

Article

Diagnosis of Tympanic Membrane Disease and Pediatric Hearing Using Convolutional Neural Network Models with Multi-Layer Perceptrons

Hongchang Lee ^{1,†}, Hyeonung Jang ^{1,†} , Wangsu Jeon ²  and Seongjun Choi ^{3,*} 

¹ Haewootech Co., Ltd., Busan 46742, Republic of Korea; hcleee@haewootech.co.kr (H.L.); hwjang@haewootech.co.kr (H.J.)

² Department Computer Engineering, Kyungnam University, Changwon 51767, Republic of Korea; jws2218@naver.com

³ Department of Otolaryngology-Head and Neck Surgery, Cheonan Hospital, Soonchunhyang University College of Medicine, Cheonan 31151, Republic of Korea

* Correspondence: akas9238@hanmail.net; Tel.: +82-42-600-9215

† These authors contributed equally to this work.

Abstract: In this study, we propose a method of classification for tympanic membrane diseases and regression of pediatric hearing, using a deep learning model of artificial neural networks. Based on the B7 Backbone model of EfficientNet, a state-of-the-art convolutional neural network model, drop connect was applied in the encoder for generalization, and multi-layer perceptron, which is mainly used in the transformer, was applied to the decoder for improved accuracy. For the training data, the open-access tympanic membrane dataset, divided into four classes, was used as the benchmark dataset, and the SCH tympanic membrane dataset with five classes of tympanic membrane diseases and pediatric hearing was also used as the training dataset. In the benchmark using the open-access tympanic membrane dataset, the proposed model showed the highest performance among the five comparative models with an average accuracy of 93.59%, an average sensitivity of 87.19%, and an average specificity of 95.73%. In the experiment trained on the SCH tympanic membrane disease dataset, the average accuracy was 98.28%, the average sensitivity was 89.66%, the average specificity was 98.68%, and the average inference time was 0.2 s. In the experiment trained on the SCH pediatric hearing dataset, the mean absolute error was 6.8678, the mean squared logarithmic error was 0.2887, and the average inference time was 0.2 s.

Keywords: deep learning; eardrum; ear diseases; EfficientNet; multi-class classification; multi-layer perceptron; pediatric hearing; regression; tympanic membrane



Citation: Lee, H.; Jang, H.; Jeon, W.; Choi, S. Diagnosis of Tympanic Membrane Disease and Pediatric Hearing Using Convolutional Neural Network Models with Multi-Layer Perceptrons. *Appl. Sci.* **2024**, *14*, 5457. <https://doi.org/10.3390/app14135457>

Academic Editor: Herbert Schneckenburger

Received: 16 May 2024

Revised: 16 June 2024

Accepted: 19 June 2024

Published: 24 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The world is seeing a steady increase in ear diseases, affecting millions of children a year, as shown in Figure 1. Among them, the most prominent is middle ear infection, accounting for 46.3%. Hearing loss and a significant impact on quality of life can be caused by otitis media, a common ear infection in children. The reason for this infection is either the inflammation or blockage of the Eustachian tube, which often occurs because of the shorter and underdeveloped Eustachian tubes in children [1–3].

The diagnosis of otitis media is challenging because symptoms are often mild and unrecognized by the patient, such as slight hearing loss and discomfort [4]. This has the potential to cause delayed treatment and potential complications [5,6]. Therefore, accurate diagnosis and early intervention are crucial to preventing hearing loss and its negative consequences on social interaction, education, employment, and overall well-being [7].

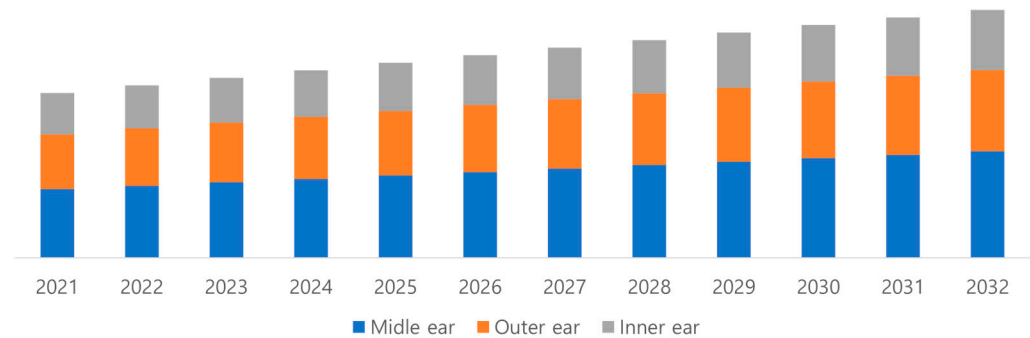


Figure 1. Ear infection treatment market, by infection type, 2021–2032 (Global Market Insights).

Different techniques are employed to diagnose the tympanic membrane, but there is a lack of objective and precise diagnostic methods. The diagnostic methods that currently exist include clinical tests, hearing tests, and imaging tests. First, a clinical test is a doctor's visual examination of the symptoms of otitis media. A clinical examination can also be subjective and less precise. Second, hearing tests are used to check pediatric hearing loss due to otitis media. Children may experience inconvenience from time-consuming hearing tests. Finally, imaging tests are methods to generate images of the middle ear cavity using techniques such as CT (computed tomography) or MRI (magnetic resonance imaging). Imaging tests can be used to identify structural changes in otitis media, but they are expensive and pose a risk of radiation exposure.

Recently, artificial intelligence technology, particularly deep learning [8,9] technology among artificial neural networks [10], has been used to develop new diagnostic methods in the medical field. Deep learning technology is capable of creating predictive models by analyzing data like images, speech, and text. Various studies [11] of deep learning are also being conducted on the middle ear disease [12], and these studies are aimed at assisting in diagnosis. The average diagnostic accuracies of doctors [13] are 73% (Otolaryngologists) and 50% (Pediatricians). Since the diagnosis results and their precision vary from doctor to doctor, and since there is a possibility of bias [14], research on diagnostic assistance using deep learning is emerging to solve this problem.

We propose a novel model for diagnosing tympanic membrane disease and predicting pediatric hearing by combining the CNN (convolution neural networks) [15] and MLP (multi-layer perceptron) [16] models. Previous research has demonstrated that CNN models are highly effective in extracting and classifying features from medical images. In contrast, MLP models are effective in learning complex nonlinear relationships.

This study covers the medical definitions of tympanic membrane diseases and pediatric hearing. First, the OME (otitis media with effusion) disease is a condition in which non-infectious fluids accumulate in the middle ear. Second, congenital cholesteatoma disease is an abnormal skin growth that occurs in the middle ear canal. Third, traumatic perforation disease is the formation of a hole in the tympanic membrane. Fourth, the COM (chronic otitis media) disease is the development of inflammation in the middle ear (middle ear) that lasts for more than 3 months. Fifth, AOM (acute otitis media) is the development of inflammation in the middle ear that lasts less than 3 weeks. Sixth, otitis externa is an inflammation that occurs in the ear canal (a passage behind the ear). Finally, pediatric hearing, which is affected by these diseases, refers to the ability to understand sounds and languages.

The main goals of this study are as follows. The first is to develop a combined CNN and MLP model for diagnosing tympanic membrane disease, the second is to develop a combined CNN and MLP model for predicting pediatric hearing, and the last is to evaluate the performance of the proposed models.

This study is anticipated to improve the accuracy of medical judgment in diagnosing tympanic membrane diseases and predicting pediatric hearing. In addition, the proposed

model is expected to play a role in showing the possibility of the combination of the CNN and the MLP in the field of medical image analysis.

2. Related Work

Recently, as the role of artificial neural networks as diagnostic tools emerges, their influence has gradually expanded, and various methods of using neural networks have been used for research on medical data.

Autoencoder, an unsupervised learning method, is one of them. Song et al. [17] studied anomaly detection, the task of identifying sample data that do not match the overall data distribution, using a variational autoencoder [18]. Because variational autoencoders make complex data such as tympanic membrane endoscopic images difficult to learn, they pre-processed tympanic membrane images using adaptive histogram equalization and canny edge detection. Then, they made the variational autoencoder learn the preprocessed data only for the normal tympanic membrane and applied the normal and abnormal tympanic membrane image anomaly scores of the distribution of the variational autoencoder to the K-nearest neighbor algorithm to classify the normal and abnormal tympanic membrane images. As a result, a total of 1232 normal and abnormal tympanic membrane images were obtained, which were classified with 94.5% accuracy, using the algorithm that applied only the normal tympanic membrane image. Studies on lightweight models are also being attempted in many directions.

Yue et al. [19] constructed the first large-scale ear endoscopy dataset consisting of eight types of ear disease and disease-free samples from two institutions. Inspired by ShuffleNetV2 [20], Best-EarNet is an ultra-fast and ultra-lightweight network that enables real-time ear disease diagnosis. Best-EarNet includes a novel local-global spatial feature fusion module and a multi-scale supervision strategy, making it easy to focus on global-local information within different levels of feature maps. Using transfer learning, the accuracy of Best-EarNet with only 0.77 M parameters achieved 95.23% (with 22,581 images inside) and 92.14% (with 1652 images outside), respectively. Specifically, the average frame per second is 80, so real-time computation was possible.

Zeng et al. [21] presented a deep learning model to automatically diagnose tympanic diseases in real time using abundant otoscope image data obtained from clinical cases. They trained nine common deep CNNs using a total of 20,542 endoscopic images and classified eight ear diseases, including normal diseases, cholesteatoma of the middle ear, chronic suppurative otitis media, external auditory canal bleeding, impacted cerumen, otomycosis external, secretory otitis media, tympanic membrane classification. A transfer learning model was selected by them to construct an ensemble model with DensNet-BC169 [22] and DensNet-BC1615, which has an average accuracy of 95.59%.

3. Materials and Methods

In this chapter, we will cover the datasets that are employed for learning and the process of preprocessing them. Our description includes the proposed model's structure, hyperparameters, and model evaluation indices.

3.1. Open-Access Tympanic Membrane Dataset

This study utilized data acquired from Kaggle in addition to using the open-access tympanic membrane dataset [23], which is an open dataset used in various papers. Normal, COM, AOM, and otitis externa are represented by 757 TIFF images in this dataset. The ratio between the training and test data is 75:25, as shown in Table 1.

Prior to training the SCH (Soonchunhyang University Hospital) tympanic membrane dataset, the performance of each model is compared with the open-access tympanic membrane dataset. There are a total of five comparative models, including MobileNet V3 [24], DenseNet 201, EfficientNet B7 [25], ConvNeXt [26], and the proposed model.

Table 1. Open-access tympanic membrane dataset.

	Training Data	Test Data
Normal	400	134
Chronic Otitis Media	47	16
Acute Otitis Media	89	30
Otitis Externa	31	10
Total	567	190

3.2. SCH Tympanic Membrane Dataset

This study uses 23,302 JPG image files provided by SCH after de-identification, which were approved by the institutional review committee of SCH. The dataset is divided into a classification dataset and a regression dataset, and each has a different training task. Usually, patients obtain their eardrum images and EAC (external auditory canal) photos via an oto-endoscopy (Pentax, Berlin, Germany) upon visit. The resolution rate of these images is 1280 (h) $\times$ 1350 (w) pixels.

The tympanic membrane disease subset in the dataset is a dataset for classification of a total of five classes: normal (completely normal eardrum, normal with healed perforation or some tympanosclerosis), OME (light yellow, orange oil or amber color, but if the liquid does not fill in the tympanic cavity, the liquid level can be seen through the tympanic membrane), cholesteatoma (loose inner pocket can be seen, and white exfoliated epithelium can be seen inside the pocket), traumatic perforation (there is perforation of the tympanic membrane, and they are not a uniform size).

In COM, the tympanic membrane may perforate due to tension and exhibit blood clumps and uneven size. Most of them are single shots. The residual tympanic membrane may have calcification, ulceration and granulation tissue growth around the perforation margin. All the image labeling was conducted by three ear specialists with more than ten years of experience.

OME was diagnosed according to the clinical otologic practice that included medical history, physical examination with otoscopes, and audiological tests (PTA [pure tone audiometry] and tympanometry). Inclusion criteria required that otoscopic images and audiological assessment results be measured at the same time and on individual OME ears. Ears with OME and a history of middle ear surgery (e.g., grommet insertion) were excluded. The pediatric hearing subset in the OME dataset is a dataset for the hearing threshold of 1 kHz in the left and right ear.

The split of the training set, validation set, and test set of the SCH tympanic membrane disease subset in the dataset was handled by SCH, and the ratio is 8:1:1. For training, a training set and a verification set were first received. The composition of the data for training is shown in Table 2. After communicating through the training set and validation set that the training was completed, the test set was received, and the test was conducted. The composition of the data that were tested is also shown in Table 2.

Table 2. SCH tympanic membrane disease subset dataset.

	Training Data	Validation Data	Test Data
Normal	11,686	1464	1464
Otitis Media with Effusion	1866	233	234
Cholesteatoma	183	23	23
Perforation	194	25	25
Chronic Otitis Media	2034	255	255
Total	15,963	2000	2001

The pediatric hearing subset in the dataset is a dataset for regression and has a certain value of dB (Decibel). The split of this dataset was also dedicated to SCH, and the ratio of training set, validation set, and test set is 8:1:1. First, training was conducted by receiving a training set and a validation set, and the composition of the data is shown in Table 3. After communicating through the training set and validation set that the learning was completed, the test set was received, and the test was conducted. The composition of the data that were tested is also shown in Table 3, and the distribution of all data in training, validation, and testing is visualized in Figure 2.

Table 3. SCH pediatric hearing subset training dataset.

	Training Data	Validation Data	Test Data
Pediatric Hearing	2670	334	334

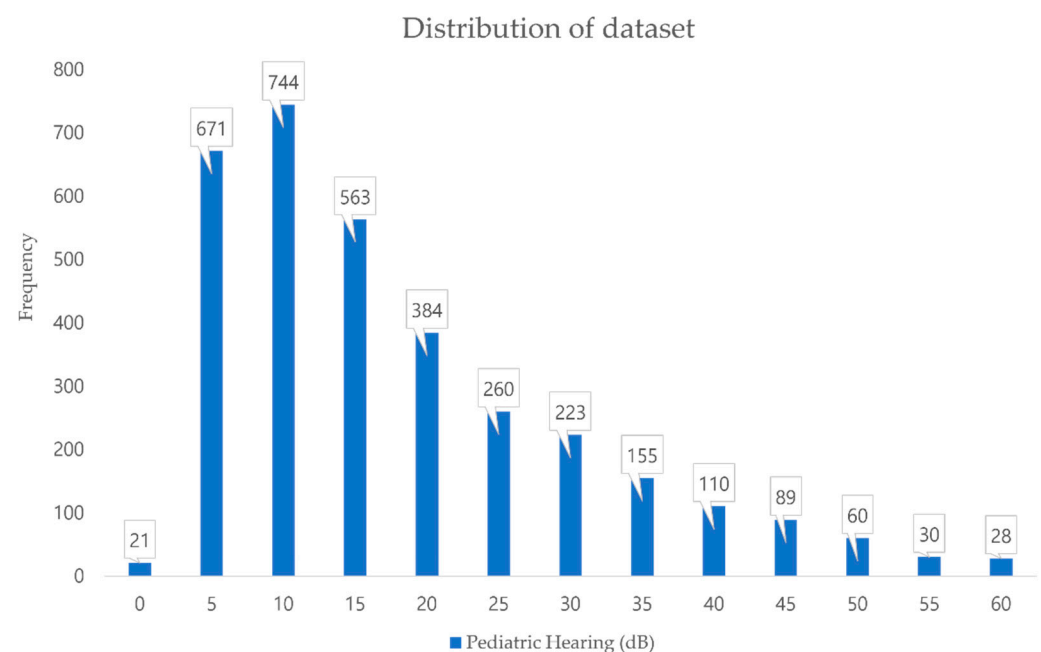


Figure 2. Distribution chart of hearing data by dB.

3.3. Data Preprocessing

A standardization layer for convergence learning was developed by EfficientNet, which resized image data from various formats to a 600×600 8-bit RGB format. Ground truth used one-hot encoding and label smoothing [27] to classify datasets like the open-access tympanic membrane dataset and the SCH tympanic membrane disease subset dataset.

In the case of label smoothing, correction is applied to prevent predictions close to 0 and 1 from becoming overly confident, and through this, neural networks are constantly focused on classes with lowered predictions through correction to improve performance. The formula for this label smoothing is shown in Equation (1), y_k is the GT value, α is the label smoothing ratio, and K is the number of classes. In the experiment, training was conducted with the smoothing ratio of $\text{Label } 1 \times 10^{-1}$, as shown in Tables 4 and 5.

$$y_{kLS} = y_k (1 - \alpha) + \alpha / K, \quad (1)$$

Table 4. The result of label smoothing on open-access tympanic membrane dataset.

	Non-Label Smoothing	Label Smoothing
Negative	0	0.025
Positive	1	0.925

Table 5. The result of label smoothing on SCH tympanic membrane disease subset dataset.

	Non-Label Smoothing	Label Smoothing
Negative	0	0.02
Positive	1	0.92

Standardization was used for regression datasets such as the SCH pediatric hearing subset dataset. The formula used in the standardization of the SCH pediatric hearing subset dataset is shown in Equation (2) below, which is the same as the formula of the standardization layer designed inside the EfficientNet. The *Mean* and *Standard Deviation* values used in the equation are shown in Table 6.

$$\text{standardization} = \frac{x_i - \bar{x}}{s} \quad (\bar{x} : \text{Mean}, s : \text{Standard Deviation}) \quad (2)$$

Table 6. The *Mean* and *Standard Deviation* on SCH pediatric hearing dataset.

	Mean	Standard Deviation
SCH pediatric hearing GT	18.0557	12.6491

3.4. Model Design

3.4.1. Backbone

The basic backbone model uses the EfficientNet model, which achieves State-of-The-Art in five dataset segments, including Flowers and CIFAR-100. EfficientNet achieved both top-1 and top-5 accuracy in ImageNet while reducing the number of parameters and attaining high accuracy, unlike the existing CNN model, which had a significant number of parameters. To improve model performance, compound scaling is essential, and the optimal values were found by organically adjusting the Width, Depth, and Resolution scaling.

As shown in Table 7, the optimized value exhibits the best performance in terms of computation and accuracy, and the compound scaling combination formula is based on Equation (3) below. In this equation, α , β , and γ are constants and are found using grid search, and ϕ is a factor that can be controlled by the user and takes an appropriate value according to the available resources.

$$\begin{aligned} \text{Depth} : d &= \alpha^\phi \\ \text{Width} : w &= \beta^\phi \\ \text{Resolution} : r &= \gamma^\phi \\ s \cdot t \cdot \alpha \cdot \beta^2 \cdot \gamma^2 &\approx 2 \\ \alpha &\geq 1, \beta \geq 1, \gamma \geq 1 \end{aligned} \quad (3)$$

Table 7. Performance with scale change at the same amount of computation.

Model	FLOPS	Top-1 Acc
EfficientNet-B0 (Baseline model)	0.4 billion	77.3%
Scale model by depth ($d = 4$)	1.8 billion	79.0%
Scale model by width ($w = 2$)	1.8 billion	78.9%
Scale model by resolution ($r = 2$)	1.9 billion	79.1%
Compound Scale ($d = 1.4, w = 1.2, r = 1.3$)	1.8 billion	81.1%

EfficientNet has a group of models such as B0, B1, B2, B3, B4, B5, B6, B7, B8, and L2 (added after the launch of EfficientNet for B8 and L2), and each model has its own Compound Scaling value. As the number of models increases, the amount of computation doubles and the intensity of regulations to prevent overfitting also increases.

3.4.2. Multi-Layer Perceptron

The MLP structure is utilized in this study to enhance the performance of EfficientNet. The MLP used is a structure that repeats fully connected with 4096 units, swish activation, and dropout [28] with a 50% probability 5 times, referring to the structure of a Transformer [29] model that utilizes MLP in various ways. This structure is used to construct the EfficientNet B7, as shown in Figure 3.

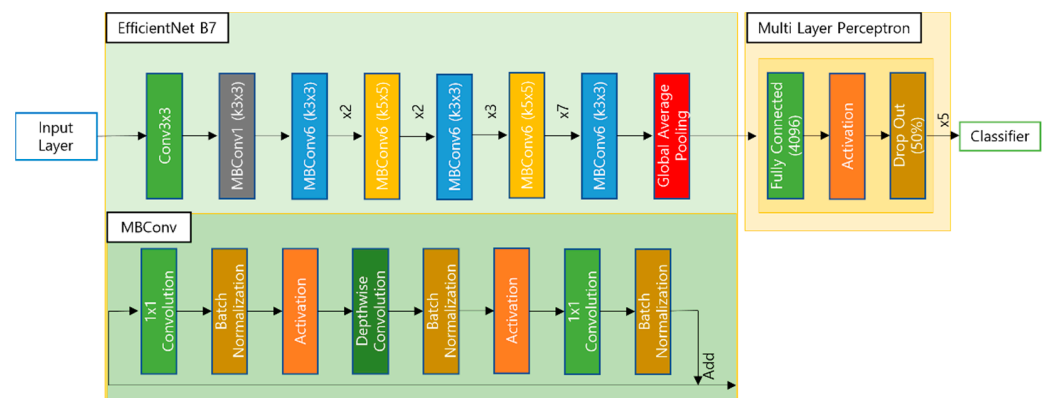


Figure 3. EfficientNet B7 model with multi-layer perceptron as a decoder.

3.4.3. Drop Connect

To prevent overfitting due to the huge size of EfficientNet B7, we applied drop connect [30]. Drop connect is a follow-up study of dropout that randomly selects nodes and turns them to zero. Unlike dropout, it is a regularization for co-adaptation prevention that deactivates weights. Dropout had previously been utilized to correct the MLP pattern, but drop connect was employed to enhance the performance and result in a 50% weight inactivation rate.

3.5. Hyper Parameters

3.5.1. Calibration Weight Classes

There are more than 1000 classes in the data class of ImageNet, and the amount of data per class is different. Most datasets are extremely rare and the data are evenly distributed for each class. As such, the problem of data imbalance by class is a very important issue in classification tasks. To solve this problem, we could consider a method of adjusting the frequency of sampling and a method of adjusting the weight by class. This paper uses the most recent method to calculate the weight of each class, which is based on Equation (4).

$$\text{Class Weight} = \frac{\text{number of samples}}{\text{number of classes} \times \text{occurrences of each class value}} \quad (4)$$

Except for the SCH pediatric hearing subset dataset, which is a regression task, the open-access tympanic membrane dataset and the SCH tympanic membrane disease subset dataset are both classification tasks, and the weights of the classes can be calculated. Basically, high weights are given to classes with limited data, low weights are given to classes with abundant data, and the calculated weights for each class are shown in Table 8 for the open-access tympanic membrane dataset and Table 9 for the SCH tympanic membrane disease subset dataset.

Table 8. Open-access tympanic membrane dataset weights by disease type.

	Training Data	Class Weight
Normal	400	0.3544
Chronic Otitis Media	47	3.0160
Acute Otitis Media	89	1.5927
Otitis Externa	31	4.5726

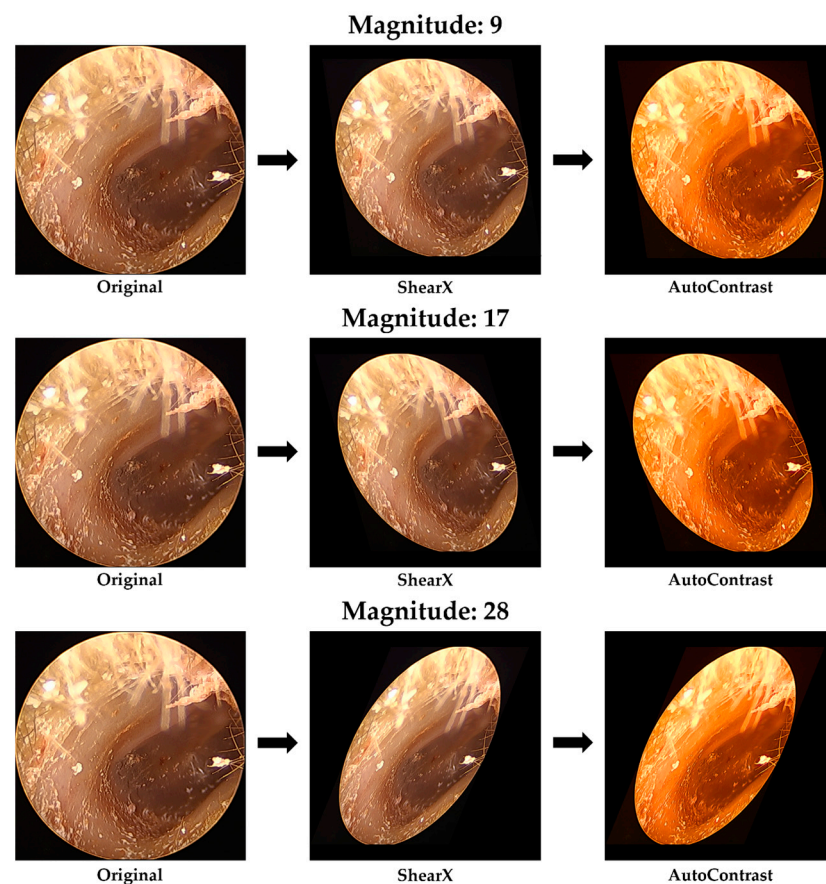
Table 9. SCH tympanic membrane disease subset dataset weights by disease type.

	Training Data	Class Weight
Normal	11,686	0.2397
Otitis Media with Effusion	1866	1.5014
Cholesteatoma	183	15.3093
Perforation	194	14.4412
Chronic Otitis Media	2034	35.4633

3.5.2. Rand Augment

The augmentation used in the training is rand augment [31], an augmentation that applies up to N random augmentations with maximum random intensity M. Rand augment is a technology that refers to fast auto augment [32] and can be applied with a very small amount of computation.

Figure 4 shows an example of rand augment, which shows the difference between the M values of 9, 17, and 28 when shearX and auto contrast were randomly selected at N = 2. As such, rand augment randomly selects N augmentation techniques for each image and applies a random magnitude between 0 and M.

**Figure 4.** Example images augmented by rand augment.

In the experiment, standard augmentations of rand augment such as flipLR, identity, auto contrast, equalize, rotate, solarize, color, posterize, contrast, brightness, sharpness, shearX, shearY, translateX, and translateY were applied. N is 2, and M is 28.

3.5.3. AdaBelief

AdaBelief [33] is an algorithm that Adam [34] uses to adjust convergence speed and generalization performance [35] by using the variance value of the slope as a replacement for Adam, which is momentum squared. Ada is derived from Adam, and Belief is named because the variance is calculated with the currently estimated momentum value and has a distance squared from the predicted slope. Despite a one-line change in the code, it still received much attention for its exceptional performance improvement. This study uses a global clip norm of 1, the learning rate of 1×10^{-4} , and weight decay of 1×10^{-4} to conduct training.

3.5.4. Mixed Precision

Mixed precision [36] is a method of converting the existing float32 operation into the float16 operation and converting the classifier back to the float32 operation, which enables twice as fast learning by reducing the burden of memory in half while maintaining accuracy. During the float16 operation, there may be losses due to values exceeding the range, which is corrected through scaling. In this study, the maximum batch size 16 was raised to 32 using mixed precision, and through this, it was possible to conduct smooth training and improve the performance by increasing the efficiency of batch normalization [37] in the model.

3.5.5. Loss

The general categorical cross-entropy was used for the loss of the classification task and the Huber loss [38], which combines the outlier robustness of the L1 loss, the fast convergence speed, and the training stability according to the differentiable of the L2 loss, was used for the loss of the regression task. In Huber loss, as shown in Equation (5), if the difference between GT and the predicted value is less than a specific threshold δ value, it follows L2 loss, and if it is large, it follows L1 loss, and in training, this δ value was set to 0.25.

$$l_n = \begin{cases} 0.5(y_n - \hat{y}_n)^2, & \text{if } |y_n - \hat{y}_n| < \delta \\ \delta \times (|y_n - \hat{y}_n| - 0.5 \times \delta), & \text{otherwise} \end{cases} \quad (5)$$

3.6. Metrics

Training evaluation is used separately for classification tasks such as the open-access tympanic membrane dataset and SCH tympanic membrane disease subset dataset and for regression tasks such as the SCH pediatric hearing subset dataset.

To calculate the metrics of classification, the confusion matrix for each class is first obtained. Based on the TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) of the classes obtained here, the Average Accuracy, Average Sensitivity, and Average Specificity of the class are obtained, and the model is evaluated with these Metrics. Each metric follows an Equations (6)–(8).

$$\text{Average Accuracy} = \frac{\sum_{class=1}^{n_{class}} (TP_{class} + TN_{class})}{\sum_{class=1}^{n_{class}} (TP_{class} + TN_{class} + FP_{class} + FN_{class})}, \quad (6)$$

$$\text{Average Sensitivity} = \frac{\sum_{class=1}^{n_{class}} TP_{class}}{\sum_{class=1}^{n_{class}} (TP_{class} + FN_{class})}, \quad (7)$$

$$\text{Average Specificity} = \frac{\sum_{class=1}^{n_{class}} TN_{class}}{\sum_{class=1}^{n_{class}} (TN_{class} + FP_{class})}, \quad (8)$$

The models of the regression task evaluate the performance with Mean Absolute Error and Mean Squared Logarithmic Error, which follow Formulas (9) and (10).

$$\text{Mean Absolute Error} = \frac{1}{n_{\text{sample}}} \sum_{i=1}^{n_{\text{sample}}} (y_i - \hat{y}_i), \quad (9)$$

$$\text{Mean Squared Logarithmic Error} = \frac{1}{n_{\text{sample}}} \sum_{i=1}^{n_{\text{sample}}} (\log(1 + y_i) - \log(1 + \hat{y}_i))^2 \quad (10)$$

4. Results

The device used in the experiment was hamoniKR 6.0, based on Linux Ubuntu 20.04. The CPU was equipped with an Intel Xeon Gold 6346 64 core, 3.10 GHz, and the GPU was equipped with four RTX 3090, 256 GB, of RAM. It installed Nvidia CUDA 11.2 and cuDNN 11.2, while Python version 3.9.13 and Anaconda 22.9 were used. The deep learning framework was based on TensorFlow 2.11.1 and Keras 2.11.0, and all experiments were conducted.

4.1. Open-Access Tympanic Membrane Dataset

In this study, benchmarks for each model were conducted using an open-access tympanic membrane dataset, and validation and performance measurements were performed as a test set. A total of five models were compared to measure the average accuracy, average sensitivity, and average specificity of normal, COM, AOM, and otitis externa, and the training graph for each model is shown in Figure 5.

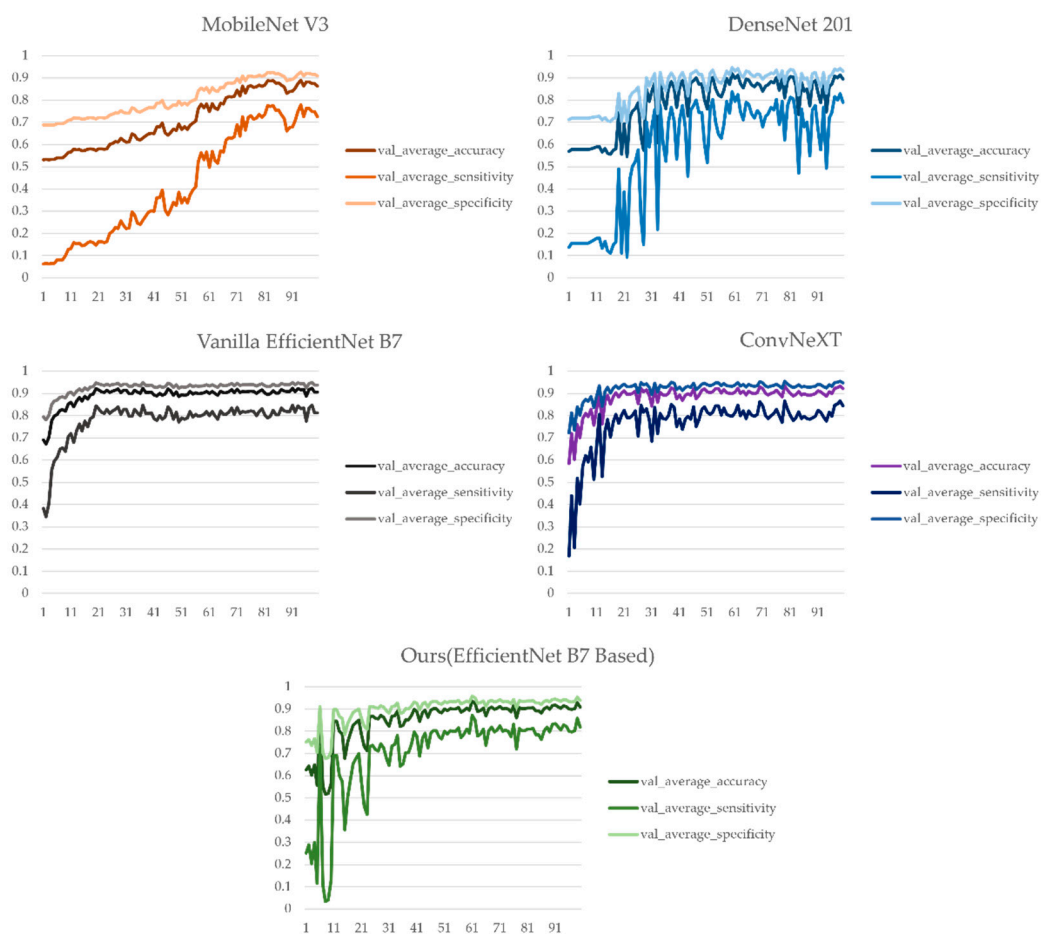


Figure 5. The visualization of validation trajectory in 100 epochs on open-access tympanic membrane dataset.

For a quick comparison, we used weights that pre-trained the dataset of ImageNet for each model, which converged all models within 100 epochs. Based on the epoch that obtained the highest performance, the model's performance was higher in the order of Our Model, ConvNeXt, Vanilla EfficientNet B7, DenseNet 201, and MobileNet V3. In particular, the fact that Our Model led ConvNeXt, which showed higher performance than Vanilla EfficientNet B7, shows that MLP and drop connect had a good effect on the performance, as shown in Table 10.

Table 10. The quantitative comparison results on the open-access tympanic membrane dataset.

Model	Average Accuracy	Average Sensitivity	Average Specificity
MobileNet V3	88.91%	77.81%	92.60%
DenseNet 201	91.88%	83.75%	94.58%
Vanilla EfficientNet B7	92.34%	84.69%	94.90%
ConvNeXt	93.28%	86.56%	95.52%
Ours (EfficientNet B7-based)	93.59%	87.19%	95.73%

4.2. SCH Tympanic Membrane Dataset

The training of the SCH tympanic membrane dataset used the proposed model, MLP, and EfficientNet B7, which connected drop connect. For better performance, the model was fine-tuned with the weight of the noisy student [39]. The weight of the noisy student refers to the additional training of the JFT-300M dataset on the ImageNet large-capacity dataset using the noisy student training method that divides the teacher and student model into non-label training.

The experimental results of the tympanic membrane disease dataset are as follows. As a result of training all 100 epochs, the weight at 50 epochs showed the highest performance, and the validation and test performance of the corresponding weight are shown in Table 11. Table 12 shows the inference time for measuring the performance of the test set received from SCH with the previous weight. Figure 6 shows a visualization of the correct answer prediction results for each class using Grad-CAM [40].

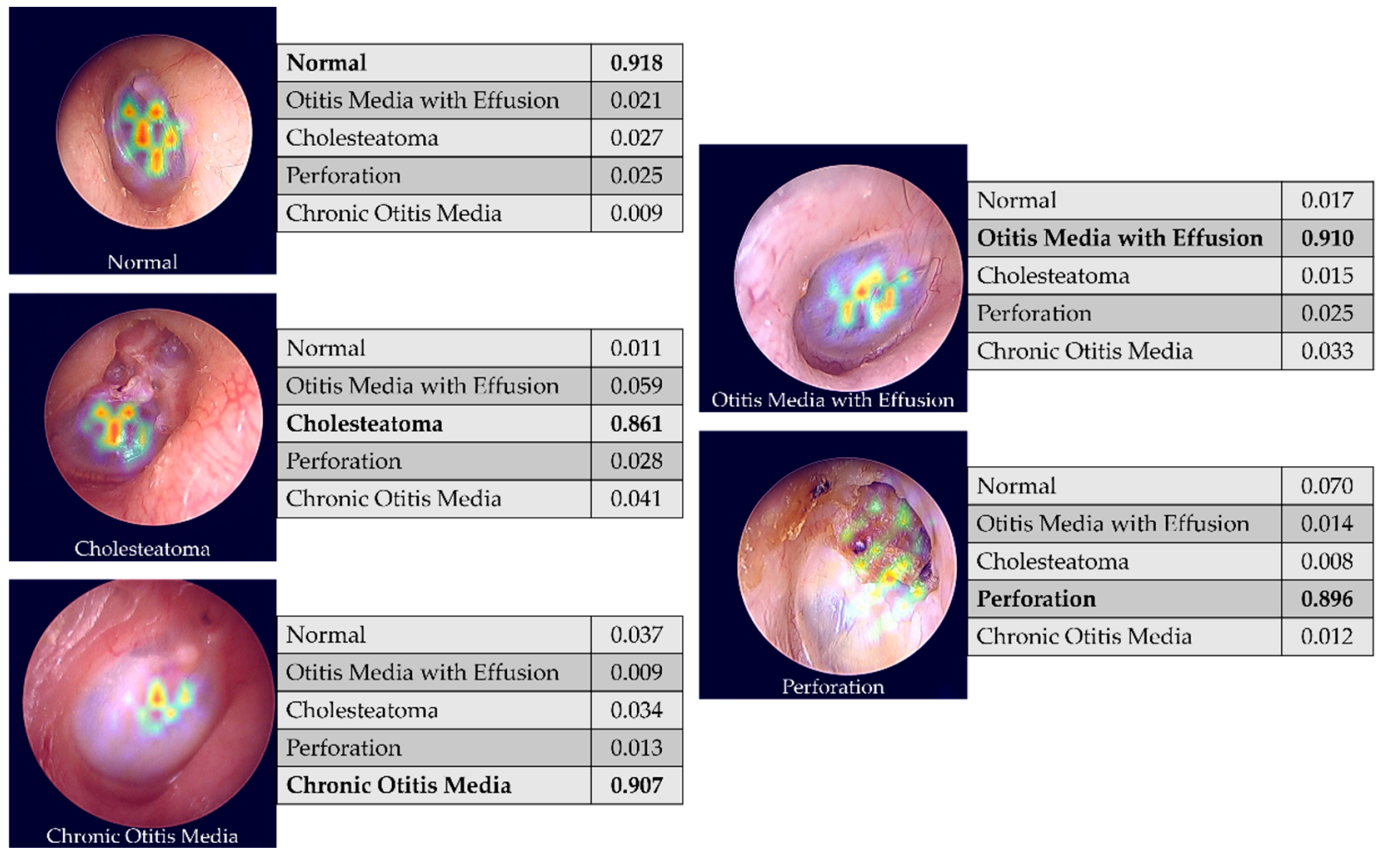
The experimental results of the pediatric hearing dataset are as follows. As a result of learning all 300 epochs, 215 epochs showed the highest performance, and the validation and test performance at these epochs are shown in Table 13. The inference time measuring the performance of the test set received from SCH is shown in Table 14, and the result of visualizing the predicted result for each dB with a difference of less than 5 between the predicted value and GT in Grad-CAM is shown in Figure 7. In the case of Table 13, additional benchmarks were executed with the same data to evaluate the regression performance of the proposed model. As a result, like the classification part, it was confirmed that the performance of our model was the most compliant among the comparative models.

Table 11. The quantitative comparison results on SCH tympanic membrane disease dataset.

	Accuracy	Sensitivity	Specificity
Validation Average	98.36%	85.88%	98.58%
Normal	96.25%	95.83%	97.39%
Otitis Media with Effusion	98.40%	96.58%	98.64%
Cholesteatoma	99.25%	82.61%	99.44%
Perforation	98.00%	97.25%	98.11%
Chronic Otitis Media	99.50%	76.03%	99.82%
Average	98.28%	89.66%	98.68%

Table 12. Inference time on the proposed model.

Inference Time	
Total	6 min
Average	0.2 s

**Figure 6.** Visualization comparison with Grad-CAM on SCH tympanic membrane disease dataset.**Table 13.** The quantitative pediatric hearing result on SCH tympanic membrane disease dataset.

	Mean Absolute Error	Mean Squared Logarithmic Error
Validation Hearing	6.9801	0.2798
MobileNet V3	7.5890	0.3260
ConvNeXt	7.4721	0.3005
Ours	6.8678	0.2887

Table 14. Inference time on the pediatric hearing model.

Inference Time	
Total	1 min
Average	0.2 s

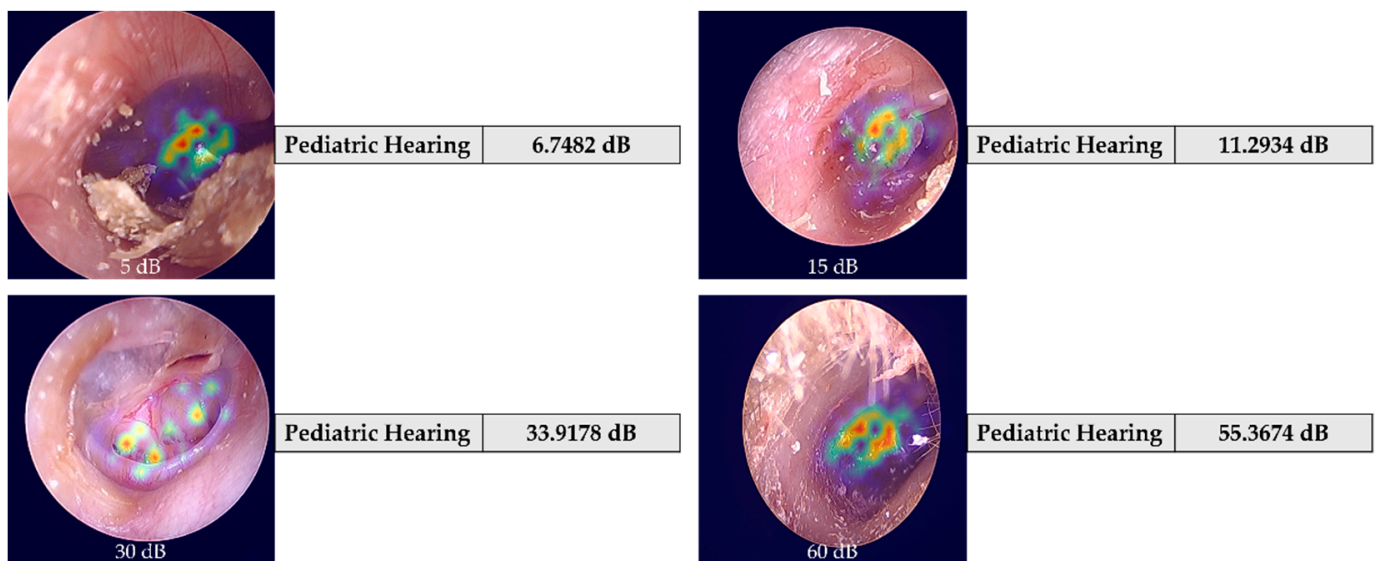


Figure 7. Visualization of the pediatric hearing using Grad-CAM.

Unlike the average performance in the experiment, both the classification model and the regression model had the problem of lowering performance in a specific class or dB. As shown in Figures 8 and 9, many such problems were seen in the cholesteatoma for the tympanic membrane disease model and 60 dB for the pediatric hearing model. This seems to be a problem caused by a data imbalance, and it was not completely overcome by a method such as class weight in the training process.

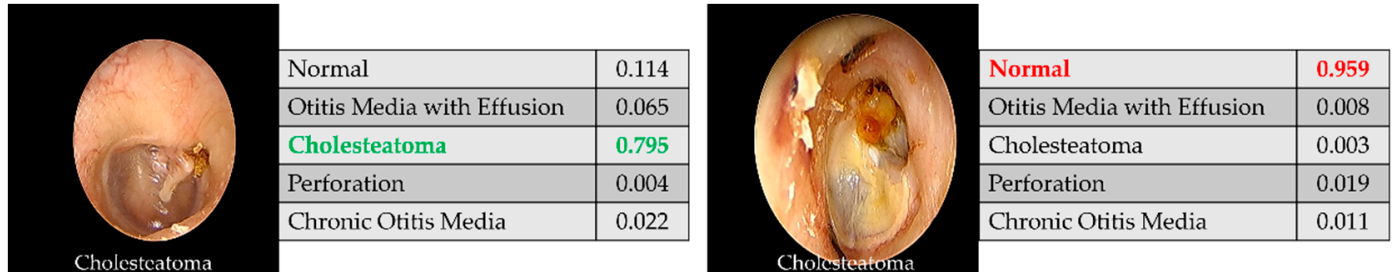


Figure 8. Correct and incorrect classification of Cholesteatoma. (Bold green letters indicate the correct prediction, and bold red letters indicate the incorrect prediction).

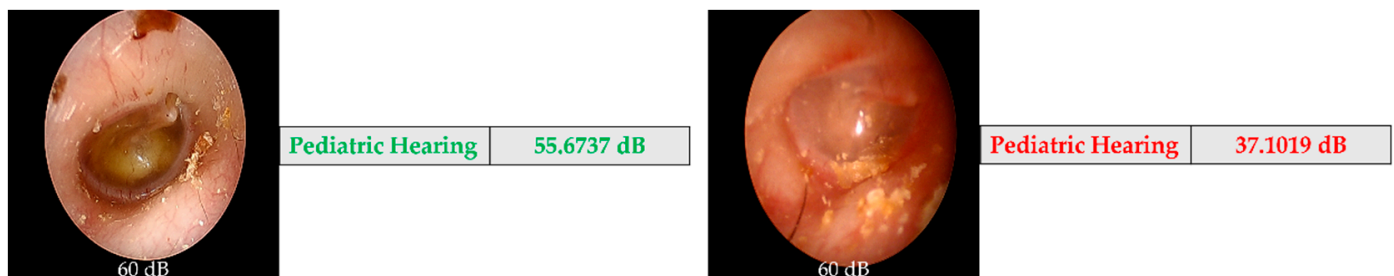


Figure 9. Small prediction error and large prediction error when GT is 60 dB. (Bold green letters indicate the correct prediction, and bold red letters indicate the incorrect prediction).

4.3. Cross Validation

We performed additional cross validation through k-fold to evaluate the performance of each class in more detail. This was performed on the previous two classification datasets, and after integrating the training set, validation set, and test set, it was proceeded by stratified sampling with 5-fold.

The results for the open-access tympanic membrane dataset are shown in Tables 15 and 16, and the visualization of this as a box plot is shown in Figure 10. Also, the results for the SCH tympanic membrane dataset are shown in Tables 17 and 18, and the visualization of this as a box plot is shown in Figure 11. From the box plot of each dataset, it can be seen that the deviation of the performance for each mold is not small compared to the average performance. These deviations are attributed to data imbalances due to the presence of data-poor classes.

Table 15. The performance of an open-access tympanic membrane disease model varies based on the number of folds.

Fold 1	Train Data	Test Data	Class Weight	Accuracy	Sensitivity	Specificity
Normal	427	107	0.3544	95.16%	88.79%	96.67%
Chronic Otitis Media	50	13	3.0040	96.05%	80.23%	98.56%
Acute Otitis Media	95	24	1.5903	88.82%	79.17%	90.63%
Otitis Externa	33	8	4.6159	96.71%	75.00%	97.92%
Fold 2	Train Data	Test Data	Class Weight	Accuracy	Sensitivity	Specificity
Normal	427	107	0.3544	96.50%	89.79%	97.44%
Chronic Otitis Media	50	13	3.0040	96.71%	84.62%	97.84%
Acute Otitis Media	96	23	1.5903	90.13%	73.91%	93.02%
Otitis Externa	32	9	4.6159	95.39%	76.67%	97.20%
Fold 3	Train Data	Test Data	Class Weight	Accuracy	Sensitivity	Specificity
Normal	427	107	0.3544	93.38%	87.20%	84.09%
Chronic Otitis Media	51	12	3.0040	98.68%	83.33%	99.99%
Acute Otitis Media	95	24	1.5903	93.38%	75.00%	96.85%
Otitis Externa	33	8	4.6159	98.68%	87.50%	99.30%
Fold 4	Train Data	Test Data	Class Weight	Accuracy	Sensitivity	Specificity
Normal	427	107	0.3544	97.42%	93.18%	98.73%
Chronic Otitis Media	51	12	3.0040	97.39%	83.33%	98.09%
Acute Otitis Media	95	24	1.5903	92.07%	83.33%	91.34%
Otitis Externa	33	8	4.6159	95.36%	72.50%	97.20%
Fold 5	Train Data	Test Data	Class Weight	Accuracy	Sensitivity	Specificity
Normal	428	106	0.3544	89.47%	83.58%	93.33%
Chronic Otitis Media	50	13	3.0040	98.01%	92.31%	98.55%
Acute Otitis Media	95	24	1.5903	90.13%	83.33%	89.53%
Otitis Externa	33	8	4.6159	96.70%	82.50%	96.50%

Table 16. The 5-fold average performance of the open-access tympanic membrane disease model.

	5-Fold Average Accuracy	5-Fold Average Sensitivity	5-Fold Average Specificity
Normal	94.39%	88.51%	94.05%
Chronic Otitis Media	97.37%	84.76%	98.61%
Acute Otitis Media	90.91%	78.95%	92.27%
Otitis Externa	96.57%	78.83%	97.62%
Average	94.81%	82.76%	95.64%

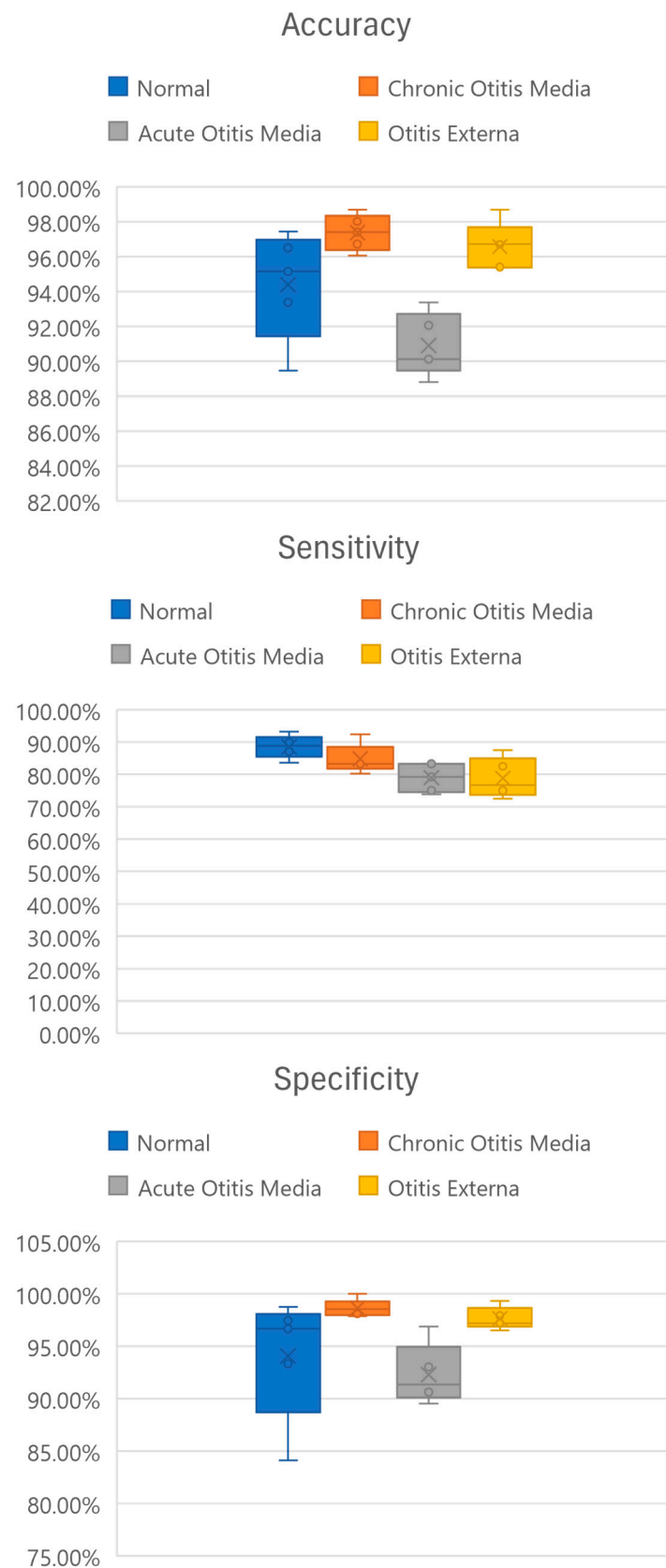


Figure 10. Visualization comparison of box plot results on open-access tympanic membrane disease model.

Table 17. The performance of an SCH tympanic membrane disease model varies based on the number of folds.

Fold 1	Train Data	Test Data	Class Weight	Accuracy	Sensitivity	Specificity
Normal	11,691	2923	0.2732	97.47%	97.84%	96.45%
Otitis Media with Effusion	1866	467	1.7114	98.35%	95.50%	98.72%
Cholesteatoma	183	46	17.4358	99.20%	30.43%	99.99%
Perforation	196	48	16.6939	99.05%	64.58%	99.47%
Chronic Otitis Media	2035	509	1.5695	97.82%	94.30%	98.34%
Fold 2	Train Data	Test Data	Class Weight	Accuracy	Sensitivity	Specificity
Normal	11,691	2923	0.2732	97.42%	98.05%	95.70%
Otitis Media with Effusion	1867	466	1.7114	98.45%	91.85%	99.32%
Cholesteatoma	183	46	17.4358	99.95%	95.65%	99.98%
Perforation	195	49	16.3639	98.90%	67.35%	99.29%
Chronic Otitis Media	2035	509	1.5695	98.17%	94.30%	98.74%
Fold 3	Train Data	Test Data	Class Weight	Accuracy	Sensitivity	Specificity
Normal	11,691	2923	0.2732	96.89%	96.41%	98.22%
Otitis Media with Effusion	1867	466	1.7114	97.82%	97.85%	97.82%
Cholesteatoma	183	46	17.4358	99.87%	91.30%	99.97%
Perforation	195	49	16.3639	99.57%	87.76%	99.72%
Chronic Otitis Media	2035	509	1.5695	98.92%	97.45%	99.14%
Fold 4	Train Data	Test Data	Class Weight	Accuracy	Sensitivity	Specificity
Normal	11,691	2923	0.2732	92.84%	91.00%	97.85%
Otitis Media with Effusion	1866	467	1.7114	97.32%	85.22%	98.92%
Cholesteatoma	183	46	17.4358	97.82%	93.48%	97.87%
Perforation	195	49	16.3639	95.37%	89.80%	95.44%
Chronic Otitis Media	2036	508	1.5695	96.32%	87.01%	97.68%
Fold 5	Train Data	Test Data	Class Weight	Accuracy	Sensitivity	Specificity
Normal	11,692	2922	0.2732	98.30%	98.70%	97.20%
Otitis Media with Effusion	1866	467	1.7114	99.20%	95.07%	99.74%
Cholesteatoma	184	45	17.4358	99.77%	95.56%	99.82%
Perforation	195	49	16.3639	99.72%	87.76%	99.87%
Chronic Otitis Media	2035	509	1.5695	98.95%	97.64%	99.14%

Table 18. The 5-fold average performance of the SCH tympanic membrane disease model.

	5-Fold Average Accuracy	5-Fold Average Sensitivity	5-Fold Average Specificity
Normal	96.58%	96.40%	97.08%
Otitis Media with Effusion	98.23%	93.10%	98.90%
Cholesteatoma	99.32%	81.28%	99.53%
Perforation	98.52%	79.45%	98.76%
Chronic Otitis Media	98.04%	94.14%	98.61%
Average	98.14%	88.87%	98.58%

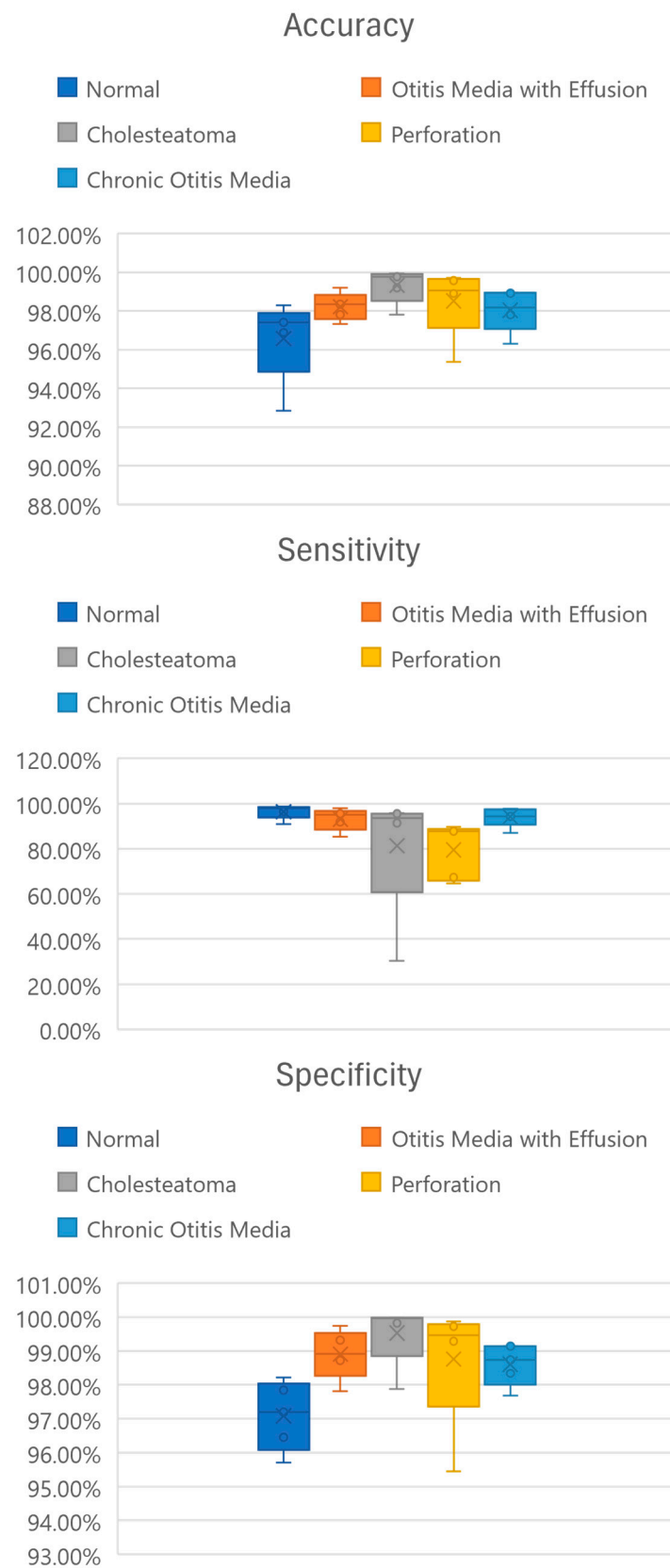


Figure 11. Visualization comparison of box plot results on SCH tympanic membrane disease model.

In the results of the open-access tympanic membrane dataset, normal in Accuracy, otitis externa in Sensitivity, and normal in Specificity had the largest deviation, and in

the results of the SCH tympanic membrane dataset, normal in Accuracy, cholesteatoma in Sensitivity, and perforation in Specificity had the largest deviation. Considering that the error of around 1–3% is generalized, it is difficult to say that the deviation of these classes is completely generalized for each class because it is outside this level, and a plan to overcome this seems necessary for future research.

5. Conclusions

In this study, we proposed the tympanic membrane disease classification and pediatric hearing prediction method of the EfficientNet B7 model using MLP and drop connect. In the process of benchmarking with the open-access tympanic membrane dataset, the proposed model, which fine-tuned the ImageNet weights, showed the best performance with an Average Accuracy of 93.59%, an Average Sensitivity of 87.19, and an Average Specificity of 95.73%. This contrasts with the lower performance of the vanilla EfficientNet B7 than ConvNeXt. In the case of the SCH tympanic membrane dataset, which fine-tuned the noisy student weights, the tympanic membrane disease model showed an Average Accuracy of 98.28%, an Average Sensitivity of 89.66%, an Average Specificity of 98.68%, and an average inference time of 0.2, and the pediatric hearing model showed a Mean Absolute Error of 6.9801, a Mean Squared Logarithmic Error of 0.2887, and an average inference time of 0.2 s.

Future research will try to find ways, such as data augmentation through GAN, a generative artificial intelligence model, or unseen data training through teacher and student model training, e.g., noisy student, to overcome the performance degradation caused by this data imbalance. In addition, we will study how to train tympanic membrane disease and a more diverse dB range of pediatric hearing data not covered in this study and study the structure of a more improved model.

Author Contributions: Conceptualization and supervision, S.C. and W.J.; data curation, methodology, and writing—original draft preparation, H.L. and H.J.; formal analysis, and writing review and editing, S.C. and W.J.; methodology and writing—review and editing, H.L. and H.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Korea Technology & Information Promotion Agency for SMEs (Project Number: 1425165869) and the Soonchunhyang Research Fund.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of Soonchunhyang University Hospital (22 February 2020).

Informed Consent Statement: Patient consent was waived due to the retrospective design of this study.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: Authors Hongchang Lee and Hyeonung Jang were employed by the company Haewootech Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Kubba, H.; Pearson, J.P.; Birchall, J.P. The aetiology of otitis media with effusion: A review. *Clin. Otolaryngol. Allied Sci.* **2000**, *25*, 181–194. [[CrossRef](#)] [[PubMed](#)]
2. 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. Clinical practice guideline: Otitis media with effusion (update). *Otolaryngol. Head Neck Surg.* **2016**, *154*, S1–S41. [[CrossRef](#)] [[PubMed](#)]
3. Vanneste, P.; Page, C. Otitis media with effusion in children: Pathophysiology, diagnosis, and treatment. A review. *J. Otol.* **2019**, *14*, 33–39. [[CrossRef](#)] [[PubMed](#)]
4. Minovi, A.; Dazert, S. Diseases of the middle ear in childhood. *GMS Curr. Top. Otorhinolaryngol. Head Neck Surg.* **2014**, *13*, Doc11. [[PubMed](#)]
5. Zielhuis, G.; Rach, G.; Van Den, B.P. Screening for otitis media with effusion in preschool children. *Lancet* **1989**, *333*, 311–314. [[CrossRef](#)] [[PubMed](#)]

6. Maw, A.R.; Bawden, R. Tympanic membrane atrophy, scarring, atelectasis and attic retraction in persistent, untreated otitis media with effusion and following ventilation tube insertion. *Int. J. Pediatr. Otorhinolaryngol.* **1994**, *30*, 189–204. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Tos, M.; Stangerup, S.E.; Holm-Jensen, S.; Sørensen, C.H. Spontaneous course of secretory otitis and changes of the eardrum. *Arch. Otolaryngol.* **1984**, *110*, 281–289. [\[CrossRef\]](#)
8. Hinton, G.E.; Osindero, S.; Teh, Y.W. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Bengio, Y.; Lamblin, P.; Popovici, D.; Larochelle, H. Greedy layer-wise training of deep networks. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 3–8 December 2007; pp. 153–160.
10. McCulloch, W.S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **1943**, *5*, 115–133. [\[CrossRef\]](#)
11. Rong, G.; Mendez, A.; Assi, E.B.; Zhao, B.; Sawan, M. Artificial intelligence in healthcare: Review and prediction case studies. *Engineering* **2020**, *6*, 291–301. [\[CrossRef\]](#)
12. Ngombu, S.; Binol, H.; Gurcan, M.N.; Moberly, A.C. Advances in Artificial Intelligence to Diagnose Otitis Media: State of the Art Review. *Otolaryngol. Head Neck Surg.* **2022**, *168*, 635–642. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Pichichero, M.E.; Poole, M.D. Assessing diagnostic accuracy and tympanocentesis skills in the management of otitis media. *Arch. Pediatr. Adolesc. Med.* **2001**, *155*, 1137–1142. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Monroy, G.L.; Won, J.; Dsouza, R.; Pande, P.; Hill, M.C.; Porter, R.G.; Novak, M.A.; Spillman, D.R.; Boppart, S.A. Automated classification platform for the identification of otitis media using optical coherence tomography. *NPJ Digit. Med.* **2019**, *2*, 22. [\[CrossRef\]](#) [\[PubMed\]](#)
15. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1989**, *1*, 541–551. [\[CrossRef\]](#)
16. McClelland, J.L.; Rumelhart, D.E.; Hinton, G.E. *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*; MIT Press: Cambridge, MA, USA, 1986; Volume 1, pp. 318–362.
17. Song, D.; Song, I.S.; Kim, J.; Choi, J.; Lee, Y. Semantic decomposition and anomaly detection of tympanic membrane endoscopic images. *Appl. Sci.* **2022**, *12*, 11677. [\[CrossRef\]](#)
18. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. In Proceedings of the 2nd International Conference on Learning Representations ICLR 2014 Conference Track Proceedings, Banff, AB, Canada, 14–16 April 2014.
19. Yue, Y.; Zeng, X.; Shi, X.; Zhang, M.; Zhang, F.; Liu, Y.; Li, Z.; Li, Y. Ear-keeper: Real-time diagnosis of ear lesions utilizing ultralight-ultrafast convnet and large-scale ear endoscopic dataset. *arXiv* **2023**, arXiv:2308.10610.
20. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. ShuffleNet V2: Practical guidelines for efficient CNN architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.
21. Zeng, X.; Jiang, Z.; Luo, W.; Li, H.; Li, G.; Shi, J.; Wu, K.; Liu, T.; Lin, X.; Wang, F.; et al. Efficient and accurate identification of ear diseases using an ensemble deep learning model. *Sci. Rep.* **2021**, *11*, 10839. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Ming, J.; Yi, B.; Zhang, Y.; Li, H. Low-dose CT image denoising using classification densely connected residual network. *KSII Trans. Internet Inf. Syst. TIIS* **2020**, *14*, 2480–2496.
23. Open-Access Tympanic Membrane Dataset. Available online: <https://www.kaggle.com/datasets/erdalbasaran/eardrum-dataset-otitis-media> (accessed on 3 January 2024).
24. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
25. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
26. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 21–24 June 2022; pp. 11976–11986.
27. Müller, R.; Kornblith, S.; Hinton, G.E. When does label smoothing help? In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019.
28. Srivastava, N.; Hinton, G.E.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
29. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
30. Wan, L.; Zeiler, M.; Zhang, S.; LeCun, Y.; Fergus, R. Regularization of neural networks using Drop Connect. In Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013; pp. 1058–1066.
31. Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q.V. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 702–703.
32. Lim, S.; Kim, I.; Kim, T.; Kim, C.; Kim, S. Fast AutoAugment. In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019.

33. Zhuang, J.; Tang, T.; Ding, Y.; Tatikonda, S.; Dvornek, N.; Papademetris, X.; Duncan, J.S. AdaBelief optimizer: Adapting stepsizes by the belief in observed gradients. In Proceedings of the Annual Conference on Neural Information Processing Systems 2020 (NIPS 2020), Virtual, 6–12 December 2020.
34. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
35. Keskar, N.S.; Socher, R. Improving generalization performance by switching from Adam to SGD. *arXiv* **2017**, arXiv:1712.07628.
36. Nishikawa, S.; Yamada, I. Studio Ousia at the NTCIR-15 SHINRA2020-ML Task. In Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo, Japan, 8–11 December 2020.
37. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015.
38. Huber, P.J. Robust estimation of a location parameter. *Ann. Mathematical Stat.* **1964**, *35*, 73–101. [[CrossRef](#)]
39. Xie, Q.; Luong, M.T.; Hovy, E.; Le, Q.V. Self-training with noisy student improves imagenet classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 May 2020; pp. 10687–10698.
40. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.