Automatic Prediction of Conductive Hearing Loss Using Video Pneumatic Otoscopy and Deep Learning Algorithm

Hayoung Byun,[1,7] Chae Jung Park,[2,7] Seong Je Oh,[3] Myung Jin Chung,[2,3,4]
Baek Hwan Cho,[3,5] and Yang-Sun Cho[6]

Objectives: Diseases of the middle ear can interfere with normal sound transmission, which results in conductive hearing loss. Since video pneumatic otoscopy (VPO) findings reveal not only the presence of middle ear effusions but also dynamic movements of the tympanic membrane and part of the ossicles, analyzing VPO images was expected to be useful in predicting the presence of middle ear transmission problems. Using a convolutional neural network (CNN), a deep neural network implementing computer vision, this preliminary study aimed to create a deep learning model that detects the presence of an air-bone gap, conductive component of hearing loss, by analyzing VPO findings.

Design: The medical records of adult patients who underwent VPO tests and pure-tone audiometry (PTA) on the same day were reviewed for enrollment. Conductive hearing loss was defined as an average air-bone gap of more than 10 dB at 0.5, 1, 2, and 4 kHz on PTA. Two significant images from the original VPO videos, at the most medial position on positive pressure and the most laterally displaced position on negative pressure, were used for the analysis. Applying multi-column CNN architectures with individual backbones of pretrained CNN versions, the performance of each model was evaluated and compared for Inception-v3, VGG-16 or ResNet-50. The diagnostic accuracy predicting the presence of conductive component of hearing loss of the selected deep learning algorithm used was compared with experienced otologists.

Results: The conductive hearing loss group consisted of 57 cases (mean air-bone gap = 25 ± 8 dB): 21 ears with effusion, 14 ears with malleus-incus fixation, 15 ears with stapes fixation including otosclerosis, one ear with a loose incus-stapes joint, 3 cases with adhesive otitis media, and 3 ears with middle ear masses including congenital cholesteatoma. The control group consisted of 76 cases with normal hearing thresholds without air-bone gaps. A total of 1130 original images including repeated measurements were obtained for the analysis. Of the various network architectures designed, the best was to feed each of the images into the individual backbones of Inception-v3 (three-column architecture) and concatenate the feature maps after the last convolutional layer from each column. In the selected model, the average performance of 10-fold cross-validation in predicting conductive hearing loss was 0.972 mean areas under the curve (mAUC), 91.6% sensitivity, 96.0% specificity, 94.4% positive predictive value, 93.9% negative predictive value, and 94.1% accuracy, which was superior to that of experienced otologists, whose performance had 0.773 mAUC and 79.0% accuracy on average. The algorithm detected over 85% of cases with stapes fixations or ossicular chain problems other than malleus-incus fixations. Visualization of the region of interest in the deep learning model revealed that the algorithm made decisions generally based on findings in the malleus and nearby tympanic membrane.

Conclusions: In this preliminary study, the deep learning algorithm created to analyze VPO images successfully detected the presence of conductive hearing losses caused by middle ear effusion, ossicular fixation, otosclerosis, and adhesive otitis media. Interpretation of VPO using the deep learning algorithm showed promise as a diagnostic tool to differentiate conductive hearing loss from sensorineural hearing loss, which would be especially useful for patients with poor cooperation.

Key words: Air bone gap, Convolutional neural network, Machine learning, Malleus incus fixation, Middle ear effusion, Ossicular fixation, Otitis media, Otosclerosis, Pneumatic otoscope, Tympanic membrane.

(Ear & Hearing 2022;43;1563–1573)

INTRODUCTION

The pneumatic otoscope was first introduced by Dr. E. Siegle in 1864 to obtain better information on the mobility of the tympanic membrane (TM) especially in poorly cooperative patients (Schwartz 1980). It recently has been widely used as a useful and easy diagnostic tool for middle ear effusions (MEE) (Mains & Toner 1989; Jones & Kaleida 2003; Rosenfeld et al. 2004; Harris et al. 2005; Cho et al. 2009; Lee et al. 2011; King & Couch 2015; Rosenfeld et al. 2016). By connecting a pneumatic otoscope to a high-resolution endoscope and video camera to perform a video pneumatic otoscopy (VPO), clinicians can share their findings on a screen and record dynamic movements as video (Cho et al. 2009; Lee et al. 2011). In addition, since the VPO findings reveal not only the presence of middle ear fluids but also dynamic movements of the TM and part of the ossicles, analyzing VPO images could also be useful for evaluating middle ear conditions other than MEE. When differentiating conductive hearing loss in patients with intact TM, for example, the diagnostic accuracy of VPO in predicting malleus-incus fixation was reported to be comparable to that of temporal bone CT (Lee et al. 2011).

Advances in machine learning technology allow computational models to learn from vast amounts of data, and their application is widening. The most commonly used algorithm in the field of computer vision is the convolutional neural network (CNN), whose architecture was inspired by the connectivity pattern of the neurons in the human visual cortex; it is especially useful for object finding and image recognition (Lecun et al. 1998; LeCun et al. 2015; Schmidhuber 2015). Since about 2012, a growing number of studies in various medical fields

[1]Department of Otorhinolaryngology-Head and Neck Surgery, Hanyang University College of Medicine, Hanyang University Medical Center, Seoul, South Korea; [2]Department of Digital Health, SAIHST, Sungkyunkwan University, Seoul, South Korea; [3]Medical AI Research Center, Samsung Medical Center, Seoul, South Korea; [4]Department of Radiology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, South Korea; [5]Department of Medical Device Management and Research, SAIHST, Sungkyunkwan University, Seoul, South Korea; and [6]Department of Otorhinolaryngology-Head and Neck Surgery, Sungkyunkwan University School of Medicine, Samsung Medical Center, Seoul, South Korea; [7]These authors contributed equally to this work.

have reported promising results on automated diagnosis from medical images using deep learning (Cha et al. 2019; Kim et al. 2019; Liu et al. 2019; Cho et al. 2020; Khan et al. 2020; Park et al. 2021; Byun et al. 2021). In a systematic review and meta-analysis, the diagnostic performance of deep learning models was reported to be equivalent to that of health-care professionals (Liu et al. 2019).

Apart from replacing human experts, we can expect to use deep learning algorithms to detect pathological conditions that cannot be recognized by human vision. In this study, we hypothesized that a deep learning algorithm would be able to interpret the VPO images to automatically detect the presence or absence of an air-bone gap (AB gap) shown in pure-tone audiometry (PTA). If the conductive component of hearing loss can be successfully distinguished from sensorineural hearing loss via office-based pneumatic otoscopy, it would be useful in clinics without facilities for bone conduction hearing tests as well as for poorly cooperative patients such as people with disabilities or pediatric patients. Hence, the aim of this pilot study was to develop a deep learning model for detecting conductive hearing loss via VPO images, and to compare its performance with that of experienced clinical otologists.

MATERIALS AND METHODS

Patients and Dataset

The medical records of patients aged 18 years and older who underwent VPO tests and PTA on the same day from 2007 to 2019 were reviewed for enrollment in Samsung Medical Center, a tertiary referral hospital. The pneumatic otoscopy was performed in patients who complained of hearing loss and showed intact TM. Conductive hearing loss was defined as an average AB gap of more than 10 dB at 0.5, 1, 2, and 4 kHz on PTA. In the cases with conductive hearing losses (CHL group), cases confirmed with a specific causative diagnosis through an otologic procedure or exploratory tympanotomy were enrolled for the analysis, as follows: MEE, ossicular fixation, otosclerosis, middle ear mass, congenital cholesteatoma, and adhesive otitis media. Since VPO can only be performed in ears with intact TMs, cases with TM perforations were not included. Incomplete VPO images with poorly visualized landmarks (i.e., cerumen covering the malleus or no visualization of posterior annulus of TM) in the video were excluded from the analysis. The control group (non-CHL group) consisted of cases with normal hearing thresholds without AB gaps.

This study was approved by the Institutional Review Board of Samsung Medical Center and performed in accordance with the Declaration of Helsinki (IRB No.2019-10-171). As it was a retrospective study using anonymous clinical data, written consent was waived since it met the document exemption requirements for informed consent.

Video Pneumatic Otoscopy and Recording

Pneumatic otoscopy was performed using a HAWKE Pneumatic adaptor (model no. 119500) distributed by Karl Storz Endoscopy-America, Inc (El Segundo, CA, USA) (**Fig. 1**). The findings were displayed with a CCD camera (OTV-SP1) and a video monitor system (LMD-2140MD) manufactured by Olympus (Tokyo, Japan), and recorded with a video capture adaptor (miroVIDEO DC 30) by Pinnacle Systems (Mountain View, CA, USA). The video files were saved at a resolution of 72 pixels per inch in MPEG format (352 × 240 pixels, 29.97 frames per second) from 2007 to 2014, and in WMV format (320 × 240 pixels, 30 fps) afterwards.

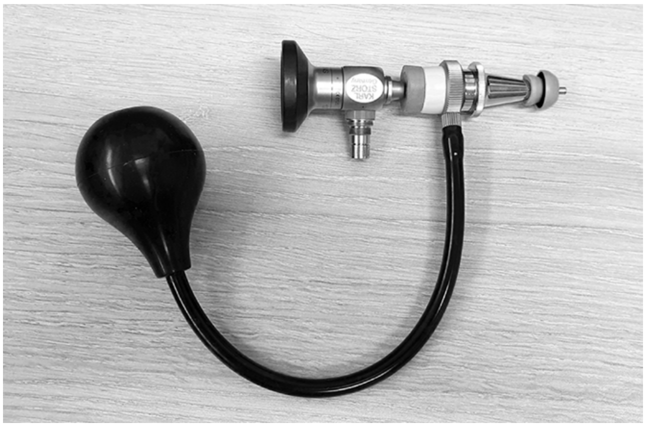


Fig. 1. Pneumatic adaptor with rubber bulb connected to an endoscope.

All included VPO was performed by a single senior otologist and involved the following: (1) positioning of the pneumatic otoscope with a half-squeezed rubber bulb, (2) airtight sealing of the ear canal with an appropriate ear speculum, (3) gently applying initial positive pressure to cause medial movement of the TM, followed by negative pressure to pull the TM laterally, (4) repeatedly applying positive and negative pressure while recording the movements of the TM.

Image Preparation

A cycle of TM motion was defined based on the point at which the TM returned to its original position after applying positive and negative pressure once each. As in previous VPO studies (Cho et al. 2009; Lee et al. 2011), two significant images in each cycle were selected, namely at the most medial position on positive pressure (POS image) and the most lateral position on negative pressure (NEG image) (**Fig. 2A**). For each VPO video, up to 5 cycles were included in the analysis, in order (**Fig. 2B**).

Selected significant frames were extracted from the video using the Imageio library for PythonTM (Python Software Foundation, Wilmington, DE, USA). The images obtained were adjusted to an RGB image with three channels with a resolution of 200 pixels per inch, and were normalized by the method of mean subtraction.

Image augmentation was performed at a scale of 60 times. For all images, rotations of 0, 5, 10, 15, 30 degrees, left-right flips, and cropping with pixel shifts of 10, 13, and 16 were applied.

Deep Learning Algorithm: Training and Cross-Validation

Pretrained versions of CNN [Inception-v3, VGG-16, and ResNet-50 (Simonyan & Zisserman 2014; Szegedy et al. 2015; He et al. 2015; Szegedy et al. 2016)] were applied to classify VPO images into two diagnostic groups: CHL and non-CHL groups.

It was natural to come up with a deep learning model architecture that takes both POS and NEG images within a cycle simultaneously as inputs because clinicians can obtain information from dynamic movements of TM between those images

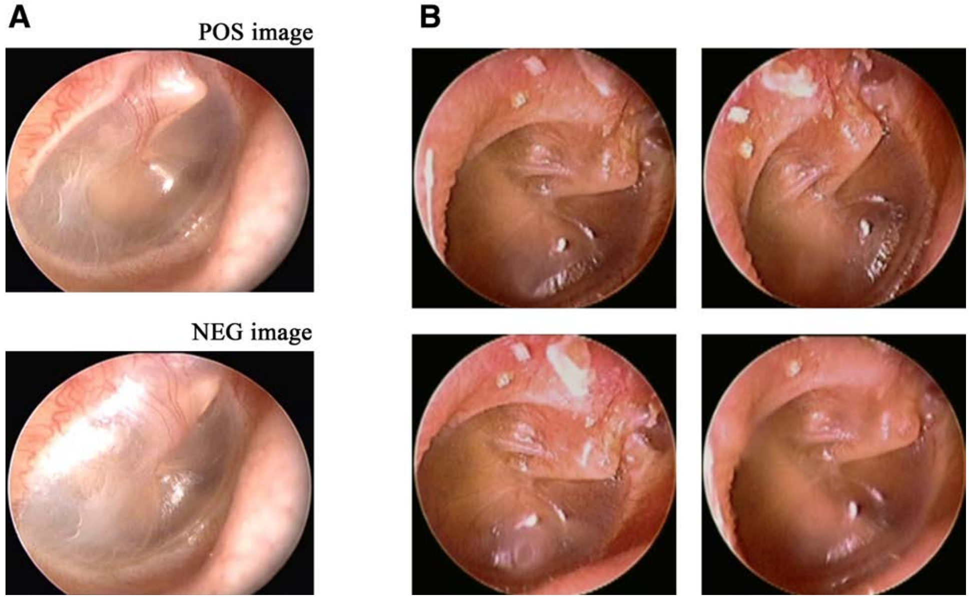


Fig. 2. Selected images from a VPO video. A, POS and NEG images of a case without conductive hearing loss. B, Multiple POS images, one for each repeated cycle, in a case with conductive hearing loss. NEG image, most lateral image in negative pressure on pneumatic otoscopy; POS image, most medial image in positive pressure on pneumatic otoscopy; VPO, video pneumatic otoscopy.

(Cho et al. 2009; Lee et al. 2011). As a method of feeding multiple images in parallel into a deep learning model, a multi-column CNN has been suggested and showed its effectiveness (Ciresan et al. 2012; Zhang et al. 2016; Jin et al. 2019; Choi et al. 2021). Therefore, we devised multi-column CNN models to train both POS and NEG images simultaneously, where the features from dynamic input images of each column are abstracted independently and combined by various arithmetic operations (concatenate, summation, or subtraction) at the near-endpoint.

To find the most effective multi-column learning model, various CNN architectures were trained using Inception-v3 (**Fig. 3**). At first, POS and NEG images were fed into the individual backbones of Inception-v3, and the feature maps after the last convolutional layers from each CNN backbone (a column) were concatenated (Model 1). Thus, Model 1 tries to learn simultaneously on multiple CNN backbones (multi-column architecture) with a single loss function. The main advantage of the multi-column architecture is that the CNN models can capture representations of the data by considering multiple image inputs simultaneously. Model 2 employed element-wise summation or subtraction using the last convolution output feature maps of each POS and NEG image column instead of the channel-wise concatenation as in Model 1. Lastly, in Model 3, an additional column of difference images from POS and NEG images was added, and the feature maps from the three columns were summed or concatenated to build a final classifier (**Fig. 3**). Then, a selected learning model was tested on additional networks of which backbones were VGG-16 and ResNet-50.

The training set included 90% of all images, and the test set contained the remaining 10%. For unbiased learning, training images were randomly selected to include diagnoses of conductive hearing loss selected as randomly as possible. Tenfold cross-validations were performed due to the small amount of input data.

Computing was performed on Linux machines with CPU Intel Xeon Processors E5-2620 v4 @ 2.10GHz and GPU NVIDIA GeForce GTX 1080Ti.

Assessment of Performance

The performance of the model was assessed by the areas under the curve (AUC) of the receiver operating characteristic (ROC) curve reflecting the sensitivity and specificity of model predictions. The 95% confidence intervals (CIs) of AUCs were also calculated. After reading the Excel file containing the prediction results into Python using the Pandas library, a ROC curve was drawn and the AUC was calculated using the Matplotlib and the sklearn package. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated from the point of the maximum Youden index (sensitivity + specificity − 1) of the ROC curves.

The diagnostic accuracy of the algorithm was compared to that of three experienced clinical otologists who had been practicing in the otology field for more than 8 years. One of these experts (Otologist 1) had been using VPO in routine clinical practice, while the others (Otologist 2 and 3) used VPO tests occasionally when needed. At first, all selected test images, with paired POS and NEG images, were randomly presented to the human experts without clinical information to record their decisions about the presence or absence of conductive hearing loss. Next, the original VPO videos of those cases were presented in a separate section in the same way, and diagnostic accuracy was calculated.

For visual interpretability, the regions of interest for the models to make predictions were visualized as heat maps using an explainable AI technique, Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al. 2017).

RESULTS

Dataset

A total of 133 ears were included in this study. There were 57 in the CHL group (mean AB gap, mABG ± SD = 25 ± 8 dB): 21 ears with MEE (mABG 25 dB), 14 with malleus-incus fixation (mABG 21 dB), 15 with stapes fixation including otosclerosis (mABG 26 dB), 1 with incus-stapes joint loosening (ABG 37

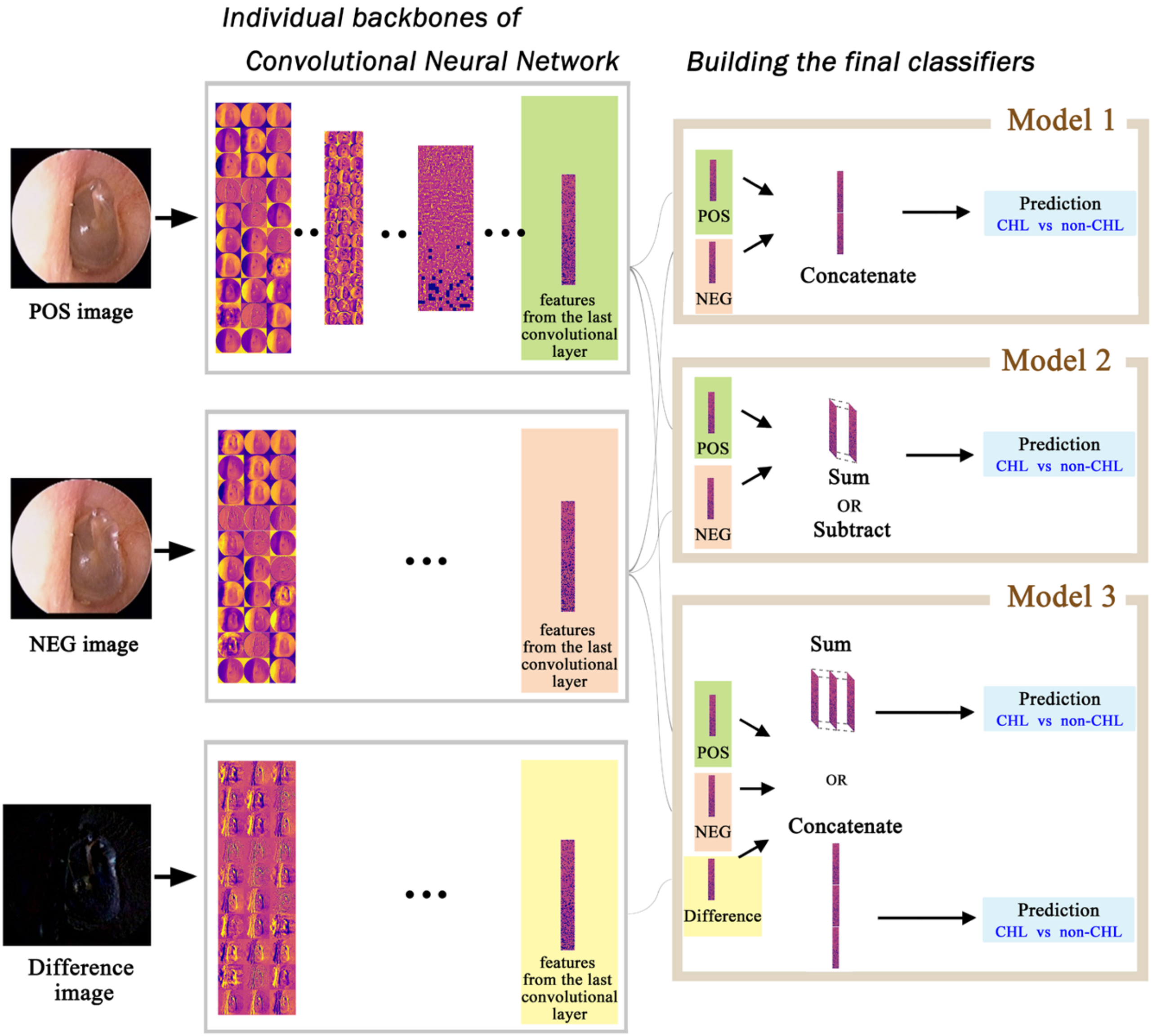


Fig. 3. Schematic diagram of multi-column deep learning models. The features of each image are abstracted independently through each column, and the last feature layers were joined arithmetically by the ways in Models 1 to 3, which were designed to identify the most effective learning model.

dB), 3 with adhesive otitis media (mABG 22 dB), and 3 with middle ear masses including congenital cholesteatoma (mABG 35 dB) (**Figs. 4, 5**). There were 76 in the non-CHL group. There was an average of 4.3 repeated measurement records for each case; 240 cycles were included in the CHL group and 325 in the non-CHL group (CHL: non-CHL = 1: 1.35). Two images, one POS and one NEG, from each cycle were extracted, and a total of 1130 original images were used in the analysis.

Performances of the Multi-Column CNN Models

The optimal hyper-parameters acquired from the experiments were 0.001 for learning rate, 16 for the batch size, and Stochastic Gradient Descent for the optimizer when trained for 30 epochs without applying further fine-tuning or learning scheduling.

The classification performances of the models, measured in mean AUCs (mAUCs) of 10-fold tests, are shown in Figure 6. In the two-column CNN algorithms (Models 1 and 2), training with subtracted features performed better in predicting CHL (mAUC 0.971 ± 0.125) than feature concatenation or summation. In the three-column approach (Model 3), concatenated feature training gave the highest mAUC (0.972 ± 0.134, 95% CI: 0.949 to 0.991) of all the models (**Fig. 6**).

Performances of the CNN Networks

The performance of Inception-v3, ResNet-50, and VGG-16 with the Model 3 concatenation approach was shown in Figure 7. The average classification performances of three deep learning networks were as follows: 0.972 ± 0.049 (95% CI: 0.949 to 0.991) with Inception-v3, 0.965 ± 0.061 (95% CI: 0.934 to 0.987) with ResNet-50, and 0.952 ± 0.070 (95% CI: 0.909 to 0.985) with VGG-16 (Fig. 7). Regarding the model complexity, number of parameters were 24 million for Inception-v3, 25 million for ResNet-50, and 138 million for VGG-16. The three-column concatenated feature training (Model 3) using Inception-v3 provided the best prediction (highest mAUC) with

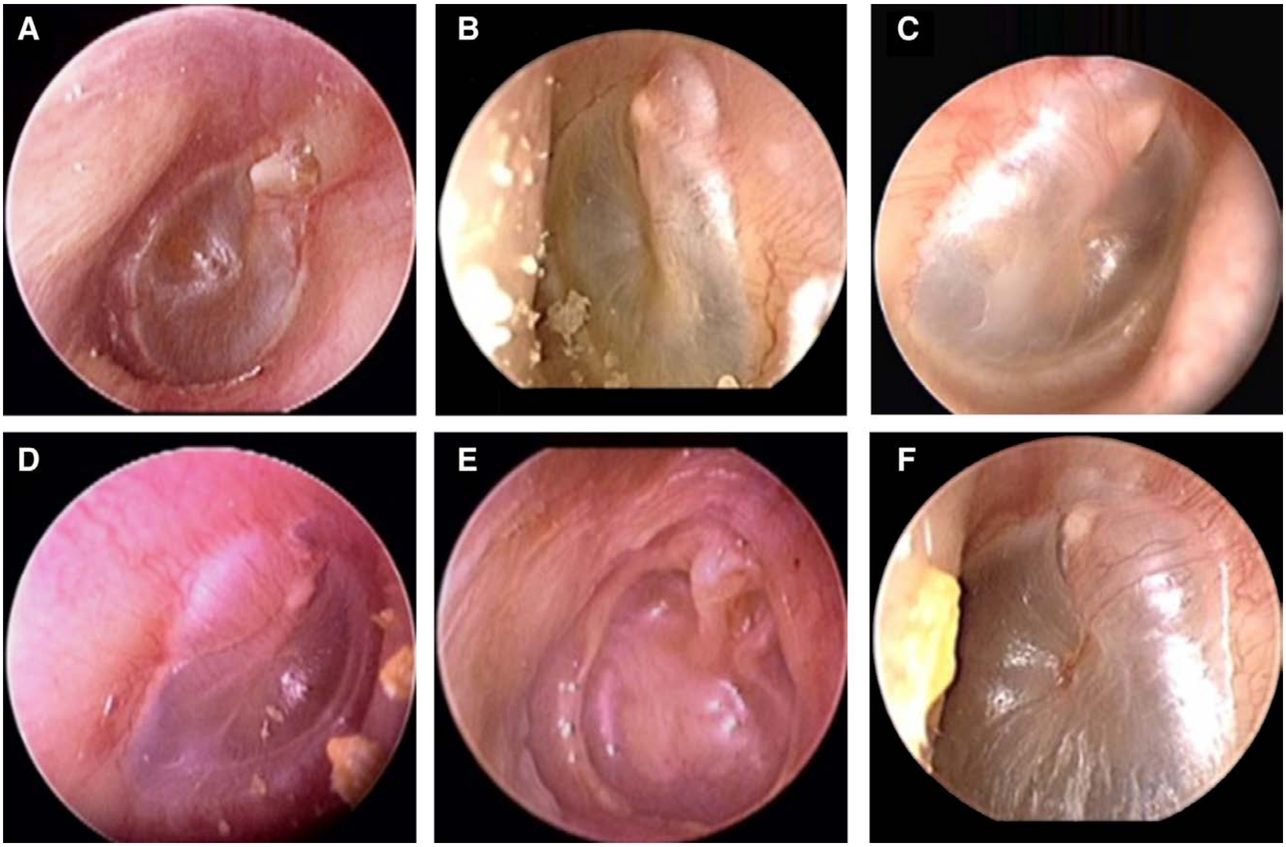


Fig. 4. Cases with conductive hearing loss at negative pressure during video pneumatic otoscopy: A, middle ear effusion; B, malleus-incus fixation; C, stapes fixation; D, loose incus-stapes joint; E, adhesive otitis media; F, middle ear mass (congenital cholesteatoma involving the ossicular chain).

the lowest complexity of all the models and was selected as a representative deep learning model.

Comparison With Experienced Otologists

The performance of the invited otologists was better with the original videos (mAUC 0.776) than with the selected still images (mAUC 0.697) (**Fig. 8**). The most accurate judgment was made by Otologist 1 using the original videos, with 98.7% specificity, 97.1% PPV, and 81.2% accuracy (**Fig. 8**) (**Table 1**). Overall, the accuracy of the deep learning algorithm was higher than that of the otologists (**Fig. 9**) (**Table 1**). The deep learning algorithm correctly predicted hearing loss in 13/15 (86.7%) of the cases with stapes fixations, while the otologists diagnosed only 1.6/15 (11.1%) of them, on average (**Fig. 9**).

Visualization of the region of interest in the deep learning model revealed that the algorithm made decisions generally based on findings for the malleus and nearby TM (**Figs. 10, 11**).

DISCUSSION

In this study, a deep learning algorithm for analyzing VPO images was developed, and its usefulness in detecting the presence of conductive hearing loss (AB gap > 10 dB in pure tone audiometry) was assessed. Two significant still VPO images, the most medial on positive pressure and the most lateral on

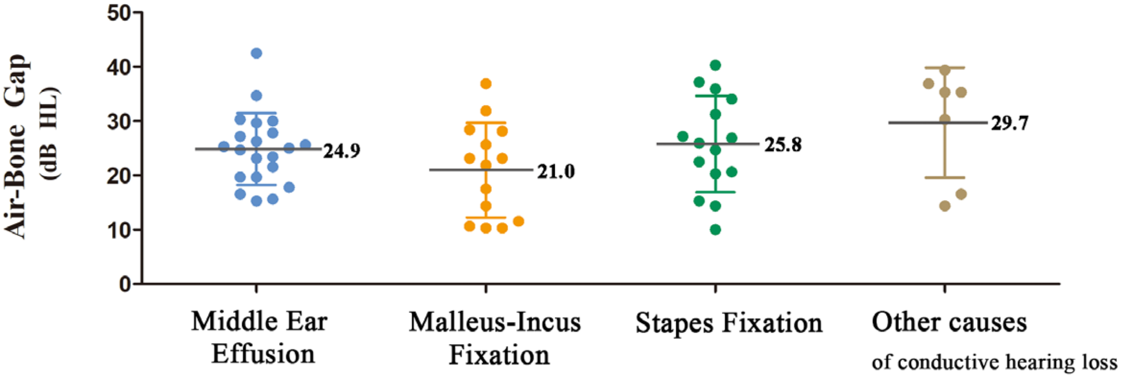
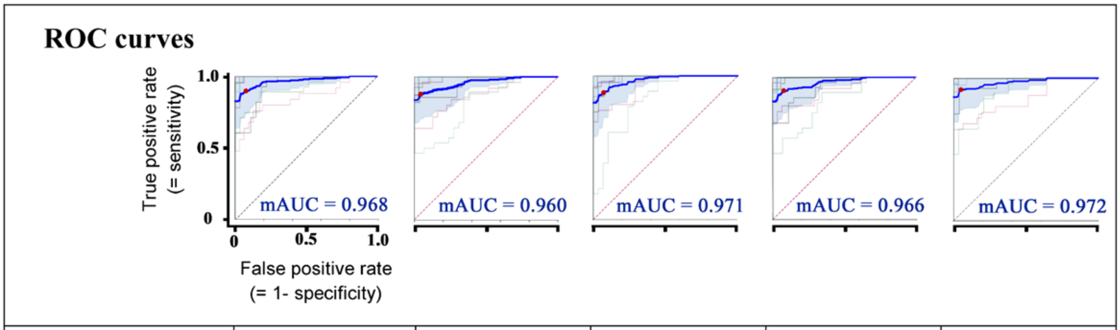
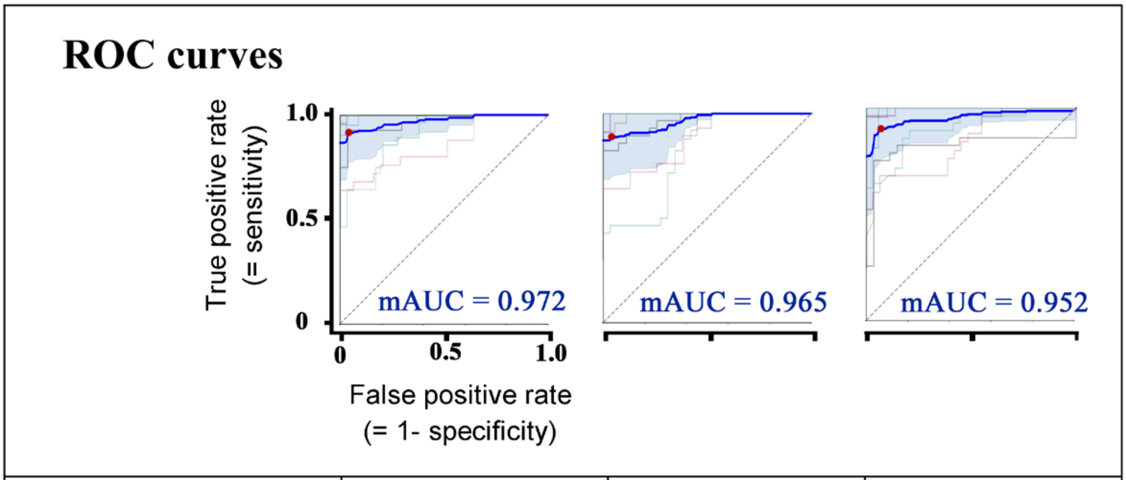


Fig. 5. Air-bone gaps of the cases according to the causative diagnosis. There was no significant difference among the diagnostic groups (p = 0.194 by one-way analysis of variance test). Black bars and numbers represent average air-bone gaps. Error bars indicate 1 SD.



	Model 1	Model 2 sum	Model 2 subtract	Model 3 sum	Model 3 concatenate
95% CI of mAUC	0.935 – 0.992	0.926 – 0.984	0.942 – 0.990	0.937 – 0.989	0.949 – 0.991
Sensitivity (%) at the maximum Youden index	90.0	88.0	86.7	87.8	91.6
Specificity (%) at the maximum Youden index	92.9	96.0	96.0	96.0	96.0

Fig. 6. The AUC of the ROC curves for the multi-column CNN models using Inception-v3 backbones. Mean AUCs were noted at the right-bottom of each graph. Sensitivity and specificity were calculated from the point of the maximal Youden index (red dots) of the ROC curves. The highest mAUC was observed in Model 3 concatenated features model. (ROC curves from 10-folds, thin lines; 1 SD, light-blue shadow; Youden index, sensitivity + specificity – 1). AUC indicates areas under the curve; CNN, convolutional neural network; ROC, receiver operating characteristic.



	Inception-v3	ResNet-50	VGG-16
95% CI of mAUC	0.949 – 0.991	0.934 – 0.987	0.909 – 0.985
Sensitivity (%) at the maximum Youden index	91.6	88.3	89.9
Specificity (%) at the maximum Youden index	96.0	96.0	92.9

Fig. 7. The AUC of the ROC curves for the CNN networks (Inception-v3, ResNet-50, and VGG-16). Mean AUCs were noted at the right-bottom of each graph. Sensitivity and specificity were calculated from the point of the maximal Youden index (red dots) of the ROC curves. The highest mAUC was noted in the Inception-v3 network. (ROC curves from 10-folds, thin lines; 1 SD, light-blue shadow; Youden index, sensitivity + specificity – 1). AUC indicates areas under the curve; CNN, convolutional neural network; ROC, receiver operating characteristic.

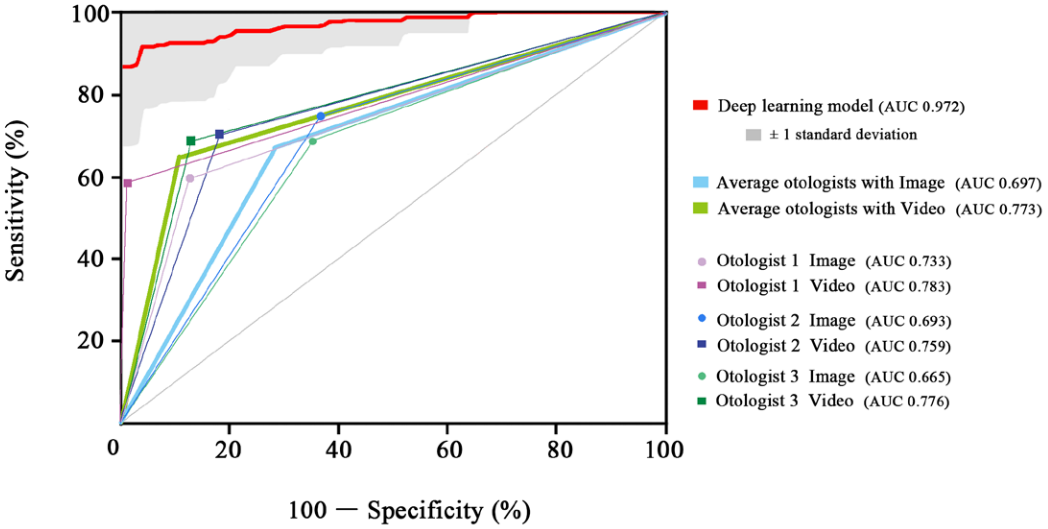


Fig. 8. The AUC of the receiver operating characteristic curves for the selected deep learning algorithm and the invited experts. The otologists were required to judge the presence or absence of conductive hearing loss from the selected still images (circles and thick sky-blue) and original VPO videos (squares and thick lines), in separate test sections. AUC indicates areas under the curve; VPO, video pneumatic otoscopy.

TABLE 1. The diagnostic accuracies of the selected deep learning model and the invited otologists in predicting the presence of conductive hearing loss via video pneumatic otoscopy video

	Deep Learning	Otologist 1	Otologist 2	Otologist 3	Otologists on Average
Sensitivity	91.6	57.9	70.2	68.4	65.5
Specificity	96.0	98.7	81.6	86.9	89.0
Positive predictive value	94.4	97.0	74.1	79.6	81.8
Negative predictive value	93.8	75.8	78.5	78.6	77.5
Accuracy	94.1	81.2	76.7	78.9	79.0

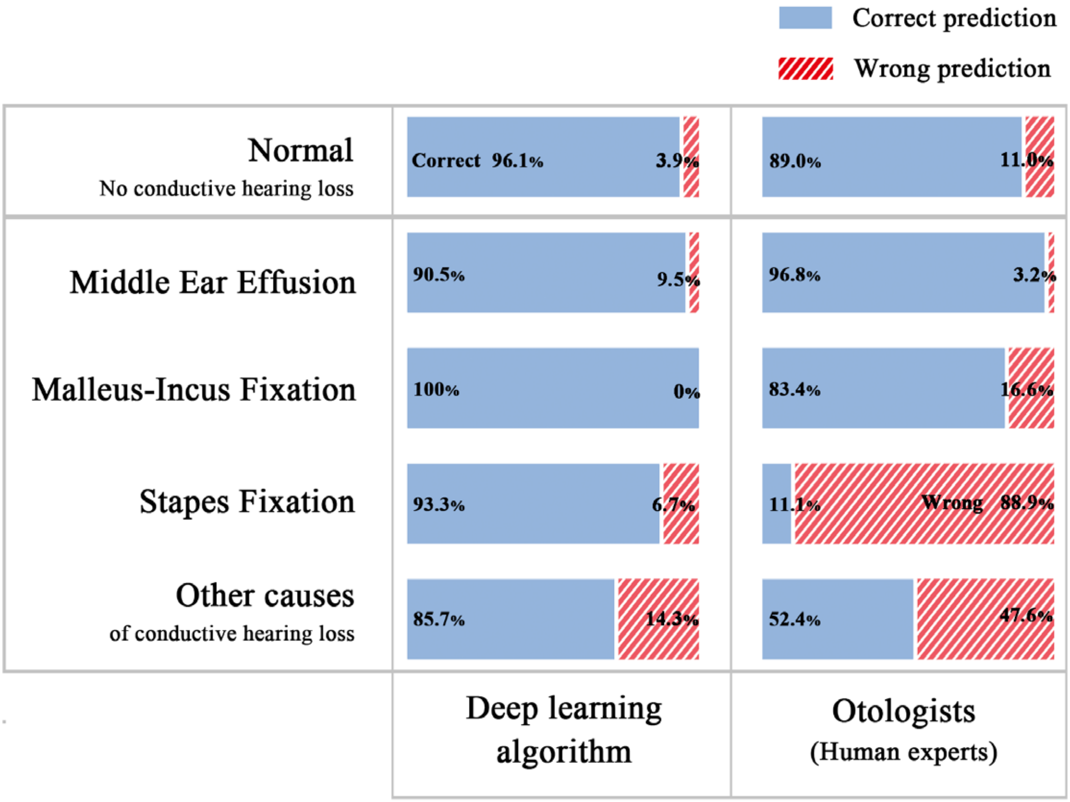


Fig. 9. Bar graphs showing the percentages of correct predictions for the presence of air-bone gaps (blue-filled), by the deep learning algorithm and the human experts, according to causative diagnosis of conductive hearing loss.

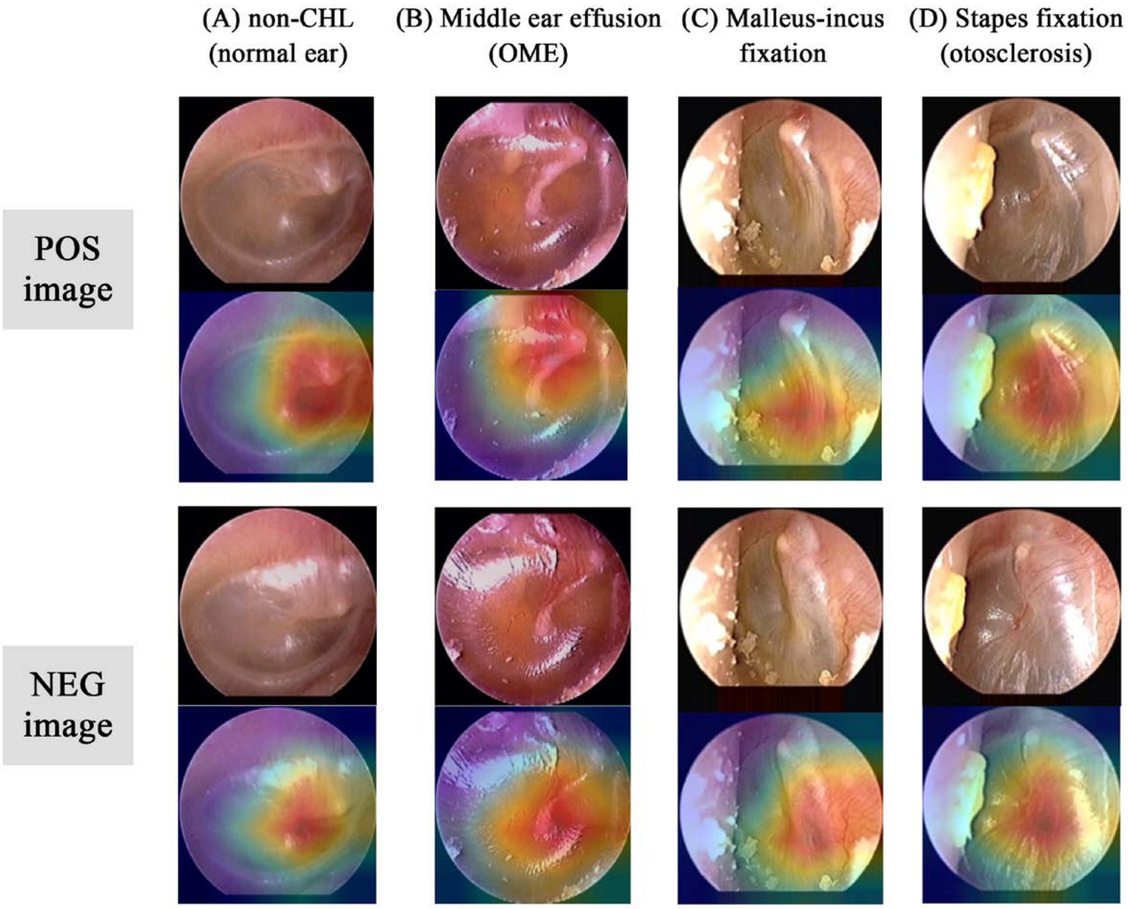


Fig. 10. Heat maps of the regions of interest of the deep learning algorithm in correctly predicted cases. OME indicates otitis media with effusion; POS image, most medial image in positive pressure on pneumatic otoscopy; NEG image, most lateral image in negative pressure on pneumatic otoscopy.

negative pressure, extracted from each of the original VPO videos were used in the analysis (**Fig. 2**). Among CNN models tested, the multi-column concatenated model with Inception-v3 (**Fig. 3**) gave the highest mAUC of 0.972 (95% CI: 0.949 to 0.991), 91.6% sensitivity, 96.0% specificity, 94.4% PPV, 93.9% NPV, and 94.1% accuracy (**Figs. 6, 7**), which was superior to that of the invited human experts, who had 0.773 mAUC and 79.0% accuracy on average (**Fig. 8**). In stapes fixations, the algorithm correctly predicted over 86% of cases as having conductive components of hearing loss (**Fig. 9**). To the best of our knowledge, the present study is the first to apply the pneumatic otoscopy to predict conductive component of hearing loss, as well as the first to develop a deep learning algorithm interpreting VPO findings.

There has been several deep learning researches on medical images in otologic field (Cho et al. 2020; Park et al. 2021; Byun et al. 2021). For TM findings, previous studies have mainly used deep learning to classify static TM images, without applying pressure changes, into disease categories suggested by human specialists (Cha et al. 2019; Khan et al. 2020; Byun et al. 2021). In those studies, however, the best performance of the deep learning algorithms did not exceed that of the human specialists who set the gold standard. The present study attempted to detect functional abnormalities causing conductive hearing loss through the VPO findings, which is challenging for human experts and possibly more useful in clinical practice.

Previous studies have examined the usefulness of pneumatic otoscopy in diagnosis of conductive hearing loss due to malleus-incus fixations as well as otitis media with effusions (Schwartz 1980; Mains and Toner 1989; Rosenfeld et al. 2004; Harris et al. 2005; Lee et al. 2011). Although predicting malleus-incus fixation by calculating geometric relations between the position of the umbo and annular rim in VPO were possible (Lee et al. 2011), it was still impossible to diagnose stapes fixations via human visual perception of the TM. In this study, we hypothesized that a deep learning model with a convolutional network could detect the presence of conductive component of hearing loss by referencing clues contained in VPO images. Interestingly, this preliminary study showed that CNN networks successfully predicted the presence of AB gaps even in otosclerosis (**Fig. 9**).

We compared the performance of different deep learning models including VGG-16, ReNet50, and Inception-v3, and the results showed Inception-v3 with three-column concatenated features model generated the highest mAUC, although differences were not statistically significant (p > 0.05 by analysis of variance tests). In our data, the model performance was not related to the model complexity, as the number of parameters was the lowest in Inception-v3. Due to the familiarity and interpretability of the deep learning models were considered similar in all our models, we selected the representative model with the highest mAUC (**Fig. 7**).

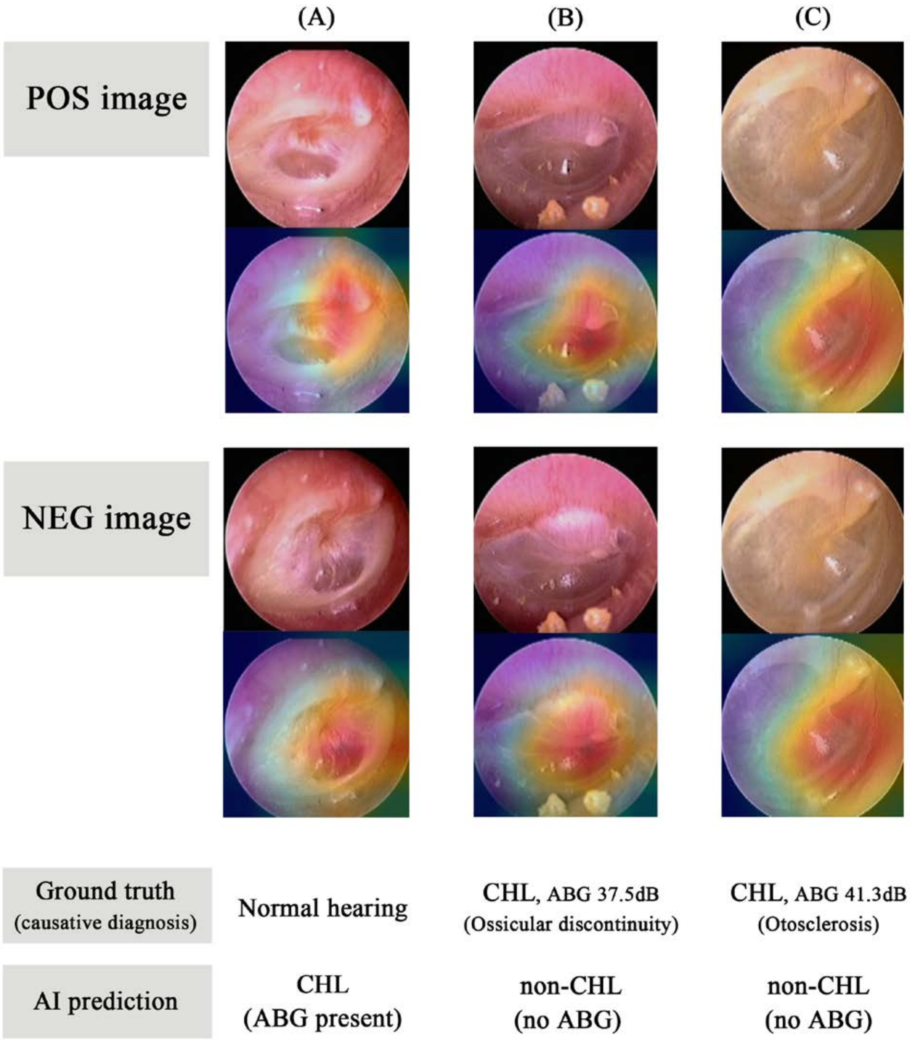


Fig. 11. Heat maps of the regions of interest of the deep learning algorithm in examples of wrong predictions. Misclassification of a non-CHL group as a CHL group was observed in a tympanic membrane with tympanosclerotic plaque and healed perforation (A). The regions of interest are also visualized in examples where diseased ears were incorrectly assigned to the non-CHL group (B–C). ABG indicates air-bone gap; CHL, conductive hearing loss; NEG image, most lateral image in negative pressure on pneumatic otoscopy; POS image, most medial image in positive pressure on pneumatic otoscopy.

Hearing loss in MEE is known to be due to a reduction of umbo velocity affecting mid-to-high frequency sounds, with the additional effect of an air-space reduction on low frequency sounds (Ravicz et al. 2004). In this study, the region of interest to the algorithm was the umbo area, especially at negative pressures (**Fig. 10B**). It is plausible to look at the umbo in NEG images, in the light of a previous report showing that the position of the umbo at negative pressure was significantly different between MEE and normal ear (Cho et al. 2009). Another difference in MEE was reported as the increased positive pressure causes the initial TM movement, but this unfortunately cannot be seen in VPO videos. This may provide a possible explanation for why areas of the malleus and pars flaccida somewhat distant from the umbo were observed during decision-making for POS images (**Fig. 10B**). If the examiner tended to increase the air pressure in performing VPO tests in MEE, those areas may have provided the deep learning algorithm with some clues.

In cases with malleus-incus fixation, the umbo movement was reported to significantly decrease compared to normal ears (Koike et al. 2005; Nakajima et al. 2005; Lee et al. 2011). As expected, the limited movement of the malleus could also be directly identified from the VPO images and was accurately predicted by both the deep learning model and the human experts (**Fig. 9**). In a case in which all the otologists incorrectly predicted no hearing loss, exploratory tympanotomy revealed a hypo-mobile malleus with absent incus. The algorithm correctly detected hearing loss in this case.

When the stapedial annular ligament is restrained, the mobility of the malleus can also be affected through changes of ossicular movement pattern and umbo velocity (Zhao et al. 2002; Nakajima et al. 2005; Kanzara & Virk 2017). In this study, the deep learning algorithm correctly predicted hearing loss in 86.7% of otosclerosis cases, while the otologists diagnosed only 11.1% of those cases on average (**Fig. 9**). The heat maps show that the model made predictions based on the broad umbo regions in both the POS and NEG images (**Fig. 10D**), which may be explained by the effect of stapes fixation on malleus mobility described above. In two cases, where the algorithm did not predict conductive hearing loss, AB gaps were 25 dB and 41.25 dB, which were not different from the correctly predicted cases (mABG 26.7 dB) ($p > 0.05$). Referring to the heat map, these wrong predictions might be caused by inadequate regions of interest (**Fig. 11C**).

The algorithm also gave a good performance in detecting the presence of conductive hearing loss caused by limitations of mobility such as in adhesive otitis media and benign tumor of the middle ear. However, cases with loose incus-stapes joint and congenital cholesteatoma eroding the incus-stapes joint were incorrectly predicted as normal hearing, presumably because of the lack of similar cases with hypermobility (**Fig. 11B**).

VPO is a practical test that can be performed in clinics by simply adding a VPO device to existing oto-endoscopic equipment. With two significant images at positive and negative pressures, the diagnostic accuracy of the deep learning algorithm was higher than that of otologists when predicting the conductive component of hearing loss due to middle ear diseases including MEE, ossicular fixation, adhesive otitis media, and middle ear mass. It can be useful in clinics without facilities for bone conduction hearing tests, as well as for pediatric patients who are not cooperative in audiometry or tuning fork tests. In patients complaining of hearing loss, the possible presence of conductive hearing loss may be assessed with VPO during the initial endoscopic evaluation before audiometry. Given the cost and unsatisfactory sensitivity of high-resolution CT in assessing ossicular chain mobility or otosclerosis, automated prediction of AB gap can be a useful supplementary tool especially for diagnosing ossicular fixations. As a new evaluation tool, possible clinical burdens of prediction errors of this application would be minimal.

Since this was an early pilot study in interpreting VPO findings using deep learning algorithms, many steps remain. First, the main limitation of the study was the unsatisfactory number of recorded VPO cases. To overcome the possibility of underfitting or overfitting, the image data extracted from repeated measurements was augmented by cropping, rotation, and flipping, and 10-fold cross-validation was adopted. In addition, wherever possible a variety of cases causing conductive hearing loss were included in the analysis. Nevertheless, there still remains a pitfall: it was difficult for the algorithm to learn rare cases that were not sufficiently included in the cases (e.g. a case of hypermobility due to ossicular chain discontinuity was misjudged as normal). Second, another limitation of VPO videos was that air pressure changes and tactile information could not be quantified or recorded as images. If the pressures used to cause the TM movements could be measured, then that would possibly improve the performance of the algorithm as well as the accuracy of human specialists. Third, there may be an issue of test standardization before the algorithm can be considered for general application. In this study, intertest variability was expected to be low as all the included tests were performed by a single senior otologist. We believe that a guideline for standard VPO testing method can be discussed and shared in academia, and resolve this issue in a short time. In addition, further research through prospective multi-center data collection is desirable for external validation, which can help overcome potential bias in modeling and show more robust results. Lastly, if multiple video frames could be put into the deep learning algorithm thanks to further increases in computing capacity, additional information regarding the temporal pattern of movement might improve its performance.

This pilot study showed that use of the deep learning algorithm to interpret VPO images could help differentiate conductive hearing loss from sensorineural hearing loss. We hope that, if it can be applied to future clinical practice, it will be particularly useful in primary clinics without facilities for bone conduction hearing tests as well as in poorly cooperative patients such as people with disabilities or pediatric patients.

ACKNOWLEDGMENTS

All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Y.-S.C. and H.B. The first draft of the manuscript was written by H.B. and C.J.P. S.J.O., M.J.C., and B.H.C. designed and performed experiment and provided critical comments. H.B. had drawn the figures. All authors commented on the previous versions of the manuscript. All authors read and approved the final manuscript.

Ethics approval: This study was approved by the Institutional Review Board of Samsung Medical Center and performed in accordance with the Declaration of Helsinki (IRB No.2019-10-171).

This research was supported by the Brain Convergence Research Program of the National Research Foundation (NRF) funded by the Korean government (MSIT; NRF-2021M3E5D2A01022399), Future Medicine 20*30 Project of the Samsung Medical Center (SMC1210791), and the Basic Science Research Program of the NRF funded by the Ministry of Education (NRF-2018R1D1A1B07048796).

The authors have no conflicts of interest to disclose.

Address for correspondence: Yang-Sun Cho, Department of Otorhinolaryngology-Head and Neck Surgery, Samsung Medical Center, Sungkyunkwan University School of Medicine, 81 Ilwon-Ro, Gangnam-Gu, Seoul, 06351, South Korea. E-mail: yscho@skku.edu; Baek Hwan Cho, Medical AI Research Center, Samsung Medical Center, Department of Medical Device Management and Research, SAIHST, Sungkyunkwan University School of Medicine, 81 Irwon-ro, Gangnam-gu, Seoul 06351, South Korea. E-mail: baekhwan.cho@samsung.com

Received March 05, 2021; accepted January 20, 2022; published online ahead of print March 29, 2022.

REFERENCES

Byun, H., Yu, S., Oh, J., Bae, J., Yoon, M. S., Lee, S. H., Chung, J. H., Kim, T. H. (2021). An assistive role of a machine learning network in diagnosis of middle ear diseases. *J Clin Med, 10*, 3198.

Cha, D., Pae, C., Seong, S. B., Choi, J. Y., Park, H. J. (2019). Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database. *EBioMedicine, 45*, 606–614.

Cho, Y. S., Cho, K., Park, C. J., Chung, M. J., Kim, J. H., Kim, K., Kim, Y. K., Kim, H. J., Ko, J. W., Cho, B. H., Chung, W. H. (2020). Automated measurement of hydrops ratio from MRI in patients with Ménière's disease using CNN-based segmentation. *Sci Rep, 10*, 7003.

Cho, Y. S., Lee, D. K., Lee, C. K., Ko, M. H., Lee, H. S. (2009). Video pneumatic otoscopy for the diagnosis of otitis media with effusion: a quantitative approach. *Eur Arch Otorhinolaryngol, 266*, 967–973.

Choi, K. J., Choi, J. E., Roh, H. C., Eun, J. S., Kim, J. M., Shin, Y. K., Kang, M. C., Chung, J. K., Lee, C., Lee, D., Kang, S. W., Cho, B. H., Kim, S. J. (2021). Deep learning models for screening of high myopia using optical coherence tomography. *Sci Rep, 11*, 21663.

Cireşan, D., Meier, U., Masci, J., Schmidhuber, J. (2012). Multi-column deep neural network for traffic sign classification. *Neural Netw, 32*, 333–338.

Harris, P. K., Hutchinson, K. M., Moravec, J. (2005). The use of tympanometry and pneumatic otoscopy for predicting middle ear disease. *Am J Audiol, 14*, 3–13.

He, K., Zhang, X., Ren S., et al. (2015). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, 770-778

Jin, H., Xie, L., Xiao, Z., et al. (2019). Classification for human balance capacity based on visual stimulation under a virtual reality environment. *Sensors (Basel)*, *19*, 2738

Jones, W. S., & Kaleida, P. H. (2003). How helpful is pneumatic otoscopy in improving diagnostic accuracy? *Pediatrics, 112*(3 Pt 1), 510–513.

Kanzara, T., & Virk, J. S. (2017). Diagnostic performance of high resolution computed tomography in otosclerosis. *World J Clin Cases, 5*, 286–291.

Khan, M. A., Kwon, S., Choo, J., Hong, S. M., Kang, S. H., Park, I. H., Kim, S. K., Hong, S. J. (2020). Automatic detection of tympanic membrane and middle ear infection from oto-endoscopic images via convolutional neural networks. *Neural Netw, 126*, 384–394.

Kim, Y., Lee, K. J., Sunwoo, L., Choi, D., Nam, C. M., Cho, J., Kim, J., Bae, Y. J., Yoo, R. E., Choi, B. S., Jung, C., Kim, J. H. (2019). Deep learning in diagnosis of maxillary sinusitis using conventional radiography. *Invest Radiol, 54*, 7–15.

King, E. F., & Couch, M. E. (2015). History, physical examination, and the preoperative evaluation. In P. W. Flint, B. H. Haughey, K. T. Robbins, et al. (Eds.), Cummings Otolaryngology Head and Neck Surgery (pp. 50). Elsevier Saunders.

Koike, T., Shinozaki, M., Murakami, S., et al. (2005). Effects of individual differences in size and mobility of the middle ear on hearing. *JSME International J Series C*, *48*, 521-528.

LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. *Nature, 521*, 436–444.

Lecun, Y., Bottou, L., Bengio, Y., et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*, 2278-2324.

Lee, J. K., Cho, Y. S., Ko, M. H., Lee, W. Y., Kim, H. J., Kim, E., Chung, W. H., Hong, S. H. (2011). Video pneumatic otoscopy for the diagnosis of conductive hearing loss with normal tympanic membranes. *Otolaryngol Head Neck Surg, 144*, 67–72.

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., Ledsam, J. R., Schmid, M. K., Balaskas, K., Topol, E. J., Bachmann, L. M., Keane, P. A., Denniston, A. K. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health, 1*, e271–e297.

Mains, B. T., & Toner, J. G. (1989). Pneumatic otoscopy: study of interobserver variability. *J Laryngol Otol, 103*, 1134–1135.

Nakajima, H. H., Ravicz, M. E., Merchant, S. N., Peake, W. T., Rosowski, J. J. (2005). Experimental ossicular fixations and the middle ear's response to sound: evidence for a flexible ossicular chain. *Hear Res, 204*, 60–77.

Park, C. J., Cho, Y. S., Chung, M. J., Kim, Y.-K., Kim, H.-J., Kim, K., Ko, J.-W., Chung, W. H., & Cho, B. H. (2021). A fully automated analytic system for measuring endolymphatic hydrops ratios in patients with ménière disease via magnetic resonance imaging: deep learning model development study. *J Med Internet Res*, *23*, e29678. https://doi.org/10.2196/29678

Ravicz, M. E., Rosowski, J. J., Merchant, S. N. (2004). Mechanisms of hearing loss resulting from middle-ear fluid. *Hear Res, 195*, 103–130.

Rosenfeld, R. M., Culpepper, L., Yawn, B., Mahoney, M. C; AAP, AAFP, AAO-HNS Subcommittee on Otitis Media with Effusion. (2004). Otitis media with effusion clinical practice guideline. *Am Fam Physician, 69*, 2776, 2778–2776, 2779.

Rosenfeld, R. M., Shin, J. J., Schwartz, S. R., Coggins, R., Gagnon, L., Hackell, J. M., Hoelting, D., Hunter, L. L., Kummer, A. W., Payne, S. C., Poe, D. S., Veling, M., Vila, P. M., Walsh, S. A., Corrigan, M. D. (2016). Clinical practice guideline: otitis media with effusion (Update). *Otolaryngol Head Neck Surg, 154*(1 Suppl), S1–S41.

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Netw, 61*, 85–117.

Schwartz, R. H. (1980). The pneumatic otoscope, a new instrument for the examination of the tympanic membrane–E. Siegle. 1984. *Int J Pediatr Otorhinolaryngol, 2*, 261–263.

Selvaraju, R. R., Cogswell, M., Das, A., et al. (2017). Grad-CAM: visual explanations from deep networks via gradient-based localization. In. The IEEE International Conference on Computer Vision (ICCV), 618-626.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR, abs/1409.1556*

Szegedy, C., Vanhoucke, V., Ioffe, S., et al. (2016). Rethinking the inception architecture for computer vision. In. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2818-2826.

Szegedy, C., Wei, L., Yangqing, J., et al. (2015). Going deeper with convolutions. In. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 1-9.

Zhang, Y., Zhou, D., Chen, S., et al. (2016). Single-image crowd counting via multi-column convolutional neural network. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 589-597).

Zhao, F., Wada, H., Koike, T., Ohyama, K., Kawase, T., Stephens, D. (2002). Middle ear dynamic characteristics in patients with otosclerosis. *Ear Hear, 23*, 150–158.