



# Multi-modal deep learning for joint prediction of otitis media and diagnostic difficulty

Josefine Vilsbøll Sundgaard PhD<sup>1</sup>  | Morten Rieger Hannemose PhD<sup>1</sup> |  
Søren Laugesen PhD<sup>2</sup>  | Peter Bray PhD<sup>3</sup> | James Harte PhD<sup>2</sup> |  
Yosuke Kamide MD<sup>4</sup> | Chiemi Tanaka PhD<sup>5</sup> | Rasmus R. Paulsen PhD<sup>1</sup> |  
Anders Nymark Christensen PhD<sup>1</sup>

<sup>1</sup>Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark

<sup>2</sup>Interacoustics Research Unit, Technical University of Denmark, Lyngby, Denmark

<sup>3</sup>Interacoustics A/S, Middelfart, Denmark

<sup>4</sup>Kamide ENT Clinic, Shizuoka, Japan

<sup>5</sup>Diatec Japan, Kanagawa, Japan

## Correspondence

Josefine Vilsbøll Sundgaard, DTU Compute, Richard Petersens Plads, Bygning 324, 2800 Kgs. Lyngby, Denmark.  
Email: [josh@dtu.dk](mailto:josh@dtu.dk)

## Funding information

William Demant Fonden

## Abstract

**Objectives:** In this study, we propose a diagnostic model for automatic detection of otitis media based on combined input of otoscopy images and wideband tympanometry measurements.

**Methods:** We present a neural network-based model for the joint prediction of otitis media and diagnostic difficulty. We use the subclassifications acute otitis media and otitis media with effusion. The proposed approach is based on deep metric learning, and we compare this with the performance of a standard multi-task network.

**Results:** The proposed deep metric approach shows good performance on both tasks, and we show that the multi-modal input increases the performance for both classification and difficulty estimation compared to the models trained on the modalities separately. An accuracy of 86.5% is achieved for the classification task, and a Kendall rank correlation coefficient of 0.45 is achieved for difficulty estimation, corresponding to a correct ranking of 72.6% of the cases.

**Conclusion:** This study demonstrates the strengths of a multi-modal diagnostic tool using both otoscopy images and wideband tympanometry measurements for the diagnosis of otitis media. Furthermore, we show that deep metric learning improves the performance of the models.

## KEYWORDS

computer-aided diagnosis, deep learning, diagnostic difficulty, otitis media

## 1 | INTRODUCTION

Automatic diagnosis of otitis media has been tackled in various ways. Previous studies have employed datasets of otoscopy images,<sup>1-4</sup>

tympanometry measurements<sup>5-7</sup> optical coherence tomography,<sup>8</sup> or computed tomography.<sup>9</sup> The approaches have utilized a variety of machine learning algorithms for the data analysis and classification task, progressing from simpler methods such as Random Forest<sup>10</sup> and Support Vector Machines,<sup>11</sup> to advanced deep neural networks.<sup>1,5,7,12,13</sup> When a doctor examines a patient, the diagnostic

Rasmus R. Paulsen and Anders Nymark Christensen are shared senior authorship.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *Laryngoscope Investigative Otolaryngology* published by Wiley Periodicals LLC on behalf of The Triological Society.

decision is rarely based solely on one modality from the clinical examination. Binol et al.<sup>14</sup> were the first to combine otoscopy images and standard 226-Hz tympanometry measurements for the classification of normal or abnormal middle ear. The standard tympanometry analysis was based on manually selected features including peak admittance, peak pressure, tympanometric width, and ear canal volume, which were fed to a Random Forest model. The otoscopy analysis was based on a pre-trained Inception-ResNet-V2 network, fine-tuned for the specific classification task. The classification decisions of these two models were fused using majority voting for the final classification. The method was demonstrated on a limited dataset of 73 cases, and the evaluation was thus performed using leave-one-out cross-validation. Furthermore, they showed that the combination of otoscopy images and standard tympanograms outperformed the classification based on the individual modalities. Wideband tympanometry (WBT) has shown to be more efficient in evaluating the condition of the middle ear, and it provides more detailed information on the mechanical and acoustic status of the middle ear than the standard 226-Hz tympanogram.<sup>15</sup> Furthermore, higher classification accuracy can be achieved using WBT measurements for the detection of otitis media, compared to both ambient absorbance and standard tympanograms.<sup>7</sup> We propose the use of WBT measurements in combination with otoscopy images for the diagnosis of otitis media in the diagnostic groups: otitis media with effusion (OME), acute otitis media (AOM), and no effusion (NOE). This is important to ensure proper treatment, as antibiotics are only needed to treat acute otitis media, not otitis media with effusion.

There has been an increasing interest in neural network-based diagnosis of otitis media, or other middle ear conditions, based on otoscopy images. Habib et al.<sup>16</sup> recently published a review on this topic, including 39 papers published over the last 10 years. They conclude that these classification models are more accurate than human assessors and that the next big task in this field is to implement these methods into a clinical tool that doctors can—and want to—use. An important aspect of this step is allowing the user of a clinical tool to learn more from the model than just the diagnosis. Several studies have employed saliency maps to allow the user to learn about the decision process of the model by identifying the most important features of the input data.<sup>7,17</sup> Another valuable output would be an estimate of the difficulty of the input case. Combining the diagnosis with the estimated difficulty allows the operator to assess the output of the model and to evaluate whether to redo the otoscopy or WBT or refer the patient to an expert ENT for further examination. Diagnostic difficulty estimation was investigated by Hannemose et al.<sup>18</sup> They present several supervised and unsupervised methods for estimating the difficulty from the distribution of the dataset in the embedding space from a metric learning-based neural network.

The goal of this work is to predict both diagnostic class and difficulty for each case, and we evaluate two methods for this task. One of the proposed methods is a deep metric learning approach, which means that the output of the neural network is a low dimensional representation of the input image, denoted an embedding, instead of a classification output. Deep metric learning has been used for learning

feature representations of otoscopy images in other studies with promising results.<sup>1,19,20</sup> The image embeddings can then be used to predict diagnostic class and using the supervised method presented by Hannemose et al.<sup>18</sup> The other approach is a multi-task network for joint prediction, which learns the prediction tasks end-to-end. We are the first to propose a purely neural network-based model for the analysis of otoscopy images and WBT measurements combined into a single model. Furthermore, our models are developed for joint prediction of otitis media and diagnostic difficulty.

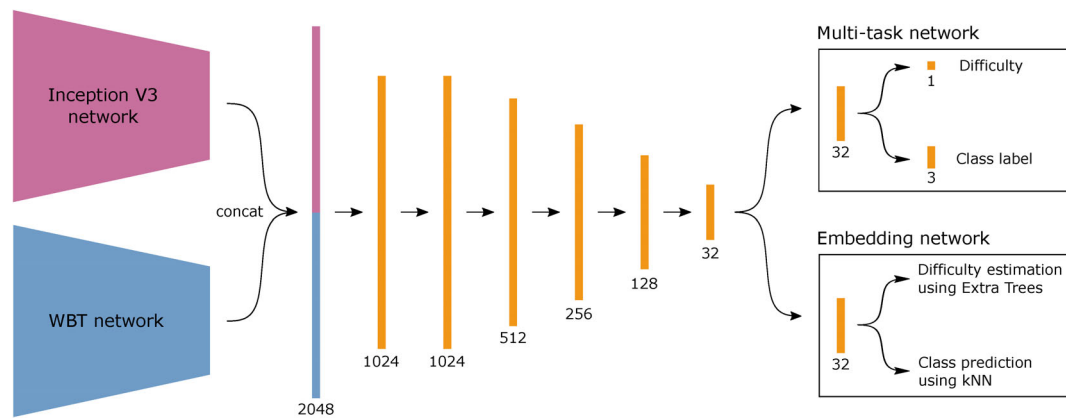
## 2 | METHODS

We propose a single network for the combined analysis of otoscopy images and WBT measurements. The network architecture, seen in Figure 1, consists of a pre-trained Inception V3<sup>21</sup> network for the otoscopy image input and a network designed specifically for the analysis of WBT measurements using the architecture proposed by Sundgaard et al.<sup>7</sup> The results presented by Sundgaard et al.<sup>7</sup> show that using the full WBT measurements for the detection of otitis media is superior to both ambient absorbance and standard tympanograms. Thus, we only use the full WBT measurements in the current study. The outputs of both these networks are feature vectors of size 1024, which are concatenated and sent through a series of fully connected layers. These fully connected layers ensure that the network learns to combine the feature vectors from the two different inputs into a single decision. The size of the layers is gradually decreased through the fully connected layers, ending at the 32-dimensional final-layer vector. The single modality models have the same linear layers after the main convolution blocks, but the first layer is of size 1024, instead of 2048.

We compare two different training procedures for the network: one based on multi-task learning, and another based on deep metric learning for embedding prediction:

**Multi-task learning:** This network is trained end-to-end for simultaneous prediction of otitis media diagnosis and diagnostic difficulty. The final layers of this network consist of two fully connected layers after the 32-dimensional output, one with a single output for the difficulty and another with a softmax output with size 3 for the classification output. During training, the loss function for this network has two terms: an L1-loss for the difficulty output and a class-weighted cross-entropy loss for the classification, using the inverse frequency of each class as weights.

**Deep metric learning:** In the deep metric neural network, the output of the network is an embedding vector representing the combination of the two inputs: image and WBT. In the proposed network architecture, this is the 32-dimensional output of the final layer in Figure 1. In deep metric learning, cases are mapped to a lower dimensional embedding space, where similar cases cluster together. During training, the network learns to move similar cases together and push dissimilar cases further apart, thus creating clusters of the different classes in the embedding space. When training a network with deep metric learning, the output of the network is a lower-dimensional representation of the input instead of a probability for a certain class.



**FIGURE 1** Network architecture of the combined otoscopy image and WBT network. Numbers below the bars indicate the size of the layer. The boxes to right show the final layers of the multi-task and the embedding network approaches respectively.

This allows us to use this embedding space for either classification or derivation of other metrics, such as diagnostic difficulty.

The deep metric learning network is trained using the multi-similarity loss function<sup>22</sup> ( $\alpha = 2$ ,  $\beta = 50$ , base = 1) and a multi-similarity miner ( $\epsilon = 0.1$ ) using cosine similarity to optimize the selection of training pairs. Classification is performed in the embedding space by predicting the class with the closest training data cluster center to the current test example. Difficulty estimation is performed with the supervised method employing Extra Trees<sup>23</sup> with both embeddings and ground truth labels as input.<sup>18</sup>

All networks were trained with Adam optimizer and a learning rate of 0.0001, decreased by a factor of 0.1 every 50th epoch. The models were trained until convergence on the training set, as it was found to achieve the best final model. During training, data augmentation of both input modalities was performed. For input images, transformations include horizontal flips, random rotation, color jitter, and random erasing. For the WBT measurements, we employ the transformations shown to improve training of the WBT network<sup>7</sup>: random Gaussian noise, noise increasing exponentially in intensity across the frequency axis, random erasing, and Gaussian hilly terrain, where a mixture of Gaussian functions with various intensities are added to the input to generate spatially correlated noise.

## 2.1 | Data

The dataset consists of 1014 pairs of otoscopy images and WBT measurements collected at Kamide ENT clinic, Shizuoka, Japan, from patients aged between 2 months and 12 years. The otoscopy images were captured with an endoscope, and the WBT measurements were performed using the Titan system (Interacoustics, Denmark). Each case was diagnosed with one of three different diagnoses: no effusion (NOE, 484 pairs), otitis media with effusion (OME, 375 pairs), and acute otitis media (AOM, 155 pairs) by an experienced ENT specialist (the 6th author) based on signs, symptoms, patient history, otoscopy examination, and WBT measurements. Examples of images and WBT

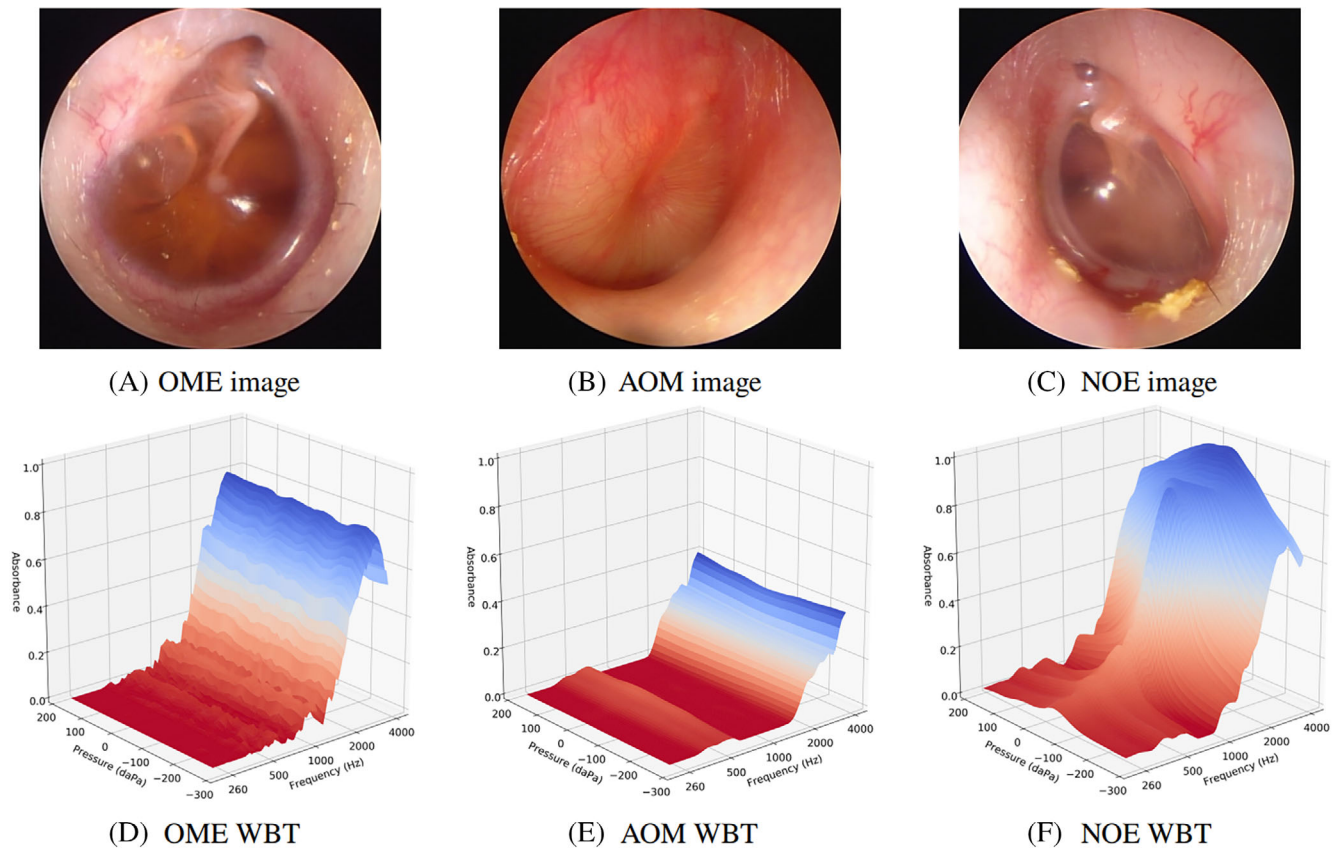
measurements from the three groups are shown in Figure 2. The data were collected and handled under the ethical approval from the Non-Profit Organization MINS Institutional Review Board (reference number 190221), with either opt-out consent or informed consent from their parent or guardian.

The otoscopy images are of size  $640 \times 480$  pixels but are cropped to a square, as the sides are black and does not contain any information, and down-sampled to  $299 \times 299$  to fit the Inception V3 architecture. Down-sampling the input images is standard procedure when employing pre-trained networks,<sup>4,12</sup> and is not expected to impact the results of the current work, as the diagnosis is based on larger features in the image that are not affected by down-sampling, such as redness of the eardrum and presence of effusion. The WBT measurements are not necessarily uniformly sampled regarding pressure, and the measured pressure values will change slightly from measurement to measurement. All measurements in the dataset were therefore resampled to a common grid using bilinear interpolation. The grid is defined from 180 daPa to  $-280$  daPa in 84 steps on a linear scale, whereas the frequency grid goes from 226 Hz to 4 kHz in 84 steps on a logarithmic scale.

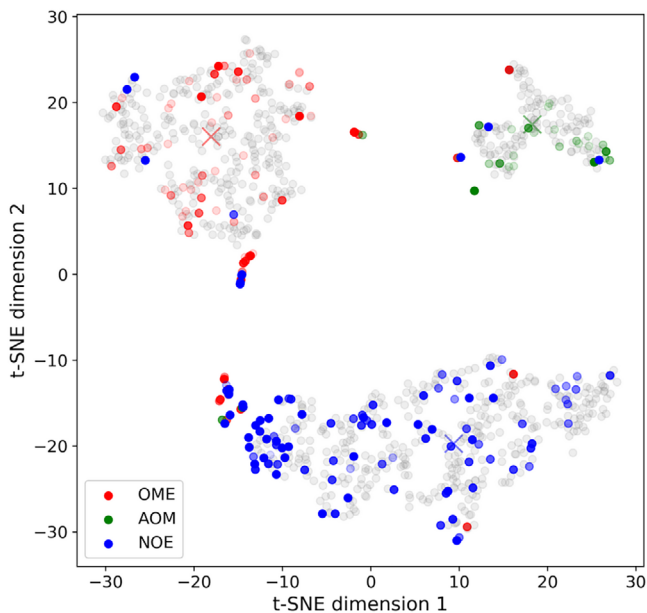
After data collection, four additional ENTs evaluated all cases in the dataset. They were shown an otoscopy image and WBT measurement pair for each patient and diagnosed with one of the three diagnoses (OME, AOM, or NOE), or “unknown.” Furthermore, they responded with their self-reported certainty on their diagnosis on the scale: very low, low, medium, moderate, or high, which was converted to a numerical scale ranging from 0 to 1. These annotations allow finding of the difficulty of each case based on the fraction of correct ENT answers (compared to the original ENT),  $\mu_{\text{correct}}$ , and the average self-evaluated certainty,  $\mu_{\text{certainty}}$ . The difficulty of each case is then given as<sup>18</sup>:

$$D = 1 - \mu_{\text{correct}} \times \mu_{\text{certainty}}$$

More details on the human inter-rater study with the four ENTs can be found in Sundgaard et al.<sup>24</sup> The ground truth diagnosis used



**FIGURE 2** Otoloscopy images and WBT measurements from patients diagnosed with otitis media with effusion (A and D), acute otitis media (B and E), and no effusion (C and F).



**FIGURE 3** Visualization of embeddings. The transparency of each point indicates the ground truth difficulty, with very transparent being the easiest. Gray points are the training cases, colored points are test cases. The center of each cluster is marked with “X.”

for training the models for this paper is based solely on the diagnosis by the original ENT, who assessed the patients in the clinic. There is, of course, a risk of human errors in the annotation process, but no other solution for achieving a ground truth diagnosis was possible in the current study. It is expected that this ENT has the best circumstances for providing the true diagnosis, compared to the four additional ENT's, who only had access to the image and WBT measurements, but not the actual patient. In the calculation of difficulty, we thus expect the original ENT to provide the best possible diagnosis, which we thus define as the true diagnosis.

Due to the limited number of cases in the dataset, all experiments were performed with five-fold cross-validation. This allows computation of performance metrics on the full dataset, instead of only a fraction of it. It was ensured that eventual multiple data pairs from one patient were only present in either a training or validation fold.

### 3 | RESULTS

Figure 3 shows the embeddings of the training and test data generated for one of the cross-validation folds for the image and WBT model in two dimensions using t-SNE dimensionality reduction.<sup>25</sup>

**TABLE 1** Performance of the proposed models.

Network	Accuracy [%]	F1-score [%]			Difficulty
		OME	AOM	NOE	Kendall's $\tau$
Image multi-task	85 ± 4	82 ± 5	78 ± 5	88 ± 3	0.39 ± 0.03
Image embed	85 ± 3	83 ± 4	77 ± 2	<b>90 ± 2</b>	0.43 ± 0.01
WBT multi-task	74 ± 2	69 ± 4	53 ± 5	87 ± 3	0.42 ± 0.03
WBT embed	68 ± 3	51 ± 10	51 ± 3	87 ± 3	0.36 ± 0.07
Image and WBT multi-task	85 ± 4	83 ± 5	77 ± 5	<b>90 ± 3</b>	0.40 ± 0.02
Image and WBT embed	<b>86 ± 2</b>	<b>84 ± 4</b>	<b>82 ± 4</b>	<b>90 ± 2</b>	<b>0.45 ± 0.02</b>

Note: Each performance metric is the average across all five cross-validation folds, with  $\pm$  indicating the standard deviation. The best performance in each column is indicated with bold.

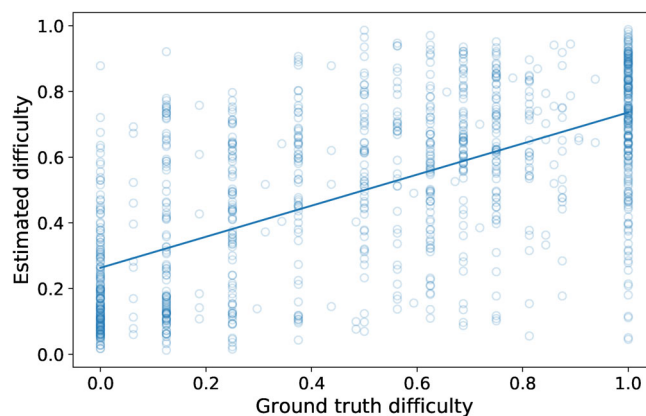
**TABLE 2** Confusion matrices for the entire dataset using the embedding models.

	Image model			WBT model			Image and WBT model		
	OME	AOM	NOE	OME	AOM	NOE	OME	AOM	NOE
OME	<b>306</b>	23	34	<b>162</b>	175	38	<b>307</b>	18	50
AOM	24	<b>117</b>	7	25	<b>121</b>	9	21	<b>124</b>	10
NOE	45	15	<b>443</b>	56	20	<b>408</b>	31	7	<b>446</b>
Total	375	155	484	243	316	455	359	149	506

Note: Rows are ground truth labels, and columns are the predictions from each model. The bold values just shows the correctly identified cases.

From these embeddings, classification and difficulty estimation was performed. It can be seen that the easy cases are in the center of the OME and AOM clusters, while the difficult cases are on the boundary of the cluster or inside other clusters. This shows that there is a relationship between the image embeddings and diagnostic difficulty, as also shown by Hannemose et al.<sup>18</sup> Table 1 shows the performance in both tasks: otitis media classification and estimation of the diagnostic difficulty for all proposed models. For classification, both accuracy and class-wise F1-scores are reported. The F1-score is the harmonic mean of the precision and recall. Kendall rank correlation coefficient,<sup>26</sup> also called Kendall's  $\tau$ , was used to evaluate the difficulty estimation. It is a non-parametric measurement of the correlation between two ranked variables. It only evaluates the ranking of cases, not the specific difficulty values. Table 2 contains confusion matrices for the three embedding models. The numbers in the table are the sum of the confusion matrices across all five test folds, such that the full dataset is represented in each table.

As seen in Table 1, the highest classification performance is achieved by the combined image and WBT model, as both accuracy and each class-wise F1-score are superior to the scores obtained with the other methods. The results also show that Kendall's  $\tau$  for difficulty estimation is increased when the model is trained on both images and WBT measurements. A Kendall's  $\tau$  of 0.45 corresponds to having ranked 73% of the cases correctly. One-way ANOVAs show a significant difference between one or more groups for accuracy ( $F = 10.97$ ,  $p = .000014$ ), AOM F1-score ( $F = 44.87$ ,  $p = 2.09e-11$ ), OME F1-score ( $F = 23.41$ ,  $p = 1.66e-8$ ), but not for NOE F1-score ( $F = 1.34$ ,  $p = .28$ ). Normal distribution the data were checked using the Shapiro-Wilks test, and homogeneity of variances across groups

**FIGURE 4** Scatter plot of ground truth difficulties and difficulties estimated with the supervised approach, together with the least-squares regression line.

was checked using Levene's test, thus fulfilling the assumptions of the ANOVA test. Tukey's post hoc tests reveal that both WBT-based models have a significantly lower performance than all image-based and combined models in both accuracy, AOM F1-score, and OME F1-score at a 0.05 significance level. There was, however, not a statistically significant difference between the embedding models and the multi-task models, or between the image-based models and the combined models.

Evaluation of Kendall's  $\tau$  show that one or more models are different from the others ( $F = 3.26$ ,  $p = .02$ ), and post-hoc analysis found that there is only a significant difference between the combined embedding model and the WBT embedding model ( $p = .01$ ).

Figure 4 shows a scatterplot of ground truth difficulties versus estimated difficulties using the combined image and WBT embedding model.

The average ground truth difficulty for the full dataset is 0.51. For the 877 correctly classified cases, the average ground truth difficulty is 0.48, while for the 137 misclassified cases, it is 0.68. Similarly, Kendall's  $\tau$  for predicting the difficulty of correctly classified cases is 0.48, corresponding to 74% correctly ranked cases, while for misclassified cases it is 0.16, corresponding to only 58% correctly ranked cases. These results show that the most difficult cases for the ENTs to diagnose are also challenging for the model to classify and that when the network fails to predict the correct class, the difficulty estimation typically also suffers. When the patients were initially diagnosed in the clinic, the ENT classified the diagnosis of AOM and OME as mild or severe, depending on the severity of the symptoms. The ground truth difficulty for mild cases is generally higher than for severe cases (0.60 vs. 0.23 for AOM, respectively, and 0.36 vs. 0.25 for OME, respectively). It is found that the classification true positive rate (TPR), or sensitivity, also differs between mild and severe cases. For AOM, the TPR is 62% and 85% for mild and severe cases, respectively, while for OME, it is 73% and 92%, respectively.

## 4 | DISCUSSION AND CONCLUSION

When inspecting the results, it is clear that the embedding networks outperform the multi-task network for both classification and difficulty estimation. Table 2 shows that, in addition to the overall superior performance, the combined embedding network manages to improve the classification of AOM, from an F1-score of 0.77 for the combined multi-task network, to 0.82. The class imbalance in the dataset makes it challenging to diagnose AOM, but these results show that deep metric learning handles this class imbalance better than a network trained with standard class-weighted cross-entropy loss functions. This confirms the findings that Sundgaard et al.<sup>1</sup> obtained for the single-modality image model. It is important to note, however, that not all increases in performance are statistically significant, as discussed in the Results section.

The confusion matrix for the WBT model in Table 2 shows that the WBT model struggles with separating AOM and OME, but it detects NOE very well. Despite this, the recall of AOM is very high, which is surprising, given the AOM and OME classification results presented in previous studies.<sup>7,27</sup> Thus, when WBT measurements and images are combined into one multi-modal model, the biggest classification improvement from the image-only model is found for the AOM class. This is an important improvement, as AOM is often difficult to diagnose, and the distinction between OME and AOM is crucial in deciding whether to prescribe antibiotics for the patient.

The results show that mild cases are more difficult to diagnose based only on the otoscopy image and WBT measurement than severe cases. This is evident for both the trained model and the four additional ENTs, as indicated by the higher ground-truth diagnostic difficulty. It shows that the mild symptoms are not well captured by

these two modalities and that more information from the patient is needed to improve the prediction. It is an important limitation of this model that symptoms must reach a certain severity or intensity before the model can detect otitis media.

The multi-modal model performs better for both classification and difficulty estimation compared to the models trained on the two modalities separately. The four ENTs in the human inter-rater study by Sundgaard et al.<sup>24</sup> achieved 64% accuracy on this dataset based on the same amount of patient information used in the multi-modal embedding model, which achieved 86%. This substantial increase in performance is very promising for a future diagnostic tool and shows the strength of deep learning models for medical image analysis.

## FUNDING INFORMATION

This study was financially supported by the William Demant Foundation.

## CONFLICT OF INTEREST STATEMENT

Søren Laugesen, Pete Bray, James Harte, and Chiemi Tanaka work for the Demant Group, which develops and manufactures otoscopy and wideband tympanometry equipment. Josefine Vilsbøll Sundgaard now works for Novo Nordisk A/S.

## ORCID

Josefine Vilsbøll Sundgaard  <https://orcid.org/0000-0003-2872-4660>

Søren Laugesen  <https://orcid.org/0000-0001-9531-9978>

## REFERENCES

1. Sundgaard JV, Harte J, Bray P, et al. Deep metric learning for otitis media classification. *Med Image Anal.* 2021;71:102034.
2. Senaras C, Aaron CM, Theodoros T, et al. Detection of eardrum abnormalities using ensemble deep learning approaches. *Medical Imaging 2018: Computer-aided diagnosis.* 2018;10575:295-300.
3. Shie CK, Chang HT, Fan FC, Chen CJ, Fang TY, Wang PC. A hybrid feature-based segmentation and classification system for the computer aided self-diagnosis of otitis media. *Annu Int Conf IEEE Eng Med Biol Soc.* 2014;4655-4658.
4. Wu Z, Lin Z, Li L, et al. Deep learning for classification of pediatric otitis media. *Laryngoscope.* 2021;131(7):E2344-E2351.
5. Grais EM, Wang X, Wang J, Zhao F, Jiang W, Cai Y. Analysing wideband absorbance immittance in normal and ears with otitis media with effusion using machine learning. *Sci Rep.* 2021;11(1):1-12.
6. Terzi S, Özgür A, Erdivanli ÖÇ, et al. Diagnostic value of the wideband acoustic absorbance test in middle-ear effusion. *J Laryngol Otol.* 2015; 129(11):1078-1084.
7. Sundgaard JV, Bray P, Laugesen S, et al. A deep learning approach for detecting otitis media from wideband tympanometry measurements. *IEEE J Biomed Heal Inform.* 2022;26:2974-2982.
8. Monroy GL, Won J, Dsouza R, et al. Automated classification platform for the identification of otitis media using optical coherence tomography. *NPJ Digit Med.* 2019;2(1):1-11.
9. Wang YM, Li Y, Cheng YS, et al. Deep learning in automated region proposal and diagnosis of chronic otitis media based on computed tomography. *Ear Hear.* 2020;0:669-677.
10. Kuruvilla A, Shaikh N, Hoberman A, Kovačević J. Automated diagnosis of otitis media: vocabulary and grammar. *Int J Biomed Imaging.* 2013; 2013.

11. Mironica I, Vertan C, Gheorghe DC. Automatic pediatric otitis detection by classification of global image features. *E-Health Bioeng Conf*. 2011;2011:1-4.
12. Khan MA, Kwon S, Choo J, et al. Automatic detection of tympanic membrane and middle ear infection from oto-endoscopic images via convolutional neural networks. *Neural Netw*. 2020;126:384-394.
13. Cha D, Pae C, Seong SB, Choi JY, Park HJ. Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database. *EBioMedicine*. 2019;45:606-614.
14. Binol H, Aaron CM, Khan NKM, et al. Decision fusion on image analysis and tympanometry to detect eardrum abnormalities. *Medical Imaging 2020: Computer-Aided Diagnosis*. Vol 11314. SPIE; 2020:375-382.
15. Harris PK, Hutchinson KM, Moravec J. The use of tympanometry and pneumatic otoscopy for predicting middle ear disease. *Am J Audiol*. 2005;14(1):3-13.
16. Habib A, Kajbafzadeh M, Hasan Z, et al. Artificial intelligence to classify ear disease from otoscopy: a systematic review and meta-analysis. *Clin Otolaryngol*. 2022;47:401-413.
17. Lee JY, Choi SH, Chung JW. Automated classification of the tympanic membrane using a convolutional neural network. *Appl Sci*. 2019;9(9):1827.
18. Hannemose MR, Sundgaard JV, Ternov NK, Paulsen RR, Christensen AN. Was that so hard? Estimating human classification difficulty. *2022 Applications of Medical Artificial Intelligence: First International Workshop, held in Conjunction with MICCAI 2022*, p. 88.
19. Camalan S, Niazi MKK, Moberly AC, et al. OtoMatch: content-based eardrum image retrieval using deep learning. *PLoS One*. 2020;15(5):1-16.
20. Chang EY. Knowledge-guided data-centric AI in healthcare: progress, shortcomings, and future directions. arXiv preprint arXiv:2212.13591 2022.
21. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception architecture for computer vision. *Proc IEEE Comp Soc Conf Comp Vis Pattern Recognit*. 2016;2818-2826.
22. Wang X, Han X, Huang W, Dong D, Scott MR. Multi-similarity loss with general pair weighting for deep metric learning. *IEEE Comp Soc Conf Comp Vision Pattern Recognit*. 2019;5017-5025.
23. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Mach Learn*. 2006;63:3-42.
24. Sundgaard JV, Värendh M, Nordström F, et al. Inter-rater reliability of the diagnosis of otitis media based on otoscopic images and wide-band tympanometry measurements. *Int J Pediatr Otorhinolaryngol*. 2022;153:111034.
25. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9(11):2579-2605.
26. Kendall MG. Rank Correlation Methods. 1948.
27. Helenius KK, Laine MK, Tähtinen PA, Lahti E, Ruohola A. Tympanometry in discrimination of otoscopic diagnoses in young ambulatory children. *Pediatr Infect Dis J*. 2012;31(10):1003-1006.

**How to cite this article:** Sundgaard JV, Hannemose MR, Laugesen S, et al. Multi-modal deep learning for joint prediction of otitis media and diagnostic difficulty. *Laryngoscope Investigative Otolaryngology*. 2023;1-7. doi:[10.1002/liv.1199](https://doi.org/10.1002/liv.1199)