


BMJ Open 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

Yuexin Cai,^{1,2} Jin-Gang Yu,³ Yuebo Chen,^{1,2} Chu Liu,^{1,2} Lichao Xiao,³ Emad M Grais,⁴ Fei Zhao,⁴ Liping Lan,^{1,2} Shengxin Zeng,^{1,2} Junbo Zeng,^{1,2} Minjian Wu,^{1,2} Yuejia Su,^{1,2} Yuanqing Li,³ Yiqing Zheng ^{1,2}

To cite: Cai Y, Yu J-G, 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. doi:10.1136/bmjopen-2020-041139

► Prepublication history for this paper is available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2020-041139>).

YXC, J-GY and YBC contributed equally.

Received 12 June 2020
Revised 18 December 2020
Accepted 28 December 2020



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to

Dr Yiqing Zheng;
zhengyiq@mail.sysu.edu.cn

ABSTRACT

Objectives This study investigated the usefulness and performance of a two-stage attention-aware convolutional neural network (CNN) for the automated diagnosis of otitis media from tympanic membrane (TM) images.

Design A classification model development and validation study in ears with otitis media based on otoscopic TM images. Two commonly used CNNs were trained and evaluated on the dataset. On the basis of a Class Activation Map (CAM), a two-stage classification pipeline was developed to improve accuracy and reliability, and simulate an expert reading the TM images.

Setting and participants This is a retrospective study using otoendoscopic images obtained from the Department of Otorhinolaryngology in China. A dataset was generated with 6066 otoscopic images from 2022 participants comprising four kinds of TM images, that is, normal eardrum, otitis media with effusion (OME) and two stages of chronic suppurative otitis media (CSOM).

Results The proposed method achieved an overall accuracy of 93.4% using ResNet50 as the backbone network in a threefold cross-validation. The F1 Score of classification for normal images was 94.3%, and 96.8% for OME. There was a small difference between the active and inactive status of CSOM, achieving 91.7% and 82.4% F1 scores, respectively. The results demonstrate a classification performance equivalent to the diagnosis level of an associate professor in otolaryngology.

Conclusions CNNs provide a useful and effective tool for the automated classification of TM images. In addition, having a weakly supervised method such as CAM can help the network focus on discriminative parts of the image and improve performance with a relatively small database. This two-stage method is beneficial to improve the accuracy of diagnosis of otitis media for junior otolaryngologists and physicians in other disciplines.

Strengths and limitations of this study

- The two-stage approach for model development attempts to replicate the human visual attention system and the procedure of clinical judgement made by an experienced otolaryngologist undertaking examination of the tympanic membrane from whole view to partial image.
- The prediction results of the two models, each having different functionality and strength, were averaged to use the advantages of each.
- Non-medical history and hearing information were provided to the deep learning model and the otolaryngologists, which may have compromised the diagnosis accuracy.
- All the otoscopic images were taken after cerumen removed if needed.

INTRODUCTION

Otitis media (OM) is a common otological disease with a number of forms; acute otitis media (AOM), otitis media with effusion (OME) and chronic suppurative otitis media (CSOM), that appropriate medical care can treat.¹ It is a primary cause for people to seek medical care, antibiotic prescription and surgery.² Without early diagnosis and appropriate treatment, deterioration and even irreversible complications may occur.³ History taking, otoscopy and otoendoscopy examination are essential first steps in the evaluation of patients with OM in Ear, Nose and Throat and Audiology Clinics. The clear view or visual image obtained from otoscopy or otoendoscopy provides important information for initial diagnosis of otological diseases.¹⁻³

Misdiagnosis is most likely to occur if the medical doctor lacks experience in the use of otoscopy or otoendoscopy.^{4,5} For example, Sorrento and Pichichero⁶ found that the correct diagnosis rate of OM by paediatricians was only 50%, in comparison with 73% by otolaryngologists. Poor diagnostic accuracy leads to misdiagnosis and delay in treatment, which may cause preventable complications.^{2,7,8} A new diagnostic strategy for patients with OM needs to be developed in order to improve diagnostic accuracy.

In recent years, Artificial Intelligence has been applied to medical image analysis to help clinical interpretation and medical diagnosis by image classification, segmentation and matching.⁷⁻¹⁰ In particular, convolutional neural networks (CNNs) have been widely used and demonstrate good performance in the automated classification of medical images, including diabetic retinopathy detection,^{9,10} skin cancer classification^{11,12} and congenital cataract detection.¹³ However, few machine-learning studies have been conducted in the field of otology for the automated diagnosis of ear diseases from otoscopic images. Myburgh *et al*¹⁴ built up a neural network using 389 images to classify five categories of video-otoscopic image, including normal tympanic membrane (TM), obstructing wax or foreign body in the external ear canal, AOM, OME and CSOM, achieving a classification accuracy of 86.84%. Recently a machine-learning model was generated by combining two of the best performing models (Inception-V3 and ResNet101).⁷ The model used 10 544 otoendoscopic images to classify six categories of ear disease; normal, attic retraction, tympanic perforation, otitis externa with myringitis, otitis externa without myringitis and tumour. This learning model achieved an accuracy of 93.67%. However, the reliability of the combined classification method appears questionable since the simple aggregation of two independent models to illustrate reliability and interpretability is hard to justify with no effective underlying rationale. Somewhat differently, Lee *et al*¹ developed a heat map using a CNN model to distinguish between the normal middle ear condition and ears with chronic OM in an inactive phase to detect the presence of perforation. The accuracy was 91.0%, however the CNN model had poor consistency in determining the perforated area of the TM, which might be due to the relatively small sample. In addition, the model has not been improved through validation.

In the present study, an algorithm that combined results from two CNNs used in two separate stages was developed for the classification of: normal eardrum, OME and CSOM in an active or inactive phase. The network in the first stage dealt with the whole image, while in the second stage the network made its decision based on a discriminative segment of the TM image. Instead of using manual annotation to locate the discriminative part, an interpretation method called Class Activation Maps (CAMs)¹⁵ was used to identify the discriminative segment automatically and without extra annotation. The rationale underlying the two-stage approach was based on the human visual

attention system together with the procedure of clinical judgement made by an experienced otolaryngologist when they undertake examination of the TM from whole view to partial image. The human visual attention system selectively concentrates on parts of the visual space to capture salient information rather than processing the whole scene.¹⁶ Otolaryngologists can easily identify the salient lesion areas in an otoscopic image, mainly due to the attention mechanism of the human visual perception system.^{17,18} Lesions in otoscopic images always attract most of the otolaryngologist's attention during the physical examination on patients. This strategy of attention-based CNN has been previously applied to several other medical image analyses.^{17,19-23} Therefore, it is logical to incorporate the lesion attention mechanism into the otoscopic image classification models.

MATERIALS AND METHODS

Otoendoscopic image data acquisition

Images were collected retrospectively from 2022 patients who had attended the Department of Otorhinolaryngology, Sun Yat-sen Memorial Hospital, University of Sun Yat-sen, China between 2015 and 2019. The anonymous images were identified and categorised by using the clinical diagnostic information. The four conditions of the middle ear included in this study were: normal middle ear; OME; CSOM in active phase (CSOMa); CSOM in inactive phase (CSOMi).

Normal TM images were obtained from healthy ears in subjects with normal hearing thresholds determined by pure tone audiometry and a type A tympanogram. All patients with OME were confirmed by a type B tympanogram. CSOMa was further confirmed if the patient reported increased otorrhoea at the time the images were taken.

TM image data acquisition

TM images were taken by otolaryngologists using a 4 mm STORZ 0° endoscope (KARL STORZ, Germany) and a video-recording system. All images were taken before any surgery. Images were saved as JPEG graphic files with pixels in the range 500×500–700×700. For each of the 2022 patients, 3 images were chosen from each patient, and a total of 6066 images were included and categorised to four conditions: 1040 images of normal TM, 2613 images of OME, 1662 images of CSOMa and 751 images of CSOMi. **Figure 1** shows examples of the TMs in each category together with a clear definition.

Labelling of images

All CSOM images were labelled according to surgical findings of TM perforation and postoperative pathological reports if available. CSOMa was further confirmed if the patient reported increased otorrhoea at the time the images were taken. OME was labelled if tympanocentesis found fluid in the tympanic cavity.

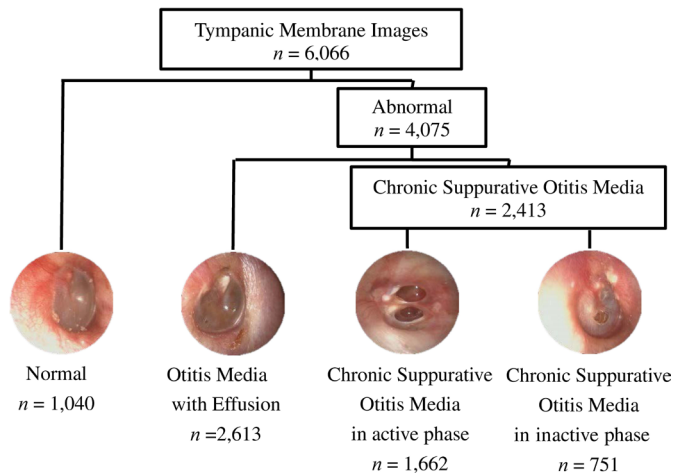


Figure 1 Classification tree for the four diagnostic classes. A total of 6066 images are included and categorised to four conditions: 1040 images of normal tympanic membrane, 2613 images of otitis media effusion, 1662 images of chronic suppurative otitis media in active phase and 751 images of chronic suppurative otitis media in inactive phase.

Two-stage model for classification

One commonly used strategy for classifying TM images is to first pretrain a CNN model on a large-scale natural image dataset, such as the most popular ImageNet,²⁴ which includes over 1 million images, and then fine tune the model on the target training dataset. However, a vital disadvantage with such a strategy is that the images in the target dataset have to be downsampled to a much lower resolution in order to fit the mandatory input size of the CNN model, for instance, a 224×224 pixel resolution is required for ResNet. The downsampling can cause severe information loss and thereby performance degradation.

In order to boost performance and make full use of all discriminative parts in the input image, a two-stage pipeline for classification was used in the present study, that resembles the attention mechanism of a human being. The pipeline involved two separated CNN models, called the main model and the focal model, providing a predicted result based on the whole image and important parts of the image.

The main model acted as a main classifier of images as well as a filter to remove irrelevant parts of an image for the focal model. By fine tuning a network pretrained on ImageNet to the task, we produced the main classifier for decision-making based on the whole image. It was then feasible to find important parts in the input image using manual annotation such as boundary boxes to train a detection network.

We were able to extract information from the trained classification network to provide weak clues as to the location of important parts automatically. The CAM¹⁵ is a simple but effective method to visualise parts of interest from a network by projecting back the weights of the output layer onto convolutional feature maps obtained from the last convolutional layer. By calculating CAMs from input images, the location of important parts could

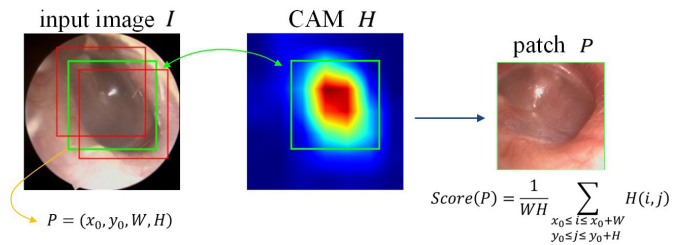


Figure 2 Every patch in the same image. The sliding-window strategy for patch selection is to take a window to scan throughout the whole image by a fixed step size, (eg, 16 pixels in both row and column directions). At each location, a patch can be cropped from the image and a score assigned to this patch according to the binarised heat map. The score of this patch is calculated by averaging the intensity values of pixels within the corresponding window in the heat map. Several different window sizes are predefined. Then, the image patches cropped at various sizes and locations (eg, the red and green boxes are ordered according to their scores, and the top ones with high scores are selected, eg, the green box). CAM, Class Activation Map.

be highlighted. After upsampling the CAMs to the same size as the original image, a discriminative patch with a higher resolution could be cropped from the original image. This contained more information and improved the classification greatly. In summary, for each input image, the main model outputs a series of scores indicating the classification result while at the same time providing a heat map from CAMs to help locate discriminative patches.

The secondary model acts as the focal classifier of images by focusing on the discriminative patches. From a single image, it was possible to extract many patches simply by random cropping, a widely used scheme for data augmentation in CNN models.^{25–26} However, patches selected under guidance have a higher confidence for relevance and discrimination, leading to a better performance. This model was fine tuned to a series of patches selected by a sliding window of different size adopted from the CAM. As shown in figure 2, the sliding-window strategy for patch selection is to use a window to scan through the whole image by a fixed step size, for example, 16 pixels in both row and column directions. At each location, a patch can be cropped from the image and a score assigned to this patch according to the binarised heat map. The score of this patch was calculated by averaging the intensity values of pixels within the corresponding window in the heat map. Several different window sizes were predefined. Then, the image patches cropped at various sizes and locations were ordered according to their scores with the top ones with high scores selected. To avoid the window sliding out of the image, locations very close to the image boundaries were not considered. In the test phase, the maximum point of the heat map indicated the most possible location of an object and a patch was selected around it.

As discussed above, the two models have different functionality and strength due to difference in scale they

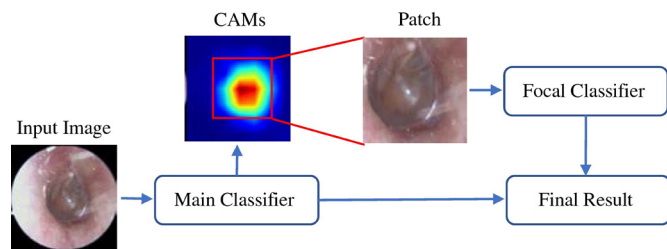


Figure 3 The complete classification pipeline of the method. The main classifier provides a global result and the focal model works on patches containing discriminative and local features. The prediction results of the two models are merged by averaging the classification output scores obtained from these two models. CAMs, Class Activation Maps.

handled. The main model provides a global result, while the focal model works on patches containing discriminative and local features. We merged the prediction results of the two models to use the advantage of each by averaging the classification output scores. The complete classification pipeline of the method is shown on [figure 3](#).

Experiment settings and procedure

All experiments were performed using the Intel Xeon E5-2620 CPU and NVIDIA TITAN Xp GPU. Python 3.6 in Keras was used as the programming language for developing the deep learning framework based on TensorFlow.

To evaluate the method, two common backbone networks, that is, ResNet²⁷ and Inception-V3,²⁶ were used in the experiments. More specifically, both networks were trained using the Adam optimiser, which runs for 30 epochs. The learning rate is initially set to be 0.0001 and decays with a factor of 0.1 for every 10 epochs.

During the training phase, some methods of online data augmentation were adopted including; random shifting, shearing, zooming and flipping. The input size of ResNet is 224×224 while Inception-V3 uses 299×299. The input image was resized to fit the input size of networks using bilinear interpolation. In addition, all the experiments were conducted using patient-level threefold cross-validation. Three images were chosen at the same time from each patient. When splitting the dataset for cross-validation based on patients rather than images, no samples from the same patient appeared in both training and testing sets. All the images chosen for further analysis were conducted in the same way. Overall accuracy was calculated as the number of correctly classified images divided by the total number of considered images. All results were the average performance of three folds. The F1 Score was used to evaluate the performance in each type of TM image with the advantage of considering both precision and recall.

The training process included three steps as follows:

First, a CNN model pretrained on ImageNet was fine tuned with the training dataset of TM images to obtain the main model. Second, a couple of local patches were located in each image by using CAM¹⁵ derived from the main model. The aggregation of all patches acquired

over the whole dataset was then taken to train another focal model, which had the same network structure as the main model. Since CAMs are class specific, and the true class labels of images in the testing phase were unavailable, we average all the CAMs of various classes to get a general attention map.

It is noteworthy that there were differences in the way we selected patches between the training and testing phases. The original heat map was continuously valued so we turned it into a binary image, that is, pixels taking the value 0 or 1 by picking up a value as threshold, and turning all pixels over this value to be 1, while those under this value to be 0. During the training phase, the general attention maps were binarised with a threshold of 0.5. A series of sliding windows were adopted selecting two patches for the size of 300, 400 and 500 pixels, in order to keep all useful information. However, in the testing phase, considering the heavy computation cost, we only selected a 400×400 patch with the centre located on the maximum point of the general attention map.

Third, a pretrained network used as the focal model was fine tuned on the selected patches, helping to achieve a better performance. The patches for training the focal model were selected from each image in the corresponding training set.

Patient and public involvement

Due to the retrospective nature of this study, patients and the public were not involved in the study design and research analysis.

RESULTS

The classification results are shown in [table 1](#). Although the accuracy of the focal classifier was lower in both backbones, the focal classifier can help the main classifier achieve better performance. The consistent performance achieved by using two different backbones indicates that our method is robust and insensitive to the choice of the backbone network.

To provide a comparison with human experts, five doctors with a variety of experience were invited to label a subset of the test dataset. A total of 270 images from 90 subjects were randomly selected for each type of image, in total 1080 images were used in this evaluation. Two associate chief doctors achieved an accuracy of 91.02% and 87.50%, while two attending doctors achieved an accuracy of 86.57% and 79.44%, respectively. A primary doctor achieved an accuracy of 79.07%. [Figure 4](#) displays the confusion matrices of our method using ResNet50 and three doctors with differing accuracies. Confusion matrices were consistent with clinical experience, indicating that it is more difficult to distinguish between cases of: normal versus OME and CSOMa versus CSOMi than normal versus pathological images, or OME versus CSOM. Using our method and experts, it was difficult to distinguish the slight differences between normal images and OME or CSOMa and CSOMi. The results obtained from

Table 1 Comparison results of various methods on our dataset

Method	Backbone	F1 Score				Overall accuracy
		Normal	OME	CSOMa	CSOMi	
Main classifier	Inception-V3	0.9178±0.0224	0.9613±0.0032	0.9028±0.0085	0.8103±0.0062	0.9219±0.0068
Focal classifier		0.9294±0.0085	0.9589±0.0031	0.8793±0.0113	0.7583±0.0444	0.9078±0.0117
Our pipeline		0.9485±0.0065	0.9470±0.0017	0.9099±0.0104	0.8180±0.0295	0.9330±0.0081
Main classifier	ResNet50	0.9033±0.0189	0.9500±0.0141	0.9133±0.0047	0.8133±0.0170	0.9162±0.0056
Focal classifier		0.9333±0.0047	0.9633±0.0094	0.8900±0.0082	0.7500±0.0283	0.9126±0.0045
Our pipeline		0.9433±0.0125	0.9684±0.0132	0.9167±0.0047	0.8237±0.0171	0.9337±0.0051

The best results in each set of experiment are in bold. All results are reported as an average with an SD of the results of three folds. We used the F1 Score to measure the performance in each type of image in order to consider both precision and recall. The overall accuracy is calculated as the ratio of the number of correct classified images and the number of total images in test set.

CSOMa, chronic suppurative otitis media in active phase; CSOMi, chronic suppurative otitis media in inactive phase; OME, otitis media with effusion.

each stage of the experimental method were combined with the three folds, that is, all the datasets were considered in calculating the confusion matrices.

The performance of the two challenging binary classification problems was further evaluated, including normal versus OME and CSOMa versus CSOMi. The receiver operating characteristic (ROC) curve is an efficient tool to evaluate the comprehensive quality of classification models in all different situations. Figure 5 shows the average ROC curves of our method with ResNet50 as backbone. True positive and false positive rates were calculated for each doctor and marked on the figure. As figure 5A shows, in assessing the images of the normal TM and OME, the performance of an inexperienced primary doctor was significantly different from the judgements reported by relatively experienced attending doctors and associate chief doctors. Moreover, in regards to assessing

the more challenging task of distinguishing the two stages of CSOM, the performance of only one associate chief doctor was better than the methods proposed in this paper.

Figure 6 shows the intermediate results of typical samples using the methods proposed in this study. The green box indicates the patch used in the test phase, and the translucent mask shows the areas used to select patches during the training phase. The core areas of TM were successfully detected with the weakly supervised approach.

DISCUSSION

In this study, we propose a two-stage CNN method for otoscopic image classification. To the best of our knowledge, this is the first time that an attention-based model

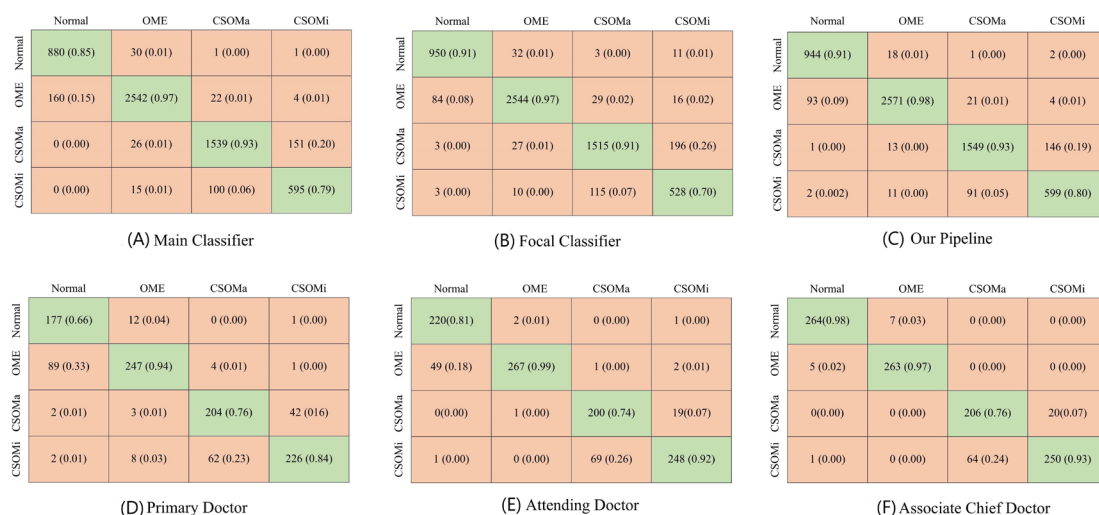


Figure 4 Confusion matrices for each stage in our method with ResNet50 and three human experts. The row axis indicates the prediction while the column axis represents for the ground truth. Results among test sets of three folds are combined to report a performance of the whole dataset. (A, B) show result of the two major classifiers of our method, while (C) reports the result of average assembling. In addition, the overall accuracies of these three experts are 79.07%, 86.57% and 91.02% (D, E, F). CSOMa, chronic suppurative otitis media in active phase; CSOMi, chronic suppurative otitis media in inactive phase; OME, otitis media with effusion.

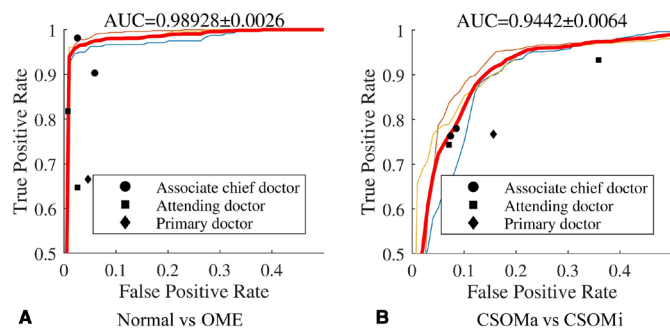


Figure 5 Receiver operating characteristic curve for classification of two challenging situations, comparing the method in ResNet50 with human experts. The red curve is the average of three folds' performance and the other curves show the result for each fold. Our method can achieve a performance similar with the associate chief doctor. AUC, area under the curve; CSOMa, chronic suppurative otitis media in active phase; CSOMi, chronic suppurative otitis media in inactive phase; OME, otitis media with effusion.

combining global and local information has been used in the field of otoscopic image analysis. This deep learning model adopts attention mechanisms to focus on salient lesion areas in the TM with correcting image alignment to reducing the impact of noise. The results show the classification performance to reach the diagnostic level of an associate professor in otolaryngology in identifying normal, OME, CSOMa, CSOMi from otoscope images.

Previous studies have built up machine-learning models for ear disease diagnosis and achieved high diagnostic accuracy.^{1 7 8 14} A recent study has compared nine models of transfer learning and assembled two of them (Inception-V3 and ResNet101) to build a deep learning model to automatically diagnose ear disease.⁷ They achieved an accuracy of 93.67% using a large database of 10544 images. However, only 86% accuracy was achieved when the number of images for training was reduced to 5000. In this study, an equivalently high accuracy of 93.37% was

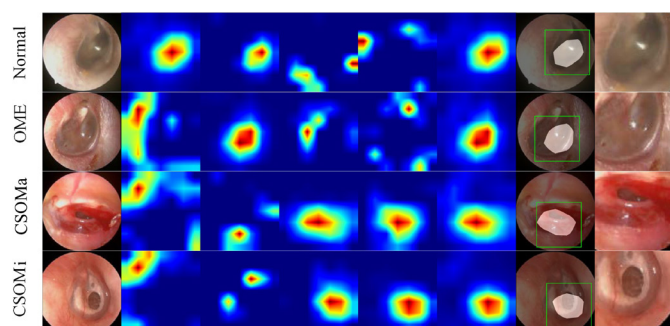


Figure 6 Typical samples of each situation, including normal, OME, CSOMa and CSOMi. From left to right, each column shows original image, the CAM of normal, the CAM of OME, the CAM of CSOMa, the CAM of CSOMi, the averages of CAM, the selected box and the patch for the focal classifier. CAM, Class Activation Map; CSOMa, chronic suppurative otitis media in active phase; CSOMi, chronic suppurative otitis media in inactive phase; OME, otitis media with effusion.

achieved using only 6066 images. Two models were assembled in different ways to that reported in a previous study.⁷ First, a main classifier was tuned on the entire TM images acquired by otoscope. Second, we calculated CAMs from the trained main classifier to locate TM details which could be used to improve performance. Finally, another pretrained network was tuned on the selected patches as the focal classifier, enabling the main classifier to perform better. As [table 1](#) shows, this assembled system improved the accuracy over the single model method.

The achievement of high accuracy with a relatively small database may be attributed to the combined use of main classifier and focal classifier with CAM. CAM highlights the important area in a trained network,¹⁵ which relates to attention mechanisms. The attention mechanism of the model was consistent with the experienced otolaryngologist concentrating on the local lesion-related areas in otoscopic examination. Based on our finding, the significant spectrum of the images for CSOM by CAM was to focus on the perforation areas of the TM; while the area of a shortened or vanished cone of light as well as colour changes in the eardrum detected by CAM presented the significant spectrum in OME images. Therefore, CAM was considered as a guide to select discriminative parts and fine tune another network to focus on those parts providing higher resolution and more information. Consequently, the focal classifier targets more results from partial images. When considering information from both whole image and partial image, the performance of two kinds of backbone networks could be enhanced by 1%~2% in overall accuracy. The further improvement in accuracy of automated detection of pathological changes in the otoscopic images is vital to facilitate the capability in clinical diagnosis of OME for the paediatricians, physicians and junior otolaryngologists.

It is of note that there are a few limitations in the present study. First, non-medical history and hearing information were provided to the deep learning model and the otolaryngologists, which may compromise the diagnosis accuracy. For example, CSOM is often accompanied by symptoms of recurrent otorrhoea and hearing loss,²⁸ and doctors can greatly improve their accuracy of diagnosis by asking for a history. In addition, all the otoscopic images used in the experiment were taken after cleaning the external auditory canal. If the model is applied at home or community hospital, it may be affected by cerumen. In further studies, more ear disease images should be collected in order to train a more robust and practical network. In addition, the current deep learning model can be improved by training with non-image information such as hearing audiometry, tinnitus, ear fullness, duration of history and presence of fever for better diagnosis accuracy.

CONCLUSION

In this study, the assembled classifier accompanying a main classifier and a focal classifier with CAMs achieves a

high accuracy in diagnosis of OM with endoscopic images based on a relatively small database. This deep learning model is useful in helping junior otolaryngologists and non-otolaryngologists to diagnose ear disease early. Further study will consider more ear diseases together with patients' information such as medical history, hearing thresholds obtained from pure tone audiometry and middle ear function assessed by using tympanometry to improve diagnostic accuracy.

Author affiliations

¹Department of Otolaryngology, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, Guangzhou, Guangdong Province, China

²Institute of Hearing and Speech-Language Science, Sun Yat-Sen University, Guangzhou, Guangdong Province, China

³Department of Automation Science and Engineering, South China University of Technology School, Guangzhou, Guangdong, China

⁴Centre for Speech and Language Therapy and Hearing Science, Cardiff School of Sport and Health Sciences, Cardiff Metropolitan University, Cardiff, UK

Acknowledgements The authors would like to acknowledge Dr Christopher Wigham for the proofreading.

Contributors YXC designed the study, YBC and CL collected and analysed the data. J-GY and YL designed deep transfer learning system and performed training with LX, LL, SZ, JZ, MW and YS performed the otoendoscopy examination. YZ, YXC, EMG and FZ drafted the manuscript, while all authors reviewed and revised the manuscript.

Funding This work was supported by the Key R&D Programme of Guangdong Province, China (Grant No. 2018B030339001), medical artificial intelligence project of Sun Yat-Sen Memorial Hospital (XYGZN201904) and the National Natural Science Foundation of China (Grant No. 81570935).

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval This study was approved by the Institutional Review Board of the Sun Yat-sen Memorial Hospital, China. Because this study was a retrospective research, the ethics committee approved to waive the need for informed consent.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. No additional data are available. However, the original data that support the findings derived from this study can be requested by emailing caiyx25@mail.sysu.edu.cn.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD

Yiqing Zheng <http://orcid.org/0000-0002-1600-8539>

REFERENCES

- Lee JY, Choi S-H, Chung JW. Automated classification of the tympanic membrane using a Convolutional neural network. *Appl Sci* 2019;9:1827.
- Schilder AGM, Chonmaitree T, Cripps AW, et al. Otitis media. *Nat Rev Dis Primers* 2016;2:16063.
- 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.
- Guan Q, Huang Y, Zhong Z. Diagnose like a radiologist: attention guided Convolutional neural network for thorax disease classification. *arXiv* 2018.
- Liskowski P, Krawiec K. Segmenting retinal blood vessels with deep neural networks. *IEEE Trans Med Imaging* 2016;35:2369–80.
- Sorrento A, Pichichero ME. Assessing diagnostic accuracy and Tympanocentesis skills by nurse practitioners in management of otitis media. *J Am Acad Nurse Pract* 2001;13:524–9.
- 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.
- Wang Y, Li Y. *Deep learning in automated region proposal and diagnosis of chronic otitis media based on computed tomography*. United States: Copyright Wolters Kluwer Health, Inc, 2020: 669–77.
- 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;316:2402–10.
- Voets M, Møllersen K, Bongo LA. *Reproduction study using public data of: development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus Photographs*. United States: Public Library of Science, 2019: e217541.
- Esteva A, Kuprel B, Novoa RA, et al. Corrigendum: Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;546:686.
- Brinker TJ, Hekler A, Enk AH, et al. Deep neural networks are superior to dermatologists in melanoma image classification. *Eur J Cancer* 2019;119:11–17.
- Long E, Lin H, Liu Z. *An artificial intelligence platform for the multihospital collaborative management of congenital cataracts*. 1. London: Nature Publishing Group, 2017.
- Myburgh HC, Jose S, Swanepoel DW, et al. *Towards low cost automated smartphone- and cloud-based otitis media diagnosis*. Elsevier Ltd, 2018: 34–52.
- Min JK, Kwak MS, Cha JM. Overview of deep learning in gastrointestinal endoscopy. *Gut Liver* 2019;13:388–93.
- Rensink RA. The dynamic representation of scenes. *Vis cogn* 2000;7:17–42.
- Fang L, Wang C, Li S, et al. *Attention to lesion: Lesion-Aware Convolutional neural network for retinal optical coherence tomography image classification*. United States: IEEE, 2019: 1959–70.
- Desimone R, Duncan J. Neural mechanisms of selective visual attention. *Annu Rev Neurosci* 1995;18:193–222.
- Kushibar K, Valverde S, González-Villà S, et al. Automated sub-cortical brain structure segmentation combining spatial and deep convolutional features. *Med Image Anal* 2018;48:177–86.
- Lisowska A, Neil A O, Dilys VValdés Hernández M, González-Castro V, eds. *Context-Aware Convolutional neural networks for stroke sign detection in Non-contrast CT scans*. Cham: Springer International Publishing, 2017: 494–505.
- Bejnordi BE, Zuidhof G, Balkenhol M, et al. Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images. *J Med Imaging* 2017;4:44504:1.
- Chen H, Qi X, Yu L, et al. DCAN: deep contour-aware networks for object instance segmentation from histology images. *Med Image Anal* 2017;36:135–46.
- Ghafoorian M, Karssemeijer N, Heskes T, et al. Deep multi-scale location-aware 3D convolutional neural networks for automated detection of lacunes of presumed vascular origin. *Neuroimage Clin* 2017;14:391–9.
- Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database. *IEEE* 2009:248–55.
- He K, Zhang X, Ren S. Deep residual learning for image recognition. *Computer vision and pattern recognition* 2016:770–8.
- Szegedy C, Vanhoucke V, Ioffe S. Rethinking the inception architecture for computer vision. *Computer vision and pattern recognition* 2016:2818–26.
- Barry KM, Paolini AG, Robertson D. *Modulation of medial geniculate nucleus neuronal activity by electrical stimulation of the nucleus accumbens*. United States: Elsevier Ltd, 2015: 1–10.
- Acuin J. Chronic suppurative otitis media. *BMJ Clin Evid* 2007;2007. [Epub ahead of print: 01 Feb 2007]. 2007.