EBioMedicine 45 (2019) 606–614

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

EBioMedicine
Published by THE LANCET

Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database

Dongchul Cha [a,1], Chongwon Pae [b,c,d,1], Si-Baek Seong [b,c], Jae Young Choi [a,c,**], Hae-Jeong Park [b,c,d,*]

[a] Department of Otorhinolaryngology, Yonsei University College of Medicine, Republic of Korea
[b] Center for Systems and Translational Brain Sciences, Institute of Human Complexity and Systems Science, Yonsei University, Republic of Korea
[c] BK21 PLUS Project for Medical Science, Yonsei University College of Medicine, Republic of Korea
[d] Department of Nuclear Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 21 May 2019
Received in revised form 19 June 2019
Accepted 25 June 2019
Available online 1 July 2019

Keywords:
Convolutional neural network
Deep learning
Otoendoscopy
Tympanic membrane
Ear disease
Ensemble learning

ABSTRACT

Background: Ear and mastoid disease can easily be treated by early detection and appropriate medical care. However, short of specialists and relatively low diagnostic accuracy calls for a new way of diagnostic strategy, in which deep learning may play a significant role. The current study presents a machine learning model to automatically diagnose ear disease using a large database of otoendoscopic images acquired in the clinical environment.
Methods: Total 10,544 otoendoscopic images were used to train nine public convolution-based deep neural networks to classify eardrum and external auditory canal features into six categories of ear diseases, covering most ear diseases (Normal, Attic retraction, Tympanic perforation, Otitis externa±myringitis, Tumor). After evaluating several optimization schemes, two best-performing models were selected to compose an ensemble classifier, by combining classification scores of each classifier.
Findings: According to accuracy and training time, transfer learning models based on Inception-V3 and ResNet101 were chosen and the ensemble classifier using the two models yielded a significant improvement over each model, the accuracy of which is in average 93·67% for the 5-folds cross-validation. Considering substantial data-size dependency of classifier performance in the transfer learning, evaluated in this study, the high accuracy in the current model is attributable to the large database.
Interpretation: The current study is unprecedented in terms of both disease diversity and diagnostic accuracy, which is compatible or even better than an average otolaryngologist. The classifier was trained with data in a various acquisition condition, which is suitable for the practical environment. This study shows the usefulness of utilizing a deep learning model in the early detection and treatment of ear disease in the clinical situation.
Fund: This research was supported by Brain Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT(NRF-2017M3C7A1049051).

1. Introduction

Ear and mastoid disease (International Statistical Classification of Diseases and Related Health Problems (ICD) code H.60-H.95) is a common disease that can easily be treated with early medical care. Nevertheless, if one does not receive timely detection and appropriate treatment, it may leave sequelae, such as hearing impairment. In the evaluation of ear and mastoid disease in the clinic, physical examination using conventional otoscopy or otoendoscopy as well as history taking is the first step. However, diagnosis by non-otolaryngologists using otoscopy or otoendoscopy is highly susceptible to misdiagnosis [1]. In a study by Pichichero, Poole [2], the correct diagnosis rate of otitis media diagnosed by 514 pediatricians using pneumatic otoscope was an average rate of 50%. The study also shows a higher (compared to pediatricians) but not a satisfactory accuracy of 73% when diagnosed by 188 otolaryngologists. This low diagnostic accuracy implies that diagnosis of ear disease without the help of additional resources such as imaging or acoustic testing is difficult even for specialists. The short of specialists in the local clinic and their relatively low diagnostic accuracy calls for a new way of diagnostic strategy, in which machine learning may play a significant role.

* Correspondence to: H. J. Park, Department of Nuclear Medicine, Yonsei University College of Medicine, 50-1, Yonsei-ro, Sinchon-dong, Seodaemun-gu, Seoul 03722, Republic of Korea.
** Correspondence to: J. Y. Choi, Department of Otorhinolaryngology, Yonsei University College of Medicine, 50-1, Yonsei-ro, Sinchon-dong, Seodaemun-gu, Seoul 03722, Republic of Korea.
E-mail addresses: jychoi@yuhs.ac (J.Y. Choi), parkhj@yonsei.ac.kr (H.-J. Park).
[1] Equally contributed first authors.

https://doi.org/10.1016/j.ebiom.2019.06.050

Research in context

Evidence before this study

Ear and mastoid disease is a common disease, which demands early and appropriate diagnosis with otoscopy or otoendoscopy, but is not trivial in local clinics and the diagnosis rate even by otolaryngologists using ear images show an unsatisfactory accuracy, as low as 73%. So far, the best known study for automatic diagnosis of ear disease using images has been done with tympanic membrane using a shallow neural network of relatively small data size (n ~ 390) with an accuracy of 86.84%, however, the previous method is only capable of partially diagnosing middle ear disease.

Added value of this study

This is the first study to utilize a deep learning scheme to classify tympanic membrane otoendoscopic images into six diagnostic categories, especially including attic retractions and tumors, using a large database (n = 10,544), and the deep learning model covers most of the ear diseases in the clinic, not only on the middle ear but also on the external ear. It also deals with an unstandardized clinical image set as-is without image quality control, which makes the current system adaptable to the real-world clinical setting. The ensemble classifier, which we propose, shows better performance than using a single transferred deep learning model with an accuracy of 93·67%.

Implications of all the available evidence

According to our evaluation on the relationship between database size and the performance of the transfer deep learning models, current study suggests the need for a sufficient size of the database for a reliable classification performance in the medical image domain.
Due to the high accuracy and the diagnostic coverage in the proposed model, clinicians with less experience in otoendoscopy, or other specialty physicians such as pediatricians, emergency, or family medicine doctors could be benefitted from the model and thus it may result in alleviating the burden of the growing number of patients with hearing impairment.

As far as we know, relatively few machine learning studies have been conducted for automated diagnosis of ear disease using otoscopic images. Myburgh and colleagues reported auto-diagnosis of otitis media, with an accuracy of 81·58% by decision tree and 86·84% by neural network method [3], which conducted a classification of tympanic membrane into five groups between normal eardrum, otitis media with perforation, acute otitis media, otitis media with effusion and cerumen impaction. However, the classification categories lack important and critical diagnosis such as attic retraction.

For clinical use, the current study is conducted to provide a reliable diagnosis of otitis media, attic retraction, atelectasis, tumors, and otitis externa, using deep learning for otoscopy photos of the eardrum and external auditory canal (EAC). These categories cover most of the domain of ear diseases that could be diagnosed using otoendoscopy in the clinics. For this, we proposed an ensemble classifier of two best-performing deep neural networks evaluated for ear images.

Deep learning or deep neural network has been introduced to various fields of medicine successfully. For example, in the field of ophthalmology, the machine learning result is comparable to a level of specialist [4–6]. Most of these studies utilize convolutional neural network (CNN), a supervised deep learning method. However, building CNN from scratch requires a large amount of dataset and computational power, which is not practical in many application areas. Instead, public CNN models pretrained for natural images could be reused and fine-tuned to a specific application, which is called transfer learning. In transfer learning, most network layers in a public network model are transferred to a new model, followed by a new fully-connected layer that classifies those features into a new set of classes. Studies with transfer learning for medical imaging showed high classification accuracy comparable to, or even better to building CNN from scratch [7,8].

This study is composed of the following three main parts. First, we evaluated the performance of nine public models to choose the best models in terms of accuracy and training time for the current application. Based on this evaluation, ensemble classifier to combine multiple models' classification results was proposed, which is expected to increase the overall classification performance than using a single classifier. Second, although transfer learning is known to be efficient in a relatively small dataset (as in labelled medical images), the dependency of the classification accuracy and model type on the size of the dataset is not exampled yet. Thus, we tested the performance of the classifier depending on the data size. We also conducted optimization of the model configuration, by assigning a hidden layer in the fully connected network layer, and changing colour channels in the image database. Finally, we showed and discussed the characteristics of the proposed model for diagnosing ear diseases in the clinical setting.

2. Materials and methods

2.1. Patient selection and data acquisition

Data from patients who visited the outpatient clinic in Severance Hospital otorhinolaryngology department from the year 2013 to 2017 were used. As a routine, patients had their otoendoscopic photo taken upon visit. Drum photos were taken with either 4 mm or 2.7 mm OTOLUX 0-degree telescope (MGB Endoskopische Geräte GmbH Berlin, Germany) tethered to Olympus OTV-SP1 video imaging system (Olympus Corporation, Japan), by otolaryngology residents, faculty or experienced nurses. The image resolution was 640 by 480 pixels. A total of 19,496 endoscope photos were reviewed for labelling. Since otoendoscopic findings of post-surgery status are mostly subjective and rely on the surgeon, 7602 photos were excluded. Additionally, 1350 photos were excluded since the photos were not appropriate for examination, for example, sites not related to eardrum or EAC, duplicates, the picture was significantly blurred due to handshakes or focus problems, or the author could not agree despite attending physician's medical records, acoustic and radiologic test results. Since photos were taken by several clinicians, and the external auditory canal is subject to individual variation, the composition of photography was not standardized; colour arrangements, white balance, eardrum size, location, rotation, angle, and light reflection in images were variable, but the photo was analyzed as-is to reflect real-life clinical setting. In addition, partially visible eardrums due to the image's field of view not containing the whole eardrum were included in the analysis. Finally, a total of 10,544 otoendoscopic images of eardrum and EAC from patients were analyzed. This retrospective study was approved by the Severance Hospital Institutional Review Boards.

2.2. Labelling of images

Photos of eardrums and its surrounding EAC were taken with otoendoscope and were labelled into six categories. The classification was done according to *Colour Atlas of Endo-Otoscopy* [9]. A normal eardrum and EAC included: 1) completely normal eardrum, 2) normal but showing healed perforation, 3) normal with some tympanosclerosis. Abnormal findings included: 1) tumorous condition which includes middle ear tumors, EAC tumors, and cerumen

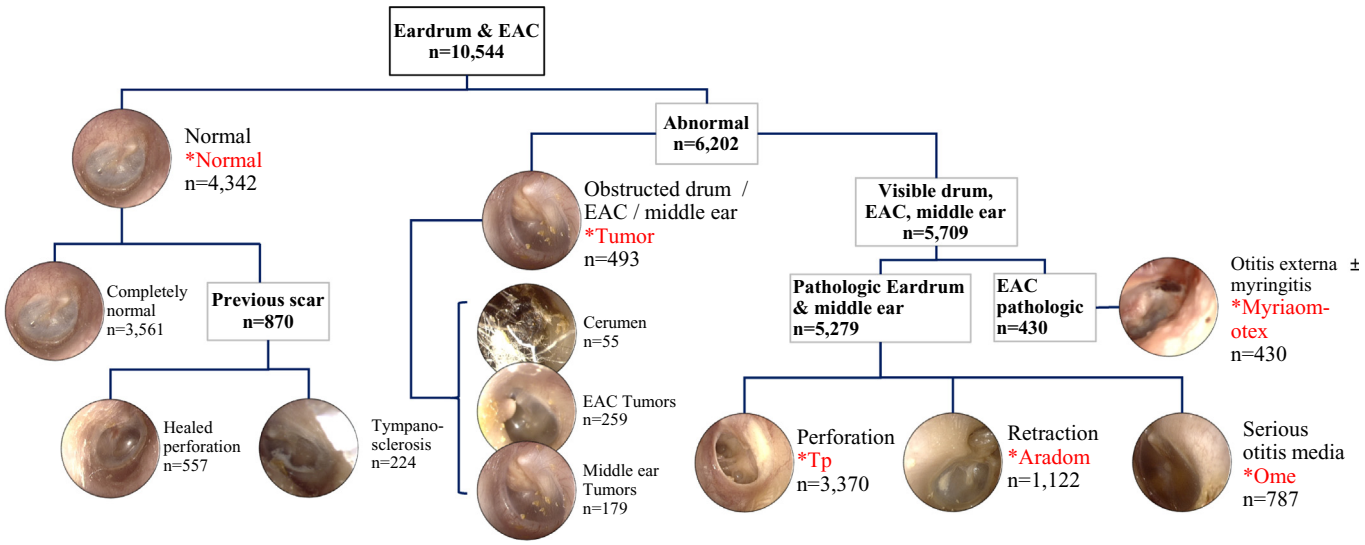

Fig. 1. Decision tree for labelling of otoendoscopy image and six diagnostic classes. Classes that were used for training are marked with an asterisk. EAC: external auditory canal.

impaction, 2) otitis media with effusion, 3) eardrum erosions, otitis externa, 4) perforation of the eardrum, 5) attic retraction/atelectasis (Fig. 1). Some classes have relatively small numbers of samples for training. In order to balance the sample size for each class, we merged several sub-classes into a class according to their similarity in diagnosis and treatment. Three normal diagnoses are trained as one big "Normal" class. "Tumor" class include cerumen impaction, EAC tumors, and middle ear tumors since they share a common property that the eardrum

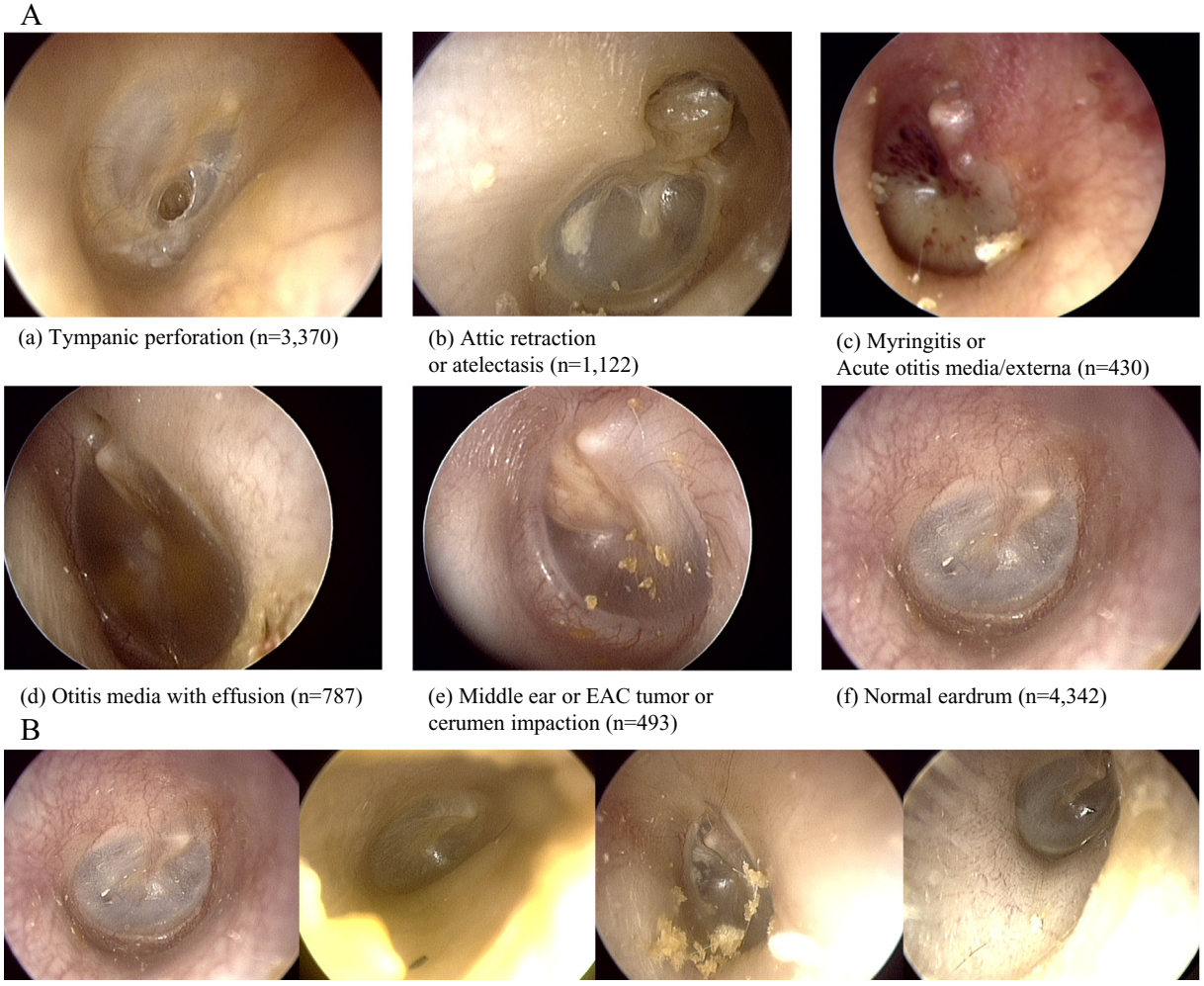

Fig. 2. A) Examples of six classifications of ear disease, sorted by labelling priority (total $n = 10{,}544$). B) Example of image diversity labelled "Normal".

is not well-visible, and since they often require surgical procedures. Attic retraction (or destruction) and eardrum atelectasis has been merged into "Aradom" class, since it shares common pathogenesis and physical findings, and often requires surgical intervention. Otitis externa and myringitis have been merged into "Myriaom-otex" class since otorrhea is the main symptom, the physical finding is similar, and first-line treatment is antibiotics.

The number of images used and samples representing each classification is shown in Fig. 2A. If there are more than one features of the ear disease, for example, tympanic perforation with attic retraction, it was labelled as tympanic perforation, according to our labelling priority. The priority is based on the certainty of the diagnosis and clinical importance, for example, requiring surgical intervention. Of note, image acquisition was not standardized in any fashion and was labelled and trained as-is. Examples of diversity in image acquisition include differences in white balance, image composition, presence of cerumen, position of the eardrum in the image. The exemplary images for diversity in the normal class are presented in Fig. 2B.

An in-house graphic user interface software implemented on MATLAB2019a® (MathWorks, Inc., Natick, Massachusetts, United States) was used for manual labelling. As mentioned above, study by Pichichero, Poole [2], confirms the limited accuracy by a single physician is below 75% at best, numerous methods were used for labelling the ground truth of otoendoscopic image. The images of eardrum and EAC were labelled by the first author, and all the images were double checked by reviewing electronic medical record written by attending physician at the time, who had at least 10 years of experience in a tertiary referral center. Since this study is retrospective, additional clinical demographic and symptomatic data was used. In addition, acoustic test results (pure tone audiometry, impedance audiometry) were often available along with computed tomography, magnetic resonance imaging results were used for more accurate labelling. For example, if pure tone audiometry and/or impedance audiometry data was available, it was used for labelling otitis media with effusion or tympanic perforation. Temporal bone computed tomography or magnetic resonance imaging results were also used if available. If the classification of otoendoscopic image could not come to an agreement even after reviewing all of the available information, it was discarded.

2.3. Training transfer learning network models

Public deep learning models pretrained with ImageNet database (http://www.image-net.org), capable of classifying 1000 natural objects, were used for training the model for otoendoscopic images. Among many deep learning models publicly available, Alexnet [10], GoogLeNet [11], ResNet [12] (ResNet18, ResNet50, ResNet101), Inception-V3 [13], Inception-ResNet-V2 [14], SqueezeNet [15], and MobileNet-V2 [16] were used and compared since these network models are known to show higher performance in the accuracy when compared with any other networks with similar prediction time. Smaller size networks models were also included to see the performance for online processing. When transferring layers in public models to new models, we replaced the last fully connected layer of each model with a new fully-connected layer with six output nodes, followed by a softmax activation function. The training was conducted using an Adaptive moment estimation (ADAM) [17] with a batch size of 50, the maximum epoch of 20 and an initial learning rate of 0·0001. The initial learning rate of 0·0001, which may seem low, was selected according to our experience that using conventional learning rate of 0·01 to 0·001 did not converge in the current application. For the fully connected layer, we assigned weight and bias learning rate factors of 10 to render faster learning in the new layer than in the transferred layers. This study was conducted using Deep Learning Toolbox in MATLAB 2019a over four graphics processing unit (GPU) in the DGX station (NVIDIA, inc., USA). To augment data, we conducted random X and Y translation of input images from −45 to 45 pixels, random rotations from −30 to 30 degrees, random scales between 0.8 and 1.2 and random left/right flips to render translation, rotation, scale and left/right invariance.

1. ***Selection of the best two models:*** the best two among nine models were selected by evaluating the performance of each model in terms of accuracy and calculation time. From a total of 10,544 otoendoscopic images, 80% of the images were used for training; 20% were left out for validation of the model. This training-validation step was done twice with different sets of training and validation data. According to the mean accuracy and calculation time, we chose two models.
2. ***Performance according to data size, and hidden layer in the fully connected layer and colour channels:*** We evaluated the performance for all nine models for the half and a quarter of the full dataset, to compare those with the performance trained with all the data. We also evaluated the performance for a model with an additional hidden layer (node size =25) between the input layer and the output layer in the fully connected neural network. We also evaluated the performance for changing colour (RGB) orders by changing R and G channels in the image data since public network models were trained with natural images different from the current ear images.
3. ***Ensemble classifier:*** We generated an ensemble classifier that combines classifiers' outputs from the best two models. Each classifier scores the probability of an input image to be one of six classes and the maximal score among all classes is chosen as a predicted label. The ensemble classifier adds the two scores from the two models for an input image and the class having a maximal score is chosen to be the image's label.
4. ***Cross-validation:*** For the best two models and the ensemble classifier, we conducted five-fold cross-validation for each classifier and evaluated the classification performance in terms of accuracy.

3. Results

Accuracy, training time (GPU time), the number of parameters of each transferred model is presented in Table 1. The number of parameters was referred from the MATLAB official web site (https://www.mathworks.com/help/deeplearning/ug/pretrained-convolutional-neural-networks.html). There was no significant improvement in the models with a hidden layer (number of nodes = 25, H25 in Table 1). The average accuracy between different sets of data size showed significant improvement according to data size - 78·88% for data set of n = 2000, 85·62% for data set of n = 5000, and 90·21% for the entire data set of n = 10,544 (Fig. 3). The performance of nine models was evaluated without a hidden layer in the fully connected neural network. The best accuracy was yielded by the Inception-ResNet-V2 (92·1%), Inception-V3 (92%) and ResNet101 (91·55%) in order. Despite its accuracy, the Inception-ResNet-V2 (33,283 s) had three times longer training time than those of Inception-V3 (11,938 s) or ResNet101 (12,215 s). Therefore, we finally chose Inception-V3 and ResNet101 as best transferred network models for the subsequent analysis.

From these two models, we generated an ensemble classifier, which decides the image label according to the sum of the two models' scores for the given image (Fig. 4). Fig. 5 shows examples of improvement using the ensemble classifier by evaluating the sum of classification scores of the two network models. Repeated measures one-way ANOVA for the 5-fold cross-validation tests (Fig. 6) showed that the ensemble model was significantly better than the other two models in accuracy [p = 0·0005, Repeated measures one-way ANOVA].

Overall, the system was able to achieve an average of 93·73% diagnostic accuracy. Fig. 7 displays the confusion matrices for Inception-V3, ResNet101, and the ensemble classifier at the fold (among 5-folds) having a maximal accuracy. Fig. 8 shows a representative figure of classification result from "InceptionV3 + ResNet101 ensemble" model.

Table 1
Performance table of training models.

Transferred models	Accuracy				GPU time (seconds)	Parameters (millions)	Number of layers
	Full	Full-H25	Quarter	Half			
SqueezeNet	85.55	85.5	73.5	82.8	4137	1.24	68
Alexnet	87.2	83.6	73.7	82.6	3805	61	25
ResNet18	90.65	90.2	83.4	86	4256	11.7	72
MobileNet-v2	90.75	89.8	79.9	84.9	7032	3.5	155
GoogLeNet	90.9	88.7	68.2	85.5	5104	7	144
Resnet50	91.2	91.4	81.3	86.3	7302	25.6	177
Resnet101	91.55	91.7	83.6	86.1	12,215	44.6	347
Inception-v3	92	92.1	84.1	89.5	11,938	23.9	316
InceptionResnet-v2	92.1	91.9	82.2	86.9	33,283	55.9	825

"Quarter" set used about 2000 images for training and validation.
"Half" set used about 5000 images for training and validation.
"Full" represents an average accuracy of twice evaluation of full dataset (80% training and 20% validation).
"Full-H25" represents adding additional 25 fully connected hidden layer to "Full" model.
GPU time represents the processing power needed for training the model.

4. Discussion

Despite many efforts to improve diagnostic accuracy, diagnosis of otitis media mainly relies on otoscopy and often relies on physician's experience [18]. Diagnosis by otoendoscopy requires expertise in image diagnosis; in a study with video-presented examination for diagnosis, otolaryngologists performed significantly better than pediatricians and general practitioners [19]. Even for otolaryngologists, diagnosis of otitis media by otoendoscopy is not trivial. In a study, a series of surveys for diagnosis of eardrum images with twelve fellowship-trained neurotologists in the United States was conducted with overall correctness of diagnosis for ear pathologies ranging from 48·6 to 100%. Along with the diagnosis, the reviewer was also asked to rate the confidence of diagnosis, which revealed overall mean 8·1 out of 10, which means even for a specialist, they are only about 80% certain about their diagnosis on average. In these situations, the current deep network model could help physicians by suggesting possible diagnosis based on otoendoscopic image, and they could achieve better diagnostic accuracy by combining clinical information along with suggestion.

The current image classification model, based on transfer learning with deep convolutional neural network, classified middle ear and EAC pathologies into six categories with a mean accuracy of 93·73%, which is unprecedented in terms of both accuracy and diagnostic diversity. This high accuracy for multiple classes is partly due to the size of the current database. In the current model, 10,544 labelled otoendoscopic images were used, which is significantly bigger than any other studies to our knowledge. Previous studies from other groups utilized 391 and 389 images and yielded 80·6% and 86·84% of five classes focused only on otitis media [3,20]. Compared to other previous studies, the advantage of the current model is that this study included almost all eardrum and EAC pathologies, especially tumors, attic retraction, and eardrum atelectasis, which is a crucial part of diagnosis in the real-world clinical setting. Retraction of the eardrum may indicate chronic otitis media, especially if retraction pocket or destruction is present in the attic area and should not be missed in the clinic. Middle ear tumors such as glomus tumor and congenital cholesteatoma are rare, and due to its rare prevalence, there is a considerable chance of missing the diagnosis unless examined by an experienced physician with clinical suspicion. The current image classification model is the first to diagnose these pathologies.

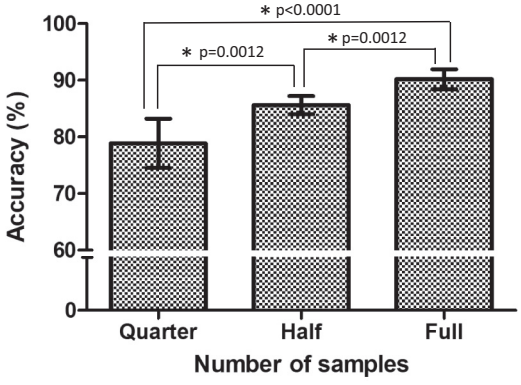

Fig. 3. Accuracy grouped by sample size. The bar represents 95% Confidence interval. Quarter: Data set of $n = 2000$. Half: Data set of $n = 5000$. Full: Data set of $n = 10{,}544$. *: statistically significant [Mann-Whitney test].

It should be noted that we intentionally included all the clinical ear images (except for no ear images), without any selection bias for training. Unlike X-ray images or histology slide images, there is no standardization for image acquisition or quality controls in the ear images. White balance is not always equal, which in turn leads to inconsistent skin or eardrum colour. Camera exposure may not optimally focus on eardrum in case of a tortuous external auditory canal or mass blocking the eardrum. Blurry or out of focus images happen quite often. The eardrum is not always in the centre of the image. Rotation, tilting of the image is inconsistent. Such example is illustrated in Fig. 2B. We included most of the images as long as a clinician could get an impression for diagnosis upon given image. We speculate that this practical image database (including uncleaned data) makes the performance of the model to be dependent on the database size.

Reducing the number of images for training to 2000 images, and 5000 images, the average accuracy was declined to a level of around 80% and 86%. Mann Whitney tests for accuracy among the three conditions showed statistical significance (Fig. 3). These results indicate that it is hard to get a satisfactory result with a small amount of data for training even in the transfer learning, at least in the current ear diagnosis. If the otoendoscopic images were acquired in a standardized setting, similar accuracy could have been achievable with fewer images. A big amount of data is advantageous for a deep network model to find features that explain various disorders regardless of clinical conditions.

Another thing to note is that as the data set gets bigger, the performance gap between training models gets reduced (Table 1). With this in mind, the efficiency of training (training time versus accuracy) for each model should be considered in choosing the best model. As for InceptionResNet-V2 model, the accuracy is only 0·5% better than Inception-V3 model, yet requiring almost 3 times as much training time, in other words, processing power. Therefore, Inception-V3 model seems to be the best choice considering the efficiency of training. When it comes to execution time, the number of parameters are related to calculation time. The MobileNet-V2 model, which has only 3·5 million parameters, achieved 90·75% accuracy, which is not best in terms of accuracy. However, given that it has fewer parameters than other competitors, this model could be more useful in devices with less

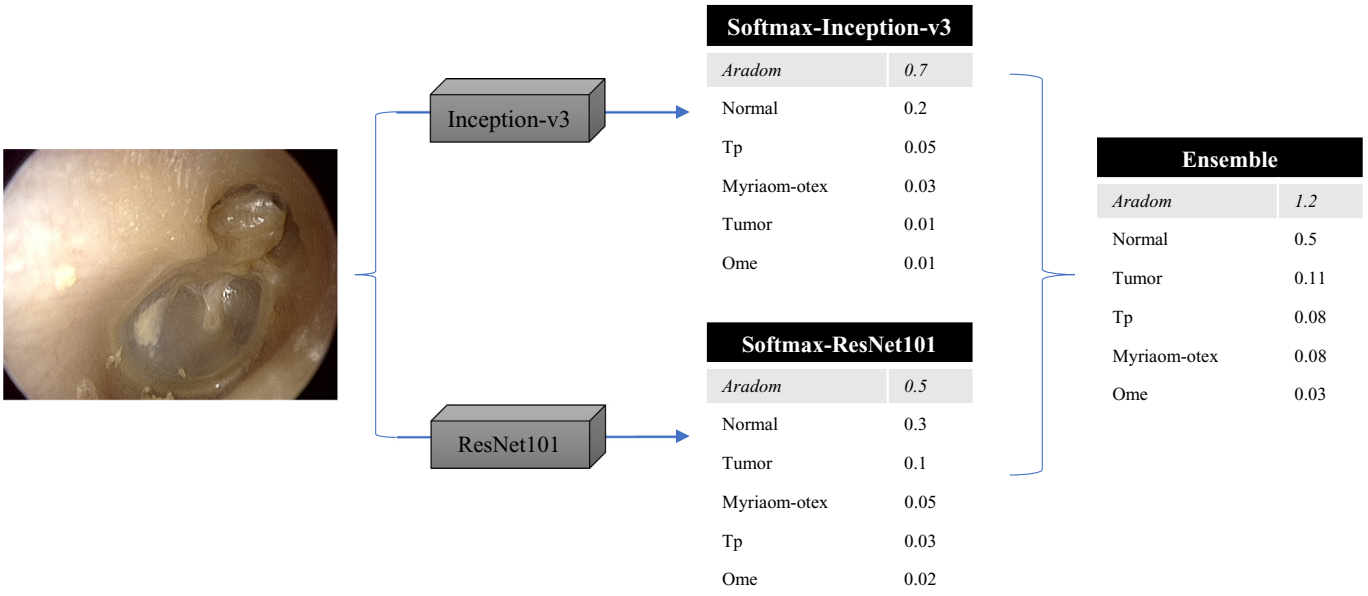

Fig. 4. Schematic example of how ensemble method classifies given otoendoscopic image. The cells shaded in gray background is the final prediction for each model.

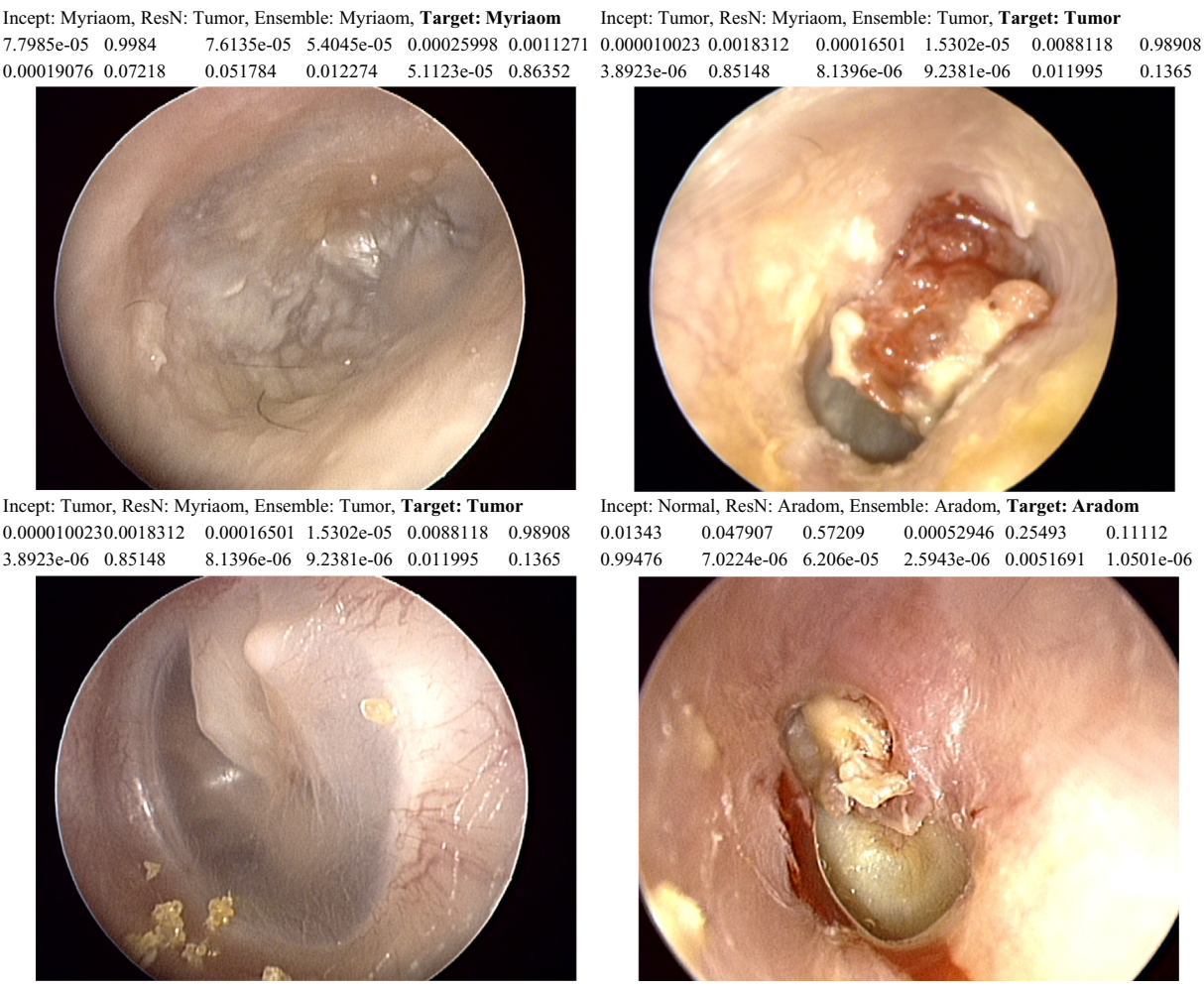

Fig. 5. Examples of inconsistencies between prediction models, and obtaining better accuracy with ensemble model. Incept: Prediction using Inception-V3 model, ResN: Prediction using ResNet101 model, Ensemble: Prediction using ensemble of both models, Target: Ground truth. The second row shows classification scores of Inception-V3 model. The third row shows classification scores of ResNet101 model. Classification scores are represented in the following order: Aradom-Myriaom-Normal-Ome-Tp-Tumor.

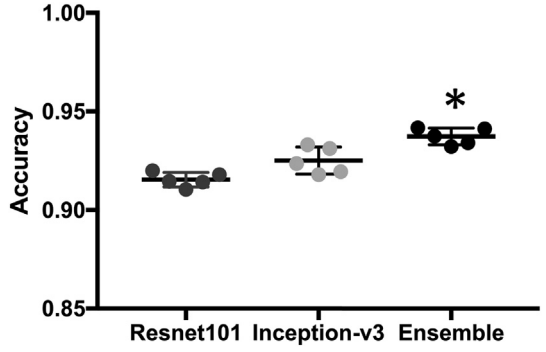

Fig. 6. Accuracy comparison between three methods. The bar represents 1 Standard deviation. Mean accuracy is 0·9154, 0·9251, 0·9373, respectively. *: statistically significant ($P = 0·0001$) [Repeated measures one-way ANOVA].

processing powers, such as mobile phones. In this case, CNN models with fewer parameters may be optimal with acceptable diagnostic accuracy, given that the model has been trained with a sufficiently large number of images.

Adding an additional fully connected layer in front of the final classifiers did not help in this study. Changing colour (RGB) orders by switching R and G channels in the image data show similar or lower accuracy than utilizing natural RGB channels. Although all these variations were not beneficial in the current study, we think it is too early to conclude the generality for these schemes in other applications.

Transfer learning method is popularly used in the medical image analysis as it makes it possible to apply deep learning techniques to a relatively small dataset without significantly sacrificing accuracy. It shows a highly reliable accuracy in various medical image diagnosis [7,8]. This study is in line with the previous studies of transfer learning with fine-tuning to make it applicable to the specific domain of medical image diagnosis. In the field of ophthalmology, transfer learning was applied to diagnose retinal optical coherence tomography (OCT) images, allowing similar accuracy to a model with the full training data with less training data [21,22]. Also, there are studies focused on microscopic histological images utilizing transfer learning for classification [23]. Since transfer learning is efficient in training time, several transfer models can be built practically with a given data set. Instead of using one model, several studies have combined different models to improve classification accuracy [24,25]. In those studies, transfer learning has been used as feature extractors. Features from each model are concatenated to train a new network. Utilizing trained models as feature extractors are limited in the fine-tuning of the transfer layers since the classifier is independently trained with feature extraction networks. Furthermore, feature sets from multiple models may contain redundant information as the number of parameters increases. In contrast to combining features to retrain a new classifier, we simply combined classification scores of each model and determined image labels according to the maximal scores (the probability of being the class). The classification using an ensemble of Inception-V3 and ResNet101 model (Fig. 4) increased diagnostic accuracy significantly (Fig. 6). Usually, transferred Inception-V3 model is a better performer, but in some cases, transferred ResNet101 is more accurate, and the ensemble method was able to take advantage of combining inconsistencies of the two models. Fig. 5 shows an example for this ensemble approach, where shows the score of each model, which is softmax value representing the probability of each classification. Upon inspection, the classifier that has a sum of classification scores close to 2 (maximum 1 for each model), which means the model is almost certain about the diagnosis, tends to be chosen by ensemble model. It resembles a case conference between two physicians, in this case, Inception-V3 and ResNet101, arguing over the right diagnosis and the one with more certainty winning the argument.

Based on the confusion matrix of the ensemble classifier, the classification system is a good performer in the diagnosis of normal, otitis media with effusion, tympanic perforation, and tumors which exceeded over 90% accuracy. Additional representative figure (Fig. 8) illustrates examples of otoendoscopic images. As for otitis externa or myringitis, accuracy is 77·91% with 89·33% sensitivity and 99·02% specificity. It often misdiagnosed as tympanic perforation or tumorous condition; in some cases of myringitis, the EAC may be whitish, wet circular fashion, with the centre being dark, mimicking large perforation. Also, it may be confused with tumors, which makes sense since it hinders the proper view of the eardrum and external auditory canal. Mostly, these images often contain discharges, crusts in the EAC, which should have been removed prior to taking images for better accuracy. Label "ARADOM" refers to attic retraction or adhesive otitis media, accuracy is 85·78% with 90·19% sensitivity and 98·25% specificity. It was commonly misdiagnosed as tympanic perforation or normal. In non-severe cases, retraction could be subtle and clinicians may find it hard to decide whether it is normal or grade I retraction by Tos or Sade classification

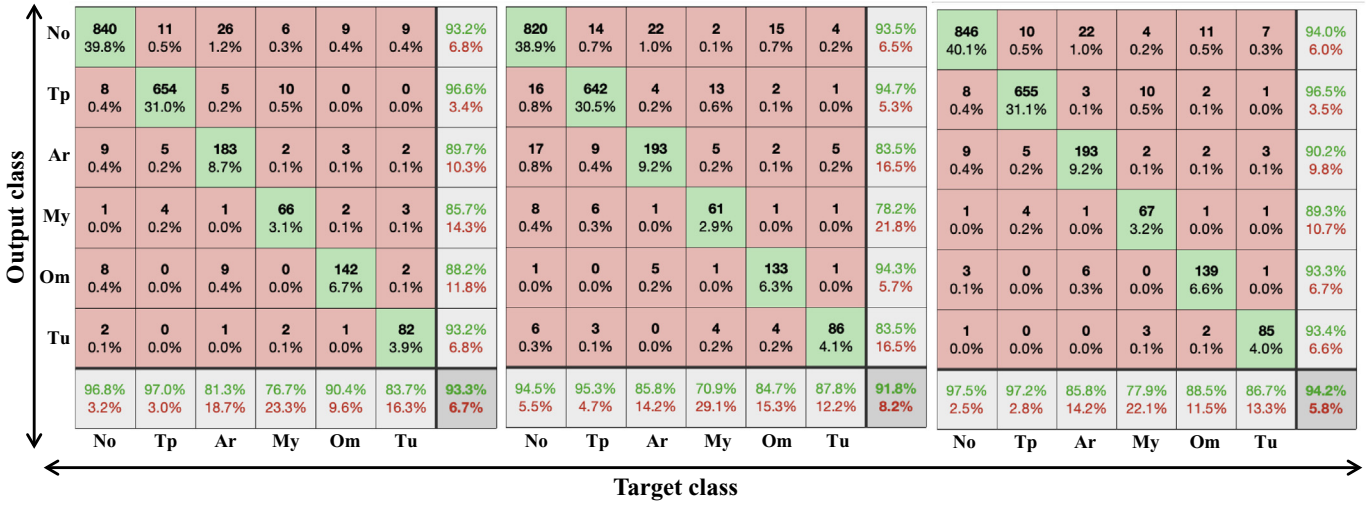

Fig. 7. Confusion matrices for Inception-V3, ResNet101, and ensemble classifier at the fold (among 5-folds) having a maximal accuracy. Target class in x axis refers to ground truth label. Output class in y axis refers to classification by InceptionV3 based model. No: Normal eardrum and external auditory canal (including some tympanosclerosis, healed perforation). Tp: tympanic perforation, Ar: Attic retraction or adhesive otitis media. My: myringitis and/or otitis externa. Om: Otitis media with effusion. Tu: middle ear or external auditory canal tumor or cerumen impaction.

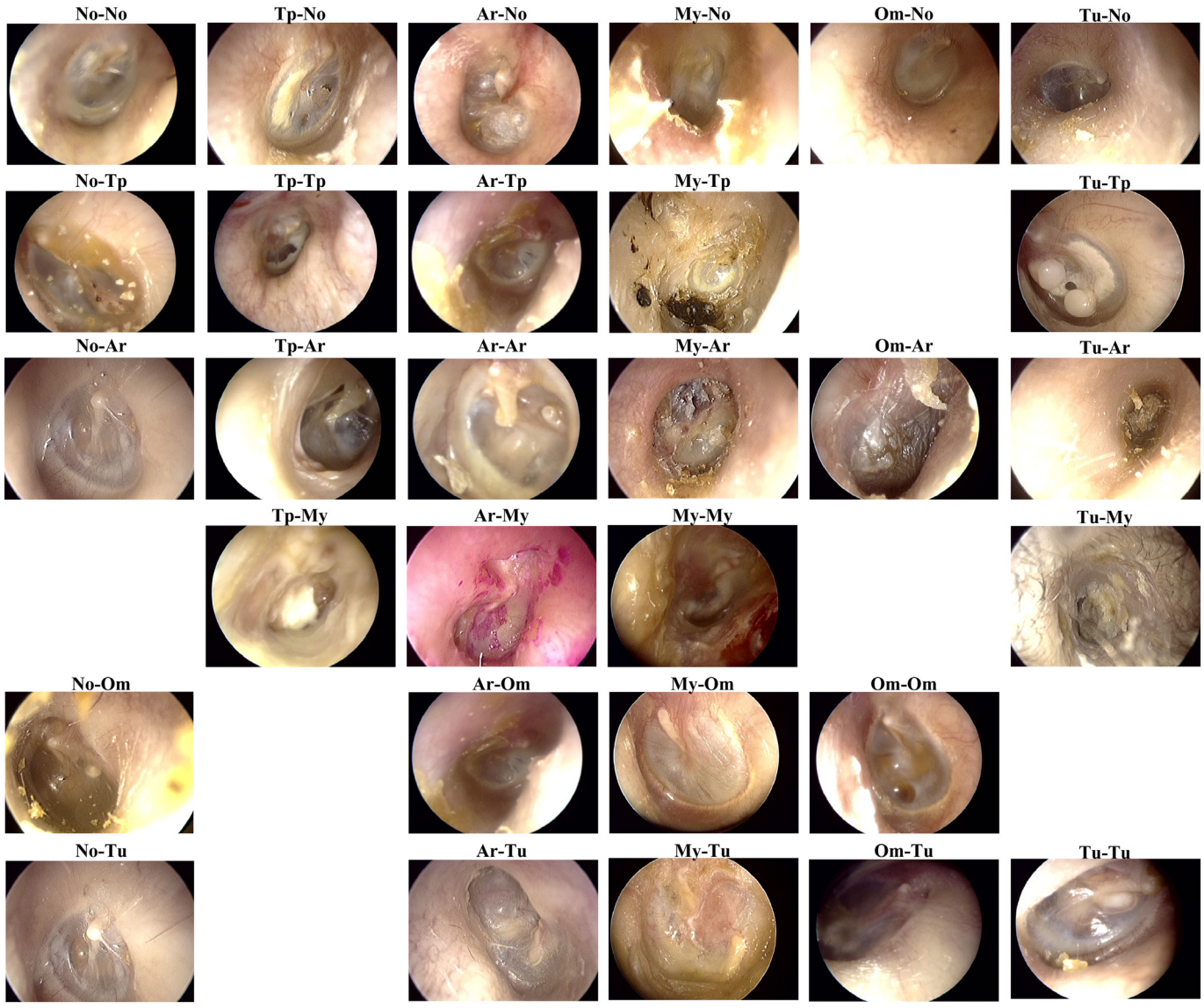

Fig. 8. Representative figure of classification results from "InceptionV3 + ResNet101 ensemble" model. Abbreviations are identical to Fig. 7. Labeling is ordered as Ground truth–Classification. Tu-No: ground truth is tumor, but the system classified as normal. Om-Tu: ground truth is otitis media with effusion, but the system classified as tumor.

[26–28]. On the other hand, severe cases of attic retraction or middle ear atelectasis often reveal the ossicles inside the tympanic membrane, sometimes making it hard to distinguish between total perforation and severe atelectasis. Detecting attic retraction is very important since it implies underlying chronic otitis media, and it often requires surgical treatment to prevent progression. This model's capability of predicting attic retraction or adhesive otitis media, with an accuracy of 85·78% is the most important and practical technologic advancement to be of use in clinics.

In terms of predicting normal and abnormal otoendoscopic findings, overall sensitivity and specificity is 93·69% and 96·82%, respectively. It shows the possibility to be used for screening of ear disease in regular routine health checkup.

Fig. 8 illustrates examples of otoendoscopic image classified in the ensemble model. Trivial cases tend to be appropriately classified, and most misclassified items have some ambiguities; the classification system tends to be not entirely wrong about the diagnosis. For example, in Fig. 8, image labelled TUM-TP, meaning ground truth is tumor, but there is also tympanic perforation present, it does have tympanic perforation, which the classification system has labelled accordingly and is also a correct diagnosis.

Acquisition of the otoendoscopic image could be easily done by non-doctors with a little degree of training, and remote diagnosis based on the otoendoscopic photo may not significantly differ whether the photo was taken by otolaryngologist or telehealth facilitator [29]. In areas short of otolaryngologists, some other speciality doctors (pediatricians, family medicine, or general practitioners), or even non-doctors could take otoendoscopic photos, and analyze images for ear disease based on our system and decide the next step. If the diagnosis is normal or otitis media with effusion, observation is recommended. Otherwise, if otitis externa or myringitis is suggested, physicians of other specialities could try antibiotics before referring the patient to an otolaryngologist for further intervention. For attic retraction, tympanic perforation, or tumors, referring the patient to otolaryngologist would be appropriate for the next step. A system based on the current study could aid early diagnosis of one of the most common childhood illness, otitis media [30], which may alleviate the burden of growing number of patients with hearing impairment.

For otolaryngology specialists, this model could be useful for generating second opinion and be used for double checking the diagnosis, especially tumors and attic retractions which could have been missed due to insufficient experience or low clinical suspicion. This model did not

take external patient factors such as age, presence of fever and otologic symptoms such as (pulsatile) tinnitus, ear fullness, otalgia, hearing loss, otorrhea, etc. In real-life clinical settings, physicians may take otoendoscopic image and correlate the current classifier's results with clinical information for diagnosis. In turn, the current deep learning classifier may be trained with these non-image information for better diagnosis rate.

Considering many previous studies regarding the diagnosis rate of ear disease, which were <80% on average [2,19,31], we carefully claim that this automated diagnosis image classification system can perform better than an average otolaryngologist specialist, and since this classification system covers most of ear disease domains including attic retraction, tumors, which is unprecedented, it is ready for use in real-world clinical settings. Ultimately, it may help the world ease the burden of hearing impairment by contributing to early diagnosis of ear disease.

Data sharing statement

The data are not available for public access because of patient privacy concerns but are available from the corresponding author on reasonable request approved by the institutional review boards of Yonsei university college of medicine.

Funding sources

This research was supported by Brain Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT(NRF-2017M3C7A1049051) and by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number HI18C0160).

Author contributions

D.C designed the study, collected & analyzed the data. C·P and H.P. designed deep transfer learning system and performed training with S.S. D.C drafted the manuscript with J.C and H.P. and all authors reviewed and revised the manuscript.

Declaration of Competing Interest

None.

Acknowledgements

None.

References

[1] Blomgren K, Pitkäranta A. Is it possible to diagnose acute otitis media accurately in primary health care? Fam Pract 2003;20:524–7.
[2] Pichichero ME, Poole MD. Assessing diagnostic accuracy and tympanocentesis skills in the management of otitis media. Arch Pediatr Adolesc Med 2001;155:1137–42.
[3] Myburgh HC, Jose S, Swanepoel DW, Laurent C. Towards low cost automated smartphone- and cloud-based otitis media diagnosis. Biomed Sig Process Control 2018;39:34–52.
[4] Grassmann F, Mengelkamp J, Brandl C, Harsch S, Zimmermann ME, Linkohr B, et al. A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. Ophthalmology 2018. https://doi.org/10.1016/j.ophtha.2018.02.037.
[5] Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402–10.
[6] Hood DC, De Moraes CG. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. Ophthalmology 2018;125:1207–8.
[7] Shin H, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 2016;35:1285–98.
[8] Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 2018; 172:1122–31 [e9].
[9] Sanna M, Russo A, Caruso A, Taibah A, Piras G. Color atlas of endo-otoscopy. Examination-Diagnosis-Treatment 2017 (Page range: pp.8-11, 14-54, 66-74, 81-92, 94-112, 118-138, 160-166, 195-197).
[10] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 2012:1097–105.
[11] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 1–9.
[12] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–8.
[13] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 2818–26.
[14] Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-resnet and the impact of residual connections on learning. Thirty-first AAAI conference on artificial intelligence; 2017.
[15] Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size; 2016.
[16] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. Mobilenetv2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 4510–20.
[17] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv Preprint 2014 arXiv:1412.6980.
[18] Marom T, Kraus O, Habashi N, Tamir SO. Emerging technologies for the diagnosis of otitis media. Otolaryngol Head Neck Surg 2019;160:447–56.
[19] 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.
[20] Myburgh HC, van Zijl WH, Swanepoel D, Hellstrom S, Laurent C. Otitis media diagnosis for developing countries using tympanic membrane image-analysis. EBioMedicine 2016;5:156–60.
[21] Karri SP, Chakraborty D, Chatterjee J. Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration. Biomed Opt Express 2017;8:579–92.
[22] Asaoka R, Murata H, Hirasawa K, Fujino Y, Matsuura M, Miki A, et al. Using deep learning and transfer learning to accurately diagnose early-onset glaucoma from macular optical coherence tomography images. Am J Ophthalmol 2019;198:136–45.
[23] Mazo C, Bernal J, Trujillo M, Alegre E. Transfer learning for classification of cardiovascular tissues in histological images. Comput Methods Programs Biomed 2018;165: 69–76.
[24] Nguyen LD, Lin D, Lin Z, Cao J. Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation. 2018 IEEE International symposium on circuits and systems (ISCAS): IEEE; 2018. p. 1–5.
[25] Cao H, Bernard S, Heutte L, Sabourin R. Improve the performance of transfer learning without fine-tuning using dissimilarity-based multi-view learning for breast cancer histology images. International conference image analysis and recognition. Springer; 2018. p. 779–87.
[26] Tos M. Incidence, etiology and pathogenesis of cholesteatoma in children. Adv Otorhinolaryngol 1988;40:110–7.
[27] Tos M, Poulsen G. Attic retractions following secretory otitis. Acta Otolaryngol 1980; 89:479–86.
[28] Sade J, Berco E. Atelectasis and secretory otitis media. Ann Otol Rhinol Laryngol 1976;85:66–72.
[29] Biagio L, de Swanepoel W, Adeyemo A, Hall 3rd JW, Vinck B. Asynchronous videootoscopy with a telehealth facilitator. Telemed J E Health 2013;19:252–8.
[30] World Health Organization. Chronic suppurative otitis media: burden of illness and management options; 2004.
[31] Moberly AC, Zhang M, Yu L, Gurcan M, Senaras C, Teknos TN, et al. Digital otoscopy versus microscopy: How correct and confident are ear experts in their diagnoses?, 24; 2018; 453–9.