



An ensemble approach for classification of tympanic membrane conditions using soft voting classifier

Kemal Akyol¹ · Emine Uçar² · Ümit Atıla³ · Murat Uçar⁴

Received: 4 October 2023 / Revised: 11 January 2024 / Accepted: 12 February 2024 /
Published online: 22 February 2024
© The Author(s) 2024

Abstract

Otitis media is a medical concept that represents a range of inflammatory middle ear disorders. The high costs of medical devices utilized by field experts to diagnose the disease relevant to otitis media prevent the widespread use of these devices. This makes it difficult for field experts to make an accurate diagnosis and increases subjectivity in diagnosing the disease. To solve these problems, there is a need to develop computer-aided middle ear disease diagnosis systems. In this study, a deep learning-based approach is proposed for the detection of OM disease to meet this emerging need. This approach is the first that addresses the performance of a voting ensemble framework that uses Inception V3, DenseNet 121, VGG16, MobileNet, and EfficientNet B0 pre-trained DL models. All pre-trained CNN models used in the proposed approach were trained using the Public Ear Imagery dataset, which has a total of 880 otoscopy images, including different eardrum cases such as normal, earwax plug, myringosclerosis, and chronic otitis media. The prediction results of these models were evaluated with voting approaches to increase the overall prediction accuracy. In this context, the performances of both soft and hard voting ensembles were examined. Soft voting ensemble framework achieved highest performance in experiments with 98.8% accuracy, 97.5% sensitivity, and 99.1% specificity. Our proposed model achieved the highest classification performance so far in the current dataset. The results reveal that our voting ensemble-based DL approach showed quite high performance for the diagnosis of middle ear disease. In clinical applications, this approach can provide a preliminary diagnosis of the patient's condition just before field experts make a diagnosis on otoscopic images. Thus, our proposed approach can help field experts to diagnose the disease quickly and accurately. In this way, clinicians can make the final diagnosis by integrating automatic diagnostic prediction with their experience.

Keywords Voting ensemble · Pre-trained deep learning model · Tympanic membrane · Otoscopy images

1 Introduction

Otitis media (OM), a global health problem, is mostly seen in children under the age of 7 and can cause severe hearing loss and speech disorders [1]. OM, which is a disease that affects approximately 1.2 billion people worldwide [2], ranks second in hearing loss and fifth in the global disease burden [2, 3]. WHO estimates that OM complications may cause 28 thousand deaths each year [4]. While OM is diagnosed by general practitioners often using a small hand-held medical device called an otoscope, otolaryngologists usually use more advanced and specialized tools such as an endoscope or microscope [5]. However, the high cost of these devices prevents the wide use of them by otolaryngologists, and this makes it difficult to diagnose the disease in hospitals that lack these devices. In addition, different interpretations that are not objective may occur in clinical examinations performed by specialists [6].

Although there are machine learning and statistical-based studies in the literature regarding hearing loss and inner ear disorder [7–16], there are still gaps in recognizing middle ear diseases causing the limited use of automatic diagnosing systems.

The need for extracting features known as hand-crafted classification tasks that are performed by traditional machine learning methods involves complex processes and requires attention. Moreover, feature extractions directly affect classification performance. Improvements in the capacities of central processing units (CPUs) and graphical processing units (GPUs) over the last decade have enabled the development of high-performance deep learning (DL) models that eliminate the difficulty of extracting hand-crafted features [17]. Complex cognitive tasks such as image and voice recognition can be performed more efficiently thanks to DL models that contain a large number of processing units and layers.

In recent years, the ability to solve complex tasks with high success by processing large data without the need for feature extraction using DL models has made the use of these models quite common in disease classification and recognition of medical images [18–23]. However, there is a limited number of studies in the literature that diagnose and classify middle ear disease using DL methods. This study uses a current public dataset prepared by Viscaino et al. [16]. The fact that only a few DL-based studies have been carried out on this dataset has been a source of motivation for us to examine the performance of different DL approaches in classifying the disease using this dataset. To fill the current gap, this study examines the effect of the ensemble voting approach on multi-class image classification performances of DL methods on the otoscopy dataset including images of TM conditions such as myringosclerosis, earwax plug, COM, and normal. The reason for adopting the ensemble approach in this study is that it includes many models, thus increasing the ability to generalize the weak and strong sides of each independently trained model to different parts of the input space and ensuring more robust models are obtained. The second reason is that ensemble approaches are less sensitive to noise. Another important point in choosing this approach is that they can achieve better results compared to single models by ensuring bias-variance balance. In this context, the following expressions summarize the significant contributions of this study.

- This study is the first that addresses the performance of a voting ensemble framework that uses Inception V3, DenseNet 121, VGG16, MobileNet, and EfficientNet B0 pre-trained DL models on the Ear Imagery dataset.
- The accuracy of TM classification was improved by utilizing hard and soft voting approaches.

- The proposed approach provides a significant enhancement in solving slow diagnosis and high-cost problems of computer-aided decision support systems in this field.

The following sections of this study are organized as follows. Section 2 presents the papers related to the classification of tympanic membrane conditions. Section 3 introduces the dataset and explains the proposed voting ensemble framework and the pre-trained CNN models included in this framework. Section 4 presents the obtained results in detail. The results are discussed in Section 5. Finally, the study is concluded in Section 6.

2 Related work

In hospitals, otolaryngologists usually examine otoscopy images to gain a more comprehensive evaluation of the disease and implement a treatment plan based on this. In recent years, the increase in the use of DL-based approaches for the classification of middle ear diseases has been noticeable due to the contribution they provide to field experts in decision-making. For detecting normal and OM cases, Lee et al. proposed a convolutional neural network (CNN) model [24]. In the feature extraction process, they analyzed the tympanic membrane (TM) perforation using a class activation map and obtained 91% accuracy. Cha et al. [25] used an auto-endoscopic image dataset including 10,544 samples in their study, where they examined the success of 9 pre-trained DL models. In their study, a multi-class image classification including attic retraction, tympanic perforation, myringitis, otitis externa, and tumor and normal conditions was performed. They composed an ensemble classifier of two pre-trained models (Inception-V3 and ResNet101) which achieved an average accuracy of 93.67% for fivefold cross-validation. Another ensemble-based DL model was proposed by Zeng et al. They trained 9 CNN models using a total of 20,542 endoscopic images to classify 8 TM conditions. Finally, they selected the two best models according to accuracy and training time and combined them in an ensemble classifier and achieved an average accuracy of 95.59% [26]. In another study, Khan et al. [27] utilized DL models such as DenseNet161, ResNet50, VGGNet16, SE_ResNet152, and Inception-ResNet_v2 on 2484 auto-endoscopic images and reported that the DenseNet161 model obtained 95% accuracy on the dataset including COM with TM perforations, OME, and normal classes. Başaran et al. [28] performed TM detection and TM classification in their study based on a faster regional convolutional neural network (Faster R-CNN) and pre-trained CNN on augmented images, respectively. In their study, they achieved the best results using Faster R-CNN and VGG16 with an average classification accuracy of 90.8% using tenfold cross-validation. In a study by Wang et al. [29] where they used computed tomography images of the temporal bone to diagnose COM, considerations of six clinicians were compared with the results of pre-trained CNN methods. They observed that the model they proposed performed superior in some cases than clinical experts. In a study that used a transfer learning approach, Zafer [30] performed TM classification using otoscopy images of normal, CSOM, AOM, and earwax diseases obtained from the same hospital that Başaran et al. [28] stated in their study. The features provided by 7 different pre-trained CNNs were fused and then fed to traditional machine learning models and it reported that the SVM model achieved the best accuracy with 99.47%. In the same study, when using tenfold cross-validation, the accuracy was reported to be 98.74%. Singh and Dutta proposed a deep learning-based

approach for the automatic detection of ear disease. They achieved the highest accuracy 96% by applying data augmentation [31]. In addition to these studies in the literature, we have recently performed a Tympanic Membrane (TM) classification in which keypoint-based deep hypercolumn features extracted from 5 different layers of the VGG16 model are classified with 99.06% accuracy using the Bi-LSTM deep learning model [32]. However, in our study, we encountered a high-level consumption of system resources due to using hypercolumn features.

In addition to the studies conducted with OM cases discussed above, other studies deal with ear problems due to other causes. For example, Zeng et al. designed a Siamese network for the classification of conductive hearing loss, in which two ResNet-101 networks share weights on raw otoscopic images and segmented images with U-Net architecture [33]. Choi et al. addressed the problem that most TM lesions have more than one diagnostic name. The authors investigated the effect of concurrent diseases on the classification performance of deep learning networks in a dataset of auto-endoscopic images with multiple diseases [34]. Habib et al. used CNN-based ResNet-50, DenseNet-161, VGG16, and Vision Transformer models on 1842 images. They used these models, which trained with the initial weights on the ImageNet dataset, to detect normal and abnormal classes. Although the authors worked on three different multiclass datasets in their study, the fact that their proposed model performs binary classification can be considered a drawback [35]. Nam et al. used the regions of interest detected from 4,808 autoscopic images with Mask R-CNN to extract the distinctive features necessary for classification. Then, by adopting an ensemble approach including EfficientNetB0 and Inception-V3 models, they achieved an average classification accuracy of 97.29% with the fivefold cross-validation technique [36]. Afify et al. created a CNN model from scratch on 880 otoscopy images and determined the model parameters using the Bayesian optimization technique. The authors achieved 98.10% classification accuracy with the model they proposed [37]. Wang et al. proposed a VGG16 model to detect cholesteatoma and chronic suppurative otitis media diseases, which was trained on 973 computed tomography images obtained from 499 patients. However, the authors reported that the model working with CT images had difficulties detecting the early stages of these two diseases [38].

As can be seen from the literature given in this section, it has been observed that the success of the ensemble approach of DL methods on TM classification has not been examined in detail. The fact that deep learning-based studies for middle ear disease recognition and classification are still limited and the need for DL-based approaches that will efficiently solve this problem motivated us to develop a DL approach for TM classification that uses system resources more efficiently.

3 Methodology

This study presents a framework that performs the classification of TM conditions including normal, earwax plug, myringosclerosis, and chronic otitis media, stably and efficiently. In this context, the predictions made by fine-tuned pre-trained CNN models on the Ear Imagery dataset were handled by a voting ensemble approach to obtain a more accurate and stable classification. A voting ensemble framework was built using both soft and hard voting approaches and the results of both approaches were examined. Figure 1 shows the overview of the proposed approach.

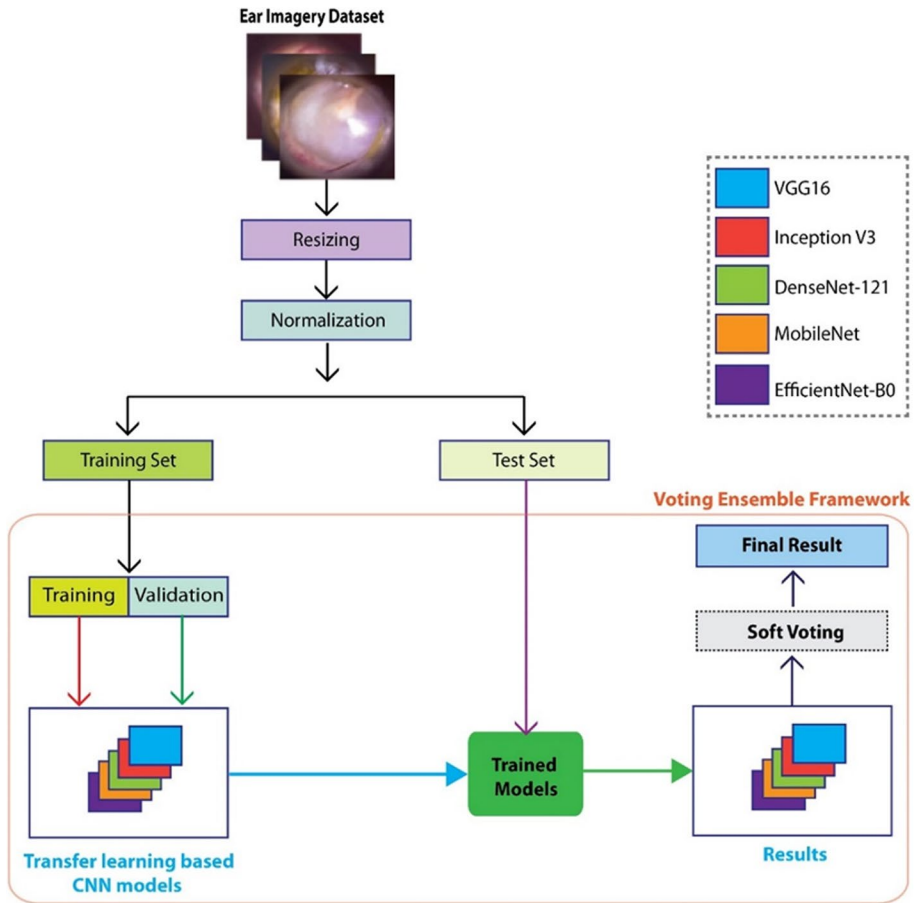


Fig. 1 Overview of the proposed voting ensemble framework

3.1 Dataset

In this study, the Ear Imagery dataset is used, which is publicly available and prepared by Viscaino et al. [16]. The images in the dataset used in this study are collected and labeled by a research group whose one of the members is an otolaryngologist in collaboration with the Department of Otolaryngology of the Clinical Hospital from Universidad de Chile. The ground truths of OM cases were determined by the otolaryngologist and approved by the scientific committee of the relevant university. In addition, this dataset was divided by its owners into train, validation, and test sets. The dataset has 880 otoscopy images with a resolution of 420×380 which belongs to 180 patients aged 7 to 65 years who applied to the otolaryngology outpatient clinic of the Universidad de Chile Clinical Hospital. The images in the dataset are divided into 4 categories chronic otitis media, myringosclerosis, earwax plug, and normal otoscopy and each category has 220 samples that provide a balanced dataset. Figure 2 demonstrates sample images of each TM condition.

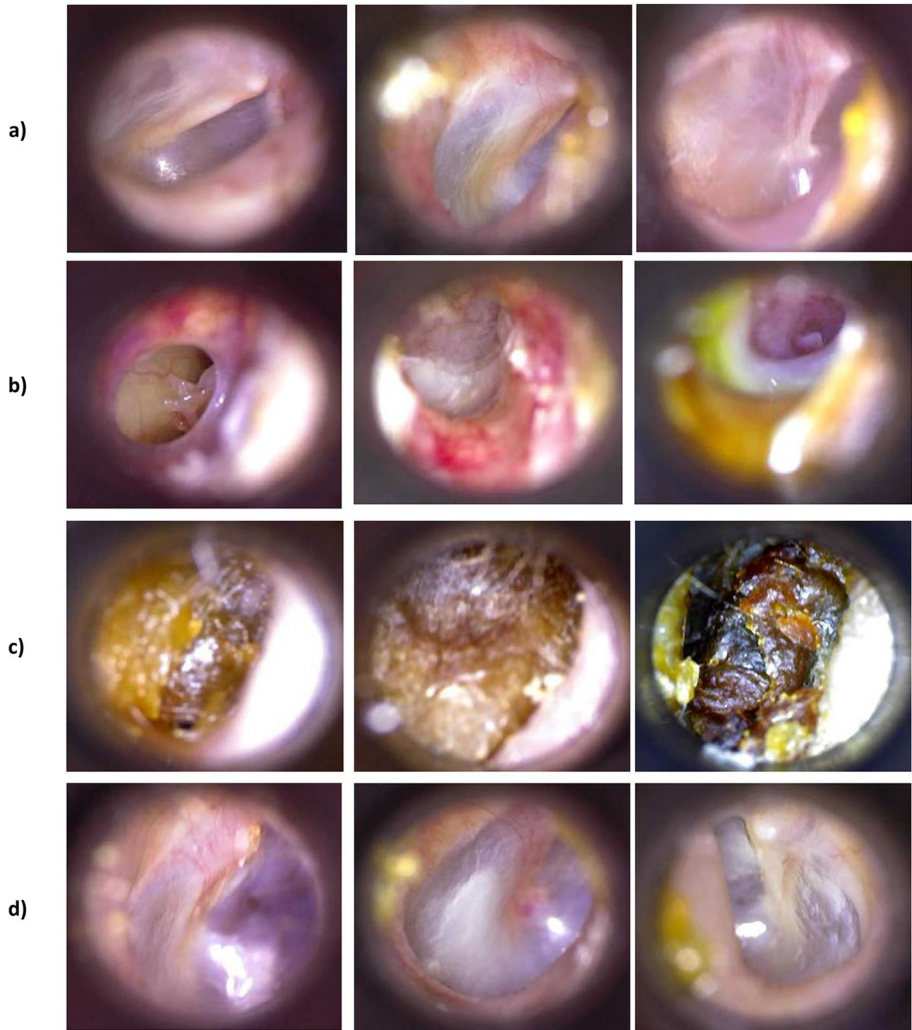


Fig. 2 Sample images belonging to different TM conditions. **a** Normal **(b)** Chronic Otitis Media **(c)** Earwax plug **(d)** Myringoesclerosis

3.2 CNN

Convolutional neural networks introduced by Yann Lecun [39] are neural networks primarily specialized for image classification, object recognition, and detection. Unlike traditional image processing methods, CNNs can learn the most appropriate features by learning on their own. Traditional CNN structure consists of a convolution layer, pooling layer, and fully connected layer (Fig. 3). The convolution layer and the pooling layer are used for feature extraction. Classification of an input image is performed in the fully connected layer. All features from previous layers are flattened into a one-dimensional

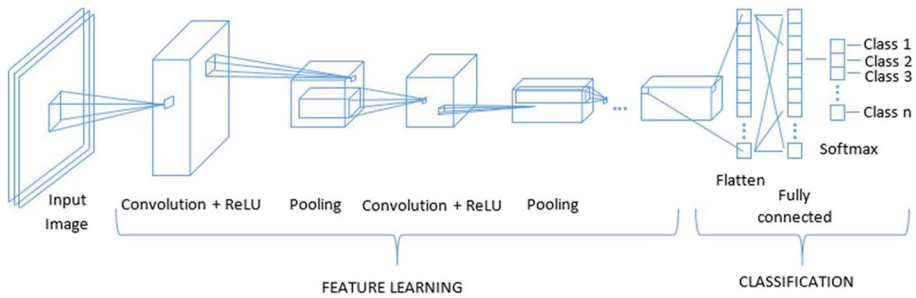


Fig. 3 Traditional CNN structure

feature vector to make the neural network suitable for classification and prediction and finally, output classification probabilities are obtained in the softmax layer.

3.3 Pre-trained CNN models

In this study, CNN-based DL models such as Inception-V3, DenseNet 121, VGG16, MobileNet, and EfficientNet B0 were used. These models were summarized in the following subsections.

3.3.1 Inception-V3

Inception V3 developed by Google is the extended network of the GoogLeNet and the third release in the Deep Learning Evolutionary Architectures series [40]. Inception-V3 is one of the most advanced architectures used in image classification. The Inception-v3 architecture proposes an initial model that combines multiple different sizes of convolutional filters in a new filter. This type of design reduces the number of parameters to be trained, thus reducing the computational complexity. Convolutions, average pooling, max pooling, concats, dropouts, and fully connected layers are parts of the symmetrical and asymmetrical building blocks that make up the model. The Softmax activation function is used in the last layer of the Inception-V3 architecture, which has 42 layers in total and an input layer that takes images of 299×299 pixels.

3.3.2 DenseNet

Convolution and sub-sampling result in a drop in feature maps during the training of neural networks, and losses in image features are also experienced during the transition between layers. Huang [41] created the DenseNet system to make better use of image information. Each layer in the system is fed forward to the other layers. A layer can thus access the property information of any layer preceding it. In addition, DenseNet has several focal points: alleviates the disappearing angle problem, includes strengthening, stimulates highlight reutilizing and generously reduces the number of parameters.

3.3.3 VGG16

The VGG16 model, which has 16 layers, was developed by Simonyan and Zisserman and is based on the CNN model [42]. It received a 92.7% accuracy score in the 2014 ImageNet Large Scale Visual Recognition Challenge (ILSVRC). First, 13 of these layers are convolutional layers with 3×3 filters and 2×2 max pooling layers. VGG16 shrinks input images across maximum pooling layers. The ReLu activation function is applied in these layers. Then, three fully connected layers contain most of the parameters of the network. In addition, the model consists of 138 million parameters. Considering the depth of the model is deeper compared to previous CNN models, it takes a long time to train such a large model.

3.3.4 MobileNet

MobileNet [43] is a lightweight deep neural network (DNN) for mobile-embedded terminals that was suggested by Google in 2017. MobileNet, which is built on modern architecture, employs a deeply separable convolution to construct a lightweight DNN. The research is aimed at model compression, and the basic notion is the skillful decomposition of the convolution kernel. It can efficiently reduce network parameters while accounting for the optimization delay.

3.3.5 EfficientNet

The EfficientNet model, which achieved 84.4% accuracy in the ImageNet classification challenge with 66 M parameters, can be thought of as a set of CNN models. The EfficientNet has 8 models ranging from B0 to B7, and as the model number increases, the estimated number of parameters does not increase significantly, but the accuracy rate does. DL architectures are designed to bring more effective ways with smaller models. Unlike other state-of-the-art models, the EfficientNet model scales in depth, width, and resolution while attempting to scale down, yielding more efficient results. The grid search is the first stage in the compound scaling approach to determine the relationship between the multiple scaling dimensions of the baseline network under a fixed resource restriction. A reasonable scaling factor for depth, width, and resolution dimensions is established in this manner. Following that, the coefficients are used to scale the baseline network [44].

3.4 Transfer learning

Transfer learning is the task of transferring the knowledge gained from a trained network to a different network created to solve a similar problem [45]. The main idea in this approach is that DL networks can learn characteristic features if they are properly trained. For general features, the last few layers of the pre-trained network are replaced with layers suitable for the new task. Using a network pre-trained with a large dataset both reduces the training time required for the new task and enables to achievement higher accuracy.

3.5 Voting ensemble approach

Rather than trying to find a single best classifier on a classification problem, stronger generalization capacity can be achieved by combining more than one classifier with ensemble methods [46]. Ensemble methods often produce more accurate results than a single classifier. The most preferred ensemble approach adopting the combination method is the voting ensemble approach [47–50], which provides models with high generalization capacity. There are two commonly used schemes among voting approaches. These are hard voting and soft voting.

Hard voting is the basic example of majority voting. This method, it is aimed to estimate the final class label by calculating the majority of labels predicted by all classifiers. The function is represented by Eq. 1.

$$\hat{y}_i = \text{mode}\{C_1(x_i), C_2(x_i), \dots, C_j(x_i), \dots, C_m(x_i)\} \quad (1)$$

where, m is the number of classifiers, x_i denotes the i^{th} sample, $C_j(x_i)$ denotes the predictions of the j^{th} classifier and the mode function is used to calculate the majority vote of all predictions [51].

Soft voting calculates the weighted sum of the prediction probabilities of all classifiers for each class for estimating the final class label as shown in Eq. 2. The label belonging to the class with the highest probability in total is selected.

$$\hat{y}_i = \arg \max_k \sum_{j=1}^m w_j p_{i,k}^j \quad (2)$$

where, m is the number of classifiers, w_j denotes the weight of j^{th} classifier, $p_{i,k}^j$ denotes the prediction probability of j^{th} classifier for assigning the i^{th} sample to k^{th} class [51].

3.6 Performance metrics

Several common performance metrics such as Accuracy (Acc), Sensitivity (Se), Specificity (Sp), and Precision (Pre) are used to measure the performance of the state-of-the-art CNN models and the voting ensemble approach including these models. These metrics are based on values of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) which are calculated using the values in the confusion matrix. Since multi-class classification is carried out in this study, TP, TN, FP, and FN values are calculated separately for each class. Thus, TP gives the correct number of classified images for the relevant class, while FP is the number of misclassified images in all other classes except the relevant class. On the other hand, while TN is the total number of correctly classified images for all other classes except the relevant class, FN is the number of incorrectly classified images for the relevant class. The performance metrics used in the study and their extended calculations for multi-class classification using macro-averaging [52, 53] are given in Eqs. 3–12.

For a class k ,

$$\text{Sen}(k) = \frac{TP(k)}{TP(k) + FN(k)} \quad (3)$$

$$\text{Spe}(k) = \frac{TN(k)}{TN(k) + FP(k)} \quad (4)$$

$$Acc(k) = \frac{TP(k) + TN(k)}{TP(k) + FN(k) + TN(k) + FP(k)} \quad (5)$$

$$Pre(k) = \frac{TP(k)}{TP(k) + FP(k)} \quad (6)$$

$$F1\ score(k) = \frac{2 \times Pre(k) \times Sen(k)}{Pre(k) + Sen(k)} \quad (7)$$

$$Average\ Sen = \frac{1}{\#classes} \sum_{k=1}^{\#classes} Sen(k) \quad (8)$$

$$Average\ Spe = \frac{1}{\#classes} \sum_{k=1}^{\#classes} Spe(k) \quad (9)$$

$$Average\ Acc = \frac{1}{\#classes} \sum_{k=1}^{\#classes} Acc(k) \quad (10)$$

$$Average\ Pre = \frac{1}{\#classes} \sum_{k=1}^{\#classes} Pre(k) \quad (11)$$

$$Average\ F1\ score = \frac{1}{\#classes} \sum_{k=1}^{\#classes} F1\ score(k) \quad (12)$$

Apart from these performance metrics, another metric used for measuring the classification performance is the Receiver Operating Characteristic (ROC) curve which demonstrates the variation of the True Positive Rate (TPR) concerning False Positive Rate (FPR). The Area Under the Curve (AUC) is the area under the ROC curve. When this area is close to 1, it indicates high success for the model, while a value close to 0 indicates low success on discrimination of classes and 0.5 indicates that the model selects a random class each time. The ratio of true positives to all positives gives TPR which is also called sensitivity. The ratio of incorrectly predicted negatives to all negatives gives the FPR, which is also defined as *1-Specificity*. The calculations of these metrics are as given in Eqs. 13–14.

For a class k ,

$$TPR(k) = Sen(k) \quad (13)$$

$$FPR(k) = 1 - Spe(k) \quad (14)$$

4 Results

4.1 Training

In this study, the Ear Imagery dataset was used in its original form determined by the publishers, without changing the number of instances for the training, testing, and validation datasets. The publishers divided the dataset containing 880 images into training,

validation, and test datasets containing 576, 144, and 160 images, respectively. This number of samples corresponds to approximately 80% and 20% for training and testing datasets respectively with reserving 20% of the training dataset for validation. The dataset includes 3 classes of OM disease (chronic otitis media, earwax plug, myringosclerosis), as well as a control class (normal), yielding in total of 4 classes. The train, validation, and test sets were balanced such that they contained an equal number of sample images for each class. For each model used in this study, the zero-center normalization method was applied to the pixels of otoscopy images. We first split the dataset into train and test sets and then applied Z-score normalization using only the training dataset. Afterward, Z-score normalization was applied to the test dataset by using the mean and variance obtained from the training dataset. In this study, pre-trained CNN models on the ImageNet dataset were used for classification by applying required modifications in the last layers of the models. The structure of the CNN layers that form the feature maps in the models was preserved and a global average pooling layer was added afterwards. The number of output neurons in the last fully connected layers of the pre-trained models was set to 4 for adapting them to the tympanic membrane classification problem. In the last layer, Softmax was used as the activation function, and categorical cross entropy was used as the loss function. The trainable parameters are set to True for all layers in the models. Hyperparameters of the CNN models such as learning rates, batch size, number of epochs, and activation function were set by trial and error and the determined parameter values were given in Table 1. The training of all CNN models used in this study was carried out in 40 epochs using Adam optimizer with batch size set to 8. To prevent overfitting, weights obtained in the epoch that achieved the best validation accuracy during training were chosen as the final model.

In this study, all codes were developed using Keras 2.3.1 in Python programming language, and experiments were carried out on a computer running a 64-bit Ubuntu 18.04.3 LTS operating system with Intel (R) Xeon (R) 2.00 GHz CPU, 12 GB RAM and NVIDIA T80 GPU with 12 GB memory.

4.2 Experimental results

In this study, TM conditions were classified using the Voting ensemble framework, and the results of the proposed framework were compared with current studies in the literature. First, as mentioned in Section 3.4, we applied the transfer learning approach and fine-tuned the pre-trained CNN models to obtain the final models. The accuracy and loss curves obtained during the training of the models are given in Fig. 4. As can be observed from the figure, the accuracies obtained with pre-trained models were quite high.

Table 1 Parameter values used for CNN models

Model name	Input size	Batch size	Learning rate	Beta 1	Beta 2	Minimum validation loss epoch
InceptionV3	299×299×3	8	0.0001	0.9	0.999	33
DenseNET 121	224×224×3	8	0.0001			16
VGG16	224×224×3	8	0.00005			6
MobileNET	224×224×3	8	0.00005			30
EfficientNET B0	224×224×3	8	0.00005			11

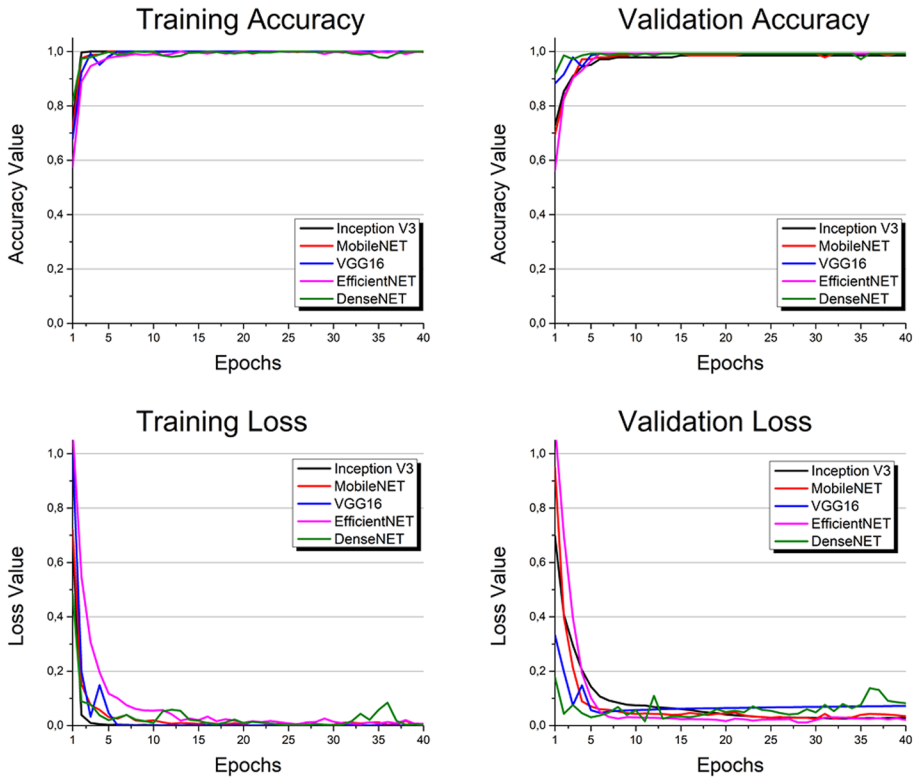


Fig. 4 Training and validation accuracy/loss curves

While classifying in the Voting ensemble framework (soft voting and hard voting) approach, the weights used for voting are the values obtained from the Softmax activation function in the output layers of the DL models used in the study. Since the dataset used in this study includes 4 classes, the average values were calculated by taking the mean of the values presented by the model concerning relevant metrics for each class. In this context,

Table 2 Average results of pre-trained DL models and voting ensembles on the test dataset

Method	Acc	Average				Training time (Sec.)	Test- ing time (Sec.)
		Sen	Spe	Pre	F1-score		
InceptionV3	0.966	0.931	0.977	0.933	0.931	360.711	3.53
DenseNet 121	0.981	0.963	0.988	0.963	0.962	421.646	4.78
VGG16	0.975	0.950	0.983	0.951	0.950	456.107	5.93
MobileNet	0.972	0.944	0.981	0.946	0.944	140.97	1.35
EfficientNet B0	0.978	0.956	0.985	0.958	0.956	290.15	3.44
Hard Voting	0.984	0.969	0.990	0.970	0.969	-	-
Soft Voting	0.988	0.975	0.991	0.976	0.975	-	-

*Sec: Second

the average Acc, Sen, Spe, and Pre values obtained on the test dataset for all models are given in Table 2. The values marked in bold in this table represent the best values obtained for the relevant performance criterion.

As seen in Table 3, while the soft and hard voting approaches showed close performances, these two approaches offered higher accuracy compared to other models. In addition, these two approaches offered the best performance in the context of Sen and Spe metrics. In particular, the soft voting approach slightly outperformed the hard voting approach. Despite close accuracy values between 96.6% and 98.8% offered by the models, when it was evaluated in the context of the precision metric, it was seen that there were more significant differences between the models. Precision indicates how many of the samples a model predicts positively are true positive, and precision increases as the number of FPs decreases. Accordingly, although the correct classification rate in the samples classified as positive by soft voting and hard voting approaches is higher than other models, soft voting provided the best performance in terms of this metric.

The challenging aspect of classification problems is the model's ability to make predictions for samples not considered in the training process and the accuracy of these predictions to be evaluated in the case of model uncertainty. The confidence interval is an appropriate way to measure the uncertainty of the estimate. To measure the confidence of the model, the model is trained multiple times and then the confidence interval is calculated by obtaining the distribution of the different predictions made by the model on the test data. Accordingly, to measure the confidence of the models used in this study, the models were trained 5 times, and mean, accuracy values, standard deviations, and the lower and upper accuracies for 95% confidence interval were calculated for each model. As seen in Table 3, the differences between the lower and upper accuracy values calculated for the models vary within a small range such as 0.12 and 0.3, which indicates that the uncertainties of the models are quite low.

In addition, the accuracies of the pre-trained models and the voting ensemble approach were given in Fig. 5 in ascending order. As can be seen from Fig. 5, performances of voting ensemble approaches were superior to pre-trained models alone for TM conditions classification. The highest accuracy was achieved with the soft voting approach with 98.8%, while the model with the lowest accuracy was Inception V3 with 96.6%.

The confusion matrices of the ensemble approaches are given in Fig. 6. The Acc, Sen, Spe, and Pre values obtained by the hard and soft voting ensemble approaches, which provide the best performances in the context of all the metrics used in this study,

Table 3 Experimental results and confidence intervals of used models

Method	Accuracy					Mean accuracy	Standard deviation	95% CI	
	Exp.1	Exp.2	Exp.3	Exp.4	Exp.5			Low	Up
InceptionV3	0.966	0.953	0.969	0.947	0.975	0.962	0.010	0.953	0.971
DenseNet 121	0.981	0.975	0.984	0.978	0.969	0.977	0.005	0.972	0.982
VGG16	0.975	0.969	0.988	0.978	0.978	0.978	0.006	0.973	0.983
MobileNet	0.972	0.984	0.981	0.975	0.975	0.978	0.004	0.974	0.982
EfficientNet B0	0.978	0.972	0.978	0.972	0.981	0.976	0.004	0.973	0.979
Hard voting	0.984	0.978	0.984	0.972	0.978	0.979	0.004	0.975	0.983
Soft voting	0.988	0.981	0.981	0.972	0.978	0.980	0.005	0.975	0.985

Bold values indicate the best results in the table

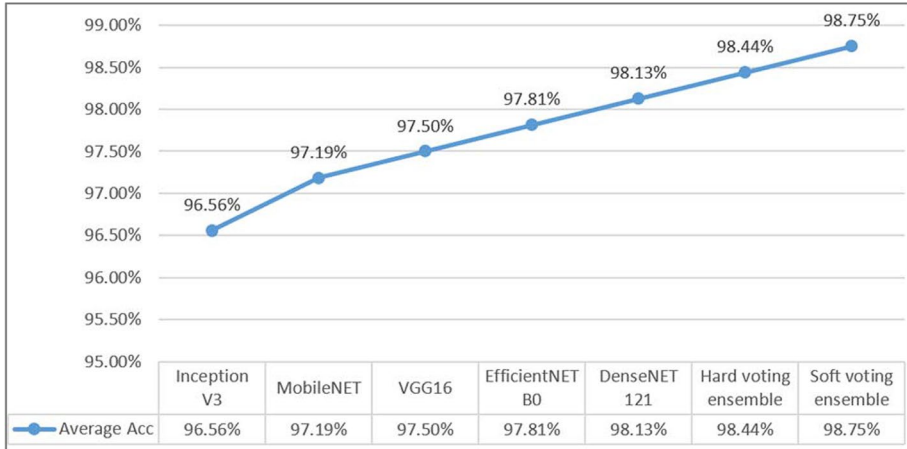


Fig. 5 Performance comparisons of the models

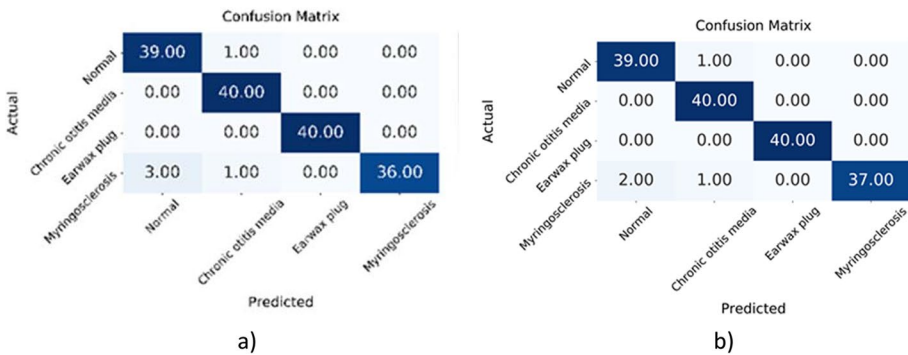


Fig. 6 Confusion Matrices (a) Hard voting model (b) Soft voting model

Table 4 Detailed results of the hard voting model

Class	Sen	Spe	Pre	Acc
Normal	0.975	0.975	0.929	0.975
Abnormal				
Chronic otitis media	1.000	0.983	0.952	0.988
Earwax plug	1.000	1.000	1.000	1.000
Myringosclerosis	0.900	1.000	1.000	0.975

Table 5 Detailed results of the soft voting model

Class	Sen	Spe	Pre	Acc
Normal	0.975	0.983	0.951	0.981
Abnormal				
Chronic otitis media	1.000	0.983	0.952	0.988
Earwax plug	1.000	1.000	1.000	1.000
Myringosclerosis	0.925	1.000	1.000	0.981

were detailed in Tables 4 and 5 for each class in the Ear Imagery dataset. Considering the classification performances of hard and soft voting approaches used in the final estimation stage, while both approaches offered the same classification performance for COM and earwax plug classes in the abnormal category, the soft voting approach was slightly better than the hard voting approach in the classification of the samples in normal class and the myringosclerosis class in the abnormal category.

Considering the precision values obtained by the soft voting approach for each class, this approach achieved values ranging from 95.1% to 100%. When the accuracy values obtained for each class were examined, it was seen that this approach made predictions with an accuracy between 98.1% and 100%. Finally, when the Sen and Spe values obtained for each class were analyzed, it was seen that the soft voting approach achieved the Sen metric between 92.5% and 100% and the Spe metric between 98.3% and 100%.

With the aim of more concretely showing the performance of DL architectures considered in this study, the number of misclassified samples, in other words, the number of FNs for each class was also indicated in Table 6. As can be observed in this table, the soft voting approach had some incorrect classifications of the myringosclerosis class. In the confusion matrix given in Fig. 6b obtained from the soft voting approach, it can be seen that there were 3 FNs including 2 normal, and 1 COM case. The soft voting approach presented the lowest number of misclassifications with 4 out of all samples. Although no model classified all samples in the normal class correctly, soft and hard voting ensemble approaches were the ones that presented the lowest number of misclassifications with 1. Among the CNN-based pre-trained models, DenseNet 121 and VGG16 were the ones that achieved this success. Hard and soft voting approaches correctly classified all the samples in the COM class at the final prediction stage. The pre-trained models that classified all the samples in this class correctly were DenseNet 121 and EfficientNet B0. Samples in the earwax plug class were correctly classified by both voting ensemble approaches and pre-trained CNN models. On the other hand, the most difficult TM condition to be distinguished among was myringosclerosis. The FP values presented by the pre-trained CNN models and the voting ensemble approach for the myringosclerosis class were higher than the other classes. In addition, the soft voting approach had the lowest number of misclassifications compared to single models and the hard voting approach. Other methods performed misclassifications with a number ranging from 4 to 6.

The ROC curves and AUC values of both voting frameworks obtained for each TM condition are shown in Fig. 7. As seen in Fig. 7, the soft voting approach was superior

Table 6 False prediction numbers of DL models

	Inception V3	DenseNet 121	VGG16	MobileNet	Efficient-Net B0	Hard voting	Soft voting
Normal	2	1	1	2	3	1	1
Chronic otitis media	3	0	2	3	0	0	0
Earwax plug	0	0	0	0	0	0	0
Myringosclerosis	6	5	5	4	4	4	3
Total	11	6	8	9	7	5	4

Bold values indicate the least number of false predictions

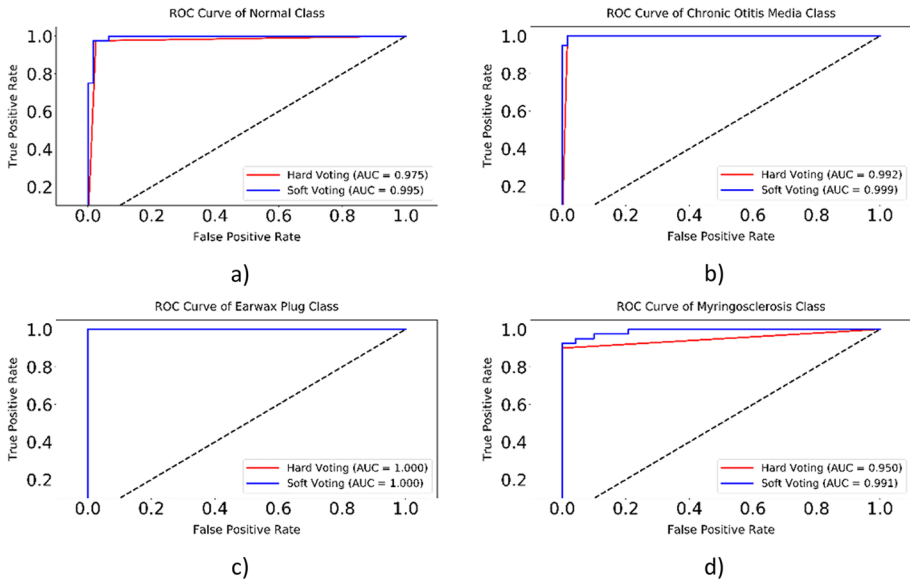


Fig. 7 ROC Curves of Voting frameworks for each class (a) COM (b) earwax plug (c) myringosclerosis (d) normal

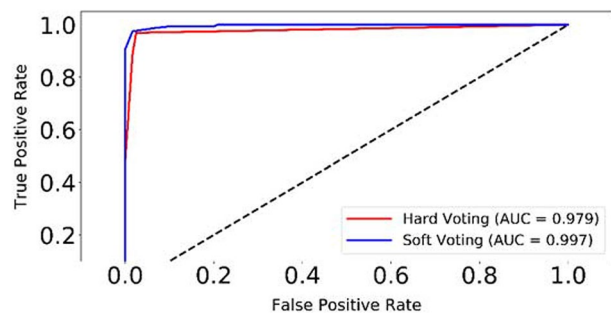
to the hard voting approach with AUC values of 0.995, 0.991, 0.999, and 1.000 for normal, myringosclerosis, COM, and earwax plug classes, respectively.

In addition, the class-based macro average ROC curves and AUC values of the hard and soft voting approaches used in the final prediction stage for the classification of TM conditions are given in Fig. 8. As seen from the ROC curves, the soft voting approach for all classes was slightly better than a hard voting approach.

5 Discussion

Studies on TM conditions classification can be divided into two groups such as traditional machine learning methods and DL methods. It has been observed that model accuracies ranging from 73 to 93% were obtained in studies where traditional machine learning methods were used on different datasets. In the last two years, DL models specialized in images,

Fig. 8 ROC curves of soft and hard voting approaches



which do not require a troublesome handcrafted feature extraction process, have been commonly used for increasing the success of TM condition classification.

The dataset used in this study is a current dataset made public by Viscaino et al. [16]. They achieved 93.90% accuracy in their study which used classical machine learning methods to classify extracted features. In this study, a voting ensemble framework containing pre-trained CNN models was proposed for the classification of TM conditions. Our proposed soft voting and hard voting ensemble approaches demonstrated their superiority over classical machine learning by achieving 98.8% and 98.4% accuracy on the dataset. The accuracy obtained by Viscaino et al. [16] fell behind the accuracy of our proposed model. In addition, the performances of single pre-trained CNN models were also examined. The performances of the pre-trained CNN models lagged behind both of the proposed voting ensemble approaches. Accordingly, among the single pre-trained CNN models, InceptionV3 has the lowest performance compared to the others in terms of all metrics. With 98.1% accuracy, DenseNet-121 both showed its superiority over other single models and offered the closest performance to hard and soft voting ensemble approaches.

Furthermore, Table 7 summarizes the TM classification results of our proposed DL-based ensemble model and other CNN-based DL studies. Except for studies of Singh and Dutta[31] and Afify et al. [37], the studies in this table used different datasets. Therefore, a direct comparison of the classification performances presented in these studies with the performance of our proposed model is inappropriate. Singh and Dutta [31] carried out a four-class classification on augmented images with their own proposed CNN architecture consisting of 6 convolutions and 2 dense layers without using a transfer learning approach. They trained the CNN model 400 epochs and achieved 96% accuracy on the test dataset. Their study is valuable in terms of showing that 96% accuracy can be achieved in TM classification with only data augmentation and without using a transfer learning approach. Since they applied data augmentation to both train and test datasets, a direct comparison of their results with our study was not suitable. On the other hand, in terms of looking at the time spent training the models, Singh and Dutta [28] did not report the training times in their study. They trained the CNN model at 400 epochs and also validated the model containing the best weights at the 367th epoch by applying the checkpoint technique. In addition, the authors needed a process that would require extra effort to apply the data augmentation technique on both train and test datasets before the training phase of their CNN model. Even though five different pre-trained models were trained in our study, using the transfer learning approach in this training enabled the trained models to reach the best generalization performance in a very short training period of 40 epochs. The obtained accuracy-loss plots in experiments support this hypothesis. On the other hand, Afify et al. proposed a CNN architecture with hyperparameters adjusted by the Bayesian optimization method. Considering the results reported by the authors, it was seen that although the model they proposed was superior to other studies using the same dataset, their proposed model fell behind our model in terms of accuracy score.

From the perspective of ensemble learning approaches, there are a limited number of studies in the literature on the classification of TM conditions. In two prominent studies [25, 26], researchers used two different trained CNN models by combining them in a soft voting ensemble and they showed that their proposed ensemble approaches increased the classification success compared to single CNN models. Our study differs from [25, 26] in terms of both the number of CNN models used and voting approaches. The model we propose in this study combines four different CNN models via both hard voting and soft voting rules. This model gave pay to an increase in the classification success compared to single CNN models in both cases.

Table 7 The results of the proposed model and other DL-based approaches

Study	Method	Number of samples	Number of classes	Acc (%)	Sen (%)	Spe (%)
Lee et al. [24]	CNN	1338	4	91.0	90.5	92.9
Cha et al. [25]	Ensemble deep learning	10,544	6	93.67	-	-
Zeng et al. [26]	Ensemble deep learning	20,542	8	95.59	-	-
Khan et al. [27]	Pre-trained CNN models	2484	3	95.0	-	-
Basaran et al. [28]	Faster R-CNN, pre-trained CNN models, VGG16	1692	2	90.48	86.84	93.50
Wang et al. [29]	Pre-trained CNN models	1147	3	76.7	-	-
Zafer [30]	Pre-trained CNN models, deep features, SVM	857	4	99.47	99.35	99.77
Singh and Dutta [31]	Data augmentation + CNN	5703	4	96.0	-	-
Habib et al. [35]	Deep Learning models	1842	2	92.0	86.0	95.0
Nam et al. [36]	Mask R-CNN + EfficientNetB0 and Inception-V3	4808	6	97.29	-	-
Afify et al. [37]	CNN-Bayesian optimization	880	4	98.1	98.11	99.36
Wang et al. [38]	VGG 16	973	3	-	96.7	97.6
Our method	Soft voting ensemble framework	880	4	98.8	97.5	99.1

6 Conclusion

In this study, the Voting ensemble framework including 5 different pre-trained models was proposed to classify different TM conditions such as normal, chronic otitis media, earwax plug, and myringosclerosis. To make the best use of the Voting ensemble approach, the performances of the soft and hard voting methods were examined. Experiments showed that the soft voting method offered the best performance with 98.8% Acc, 97.5% Sen, and 99.1% Spe compared to the state-of-the-art pre-trained models. Moreover, the classification performance of the proposed method was superior to other studies performed on the same dataset in the literature. Considering the high classification accuracy achieved, the proposed soft voting ensemble framework can play an important role in the development of expert systems to be used in real clinical environments. A low-cost system based on DL will help to diagnose TM conditions more accurately in a clinical environment where there is no field specialist, and on the other hand, such a system will decrease the variability among the visual examination results of observers which is error-prone.

On the other hand, the fact that this study has been performed on a small dataset can be considered a limitation. However, it should be noted that insufficiency in the number of publicly available datasets is a matter to be considered. Finally, in our future studies, we intend to use domain adaptation by conducting experimental studies on different datasets with more cases of middle ear disease, and it is planned to obtain a model that can generalize to more than one dataset.

Acknowledgements The authors would like to thank Viscaino et al. for providing the public Ear Imagery dataset.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK). No funds, grants, or other support was received.

Data availability This study uses a public dataset provided by Viscaino et al. available at: <https://doi.org/10.6084/m9.figshare.11886630>.

Declarations

Conflicts of interest The authors declare that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Kørvel-Hanquist A, Koch A, Lous J et al (2018) Risk of childhood otitis media with focus on potentially modifiable factors: a Danish follow-up cohort study. *Int J Pediatr Otorhinolaryngol* 106:1–9. <https://doi.org/10.1016/J.IJPORL.2017.12.027>

2. Morris PS, Leach AJ (2009) Acute and chronic otitis media. *Pediatr Clin North Am* 56:1383–1399. <https://doi.org/10.1016/j.pcl.2009.09.007>
3. Rovers MM, Schilder AGM, Zielhuis GA, Rosenfeld RM (2004) Otitis media. *Lancet* 363:465–473. [https://doi.org/10.1016/S0140-6736\(04\)15495-0](https://doi.org/10.1016/S0140-6736(04)15495-0)
4. World Health Organization (2004) Chronic suppurative otitis media - burden of illness and management options. <https://iris.who.int/handle/10665/42941>
5. Sundgaard JV, Harte J, Bray P et al (2021) Deep metric learning for otitis media classification. *Med Image Anal* 71:102034. <https://doi.org/10.1016/j.media.2021.102034>
6. Pichichero ME (2003) Diagnostic accuracy of otitis media and tympanocentesis skills assessment among pediatricians. *Eur J Clin Microbiol Infect Dis* 22:519–524. <https://doi.org/10.1007/s10096-003-0981-8>
7. Bing D, Ying J, Miao J et al (2018) Predicting the hearing outcome in sudden sensorineural hearing loss via machine learning models. *Clin Otolaryngol* 43:868–874. <https://doi.org/10.1111/coa.13068>
8. Chao T-K, Hsiu-Hsi Chen T (2010) Predictive model for improvement of idiopathic sudden sensorineural hearing loss. *Otol Neurotol* 31:385–393. <https://doi.org/10.1097/MAO.0b013e3181cdd6d1>
9. Suzuki H, Mori T, Hashida K et al (2011) Prediction model for hearing outcome in patients with idiopathic sudden sensorineural hearing loss. *Eur Arch Oto-Rhino-Laryngol* 268:497–500. <https://doi.org/10.1007/s00405-010-1400-2>
10. Suzuki H, Tabata T, Koizumi H et al (2014) Prediction of hearing outcomes by multiple regression analysis in patients with idiopathic sudden sensorineural hearing loss. *Ann Otol Rhinol Laryngol* 123:821–825. <https://doi.org/10.1177/0003489414538606>
11. Kuruvilla A, Shaikh N, Hoberman A, Kovačević J (2013) Automated diagnosis of otitis media: vocabulary and grammar. *Int J Biomed Imaging* 2013:1–15. <https://doi.org/10.1155/2013/327515>
12. Shie C-K, Chang H-T, Fan F-C et al (2014) A hybrid feature-based segmentation and classification system for the computer aided self-diagnosis of otitis media. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, pp 4655–4658. <https://doi.org/10.1109/EMBC.2014.6944662>
13. Mironică I, Vertan C, Gheorghe DC (2011) Automatic pediatric otitis detection by classification of global image features. In: 2011 E-Health and Bioengineering Conference (EHB). pp 1–4. <https://api.semanticscholar.org/CorpusID:41625076>
14. Myburgh HC, van Zijl WH, Swanepoel D et al (2016) Otitis media diagnosis for developing countries using tympanic membrane image-analysis. *EBioMedicine* 5:156–160. <https://doi.org/10.1016/J.EBIOM.2016.02.017>
15. Myburgh HC, Jose S, Swanepoel DW, Laurent C (2018) Towards low cost automated smartphone- and cloud-based otitis media diagnosis. *Biomed Signal Process Control* 39:34–52. <https://doi.org/10.1016/J.BSPC.2017.07.015>
16. Viscaino M, Maass JC, Delano PH et al (2020) Computer-aided diagnosis of external and middle ear conditions: a machine learning approach. *PLoS ONE* 15:1–18. <https://doi.org/10.1371/journal.pone.0229226>
17. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521:436–444. <https://doi.org/10.1038/nature14539>
18. Ciompi F, de Hoop B, van Riel SJ et al (2015) Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box. *Med Image Anal* 26:195–202. <https://doi.org/10.1016/J.MEDIA.2015.08.001>
19. Gao M, Bagci U, Lu L et al (2018) Holistic classification of CT attenuation patterns for interstitial lung diseases via deep convolutional neural networks. *Comput Methods Biomech Biomed Eng Imaging Vis* 6:1–6. <https://doi.org/10.1080/21681163.2015.1124249>
20. Kleesiek J, Urban G, Hubert A et al (2016) Deep MRI brain extraction: a 3D convolutional neural network for skull stripping. *Neuroimage* 129:460–469. <https://doi.org/10.1016/J.NEUROIMAGE.2016.01.024>
21. Moeskops P, Viergever MA, Mendrik AM et al (2016) Automatic segmentation of MR brain images with a convolutional neural network. *IEEE Trans Med Imaging* 35:1252–1261. <https://doi.org/10.1109/TMI.2016.2548501>
22. Plis SM, Hjelm DR, Salakhutdinov R et al (2014) Deep learning for neuroimaging: a validation study. *Front Neurosci* 8:229. <https://doi.org/10.3389/fnins.2014.00229>
23. Suk H-I, Lee S-W, Shen D (2015) Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. *Brain Struct Funct* 220:841–859. <https://doi.org/10.1007/s00429-013-0687-3>
24. Lee JY, Choi SH, Chung JW (2019) Automated classification of the tympanic membrane using a convolutional neural network. *Appl Sci* 9:1827. <https://doi.org/10.3390/app9091827>

25. Cha D, Pae C, Seong S-B et al (2019) Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database. *EBioMedicine* 45:606–614. <https://doi.org/10.1016/J.EBIOM.2019.06.050>
26. Zeng X, Jiang Z, Luo W et al (2021) Efficient and accurate identification of ear diseases using an ensemble deep learning model. *Sci Rep* 11(11):1–10. <https://doi.org/10.1038/s41598-021-90345-w>
27. Khan MA, Kwon S, Choo J et al (2020) Automatic detection of tympanic membrane and middle ear infection from oto-endoscopic images via convolutional neural networks. *Neural Netw.* <https://doi.org/10.1016/J.NEUNET.2020.03.023>
28. Başaran E, Cömert Z, Çelik Y (2020) Convolutional neural network approach for automatic tympanic membrane detection and classification. *Biomed Signal Process Control* 56:101734. <https://doi.org/10.1016/J.BSPC.2019.101734>
29. Wang Y-M, Li Y, Cheng Y-S et al (2019) Deep learning in automated region proposal and diagnosis of chronic otitis media based on computed tomography. *Ear Hear.* 41(3):669–677. <https://doi.org/10.1097/AUD.0000000000000794>
30. Zafer C (2020) Fusing fine-tuned deep features for recognizing different tympanic membranes. *BioCybern Biomed Eng* 40:40–51. <https://doi.org/10.1016/j.bbe.2019.11.001>
31. Singh A, Dutta MK (2021) Diagnosis of ear conditions using deep learning approach. *ICCISC 2021 - 2021 Int Conf Commun Control Inf Sci Proc.* <https://doi.org/10.1109/ICCISC52257.2021.9484919>
32. Uçar M, Akyol K, Atilla Ü, Uçar E (2021) Classification of different tympanic membrane conditions using fused deep hypercolumn features and bidirectional LSTM. *IRBM.* <https://doi.org/10.1016/j.irbm.2021.01.001>
33. Zeng J, Kang W, Chen S et al (2022) A deep learning approach to predict conductive hearing loss in patients with otitis media with effusion using otoscopic images. *JAMA Otolaryngol Neck Surg* 148:612–620. <https://doi.org/10.1001/JAMAOTO.2022.0900>
34. Choi Y, Chae J, Park K et al (2022) Automated multi-class classification for prediction of tympanic membrane changes with deep learning models. *PLoS ONE* 17:e0275846. <https://doi.org/10.1371/JOURNAL.PONE.0275846>
35. Habib AR, Xu Y, Bock K et al (2023) Evaluating the generalizability of deep learning image classification algorithms to detect middle ear disease using otoscopy. *Sci Rep* 13(13):1–9. <https://doi.org/10.1038/s41598-023-31921-0>
36. Nam Y, Choi SJ, Shin J, Lee J (2023) Diagnosis of middle ear diseases based on convolutional neural network. *Comput Syst Sci Eng* 46:1521–1532. <https://doi.org/10.32604/CSSE.2023.034192>
37. Afify HM, Mohammed KK, Hassanien AE (2023) Insight into automatic image diagnosis of ear conditions based on optimized deep learning approach. *Ann Biomed Eng* 1:1–12. <https://doi.org/10.1007/S10439-023-03422-8/TABLES/7>
38. Wang Z, Song J, Su R et al (2022) Structure-aware deep learning for chronic middle ear disease. *Expert Syst Appl* 194:116519. <https://doi.org/10.1016/J.ESWA.2022.116519>
39. Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. *Proc IEEE* 86:2278–2324. <https://doi.org/10.1109/5.726791>
40. Szegedy C, Vanhoucke V, Ioffe S et al (2016) Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* pp 2818–2826. <https://doi.org/10.1109/CVPR.2016.308>
41. Huang G, Liu Z, van der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* <https://doi.org/10.1109/CVPR.2017.243>
42. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. *arXiv Prepr arXiv14091556*
43. Howard AG, Zhu M, Chen B et al (2017) Mobilenets: efficient convolutional neural networks for mobile vision applications. *arXiv Prepr arXiv170404861*
44. Tan M, Le QV (2019) EfficientNet: rethinking model scaling for convolutional neural networks. *36th Int Conf Mach Learn ICML 2019 June*:10691–10700
45. Weiss K, Khoshgoftaar TM, Wang D (2016) A survey of transfer learning. *J Big Data* 3:9. <https://doi.org/10.1186/s40537-016-0043-6>
46. Dietterich TG (2000) Ensemble methods in machine learning. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* 1857 LNCS:1–15. https://doi.org/10.1007/3-540-45014-9_1/COVER
47. Manconi A, Armano G, Gnocchi M, Milanese L (2022) A Soft-voting ensemble classifier for detecting patients affected by COVID-19. *Appl Sci* 12. <https://doi.org/10.3390/app12157554>

48. Kumari S, Kumar D, Mittal M (2021) An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *Int J Cogn Comput Eng* 2:40–46. <https://doi.org/10.1016/J.IJCCCE.2021.01.001>
49. Chandra TB, Verma K, Singh BK et al (2021) Coronavirus disease (COVID-19) detection in Chest X-Ray images using majority voting based classifier ensemble. *Expert Syst Appl* 165:113909. <https://doi.org/10.1016/J.ESWA.2020.113909>
50. Saha S, Ekbal A (2013) Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition. *Data Knowl Eng* 85:15–39. <https://doi.org/10.1016/J.DATAK.2012.06.003>
51. Yu X, Zhang Z, Wu L et al (2020) Deep ensemble learning for human action recognition in still images. *Complexity* 2020:9428612. <https://doi.org/10.1155/2020/9428612>
52. Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. *Inf Process Manag* 45:427–437. <https://doi.org/10.1016/J.IPM.2009.03.002>
53. Al Afandy KA, Omara H, Lazaar M, Al Achhab M (2022) Deep learning. *Approaches Appl Deep Learn Virtual Med Care*:127–166. <https://doi.org/10.4018/978-1-7998-8929-8.CH006>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Kemal Akyol¹  · Emine Uçar²  · Ümit Atıla³  · Murat Uçar⁴ 

✉ Murat Uçar
murat.ucar@bakircay.edu.tr

Kemal Akyol
kakyol@kastamonu.edu.tr

Emine Uçar
emine.ucar@bakircay.edu.tr

Ümit Atıla
umitatila@gazi.edu.tr

¹ Department of Computer Engineering, Faculty of Engineering and Architecture, Kastamonu University, Kastamonu, Turkey

² Department of Management Information Systems, Faculty of Economics and Administrative Sciences, İzmir Bakırçay University, İzmir, Turkey

³ Department of Computer Engineering, Faculty of Engineering, Gazi University, Ankara, Turkey

⁴ Department of Computer Engineering, Faculty of Engineering and Architecture, İzmir Bakırçay University, İzmir, Turkey