Received: 16 November 2021 | Accepted: 27 February 2022

DOI: 10.1111/coa.13925

META-ANALYSIS

WILEY

Artificial intelligence to classify ear disease from otoscopy: A systematic review and meta-analysis

Al-Rahim Habib[1,2,3] | Majid Kajbafzadeh[1] | Zubair Hasan[3] | Eugene Wong[3] | Hasantha Gunasekera[1,4] | Chris Perry[2,5] | Raymond Sacks[1] | Ashnil Kumar[6] | Narinder Singh[1,3]

[1]Faculty of Medicine and Health, University of Sydney, Sydney, New South Wales, Australia

[2]Department of Otolaryngology – Head and Neck Surgery, Princess Alexandra Hospital, Woolloongabba, Queensland, Australia

[3]Department of Otolaryngology - Head and Neck Surgery, Westmead Hospital, Westmead, New South Wales, Australia

[4]The Children's Hospital at Westmead, Westmead, New South Wales, Australia

[5]University of Queensland Medical School, Brisbane, Queensland, Australia

[6]School of Biomedical Engineering, Faculty of Engineering, University of Sydney, Sydney, New South Wales, Australia

Correspondence
Al-Rahim Habib, Faculty of Medicine and Health, University of Sydney, Camperdown, NSW, Australia.
Email: al-rahim.habib@sydney.edu.au

Funding information
Research Scholarship from the Garnett Passe and Rodney Williams Memorial Foundation. Avant Foundation Doctor-in-Training Research Grant

Abstract

Objectives: To summarise the accuracy of artificial intelligence (AI) computer vision algorithms to classify ear disease from otoscopy.

Design: Systematic review and meta-analysis.

Methods: Using the PRISMA guidelines, nine online databases were searched for articles that used AI computer vision algorithms developed from various methods (convolutional neural networks, artificial neural networks, support vector machines, decision trees and k-nearest neighbours) to classify otoscopic images. Diagnostic classes of interest: normal tympanic membrane, acute otitis media (AOM), otitis media with effusion (OME), chronic otitis media (COM) with or without perforation, cholesteatoma and canal obstruction.

Main outcome measures: Accuracy to correctly classify otoscopic images compared to otolaryngologists (ground truth). The Quality Assessment of Diagnostic Accuracy Studies Version 2 tool was used to assess the quality of methodology and risk of bias.

Results: Thirty-nine articles were included. Algorithms achieved 90.7% (95%CI: 90.1–91.3%) accuracy to difference between normal or abnormal otoscopy images in 14 studies. The most common multiclassification algorithm (3 or more diagnostic classes) achieved 97.6% (95%CI: 97.3–97.9%) accuracy to differentiate between normal, AOM and OME in three studies. AI algorithms outperformed human assessors to classify otoscopy images achieving 93.4% (95%CI: 90.5–96.4%) versus 73.2% (95%CI: 67.9–78.5%) accuracy in three studies. Convolutional neural networks achieved the highest accuracy compared to other classification methods.

Conclusion: AI can classify ear disease from otoscopy. A concerted effort is required to establish a comprehensive and reliable otoscopy database for algorithm training. An AI-supported otoscopy system may assist health care workers, trainees and primary care practitioners with less otology experience identify ear disease.

KEYWORDS
artificial intelligence, computer vision, diagnosis, machine learning, otoscopy

Meeting information: Verbal presentation at the 2021 Australian Society of Otolaryngology – Head and Neck Surgery Annual Scientific Meeting on Friday, 17 September 2021, in Melbourne, Victoria, Australia and 2021 New Zealand Society of Otolaryngology - Head & Neck Surgery Annual Scientific Meeting virtual conference on Friday, 25 February 2022.

1 | INTRODUCTION

Otoscopy is a routine component of the ear assessment that visualises the external auditory canal and tympanic membrane (TM). It is used to identify common conditions, including infection (otitis externa, otitis media—acute and chronic), otitis media with effusion (OME), perforation, cholesteatoma, tympanosclerosis, foreign body, tympanostomy tube presence/position and cerumen impaction. Ear examinations are often conducted in primary care settings by local community health workers, nurses, medical students, general practitioners and emergency physicians. Concerning findings typically then lead to intervention or to specialist referral, where an otolaryngologist will examine the ear.

Otoscopy accuracy in primary care settings varies based on user training and experience. However, the reported literature describes diagnostic accuracy estimates between 30% and 67.5% when compared to otolaryngologists (ground truth) for specific diagnoses.[1] Previous efforts to improve performance have focussed on educational techniques, including online teaching modules comparing normal and abnormal otoscopic TM images, practical tutorials conducted by otolaryngologists and simulation with artificial ear models.[2,3] Comparisons made before and after education sessions demonstrate short-term improvements in otoscopy performance, although benefits are not sustained long-term and can decrease from initial assessments.[3] The frequency of otoscopy in primary care settings and the observed performance inconsistencies may provide an opportunity for artificial intelligence (AI) to assist in the accurate identification of ear disease.

AI can replicate the ability of human cognition to recognise patterns, identify anomalies and construct rational solutions to potential obstacles.[4] Popularised applications of AI in health care include the use of computer vision to differentiate benign versus malignant skin lesions, identify diabetic retinopathy from fundoscopic images, assist radiologists to interpret chest x-rays and predict infectious disease outbreaks, as in the case of the novel coronavirus (COVID-19) pandemic.[5–8]

AI-based computer vision algorithms are an emerging technology that can be used to classify ear disease using otoscopic images.[9] The use of AI-based computer vision algorithms as an adjunct to otoscopy performed in primary care settings may be most relevant in rural and remote areas where access to otolaryngologists is limited. In these scenarios, ear examinations are often performed by nurses and community health workers with less clinical experience than otolaryngologists in accurately recognising ear disease. In rural and remote areas, telemedicine initiatives, such as tele-otoscopy, are feasible strategies to capture otoscopic images for subsequent metropolitan specialist review but are often disadvantaged by delays in clinical decision making and implementation of interventions.[10]

The aim of this review was to evaluate the performance of AI-based computer vision algorithms to classify ear disease from otoscopy. Our objectives were to synthesise existing literature related to the use of AI-based computer vision algorithms for otoscopy, assess the performance of existing models and propose a guide for future algorithm development.

Key Points

1. AI-based computer vision algorithms can differentiate between binary (2 diagnosis options) and multiple ear disease diagnoses (3 or more diagnosis options).
2. AI-based computer vision algorithms have been shown to classify otoscopy images more accurately than human, nonexpert assessors.
3. AI-based computer vision algorithms have been developed using various machine learning techniques of which, models using CNNs achieve the greatest classification accuracy.
4. Substantial heterogeneity was found between studies reflecting, in part, diverse sources of images and collection practices, machine learning methods, diagnostic classes and ground-truth definitions.
5. Future efforts are needed to establish a standardised, comprehensive and validated database to develop clinically relevant AI-based computer vision algorithms for otoscopy.

2 | METHODS

The present systematic review was conducted in accordance with the PRISMA guidelines[11] and registered with the International prospective register of systematic reviews (PROSPERO) on 18 February 2021 (ID number: CRD42021202594). Ethics approval and patient consent were not required for this review.

2.1 | Literature search

A systematic search of online databases (including Google Scholar, MEDLINE, Embase, PubMed, Scopus, ProQuest, ACP Journal Club, Health Technology Assessment and the Cochrane Library) for articles, abstracts or conference proceedings published in the past 10 years through 31 October 2021 that used AI-based computer vision approaches to classify otoscopic TM images was conducted. Searches were limited to those involving human subjects and those published in the English Language.

Medical subject headings (MeSH) terms and non-MeSH terms related to AI approaches included: 'artificial intelligence', 'machine learning', 'deep learning', 'convolutional neural networks', 'support vector machines', 'image recognition', 'image classification', 'object detection' and 'computer-assisted diagnosis'. MeSH terms and keywords related to otoscopic TM images included: 'otoscopy', 'ear', 'eardrum', 'tympanic membrane', 'ear disease', 'acute otitis media',

'otitis media', 'chronic otitis media' and 'chronic suppurative otitis media'.

2.2 | Selection criteria

Titles and abstracts were reviewed for eligibility by two independent investigators (ARH and MK). Discrepancies between the two investigators were resolved by the senior author (NS), a board-certified otolaryngologist. Reference lists of available full-text articles were also manually screened for further studies eligible for inclusion in this review. Diagnostic observational studies describing the development of an autonomous, supervised algorithm to classify otoscopic TM images using the AI approaches described above were included. Articles that were excluded consisted of those that did not use AI approaches of interest, utilised imaging modalities other than otoscopy or were review articles or editorials. Study inclusion/exclusion is summarised in a PRISMA flow diagram (Figure 1).

2.3 | Data extraction

Two investigators (ARH and MK) independently extracted data from included studies for analysis. The following characteristics were extracted from included studies: primary author, year, study objective, AI technique to achieve study objective, data source, otoscope type and manufacturer, image labelling method, source for ground-truth classification, diagnostic categories of interest, image size and quality, number and distribution of training images, number and distribution of test images, number and distribution of validation images, ratio of training/test/validation images by diagnostic categories and performance characteristics. Additional characteristics extracted included the type of deep learning models, batch size, learning rate, method used to standardise input images and use of image augmentation methods, segmentation, hyperparameter tuning and cross-validation techniques.

2.4 | Critical appraisal and risk of bias assessment

The quality of included studies was assessed using the Quality Assessment and Diagnostic Accuracy Studies-2 (QUADAS-2) tool, as per the Cochrane Collaboration for critical appraisal of diagnostic test accuracy evaluations.[12] The QUADAS-2 assessment tool is composed of patient selection, index test, reference standard and flow and timing. Study quality was assessed by two independent investigators (ARH and MK). Uncertainties or discrepancies were discussed with the senior author (NS) to achieve consensus. As described by Whiting et al.,[12] applicability concerns were determined for patient selection, the index test and the reference standard. For patient selection, reviewers considered whether the subjects recruited for the study and used to train the algorithm were applicable to the target population where algorithms would be implemented. For the index test, reviewers considered whether the classification categories of the algorithm were applicable to its

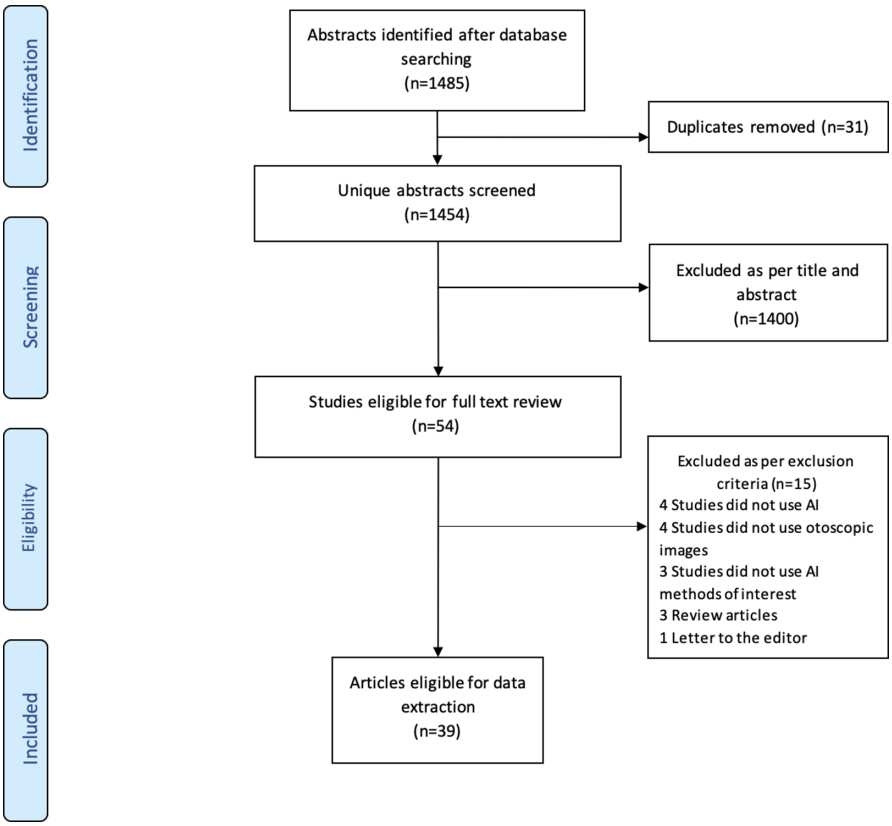

FIGURE 1 PRISMA study flow diagram [Colour figure can be viewed at wileyonlinelibrary.com]

intended use. For the reference standard, reviewers considered whether the method used reflected clinical practice for future applications.

2.5 | Outcomes

A systematic review and meta-analysis were performed with the primary outcome assessed being algorithm diagnostic accuracy in classifying ear disease from otoscopic images. Secondary outcomes included sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1 score and area under the curve (AUC).

2.6 | Statistical analysis

The statistical analysis was performed using ReviewManager (RevMan 5.3, Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2014) in accordance with the Cochrane Handbook[13] and Stata 17 (StataCorp LLC, 2021). Heterogeneity of model performance was summarised with the I^2 statistic, H^2 statistic, Cochrane's Q statistic and chi-square test. Heterogeneity values >75% were considered as substantial heterogeneity. Inverse variance weighting was used for pooling model performance, and fixed effects were applied.

3 | RESULTS

3.1 | Study selection

The search strategy yielded a total of 1485 relevant articles and abstracts. Following full-text review, 39 articles met the inclusion criteria (Figure 1).

3.2 | Characteristics of included studies

Table 1 provides a summary of characteristics for included studies. Thirty studies collected images from outpatient and inpatient primary care settings, six studies combined images from primary care settings and online sources (Google Images), one study collected images from Google Images alone and two studies did not report the source of images.

3.3 | Risk of bias of included studies

Figure 2 and Figure S1 illustrate the risk of bias and applicability concerns of included studies using the QUADAS-2 tool. High risk of bias (37 of 39 studies) was identified in patient selection criteria due to failure to utilise consecutive or random sampling. High risk of bias (15 of 39 studies) was also observed in use of the reference standard (ground truth), as these articles did not utilise more than 1 otolaryngologist to review otoscopic images to confirm class labels.

3.4 | Binary classification algorithms

Nine unique binary classification algorithms to classify otoscopy images were reported (Table S1).

3.4.1 | Normal versus abnormal

AI algorithms achieved a pooled accuracy of 90.7% (95%CI: 90.1–91.3%) to difference between normal or abnormal otoscopy images with substantial heterogeneity between studies (n = 14 studies, I^2 = 96.9%, p = .001, Figure 3).

Four studies[14–17] used the Özel Van Akdamar Hospital otoscopic image database to train binary algorithms using various classification techniques. Simon et al.[14] demonstrated that pretrained convolutional neural networks (CNNs) and support vector machines (SVMs) could achieve greater classification accuracy than k-nearest neighbours (k-NNs), artificial neural networks (ANNs), decision trees (DTs) or the Naïve Bayes technique. Basaran et al.[15] utilised multiple pretrained CNNs to evaluate the effect of segmentation, distribution between training and test data and cross-validation on algorithm performance. In this study, pretrained CNNs achieved enhanced accuracy by applying basic image augmentation techniques and using region of interest patches, rather than full otoscopic images.[15] The pretrained CNNs VGGNet-16 and VGGNet-19 achieved the greatest classification accuracy to differentiate normal from abnormal otoscopic images (90.5% and 90.1% respectively).[15] Mironica et al.[18] demonstrated that the greatest classification accuracy was achieved by CNNs and SVMs in 186 images collected from outpatient primary care settings. Enhanced performance was achieved by adding a colour coherence vector (CCV) to the algorithms (models with CCV vs without: CNN – 73.1% vs. 68.8%, SVM – 72.0% vs. 64.5%).[18]

Using intraoperative assessment of children taken to the operating room with the intent of myringotomy to determine the ground truth, Crowson et al.[19] differentiated between normal and OME with 83.8% accuracy using ResNet-34 and Monte Carlo cross-validation resampling with 5 repetitions.

3.5 | Multiclassification algorithms

Seventeen unique multiclassification algorithms to classify otoscopy images were identified (Table S2). Overall, multiclassification algorithms achieved a pooled accuracy of 96.2% (95%CI: 96.1–96.4%) with substantial heterogeneity between studies (n = 18 studies, I^2 = 98.8%, p = .001, Figure 4).

TABLE 1 Summary of included studies

Author/year	Image source	Ground truth	Classes	Total images	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F1 score	AUC
Wang 2021[14]	H	2 otolaryngologists	2	100	81.7	83.3	80.0	80.6	NR	NR	0.88
Sundgaard 2021[15]	H	1 otolaryngologist	3	1336	86.0	NR	NR	NR	NR	NR	NR
Tsutsumi 2021[16]	H[1], O	O: NR, H[1]: 2 otolaryngologists, 1 paediatrician	2	400	77.0	70.0	84.4	81.4	73.7	NR	0.90
Tsutsumi 2021[16]	H[1], O	O: NR, H[1]: 2 otolaryngologists, 1 paediatrician	5	400	66.0	55.4	NR	79.8	NR	NR	0.88
Byun 2021[17]	H	3 otolaryngologists	4	2372	97.2	NR	NR	NR	NR	NR	NR
Zeng 2021[18]	H	6 otologists	8	20542	95.5	NR	NR	NR	NR	NR	0.99
Crowson 2021[19]	H	1 of 5 paediatric otolaryngologists	2	338	83.8	NR	NR	NR	NR	80.0	0.93
Alhudhaif 2021[20]	H[1]	2 otolaryngologists, 1 paediatrician	4	857	98.2	97.7	99.3	NR	NR	96.9	NR
Cai 2021[21]	H	otolaryngologists (number not reported).	4	6066	93.4	NR	NR	NR	NR	96.8	0.98
Camalan 2021[22]	H[1]	1 otologist, 1 paediatric otolaryngologist	3	300	85.8	NR	NR	NR	NR	NR	NR
Ucar 2021[23]	H[1]	1 otolaryngologist	4	880	98.1	98.1	99.4	98.2	NR	NR	0.99
Camalan 2020[24]	H[1]	NR	3	454	88.1	NR	NR	NR	NR	NR	NR
Wu 2020[25]	H	2 otologists	3	12203	97.8	96.8	98.0	96.9	98.4	NR	0.99
Basaran 2020[26]	H[1]	2 otolaryngologists, 1 paediatrician	2	282	90.4	86.8	93.5	NR	NR	87.3	0.95
Goshtasbi 2020[27]	H[1], O	O: NR, H[1]: 2 otolaryngologists, 1 paediatrician	2	400	77.0	70.0	84.0	81.0	74.0	NR	0.89
Goshtasbi 2020[27]	H[1], O	O: NR, H[1]: 2 otolaryngologists, 1 paediatrician	5	400	71.0	NR	NR	NR	NR	NR	0.91
Habib 2020[28]	O	2 otolaryngologists	2	233	76.0	76.0	76.0	76.0	76.0	NR	0.87
Khan 2020[29]	H	2 otolaryngologists	3	2484	87.0	95.0	NR	95.2	NR	95.1	0.99
Simon 2020[30]	H[1]	2 otolaryngologists, 1 paediatrician	2	956	81.4	83.6	83.8	NR	NR	NR	0.89
Viscaino 2020[31]	H[1]	1 otolaryngologist	4	720	88.1	87.8	95.9	87.7	NR	NR	1.00
Cömert 2020[32]	H[1]	2 otolaryngologists, 1 paediatrician	4	857	99.4	99.4	99.8	NR	NR	99.3	NR
Basaran 2019[33]	H[1]	2 otolaryngologists, 1 paediatrician	2	598	97.9	99.1	98.5	NR	NR	NR	NR
Basaran 2019[34]	H[1]	2 otolaryngologists, 1 paediatrician	2	223	76.1	70.8	80.1	NR	NR	NR	0.81
Cha 2019[35]	H	1 otologist, 1 physician	6	10544	94.2	93.7	96.8	NR	NR	NR	NR
Lee 2019[36]	H	2 otologists	2	1338	91.0	90.5	92.9	98.0	72.3	NR	0.92
Livingstone 2019[37]	H, O	1 otologist, 1 otolaryngology resident	14	1366	88.7	86.1	NR	90.9	NR	NR	NR

(Continues)

TABLE 1 (Continued)

Author/year	Image source	Ground truth	Classes	Total images	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F1 score	AUC
Livingstone 2019[38]	H	2 otolaryngologists	3	529	84.4	NR	NR	NR	NR	NR	NR
Seok 2019[39]	H	2 otolaryngologists	2	920	92.9	NR	NR	92.9	NR	NR	NR
Huang 2018[40]	NR	NR	3	20	70.0	NR	NR	NR	NR	NR	NR
Kasher 2018[41]	H, O	NR	2	108	82.1	NR	NR	NR	NR	NR	NR
Myburgh 2018[42]	NR	2 otologists	5	389	86.2	86.8	96.4	87.4	96.4	NR	NR
Senaras 2018[43]	H, O	NR	2	1082	84.3	86.8	81.9	NR	NR	NR	NR
Tran 2018[44]	H	NR	2	214	91.4	89.5	93.3	91.9	NR	NR	0.92
Senaras 2017[45]	H	otolaryngologists (number not reported)	2	247	84.6	87.3	81.4	NR	NR	NR	NR
Myburgh 2016[46]	H	2 otologists	5	486	80.6	80.6	94.4	81.0	94.6	NR	NR
Wang 2015[47]	H, O	NR	2	215	90.0	85.0	92.0	NR	NR	NR	NR
Shie 2014[48]	H	otolaryngologists (number not reported)	4	865	88.0	91.6	79.9	NR	NR	NR	0.91
Kuruvilla 2013[49]	H	3 otologists	3	181	89.9	NR	NR	NR	NR	NR	NR
Kuruvilla 2012[50]	H	3 otologists	3	135	84.0	NR	NR	NR	NR	NR	NR
Mironica 2011[51]	H	1 otolaryngologist	2	186	73.1	NR	NR	NR	NR	NR	NR
Vertan 2011[52]	H	1 otolaryngologist	3	100	59.9	NR	NR	NR	NR	NR	NR

Note: Classes refers to the number of diagnostic categories included in the algorithm; (1) online data set.

Abbreviations: AUC—area under the curve; H—primary care hospital database; NPV—negative predictive value; NR—not reported; O—online repository; PPV—positive predictive value.

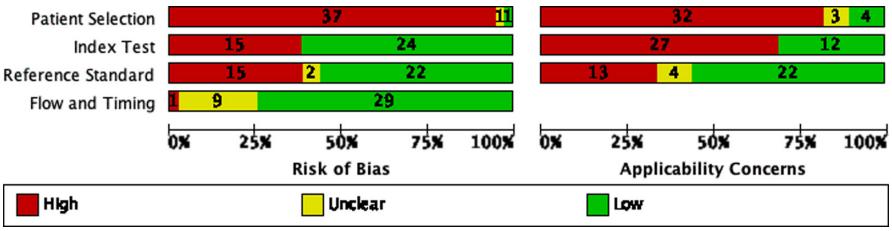

FIGURE 2 Summary of risk of bias and applicability of concerns graph using the Quality Assessment of Diagnostic Accuracy Studies tool, version 2 (QUADAS-2) [Colour figure can be viewed at wileyonlinelibrary.com]

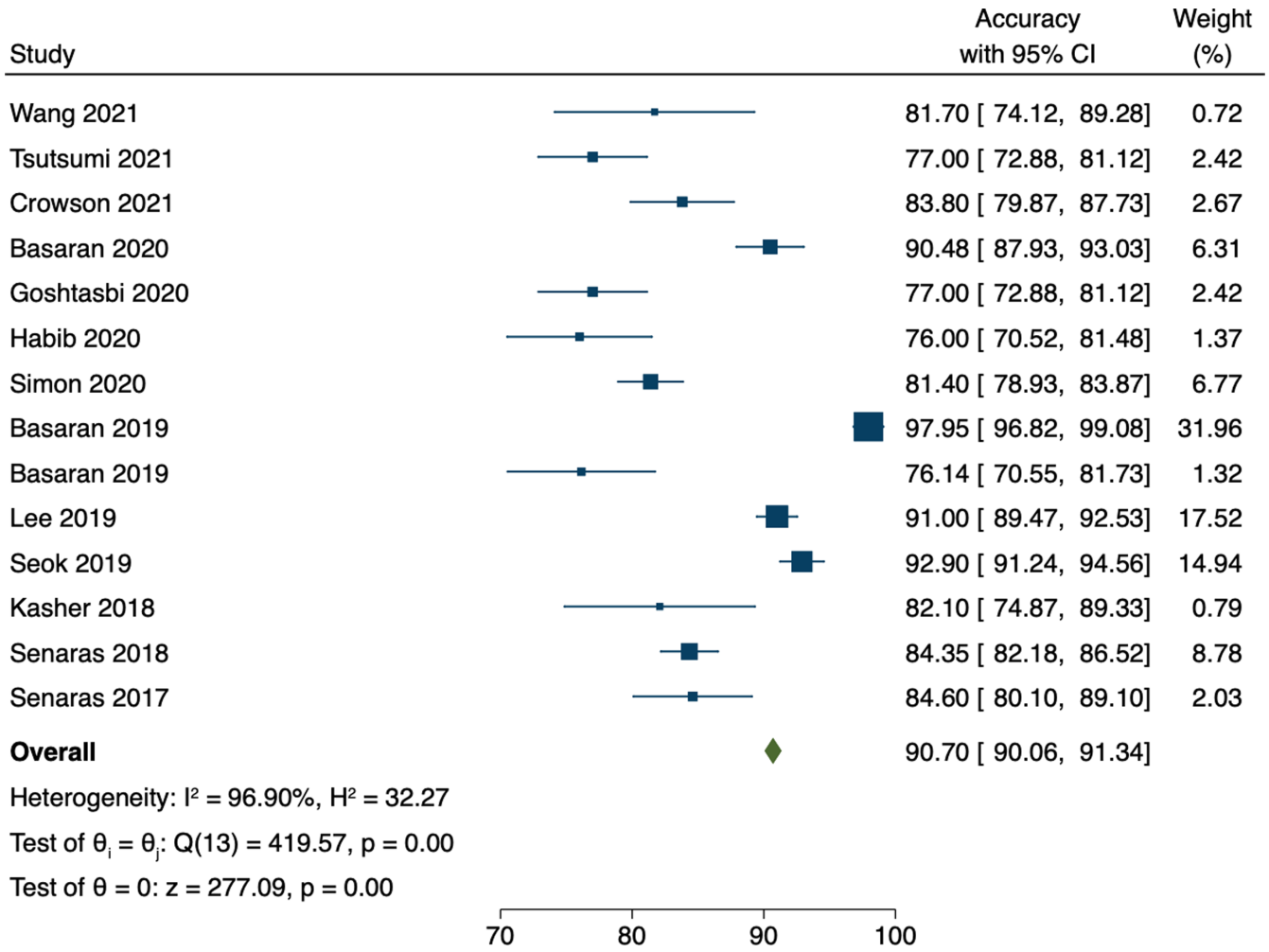

FIGURE 3 Forest plot comparing accuracy of AI algorithms to classify normal versus abnormal otoscopy images [Colour figure can be viewed at wileyonlinelibrary.com]

3.5.1 | Normal, AOM and OME

The most common multiclassification model differentiated between normal, AOM and OME in five studies[20–24] Overall, the algorithms achieved an accuracy of 97.6% (95%CI: 97.3–97.9%) with substantial heterogeneity between studies ($n = 3$ studies, $I^2 = 98.7\%$, $p = .001$, Figure S2).[20,21,24] Two studies were excluded from the meta-analysis because the number of test images was not reported.[22,23]

Wu et al.[20] used 12 203 otoscopic images and applied the Xception and MobileNets-V2 pretrained CNNs for classification. Image augmentation was used during the preprocessing phase consisting of rotation, width shift, height shift, shearing, zooming and horizontal flip. Hyperparametric tuning was used to establish the optimal classification algorithm (epochs: 100, initial learning rate: 0.001, batch size: 14 [Xception] and 64 [MobileNets-V2]). Between diagnostic categories, OME was classified less accurately (96.7%) than normal (98.3%) or AOM (98.5%).

3.6 | AI versus human classification

Five studies[25–29] compared the performance of image classification algorithms to human assessors. Of these, three studies compared the performance of the AI algorithm to human assessors using the same test set.[25,26,28] Restricting the meta-analysis to these three studies demonstrated that AI algorithms outperformed human assessors

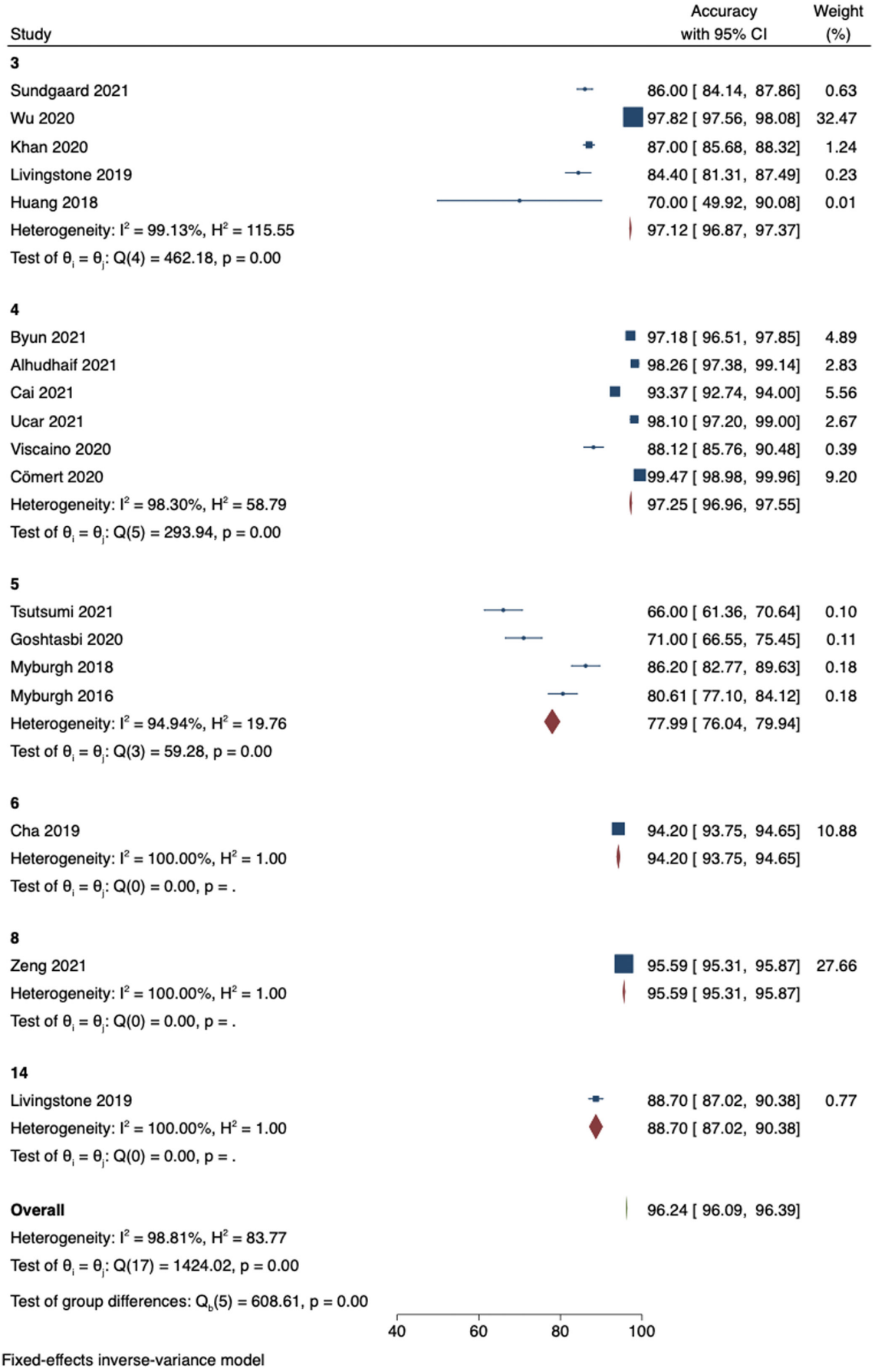

FIGURE 4 Forest plot comparing accuracy of multiclassification AI algorithms to classify ear disease from otoscopy, stratified by number diagnostic classes [Colour figure can be viewed at wileyonlinelibrary.com]

having achieved an accuracy of 93.4% (95%CI: 90.5–96.4%) versus 73.2% (95%CI: 67.9–78.5%) with substantial heterogeneity between studies (AI: $n = 3$ studies, $I^2 = 78.6\%$; human assessors: $n = 3$ studies, $I^2 = 83.7\%$, $p = .001$, Figure 5, Table S3).

Byun et al.[28] found that a 4-class image classification algorithm using ResNet-18 and a shuffle attention model outperformed 10 nonexperts (first- and second-year resident physicians) (97.1% vs. 82.9%) to classify normal, OME, COM or cholesteatoma. Otolaryngology experience of the resident physicians was not reported.

Livingstone et al.[25] recruited 10 nonexperts (general practitioners and trainees from paediatrics, emergency medicine and otolaryngology) to review a test set of 89 images for 14 diagnostic classes. Human, nonexpert assessors were summarised together, and stratification by specialty was not provided. Overall, the algorithm outperformed all the human assessors in 12 out of 14 categories (algorithm: 88.7% vs. human assessors: 58.9%). The algorithm classified cholesteatoma and otomycosis less accurately than human assessors (cholesteatoma – 50.0% vs. 55.0%, otomycosis – 0.0% vs. 40.0% respectively).

Khan et al.[26] recruited 17 human expert and nonexpert assessors (7 specialist otolaryngologists and 10 nonexperts) to review 100 test images for three diagnostic classes (normal, OME, COM). Overall, the algorithm achieved a classification accuracy of 87.0% compared to 74.0% for the pooled human expert and nonexpert assessors. Stratification of accuracy results by diagnostic class or assessor seniority was not provided.

4 | DISCUSSION

Performance in diagnosing ear disease from otoscopy varies by user training and experience.[30] This study summarises the performance of AI-based computer vision algorithms to diagnosis of ear disease from otoscopy. Thirty-nine studies were included in this review, evaluating the performance of 9 binary and 17 multiclassification algorithms. AI-based computer vision algorithms achieved 90.7% (14 studies) accuracy to differentiate between normal or abnormal otoscopy images and 97.6% (3 studies) to differentiate between normal, AOM or OME. Compared to manual classification, AI-based computer vision algorithms outperformed human assessors to classify otoscopy images (93.4% vs. 73.2% accuracy, respectively) in three studies. Substantial heterogeneity in performance was identified between studies.

AI-based computer vision algorithms with CNNs achieved greater accuracy in binary and multiclassification categories. Harnessing transfer learning by freezing the final layers of a CNN's architecture is advantageous by using established computer vision performance to evaluate new tasks of interest. CNNs may have the potential to

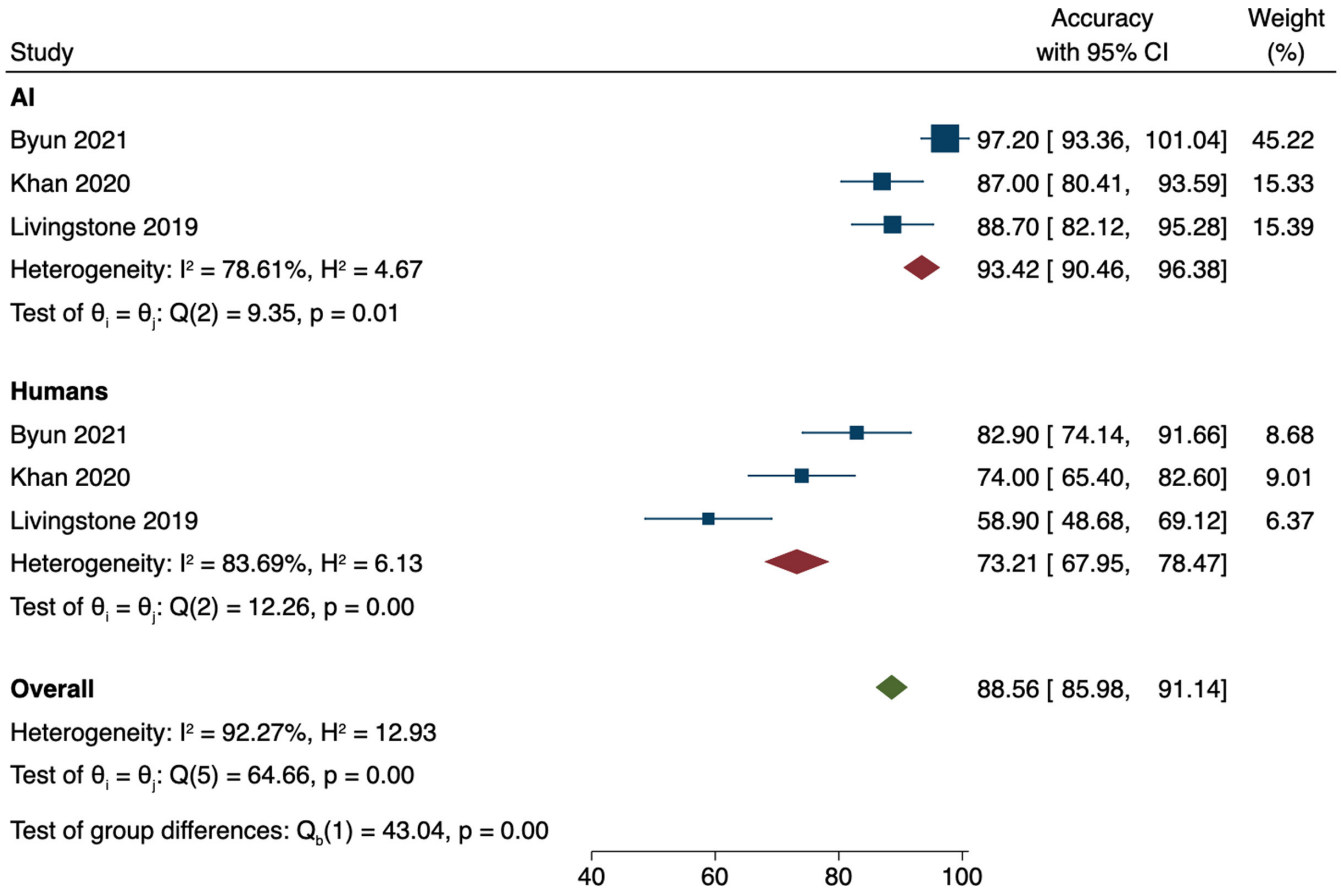

FIGURE 5 Forest plot comparing ear disease classification accuracy from otoscopy between AI algorithms and human assessors [Colour figure can be viewed at wileyonlinelibrary.com]

achieve greater accuracy than ANNs, SVMs or DTs due in part to their architecture, weighting system and features extracted in hidden layers.[20,31] However, CNNs vary in relation to the amount of computational power, speed and size of models produced. For example, Inception-V3 yielded greater accuracy than MobileNet-V2 in multiclassification but produced a larger size model.[32] This is relevant whether considering deployment of the model into an edge device (e.g. otoscope or smartphone) versus workstations with graphics processing units (GPUs) or cloud-based platforms.[32] Focus on the TM and localising anatomical or pathological characteristics with segmentation models has achieved substantial improvements in accuracy while considering various middle ear conditions.[33]

This review has several strengths. Firstly, broad inclusion criteria were used to explore articles clinically relevant to the study question. This identified algorithms with various combinations of diagnostic classes and classification approaches. Secondly, the breadth of diagnostic classes is important for clinicians to appreciate the strengths and weaknesses of computer vision algorithms and highlight potential for future exploration. Thirdly, the QUADAS-2 assessment tool was used to identify risk of biases and applicability concerns. Before image classification algorithms are introduced into routine clinical practice, rigorous efforts are required to evaluate performance and reliability.[34]

This systematic review has limitations. Firstly, this review is limited by the variability between eligible studies in terms of data sources, image capture devices and machine learning methods. Studies included used various patient selection criteria, classification categories and ground-truth definitions. As a result, this could plausibly introduce selection and measurement bias in the outcomes reported and overall conclusions drawn from the results. The heterogeneity between classification categories suggests a lack of standardisation or quality control in terms of image acquisition. It is important to acknowledge the clinical need for an AI-based computer vision algorithm and apply this technology accordingly.

AI-based computer vision algorithms depend on high-quality training data with accurate and reliable ground-truth labels. Ideally ground-truth labels should be determined by consensus of multiple independent experts and/ or by additional evidence (e.g. clinical history, histopathology or independent investigation results).[35,36] For otoscopy, ground truth may be based on expert review of images along with clinical history, tympanometry, audiometry or myringotomy results. In this review, most studies used single otolaryngologists to classify otoscopic images during routine outpatient examinations. High risk of bias was identified in patient and image selection as most sampling methods for training images were not randomised or did not use consecutive recruitment. Furthermore, ground-truth assessment was typically not validated by multiple independent experts or via other means. Crowson et al.[19] uniquely used intraoperative otoscopic images of children undergoing myringotomy for recurrent AOM or OME. The authors suggested that this approach may represent the gold standard of detecting OME. Despite this, the authors did not report whether myringotomy was performed in all children, even those with low pretest probability of disease. Furthermore, positive pressure ventilation may displace middle ear fluid immediately prior to myringotomy in up to 15% of children with OME.[37] While myringotomy results may add useful additional evidence, this approach significantly reduces the number of test images available. Only those patients proceeding to intervention can be included, effectively excluding all patients with suspected normal ears (that do not proceed to intervention) from the training dataset. Accordingly, myringotomy results cannot be used as the basis for ground-truth labelling in large training datasets.

To create large datasets for training AI-based computer vision algorithms, we recommend that ground-truth labelling for otoscopic images be based primarily on the consensus of multiple independent experts along with clinical history, and tympanometry and audiometry results. This may reflect real-world practices, where clinical suspicion from experts (e.g. otolaryngologists with subspeciality interest in otology) is used to determine which patient is appropriate for invasive interventions. Tele-otoscopy databases may be a suitable source for this research as images have been collected by clinical staff with experience in performing otoscopy and have undergone quality control, vetting by an otolaryngologist, and often review in conjunction with clinical history, tympanometry and audiometry to establish a diagnosis and treatment plan.

As identified in this review, the optimal approach to develop an AI-based computer vision algorithm can use data augmentation, CNNs (DenseNet, Xception or Inception ResNet-V2), ensemble models combining multiple classifying features, sequential multistep segmentation to localise the TM, hyperparameter tuning and cross-validation resampling to yield the greatest performance. It is an important methodological consideration to partition data into training, validation and test groups to minimise the risk of overfitting, limit training bias and consider the generalisability of model predictions to independent or heterogenous data.[38] Training a model and then testing, it on the same data would overestimate prediction performance and misrepresent its function on unseen data. However, partitioning data into discrete groups for training, validation and testing can reduce the number of samples used for model learning. To address this, cross-validation techniques can be applied to divide the training group into sets (i.e. folds) and hold-out sets to validate the model by averaging performance on a predetermined number of loops.[39,40] Most studies in this review applied cross-validation techniques. Basaran et al. (2020) demonstrated that 10-fold cross-validation yielded higher accuracy, sensitivity and specificity results than utilising 50% of the data source for training and reserving 50% for testing.[15] The collection of otoscopic images requires time, equipment, adequate storage, staff training and consideration of privacy, patient confidentiality and security. In this scenario, efforts to include a greater number of images for training and limiting bias may be important to maximise generalisability and clinical value. Attention models can be applied to visualise important areas used to establish predictions. Future algorithms may incorporate specific segmentation models using the U-Net architecture. For example, Pham et al. (2021) proposed EAR-UNet, an automatic segmentation model for TMs from video-otoscopic images integrating pretrained

CNNs (EfficientNet-B4 and ResNet), achieving 96% accuracy to localise the TM in normal, AOM, OME and COM otoscopic images.[33]

Transitioning the AI-based computer vision algorithm for otoscopy from a virtual environment to the clinical frontline will depend on real-world test performance and applicability in daily clinical practice. This technology has the potential to inform judgement, improve triage and save time and resources for batch screening. Despite this, performance of the algorithm depends on training data. Bias in patient selection, ground-truth labelling and diagnostic classes may impact generalisability. Implementation in real-world settings may depend on desirably, feasibility and viability of this technology as an adjunct to existing clinical practices. Further efforts to progress the application of AI for otoscopy may be directed at establishing a comprehensive, publicly available, open access otoscopy database with associated symptoms, tympanometry, pneumatic otoscopy and audiometry findings, validated by multiple otologists. Future studies may build on previous studies evaluating the concordance between expert and nonexpert assessments,[40] by exploring the use of AI as an adjunct to clinical decision making. Initial efforts for real-world applications may be targeted at identifying poor quality otoscopic images that limit accurate assessment (e.g. blurriness, over saturation, lack of white balance, moisture, TM not visualised and wax obstruction). Delineating the strengths and limitations of AI to autonomously classify otoscopic images is necessary for real-world applications.

5 | CONCLUSION

In this review, 39 articles explore the role of AI-based computer vision algorithms for otoscopy. AI algorithms achieved 90.7% accuracy to differentiate between normal or abnormal otoscopy images and 97.6% to differentiate between normal, AOM or OME. Compared to manual classification, AI algorithms outperformed human assessors to classify otoscopy images (93.4% versus 73.2% accuracy respectively). However, substantial heterogeneity in performance was identified between studies. A concerted effort is warranted to establish a standardised, robust, comprehensive and reliable database to develop clinically relevant computer vision algorithms for otoscopy. An AI-based computer vision algorithm for otoscopy has the potential to support health care workers and primary care practitioners with less otology experience to identify and manage ear conditions early to minimise the risk of sequelae from untreated disease.

ACKNOWLEDGEMENTS

We acknowledge support from the Garnett Passe and Rodney Williams Memorial Foundation Research Scholarship and the Avant Foundation Doctor-in-Training Research Grant. The authors would like to acknowledge Dr Mukesh Prasad for proofreading the manuscript. Open access publishing facilitated by The University of Sydney, as part of the Wiley - The University of Sydney agreement via the Council of Australian University Librarians. [Correction added on 24 May 2022, after first online publication: CAUL funding statement has been added.]

CONFLICT OF INTEREST

Dr Al-Rahim Habib receives grant funding from the Garnett Passe and Rodney Williams Memorial Foundation, Microsoft AI for Humanitarian Action Grant and the Avant Foundation Doctor-in-Training Research Grant. Professor Raymond Sacks is a consultant for Medtronic. Associate Professor Narinder Singh receives grant funding from the Garnett Passe and Rodney Williams Memorial Foundation and Microsoft AI for Humanitarian Action Grant. He is a consultant for ENT Technologies, Optinose, Nasus Medical and ResMed, and a principal investigator for drug/ device trials for GSK, Lyra and Covance.

ETHICS STATEMENT

Ethics approval and patient consent were not in need for this study.

AUTHOR CONTRIBUTION

ARH, AK and NS designed the study; ARH and MK extracted and analysed the data; ARH prepared the manuscript; ARH, MK, ZH, EW, HG, CP, RS, AK, and NS involved in synthesis, revision and approval of manuscript. All authors agreed to be accountable for all aspects of the work.

[Correction added on March 21, 2022, after first online publication: Peer review history is not available for this article, so the peer review history statement has been removed.]

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Al-Rahim Habib https://orcid.org/0000-0001-6327-4648
Eugene Wong https://orcid.org/0000-0002-5799-5083

REFERENCES

1. Buchanan CM, Pothier DD. Recognition of paediatric otopathology by General Practitioners. *Int J Pediatr Otorhinolaryngol*. 2008;72(5):669-673.
2. Wormald PJ, Browning GG, Robinson K. Is otoscopy reliable? A structured teaching method to improve otoscopic accuracy in trainees. *Clin Otolaryngol Allied Sci*. 1995;20(1):63-67.
3. Oyewumi M, Brandt MG, Carrillo B, et al. Objective evaluation of otoscopy skills among family and community medicine, pediatric, and otolaryngology residents. *J Surg Educ*. 2016;73(1):129-135.
4. Lecun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444.
5. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115-118.
6. Raumviboonsuk P, Krause J, Chotcomwongse P, et al. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *NPJ Digit Med*. 2019;2(1):10. 10.1038/s41746-019-0099-8
7. Bogoch II, Watts A, Thomas-Bachli A, Huber C, Kraemer MUG, Khan K. Pneumonia of unknown aetiology in Wuhan, China: potential for international spread via commercial air travel. *J Travel Med*. 2020;27(2):1-3.

8. Seah JCY, Tang CHM, Buchlak QD, et al. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *Lancet Digit Health*. 2021;7500(21):1-11.
9. Marom T, Kraus O, Habashi N, Tamir SO. Emerging technologies for the diagnosis of otitis media. *Otolaryngol Head Neck Surg*. 2019;160(3):447-456.
10. Smith AC, Armfield NR, Wu WI, Brown CA, Perry C. A mobile telemedicine-enabled ear screening service for Indigenous children in Queensland: activity and outcomes in the first three years. *J Telemed and Telecare*. 2012;18(8):485-489.
11. Moher D, Liberati A, Tetzlaff J, Altman DG, Group TP. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Medicine*. 2009;6(7):e1000097.
12. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(4):529-536.
13. Higgins J & Green S. Cochrane handbook for systematic reviews of interventions version 5.1.0 [Internet]. 2011. Accessed February 18, 2021. www.cochrane-handbook.org
14. Simon C, Caorsi V, Campillo C. Chapter 14: otitis media diagnosis model for tympanic. *Modelling and analysis of active biopotential signals in healthcare*, 2nd ed. Institute of Physics Publishing Ltd; 2020.
15. Basaran E, Comert Z, Celik Y. Convolutional neural network approach for automatic tympanic membrane detection and classification. *Biomed Signal Process Control*. 2020;56(1):1-14.
16. Goshtasbi K. Machine learning models to predict diagnosis and surgical outcomes in otolaryngology [Internet]. University of California Irvine; 2020. Accessed January 3, 2021. https://escholarship.org/uc/item/1tr0c2p0
17. Tsutsumi K, Goshtasbi K, Risbud A, et al. A web-based deep learning model for automated diagnosis of otoscopic images. *Otol Neurotol*. 2021;42(9):e1382-e1388.
18. Mironică I, Vertan C, Gheorghe DC. Automatic pediatric otitis detection by classification of global image features. 2011 E-Health and Bioengineering Conference, EHB 2011. 2011;1-4.
19. Crowson MG, Hartnick CJ, Diercks GR, et al. Machine learning for accurate intraoperative pediatric middle ear effusion diagnosis. *Pediatrics*. 2021;147(4):e2020034546.
20. Wu Z, Lin Z, Li L, et al. Deep learning for classification of pediatric otitis media. *Laryngoscope*. 2020;131:1-8.
21. Huang YK, Huang CP. A depth-first search algorithm based otoscope application for real-time otitis media image interpretation. Parallel and Distributed Computing, Applications and Technologies, PDCAT Proceedings. 2018;2017-Decem:170–5.
22. Kuruvilla A, Shaikh N, Hoberman A, Kovačević J. Automated diagnosis of otitis media: vocabulary and grammar. *Int J Biomed Imaging*. 2013;2013:1-15.
23. Kuruvilla A, Li J, Yeomans PH, Quelhas P, Instituto I, Biomedica DE. Otitis media vocabulary and grammar. *Proc Int Conf Image Proc*. 2012;2012:2845-2848.
24. Sundgaard JV, Harte J, Bray P, et al. Deep metric learning for otitis media classification. *Med Image Anal*. 2021;71:102034.
25. Livingstone D, Chau J. Otoscopic diagnosis using computer vision: an automated machine learning approach. *Laryngoscope*. 2020;130:1408-1413.
26. Khan MA, Kwon S, Choo J, et al. Automatic detection of tympanic membrane and middle ear infection from oto-endoscopic images via convolutional neural networks. *Neural Netw*. 2020;126:384-394. 10.1016/j.neunet.2020.03.023
27. 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(1):1-7.
28. 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(15):3198.
29. Wang W, Tamhane A, Santos C, et al. Pediatric otoscopy video screening with shift contrastive anomaly detection. *Front Digit Health*. 2022;3:810427.
30. Pichichero ME, Poole MD. Comparison of performance by otolaryngologists, pediatricians, and general practioners on an otoendoscopic diagnostic video examination. *Int J Pediatr Otorhinolaryngol*. 2005;69(3):361-366.
31. Geron A, Géron A. Hands-on machine learing with scikit-learn & tensor flow [Internet]. O'Reilly Media; 2017:760. Accessed March 18, 2021. http://arxiv.org/abs/1412.3919
32. Kasher MS. Otitis media analysis-an automated feature extraction and image classification system. 2018. Accessed November 10, 2018. https://www.theseus.fi/handle/10024/144562
33. 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.
34. Pichichero ME. Can machine learning and AI replace otoscopy for diagnosis of otitis media? *Pediatrics*. 2021;147(4):e2020049584.
35. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402.
36. Krause J, Gulshan V, Rahimy E, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*. 2018;125(8):1264-1272.
37. Gates GA, Cooper JC. Effect of anesthetic gases on middle ear pressure in the presence of effusion. *Ann Otol Rhinol Laryngol Suppl*. 1980;89(3 Pt 2):62-64.
38. Ghojogh B, Crowley M. The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial. *arXiv*. 2019;1905.12787. http://arxiv.org/abs/1905.12787
39. Hawkins DM, Basak SC, Mills D. Assessing model fit by cross-validation. *J Chem Inf Comput Sci*. 2003;579-586.
40. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. *Stat Surv*. 2010;4:40-79.
41. Alenezi EMA, Jajko K, Reid A, et al. The reliability of video otoscopy recordings and still images in the asynchronous diagnosis of middle-ear disease. *Int J Audiology*. 2021. doi:10.1080/14992027.2021.1983217.
42. 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.
43. Senaras C, Moberly AC, Teknos T, Essig G, Elmaraghy C, Taj-Schaal N, et al. Detection of eardrum abnormalities using ensemble deep learning approaches. In: Mori K, Petrick N, eds. *Medical Imaging 2018: Computer-Aided Diagnosis* [Internet]. SPIE; 2018:45. Accessed February 26, 2019. https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10575/2293297/Detection-of-eardrum-abnormalities-using-ensemble-deep-learning-approaches/10.1117/12.2293297.full
44. Tran TT, Fang TY, Pham VT, Lin C, Wang PC, Lo MT. Development of an automatic diagnostic algorithm for pediatric otitis media. *Otol Neurotol*. 2018;39(8):1060-1065.
45. Senaras C, Moberly AC, Teknos T, et al. Autoscope: automated otoscopy image analysis to diagnose ear pathology and use of clinically motivated eardrum features. *Med Imaging 2017: Compu-Aided Diagnosis*. 2017;10134(March 2017):101341X.
46. Myburgh HC, van Zijl WH, Swanepoel DW, Hellström S, Laurent C. Otitis media diagnosis for developing countries using tympanic membrane image-analysis. *EBioMedicine*. 2016;5:156-160.
47. Wang X, Valdez TA, Bi J. Detecting tympanostomy tubes from otoscopic images via offline and online training. *Comput Biol Med*. 2015;61:107–118. 10.1016/j.compbiomed.2015.03.025
48. Shie CK, Chang HT, Fan FC, et al. A hybrid feature-based segmentation and classification system for the computer aided self-diagnosis of otitis media. In: 2014 36th Annual International Conference of

the IEEE Engineering in Medicine and Biology Society, EMBC 2014. 2014:4655–4658.
49. Kuruvilla A, Shaikh N, Hoberman A, Kovačević J. Automated diagnosis of otitis media: vocabulary and grammar. *Int J Biomed Imaging*. 2013;2013:1–15.
50. Kuruvilla A, Li J, Yeomans PH, Quelhas P, Instituto I, Biomedica DE. Otitis media vocabulary and grammar. *IEEE*. 2012;2845–2848.
51. Mironică I, Vertan C, Gheorghe DC. Automatic pediatric otitis detection by classification of global image features. 2011 E-Health and Bioengineering Conference, EHB 2011. 2011:1–4.
52. Vertan C, Gheorghe DC, Ionescu B, Eardrum color content analysis in video-otoscopy images for the diagnosis support of pediatric otitis. ISSCS 2011 - International Symposium on Signals, Circuits and Systems, Proceedings. 2011;129–32.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Habib A-R, Kajbafzadeh M, Hasan Z, 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