A Deep Learning Approach for Detecting Otitis Media From Wideband Tympanometry Measurements

Josefine Vilsbøll Sundgaard, Peter Bray, Søren Laugesen, James Harte, Yosuke Kamide,
Chiemi Tanaka, Anders Nymark Christensen, and Rasmus R. Paulsen

***Abstract*—Objective: In this study, we propose an automatic diagnostic algorithm for detecting otitis media based on wideband tympanometry measurements. Methods: We develop a convolutional neural network for classification of otitis media based on the analysis of the wideband tympanogram. Saliency maps are computed to gain insight into the decision process of the convolutional neural network. Finally, we attempt to distinguish between otitis media with effusion and acute otitis media, a clinical subclassification important for the choice of treatment. Results: The approach shows high performance on the overall otitis media detection with an accuracy of 92.6%. However, the approach is not able to distinguish between specific types of otitis media. Conclusion: Out approach can detect otitis media with high accuracy and the wideband tympanogram holds more diagnostic information than the commonly used techniques wideband absorbance measurements and simple tympanograms. Significance: This study shows how advanced deep learning methods enable automatic diagnosis of otitis media based on wideband tympanometry measurements, which could become a valuable diagnostic tool.**

***Index Terms*—Computer-aided diagnosis, convolutional neural network, deep learning, wideband tympanometry.**

I. Introduction

OTITIS media (OM) is an inflammation in the middle ear. The condition is divided clinically into two diagnostic groups: acute otitis media (AOM) and otitis media with effusion (OME). Acute otitis media is characterized by an acute infection with a rapid onset, while OME is characterized by the presence of fluid in the middle ear. Both types are extremely common among children, and OM is one of the most common reasons for medical consultations for children at primary-care physicians [1].

Even though AOM and OME are similar, their clinical classification is important because antibiotics are only recommended for the treatment of AOM, which is caused by infections. Antibiotics are not used to treat OME as it is self-limiting and is not an infection. Diagnosing which type of OM a patient has is challenging. The condition is usually assessed with an otoscope that allows the doctor to obtain a visual impression of the patient's eardrum. This technique requires specific training and diagnosis has been shown to be highly subjective [2]. In response to these challenges, the present authors have previously demonstrated the advantages of applying deep learning methods for automatic identification of otitis media in otoscopy images [3].

In this paper, we turn our attention to another technique that can be used to diagnose middle ear conditions - tympanometry. This technique characterizes the ear canal acoustically by using a range of positive and negative pressure offsets. From this, one can derive conclusions about both eardrum mobility and middle ear condition. Tympanometry objectively evaluates the energy transmission through the middle ear without assessing the sensitivity of hearing.

Standard absorbance tympanometry is performed by using an acoustic probe with an airtight seal in the ear canal, as shown in Fig. 1. This probe presents a tone into the ear canal, typically at a frequency of 226 Hz or 1 kHz and around 85 dB SPL (sound pressure level), and uses a microphone to measure the sound. The choice of frequency depends on the patient, 226 Hz is used for adults, whereas 1 kHz is used in pediatric tympanometry. The resultant sound pressure level in the ear canal is determined by the relative proportions of absorbed and reflected sound energy. During the measurement, the instrument changes the pressure in the ear canal, typically from +200 to -400 daPa. The proportion of absorbed energy changes as the changes in pressure alter the eardrum tension and displace the attached middle ear structures. These changes are typically plotted as a tympanogram [4], which is a graph of admittance versus pressure, since this provides the greatest diagnostic utility.

Manuscript received September 7, 2021; revised January 21, 2022; accepted March 4, 2022. Date of publication March 15, 2022; date of current version July 4, 2022. This work was supported by William Demant Foundation. *(Anders Nymark Christensen and Rasmus R. Paulsen are co-first authors.) (Corresponding author: Josefine Vilsbøll Sundgaard.)*

Josefine Vilsbøll Sundgaard, Anders Nymark Christensen, and Rasmus R. Paulsen are with the Department of Applied Mathematics and Computer Science, Technical University of Denmark, 2800 Kongens Lyngby, Denmark (e-mail: josh@dtu.dk; anym@dtu.dk; rapa@dtu.dk).

Peter Bray is with Interacoustics A/S, 5500 Middelfart, Denmark (e-mail: pebr@dgs.com).

Søren Laugesen and James Harte are with the Interacoustics Research Unit C/O, Technical University of Denmark, 2800 Kongens Lyngby, Denmark (e-mail: slau@iru.interacoustics.com; jmha@eriksholm.com).

Yosuke Kamide is with the Kamide ENT Clinic, Shizuoka, Japan (e-mail: dr.ent@icloud.com).

Chiemi Tanaka is with Diatec Japan K.K., Kanagawa, Japan (e-mail: ceta@demant.com).

Digital Object Identifier 10.1109/JBHI.2022.3159263

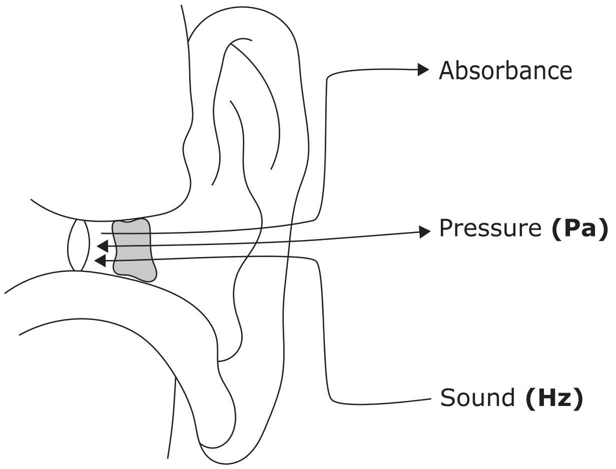


Fig. 1. Measurement of a WBT. The pressure in the middle ear is changed while a sound at specific frequencies is presented. The instrument then records the reflected sound from the eardrum and thus computes the absorbance.

Wideband tympanometry (WBT) is an extension to standard tympanometry in that it measures the ear canal's acoustic properties over a range of frequencies [5], [6]. The use of a wideband stimulus (i.e., short duration rectangular pulse or a chirp covering the range of 226 Hz to 8000 Hz) has been shown to be more efficient and precise for middle ear assessment [7]–[11] than a normal 226 Hz or 1 kHz tympanogram, since it simultaneously determines the characteristics of the middle ear over the full range of the audiometrically most important frequencies. Because of the presence of multiple frequencies in the transient stimuli, WBT is less susceptible to myogenic noise, which originates from the patient's movements [4].

Assessment of middle ear function over this broad bandwidth provides detailed information on the middle ear status and can assist considerably with diagnosis. Higher absorbance values suggest a more efficient middle ear transmission of sound, as shown in Fig. 2(c). Fig. 2(a) and (b) show how lower values mean that the eardrum cannot move properly, which could be caused by increased stiffness in the ossicular chain, or a fluid-filled middle ear. Fig. 2(c) shows a WBT of a patient with no effusion (NOE), and thus a healthy middle ear. The average NOE WBT shows change in absorbance on the pressure axis. Fig. 2(a) and (b) presents with a flat absorbance across various pressure values, indicating reduced eardrum mobility due to otitis media.

Clinical assessment of OM using WBT could benefit from an automatic diagnostic system designed to assist medical experts when diagnosing patients. As described above, WBT is an objective measurement, and it has been established that it can be successfully used to diagnose OM. Further, its traditional use requires specific training of hearing care professionals to allow them to interpret WBT results to diagnose OM. Thus an automatic diagnostic system could prove a useful clinical tool.

The contributions of this paper include the development of a 2D convolutional neural network designed and trained to perform fully automatic classification of OM from WBT measurements. The analysis is conducted on the full WBT without the need for any manual feature extraction. We compare the diagnostic value of the full WBT measurements with that of the more traditional measurements: ambient absorbance and the 0.375-2 kHz averaged tympanogram.

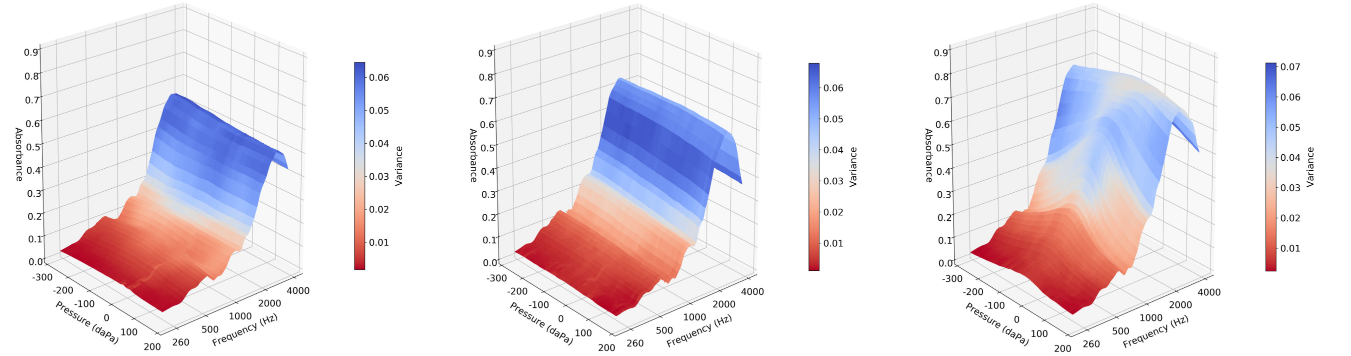
We are the first to include AOM in our classification pipeline, and our proposed approach outperforms previous state-of-the-art methods for binary classification of OM and NOE. We compute saliency maps for the WBT classification to investigate the most important features of the WBT for the diagnosis of OM and compare the key regions with the findings in previous studies. The tools we present in this paper can be used by clinicians to diagnose OM with 92.6% accuracy. Furthermore, by inspecting the saliency maps, clinicians can gain valuable insights into the decision process of the neural network.

A. Related Works

Tympanometry provides quantitative information about the presence of fluid in the middle ear, about the mobility of the tympanic-ossicular system, and about the volume of the external auditory canal. The standard tympanometry method has limitations, including lack of specific norms for different population types (children, infants, adults), as the eardrum and external ear canal are anatomically different in children and adults [4], and specific norms for different diagnostic conditions such as OM. The accuracy of tympanometry in detecting OME has been examined by Palmu *et al.* [12] and Harris *et al.* [13]. Both studies concluded that tympanometry has both high sensitivity to and specificity for OME. [13] has shown that WBT provides more detailed information on the mechanical and acoustic status of the middle ear than the standard 226 Hz tympanogram. Terzi *et al.* [10] employed a receiver operating characteristic (ROC) test to distinguish between NOE and OME cases based on WBT measurements from pediatric patients, and compared the diagnostic value of averaging the absorbance values centered at different frequencies and using different frequency ranges. The highest diagnostic value was found for the 0.375-2 kHz average, followed by the 1 kHz mean and the 1.5 kHz mean. Ellison *et al.* [8] analyzed measurements only at ambient pressure using a likelihood-ratio classifier and found that the absorbance is sensitive to middle ear stiffness and middle ear effusion. They found that the highest classification performance was achieved when employing the full frequency range (0.25 Hz to 8 kHz), while the bandwidth of frequencies from 800 Hz to 2 kHz was the one most affected by eardrum stiffness. Aithal *et al.* [14] showed that wideband absorbance at ambient pressure and tympanometry peak pressure can successfully be used to detect OME, although not significantly better than a 226 Hz tympanogram.

Recent studies have thus shown an interest in automatic classification of these measurements. So far, this has been limited to the binary classification of OM and NOE. Merchant *et al.* [15] created a multivariate prediction model based on the three first principal components using logistic regression, showing good results. Their study concludes that wideband absorbance is a strong and sensitive indicator of the effusion volume.

More advanced machine learning and, in particular, deep learning models are the state of the art for most classification tasks in all data domains, as seen in the current literature [16]–[18]. This development is also seen in the field of tympanometry classification. Binol *et al.* [19] automatically detected



(a) AOM (b) OME (c) NOE

Fig. 2. Average WBT across all subjects in the dataset: acute otitis media (a), otitis media with effusion (b), and no effusion (c) cases. Color scale shows the variance across the measurements within each class.

NOE or OME based on a combination of otoscopy imaging and tympanograms. Their analysis used a random forest classifier on selected features (peak admittance, peak pressure, width of the tympanogram, and ear canal volume) from a standard 226 Hz tympanogram, which was combined using majority voting with the output of a convolutional neural network predicting diagnosis based on the otoscopy image of the patient. Grais *et al.* [20] employed several machine learning methods to analyze the WBT measurements, and found the convolutional neural network to be the best performing approach. They also used a random forest model to produce class activation maps that were used to interpret the diagnostic decision.

II. Data

The data used for this study include WBT measurements collected at Kamide ENT clinic, Shizouka, Japan, from patients aged between 2 months and 12 years. The data collection had ethical approval from the Non-Profit Organization MINS Institutional Review Board (reference number 190221). The measurements were performed using the Titan system (Interacoustics, Denmark). Similarly to standard absorbance tympanometry, a WBT measurement is performed by inserting, and hermetically sealing, an acoustic probe with an appropriately sized silicone ear tip into the patient's ear canal. The probe repeatedly presents a transient stimulus with a frequency range encompassing 226 Hz to 8 kHz while modifying the pressure in the external acoustic canal from 200 to -300 daPa [4]. Diagnosis was decided by an experienced ear-nose-throat (ENT) specialist based on signs, symptoms, patient history, otoscopy examination, and the WBT measurement.

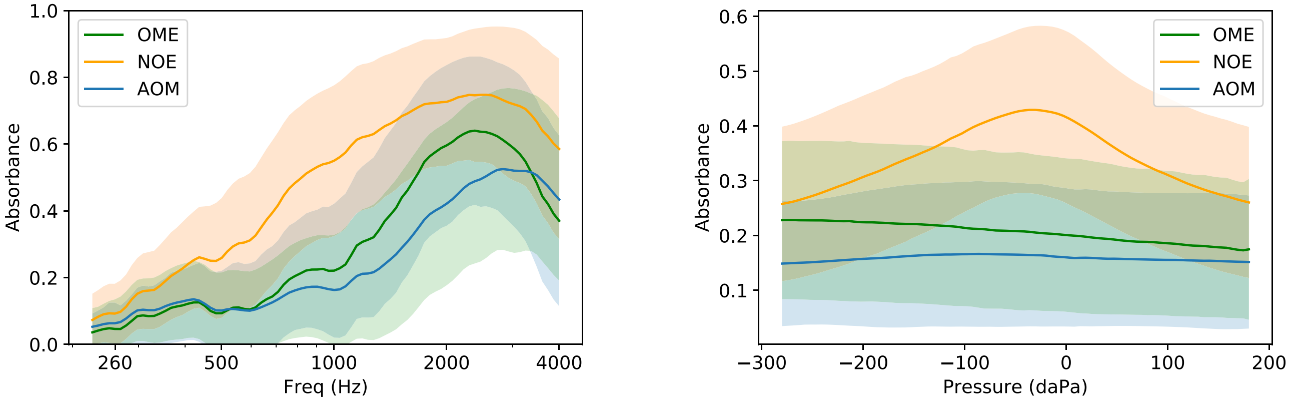
A WBT measurement was excluded from the dataset if the minimum pressure was above -280 daPa, or the maximum pressure was below 180 daPa, or if the measurement consisted of less than 20 pressure samples. If these conditions were not met, it was assumed that there had been an air leak between the probe and the ear canal during measurement, and the pressurization therefore failed. Across WBT measurements, pressure intervals are not uniformly sampled, as a pressure sweep (gradual increase and then decrease) is applied while acoustic stimuli are presented in series. The total number of measurements on the pressure axis therefore varies between measurements. For the purpose of analysis, the frequency axis sampled regularly on a logarithmic scale for each measurement. Measurements above 4 kHz are very prone to noise, and little diagnostic value is found in this high frequency range [21]. A common grid is therefore defined from 180 daPa to -280 daPa in 84 steps on a linear scale, and from 226 Hz to 4 kHz in 84 steps sampled on a logarithmic scale. All WBT measurements are resampled to fit this grid using bilinear interpolation.

The dataset thus consists of 1014 WBT measurements from both left and right ears, separated into the three diagnostic groups: no effusion (NOE, 488 measurements), otitis media with effusion (OME, 372 measurements), and acute otitis media (AOM, 154 measurements). The average WBT measurements for each diagnostic group and variance within each group are shown in Fig. 2. The dataset was split into training (80%) and test (20%) sets, and the training set was further split into a training (80%) and validation (20%) set. It was ensured that data from each patient were only used for either training, validation, or testing.

From the WBT measurement, it is possible to extract a simple tympanogram and an absorbance measurement, which are also commonly used to assess middle ear conditions. The absorbance measurement is extracted at ambient pressure and displays the absorbance across frequency without pressure alterations. A simple tympanogram shows the absorbance change as a function of the pressure variation in the middle ear at a certain frequency. Based on the findings from [10], the average absorbance over the range 0.375-2 kHz was selected to create the averaged tympanogram. These two measures were extracted from all WBT measurements in the dataset after preprocessing. Fig. 3 shows the average ambient absorbance and averaged tympanogram for each of the three diagnostic groups together with the standard deviation within each group, showing considerable overlap across all frequencies, but clear morphological differences.

III. Methods

The first approach is developed to classify no effusion (NOE) and otitis media (a combined group of AOM and OME, denoted OM). The conditions AOM and OME show considerable overlap and similarities, and we therefore start by separating the overall groups NOE and OM. Later, we will attempt to



(a) Ambient absorbance

(b) 0.375-2 kHz averaged tympanogram

Fig. 3. Average ambient absorbance measurements (a) and 0.375-2 kHz averaged tympanogram (b) of each diagnostic group: OME (green), NOE (orange), and AOM (blue). The faded background curves show the standard deviation of each group.

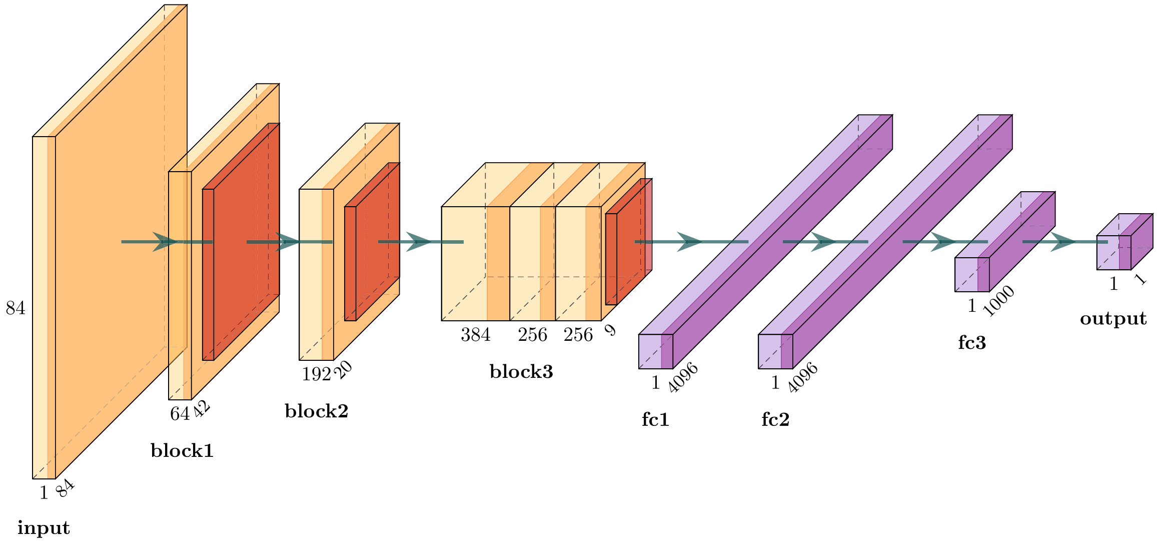


Fig. 4. 2D network architecture. The first number at the bottom of each block is the number of features, the second number shows the dimension (the dimension is the same for height and width of the feature maps). Details about each layer are provided in Table I.

automatically distinguish between AOM and OME. This section is divided into the following parts: WBT classification using a 2D convolutional neural network; ambient absorbance and averaged tympanogram classification using a 1D convolutional neural network; data augmentation; comparison with related methods; saliency maps for WBT classification; and finally, classification of AOM, OME and NOE.

A. WBT Classification

A 2D convolutional neural network is employed for the classification of NOE and OM. The network structure is shown in Fig. 4, and more details about each layer are presented in Table I. The input to the network is the one-channel 84×84 WBT. Through repeated 2D convolution and max pooling, features are extracted from the WBT, and finally the output of the network

TABLE I
2D Neural Network Structure

	Output size (ch, w, h)	Details
Input	(1, 84, 84)	
2D convolution	(64, 42, 42)	Kernel: 5x5, stride: 2, pad: 2
Max pooling	(64, 20, 20)	Kernel: 3x3, stride: 2
2D convolution	(192, 20, 20)	Kernel: 5x5, stride: 2, pad: 2
Max pooling	(192, 9, 9)	Kernel: 3x3, stride: 2
2D convolution	(384, 9, 9)	Kernel: 3x3, stride: 1, pad: 1
2D convolution	(256, 9, 9)	Kernel: 3x3, stride: 1, pad: 1
2D convolution	(256, 9, 9)	Kernel: 3x3, stride: 1, pad: 1
Max pooling	(256, 7, 7)	Kernel: 3x3, stride: 1
Dropout + Linear layer + ReLu	(4096)	Dropout: 0.5
Dropout + Linear layer + ReLu	(4096)	Dropout: 0.5
Dropout + Linear layer + ReLu	(1000)	Dropout: 0.5
Linear layer	(1)	Dropout: 0.5

indicates the probability of OM presence. The architecture of the network was designed specifically for the characteristics of the WBT measurements, with inspiration from the AlexNet architecture [22]. State-of-the-art convolutional neural networks such as ResNet [23], VGG [24], or Inception V3 [25] are all large-scale networks for image classification. PyTorch provides pre-trained versions of these networks, trained on the ImageNet database [26] with input dimensions of 224 × 244, or 299 × 299, depending on the network architecture. This is helpful when limited data are available for training for an image classification task. However, the WBT data are of a completely different nature than the images of the ImageNet database, as the WBT measurements are measured signals, not images. Furthermore, the WBT data are rather simple compared to an image, and do not require a large-scale network for classification. The input dimensions are much lower (84 × 84), the input only consists of one channel, and the measurements consist of fewer details compared to images, as seen in Fig. 2. It is therefore not feasible, nor necessary, to employ a pre-trained network for this task. Since the network employed for this classification task has to be trained end-to-end, we need to limit the amount of parameters, and thus the size of the model. We have therefore designed a 2D convolutional neural network for this specific classification task for WBT measurements, customized to the input WBT size and requirements of this data type.

The neural network is trained end-to-end with a binary cross entropy loss function using the Adam optimizer [27] with a learning rate of 0.0001, which is decreased with a multiplicative factor of 0.1 every 8$^{\text{th}}$ epoch. Batch size is set to 16, all training inputs are shuffled for each epoch, and early stopping is employed with a patience of 20 epochs. The final classification is obtained from the probability output with a threshold value of 0.5.

B. Absorbance and Tympanogram Classification

Two 1D convolutional neural networks with a similar structure to the 2D networks for WBT classification are employed for the classification of ambient absorbance measurements and 0.365-2 kHz averaged tympanograms. Two 1D networks are trained separately for the two tasks. The networks have the same architecture as shown in Table I, only using 1D operations instead of 2D operations. The input is a (1, 84) tensor (absorbance or tympanogram), and thus all output sizes in the table are the same, except using only one dimension instead of two. The last linear layers have output dimensions (1024), (1024), (1000), and (1) due to the reduced input dimensions. The training parameters are also the same as for the WBT neural network.

C. Data Augmentation

Extensive data augmentation is employed to improve training and to avoid overfitting [28]. When performing image classification using convolutional neural networks, data augmentation usually consists of geometric transformations. However, the WBT measurements will always be specified on the same grid, i.e., the features of the WBT will be in the same location of the measurement across different measurements. Geometric transformations such as rotation and translation are therefore not appropriate for this application. Instead, various types of noise and other distortions are generated: Random Gaussian noise is added to the input with intensities up to 0.1 of the maximum value in the measurement; exponential noise with exponentially increasing intensity across the frequency axis, and with no change across the pressure axis; intensity shift, where a constant between -0.2 and 0.2 is added to all intensities in the input; intensity manipulation, where the input is multiplied with a constant between 0.8 and 1.2; random erasing, where a randomly selected region of the input is erased by setting all values in the region to the mean value of the input measurement [29]; and Gaussian hilly terrain, where a mixture of Gaussian functions with various intensities are added to the input to generate noise affecting a larger area in the input than the random noise. Note that Gaussian hilly terrain changes the landscape of the input to a larger extent than the other distortion methods.

Each of the distortion methods are added to the measurements during training with a probability of 0.5. After performing the augmentation, the intensity of the input is ensured to be between 0 and 1, which are the natural boundaries of WBT. The various types of data augmentation can be performed in both 2D and 1D, and are therefore employed during training for all classification networks. It is, however, unknown whether all types of data augmentation increase performance in both 1D and 2D. Experiments were therefore run with all three networks, examining each type of data augmentation.

D. Comparisons

Besides our proposed methods, we have also run experiments with the methods proposed by Merchant *et al.* [15] and Grais *et al.* [20] for comparison. These methods were trained and tested using our dataset to ensure a proper comparison. Merchant *et al.* [15] propose an approach based on a multivariate logistic classification model based on the three first principal components of the WBT measurements. We trained the binary classification model to predict OM or NOE, and tested it on our test dataset. Grais *et al.* [20] compared several machine learning methods for the classification of OM and NOE based on WBT measurements. They show that the CNN is superior to a fully connected neural network, random forest model, support vector machine, and a k-NN. Since they have provided this detailed comparison with other machine learning algorithms, we will refrain from performing the same experiments, and compare our approach with their best performing CNN. The CNN is implemented as described in the paper, and trained and tested on our dataset.

E. Saliency Maps

A saliency map is a representation of the unique importance of each pixel or neuron in the network input. The purpose of these maps is to visualize the feature maps of a neural network, and thus use the visual representation to interpret the decision process of a neural network. This attempt to interpret and analyze the output of a neural network can build trust in the model amongst its users, enable understanding of the model, and ease the integration of systems such as this into, for example, clinical practice.

A variety of methods for output explanation from deep neural networks exist, as seen in the survey by Singh *et al.* [30]. For this pipeline, the widely used method of GradCAM [31] is implemented and applied to the WBT classification network. GradCAM is a generalization of class activation maps (CAM), in which gradient information from the last convolutional layer of the convolutional neural network is used to understand the importance of each neuron in the feature maps. Convolutional neural networks retain spatial information throughout the network until it is lost in the final fully connected layers. The last convolutional layer will therefore have the best trade-off between high-level features and detailed spatial information.

The saliency maps are generated in several steps. The first step is to compute the gradient of the class score for each feature map in the last convolutional layer. A weighted combination of all feature maps is computed using the class scores as weights, and finally, a ReLU activation is performed to ensure that only positive influences on the output class are included. This results in a coarse saliency map of the same size as the feature maps in the last convolutional layer (in this case 9×9). The coarse map is upsampled using bilinear interpolation to obtain a full input size heat map of 84×84.

F. Classification of AOM and OME

Finally, an approach to distinguish between AOM, OME, and NOE based on the full WBT measurement is investigated. It has not previously been shown or demonstrated that it is possible use WBT to distinguish the two types of otitis media. Other studies such as [8], [10], [20] only include OME cases, and not AOM. Helenius *et al.* [32] investigated discrimination of diagnosis based on standard 226 Hz tympanometry, and found that this measurement can be used to distinguish between NOE and OM cases, but not to diagnose specific types of OM. The present study therefore examines if the additional information provided by WBT (compared to a 226 Hz tympanogram) allows for a specific diagnosis of types of OM.

This approach follows the same architecture as the binary classification network for WBT classification described in Section III-A. The only changes are the input data, which are now from three different classes, because the OM class is divided into OME and AOM, and the class-weighted cross-entropy loss function is utilized during training to cope with the imbalance in the dataset due to fewer AOM cases.

IV. Results

The performance of OM detection on the test set with the three different models is presented in Table II. The performance metrics include accuracy, area under the curve (AUC) (which shows how well the model separates the two classes), sensitivity, specificity, and F1-score. Since sensitivity and specificity are inversely proportional to each other, there is always a trade-off between the two measures. The F1-score (the harmonic mean of the precision and recall of a test) is therefore computed to ease comparison. The models were trained using the best-suited data augmentation methods for each method, as shown in Table III, and for the full WBT CNN, the performance results in Table II are shown both with and without augmentation. The same comparison can be found in Table III for the 1D networks. The rest of the presented results are generated with the full WBT approach, as this approach shows the highest performance. Examples of misclassified measurements are shown in Fig. 5, separated into false positives (representative selection from eight measurements) and false negatives (representative selection from nine measurements).

TABLE II
Otitis Media Classification Performance for WBT, Ambient Absorbance (absorb.), and Averaged Tympanogram (tymp.) Networks on the Test Set

	Acc.	AUC	Sens.	Spec.	F1-score
Merchant et al. [15]	73.4%	0.84	73.3%	73.5%	73.6%
Grais et al. [20]	88.2%	0.92	87.9%	88.5%	88.3%
Ambient absorb.	86.5%	0.94	91.4%	81.4%	87.2%
Averaged tymp.	90.0%	0.96	**92.2%**	87.6%	90.3%
WBT w/o aug.	90.0%	**0.97**	88.8%	91.2%	90.0%
WBT	**92.6%**	**0.97**	**92.2%**	**92.9%**	**92.6%**

Performance for approaches proposed by merchant *et al.* [15] and grais *et al.* [20] on the test set are also included. Bold font marks the highest performance within each metric.

TABLE III
Effect on Classification Accuracy of Various Types of Data Augmentation on the Three Neural Networks: WBT With 2D Augmentation, Ambient Absorbance and Averaged Tympanogram (tymp.) With 1D Augmentation

	WBT	Ambient absorbance	Averaged tymp.
No aug.	90.0%	85.2%	88.2%
Random noise	90.0% *	85.2% *	88.2% *
Exp. noise	90.1% *	84.3%	88.6% *
Intensity man.	88.6%	86.0% *	88.2% *
Random erasing	91.7% *	85.2% *	90.4% *
Intensity shift	89.5%	84.8%	88.2% *
Hilly terrain	90.3% *	85.6% *	89.0% *
* together	92.6%	86.5%	90.0%

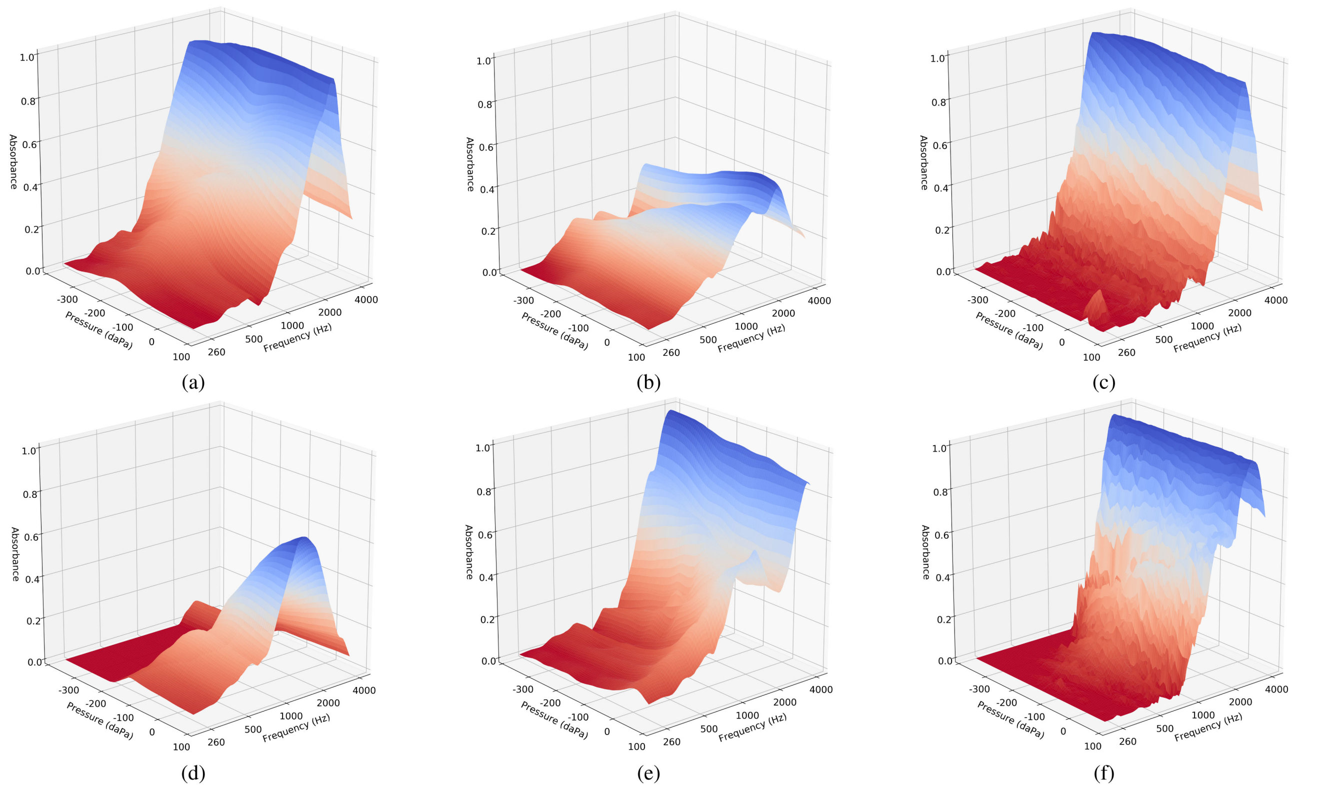
TABLE IV
Performance of Multi-Class Classification (NOE, AOM, and OME)

NOE		AOM		OME		Acc.
Recall	Precision	Recall	Precision	Recall	Precision	
90.3%	90.3%	36.4%	52.2 %	80.7%	72.0%	79.0%

The table shows recall and precision for each class and the overall accuracy

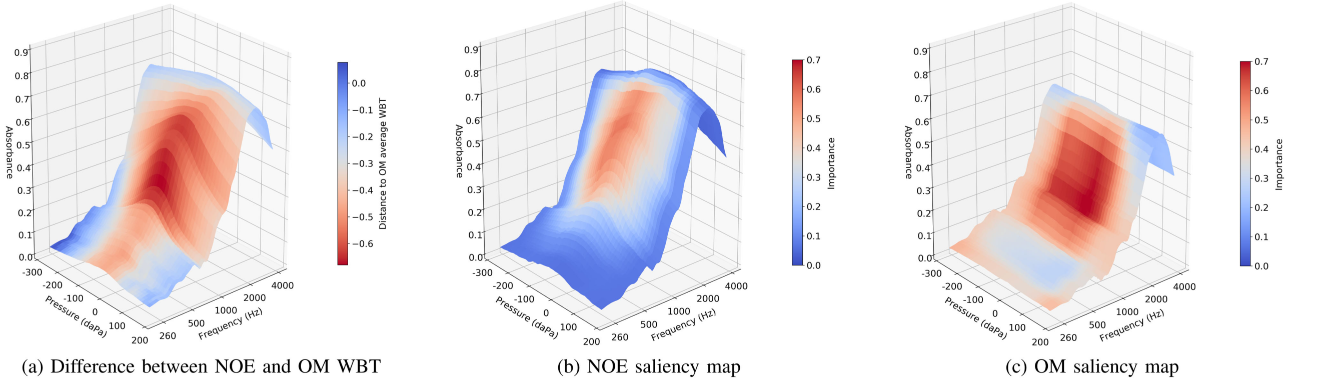
Table III shows the effect of the different types of data augmentation employed during training of the three different neural networks. For each classification approach, the augmentation methods that improve the performance are marked with *. The last row shows the final performance for each of the three with a combination of the augmentation types best suited for the particular network (those marked with *). This shows how the combination of various types of augmentation outperforms each individual type of augmentation. The final combination of augmentation is used for the results presented in both Table II and IV.

Saliency maps are generated for each WBT measurement in the test set using the 2D network for binary classification. An average saliency map is then generated for each class (NOE and



(a) (b) (c) (d) (e) (f)

Fig. 5. Examples of false positive i.e. NOE classified as OM (a, b, c) and false negative i.e. OM classified as NOE (d, e, f) measurements.



(a) Difference between NOE and OM WBT (b) NOE saliency map (c) OM saliency map

Fig. 6. Saliency maps for otitis media classification network. (a) Shows the average NOE WBT with a color map showing the relative difference between average NOE and OM WBTs. (b) and (c) shows the saliency map projected onto the average WBT for each of the two classes. Red areas indicate high importance areas, while blue indicates low importance.

OM) to evaluate the most important features for each diagnostic group. This would not be possible in normal image classification networks, since the object in a natural image can be positioned in various locations in the image. The WBT measurements are however resampled to the same grid, and the features will thus be in the same position across measurements. This means we can compare the saliency maps directly. Fig. 6 shows the average saliency maps for NOE measurements (b) and OM measurements (c). The saliency maps are projected onto the average WBT of each class to ease interpretation of the most important features.

The final approach described in Section III-F attempts to separate the OM classification into either OME or AOM. The performance is shown in Table IV and shows the precision and recall for each class and the overall classification accuracy. The results clearly show how challenging it is to distinguish between AOM and OME based on only the WBT measurement from a patient.

V. Discussion

The classification results in Table II show very high performance in all performance metrics for the WBT approach to classifying NOE from OM cases. The averaged tympanogram and ambient absorbance approaches are inferior to WBT, except for sensitivity, where the WBT and averaged tympanogram approaches are tied. It is clear from the F1-score that the WBT approach has the highest overall performance. The AUC summarizes the overall diagnostic accuracy, and an AUC above 0.9 is considered outstanding [33].

The method proposed by Merchant *et al.* [15] has the lowest performance, and is also the simplest method, as it is based on principal component analysis and logistic regression. The performance of the 2D CNN for WBT classification proposed by Grais *et al.* [20] is comparable to our performance, but still lower. The proposed CNN architecture is simpler than ours, as they employ fewer layers (both convolutional and fully connected layers) and larger convolution kernels in each layer. We show that even without our extensive use of augmentation, our network architecture has a higher performance.

The false positive and negative examples in Fig. 5 show a selection of challenging WBT measurements. These examples show that not all WBT measurements look like the average WBT measurements presented in Fig. 2, and that WBT measurements can have unusual shapes. For example, Fig. 5(c) and (f) look quite similar, but are annotated differently by the ENT. This could indicate that in (f), the primary signs of OM were found in the additional patient data available, such as the otoscopy examination or the patient-reported symptoms, and that WBT does not provide enough information for that particular diagnosis.

Deep learning is generally considered a ‘black box’ approach for classification problems, yet there are several methods that allow users to interpret the decision making behind the results. This is particularly important when developing a diagnostic tool for clinical professionals, to allow them to understand the decision process and trust the decisions made by the neural network. The saliency maps in Fig. 6 introduce valuable insight into the decision strategy of the trained neural network. The average NOE saliency map in Fig. 6(b) clearly shows that the region between 1 and 2 kHz is the key area for a normal WBT measurement, which coincides with the findings in [8], [10], [20]. This corresponds with the physiological resonance frequency of the eardrum around 1 kHz [34], which is affected by membrane stiffness and middle ear fluid present in otitis media cases. Thus, this peak in importance between 1 and 2 kHz can be used to distinguish between a healthy and unhealthy eardrum. The average OM saliency map in Fig. 6(c) shows that the frequency region from 500 Hz to 2 kHz is a key area for this class in a large area on the pressure axis as well, compared to the NOE saliency map. It is clear that abnormal WBT measurements have a much flatter appearance across the pressure axis, together with generally lower absorbance levels, compared to the normal WBT measurement. From the OM saliency map it is clear that the neural network determines the diagnosis of OM from the changes on the pressure axis and the slope from the low to high frequencies.

Heat maps like these allow the expert ENT to evaluate every decision made by the model, and to check that the highlighted regions correspond to the clinical findings. The heat maps can also be used as a training tool for new ENTs or primary-case physicians to learn how to analyze WBT measurements. There are therefore many possible applications of these heat maps.

The results from these heat maps could also explain the lower performance of the tympanometry and wideband absorbance approaches. Since the wideband absorbance measurement does not include knowledge about the variation across pressure, valuable information is missing that is important for the classification. The type of tympanometry considered in this study includes this variation across pressure because it is calculated as an average from 0.375 to 2 kHz, and is also the highest-performing approach of the two 1D approaches. These results show that pressurization during measurement is very valuable and adds diagnostic value to the test.

Our final experiment shows that there are limitations to the diagnostic value of a WBT measurement. While the performance of binary NOE/OM classification is very high, the neural network is challenged when attempting to distinguish between AOM and OME, as seen in Table IV. It is not surprising that this is a difficult task, as indicated by the plots shown in Fig. 3. The plots clearly show that there is substantial overlap between all three groups, but especially the AOM and OME groups have a major overlap. In the lower frequencies of the absorbance measure, the two groups are almost identical, and only a slight difference is seen from 1 to 2 kHz. A similar picture is seen in the averaged tympanograms, where they are both flat but with a slightly different mean absorbance level. A similar result was also found by Helenius *et al.* [32], who only evaluated 226 Hz tympanograms. The results of the present study show that WBT does not demonstrate high performance in diagnosing specific types of OM despite the fact that WBT introduces new information to the diagnosis process. It is, however, satisfying that the neural network has not just over-fitted to the dataset by finding hidden features and creating complex decision strategies in order to perform the classification, when it is clinically questionable that it is possible.

As previously mentioned, WBT measurements will vary between patients of different ages, as the ear structures develop with age. Our dataset covers children from 2 months to 12 years and will thus include different age profiles. It is expected that the neural network learns to model these variations and differences between age groups, and thus incorporates them into the model. It was investigated whether there is a correlation between misclassifications and a certain age group, but none was found. The misclassifications are randomly distributed across ages. It is therefore concluded that age-related changes are not an issue for our approach.

VI. Conclusion

The results of this study show that WBT measurements can be used to determine whether OM is present. The classification results show very high performance, and since this approach is fully automatic with no human input, this bodes well for applying the approach in an automatic diagnostic tool for OM detection. Our study shows that WBT measurements provide more diagnostic information than both the ambient absorbance measure and the 0.375-2 kHz averaged tympanogram. As expected on the basis of clinical practice and pathological studies related to OM, we found that WBT has to be combined with other sources of information about the patient to diagnose specific types of OM.

References

[1] G. Worrall, “Acute otitis media,” *Can. Fam. Physician*, vol. 53, no. 12, pp. 2147–2148, 2007.

[2] M. E. Pichichero and M. D. Poole, “Assessing diagnostic accuracy and tympanocentesis skills in the management of otitis media,” *Arch. Pediatrics Adolesc. Med.*, vol. 155, no. 10, pp. 1137–1142, 2001.

[3] J. V. Sundgaard *et al.*, “Deep metric learning for otitis media classification,” *Med. Image Anal.*, vol. 71, 2021.

[4] T. A. D. Hein, S. Hatzopoulos, P. H. Skarzynski, and M. F. Colella-Santos, “Wideband tympanometry,” *Adv. Clin. Audiol.*, IntechOpen, 2017, doi: 10.5772/67155.

[5] A. Biswas and N. Dutta, “Wideband tympanometry,” *Ann. Otol. Neurotol.*, vol. 1, no. 2, pp. 126–132, 2018.

[6] C. A. Sanford, L. L. Hunter, M. Patrick Feeney, and H. H. Nakajima, “Wideband acoustic immittance: Tympanometric measures,” *Ear Hear.*, vol. 34, no. 1, pp. 65–71, 2013.

[7] A. N. Beers, N. Shahnaz, B. D. Westerberg, and F. K. Kozak, “Wideband reflectance in normal Caucasian and Chinese school-aged children and in children with otitis media with effusion,” *Ear Hear.*, vol. 31, no. 2, pp. 221–233, 2010.

[8] J. C. Ellison, M. Gorga, E. Cohn, D. Fitzpatrick, C. A. Sanford, and D. H. Keefe, “Wideband acoustic transfer functions predict middle-ear effusion,” *Laryngoscope*, vol. 122, no. 4, pp. 887–894, 2012.

[9] D. H. Keefe and J. L. Simmons, “Energy transmittance predicts conductive hearing loss in older children and adults,” *J. Acoustical Soc. Amer.*, vol. 114, no. 6, pp. 3217–3238, 2003.

[10] S. Terzi *et al.*, “Diagnostic value of the wideband acoustic absorbance test in middle-ear effusion,” in *Laryngol. Otol.*, vol. 129, no. 11, 2015, pp. 1078–1084.

[11] L. Stuppert, S. Nospes, A. Bohnert, A. K. Läßig, A. Limberger, and T. Rader, “Clinical benefit of wideband-tympanometry: A pediatric audiology clinical study,” *Eur. Arch. Oto-Rhino-Laryngol.*, vol. 276, no. 9, pp. 2433–2439, 2019.

[12] A. Palmu, H. Puhakka, T. Rahko, and A. K. Takala, “Diagnostic value of tympanometry in infants in clinical practice,” *Int. J. Pediatr. Otorhinolaryngol.*, vol. 49, no. 3, pp. 207–213, 1999.

[13] P. K. Harris, K. M. Hutchinson, and J. Moravec, “The use of tympanometry and pneumatic otoscopy for predicting middle ear disease,” *Amer. J. Audiol.*, vol. 14, no. 1, pp. 3–13, 2005.

[14] V. Aithal, S. Aithal, J. Kei, S. Anderson, and D. Wright, “Predictive accuracy of wideband absorbance at ambient and tympanometric peak pressure conditions in identifying children with surgically confirmed otitis media with effusion,” *J. Amer. Acad. Audiol.*, vol. 31, no. 7, pp. 471–484, 2020.

[15] G. R. Merchant, S. Al-Salim, R. M. Tempero, D. Fitzpatrick, and S. T. Neely, “Improving the differential diagnosis of otitis media with effusion using wideband acoustic immittance,” *Ear Hear.*, vol. 42, no. 5, pp. 1183–1194, 2021.

[16] Z. Guo, L. Tang, T. Guo, K. Yu, M. Alazab, and A. Shalaginov, “Deep graph neural network-based spammer detection under the perspective of heterogeneous cyberspace,” *Future Gener. Comput. Syst.*, vol. 117, pp. 205–218, 2021.

[17] L. Cai, J. Gao, and D. Zhao, “A review of the application of deep learning in medical image classification and segmentation,” *Ann. Transl. Med.*, vol. 8, no. 11, pp. 713–713, 2020.

[18] A. Raza, A. Mehmood, S. Ullah, M. Ahmad, G. S. Choi, and B. W. On, “Heartbeat sound signal classification using deep learning,” *Sensors (Switzerland)*, vol. 19, no. 21, pp. 1–15, 2019.

[19] H. Binol *et al.*, “Decision fusion on image analysis and tympanometry to detect eardrum abnormalities,” in *Med. Imag. 2020: Comput.-Aided Diagnosis, Int. Soc. Opt. Photon.*, vol. 11314, pp. 375–382, 2020

[20] E. M. Grais, X. Wang, J. Wang, F. Zhao, W. Jiang, and Y. Cai, “Analysing wideband absorbance immittance in normal and ears with otitis media with effusion using machine learning,” *Sci. Rep.*, vol. 11, no. 1, pp. 1–12, 2021.

[21] K. R. Nørgaard, K. K. Charaziak, and C. A. Shera, “A comparison of ear-canal-reflectance measurement methods in an ear simulator,” *J. Acoustical Soc. Amer.*, vol. 146, no. 2, pp. 1350–1361, 2019.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Adv. Neural Inf. Process. Syst.*, vol. 25, 2012.

[23] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Dec. 2016, pp. 770–778.

[24] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, *arXiv preprint arXiv:1409.1556*.

[25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.

[26] J. Deng, W. Dong, R. Socher, Li-Jia Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,”, *IEEE conf. comput. vision pattern recognition*, pp. 248–255, 2009.

[27] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” 2014, *arXiv preprint arXiv:1412.6980*.

[28] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, “Understanding data augmentation for classification: When to warp,” in *Int. Conf. Digit. Image Comput., Techn. Appl.*, 2016, pp. 1–6.

[29] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, “Random erasing data augmentation,” *Proc. AAAI Conf. Artificial Intell.*, vol. 34, no. 7, pp. 13001–13008, 2020

[30] A. Singh, S. Sengupta, and V. Lakshminarayanan, “Explainable deep learning models in medical image analysis,” *J. Imag.*, vol. 6, no. 6, pp. 1–19, 2020.

[31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[32] K. K. Helenius, M. K. Laine, P. A. Tähtinen, E. Lahti, and A. Ruohola, “Tympanometry in discrimination of otoscopic diagnoses in young ambulatory children,” *Pediatr. Infect. Dis. J.*, vol. 31, no. 10, pp. 1003–1006, 2012.

[33] J. N. Mandrekar, “Receiver operating characteristic curve in diagnostic test assessment,” *J. Thoracic Oncol.*, vol. 5, no. 9, pp. 1315–1316, 2010.

[34] G. Volandri, F. Di Puccio, P. Forte, and C. Carmignani, “Biomechanics of the tympanic membrane,” *J. Biomech.*, vol. 44, no. 7, pp. 1219–1236, 2011.