



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

의학박사 학위논문

딥러닝 모델의 자동 다단계 분류를
활용한 고막의 변화 예측

Automated Multi-class Classification for Prediction of
Tympanic Membrane Changes with
Deep Learning Models

울산대학교 대학원

의학과

최연주

딥러닝 모델의 자동 다단계 분류를
활용한 고막의 변화 예측

지도교수 안중호

이 논문을 의학박사 학위 논문으로 제출함.

2024년 2월

울산대학교 대학원

의학과

최연주

최연주의 의학박사 학위 논문을 인준함.

심사위원 정 종 우 (인)

심사위원 박 홍 주 (인)

심사위원 안 중 호 (인)

심사위원 강 우 석 (인)

심사위원 임 기 정 (인)

울 산 대 학 교 대 학 원

2024년 2월

영문 요약

Backgrounds and Objective

Evaluating the tympanic membrane (TM) using an otoendoscope is the first and most important step in various clinical fields. Unfortunately, most lesions of TM have more than one diagnostic name. Therefore, we built a database of otoendoscopic images with multiple diseases and investigated the impact of concurrent diseases on the classification performance of deep learning networks.

Study Design

This retrospective study investigated the impact of concurrent diseases in the tympanic membrane on diagnostic performance using multi-class classification. A customized architecture of EfficientNet-B4 was introduced to predict the primary class (otitis media with effusion (OME), chronic otitis media (COM), and 'None' without OME and COM) and secondary classes (attic cholesteatoma, myringitis, otomycosis, and ventilation tube).

Results

Deep-learning classifications accurately predicted the primary class with dice similarity coefficient (DSC) of 95.19%, while misidentification between COM and OME rarely occurred. Among the secondary classes, the diagnosis of attic cholesteatoma and myringitis achieved a DSC of 88.37% and

88.28%, respectively. Although concurrent diseases hampered the prediction performance, there was only a 0.44% probability of inaccurately predicting two or more secondary classes (29/6,630). The inference time per image was 2.594 ms on average.

Conclusion

The algorithm presented in this study demonstrated the ability to accurately predict TM lesions even in situations with multiple concurrent diseases. This finding is expected to contribute to clinical decision-making in the future, providing valuable assistance in managing cases involving complex medical conditions.

차 례

영문 요약.....	i
표, 그림 목차.....	iv
서론.....	1
연구대상 및 방법.....	5
결과.....	10
고찰.....	14
결론.....	20
참고 문헌.....	31
국문 요약.....	35

Table Contents

Table 1. Preliminary test results of DSC values from multiple deep-learning models for tympanic membrane changes. OME, otitis media with effusion; COM, chronic otitis media.....	21
Table 2. Prediction performance of combined model for primary and secondary classes. McNemar test was applied for the comparison with separate models, denoted with the subscript 'sep'. OME, otitis media with effusion; COM, chronic otitis media	22
Table 3. Comparison of prediction accuracy between combined and separate models according to the number of positives in the secondary classes.....	23
Table 4. Computational cost and inference time for application of deep-learning classification for tympanic membrane changes	24

Figure Contents

Figure 1. Classification of otoendoscopic images by primary and secondary classes with representative examples. OME, otitis media with effusion; COM, chronic otitis media..... 25

Figure 2. (a) Schematic diagram of deep learning network for multi-class classification of otoendoscopic images; Images for deep learning were pass through pre-processing procedures including thresholding, circular cropping, moving to center and image size reformatting. Input images pass through shared layer and task-specific layers combinely or separately depending on the types of models (b) Labeling examples; For a normal tympanic membrane (TM), the otoendoscopic image was labeled as 'None' for the primary class and 'False' for the secondary classes (attic cholesteatoma, myringitis, otomycosis and ventilation tube). When TM was diseased as one of the secondary classes without otitis media with effusion (OME) and chronic otitis media (COM), the primary class was given as 'None' for the otoendoscopic image. For example, when only otomycosis is identified, labelling is done with 'none' for primary class, 'true' for otomycosis, and 'false' for other secondary classes .. 26

Figure 3. Confusion matrix of combined model in 5-fold cross validation for the prediction of primary and secondary classes. GT, ground truth; OME, otitis media with effusion;

COM, chronic otitis media 28

Figure 4. Receiver operating characteristics (ROC) curves and AUC values for primary and secondary classes for combined model and separate model. Micro-average was applied to evaluate the overall predictability of deep learning model for the primary class. AUC, area under the ROC curve; OME, otitis media with effusion; COM, chronic otitis media..... 29

Figure 5. Grad-CAM visualization of representative examples for combined (upper row) and separate (lower row) models. The red area refers to the part of the model where the attention is strong. GT, ground truth; OME, otitis media with effusion. 30

INTRODUCTION

In the otologic field, evaluating the tympanic membrane (TM) and the middle ear via endoscopic evaluation is usually the first step for patients complaining of earache or other problem such as hearing loss, dizziness, or facial palsy(1). To evaluate otologic diseases such as acute/chronic otitis externa or acute/chronic otitis media, it is important to examine the state of the external auditory canal (EAC) and TM using common tools like the otoscope, which allows for simple observation and diagnosis. Apart from being a primary diagnostic step, an accurate otoscopic exam can also guide the correct course of treatment during the follow up period. Given how important it is to diagnose and evaluate accurately the state of disease during the follow up period, intensive training is required before being able to accurately diagnose the condition (1). Unfortunately, misdiagnosis in the clinical field is still fairly common.

One study reported that diagnostic accuracy varied among physicians, including otolaryngologists, pediatricians, and family medicine doctors(2). Another study reported that otolaryngologists diagnosed these otologic diseases with 73% accuracy while pediatricians and general practitioners had an accuracy rate of 50% and 64%, respectively(3). Therefore, even though there is a glaring need for trained otolaryngologists to make accurate diagnoses, the limited number of specialists makes it impossible(4). Therefore, there is a need to develop a modality that can accurately evaluate the status of EAC and TM to support the diagnostic system. Specifically, there is a need for an image-based

diagnostic algorithm based on otoscopic images.

In recent years, advances in image classification using deep learning networks have been proven to improve the diagnosis performance of middle ear diseases(5-12). Khan et al.(9) reported that classification accuracy of deep network reached 94.9% in the classification of normal, chronic otitis media (COM) with TM perforation, and otitis media with effusion (OME). Detection of tympanic perforation had an accuracy rate of 91%(10). The ensemble approach, which combines the outputs of multiple networks, enhanced predictability in the categorical classification of otoendoscopic images(11, 12). Deep learning prediction can help clinicians make more accurate decisions(13). Although previous studies showed the potential applicability of deep learning-based diagnosis, otoendoscopic images of multiple diseases that could hamper diagnostic accuracy were excluded from the prediction.

Convolutional neural networks (CNNs), a category of deep neural networks, have demonstrated significant effectiveness in domains like image recognition and classification. This resembles the reaction of a neuron in the visual cortex when exposed to a particular stimulus(14). These networks are specifically crafted to autonomously and flexibly acquire spatial hierarchies of features from input data. CNNs excel in handling tasks that deal with data organized in a grid-like fashion, such as images(15). They have achieved notable success in various applications like image classification, object detection, facial recognition, and other computer vision assignments(16-18). Their strength lies in the capacity to autonomously acquire hierarchical features, making them potent for discerning

meaningful patterns within intricate datasets(19).

EfficientNet architecture is a family of CNNs designed to achieve better performance with fewer parameters and computations compared to traditional CNN architectures. In 2019, Tan and Le introduced a neural network architecture called EfficientNet. The EfficientNet family consisted of various versions, including B0, B1, B2, B3, B4, B5, and B6, each characterized by distinct levels of model complexity and computational demands. EfficientNet achieved notable improvements in performance over ResNet while simultaneously reducing computational complexity. This architecture demonstrated high efficacy in tasks such as image classification and object recognition, showcasing a superior balance between accuracy and computational efficiency. EfficientNet-B4 is one of a specific variant of the EfficientNet architecture. Generally, larger variants like B4 tend to provide increased accuracy but necessitate more resources for both training and inference. The selection of a particular variant is contingent upon the available resources and the specific needs of the given task(20).

Pytorch is an open-source framework designed for deep learning, offering a dynamic and adaptable computational graph(21). This feature makes it highly suitable for research and development in the realm of artificial intelligence. Pytorch finds extensive application in tasks like deep learning, machine learning, and computer vision. Pytorch uses a dynamic computational graph, which makes it easier to work with variable-length sequences and dynamic inputs. Pytorch introduces a multi-dimensional array called a tensor, which is similar to NumPy arrays. Tensors in Pytorch can be used

for a wide range of mathematical operations and are fundamental to building neural networks.

Pytorch includes a built-in automatic differentiation library called Autograd. This enables automatic computation of gradients, which is crucial for training neural networks using gradient-based optimization algorithms(22-24). Pytorch and TensorFlow are commonly compared as two of the most popular open-source deep learning frameworks. The debate over which framework is superior has been ongoing, with TensorFlow often recognized as an industry-centric framework, while Pytorch is renowned for its prominence in research and academia(25).

In this study, we built a database of otoendoscopic images containing multiple diseases using the most efficient network system to investigate the impact of concurrent diseases on the classification performance of deep learning networks.

MATERIALS AND METHODS

Data description

Otoendoscopic images of TM were collected from patients who visited the otologic clinic in Asan Medical Center from Jan 2018 to Dec 2020. In clinical practice, the otoendoscopic video sequence was taken for diagnostic examination and an image frame visualizing the whole TM was stored in the hospital system without patient-identifiable information. Otoendoscopic images enrolled based on the date of visit were completely anonymized before being provided by the hospital system. The collected images were classified into one primary class and four secondary classes according to their diagnostic classification. The categories of each image were blindly annotated by two otologists with 26 and 5 years of experience, respectively. A total of 6,630 otoendoscopic images labeled identically by two annotators were included in this study. Figure 1 demonstrated the classification of otoendoscopic images by primary and secondary classes with representative examples. The primary class was annotated as one of otitis media with effusion (OME, 1,630 images), chronic otitis media (COM, 1,534 images), and 'None' (3,466 images) – meaning the absence of OME and COM. OME refers to effusions in the middle ear cavity, which manifest in the air-fluid level or as an amber-like color change of TMs. COM refers to a perforated TM. Binary labels were given for the secondary classes of attic cholesteatoma (893 images), myringitis (1,083 images), otomycosis (181 images), and ventilation tube (1,676 images) (Fig. 1). Attic cholesteatoma refers to any sign of retraction

pocket in attic or visible attic destruction. Myringitis is defined as any inflammation of the tympanic membrane, including acute otitis media (AOM). Otomycosis refers to a fibrinous accumulation of debris or visible pores of fungus in the external auditory canal. Ventilation tube refers to an inserted tube across the TM. For example, when a TM was normal, the primary class was 'None' and the secondary classes were 'False' for attic cholesteatoma, myringitis, otomycosis, and ventilation tube (Fig. 2b). An otoendoscopic image with only otomycosis was assigned 'None' for the primary class, 'True' for otomycosis, and 'False' for the other secondary classes. For 3,508 images, one or more secondary classes were positive. The present study is in compliance with the Declaration of Helsinki and research approval was granted from the Institutional Review Board of the Asan Medical Center with a waiver of research consent (IRB no. 2021-0837).

Deep learning network

The architecture of EfficientNet-B4(20) was customized to have shared and task-specific layers for the multi-task learning (Fig. 2a). Initially, we conducted several preliminary experiments using relatively big classification models such as EfficientNet-B7, InceptionResNet-V2, Inception-V3, DenseNet201. Among the models, EfficientNet-B7 exhibited the best performance (Table1). When evaluated on B4, which belongs to the same family, the performance difference compared to B7 was negligible. Considering the similar performance and a smaller model size, we have decided to use EfficientNet-B4 as the final architecture. The task-specific layers consisted of five shallow classifiers

corresponding to the primary class and four secondary classes ('combined model'). Parameters between the classifiers were not shared.

As an input to deep networks, RGB images reformatted into $256 \times 256 \times 3$ with circular cropping were used (Fig. 2a). Data augmentation was performed by randomly applying rotation (-90° to 90°), translation shift (0–20% of image size in horizontal and vertical axes), zoom (0–20%), horizontal flip, brightness change (0–20%) and downscale (0–50%). The pre-trained weight from ImageNet was applied for transfer learning. Categorical cross-entropy loss was adopted to train the models for multi-class classification, which is defined as,

$$L_{CCE} = -\frac{1}{N} \sum_i^N \sum_c^M t_{i,c} \log(p_{i,c})$$

where N is the number of training samples, M is the number of classes, $t_{(i,c)}$ is the ground truth, and $p_{(i,c)}$ is the output probability. The final output was determined as the primary rank of the softmax value.

Training setup and Evaluation metrics

The deep learning model implemented using Pytorch which was known to the most commonly used library for deep learning networks was trained. Pytorch is an open-source library designed for deep learning, offering a dynamic and adaptable computational graph. It was adapted in this study due to its well-known stability and efficacy, especially in academic fields. This training was performed on

a workstation with AMD Ryzen 7 5800X CPU 3.8 GHz, 128 GB RAM, and two NVIDIA Geforce RTX 3090 Ti GPUs. The model training was conducted for 200 epochs at maximum with a mini-batch size of 32. For training, an Adam optimizer was applied with $\beta_1 = 0.9$ and $\beta_2 = 0.9999$. The learning rate was initially set as 10^{-3} and was reduced by half with a saturation criteria of 50 epochs. The evaluation metrics for each label were precision, sensitivity (recall), specificity, and dice similarity coefficient (DSC), which were defined as $\text{precision} = \text{TP} / (\text{TP} + \text{FP})$, $\text{sensitivity} = \text{TP} / (\text{TP} + \text{FN})$, $\text{specificity} = \text{TN} / (\text{FP} + \text{TN})$ and $\text{DSC} = 2 \times \text{precision} \times \text{recall} / (\text{precision} + \text{recall})$, where TP is true positive, FP is false positive, and FN is false negative. The per-class accuracy was calculated by dividing the sum of TPs and TNs with the total number of images in a fold. For 5-fold cross validation, the dataset was divided so that each fold contained an equal number of images ($n = 1,326$). The fold proportion of training, validation, and test sets was fixed at 3:1:1 and their compositions were changed under cyclic permutation.

Combine prediction and Separate prediction for single class as reference

To evaluate the performance of multi-class classification, the deep learning models for the prediction of each class were separately trained ('separate model') as well as combinely trained ('combined model'). In separate model setting, only one classifier for the target class remained in the task-specific layers. In combined model setting, whole classifiers including primary classes and all secondary classes pass through task-specific layers at the same time (Fig.2a).

Visual setup

In the realm of medical imaging, it is imperative to verify whether decisions are influenced by accurate regions within the image. The utilization of class activation mapping (CAM) techniques, particularly Grad-CAM in this study, proves beneficial. Grad-CAM, an extended version of CAM, utilizes class-specific gradient information from the final convolutional layer to generate a coarse localization map, highlighting significant areas in the image comparing both in combined and separate models.

Statistical analysis

Categorical variables are presented as numbers and percentages. The McNemar test was applied to compare DSC values between combined and separate models. Statistical analyses were performed using R package (version 4.3.1).

RESULTS

Classification performance of combined model

In the prediction of the primary class, the overall DSC was 95.19%, with COM achieving the highest DSC of 96.09%. Followed by COM, 'none' accounted for 95.68% and OME accounted for 93.80% (Table 2). Figure 3 demonstrated the confusion matrix which displayed the probabilities associated with predicting each class compared to the ground truth. Misidentification between COM and OME rarely occurred (7 images), and most of the prediction errors appeared as false positives and false negatives in the 'None' class (Figure 3). Among the secondary classes, the ventilation tube was most accurately diagnosed (DSC = 98.89%), followed by attic cholesteatoma and myringitis with DSCs of 88% or higher; 88.37% and 88.28%, respectively (Table 2). Otomycosis, which trained with fewer positive cases, had lower predictive accuracy, as DSC value 72.38%, than other classes. In confusion matrix, true positive and false negative were highest in ventilation tube (Figure 3). Figure 4 demonstrated the receiver operating characteristics (ROC) curves and area under curve (AUC) values for primary classes. The AUC values for the primary and secondary classes were ≥ 0.9925 in both combined and separate models. When examining the AUC values, we observed slight differences across various categories, but it was evident that the combined model and the separate model exhibited almost similar performance.

Impact of concurrent diseases

Table 3 demonstrated the comparison of prediction accuracy between combined and separate models according to the number of positives in the secondary classes. With a greater number of positive secondary classes, the probability of accurate prediction for all classes gradually decreased from 92.57% to 14.29% in combined model. Separate model also demonstrated in similar decrease, from 91.51% to 57.14%. Overall, it appeared that the combined model outperformed the separate model in terms of prediction accuracy, slightly. However, in cases where there were three secondary classes, the separate model seemed somewhat higher. This might be attributed to the limited number of images with three or more secondary classes (n=7) during the training and validation, resulting in such outcomes.

When the number of positives in the secondary classes ≥ 2 , the proportion of images with at least one false prediction was over 40%. Nonetheless, the combined model had only a 0.32% probability of inaccurately predicting two or more secondary classes (21/6,630), whereas separate model demonstrated 0.44% probability of inaccurately predicting two or more secondary classes (29/6,630).

Comparison with separate models with combine models

Table 2 demonstrated the prediction performance of combined model for primary and secondary classes. Compared to the separate models, the combined model slightly improved the predictability of the deep learning models. In primary class, difference of DSC values between combined model

and separate model was 0.29% ($p=0.360$), whereas in secondary class, attic cholesteatoma accounted for 0.62% ($p=0.663$), otomycosis accounted for 4.12% ($p=0.030$), and ventilation tube accounted for 0.21% ($p=0.879$). Only for myringitis demonstrated slightly higher predictability in separate model than combined model, albeit not in a statistically significant way ($p=0.404$). The combined model provided correct diagnoses for all classes in 88.1% of the images (5,841/6,630), which was 0.98% higher than the separate models (Table 3, $p = 0.009$).

Table 4 demonstrated computational cost and inference time for application of deep-learning classification for tympanic membrane changes. Overall, when comparing the combined model and the separate model, the combined model demonstrated a slightly higher prediction accuracy, although the difference was not very significant. However, upon comparing the training and inference times, it was evident that the combined model required considerably less time. Specifically, the interpretation time per image was approximately 2.594 ms for the combined model and 12.893 ms for the separate model, indicating about a 5-fold difference in processing time. This observation suggested that when making deep learning predictions, the combined model not only maintained the overall quality of cochlear evaluations but also provided more efficient prediction times, highlighting its advantages.

Grad-CAM visualization

Figure 5 demonstrated Grad-CAM visualization of representative examples for combined and

separate models. The red area in the figure highlighted the part of the model where the attention is strong. It presented that the combined model made its prediction by comprehensively observing the entire tympanic membrane than separate model.

DISCUSSION

In real practice, it is not easy to examine the status of TM and reach an accurate diagnosis of the middle ear in crying children or non-cooperative patients in a short time. Additionally, in situations where a skilled otologist is not available, there is likely to be an incorrect diagnosis, which leads to malpractice. Although diagnostic rates have dramatically increased since the otoendoscopy was introduced, diagnostic accuracy still differs among physicians(2), while even otolaryngologists can sometimes produce inaccurate diagnoses(3). Therefore, many researchers have worked on various deep learning models for the effective diagnosis of middle ear diseases.

Previous studies have shown that deep-learning classification can accurately predict the diagnosis of otitis media, up to almost 98.26% of the time(8-12). Alhudhaif et al.(8) analyzed a total 956 otoendoscopic images. A newly designed computer-aided decision support model leveraging CNN has been implemented. To enhance the overall effectiveness of the proposed model, a fusion of channel and spatial attention model (CBAM), residual blocks, and the hypercolumn technique has been integrated. They divided into five classes consisting of otitis externa, ear ventilation tube, foreign bodies in the ear, pseudo-membranes, and tympanosclerosis with an overall accuracy rate of 98.26%. Khan et al.(9) analyzed 2,484 otoendoscopic images. This paper presented a novel application of state-of-the-art CNN models, DenseNet. One layer collected and combined all the outputs from preceding layers by concatenating them along the depth dimension. This design results

in a superior accuracy on the ImageNet dataset. They divided into three classes consisting of normal, perforation, and middle ear effusion with an overall accuracy rate of 95%. Lee et al.(10) analyzed 1338 otoendoscopic images. The Python programming language was employed to construct the CNN model designed for identifying the status of the TM and the presence of perforation. This model was comprised of two convolutional layers, two max pooling layers, and two fully connected layers. In the test dataset, normal TM and perforated TM were labelled, and the CNN model achieved an accuracy of 97.9% in detecting the status of the TM and 91.0% in identifying the presence of perforation. Cha et al.(11) analyzed 10,544 otoendoscopic images. They initially utilized nine public convolution-based deep neural networks; SqueezeNet, Alexnet, ResNet18, MobileNet-V2, GoogLeNet, Resnet50, Resnet101, Inception-V3, InceptionResnet-V2, and finally two best-performing model, Inception-V3 and Resnet101, were adopted. They categorized characteristics of the TM and external auditory canal into six groups corresponding to various ear conditions, encompassing a broad range of ear diseases such as normal, attic retraction, tympanic perforation, otitis externa \pm myringitis, and tumor. Overall, the system was able to achieve an average of 93% diagnostic accuracy. They implied that it was noteworthy that as the dataset size increased, the performance difference between training models diminished. Therefore, choosing a model that struck a balance between efficacy and accuracy was emphasized as crucial. Zeng et al.(12) analyzed 20,542 otoendoscopic images. ResNet19 (including ResNet50 and ResNet101), DensNet-BC20 and BC21 (comprising DensNet-BC121, DensNet-BC161, and DensNet-BC169), InceptionV322 and V423,

Inception-ResNet-V223, and MobileNet-V224 and V325 were instantiated and performance data were compared across different releases. In this procedure, they opted to replace the fully connected layers in each model with global average pooling, resulting in the generation of eight output nodes employing a softmax activation function. They divided into eight classes consisting of normal, cholesteatoma of the middle ear, COM, external auditory canal bleeding, impacted cerumen, otomycosis external, secretory otitis media, and TM calcification with an overall accuracy rate of 95.59%. Study from Wu et al.(5) performed the deep learning classification of pediatric otitis media. They utilized two most widely used CNN architectures named Xception and MobileNet-V2 for deep learning networks. They divided into three classes including, AOM, OME, and normal. Out of the eligible otoscopic images, 10,703 were employed for the training set, and 1,500 images were allocated to the testing set. The Xception model and the MobileNet-V2 model exhibited comparable overall accuracies, achieving 97.45% and 95.72%, respectively. However, these studies were limited by the fact that only one diagnostic label per image was assigned for deep-learning prediction, despite the fact that multiple diseases can be detected simultaneously in real practice. For example, some patients with attic cholesteatoma can have ventilation tube for prevention of TM retraction, while we can also diagnose myringitis in a patient who has tympanic perforation with or without tympanosclerosis.

In this study, we proposed a deep-learning method that can predict the diagnosis of TM changes for two non-coexisting diseases (OME and COM) and four concurrently detectable categories (attic

cholesteatoma, myringitis, otomycosis and ventilation tube) with a single network. Our deep-learning classification demonstrated high predictive performance using a database including TMs with up to 4 diseases at the same time. The DSC value of the primary class was greater than 95%, with COM achieving the highest value. In terms of secondary classes, the ventilation tube was rarely misidentified (DSC = 98.89%). Therefore, the multi-class classification for TM changes may have potential for higher clinical applicability than previous approaches in which all images were single labeled.

The combined model for predicting multiple classes at the same time produced better outcomes and required less inference time than the separate models that required a per-class training. The combined model also finished the prediction in 1/5 of the training and inference time required for separate models (Table 4). In our research, we were able to perform analysis and predictions at a rate of over 100 images per second using the performance of a CPU i9-10920X, RAM 128GB, and GPU RTX2080TI. Even the analyzed time in this algorithm was fast enough to utilize in real world, the most direct approach to reducing computational workload is to utilize superior hardware and optimize the model. These advantages of deep-learning prediction can help improve the overall diagnostic quality for TM changes. Due to their high predictability, the deep learning models can also support clinical decision-making for inexperienced clinicians and be utilized as a training tool for medical staff. The reduced analysis time of the deep learning models can also make real-time application more feasible. In the same regard, deep learning prediction can help with more accurate diagnoses

beyond the constraints of time and space through tele-medicine. Finally, their high reproducibility can enhance the reliability and objectivity of the analysis tool for diagnosis.

However, there were still some limitations on this study. First, even though large amount of samples were collected for analysis, the deep learning dataset was collected from a single center. As the data were collected from a single center, imaging was performed using endoscopic equipment from a specific company rather than various companies. Consequently, the collected data maintain a consistent resolution and image quality. However, it should be noted that the dataset may not reflect a diverse range of shooting qualities as it is limited to the specifications and characteristics of the equipment from particular company. Second, a small sample size of otomycosis resulted in fewer training opportunities, thus impairing its predictability. Third, as the number of positives in the secondary classes increases, the number of the secondary classes correctly predicted decreased, even in multi-class classification. An extended dataset with diverse disease patterns can be used to validate the generality and robustness of our classification and improve the prediction performance of TM changes. In the same vein, when applied to otoendoscopic video sequences(26), it can help overcome the bias of still image-based prediction. Cerumen, which was not included in this study, may limit the information on TMs required for diagnosis. As part of the pre-diagnosis evaluation process, quantifying the amount of cerumen using deep-learning segmentation would be helpful to determine whether cleaning of external acoustic meatus is necessary for accurate diagnosis. Ultimately, it is necessary to develop diagnostic tools that anyone can use in the EAC to easily diagnose otologic

diseases.

CONCLUSION

In the present study, we developed a multi-class classification method for predicting TM changes using deep-learning. The deep-learning algorithm accurately diagnosed the TM changes on otoendoscopic images, even for multiple concurrent diseases. Using the combined model, the inference time per image was reduced to 2.594 ms (more than 380 images can be processed per second), which indicates that deep-learning prediction can be applicable in real-time. Therefore, deep-learning classification can support clinical decision-making by accurately and reproducibly predicting tympanic membrane changes in real time, even in the presence of multiple concurrent diseases.

Table 1. Preliminary test results of DSC values from multiple deep-learning models for tympanic membrane changes. OME, otitis media with effusion; COM, chronic otitis media.

	DSC value (%)				
	Normal	OME	Myringitis	COM	Total
EfficientNet-B7	95.36 ± 1.81	93.59 ± 2.69	96.19 ± 2.75	96.12 ± 3.33	95.31 ± 0.55
InceptionResNet-V2	93.99 ± 3.30	91.26 ± 4.59	96.09 ± 1.89	96.88 ± 1.44	94.55 ± 1.76
Inception-V3	94.37 ± 4.05	90.59 ± 5.30	92.12 ± 2.34	93.51 ± 1.13	92.65 ± 2.73
DenseNet201	95.90 ± 1.53	92.09 ± 2.84	94.06 ± 1.86	96.77 ± 0.62	94.70 ± 1.62

Table 2. Prediction performance of combined model for primary and secondary classes. McNemar test was applied for the comparison with separate models, denoted with the subscript 'sep'. OME, otitis media with effusion; COM, chronic otitis media.

	DSC	Accuracy	Sensitivity	Precision	Specificity	DSC_{sep}	DSC - DSC_{sep}	p-value
Primary class (P)	95.19%	95.32%	95.38%	95.32%	94.65%	94.90%	0.29%	0.360
None	95.68%	-	96.91%	94.49%	93.58%	95.49%	0.19%	-
OME	93.80%	-	91.90%	95.78%	96.44%	93.76%	0.04%	-
COM	96.09%	-	95.37%	96.82%	95.31%	95.46%	0.63%	-
Attic cholesteatoma (S1)	88.37%	96.97%	85.54%	91.39%	98.74%	87.75%	0.62%	0.663
Myringitis (S2)	88.28%	96.21%	87.26%	89.32%	97.96%	88.58%	-0.30%	0.404
Otomycosis (S3)	72.38%	98.69%	62.98%	85.07%	99.69%	68.26%	4.12%	0.030
Ventilation tube (S4)	98.89%	99.44%	98.57%	99.22%	99.74%	98.68%	0.21%	0.879

DSC(Dice similarity coefficient) = $2 \times \text{precision} \times \text{recall} / (\text{precision} + \text{recall})$; Accuracy = $TP / (TP + FP)$; Sensitivity(recall) = $TP / (TP + FN)$; Specificity = TN

$/ (FP + TN)$; TP, true positive; FP, false positive; TN, true negative; FN, false negative

Table 3. Comparison of prediction accuracy between combined and separate models according to the number of positives in the secondary classes.

	Number of positives in the secondary classes	Primary class correct					Primary class incorrect				
		Number of the secondary classes incorrectly predicted					Number of the secondary classes incorrectly predicted				
		4	3	2	1	0	4	3	2	1	0
Combined model	0 (n = 3,122)	-	-	-	127 (4.07%)	2,890 (92.57%)	-	-	-	14 (0.45%)	91 (2.91%)
	1 (n = 3,190)	-	1 (0.03%)	8 (0.25%)	222 (6.96%)	2,777 (87.05%)	-	-	3 (0.09%)	33 (1.03%)	146 (4.58%)
	2 (n = 311)	-	1 (0.32%)	10 (3.22%)	104 (33.44%)	173 (55.63%)	-	1 (0.32%)	2 (0.64%)	12 (3.86%)	8 (2.57%)
	3 (n = 7)	-	-	3 (42.86%)	3 (42.86%)	1 (14.29%)	-	-	-	-	-
	Sum (n = 6,630)	-	2 (0.03%)	21 (0.32%)	456 (6.88%)	5,841 (88.10%)	-	1 (0.02%)	5 (0.08%)	59 (0.89%)	245 (3.70%)
Separate model	0 (n = 3,122)	-	-	4 (0.13%)	122 (3.91%)	2,857 (91.51%)	-	-	-	16 (0.51%)	123 (3.94%)
	1 (n = 3,190)	-	-	13 (0.41%)	269 (8.43%)	2,743 (85.99%)	-	-	1 (0.03%)	16 (0.50%)	148 (4.64%)
	2 (n = 311)	-	-	10 (3.22%)	106 (34.08%)	172 (55.31%)	-	1 (0.32%)	-	14 (4.50%)	8 (2.57%)
	3 (n = 7)	-	-	2 (28.57%)	1 (14.29%)	4 (57.14%)	-	-	-	-	-
	Sum (n = 6,630)	-	-	29 (0.44%)	498 (7.51%)	5,776 (87.12%)	-	1 (0.02%)	1 (0.02%)	46 (0.69%)	279 (4.21%)

Table 4. Computational cost and inference time for application of deep-learning classification for tympanic membrane changes.

Model		Number of parameters (million)	Training time (s)	Training time per epoch (s)	Inference time per image (ms)
Combined		17.57	10,166	50.83	2.594
Separated	Primary class	17.55	8,654	43.27	2.616
	Attic cholesteatoma	17.55	11,365.2	56.83	2.570
	Myringitis	17.55	11,147.2	55.74	2.558
	Otomycosis	17.55	11,272	56.36	2.572
	Ventilation tube	17.55	7,087.8	35.44	2.577
	Sum		49,526.2	247.64	12.893

Figure 1. Classification of otoendoscopic images by primary and secondary classes with representative examples. OME, otitis media with effusion; COM, chronic otitis media.








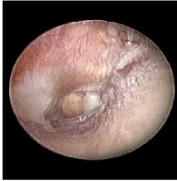










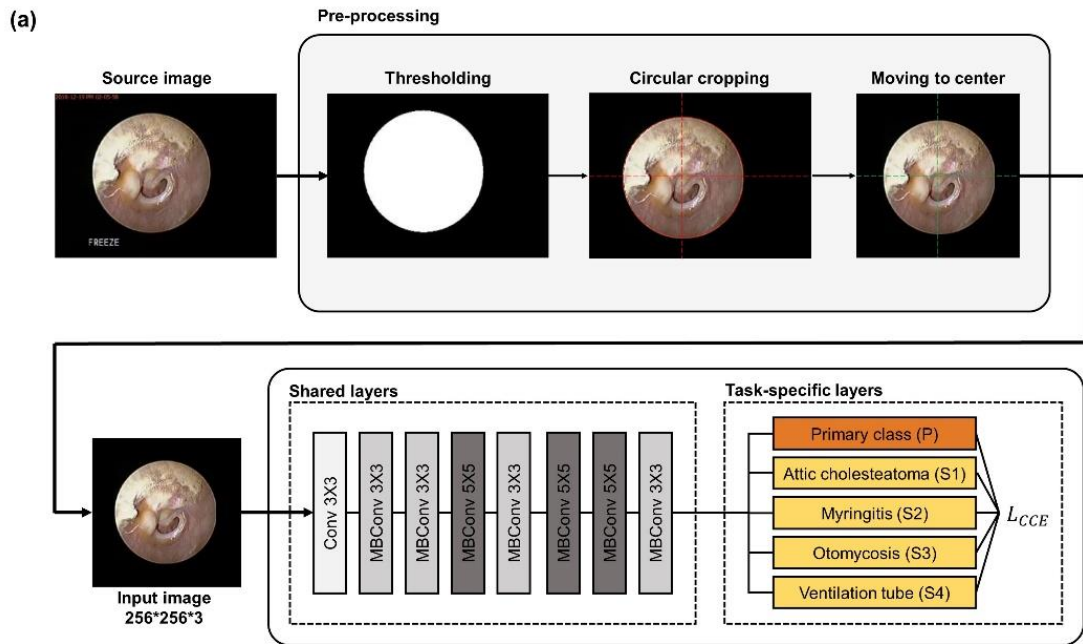
Primary class \ Secondary class	None n = 3,466	OME n = 1,630	COM n = 1,534
No secondary class n = 3,122	 n = 692	 n = 1,433	 n = 997
Attic cholesteatoma n = 744	 n = 610	 n = 95	 n = 39
Myringitis n = 842	 n = 439	 n = 17	 n = 386
Otomycosis n = 58	 n = 18	 n = 3	 n = 37
Ventilation tube n = 1,546	 n = 1,468	 n = 74	 n = 4
Multiple secondary classes n = 318	 n = 239	 n = 8	 n = 71

Figure 2. (a) Schematic diagram of deep learning network for multi-class classification of otoendoscopic images; Images for deep learning were pass through pre-processing procedures including thresholding, circular cropping, moving to center and image size reformatting. Input images pass through shared layer and task-specific layers combinely or separately depending on the types of models (b) Labeling examples; For a normal tympanic membrane (TM), the otoendoscopic image was labeled as 'None' for the primary class and 'False' for the secondary classes (attic cholesteatoma, myringitis, otomycosis and ventilation tube). When TM was diseased as one of the secondary classes without otitis media with effusion (OME) and chronic otitis media (COM), the primary class was given as 'None' for the otoendoscopic image. For example, when only otomycosis is identified, labelling is done with ‘none’ for primary class, ‘true’ for otomycosis, and ‘false’ for other secondary classes.



(b)

	Normal	Otitis media with effusion	OME	OME & Myringitis	COM, Myringitis & Ventilation tube
Primary class (P)	None	None	OME	OME	COM
Attic cholesteatoma (S1)	False	False	False	False	False
Myringitis (S2)	False	False	False	True	True
Otitis media with effusion (S3)	False	True	False	False	False
Ventilation tube (S4)	False	False	False	False	True

Figure 3. Confusion matrix of combined model in 5-fold cross validation for the prediction of primary and secondary classes. GT, ground truth; OME, otitis media with effusion; COM, chronic otitis media.

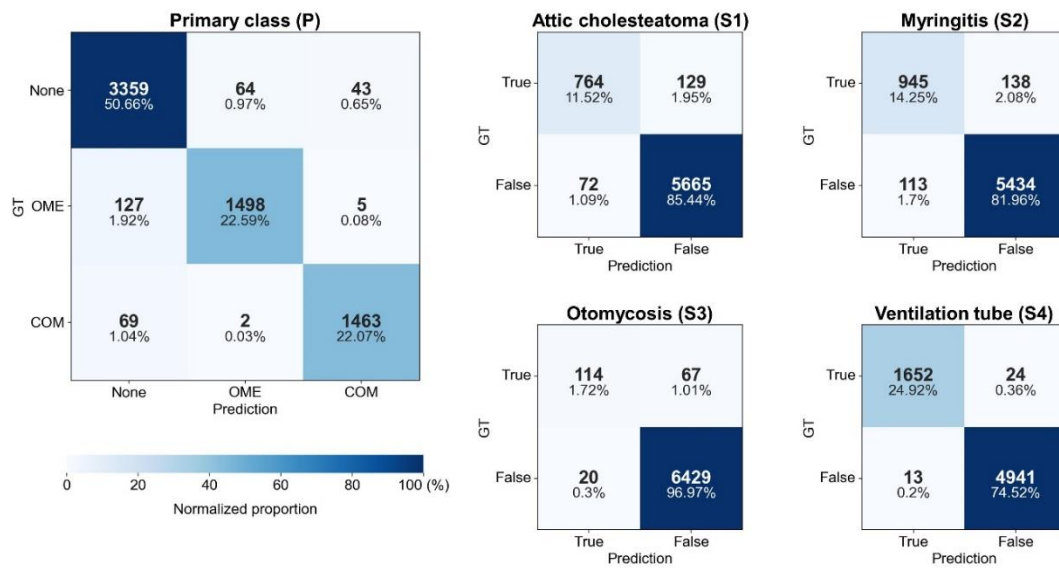


Figure 4. Receiver operating characteristics (ROC) curves and AUC values for primary and secondary classes for combined model and separate model. Micro-average was applied to evaluate the overall predictability of deep learning model for the primary class. AUC, area under the ROC curve; OME, otitis media with effusion; COM, chronic otitis media.

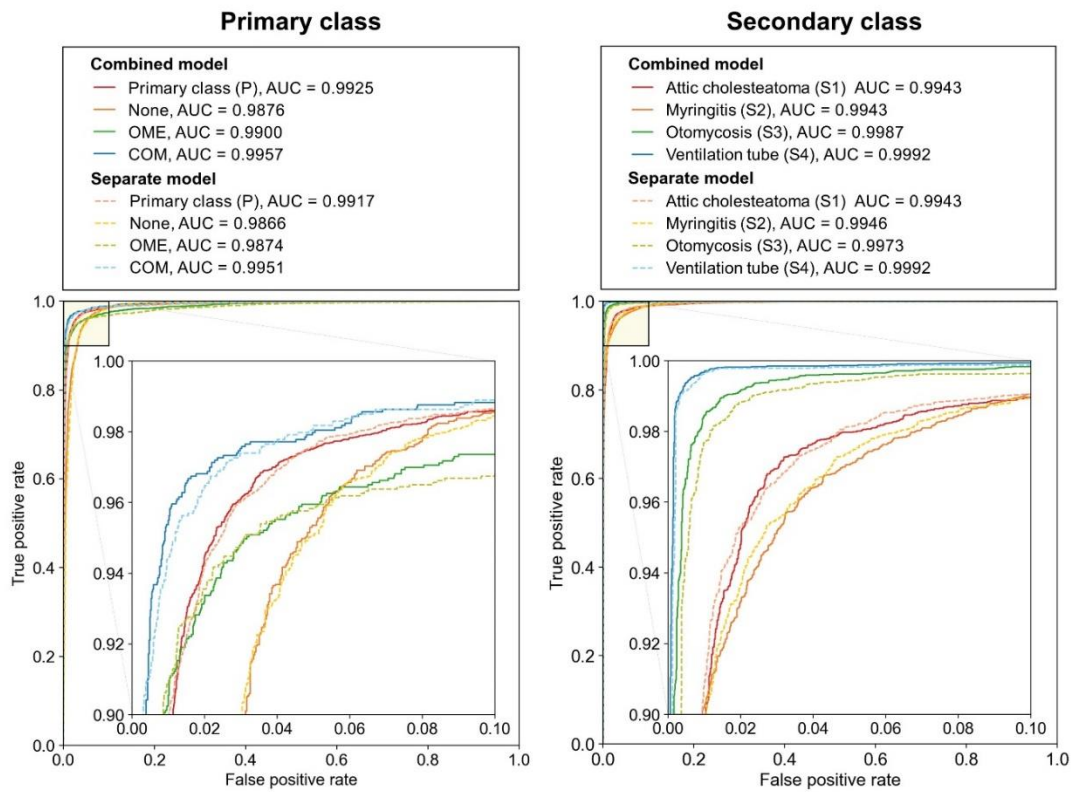
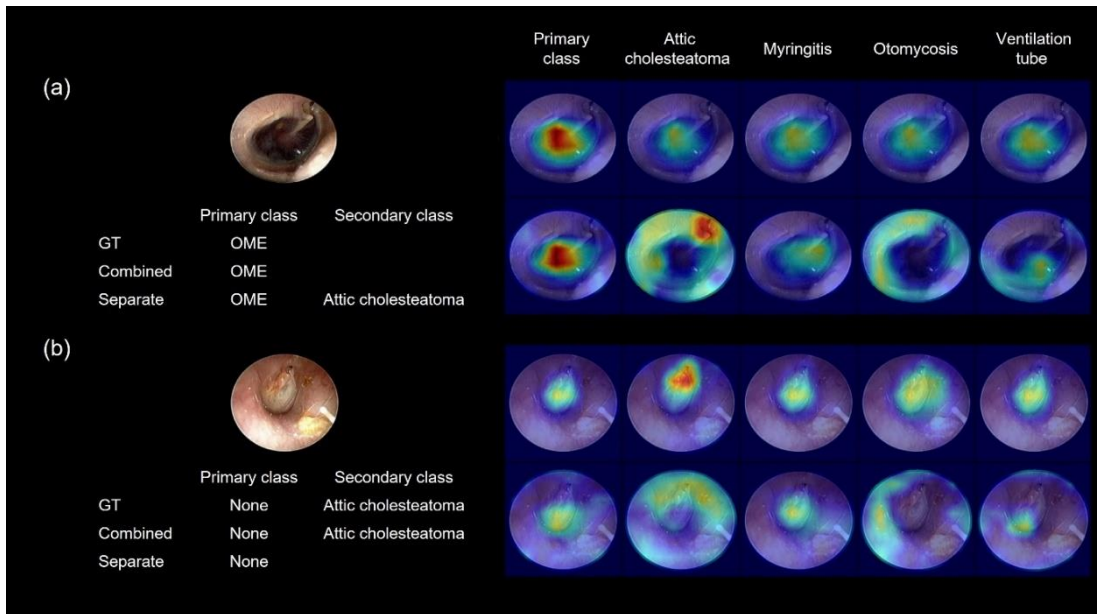


Figure 5. Grad-CAM visualization of representative examples for combined (upper row) and separate (lower row) models. The red area refers to the part of the model where the attention is strong.

GT, ground truth; OME, otitis media with effusion.



REFERENCES

1. Davies J, Djelic L, Campisi P, Forte V, Chiodo A. Otoscopy simulation training in a classroom setting: a novel approach to teaching otoscopy to medical students. *Laryngoscope*. 2014;124(11):2594-7.
2. Oyewumi M, Brandt MG, Carrillo B, Atkinson A, Iglar K, Forte V, Campisi P. Objective Evaluation of Otoscopy Skills Among Family and Community Medicine, Pediatric, and Otolaryngology Residents. *J Surg Educ*. 2016;73(1):129-35.
3. Pichichero ME, Poole MD. Assessing diagnostic accuracy and tympanocentesis skills in the management of otitis media. *Arch Pediat Adol Med*. 2001;155(10):1137-42.
4. Monasta L, Ronfani L, Marchetti F, Montico M, Brumatti LV, Bavcar A, et al. Burden of Disease Caused by Otitis Media: Systematic Review and Global Estimates. *Plos One*. 2012;7(4):e36226.
5. Wu ZB, Lin ZQ, Li L, Pan HG, Chen GW, Fu YQ, Qiu QH. Deep Learning for Classification of Pediatric Otitis Media. *Laryngoscope*. 2021;131(7):E2344-E51.
6. Cömert Z. Original Fusing fine-tuned deep features for recognizing different tympanic membranes. *Biocybern Biomed Eng*. 2020;40(1):40-51.
7. Sundgaard JV, Harte J, Bray P, Laugesen S, Kamide Y, Tanaka C, et al. Deep metric learning for otitis media classification. *Med Image Anal*. 2021;71.

8. Alhudaif A, Cömert Z, Polat K. Otitis media detection using tympanic membrane images with a novel multi-class machine learning algorithm. *Peerj Comput Sci.* 2021;23(7):e405.
9. Khan MA, Kwon S, Choo J, Hong SM, Kang SH, Park IH, et al. Automatic detection of tympanic membrane and middle ear infection from oto-endoscopic images via convolutional neural networks. *Neural Networks.* 2020;126:384-94.
10. Lee JY, Choi SH, Chung JW. Automated Classification of the Tympanic Membrane Using a Convolutional Neural Network. *Appl Sci-Basel.* 2019;9(9):1827.
11. Cha D, Pae C, Seong SB, Choi JY, Park HJ. Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database. *Ebiomedicine.* 2019;45:606-14.
12. Zeng XY, Jiang ZF, Luo W, Li HG, Li HY, Li G, et al. Efficient and accurate identification of ear diseases using an ensemble deep learning model. *Sci Rep-Uk.* 2021;11(1):10839.
13. Byun H, Yu S, Oh J, Bae J, Yoon MS, Lee SH, et al. An Assistive Role of a Machine Learning Network in Diagnosis of Middle Ear Diseases. *J Clin Med.* 2021;10(15):3198.
14. Fukushima K. Self-organizing neural network models for visual pattern recognition. *Acta Neurochir Suppl (Wien).* 1987;41:51-67.
15. Valueva MV, Nagornov NN, Lyakhov PA, Valuev GV, Chervyakov NI. Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. *Math Comput Simulat.* 2020;177:232-43.

16. Lawrence S, Giles CL, Tsoi AC, Back AD. Face recognition: a convolutional neural-network approach. *IEEE Trans Neural Netw.* 1997;8(1):98-113.
17. Le Callet P, Viard-Gaudin C, Barba D. A convolutional neural network approach for objective video quality assessment. *Ieee T Neural Networ.* 2006;17(5):1316-27.
18. Matsugu M, Mori K, Mitari Y, Kaneda Y. Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Networks.* 2003;16(5-6):555-9.
19. Taherdoost H. Deep Learning and Neural Networks: Decision-Making Implications. *Symmetry.* 2023;15(9):1723.
20. Tan MX, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Pr Mach Learn Res.* 2019;97.
21. Ketkar N. *Deep Learning with Python.* Berkley, CA: Apress; 2017:195-208.
22. An Introudction to PyTorch - A simple yet Powerful Deep Learning Library [updated 2023.11.15. Available from: analyticsvidhya.com.
23. Introducing Accelerated PyTorch Training on Mac [updated 2023.11.15. Available from: pytorch.org.
24. Suh HS, Kweon C, Lester B, Kramer S, Sun WC. A publicly available PyTorch-ABAQUS UMAT deep-learning framework for level-set plasticity. *Mech Mater.* 2023;184:104682.
25. Novac OC, Chirodea MC, Novac CM, Bizon N, Oproescu M, Stan OP, Gordan CE.

Analysis of the Application Efficiency of TensorFlow and PyTorch in Convolutional Neural

Network. *Sensors-Basel*. 2022;22(22):8872.

26. Viscaino M, Maass JC, Delano PH, Cheein FA. Computer-Aided Ear Diagnosis System

Based on CNN-LSTM Hybrid Learning Framework for Video Otoscopy Examination. *Ieee Access*.

2021;9:161292-304.

국문 요약

배경

이내시경을 사용하여 고막을 평가하는 것은 임상 분야에서 첫 번째이자 가장 중요한 단계이다. 임상 현장에서 대부분의 고막 병변은 단일 질환만 있는 것이 아니라 둘 이상의 진단명을 가지게 되는 경우가 있다. 따라서 우리는 다양한 질병을 포함한 이내시경 이미지 데이터베이스를 구축하고, 딥러닝 네트워크를 이용하여 다양한 고막의 병변을 분류해 내는 능력에 대해 알아보려고 하였다.

방법

본 연구는 다중 클래스 분류를 사용하여 고막 질환의 진단 성능에 대하여 알아보았다. EfficientNet-B4 을 사용하여 주된 클래스(삼출성 중이염, 만성 중이염, '없음')와 부가 클래스(상고실 진주종, 고막염, 이진균증, 환기관 삽입)를 예측하였다.

결과

딥러닝 분류는 주된 클래스에 대해 95.19%의 정확도로 예측이 되었으며, 삼출성 중이염과 만성 중이염 사이에 잘못된 식별은 거의 발생하지 않았다. 부가 클래스 중에서는 상고실 진주종과 고막염의 진단이 각각 88.37%와 88.28%의 정확도를 보였다. 동시 질병의 경우 예측 성능이 비교적 낮았으나, 두 개 이상의 부가 클래스를 부정확하게 예측할 확률은 0.44%에 불과했다

(22/6,630). 이미지 당 추론 시간은 평균 2.594ms 였다.

결론

본 연구에서 제시된 알고리즘을 통하여 다중 동시 질병이 있는 상황에서도 정확하게 고막 병변을 예측할 수 있는 능력을 확인하였다. 이 결과는 추후 임상현장에서 의사 결정 과정에 긍정적으로 기여할 것으로 기대되며, 복잡한 의료 상황을 포함하는 사례들을 관리하는 데 유용도록 도움을 줄 수 있을 것으로 예상된다.