IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 October 2023, accepted 13 October 2023, date of publication 17 October 2023, date of current version 26 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3325346

RESEARCH ARTICLE

Toward Better Ear Disease Diagnosis: A Multi-Modal Multi-Fusion Model Using Endoscopic Images of the Tympanic Membrane and Pure-Tone Audiometry

TAEWAN KIM[1], SANGYEOP KIM[2], JAEYOUNG KIM[3,4,5], YEONJOON LEE[1], AND JUNE CHOI[2]

[1]Major in Bio Artificial Intelligence, Department of Applied Artificial Intelligence, Hanyang University, Ansan 15588, Republic of Korea
[2]Department of Otorhinolaryngology-Head and Neck Surgery, Ansan Hospital, Korea University College of Medicine, Ansan 15355, Republic of Korea
[3]Core Research and Development Center, Korea University Ansan Hospital, Ansan 15355, Republic of Korea
[4]Department of Dermatology and Skin Sciences, The University of British Columbia, Vancouver, BC V6T 1Z1, Canada
[5]Cancer Control Research Program, BC Cancer, Vancouver, BC V5Z 1L3, Canada

Corresponding authors: Yeonjoon Lee (yeonjoonlee@hanyang.ac.kr) and June Choi (mednlaw@korea.ac.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) funded by the Korean Government for the Ministry of Science and ICT (MSIT), South Korea, through the Artificial Intelligence Convergence Innovation Human Resources Development funded by Hanyang University (ERICA) under Grant RS-2022-00155885; in part by the Artificial Intelligence Convergence Research Center, Hanyang University (ERICA), under Grant 2020-0-01343; in part by the National Research Foundation of Korea (NRF) funded by the Korean Government (MSIT) under Grant NRF-2022R1F1A1074999; in part by the Korea University Grant and the Medical Data-Driven Hospital Support Project through the Korea Health Information Service (KHIS) funded by the Ministry of Health and Welfare, Republic of Korea; and in part by MSIT under the ICT Challenge and Advanced Network (ICAN) of HRD Program Supervised by IITP under Grant IITP-2022-RS-2022-00156439.

ABSTRACT Chronic otitis media is characterized by recurrent infections, leading to serious complications, such as meningitis, facial palsy, and skull base osteomyelitis. Therefore, active treatment based on early diagnosis is essential. This study developed a multi-modal multi-fusion (MMMF) model that automatically diagnoses ear diseases by applying endoscopic images of the tympanic membrane (TM) and pure-tone audiometry (PTA) data to a deep learning model. The primary aim of the proposed MMMF model is adding ''normal with hearing loss'' as a category, and improving the diagnostic accuracy of the conventional four ear diseases: normal, TM perforation, retraction, and cholesteatoma. To this end, the MMMF model was trained on 1,480 endoscopic images of the TM and PTA data to distinguish five ear disease states: normal, TM perforation, retraction, cholesteatoma, and normal (hearing loss). It employs a feature fusion strategy of cross-attention, concatenation, and gated multi-modal units in a multi-modal architecture encompassing a convolutional neural network (CNN) and multi-layer perceptron. We expanded the classification capability to include an additional category, normal (hearing loss), thereby enhancing the diagnostic performance of extant ear disease classification. The MMMF model demonstrated superior performance when implemented with EfficientNet-B7, achieving 92.9% accuracy and 90.9% recall, thereby outpacing the existing feature fusion methods. In addition, five-fold cross-validation experiments were conducted, in which the model consistently demonstrated robust performance when endoscopic images of the TM and PTA data were applied to the deep learning model across all datasets. The proposed MMMF model is the first to include a category of normal ear disease state with hearing loss. The developed model demonstrated superior performance compared to existing CNN models and feature fusion methods. Consequently, this study substantiates the utility of simultaneously applying PTA data and endoscopic images of the TM for the automated diagnosis of ear diseases in clinical settings and validates the usefulness of the multi-fusion method.

INDEX TERMS Artificial intelligence, biomedical imaging, classification algorithms, computer aided diagnosis, convolutional neural networks, deep learning, electronic medical records.

The associate editor coordinating the review of this manuscript and approving it for publication was Sangsoon Lim.

I. INTRODUCTION

Chronic otitis media (COM) with or without cholesteatoma is a significant public health issue affecting 0.5%–30% of the

population and can lead to severe complications due to its characteristic recurrent infections [1]. Cholesteatoma causes include congenital or chronic ear infections and even trauma. Additionally, cholesteatoma can have severe consequences, such as hearing loss, facial paralysis, and intracranial complications [2]. Otolaryngology involves various diagnostic methods for ear diseases, such as computed tomography (CT) and endoscopy analysis [3]. However, COM is often difficult to diagnose because it shows various signs and symptoms, such as middle ear inflammation, TM perforation, and retraction [4]. Moreover, because the diagnosis of ear diseases primarily relies on visual data, such as endoscopy or CT images related to the eardrum, the accuracy of this diagnosis may be limited by the clinician's experience [5]. To address these problems, we applied artificial intelligence (AI) models in otolaryngological examinations to increase the accuracy of ear disease diagnosis.

Amidst recent advancements in deep learning technology, numerous studies have applied AI in the medical domain [6]. In otolaryngology, research leveraging deep learning technology to diagnose middle ear diseases has attracted increasing attention [7]. In an earlier study, Shie et al. deployed 865 otorhinolaryngological images in an AdaBoost model and discerned four diagnostic categories based on ear diseases: normal, acute otitis media, otitis media with effusion (OME), and COM, achieving an accuracy rate of 88.06% [8]. Khan et al. utilized 2,484 endoscopic images of the TM in a DenseNet-161 model, classifying ear diseases into three categories: normal, COM with perforation, and OME, and achieved an accuracy of 94.9% [9].

Endoscopic images of the TM are vital for diagnosing ear diseases because they qualitatively offer various visual markers indicative of ear pathologies, such as the color and transparency of the TM and the presence of middle ear effusion [5]. Consequently, the most prevalent method for automatically diagnosing ear diseases involves analyzing endoscopic images of the TM using AI [7], [10], [11]. However, the diagnostic performance of AI decreases when classes such as OME, which visually resembles normal TM, are included [12]. In addition, when solely relying on endoscopic images of the TM and not pure tone audiometry for diagnosing ear diseases, abnormal eardrum images with only subtle visual deviations may lead not only to misdiagnosis but also to disease progression. For example, when subtle otitis media is overlooked, symptoms such as hearing loss and ear fullness may worsen, with the possibility of newly formed chronic otitis media and cholesteatoma [13].

Studies have suggested that ear diseases can be diagnosed using pure-tone audiometry (PTA) data in conjunction with endoscopic images of the TM [1], [14], [15]. PTA measures air and bone conduction. The PTA air conduction threshold is ascertained using headsets or earphones, whereas bone conduction is assessed by vibrating the skull to stimulate the inner ear using a bone vibrator [16]. Both methods determine the patient's decibel threshold for each frequency band [17]. The discrepancy between PTA air and bone conduction thresholds, known as the air-bone gap (ABG), can help differentiate between normal and abnormal conditions of the middle ear [1], [16]. Numerous studies have proposed the fusion of images and electronic health record (EHR) data in the medical field [18]. However, to the best of our knowledge, no previous studies have integrated PTA data with endoscopic images of the TM for AI applications.

Ongoing efforts are being made to overcome the limitations of image-only models by fusing medical images with EHR data [18]. In prior research, Prabhu et al. applied MRI images combined with EHR data to Multi-Modal Deep Learning Models for the classification of Alzheimer's Disease [19]. Jabbour et al. diagnosed acute respiratory failure (ARF) by applying chest X-rays and EHR data to CNN and ANN models, respectively [20]. Additionally, besides EHR data, Kumar et al. analyzed patients' cough sounds, leveraging deep learning models to recognize pulmonary diseases [21]. Our model employs late fusion to integrate the endoscopic images of the TM and PTA data. However, while endoscopic images of the TM are high-dimensional data containing considerable information on ear diseases, PTA data contain significantly fewer details, resulting in an imbalance in the information between the two data types. Using a single-fusion approach may not deliver optimal performance [18]. To address these challenges, we propose a multi-modal multi-fusion (MMMF) model that employs multiple fusion methods rather than relying on a single method.

Cross-attention captures the interplay between two datasets [22]. Ying et al. [23] used a cross-attention method to fuse features from text and image data on social media platforms to detect fake news. Consequently, they surpassed the performance of existing state-of-the-art models. Concatenation is a straightforward feature fusion method that links features. Although simple, this method allows the fusion of different features without compromising the original state of each feature. Hilmizen et al. concatenated the features extracted from CT-scan and X-ray images to diagnose COVID-19 pneumonia, and the performance of this approach was superior to that of other approaches [24]. Gated multimodal units (GMUs) are modules designed to identify intermediate representations based on various feature combinations, enabling the learning of hidden, latent variables by fusing each feature [25]. Arevalo et al. developed a GMU fusion method to classify movie genres by fusing features from movie posters and plot data [25]. They outperformed other fusion methods, including a mixture of expert models. We implemented a multi-fusion method that applies feature fusion techniques, including cross-attention, concatenation, and GMUs.

The primary contributions of this work are as follows:

- We propose an MMMF model that automatically diagnoses ear diseases using semantic information from endoscopic images of the TM and PTA data.

- Our model employs a multi-fusion approach (cross-attention, concatenate, and GMU), incorporating each feature

fusion method rather than depending on a single-feature fusion method to fuse the information extracted from the convolutional neural network (CNN) and multi-layer perceptron (MLP) models.

- We improved the diagnostic performance of ear diseases by fusing information derived from endoscopic images of the TM and PTA data. Furthermore, we expanded the classification capability beyond the typical categories of normal, TM perforation, retraction, and cholesteatoma by introducing an additional class: normal (hearing loss).

- We have verified the efficacy of the multi-fusion method. Moreover, we applied the proposed MMMF model to the endoscopic images of the TM and PTA data, validating its superior diagnostic performance for ear diseases.

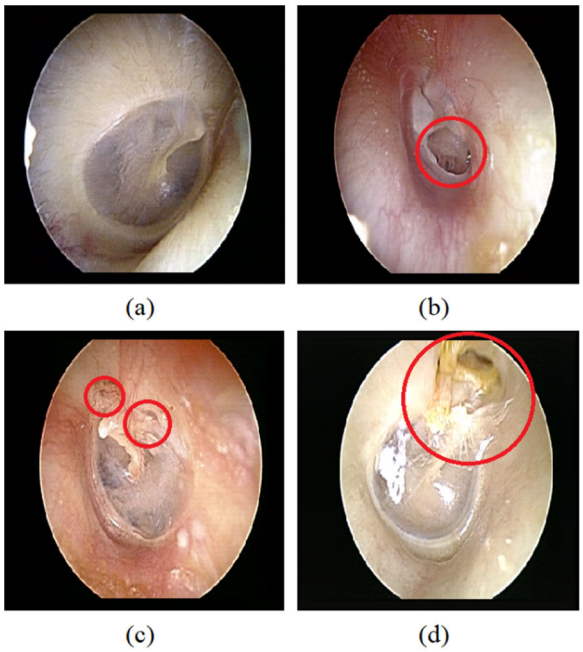

FIGURE 1. Types of ear diseases in the collected dataset. The red circles indicate the locations representing specific characteristics of each disease. (a) Normal TM. (b) Marginal perforation of TM. (c) Attic retraction. (d) Attic destruction with cholesteatoma.

II. MATERIAL AND METHODS

A. PATIENT SELECTION AND DATA ACQUISITION

We collected and analyzed 1632 TM endoscopic images from Korea University Ansan Hospital. Among these, 1,480 endoscopic images of the TM included 330 normal images, 554 images of TM with perforations, 300 retraction images, 159 cholesteatoma images, and 137 images obtained from patients categorized as normal (hearing loss). The remaining 152 images were excluded because they exhibited severe swelling, bleeding, indistinguishable diseases, and overlapping or blurred foci. Furthermore, the TM size, angle, location, rotation, light reflection, and smudging varied across the endoscopic images of the TM; however, these images were analyzed without filters, mirroring real-world clinical situations. The endoscopic photography equipment was replaced midway through the data acquisition process. Consequently, because the image resolutions varied between 1920 × 1080 and 640 × 480, we adjusted the image size to a uniform size of 384 × 384 pixels.

The clinical features of ear diseases are shown in Figure 1, with red circles indicating the location of a specific feature. TM perforation, TM or attic retraction, and cholesteatoma are features observed in individuals with hearing impairment. The presence of TM retraction indirectly shows the patient's Eustachian tube function while suggesting potential cholesteatoma formation in the middle ear. Cholesteatomas can induce symptoms such as hearing loss, otorrhea, vertigo, and headache.

TABLE 1. PTA data characteristics for patients included within the dataset used to train the model.

Characteristic	Frequency	Range
Sex	-	0-1
Age	-	1-80
PTA air conduction		
PTA bone conduction	0.25, 0.5, 1, 2, 3, 4kHz and Avg	−10-120
PTA air-bone gap		
Sexenary average [26]	-	0-120
Hearing loss flag	-	0, 100

The sex of the patients was denoted as 0 for males and 1 for females. Avg denotes the average decibel values at 0.25, 0.5, 1, 2, 3, and 4 kHz. The hearing loss flag was indicated as 100 when the sexenary average value was equal to or greater than 26 or when there was a difference of 11 or more between the air-bone gap (ABG) and PTA-air average values. Otherwise, it was represented as zero.

Table 1 lists the characteristics of the patients from whom PTA data were collected. Previous studies have suggested that hearing levels vary according to sex and age [27]; therefore, we included both factors in our PTA data. We collected PTA decibel thresholds at 0.25, 0.5, 1, 2, 3, and 4 kHz frequencies and computed the average decibel thresholds across the entire frequency band using two testing methods: air and bone conduction. Given that ABG can help differentiate normal and abnormal TM [1], [16], we included ABG in our analysis of patient characteristics. We also included the values calculated using the sexenary average formula to determine hearing loss grades [26]. The sexenary average is calculated as follows:

$$\textit{Sexenary average} = \frac{0.5k + 2 \times 1k + 2 \times 2k + 4k}{6} \quad (1)$$

When endoscopic images of the TM show normal eardrums but the sexenary average exceeds 25 dB, diseases such as sudden sensorineural hearing loss, congenital middle ear anomalies, and otosclerosis may cause hearing loss [28].

In addition, normal eardrums with ABG of 11 or more may indicate otosclerosis, bone anomaly, and inner ear disorders [29]. Therefore, we included a hearing loss flag in our data because a difference of 11 or more between the ABG and the average value of PTA air, or a sexenary average value of 26 or more, significantly increases the likelihood of hearing loss. Consequently, we constructed one-dimensional data with a length of 25. The patients' ages ranged from 0–80, and the decibel levels varied between −10 and 120 dB, a range. In addition, we classified the PTA dataset into 330 normal and 1150 abnormal cases (perforation, retraction, cholesteatoma, and normal (hearing loss)) based on the analysis of endoscopic images of the TM.

The endoscopic images of the TM and PTA data were randomly divided into five distinct datasets, each representing 20% of the total data per disease category, with no overlap. Four datasets, comprising 80% of the images (1184), were employed for training, whereas the remaining dataset, containing 20% of the images (296), was used for validation. In addition, all procedures in this study were performed following the rules of the 1975 Helsinki Declaration, and the use of the data was approved by the IRB (2021AS0329) of Korea University Ansan Hospital. The ethics committee waived informed consent because of the retrospective nature of the study.

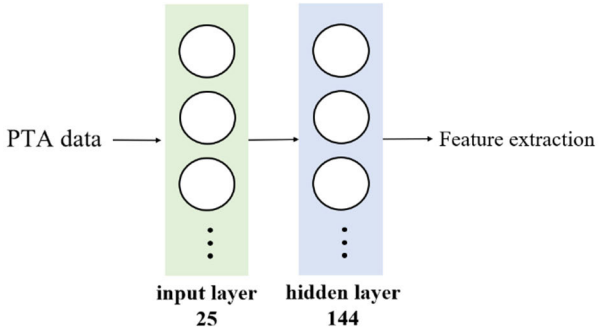

FIGURE 2. MLP architecture for extracting meaningful features from PTA data.

B. MLP MODEL FOR EXTRACTING SEMANTIC INFORMATION FROM PTA DATA

Images are high-dimensional data containing a wealth of information [30]. Consequently, they can accurately identify the properties of normal, perforation, retraction, and cholesteatoma-affected eardrums. While one-dimensional PTA data can classify eardrums into normal and abnormal categories, identifying the specific characteristics of retraction, perforations, and cholesteatoma using PTA data presents a challenge. Hence, we applied PTA data to an MLP to develop a simple PTA model that extracts information regarding normal and abnormal TM states. The architecture of the proposed PTA model is shown in Figure 2. We constructed an MLP comprising an input layer with 25 nodes and a hidden layer with 144 nodes.

C. CNN MODELS FOR EXTRACTING SEMANTIC INFORMATION FROM ENDOSCOPIC IMAGES OF THE TM

We extracted endoscopic image information using a pre-trained public CNN model validated using the ImageNet database. The CNN model was trained to classify images into 1,000 categories. Therefore, the ImageNet CNN model included a fully connected (FC) layer of 1,000 nodes. As we used the CNN model solely to extract image features, we employed it with the FC layer removed.

D. MMMF MODELS FOR THE AUTOMATIC DIAGNOSIS OF EAR DISEASES USING ENDOSCOPIC IMAGES OF THE TM AND PTA DATA

The architecture of the MMMF model is shown in Figure 3. First, the MLP model was applied to PTA data to extract MLP features related to hearing. In addition, endoscopic images of the TM were applied to the CNN model to extract the features of the TM. The CNN features were extracted both before and after average pooling. The CNN feature map before Avg pooling was reshaped into channel × (width × height) images. The MLP feature map was then expanded by the number of channels in the CNN feature map to obtain CNN and MLP feature maps of the same size. Cross-attention was used to generate the CA features to fully integrate the information from the endoscopic images of the TM and PTA data. To preserve feature information from both models, the CNN feature map after Avg pooling was concatenated with the feature map extracted from the MLP, resulting in CC features. Finally, the GMU [25] module fused the CA and CC features to produce multi-fusion features. An FC layer was then used to classify the multi-fusion data into the following classes: normal, perforation, retraction, cholesteatoma, and normal (hearing loss).

E. MULTI-FUSION METHOD

First, we applied the cross-attention mechanism in our model to mix information from the endoscopic images of the TM and PTA data. The cross-attention structure we used aligned with the scaled dot-product attention structure proposed in the transformer [31]. To learn the correlation between the information in endoscopic images of the TM and PTA data, each feature sequence should be utilized as three variables: query, key, and value. The CNN feature map was employed, which contains information from endoscopic images of the TM as the query, and the MLP feature map, which contains PTA data information, as the key and value.

We computed the correlation between the endoscopic images of the TM and the PTA data using dot products of the query and key. By scaling this with the softmax function, we derived the attention weights for both datasets. These weights were then multiplied with the value to produce the CA feature, used for the GMU module. The operation process for cross-attention is as follows:

$$Attension(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \qquad (2)$$

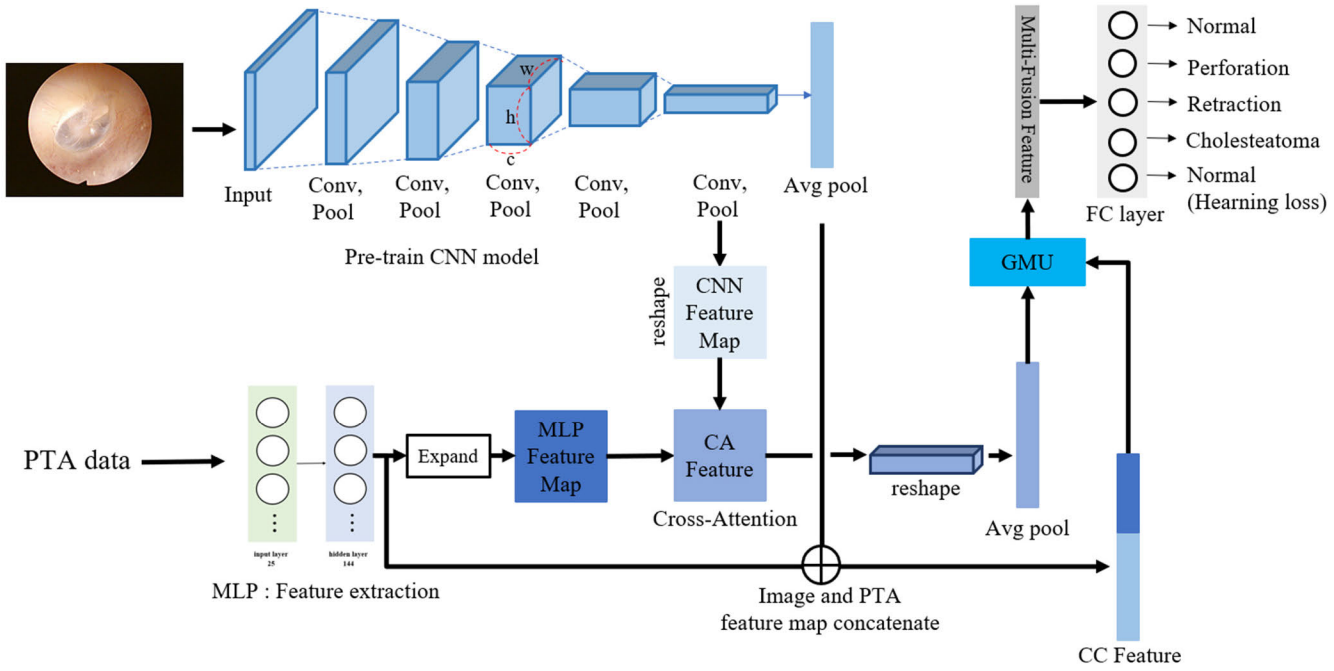

FIGURE 3. MMMF model architecture. Information extracted from endoscopic images of the TM and PTA data through CNN and MLP is integrated using the cross-attention, concatenate, and GMU fusion methods. Following fusion, the combined features are forwarded through the FC layer to classify five ear diseases (normal, perforation, retraction, cholesteatoma, and normal (hearing loss)). PTA, Pure-tone audiometry. CNN, Convolutional neural networks. MLP, Multi-layer perceptron. CA, Cross-attention. CC, Concatenate. GMU, Gated multimodal units. Avg, Average. FC layer, Fully connected layer. c, Channel size. w, Width size. h, Height size.

where Q represents the query, K represents the key, and V represents the value, while d_k signifies the dimensions of the queries and keys. The dot products of the query and key are computed, each divided by $\sqrt{d_k}$, and the softmax function is then applied to obtain the weights for the values.

Second, the concatenation method was used to focus on the interactions between the features extracted from the CNN and MLP, and these features were fused without damaging their original state. The structure of the GMU module is shown in Figure 4. We used the GMU module to extract a multi-fusion feature that learned the intermediate representation and hidden, latent variables of the cross-attention-fused CA feature and CC feature. This CA feature contained information about the interrelation between the two datasets, and the CC feature retained the original forms of the endoscopic images of the TM and PTA data.

Within the GMU module, our first step extract hidden features from the CA features and CC features. Subsequently, employing the concatenation operator and the sigmoid activation function, we derive the z activation function from the two features. Then utilized z activation function and the hidden features of CA features and CC features, resulting in a multi-fusion feature used for ear disease classification. The GMU operational process is as follows:

$$h_v = \tanh(W_v \cdot x_v) \tag{3}$$

$$h_t = \tanh(W_t \cdot x_t) \tag{4}$$

$$z = \sigma(W_z \cdot [x_v, x_t]) \tag{5}$$

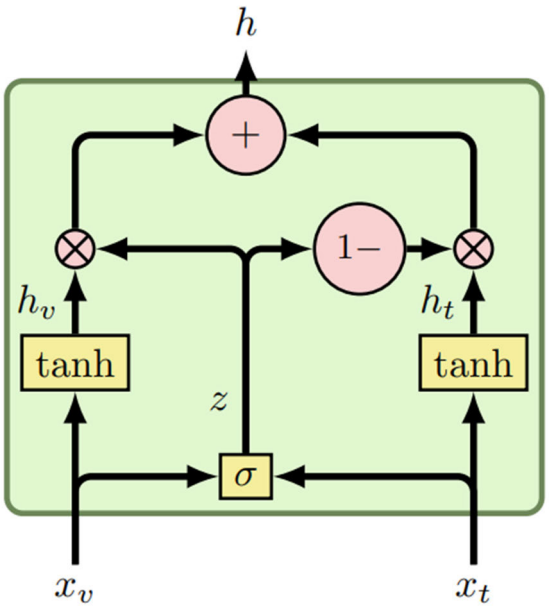

FIGURE 4. GMU architecture [25].

$$h = \mathrm{z} \times \mathrm{h}_v + (1 - z) \times h_t \tag{6}$$

$$\theta = \{W_v, W_t, W_z \tag{7}$$

where θ represents the parameter to be learned, and $[\cdot, \cdot]$ denotes the concatenation operator; both operations are

differentiable. This fusion method can be easily integrated with other neural network architectures and trained using stochastic gradient descent.

F. TRAINING DETAILS

We evaluated the usefulness of the features extracted from the MLP model described in Figure 2, in which we added an FC layer with two output nodes to the MLP model. To extract features from endoscopic images of the TM, we compared seven CNN models from ImageNet. The models include Vgg-19 [32], ResNet-152 [33], GoogleNet [34], DenseNet-161 [35], Inception-V3 [36], Inception+ResNet-V2 [37], and EfficientNet-B7 [38]. In addition, we added an FC layer with four output nodes to each CNN model. Following this, we adopted the CNN model that exhibited the best performance and compared and evaluated our proposed MMMF model with conventional feature fusion methods. All models were trained using a batch size 16, a learning rate 1e−4, an Adam optimizer, and a cross-entropy loss function. Moreover, owing to a class imbalance in the data, loss weight was applied by calculating the ratio of the number of data points for each disease. All experiments in this study were conducted on a deep learning server equipped with eight NVIDIA GeForce RTX 3080 12GB graphic processing units.

G. EVALUATION PROTOCOLS

We evaluated the performances of all the models using accuracy and recall indicators. Accuracy, which represents the proportion of correctly predicted data across an entire dataset, is the most commonly used performance metric. Recall indicates the proportion of data correctly predicted to belong to the actual class from the entire dataset of that class. The formulas for these evaluation metrics are as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

where TP, TN, FP, and FN represent the true positives, true negatives, false positives, and false negatives, respectively. The higher the value of each metric, the better the classification performance.

III. RESULTS

A. PTA MODEL FEATURE EXTRACTION PERFORMANCE

As depicted in Figure 3, to implement cross-attention between the semantic information from endoscopic images of the TM and PTA, a feature map of the same size is required. Therefore, we generated a feature map with a length of 144, corresponding to one channel size of the CNN feature map, to represent the semantic information of PTA data. In addition, it is essential to ascertain whether the features of PTA data can distinguish between the normal and abnormal states of the eardrum. We trained the model by adding an FC layer with two output nodes to the MLP model, as shown in Figure 2. The MLP model demonstrated an accuracy of 90.5%, recall of 94.0%, and loss of 0.221. This indicates that when the PTA data were applied to our proposed MLP model, the features of the normal and abnormal eardrum conditions were effectively differentiated.

B. CNN MODEL CLASSIFICATION PERFORMANCE

The performances of seven CNN models pre-trained on ImageNet were compared for use in our MMMF model. Our endoscopic images of the TM showed the characteristics of four ear diseases: normal, perforation, retraction, and cholesteatoma. Therefore, the CNN models were classified into four classes. Table 2 compares the results. The CNN models displayed an average accuracy of 89.9%, recall of 85.9%, and loss of 0.565 for identifying ear diseases. Furthermore, the EfficientNet-B7 model, which exhibited superior performance, showed an accuracy of 91.5% and a recall of 87.6%, indicating that applying endoscopic images of the TM to the CNN model enables the precise extraction of features corresponding to the four ear diseases. Additionally, we adopted EfficientNet-B7 for the proposed MMMF model.

TABLE 2. Performance comparison of CNN models.

Model	Accuracy	Recall	Loss
Vgg19-bn	86.1	79.9	0.752
ResNet-152	89.9	86.7	0.375
Inception-resNet-V2	90.8	87.5	0.436
Inception-V3	90.2	87.4	0.763
GoogleNet	90.5	87.0	0.379
EfficienNet-B7	91.5	87.6	0.727
DenseNet-161	90.2	85.2	0.517
Average	89.9	85.9	0.565

The models were trained to classify four types of ear conditions: normal, perforation, retraction, and cholesteatoma.

C. MULTI-FUSION AND SINGLE-FUSION COMPARISON RESULTS

We extended the classification of ear diseases into five categories: normal, perforation, retraction, cholesteatoma, and normal (hearing loss). The latter was identified using features of abnormal hearing extracted from the PTA data in conjunction with features of the normal class extracted from endoscopic images of the TM. Table 3 compares the performance of the proposed MMMF model with that of conventional single-feature fusion methods. Our model exhibited superior performance with an accuracy of 92.9%, recall of 90.9%, and loss of 0.671. Our proposed model also outperformed the EfficientNet-B7 model, which was trained to classify four classes, despite adding a fifth class: normal (hearing loss). Figure 5 shows cases of improved diagnosis of ear diseases

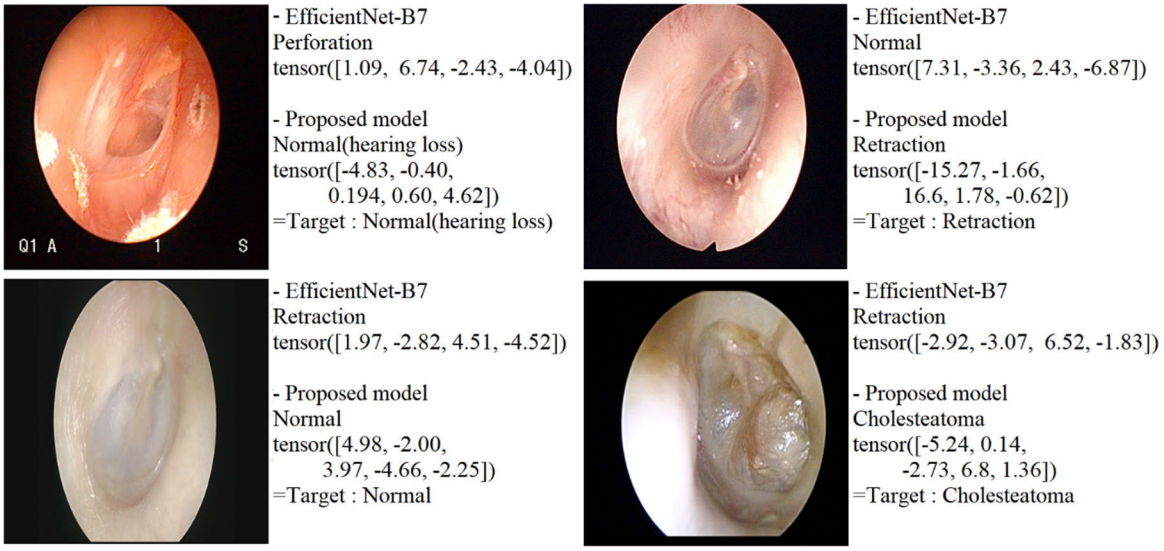

FIGURE 5. Examples of improved misclassification results of the EfficientNet-B7 model when using the proposed model. Tensor is the predicted value for each class by the model. For the EfficientNet-B7 model, the order is normal, perforation, retraction, and cholesteatoma. The order of the proposed model is normal, perforation, retraction, cholesteatoma, and normal (hearing loss). The target is ground truth.

using our proposed MMMF model compared to using only endoscopic images of the TM in the EfficientNet-B7 model. This proves that our model not only effectively classifies the additional normal (hearing loss) class but also enhances the diagnostic performance of the original four types of ear diseases.

TABLE 3. Performance comparison between the proposed MMMF model and conventional single-feature fusion methods.

Fusion methods	Accuracy	Recall	Loss
Concatenate	83.1	75.8	1.048
Linear Sum	84.1	76.5	1.167
Cross-Attention	79.7	67.9	1.575
Gated multimodal units (GMU)	86.4	81.2	0.801
Proposed	92.9	90.9	0.671

The models were trained to classify five types of ear conditions: normal, perforation, retraction, cholesteatoma, and normal (hearing loss).

D. MMMF MODEL FIVE-FOLD CROSS-VALIDATION RESULT

Table 4 summarizes the results of the five-fold cross-validation of the proposed MMMF model. Our model demonstrated consistent performance across all datasets, achieving an average accuracy of 89.7%, recall of 87.2%, and loss of 0.852. Confusion matrices depict the rate of correct predictions for each class, as well as the proportions at which certain classes are mistakenly identified as others. Within these matrices, the diagonal elements represent the correct prediction rate for each specific class. Aggregating the correct prediction rates from all classes results in the overall accuracy value. Furthermore, elements outside the diagonal provide insights into which specific classes were commonly misclassified. Figure 6 shows the confusion matrix results of all datasets for our proposed model, demonstrating the performance of our model in accurately classifying all five types of ear diseases across all datasets. This confirms the capability of the proposed MMMF model to accurately categorize the added normal (hearing loss) class across all datasets.

TABLE 4. Five-fold cross-validation performance results of the proposed MMMF model.

Dataset	Accuracy	Recall	Loss
Dataset1	92.9	90.9	0.671
Dataset2	89.1	87.4	1.016
Dataset3	88.8	87.0	1.017
Dataset4	91.5	89.5	0.622
Dataset5	86.1	81.4	0.932
Average	89.7	87.2	0.852

Each validation dataset was divided without overlap and comprised 20% of the data for each ear condition.

E. GRAD-CAM ANALYSIS

Grad-CAM is a common method for visualizing the areas in an image that the CNN model focuses on during classification. Figure 7 shows the Grad-CAM outputs for each disease type for the MMMF and EfficientNet-B7 models. EfficientNet-B7, when integrated into the proposed MMMF model, displays a heat map of the precise eardrum and affected area locations, resembling the findings from the standalone EfficientNet-B7 model. This demonstrates that the EfficientNet-B7 model incorporated into the MMMF

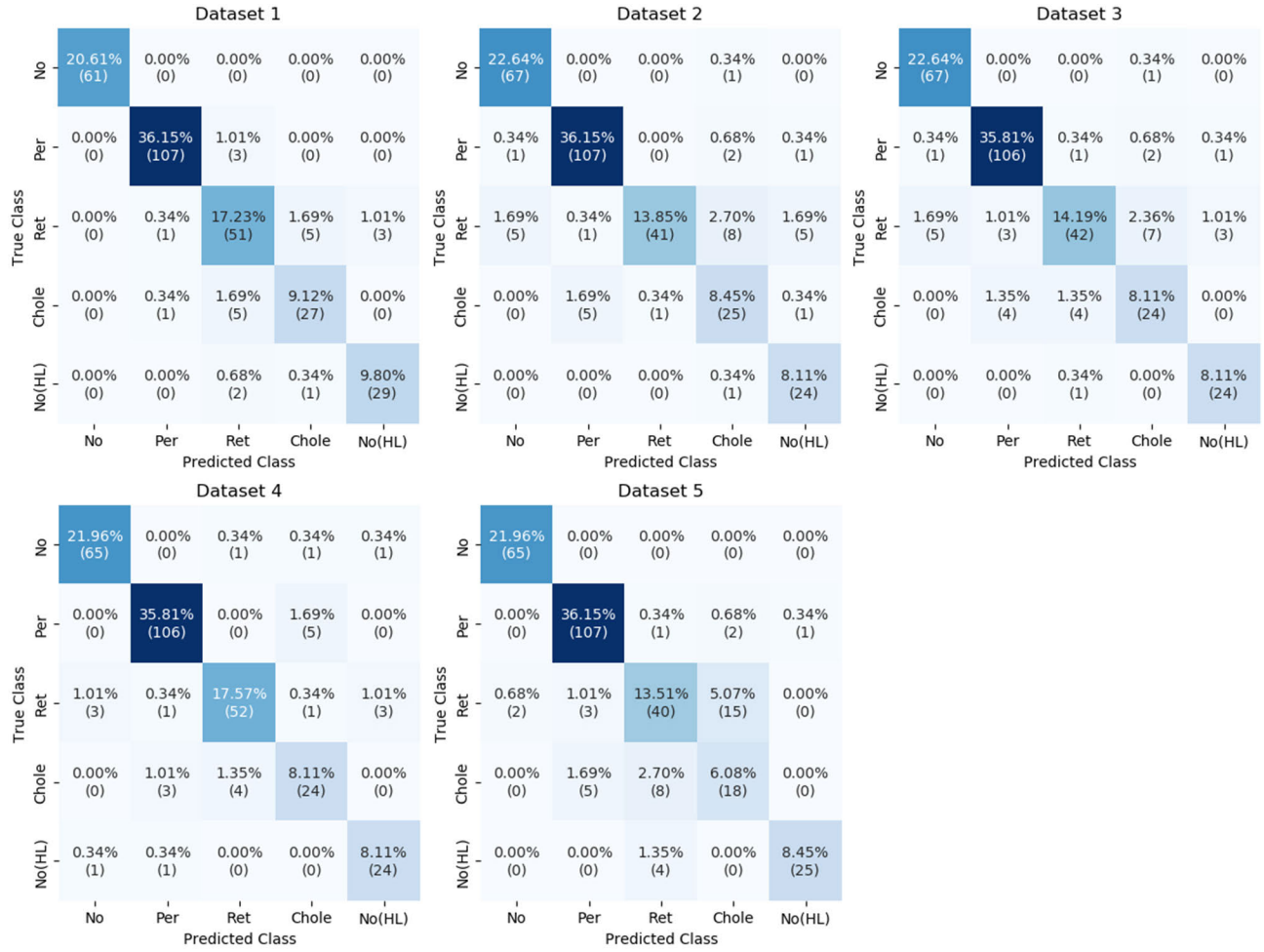

FIGURE 6. Confusion matrix for each dataset of a MMMF model. No, normal. Per, perforation. Ret, retraction. Chole, cholesteatoma. No (HL), normal (hearing loss).

model classifies ear diseases by focusing on exact eardrum locations.

IV. DISCUSSION

The diagnosis of ear diseases predominantly depends on ear examinations and clinician expertise [3], [5]. Moreover, diagnostic accuracy can vary based on the clinician's experience, as the primary basis of diagnosis lies in visual data, such as endoscopic, CT, and MRI images [5]. Studies have indicated that the average diagnostic accuracy of pediatricians and otolaryngologists is <70%, indicating that their diagnostic precision is not high [39]. Hence, incorporating AI technology into otolaryngology could assist in making objective judgments when diagnosing ear diseases. Studies on the automated diagnosis of ear diseases suggest that deep learning models based on endoscopic images of the TM can considerably assist physicians [40]. Most AI research in otolaryngology utilizes eardrum images [7], [10], [11]. However, if only endoscopic images of the TM are used in a deep-learning model, there is a risk of misdiagnosing ear diseases that appear similar to a normal eardrum [12]. In previous research, the fusion of medical images and EHR has been employed to overcome the challenge of diagnosing complex clinical situations using only medical images [18]. In otolaryngology, eardrum images and PTA data can be used to diagnose ear diseases [1], [14], [15]. Therefore, applying PTA data and endoscopic images of the TM to our deep learning model, we solved problems such as the misdiagnosis of images similar to normal eardrums or patients with hearing problems who exhibited normal eardrum images, as shown in Figure 5. In addition, our MMMF model achieved superior diagnostic performance compared with traditional deep learning models that only use conventional feature fusion methods and endoscopic images of the TM.

Our proposed MMMF model, using 1,480 endoscopic images of the TM and PTA data, automated the diagnosis of five ear diseases with an accuracy of 92.9% and a recall of 90.9%. These results demonstrate improvements of 1.4% and 3.3% over those of the CNN models that used only endoscopic images of the TM for diagnosing the four ear diseases. In our proposed model, the EfficientNet-B7 model accurately classified specific ear diseases by pinpointing the exact location of the eardrum and affected area in the otoendoscopic image. The PTA model accurately classified a patient's

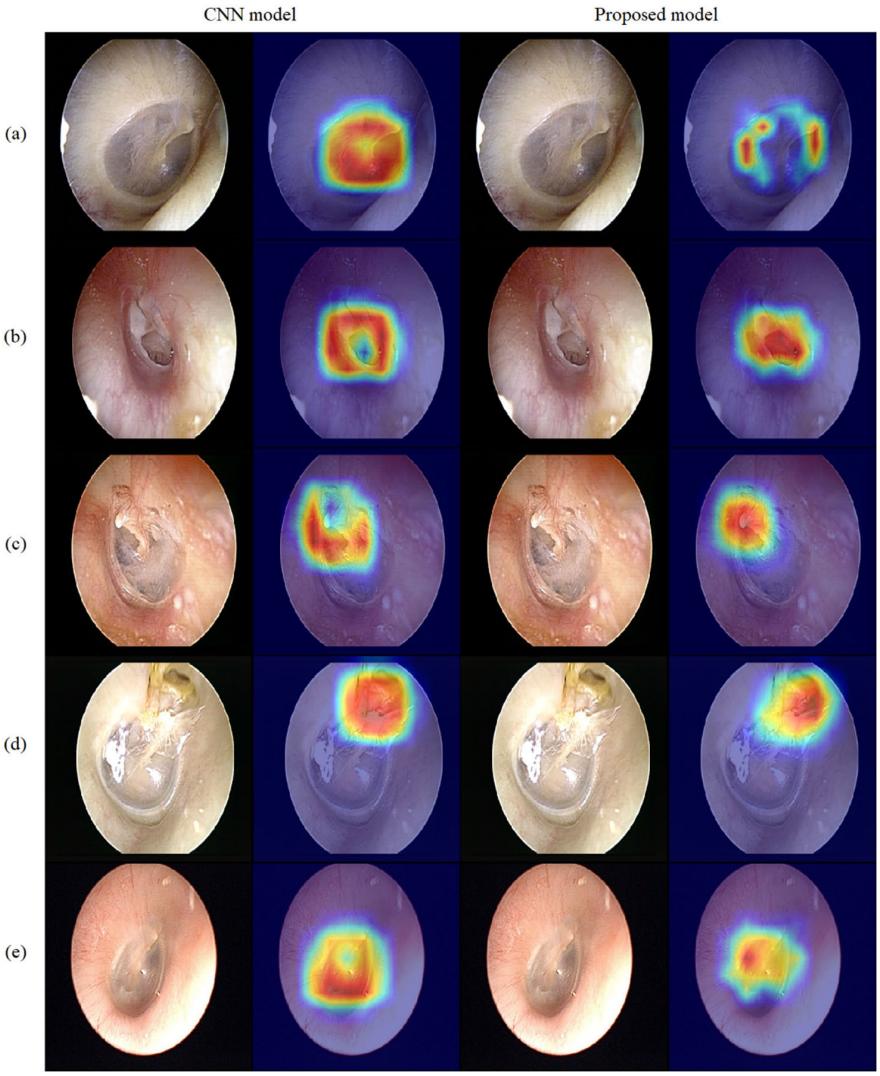

FIGURE 7. Comparison of heat maps of the Grad-CAM of the EfficientNet-B7 and MMMF models for each disease condition. The closer the color is to red, the greater the influence on the model's classification of ear diseases. Heat maps for the (a) normal, (b) perforation, (c) retraction, (d) cholesteatoma, and (e) normal (hearing loss) classes.

eardrum condition as normal or abnormal. Therefore, we not only classified an additional normal (hearing loss) class by merging the CNN and PTA models but also enhanced the performance of automatic ear disease diagnosis. Furthermore, our proposed method represents the first attempt to apply a multi-fusion method in otolaryngology and is also an inaugural study to classify the normal (hearing loss) class of ear disease. This validates the capabilities of our MMMF model and multi-fusion method to diagnose a broader spectrum of ear diseases more efficiently than previous diagnostic strategies. Moreover, it validates the enhanced utility of concurrently using endoscopic images of the TM and PTA data over solely relying on endoscopic images of the TM in deep learning models.

The medical significance of this study first, lies in the fact that it is a study of this type in developing countries, where otologists are not readily available for patients' hospital care. Using both PTA and endoscopic images of the TM to increase diagnostic rates may also be employed in telemedicine in developing countries, while endoscopic images of the TM and PTA data may be sent for analysis and diagnosis in a cheaper and cost-efficient manner. Moreover, physicians with endoscopic systems but different specialties, such as pediatrics, internal medicine, or family medicine, may examine features that signal chronic ear diseases, such as attic destruction or minimal TM perforations. The automatic diagnosis of chronic otitis media may assist medical doctors in diagnosing patients with such features. Finally, the current study not only enables the discrimination of normal TM from hearing loss but also facilitates the diagnosis of the aforementioned diseases in clinical situations.

However, our proposed method has the limitation of not implementing data augmentation. Moreover, despite the concurrent use of endoscopic images of the TM and PTA data in our proposed MMMF model, it only classifies five ear disease categories. Previous studies indicated that increased amounts

of data could enhance the accuracy of automatic diagnostic systems for ear diseases [40]. Nevertheless, patient medical data collection is challenging owing to various regulations such as the Privacy Act [41]. Hence, previous otolaryngology studies have used data augmentation techniques such as image rotation and flipping [9], [40], [42]. In our approach, we similarly augmented the eardrum endoscopy data. However, for the PTA data, data augmentation generates duplicate PTA data, equivalent to the number of augmented eardrum endoscopic images. Therefore, we could not perform data augmentation. Despite this, our study demonstrated excellent performance with a relatively small dataset of 1,480 samples. If we manage to amass a larger dataset consisting of more eardrum endoscopy images and PTA data, our proposed MMMF model can potentially be used to diagnose more than six classes. Therefore, future studies should aim to collect more data and conduct tests to diagnose various ear diseases. Furthermore, a more detailed exploration of the relationship between the endoscopic images of the TM and PTA data, or the adoption of state-of-the-art technologies, might prove beneficial in classifying not just the 'normal with hearing loss' category but also in introducing new categories for diagnosis.

V. CONCLUSION

In this study, we developed an MMMF model for automatically diagnosing five ear disease classes: normal, perforation, retraction, cholesteatoma, and normal (hearing loss), employing endoscopic images of the TM and PTA data. Our model demonstrated the best performance when EfficientNet-B7 was applied, with an accuracy of 92.9% and recall of 90.9%. Furthermore, the proposed multi-fusion method exhibited superior performance over the single-fusion method, and our model demonstrated excellent results across all datasets in a five-fold cross-validation. Despite using a feature fusion method, our model categorizes ear diseases by referring to the precise locations of the eardrum and affected areas in the endoscopic images of the TM. Thus, the proposed model outperformed traditional single-feature fusion methods, and CNN models that solely utilize endoscopic images of the TM. Considering that our PTA model was integrated with a conventional CNN model in this study, the additional classification of the normal (hearing loss) class could be attributed to the PTA model. Furthermore, considering the application of the multi-fusion method to multi-modal data, the enhanced performance of ear disease diagnosis can be attributed to the multi-fusion method. Consequently, the proposed method demonstrates that deep learning models can leverage new semantic information from eardrum endoscopic images of the TM and PTA data to diagnose complex ear diseases further, thereby achieving high diagnostic performance. This indicates the potential benefits for future clinical scenarios, such as telemedicine and diagnostic support systems.

ACKNOWLEDGMENT

(Taewan Kim and Sangyeop Kim are co-first authors.)

REFERENCES

[1] M. Luntz, N. Yehudai, M. Haifler, G. Sigal, and T. Most, "Risk factors for sensorineural hearing loss in chronic otitis media," *Acta Oto-Laryngol.*, vol. 133, no. 11, pp. 1173–1180, Nov. 2013, doi: 10.3109/00016489.2013.814154.

[2] L. Louw, "Acquired cholesteatoma pathogenesis: Stepwise explanations," *J. Laryngol. Otol.*, vol. 124, no. 6, pp. 587–593, Jun. 2010, doi: 10.1017/S0022215109992763.

[3] T. Marom, O. Kraus, N. Habashi, and S. O. Tamir, "Emerging technologies for the diagnosis of otitis media," *Otolaryngol.-Head Neck Surg.*, vol. 160, no. 3, pp. 447–456, Mar. 2019, doi: 10.1177/0194599818809337.

[4] N. S. Tsilis, P. V. Vlastarakos, V. F. Chalkiadakis, D. S. Kotzampasakis, and T. P. Nikolopoulos, "Chronic otitis media in children: An evidence-based guide for diagnosis and management," *Clin. Pediatrics*, vol. 52, no. 9, pp. 795–802, Sep. 2013, doi: 10.1177/0009922813482041.

[5] G. L. Monroy, J. Won, R. Dsouza, P. Pande, M. C. Hill, R. G. Porter, M. A. Novak, D. R. Spillman, and S. A. Boppart, "Automated classification platform for the identification of otitis media using optical coherence tomography," *npj Digit. Med.*, vol. 2, no. 1, p. 22, Mar. 2019, doi: 10.1038/s41746-019-0094-0.

[6] G. Rong, A. Mendez, E. B. Assi, B. Zhao, and M. Sawan, "Artificial intelligence in healthcare: Review and prediction case studies," *Engineering*, vol. 6, no. 3, pp. 291–301, Mar. 2020, doi: 10.1016/j.eng.2019.08.015.

[7] S. Ngombu, H. Binol, M. N. Gurcan, and A. C. Moberly, "Advances in artificial intelligence to diagnose otitis media: State of the art review," *Otolaryngol.-Head Neck Surg.*, vol. 168, no. 4, pp. 635–642, Apr. 2023, doi: 10.1177/01945998221083502.

[8] C.-K. Shie, H.-T. Chang, F.-C. Fan, C.-J. Chen, T.-Y. Fang, and P.-C. Wang, "A hybrid feature-based segmentation and classification system for the computer aided self-diagnosis of otitis media," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2014, pp. 4655–4658, doi: 10.1109/EMBC.2014.6944662.

[9] M. A. Khan, S. Kwon, J. Choo, S. M. Hong, S. H. Kang, I.-H. Park, S. K. Kim, and S. J. Hong, "Automatic detection of tympanic membrane and middle ear infection from oto-endoscopic images via convolutional neural networks," *Neural Netw.*, vol. 126, pp. 384–394, Jun. 2020, doi: 10.1016/j.neunet.2020.03.023.

[10] A. M. Bur, M. Shew, and J. New, "Artificial intelligence for the otolaryngologist: A state of the art review," *Otolaryngol.-Head Neck Surg.*, vol. 160, no. 4, pp. 603–611, Apr. 2019, doi: 10.1177/0194599819827507.

[11] G. Chawdhary and N. Shoman, "Emerging artificial intelligence applications in otological imaging," *Current Opinion Otolaryngol. Head Neck Surg.*, vol. 29, no. 5, pp. 357–364, 2021, doi: 10.1097/MOO.0000000000000754.

[12] A.-R. Habib, G. Crossland, H. Patel, E. Wong, K. Kong, H. Gunasekera, B. Richards, L. Caffery, C. Perry, R. Sacks, A. Kumar, and N. Singh, "An artificial intelligence computer-vision algorithm to triage otoscopic images from Australian aboriginal and torres strait islander children," *Otol. Neurotol.*, vol. 43, no. 4, pp. 481–488, 2022, doi: 10.1097/MAO.0000000000003484.

[13] K. A. Daly, L. L. Hunter, and G. S. Giebink, "Chronic otitis media with effusion," *Pediatrics Rev.*, vol. 20, no. 3, pp. 85–94, Mar. 1999, doi: 10.1542/pir.20-3-85.

[14] J. Roberts, L. Hunter, J. Gravel, R. Rosenfeld, S. Berman, M. Haggard, J. Hall, C. Lannon, D. Moore, L. Vernon-Feagans, and I. Wallace, "Otitis media, hearing loss, and language learning: Controversies and current research," *J. Develop. Behav. Pediatrics*, vol. 25, no. 2, pp. 110–122, Apr. 2004, doi: 10.1097/00004703-200404000-00007.

[15] C. Macandie and B. F. O'Reilly, "Sensorineural hearing loss in chronic otitis media," *Clin. Otolaryngol. Allied Sci.*, vol. 24, no. 3, pp. 220–222, Jun. 1999, doi: 10.1046/j.1365-2273.1999.00237.x.

[16] W. Mohamad, W. Najibah, Romli, Maziah, Awang, M. Almyzan, Lih, A. Cheu, Abdullah, Rosninda, Zakaria, M. Normani, "The presence of unusual bone conduction thresholds in pure tone audiometry," *Indian J. Otol*, vol. 26, no. 1, pp. 54–57, 2020, doi: 10.4103/indianjotol.INDIANJOTOL_99_19.

[17] *Acoustics- Audiometric Test Methods Part 1: Basic Pure Tone Air and Bone Conduction Threshold Audiometry*, ISO, Geneva, Switzerland, 1989, pp. 8251–8253.

[18] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines," *npj Digit. Med.*, vol. 3, no. 1, p. 136, Oct. 2020, doi: 10.1038/s41746-020-00341-z.

[19] S. S. Prabhu, J. A. Berkebile, N. Rajagopalan, R. Yao, W. Shi, F. Giuste, Y. Zhong, J. Sun, and M. D. Wang, "Multi-modal deep learning models for Alzheimer's disease prediction using MRI and EHR," in *Proc. IEEE 22nd Int. Conf. Bioinf. Bioeng. (BIBE)*, Nov. 2022, pp. 168–173, doi: 10.1109/BIBE55377.2022.00044.
[20] S. Jabbour, D. Fouhey, E. Kazerooni, J. Wiens, and M. W. Sjoding, "Combining chest X-rays and electronic health record (EHR) data using machine learning to diagnose acute respiratory failure," *J. Amer. Med. Inform. Assoc.*, vol. 29, no. 6, pp. 1060–1068, May 2022, doi: 10.1093/jamia/ocac030.
[21] A. Kumar, K. Abhishek, C. Chakraborty, and N. Kryvinska, "Deep learning and Internet of Things based lung ailment recognition through coughing spectrograms," *IEEE Access*, vol. 9, pp. 95938–95948, 2021, doi: 10.1109/ACCESS.2021.3094132.
[22] L. Sun, B. Liu, J. Tao, and Z. Lian, "Multimodal cross- and self-attention network for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 4275–4279, doi: 10.1109/ICASSP39728.2021.9414654.
[23] L. Ying, H. Yu, J. Wang, Y. Ji, and S. Qian, "Multi-level multi-modal cross-attention network for fake news detection," *IEEE Access*, vol. 9, pp. 132363–132373, 2021, doi: 10.1109/ACCESS.2021.3114093.
[24] N. Hilmizen, A. Bustamam, and D. Sarwinda, "The multimodal deep learning for diagnosing COVID-19 pneumonia from chest CT-scan and X-ray images," in *Proc. 3rd Int. Seminar Res. Inf. Technol. Intell. Syst. (ISRITI)*, Dec. 2020, pp. 26–31, doi: 10.1109/ISRITI51436.2020.9315478.
[25] J. Arevalo, T. Solorio, M. Montes-y-Gómez, and F. A. González, "Gated multimodal units for information fusion," 2017, *arXiv:1702.01992*, doi: 10.48550/arXiv.1702.01992.
[26] H.-Y. Chuang, C.-H. Kuo, Y.-W. Chiu, C.-K. Ho, C.-J. Chen, and T.-N. Wu, "A case-control study on the relationship of hearing function and blood concentrations of lead, manganese, arsenic, and selenium," *Sci. Total Environ.*, vol. 387, nos. 1–3, pp. 79–85, Nov. 2007, doi: 10.1016/j.scitotenv.2007.07.032.
[27] J. F. Corso, "Age and sex differences in pure-tone thresholds," *J. Acoust. Soc. Amer.*, vol. 31, no. 4, pp. 498–507, Apr. 1959, doi: 10.1121/1.1907742.
[28] C. Tang, J. Zhang, W. Han, W. Shen, J. Liu, Z. Hou, P. Dai, S. Yang, and D. Han, "Analyses of the clinical characteristics of unilateral conductive hearing loss with intact tympanic membrane," *Zhonghua Er Bi Yan Hou Tou Jing Wai Ke Za Zhi*, vol. 51, pp. 348–354, May 2016, doi: 10.3760/cma.j.issn.1673-0860.2016.05.007.
[29] A. Scarpa, M. Ralli, C. Cassandro, F. M. Gioacchini, A. Greco, A. D. Stadio, M. Cavaliere, D. Troisi, M. de Vincentiis, and E. Cassandro, "Inner-ear disorders presenting with air–bone gaps: A review," *J. Int. Adv. Otology*, vol. 16, no. 1, pp. 111–116, Apr. 2020, doi: 10.5152/iao.2020.7764.
[30] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, "Artificial intelligence in healthcare: Past, present and future," *Stroke Vascular Neurol.*, vol. 2, no. 4, pp. 230–243, Dec. 2017, doi: 10.1136/svn-2017-000101.
[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.
[32] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*, doi: 10.48550/arXiv.1409.1556.
[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.
[35] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708, doi: 10.48550/arXiv.1608.06993.
[36] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.
[37] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf., Artif. Intell*, vol. 31, 2017, pp. 4278–4284, doi: 10.48550/arXiv.1602.07261.
[38] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114, doi: 10.48550/arXiv.1905.11946.
[39] M. G. Crowson, C. J. Hartnick, G. R. Diercks, T. Q. Gallagher, M. S. Fracchia, J. Setlur, and M. S. Cohen, "Machine learning for accurate intraoperative pediatric middle ear effusion diagnosis," *Pediatrics*, vol. 147, no. 4, Apr. 2021, doi: 10.1542/peds.2020-034546.
[40] D. Cha, C. Pae, S.-B. Seong, J. Y. Choi, and H.-J. Park, "Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database," *EBioMedicine*, vol. 45, pp. 606–614, Jul. 2019, doi: 10.1016/j.ebiom.2019.06.050.
[41] J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang, "The practical implementation of artificial intelligence technologies in medicine," *Nature Med.*, vol. 25, no. 1, pp. 30–36, Jan. 2019, doi: 10.1038/s41591-018-0307-0.
[42] X. Zeng, Z. Jiang, W. Luo, H. Li, H. Li, G. Li, J. Shi, K. Wu, T. Liu, X. Lin, F. Wang, and Z. Li, "Efficient and accurate identification of ear diseases using an ensemble deep learning model," *Sci. Rep.*, vol. 11, no. 1, p. 10839, May 2021, doi: 10.1038/s41598-021-90345-w.

TAEWAN KIM received the B.S. degree from the Department of Computer Science, Inha Technical College, Republic of Korea. He is currently pursuing the master's degree with the Department of Applied Artificial Intelligence, Hanyang University, Republic of Korea. His research interests include artificial intelligence and medical image processing.

SANGYEOP KIM received the B.S. and M.D. degrees from the College of Medicine, Korea University. He is currently a Resident with the Department of Otorhinolaryngology-Head and Neck Surgery, Korea University Ansan Hospital.

JAEYOUNG KIM received the Ph.D. degree from the Department of Medicine, Korea University, South Korea. He is currently a Research Professor with the Department of Medicine, Korea University Ansan Hospital, Republic of Korea. His research interests include optical imaging, AI-based clinical imaging analysis, and bioinformatics.

YEONJOON LEE received the B.S. degree from Hanyang University, in 2012, and the Ph.D. degree in security informatics from Indiana University Bloomington. He is currently an Assistant Professor with the College of Computing, Hanyang University. His research interests include mobile security, the IoT security, network security, and artificial intelligence applications in medicine and healthcare.

JUNE CHOI received the Ph.D. degree from the Department of Medicine, Korea University, Republic of Korea. He is currently a Professor with the Department of Otorhinolaryngology-Head and Neck Surgery, Ansan Hospital, Korea University. His research interests include AI-based new drug development/chemicals or drug toxicity screening systems and virtual reality assisted pre-operative surgery.