol. *Diagn Progn Res*. 2019;3:1-8.
14. McInnes MDF, Moher D, Thombs BD, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. *JAMA*. 2018;319:388-396.
15. Salameh J-P, Bossuyt PM, McGrath TA, et al. Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. *BMJ*. 2020;370:m2632.
16. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155:529-536.
17. Moons KG, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann Intern Med*. 2019;170:W1-W33.
18. Dwamena B. MIDAS: Stata Module for Meta-Analytical Integration of Diagnostic Test Accuracy Studies. 2009.
19. Huedo-Medina TB, Sánchez-Meca J, Marin-Martinez F, Botella J. Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychol Methods*. 2006;11:193-206.
20. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol*. 2005;58:882-893.
21. Alhudhaif A, Comert Z, Polat K. Otitis media detection using tympanic membrane images with a novel multi-class machine learning algorithm. *PeerJ Comput Sci*. 2021;7:e405.
22. Başaran E, Cömert Z, Çelik Y. Convolutional neural network approach for automatic tympanic membrane detection and classification. *Biomed Signal Process Control*. 2020;56:101734.
23. Byun H, Yu S, Oh J, et al. An assistive role of a machine learning network in diagnosis of middle ear diseases. *J Clin Med*. 2021;10:3198.
24. 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.
25. Crowson MG, Hartnick CJ, Diercks GR, et al. Machine learning for accurate intraoperative pediatric middle ear effusion diagnosis. *Pediatrics*. 2021;147:e2020034546.
26. Habib AR, Wong E, Sacks R, Singh N. Artificial intelligence to detect tympanic membrane perforations. *J Laryngol Otol*. 2020;134:311-315.
27. Livingstone D, Chau J. Otoscopic diagnosis using computer vision: an automated machine learning approach. *Laryngoscope*. 2020;130:1408-1413.
28. Livingstone D, Talai AS, Chau J, Forkert ND. Building an Otoscopic screening prototype tool using deep learning. *J Otolaryngol Head Neck Surg*. 2019;48:66.

29. Sundgaard JV, Harte J, Bray P, et al. Deep metric learning for otitis media classification. *Med Image Anal*. 2021;71:102034.
30. Uçar M, Akyol K, Atila Ü, Uçar E. Classification of different tympanic membrane conditions using fused deep hypercolumn features and bidirectional LSTM. *IRBM*. 2021;43:187-197.
31. Viscaino M, Maass JC, Delano PH, Torrente M, Stott C, Auat CF. Computer-aided diagnosis of external and middle ear conditions: a machine learning approach. *PLoS One*. 2020;15:e0229226.
32. Wu Z, Lin Z, Li L, et al. Deep learning for classification of pediatric otitis media. *Laryngoscope*. 2021;131:E2344-E2351.
33. Zeng X, Jiang Z, Luo W, et al. Efficient and accurate identification of ear diseases using an ensemble deep learning model. *Sci Rep*. 2021;11:10839.
34. Barua S, Islam MM, Yao X, Murase K. MWMOTE--majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Trans Knowl Data Eng*. 2012;26:405-425.
35. Bae J, Yu S, Oh J, et al. External validation of deep learning algorithm for detecting and visualizing femoral neck fracture including displaced and non-displaced fracture on plain X-ray. *J Digit Imaging*. 2021;34:1099-1109.
36. Sampalis JS, Watson J, Boukas S, et al. Navigating the clinical trial pathway: conception, design, execution, and results dissemination. *Surgery*. 2017;161:576-583.
37. Cai Y, Zeng J, Lan L, et al. Expert recommendations on collection and annotation of otoscopy images for intelligent medicine. *Intell Med*. Forthcoming 2022.
38. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data*. 2019;6:1-48.
39. Chen H, Qi X, Yu L, Dou Q, Qin J, Heng P-A. DCAN: deep contour-aware networks for object instance segmentation from histology images. *Med Image Anal*. 2017;36:135-146.
40. Nie L, Li C, Marzani F, Wang H, Thibouw F, Grayeli AB. Classification of wideband tympanometry by deep transfer learning with data augmentation for automatic diagnosis of Otosclerosis. *IEEE J Biomed Health Inform*. 2021;26:888-897.
41. Deepak S, Ameer P. Brain tumor classification using deep CNN features via transfer learning. *Comput Biol Med*. 2019;111:103345.
42. Shi Z, Hao H, Zhao M, et al. A deep CNN based transfer learning method for false positive reduction. *Multimed Tools Appl*. 2019;78:1017-1033.
43. Wilson BS, Tucci DL, Merson MH, O'Donoghue GM. Global hearing health care: new findings and perspectives. *Lancet*. 2017;390:2503-2515.
44. Mulwafu W, Tataryn M, Polack S, Viste A, Goplen FK, Kuper H. Children with hearing impairment in Malawi, a cohort study. *Bull World Health Organ*. 2019;97:654-662.
45. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? *BMJ Glob Health*. 2018;3:e000798.
46. Alami H, Rivard L, Lehoux P, et al. Artificial intelligence in health care: laying the foundation for responsible, sustainable, and inclusive innovation in low-and middle-income countries. *Glob Health*. 2020;16:1-6.