Artificial intelligence to diagnose ear disease using otoscopic image analysis: a review

Therese L Canares,[1] Weiyao Wang,[2] Mathias Unberath,[2] James H Clark [3]

[1]Johns Hopkins University School of Medicine, Baltimore, Maryland, USA
[2]Johns Hopkins University Whiting School of Engineering, Baltimore, Maryland, USA
[3]Otolaryngology-HNS, Johns Hopkins Medicine School of Medicine, Baltimore, Maryland, USA

Correspondence to
Dr James H Clark, Otolaryngology-HNS, Johns Hopkins Medicine School of Medicine, Baltimore, MD 21224, USA; jclark79@jhmi.edu

Accepted 27 July 2021

To cite: Canares TL, Wang W, Unberath M, *et al*. *J Investig Med* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/jim-2021-001870

ABSTRACT

AI relates broadly to the science of developing computer systems to imitate human intelligence, thus allowing for the automation of tasks that would otherwise necessitate human cognition. Such technology has increasingly demonstrated capacity to outperform humans for functions relating to image recognition. Given the current lack of cost-effective confirmatory testing, accurate diagnosis and subsequent management depend on visual detection of characteristic findings during otoscope examination. The aim of this manuscript is to perform a comprehensive literature review and evaluate the potential application of artificial intelligence for the diagnosis of ear disease from otoscopic image analysis.

INTRODUCTION

Ear-related symptoms are the leading health-related concern expressed by parents in relation to their child's general health.[1] Even in the absence of ear-specific symptoms, parents frequently attribute behavioral changes in their child such as increased irritability and disrupted sleep to ear disease.[2] It is therefore unsurprising that ear-related concerns constitute the leading cause for seeking pediatric healthcare attention.[1]

Disease of the middle ear and external auditory canal represent a heterogeneous spectrum of pathological entities that beyond having some shared symptomatic overlap, can also present with constitutional symptoms such as fever, nausea or abdominal pain.[3] Clinical history may therefore be unrevealing in terms of underlying otological etiologies.[4] The current diagnostic 'gold standard' is highly reliant on the identification of pathognomonic findings during otoscopic examination given the absence of cost-effective clinical test. The diagnostic accuracy of ear disease is directly dependent on the exam proficiency and diagnostic skill and interpretative expertise of the otoscope operator.[5] The American Academy of Pediatric therefore stresses the importance of ensuring proficiency in ear exam, recommending that otoscopic training be initiated early during medical school and continuing throughout post-graduate training.[6] Medical student and junior physicians however have frequently been found to report lack of confidence in their ability to both examine and diagnose ear pathology.[7–10] Pichichero *et al* investigated diagnostic performance based on otoscope exam among a sizeable cohort of US pediatricians (n=2190) and general practitioners (n=360) and found to be 51% (±11) and 46% (±26), respectively ($p<0.0001$). Findings from this study further demonstrated a clear bias towards overdiagnosis of pathological ear disease.[3] Similar diagnostic performance has subsequently been replicated by a number of studies.[6 11–14]

In alignment with the bias toward overdiagnosis of ear disease, it is currently estimated that between 25% and 50% of all antibiotics prescribed for ear disease are not indicated.[13–15] Beyond risking unnecessary medical complications and the downstream unintended consequence of potential antibiotic resistance, overdiagnosis of ear disease adds an estimated US$59 million in unnecessary healthcare spending in the USA per annum.[16] In an effort to standardize appropriate diagnosis and treatment of pathological ear disease, a number of initiatives have been implemented, the most notable of which was the development of societal guidelines across otolaryngology and pediatrics for commonly encountered ear disease.[16–18] While the publication of clinical guidelines has provided much-needed evidence-based consensus relating to standardization of care, these guidelines have had limited impact on everyday clinical practice.[19–21] Actualizing change in clinical practice presents considerable challenges and relates to several reciprocal factors including clinicians' lack of awareness, familiarity, agreement, self-efficacy and outcome expectancy, in addition to the inertia of previous practice, and presence of external system barriers.[22] These factors lay the exciting groundwork for the role of artificial intelligence (AI), an emerging tool that may provide technological capacity to overcome these challenges by providing clinicians with direct medical decision guidance and feedback, thereby minimizing treatment variation and ensuring high-quality care delivery.[23]

AI relates broadly to the science of developing computer systems to imitate human intelligence, thus allowing for the automation of tasks that would otherwise necessitate human cognition.[24 25] While contemporary technology lacks the capacity to match or surpass general human intelligence, a form of AI known as narrow

artificial intelligence (NAI) has demonstrated proficiency to complete well-circumscribed subtasks without needing external (human) input.[26 27] Machine learning (ML) algorithms are among the most commonly applied form of NAI and will constitute the focus of this review.

ML algorithms are data analytic models that can learn automatically from previous experience without need for external input.[28] This functionality enables ML algorithms to be deployed to infer meaning or categorize data according to specific data traits, within structured data sources such as images.[29] The mathematical framework coded for by an ML algorithm is explicit but can be trained to process any presented data that is compatible. In fact, this generalizability has enabled the release of numerous open-source ML algorithm models online. Interested users can therefore develop their own AI tools simply by uploading training data to one of these open-source ML algorithms.[30 31]

During ML algorithm training, the parameters of the framework are fitted to the desired function, thereby enables the ML algorithm to infer meaning or categorize unseen data during deployment.[21] Training can be performed using a supervised, unsupervised or reinforced learning approach.[32] In supervised learning, ML training is performed using labeled datasets. Using a trial and error method, the ML algorithm learns to recognize the correct data trait necessary for the desired task. Unsupervised learning, in contrast, relies on the ML algorithm analyzing unlabeled data and categorizing the data according to inherent traits discovered within the training. Reinforced learning represents a hybrid approach, which relies on using both labeled and unlabeled data for training.

Contemporary ML algorithms have demonstrated functional capacity to equal that of human for tasks of image recognition and at times have even exceeded it.[26] This has motivated clinical application of this technology with ML algorithms being developed to automate numerous medical tasks, such as the reading of ECGs, the interpretation of radiological images and the diagnosis of skin lesions.[28 33 34] The aim of this manuscript is to perform a comprehensive literature review and evaluate the potential application of such ML algorithm for the diagnosis of ear disease from otoscopic image analysis.

METHODS

Search strategy

A literature search was conducted using PubMed (1953–2020), EMBASE (1974–2010), CINAHL (1982–2020), PsycINFO (1887–2020) and Web of Science (1945–2020), using the search strings: (Artificial Intelligence) AND (Ear Disease).

Study selection

After removing duplicated cases, the search results were imported into a reference management tool (Zotero, 5.0.96). The first author screened all titles and abstract. Inclusion criteria were titles and/or abstract containing the words “Artificial Intelligence” and terms related to “Middle or External Ear Disease”. The exclusion criterion included were non-English language, not peer reviewed, not using image analysis from clinical examination, or articles not presenting primary data. The references of all included articles were inspected for any relevant citations not discovered with our search strategy.

Data extraction and quality assessment

Data extraction and quality assessment was performed in accordance with Lou *et al*. Guidelines for developing and reporting ML predictive models in biomedical research: a multidisciplinary view.[35] Full and comprehensive review was sequentially completed by authors JHC and WW of all articles meeting criteria for inclusion.

Data synthesis and analysis

Each article was summarized in a Microsoft Word table detailing article type, data input, ML design, diagnosis used, image capture device, training of and number of image annotators, image pixel size, size of training dataset, reported diagnostic performance and area under the receiver operating characteristic curve (AUROC). The ad hoc nature of reported outcomes prevented further analysis beyond description.

RESULTS

The literature search strategy yielded 1862 citations, of which 9 manuscripts were eligible for review (figure 1). All included manuscripts detail the development of AI algorithms with the capacity to diagnose ear disease from a single photographic image without needing external input. The disease processes that the algorithms were trained to diagnose varied considerably between groups (table 1).

Selection of AI method

ML, a class of AI algorithms, was used in the development of all nine diagnostic algorithms.[36–44] Of the nine algorithms, six were developed using a form of ML known as Deep Neural Networks.[36 38–42] The remaining three algorithms were developed using a variety of commonly used forms of ML models (Support Vector Machine, k-Nearest Neighbor and Decision Trees).[37 43 44]

AI algorithm training

All algorithms were trained using a similar method which necessitates the creation of image databases. Database images consisted of representative images of the chosen trait and have been annotated as being that diagnosis. Multiple strategies were applied for image collection, with five of the groups collecting these data in a prospective fashion[36–38 40 44] and three relying on previously established image databases.[41–43] One group relied entirely on images from Google Image search to create their database while Livingstone and Chau supplemented their database with images collected from Google Search and Textbooks.[38 39] A variety of devices including digital otoscopes and endoscopes were used for image capture (table 1). Image size was stated in four manuscripts, and ranged from 224×224 pixels to 486×486 pixels.[36 37 42 44] Annotation of training data was performed by ear specialists, which was defined consistently as being an otolaryngologist or an otologist.[36–44] A cohort consisting of two ear specialists was used in seven manuscripts and image inclusion to the database required diagnostic agreement by both ear specialist.[36 38–41 43 44] In the remaining two manuscripts, annotation was performed

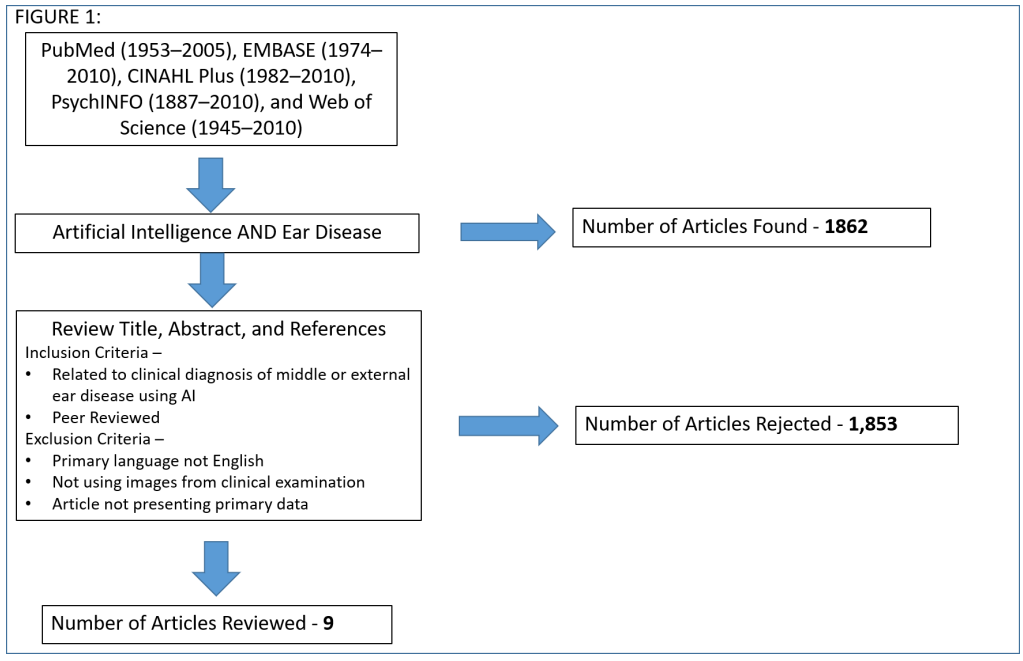

Figure 1 Flow chart of article selection from the literature search strategy.

by a single otolaryngologist.[37 42] The size of the database used for training varied between manuscripts ranging from 183 to 8435 images.[39 42]

AI algorithm testing

In eight of the manuscripts, a cohort of representative, non-annotated images were reserved for testing and not included in algorithm training.[36–38 40–44] The cohort of images was independently presented for inference of diagnosis by both the AI algorithms and the same cohort of ear specialists used to annotate the training data (see table 1 for list of diagnosis used within each manuscript). The AI algorithm's diagnostic performance was then rated by comparing the algorithms inferred diagnostic results with those of the ear specialist. Using this methodology, the diagnostic accuracy performance of the eight ML algorithms was reported for a variety of trained diagnosis ranging between 80.6% and 98.7%.[41 44] There was considerable variation in the sensitivity (recall) and the positive predictive value (precision) among the algorithms for their selected diagnosis, which ranged between 50.0%–100% and 14.3%–100%, respectively.[38] Of these eight algorithms, the AUROC score was reported in four and ranged between 0.91 and 1.0.[37 43]

Habib *et al* tested their AI algorithm using an image database that was not used during training. Despite the variation in image quality between the images used in training and those used for testing the algorithm achieved a diagnostic accuracy performance average of 76% with an AUROC of 0.86.[39]

AI algorithm comparison with non-ear specialist

Two manuscripts further tested their AI by comparing the diagnostic performance of the algorithm in comparison to a cohort of non-ear specialist clinicians. Both manuscripts report their AI algorithm's as surpassing the diagnostic performance of the non-ear specialist cohort (table 2).[36 38]

Using a different approach, Myburgh *et al* created a rudimentary but cost-effective video-otoscope that was deployed with an experienced general practitioner to trial during routine shifts in a South African emergency room. Captured images from this device were then transferred for independent analysis by a computer with the AI algorithm. The results of the algorithms were then compared with the correct result, which the group defined as the diagnosis inferred by the general practitioner. In this small pilot study, the diagnostic accuracy of the AI algorithm was determined to be 78.7%.[44]

DISCUSSION

In this review, we identified nine manuscripts that provide small proof-of-concept studies for the application of ML algorithms in the diagnosis of ear disease from an image captured during an otoscopic exam. The study designs of the manuscripts however largely fail to demonstrate meaningful performance validation of the ML algorithms, and include a lack of comparison of the ML algorithms with current care standards in a clinical setting. Attempts to use this literature to contemplate the clinical potential of such ML algorithms are therefore significantly hampered by the paucity of details relating to pathways for scaling the technology. Furthermore, and perhaps more significantly there is a fundamental failure in providing a specific outline for how such technology will fit within the current model of healthcare delivery.

The manuscripts included in this review relied on an ML method to algorithm development and a supervised learning approach to training. In this approach, the algorithms were presented a group of annotated images depicting the pathognomonic appearance of a specific diagnosis. To adhere to ML terminology, henceforth the term 'domain' will replace the word 'diagnosis' in a synonymous fashion. The term 'ground truth' is often the nomenclature used to describe

Table 1 Design and outcome reported in the literature in relation to the application of artificial intelligence (AI) to diagnose ear disease using otoscope image analysis[36–44]

Study	Data input	ML design	Diagnosis	Capture device	Annotator (n)	Pixels	Training dataset size (N)	Accuracy (%)	Precision (%)	Recall (%)	AUROC
Khan *et al*[36]	Single image	Deep Neural Networks	Normal exam Chronic otitis media Otitis media with effusion	Endoscope	Otolaryngologist[2]	224×224	2484	96.2 for normal exam 96.2 for chronic otitis media 92.6 for otitis media with effusion	94.6 for normal exam 96.2 for chronic otitis media 94.9 for otitis media with effusion	96.2 for normal exam 96.2 for chronic otitis media 92.6 for otitis media with effusion	0.99
Viscaino *et al*[37]	Single image	SVM (k-NN/ Decision Tree)	Normal exam Myringosclerosis Earwax plug Chronic otitis media	Digital otoscope	Otolaryngologist[1]	420×380	576	Mean 93.9	Mean 87.7	Mean 87.8	1.00
Livingstone and Chau[38]	Single image	Deep Neural Networks	Normal Acute otitis media Cerumen Cholesteatoma Exostoses Myringitis Myringosclerosis Otitis externa Otomycosis Serous otitis media TM perforation TM retraction PE tube in position PE tube extruded	Endoscope/ Google /textbooks/ databases	Otologist[1] PGY-5 Otolaryngology Resident[1]	N/A	1277	N/A	93.1 for normal 100 for acute otitis media 90 for cerumen 100 for cholesteatoma 100 for exostoses 100 for myringitis 100 for myringosclerosis 91.7 for otitis externa 14.3 for otomycosis 90.9 for serous otitis media 83.3 for TM perforation 100 for TM retraction 100 for PE tube in position 100 for PE tube in position	93.1 for normal 75.0 for acute otitis media 93.1 for cerumen 50.0 for cholesteatoma 100 for exostoses 100 for myringitis 72.2 for myringosclerosis 100 for otitis externa 100 for otomycosis 90.9 for serous otitis media 55.6 for TM perforation 88.9 for TM retraction 60.0 for PE tube in position 100 for PE tube extruded	N/A
Habib *et al*[39]	Single image	Deep Neural Networks	Intact tympanic membrane Small TM perforation Medium TM perforation Large TM perforation	Google	Otolaryngologist[2]	N/A	183	76.0 for intact TM perforation 85.7 for small TM perforation 85.7 for medium TM perforation 63.6 large TM perforation	N/A	N/A	0.87
Livingstone *et al*[40]	Single image	Deep Neural Networks	Normal PE tube Cerumen impaction	Digital otoscope	Otolaryngologist[2]	N/A	529	Mean 84.4	N/A	N/A	N/A

Continued

Table 1 Continued

Study	Data input	ML design	Diagnosis	Capture device	Annotator (n)	Pixels	Training dataset size (N)	Accuracy (%)	Precision (%)	Recall (%)	AUROC
Lee *et al*[41]	Single image	Deep Neural Networks	Laterality	Endoscope	Otologist[2]	N/A	1338	97.9 for correct laterality	96.9 for correct laterality	93.3 for correct laterality	0.92 (0.98*)
			TM perforation No TM perforation					91.0 for presence or absence of perforation	98.0 for presence or absence of perforation	90.5 for presence or absence of perforation	
Cha *et al*[42]	Single image	Deep Neural Networks	Normal	Endoscope	Otolaryngologist[2]	640×480	8435	97.5 for normal	N/A	94.0 for normal	N/A
			TM attic retraction					85.8 for TM attic retraction		90.2 for TM attic retraction	
			Tympanic perforation					97.2 for TM perforation		96.5 for TM perforation	
			Otitis externa with or without myringitis					77.9 for otitis externa with or without myringitis		89.3 for otitis externa with or without myringitis	
			Otitis media with effusion					88.5 for otitis media with effusion		93.3 for otitis media with effusion	
			Tumor					86.7 for tumor or cerumen		93.4 for tumor or cerumen	
Tran *et al*[43]	Single image	MTJSRC	Acute otitis media	Digital otoscope	Otologist[2]	N/A	171	91.9 for acute otitis media	N/A	89.4 for acute otitis media	0.91
			Otitis media with effusion					91.3 for otitis media with effusion		93.3 for otitis media with effusion	
Myburgh *et al*[44]	Single image	Decision Tree	Normal exam	Digital otoscope	Otologist[2]	500×500	391	80.6 (78.7†)	83 for normal exam	80 for normal exam	N/A
			Earwax plug						89 for earwax/foreign body	79 for earwax/foreign body	
			Foreign body						75 for acute otitis media	81 for acute otitis media	
			Acute otitis media						80 for otitis media with effusion	81 for otitis media with effusion	
			Otitis media with effusion						78 for chronic suppurative otitis media	82 for chronic suppurative otitis media	
			Chronic suppurative otitis media								

*AUROC performance relative to tympanic membrane laterality.
†Testing accuracy using self-made otoscope.
AUROC, area under the receiver operating characteristic curve; k-NN, k-Nearest Neighbor; MTJSRC, Multitask Joint Sparse Representation-Base Classification; N/A, not available; PE tube, pressure equalizing tubes; SVM, Support Vector Machine; TM, tympanic membrane.

Table 2 Diagnostic performance of non-ear specialist versus machine learning (ML) algorithm[36 38]

Study	Clinician cohort (n)	Non-ear specialist diagnostic performance	ML algorithm diagnostic performance
Khan *et al*[36]	Specialists (7)* Residents (6)* Interns (4)*	74.0%	87.0%
Livingstone and Chau[38]	Otolaryngology residents (5) Pediatric medicine resident (1) Internal medicine resident (1) Emergency medicine resident (1) General practitioners (2)	58.9%	88.7%

*No further detail provided in manuscript.

the labeled data used to train the algorithm to recognize a characteristic data pattern that occurs within the data that is specific to that domain. Once the ML algorithm is trained, it codes a computer program that provides a mathematical framework that enables computer systems to analyze previously unseen, non-labeled data. Running such a program enables computation either of the 'presence' or 'absence' of a trained domain in the case of a non-predictive model such as the Decision Tree, or the 'statistical likelihood' of a trained domain occurring within the data in case of a predictive model such as deep learning (DL). As demonstrated in this review, despite ML algorithms coding for different mathematical frameworks the various designs can be trained to perform the same functional task. DL is the most contemporary form of ML, and represents the design most commonly selected by the included manuscripts. As such, this form of ML algorithm will be described in greater detail.

DL designs demonstrate a great capacity for discovering intrinsic patterns within structured data and can be used to directly analyze pixel intensity. Deploying a DL design (using a supervised learning approach to training) therefore enables the algorithm to simply be presented with the desired ground truth which in this case would consist of otoscopic images stratified and labeled according to a specific domain. Next, without external input, the algorithm performs image analysis, which enables it to discover intrinsic pixel patterns within the image, specific to that domain. The advantage of this is that it negates the previous requirement for ML developers to manually abstract a data pattern for the algorithm to use. For tasks relating to image recognition, one of the most common forms of DL deployed is called Convolutional Neural Networks (CNN).[45] CNN performs image recognition tasks by extracting hierarchical features from images in segments termed, convolutional layers. The network is composed of learnable parameters (eg, filters in the convolutional layers) that are developed during training with labeled data. Once trained, the accuracy of the algorithm is further refined by presenting unlabeled training data and adding weights within the model that serve to increase the likelihood of the algorithm conferring the correct diagnosis.[46] The development of these models, therefore, requires a large quantity of data. A further disadvantage of DL models is that training and refinement can be technically challenging. Within training data, there are multiple millions of trainable parameters, presenting significant challenges for a developer to know whether the algorithm is using the optimal parameter within the data. There is also a need to balance the number of convolutional layers used with the targeted algorithm performance. For example, with a properly engineered structure, a larger number of convolutional layers can potentially improve the prediction accuracy and increase the training and processing time for images because more computation is necessary. More traditional ML algorithm designs that were also included in this review provide ML developers with differing advantages and disadvantages as outlined in table 3.

The selection of an ML algorithm design depends on a multitude of factors including the data being used (format, complexity and quantity), the planned approach to training, and most importantly, the algorithm's performance accuracy in predicting the desired outcome.[45]

The diagnostic performance data in a controlled setting was provided by all the included manuscripts. Study design for algorithm testing was uniformed across manuscripts with accuracy, precision and recall being determined by comparing domain prediction of the ML algorithm (for previously unseen and unlabeled images) against those domains provided for the same images by a cohort of ear specialists. It should be noted that in all groups the cohort of ear specialists was the same for both algorithm training and testing. Using these performance metrics, the ML models demonstrated a high level of diagnostic accuracy (76%–95%), precision (83%–95%) and recall (79%–95%) for certain trained domains (table 1). In the process of developing their algorithm, Viscaino *et al* noted that the trial of three different ML algorithm designs, with Support Vector Machine and k-Nearest Neighbor demonstrated superior performance compared with that of the Decision Tree classification.[37] Five of the included manuscripts also reported AUROC scores. An AUROC score serves to chracterize the ML algorithm's capacity to distinguish between a non-disease state and a disease state under tradeoff between sensitivity and specificity using different decision thresholds. The AUROC scores reported in this review range between 0.86 and 0.99. The closer an AUROC score is to 1.00, the greater the discriminatory ability the ML algorithm has, with 1.0 meaning that the algorithm is able to reach 1.0 sensitivity and 1.0 specificity at the same time.[37 39] Performance comparison between the nine diagnostic ML algorithms is inappropriate given that each was developed using different data quality and diagnosis selection. As a result, each of the ML algorithms should be considered as performing a function unique to itself, which will differ in the level of complexity relative to the function of the other included ML algorithms. In addition to algorithm testing data, two manuscripts compared the diagnostic performance of non-ear specialists with their ML algorithms in a

Table 3 Basic principle and comparison of included machine learning (ML) model design

ML model	Basic principle	Advantage	Disadvantage
Convolutional Neural Networks	Performs data recognition by extracting hierarchical features from data with convolutional layers. The network is composed of trained parameters that are learnt from labeled data.	1. No input abstraction required 2. Applied directly to pixels	1. Requires large amount of data 2. Complex to training and refine 3. Millions of trainable data parameters
Decision Tree classification	Relies on a branching structure where each branch (node), in a binary fashion, directs to a specific outcome (leaf).	1. Normalization or scaling of data non- needed 2. No considerable impact of missing values 3. Easy to explain and visualize	1. Quickly becomes overly complex with multiple domains 2. Challenging to manage data outliers
k-Nearest Neighbor	Determine the distance between plotted unlabeled data relative to labeled data. The unlabeled data are classified to share the domain of the nearest labeled neighboring data.	1. Very simple 2. No assumption about data 3. Can solve multidomain problems	1. Cumbersome on large datasets 2. Suboptimal for datasets with a large number of domains 3. Sensitive to data outliers
Support Vector Machine	Labeled data are plotted with each domain being represented by a particular set of coordinates. The algorithm then calculates optimal hyperplanes, which function as lines that best separate the domains. Then depending on the laterality unlabeled data fall relatively to this line serves to determine the domain the data are classified as representing.	1. Perform well with multiple different data domains present 2. Data outliers have little impact on performance	1. For large dataset, requires a significant amount of processing time 2. Poor performance if domains overlap is present 3. Training can be challenging

non-clinical setting. This was accomplished in both manuscripts by comparing both the non-ear specialist group's and algorithm's diagnostic inference from otoscopic captured images with that of an ear specialist cohort who served as the control. Both manuscripts report that the ML algorithm diagnostic inference outperforms that of the non-ear specialist group.[36 37] Caution, however, is needed in the interpretation of these findings, as the design of the study was suboptimal due to it being performed in a non-clinical setting, and with a reliance on small sample size and incomplete data. Furthermore, the non-specialist cohorts demonstrated significant variation in the level of both training and experience resulting in considerable data spread.

On review of the reported testing data, an argument could be made that there is literature to support that these ML algorithms already demonstrate the capacity to outperform the diagnostic efforts of a non-specialist. In particular, Pitchichero *et al* investigated the diagnostic accuracy of pediatricians and general practitioners for a normal ear exam, acute otitis media or otitis media with effusion after viewing an otoscopic exam video. This study found a fair diagnostic accuracy of 51% (±11) and 46% (±21) for pediatricians and general practitioners, respectively.[3] When comparing this with the reported diagnostic performance of the ML algorithms trained to recognize these three diagnoses, the results (78%–95%) surpass those of both pediatricians and general practitioners.[36 38 43 44] The validity of this argument is uncertain however as the study by Pitchichero *et al* and the ML algorithms were performed in controlled settings that are unlikely to be encountered in clinical practice. Furthermore, the success of an ML algorithm is dependent on many factors in addition to diagnostic performance. This is clearly demonstrated when considering that despite generating considerable excitement within healthcare and the widespread, rapid emergence of increasingly accurate ML algorithms, the clinical adoption of such technology has not occurred at nearly the same pace.[47 48]

One of the large challenges with developing ML algorithms is the process of scaling the technology beyond the laboratory. As previously described, the functionality of an ML algorithm is dependent on adherence to a predefined model. The models are fitted during the training of the ML algorithm and enable inference of specific domains once deployed. Hence, meticulous attention and foresight is therefore required at this stage to ensure that the characteristics patterns used for the training of the ML algorithm are universally agreed on as being representative of the selected chosen domain and that the characteristics patterns are representative of the selected diagnosis typically encountered in the clinical setting.[49] As well as not detailing a pathway to scaling, the non-standardization approach to data collection and the methodology employed by the reviewed studies also increase the risk of the reported ML algorithm demonstrating limited widespread applicability. In addition, scaling this technology beyond the laboratory is likely to face further challenges during deployment if the data used for algorithm training are not of comparable quality with that captured in a clinical setting. Given that ML algorithms rely on the characteristics of the elements that make up an image (pixels) to infer a diagnosis, any change due to variation in image captures such as using a different image capturing system or variable image capture settings, will adversely impact the algorithm's performance. This could represent considerable challenges to the ability to scale this ML algorithm-based technology given that a large percentage of clinicians remain without access to digital otoscopes, and if digital otoscopes are being used there is still the inherent risk of variation in image acquisition and quality, which would confound diagnostic accuracy.

Beyond functionality and challenges of technological scalability, perhaps the more fundamental, unanswered question that remains is how such technology will integrate with the current healthcare delivery model. To date, a common crux within AI development is related to

innovation, which remains outside of the core processes that drive care delivery.[48 50] For example, it remains to be determined whether the relatively poor diagnostic accuracy and excessive antibiotic prescribing practices are important enough to practitioners to motivate widespread adoption of this emerging technology and investment of the associated monetary costs.[51] There is also a need to better understand any objective factors that influence why clinicians make decisions as this will also impact the value of this technology. For example, if clinicians, in part, prescribe antibiotics due to the expectation of a concerned parent, then it is unlikely that this practice will change even if this technology is implemented. The recent trend towards telemedicine is also likely to present uncertainty for the successful implementation of this technology, as this will require ear exams to be performed by either a parent or guardian.

Several limitations of this review should be considered. First, the manuscripts included in this review use relatively small sample sizes, ad hoc methodology and variable outcomes, which limit the ability to generalize findings. Second, as highlighted above, the performance of the algorithms is specific to a controlled setting and might not represent actual clinical performance. Third, the method by which such technology can be clinically deployed is influenced by a number of variable factors, and the role of AI diagnostic tools within the current healthcare workflow remains unknown.

CONCLUSION

The current literature provides some proof of evidence supporting the capacity of AI to diagnose ear disease with otoscope image analysis. This work, however, remains in its infancy, and there is a need for well-designed prospective clinical studies before the potential of such AI technology can fully be elucidated.

Contributors All listed authors were involved in designing and writing of the manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Commissioned; externally peer reviewed.

ORCID iD

James H Clark http://orcid.org/0000-0002-8841-3402

REFERENCES

1 Statistical brief #228: ear infections (Otitis Media) in children (0-17): use and expenditures, 2006. Available: https://www.meps.ahrq.gov/data_files/publications/st228/stat228.shtml [Accessed 23 Jun 2020].
2 Meherali S, Campbell A, Hartling L, *et al*. Understanding parents' experiences and information needs on pediatric acute otitis media: a qualitative study. *J Patient Exp* 2019;6:53–61.
3 Pichichero ME, Poole MD. Comparison of performance by otolaryngologists, pediatricians, and general practioners on an otoendoscopic diagnostic video examination. *Int J Pediatr Otorhinolaryngol* 2005;69:361–6.
4 Minovi A, Dazert S. Diseases of the middle ear in childhood. *GMS Curr Top Otorhinolaryngol Head Neck Surg* 2014;13:Doc11.
5 Guldager MJ, Melchiors J, Andersen SAW. Development and validation of an assessment tool for technical skills in handheld Otoscopy. *Ann Otol Rhinol Laryngol* 2020;129:715–21.
6 Paul CR, Keeley MG, Rebella G, *et al*. Standardized checklist for Otoscopy performance evaluation: a validation study of a tool to assess pediatric Otoscopy skills. *MedEdPORTAL* 2016;12:10432.
7 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;73:129–35.
8 Paul CR, Higgins Joyce AD, Beck Dallaghan GL, *et al*. Teaching pediatric otoscopy skills to the medical student in the clinical setting: preceptor perspectives and practice. *BMC Med Educ* 2020;20:429.
9 Higgins Joyce A, Raman M, Beaumont JL, *et al*. A survey comparison of educational interventions for teaching pneumatic otoscopy to medical students. *BMC Med Educ* 2019;19:79.
10 Paul CR, Keeley MG, Rebella GS, *et al*. Teaching pediatric Otoscopy skills to pediatric and emergency medicine residents: a Cross-Institutional study. *Acad Pediatr* 2018;18:692–7.
11 Buchanan CM, Pothier DD. Recognition of paediatric otopathology by general practitioners. *Int J Pediatr Otorhinolaryngol* 2008;72:669–73.
12 You P, Chahine S, Husein M. Improving learning and confidence through small group, structured otoscopy teaching: a prospective interventional study. *J Otolaryngol Head Neck Surg* 2017;46:68.
13 Gurnaney H, Spor D, Johnson DG, *et al*. Diagnostic accuracy and the observation option in acute otitis media: the capital region otitis project. *Int J Pediatr Otorhinolaryngol* 2004;68:1315–25.
14 Brinker DL, MacGeorge EL, Hackman N. Diagnostic accuracy, prescription behavior, and watchful waiting efficacy for pediatric acute otitis media. *Clin Pediatr* 2019;58:60–5.
15 Poole NM, Shapiro DJ, Fleming-Dutra KE, *et al*. Antibiotic prescribing for children in United States emergency departments: 2009-2014. *Pediatrics* 2019;143:e20181056.
16 Rosenfeld RM, Shin JJ, Schwartz SR, *et al*. Clinical practice guideline: otitis media with effusion (update). *Otolaryngol Head Neck Surg* 2016;154:S1–41.
17 Lieberthal AS, Carroll AE, Chonmaitree T, *et al*. The diagnosis and management of acute otitis media. *Pediatrics* 2013;131:e964–99.
18 American Academy of Otolaryngology–Head and Neck Surgery. Clinical practice guideline: acute otitis externa [Internet], 2014. Available: https://www.entnet.org/content/clinical-practice-guideline-acute-otitis-externa [Accessed 07 May 2021].
19 Mildenhall N, Honeybrook A, Risoli T, *et al*. Clinician adherence to the clinical practice guideline: acute otitis externa. *Laryngoscope* 2020;130:1565–71.
20 Haggard M. Poor adherence to antibiotic prescribing guidelines in acute otitis media--obstacles, implications, and possible solutions. *Eur J Pediatr* 2011;170:323–32.
21 Forrest CB, Fiks AG, Bailey LC, *et al*. Improving adherence to otitis media guidelines with clinical decision support and physician feedback. *Pediatrics* 2013;131:e1071–81.
22 Cabana MD, Rand CS, Powe NR, *et al*. Why don't physicians follow clinical practice guidelines? A framework for improvement. *JAMA* 1999;282:1458–65.
23 Fogel AL, Kvedar JC. Artificial intelligence powers digital medicine. *NPJ Digit Med* 2018;1:1–4.
24 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25:44–56.
25 Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J* 2019;6:94–8.
26 Grant AE, Meadows JH. *Communication technology update and fundamentals*. 17 edn. Routledge, 2020: 535.
27 Ahmad HM, Khan MJ, Yousaf A, *et al*. Deep learning: a breakthrough in medical imaging. *Curr Med Imaging* 2020;16:946–56.
28 Chartrand G, Cheng PM, Vorontsov E, *et al*. Deep learning: a primer for radiologists. *Radiographics* 2017;37:2113–31.
29 Goldberg SB, Flemotomos N, Martinez VR, *et al*. Machine learning and natural language processing in psychotherapy research: alliance as example use case. *J Couns Psychol* 2020;67:438–48.
30 Apple Developer. Machine Learning [Internet]. Available: https://developer.apple.com/machine-learning/ [Accessed 31 Jan 2021].
31 Google AI. Tools [Internet]. Available: https://ai.google/tools/ [Accessed 31 Jan 2021].
32 Molnar C. *Interpretable machine learning*. Lulu.com, 2020: 320.
33 Willems JL, Abreu-Lima C, Arnaud P, *et al*. The diagnostic performance of computer programs for the interpretation of electrocardiograms. *N Engl J Med* 1991;325:1767–73.
34 De A, Sarda A, Gupta S, *et al*. Use of artificial intelligence in dermatology. *Indian J Dermatol* 2020;65:352–7.
35 Luo W, Phung D, Tran T, *et al*. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res* 2016;18:e5870.
36 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;126:384–94.

37 Viscaino M, Maass JC, Delano PH, *et al*. Computer-Aided diagnosis of external and middle ear conditions: a machine learning approach. *PLoS One* . 2020;15:e0229226.
38 Livingstone D, Chau J. Otoscopic diagnosis using computer vision: an automated machine learning approach. *Laryngoscope* 2020;130:1408–13.
39 Habib A-R, Wong E, Sacks R, *et al*. Artificial intelligence to detect tympanic membrane perforations. *J Laryngol Otol* 2020;134:311–5.
40 Livingstone D, Talai AS, Chau J, *et al*. Building an Otoscopic screening prototype tool using deep learning. *J Otolaryngol Head Neck Surg* 2019;48:1–5.
41 Lee JY, Choi S-H, Chung JW. Automated classification of the tympanic membrane using a Convolutional neural network. *Appl Sci* 2019;9:1827.
42 Cha D, Pae C, Seong S-B, *et al*. Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database. *EBioMedicine* 2019;45:606–14.
43 Tran T-T, Fang T-Y, Pham V-T, *et al*. Development of an automatic diagnostic algorithm for pediatric otitis media. *Otol Neurotol* 2018;39:1060–5.
44 Myburgh HC, van Zijl WH, Swanepoel D, *et al*. Otitis media diagnosis for developing countries using tympanic membrane Image-Analysis. *EBioMedicine* 2016;5:156–60.
45 Murray NM, Unberath M, Hager GD, *et al*. Artificial intelligence to diagnose ischemic stroke and identify large vessel occlusions: a systematic review. *J Neurointerv Surg* 2020;12:156–64.
46 Brownlee J. *Better deep learning: train faster, reduce Overfitting, and make better predictions*. Machine Learning Mastery, 2018: 575.
47 Emanuel EJ, Wachter RM. Artificial intelligence in health care: will the value match the hype? *JAMA* 2019;321:2281–2.
48 Li RC, Asch SM, Shah NH. Developing a delivery science for artificial intelligence in healthcare. *NPJ Digit Med* 2020;3:107.
49 Unberath M, Ghobadi K, Levin S, *et al*. Artificial Intelligence-Based clinical decision support for COVID-19-Where art thou? *Adv Intell Syst* 2020:2000104.
50 Schulman KA, Richman BD. Toward an effective innovation agenda. *N Engl J Med* 2019;380:900–1.
51 Pichichero ME. Can machine learning and AI replace Otoscopy for diagnosis of otitis media? *Pediatrics* 2021;147:e2020049584.