



## Deep learning methods for pediatric middle ear diagnostics

Sundgaard, Josefine Vilsbøll

*Publication date:*  
2022

*Document Version*  
Publisher's PDF, also known as Version of record

[Link back to DTU Orbit](#)

*Citation (APA):*  
Sundgaard, J. V. (2022). *Deep learning methods for pediatric middle ear diagnostics*. Technical University of Denmark.

---

### General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Ph.D. Thesis  
Doctor of Philosophy

 **DTU Compute**  
Department of Applied Mathematics and Computer Science

# Deep learning methods for pediatric middle ear diagnostics

Josefine Vilsbøll Sundgaard

Kongens Lyngby 2022



**DTU Compute**

**Department of Applied Mathematics and Computer Science  
Technical University of Denmark**

Richard Petersens Plads

Building 324

2800 Kongens Lyngby, Denmark

[www.compute.dtu.dk](http://www.compute.dtu.dk)

# Summary

---

Middle ear infection, also called otitis media, is extremely common in children with around 80% having a case before school age. It is challenging even for trained specialists to diagnose otitis media, especially in the subclassifications acute otitis media and otitis media with effusion. Untreated otitis media can cause hearing loss, delays in language acquisition, poor school performance, and behavioral problems. At the same time, otitis media is the leading contributor to antibiotic prescriptions and medical costs in children. Historically, there has been a global tendency to over-prescribe antibiotics in cases where middle ear effusion is present, but it is not clear if there is an infection.

This PhD project aims to address the challenges of diagnosing otitis media by developing deep learning methods for automatic diagnosis. The work is based on a clinical dataset consisting of otoscopy images of the eardrum and wideband tympanometry measurements, which are objective measurements of the acoustic function of the middle ear.

The contributions of this thesis are manifold. Three classification models are presented; one for the analysis of otoscopy images, one for the analysis of wideband tympanometry measurements, and a final approach based on a combination of the two modalities. It is shown that it is possible to determine the diagnosis based on these two different types of patient data using a deep learning model.

Next, a generative model was developed for the generation of new artificial data from both modalities in the dataset. Additionally, it was examined how to employ a generative model to eliminate domain shifts in a medical image dataset. Domain shifts can occur when, e.g., data is collected in different hospitals or using different equipment.

Furthermore, the human inter-rater variability of the diagnosis of the cases in the dataset was investigated. Each case was additionally diagnosed by four Ear-Nose-and-Throat specialists based on the otoscopy images and wideband tympanometry measurements from the patients. This allowed for the determination of the diagnostic difficulty of each of the cases. A deep learning-based method for automatic estimation of the diagnostic difficulty was then developed.

The methods presented in this thesis could potentially be used as a diagnostic tool to assist medical professionals in the assessment of the condition of the eardrum, and thus improve the diagnosis of otitis media in the future.



# Resumé

---

Mellemørebetændelse er meget almindelig hos børn, hvor omkring 80% har et tilfælde før de starter i skole. Det er udfordrende selv for specialister at diagnosticere mellemørebetændelse, især i subklassifikationen akut mellemørebetændelse og sekretorisk mellemørebetændelse. Ubehandlet mellemørebetændelse kan forårsage høretab, forsinket udvikling af sprog, dårlig præstation i skolen og adfærdsproblemer. Samtidig er mellemørebetændelse den førende bidragsyder til brug af antibiotika og medicinske omkostninger hos børn.

Dette Ph.d.-projekt adresserer udfordringerne ved diagnosticering af mellemørebetændelse ved at udvikle metoder inden for dyb læring til automatisk diagnosticering. Arbejdet er baseret på et klinisk datasæt bestående af otoskopibilleder af trommehinden og bredbåndstympanometrimålinger, som er objektive målinger af den akustiske funktion af mellemøret.

Bidragene præsenteret i denne afhandling er mangfoldige. Der præsenteres tre klassifikationsmodeller: én til analyse af otoskopibilleder, én til analyse af bredbåndstympanometrimålinger og en model baseret på en kombination af de to modaliteter. Det bliver påvist, at det er muligt at diagnosticere mellemørebetændelse ud fra disse to forskellige typer patientdata ved hjælp af dyb læring.

Derudover bliver der præsenteret en generativ model til generering af nye kunstige data fra begge modaliteter i datasættet. Det bliver også undersøgt, hvordan man kan anvende en generativ model til at eliminere domæneskift i et medicinsk billeddatasæt. Domæneskift kan f.eks. forekomme, når data indsamles på forskellige hospitaler eller ved brug af forskelligt udstyr.

Endvidere bliver interobservatør variabiliteten for diagnosen af patienterne i datasættet undersøgt. Hver patient blev diagnosticeret af yderligere fire øre-næse-hals specialister baseret på otoskopibilleder og bredbåndstympanometrimålinger. Ud fra disse ekstra annoteringer, kan den diagnostiske sværhedsgrad for hver af patienterne bestemmes. En metode baseret på dyb læring til automatisk estimering af den diagnostiske sværhedsgrad blev derefter udviklet.

Metoderne præsenteret i denne afhandling kan potentielt bruges som et diagnostisk værktøj til at hjælpe læger med at vurdere tilstanden af trommehinden og dermed forbedre diagnosen af mellemørebetændelse i fremtiden.



# Preface

---

This thesis was prepared at Section for Visual Computing at the Department for Applied Mathematics and Computer Science at the Technical University of Denmark (DTU) in fulfillment of the requirements for acquiring the PhD degree in medical image analysis. The work presented in this thesis was funded by William Demant Fonden. The project was carried out in collaboration with Interacoustics Research Unit, and the research was conducted from May 2019 to April 2022.

The thesis deals with the development of an automatic diagnostic tool for otitis media in children. Eleven papers were written in the process of completing this PhD thesis. An overview of the papers is found on page ix. Seven papers are included in this thesis and can be found in the appendix. Paper A, B, and E are peer-reviewed journal papers, Paper C and D are technical reports, Paper F is in submission, and Paper G is still in preparation and is expected to be submitted to a peer-reviewed journal soon. The other four papers, including three peer-reviewed conference proceedings and one paper in submission, are left out of the thesis, as they deal with different topics.

The project was supervised by Rasmus Reinhold Paulsen from DTU Compute, Anders Nymark Christensen from DTU Compute, Søren Laugesen from Interacoustics Research Unit, Peter Bray from Interacoustics, and James Harte from Eriksholm Research Center. The research activities have taken place at DTU Compute at the Section for Visual Computing and at Interacoustics Research Unit.

Kongens Lyngby, April 30, 2022



Josefine Vilsbøll Sundgaard





# Acknowledgements

---

I would like to thank my great group of supervisors, Rasmus R. Paulsen, Anders Nymark Christensen, Søren Laugesen, Pete Bray, and James Harte for their support and encouragement during the project. It has been a pleasure to have such a broad spectrum of competences and knowledge in my group of supervisors.

This project would not have been possible without the help of Dr. Yosuke Kamide from Kamide ENT clinic in Shizouka, Japan, and Chiemi Tanaka from Diatec in Kanagawa, Japan. Thank you for providing me with data for this project and for sharing your expert knowledge about the clinical aspects of my PhD project. Finally, thank you very much for your hospitality during my visit to Japan.

Thank you to all my colleagues at Section for Visual Computing at DTU Compute, both past and present. It has been such a joy to share this experience with you, and I greatly appreciate all the discussions, coffee breaks, and social events we have shared in the last three years. Special thanks go to Morten Rieger Hannemose, Kristine Aavild Sørensen, and Paula López Díez for many great discussions and interesting collaborative work.

Thank you to everyone at Interacoustics Research Unit for your warm welcome during my COVID-friendly external stay in your office, and for allowing me to continue to visit throughout my PhD. I truly enjoyed spending time with you all.

I also want to thank my family for their endless support and for always believing in me. A big thank you to my friends for their encouragement during these three years. I owe a huge thank you and all the gratitude in the world to my husband Simon, who is a constant rock in my life. Thank you for your love and support during my PhD, for always motivating me, and for keeping me sane through all those months of working from home.



# List of contributions

---

## Peer-reviewed journal papers

- **Paper A: Josefine Vilsbøll Sundgaard**, James Harte, Peter Bray, Søren Laugesen, Yosuke Kamide, Chiemi Tanaka, Rasmus R. Paulsen, and Anders Nymark Christensen. Deep metric learning for otitis media classification. *Medical Image Analysis*, 2021. DOI: 10.1016/j.media.2021.102034 [68]
- **Paper B: Josefine Vilsbøll Sundgaard**, Peter Bray, Søren Laugesen, James Harte, Yosuke Kamide, Chiemi Tanaka, Anders Nymark Christensen, and Rasmus R. Paulsen. A deep learning approach for detecting otitis media in wideband tympanometry measurements. *IEEE Journal of Biomedical and Health Informatics*, 2022. DOI: 10.1109/JBHI.2022.3159263 [66]
- **Paper E: Josefine Vilsbøll Sundgaard**, Maria Värendh, Franziska Nordström, Yosuke Kamide, Chiemi Tanaka, James Harte, Rasmus R. Paulsen, Anders Nymark Christensen, Peter Bray, and Søren Laugesen. Inter-rater reliability of the diagnosis of otitis media based on otoscopic images and wideband tympanometry measurements. *International Journal of Pediatric Otorhinolaryngology*, 2022. DOI: 10.1016/j.ijporl.2021.111034 [71]

## Peer-reviewed conference papers

- **Josefine Vilsbøll Sundgaard**, Kristine A. Juhl, Klaus Fuglsang Kofoed, Rasmus R. Paulsen. Multi-planar whole heart segmentation of 3D CT images using 2D spatial propagation CNN. *Proc. of SPIE Medical Imaging 2020: Image Processing*, 2020. DOI: 10.1117/12.2548015 [69]
- Paula López Diez, **Josefine Vilsbøll Sundgaard**, François Patou, Jan Margeta, Rasmus Reinhold Paulsen. Facial and cochlear nerves characterization using deep reinforcement learning for landmark detection. *Proc. of Medical Image Computing and Computer Assisted Intervention*, 2021. DOI: 10.1007/978-3-030-87202-1\_50 [43]
- Paula López Diez, Kristine A. Juhl, **Josefine Vilsbøll Sundgaard**, Hassan Diab, Jan Margeta, François Patou, Rasmus R. Paulsen. Deep reinforcement

---

learning for detection of abnormal anatomies. *Proc. of Northern Lights Deep Learning Workshop*, 2022. DOI: 10.7557/18.6280 [16]

## Technical reports

- **Paper C: Josefine Vilsbøll Sundgaard\***, Kristine Aavild Juhl\*, and Jakob Mølkjær Slipsager. EyeLoveGAN: Exploiting domain-shifts to boost network learning with cycleGANs. *arXiv preprint arXiv:2203.05344*, 2022 [70]
- **Paper D: Josefine Vilsbøll Sundgaard**, Morten Rieger Hannemose, Søren Laugesen, Peter Bray, James Harte, Yosuke Kamide, Chiemi Tanaka, Rasmus R. Paulsen, and Anders Nymark Christensen. Multi-modal data generation with a deep metric variational autoencoder. *arXiv preprint arXiv:2202.03434*, 2022 [67]

## In submission

- **Paper F: Morten Rieger Hannemose\***, **Josefine Vilsbøll Sundgaard\***, Niels Kvorning, Rasmus R. Paulsen, and Anders Nymark Christensen. Was that so hard? Finding the likelihood of incorrect classification by humans. *arXiv preprint arXiv:2203.11824*, 2022 [25]
- Paula López Diez, Kristine A. Juhl, **Josefine Vilsbøll Sundgaard**, Hassan Diab, Jan Margeta, François Patou, Rasmus R. Paulsen. Deep reinforcement learning for detection of inner ear abnormal anatomy in computed tomography

## In preparation

- **Paper G: Josefine Vilsbøll Sundgaard**, Morten Rieger Hannemose, Søren Laugesen, Peter Bray, James Harte, Yosuke Kamide, Chiemi Tanaka, Rasmus R. Paulsen, and Anders Nymark Christensen. Multi-modal deep learning for joint prediction of otitis media and diagnostic difficulty

## Supervised student projects

- Paula López Diez, M.Sc. Deep learning for landmark detection and segmentation with geodesic path finding of facial and cochlear nerves, 2021 [42]
- Freja Rindel Peulicke, B.Sc. Image-based quality evaluation of otoscopy images, 2021 [52]

---

\*Authors contributed equally

# Contents

---

<b>Summary</b>	<b>i</b>
<b>Resumé</b>	<b>iii</b>
<b>Preface</b>	<b>v</b>
<b>Acknowledgements</b>	<b>vii</b>
<b>List of contributions</b>	<b>ix</b>
<b>Contents</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Thesis outcome . . . . .	2
<b>2 Background</b>	<b>5</b>
2.1 Clinical background . . . . .	5
2.1.1 Diagnostic tools for ear examination . . . . .	6
2.1.2 Otitis media . . . . .	9
2.2 Technical background . . . . .	12
2.2.1 Metric learning . . . . .	12
2.2.2 Generative models . . . . .	15
<b>3 Related works</b>	<b>19</b>
3.1 Clinical practise . . . . .	20
3.1.1 Diagnosis of AOM . . . . .	20
3.1.2 Diagnosis of OME . . . . .	21
3.2 Automatic otitis media detection . . . . .	22
3.2.1 Otoscopy analysis . . . . .	23
3.2.2 WBT analysis . . . . .	24
<b>4 Contributions</b>	<b>25</b>
4.1 Computer-aided diagnosis of otitis media . . . . .	25
4.1.1 Otoscopy classification (Paper A) . . . . .	25
4.1.2 WBT classification (Paper B) . . . . .	27
4.2 Generative models . . . . .	28

---

4.2.1	Domain shifts (Paper C) . . . . .	29
4.2.2	Data augmentation (Paper D) . . . . .	30
4.3	Diagnostic difficulty . . . . .	32
4.3.1	Human inter-rater study (Paper E) . . . . .	32
4.3.2	Estimating human annotation difficulty (Paper F) . . . . .	33
4.4	Combining it all (Paper G) . . . . .	35
4.5	Further challenges . . . . .	37
<b>5</b>	<b>Conclusion</b>	<b>39</b>
	<b>Bibliography</b>	<b>41</b>
<b>A</b>	<b>Deep metric learning for otitis media classification</b>	<b>51</b>
<b>B</b>	<b>A deep learning approach for detecting otitis media in wideband tympanometry measurements</b>	<b>61</b>
<b>C</b>	<b>EyeLoveGAN: Exploiting domain-shifts to boost network learning with cycleGANs</b>	<b>71</b>
<b>D</b>	<b>Multi-modal data generation with a deep metric variational autoencoder</b>	<b>79</b>
<b>E</b>	<b>Inter-rater reliability of the diagnosis of otitis media based on otoscopic images and wideband tympanometry measurements</b>	<b>87</b>
<b>F</b>	<b>Was that so hard? Estimating human classification difficulty</b>	<b>95</b>
<b>G</b>	<b>Multi-modal deep learning for joint prediction of otitis media and diagnostic difficulty</b>	<b>107</b>

# CHAPTER 1

## Introduction

---

Young children are prone to middle ear infections, called otitis media. This group of diseases has a high incidence, and it is the leading contributor to antibiotic prescriptions and medical costs in children. Therefore, they play a key role in the daily practise of Ear, Nose, and Throat (ENT) specialists and other healthcare professionals. It is, however, challenging even for specialists to correctly diagnose otitis media in the two subgroup diagnoses: acute otitis media and otitis media with effusion. A specific diagnosis is needed to provide proper treatment, and to avoid over-prescription of antibiotics for cases where it is not needed. On the other hand, complications of mistreated otitis media include hearing loss, ruptured tympanic membrane, speech and language learning delays, and meningitis.

In the clinical practise, the diagnosis of otitis media is based on otoscopy; a visual examination of the middle ear. The otoscope is a handheld device that can be inserted into the ear canal, allowing the doctor to get a visual impression of the condition of the middle ear. Modern otoscopes are digital and can capture videos and still images of the eardrum. The interpretation of the otoscopy is subjective and it can be challenging to correctly diagnose based solely on this examination. In addition to otoscopy, another commonly used diagnostic tool is tympanometry. Tympanometry is an objective measurement of the condition of the middle ear and the mobility of the tympanic membrane. It is measured by inserting an acoustic probe into the ear canal, which allows the device to alter the pressure in the middle ear and at the same time present an acoustic stimulus. The device records the reflection of the stimulus from the tympanic membrane, and this measurement can therefore evaluate the transmission of acoustic energy through the middle ear. In recent years, this diagnostic tool has been improved by employing a broadband stimulus. This is called wideband tympanometry (WBT) and provides an analysis of the acoustic absorbance over a wide frequency range, which encompasses the most functionally important audiometric frequencies.

These two clinical instruments require specialized training for both operation of the equipment and interpretation of the results. Previous studies show that pediatricians distinguished correctly between normal, otitis media with effusion, and acute otitis media 50% of the time, while the accuracy of ENTs was 75% [53], and that an ENT is less likely to diagnose acute otitis media compared with the general practitioner (44% compared to 64% of the cases) [8]. With these low accuracy rates for medical professionals without specific ENT training, an automatic diagnostic tool would be of great value in the clinical practise.



Deep learning methods are widely used in medical image analysis and specifically for computer-aided diagnosis. Deep neural networks can learn the characteristics of the diagnostic groups in a dataset, and then perform classification of unseen cases. A computer-aided diagnostic system can thus be used to quickly analyse biomedical data and help physicians make the correct medical decisions. There are, however, many other aspects of a computer-aided diagnostic system besides the diagnostic classification output. It is, for example, valuable for the medical professional to gain insight into the decision process of the neural network to gain trust in the diagnostic tool. This could be with the use of explainability methods, such as classification saliency maps. Additionally, it would be helpful to estimate the diagnostic difficulty of a certain case. This allows the user of the diagnostic tool to assess whether this is a standard case, which is easy to diagnose, or a difficult case. This could be used to refer difficult cases to an ENT specialist for further examination. A diagnostic tool like this can therefore have two different fields of application. First, it can assist the medical professional in making a diagnostic decision and, second, it can work as an educational tool for untrained practitioners with the use of explainability methods or a difficulty estimate for each case.

This PhD project aims to address the challenges of diagnosing otitis media by developing deep learning methods for the analysis of the two modalities: otoscopy images and WBT measurements. In this project, various deep learning methods will be employed, with the main focus on deep metric learning. However, the contributions of this thesis also include classification networks and generative models, both variational autoencoders and generative adversarial networks. Through the seven included contributions, the above-mentioned aspects and challenges of a computer-aided diagnostic tool are addressed from both a technical and clinical perspective.

## 1.1 Thesis outcome

The outcome of this PhD project is two-fold: the academic outcome and the industrial outcome related to a potential diagnostic tool using the developed methods. The academic outcome is presented in seven publications, of which at the time of writing three are published journal papers, two are technical reports, one is in submission, and one is in preparation. These publications all contribute to the overall objectives of this project of developing automatic tools for the diagnosis of otitis media.

This project was conducted in collaboration with the company Interacoustics, which develops diagnostic equipment for the assessment of hearing and balance. The models and methods developed in this project have been transferred to the research and development department at the company. There is a significant interest in deep learning for diagnosing ear and hearing-related conditions at Interacoustics, and this PhD project has laid the ground for further investigations within this field.

The publications included in this thesis are separated into sections based on topics, which are also used in Chapter 4. The following is a brief description of the papers.

### *Computer-aided diagnosis of otitis media*

#### **Paper A: Deep metric learning for otitis media classification**

This paper presents work on the analysis of otoscopy images using deep neural networks. The class imbalance in the dataset is overcome with the use of deep metric learning, and it is shown that it is possible to distinguish between acute otitis media and otitis media with effusion based on otoscopy images. In the paper, different loss functions for the classification task are compared, and it is demonstrated that deep metric learning achieves the best performance, while allowing interpretation of the model output.

#### **Paper B: A deep learning approach for detecting otitis media in wideband tympanometry measurements**

In this paper, a deep learning model is presented for the prediction of otitis media based on WBT measurements. The use of various types of augmentation for biomedical data to improve classification performance is demonstrated. Classification saliency maps are computed, allowing interpretation of the classification output of each measurement. It is demonstrated how the WBT measurements can be used to distinguish between otitis media and no effusion, but that more clinical information is needed for the subclassification of otitis media.

### *Generative models*

#### **Paper C: EyeLoveGAN: Exploiting domain-shifts to boost network learning with cycleGANs**

This paper shows how to handle data from different data sources when training a deep neural network - more specifically when applying a model trained on images from one camera to a test dataset of images from another camera. CycleGANs are employed for domain shifts, and artificially generated data are used to train neural networks for three different tasks: segmentation, point detection, and classification. This paper documents my participation in the Retinal Fundus Glaucoma challenge at the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2020.

#### **Paper D: Multi-modal data generation with a deep metric variational autoencoder**

In this paper, a multi-modal generative model is presented for the generation of pairs of otoscopy images and WBT measurements from the three diagnostic groups. A variational autoencoder architecture is employed and metric learning in the latent space of the model is introduced to allow for conditional data generation. The use of the model is demonstrated with generated pairs of otoscopy images and WBT measurements from each of the three diagnostic groups.

### *Diagnostic difficulty*

#### **Paper E: Inter-rater reliability of the diagnosis of otitis media based on otoscopic images and wideband tympanometry measurements**

This paper presents a study conducted with the help of four ENTs from Skåne University Hospital. In the study, the four ENTs diagnosed the cases in our clinical dataset based on otoscopy images and WBT measurements and rated their certainty of each diagnosis. The paper presents an analysis of the annotations, investigating the agreement between the ENTs, the self-reported certainty, and the diagnostic value of the WBT measurements in addition to the otoscopy images.

#### **Paper F: Was that so hard? Estimating human classification difficulty**

Based on the annotations from Paper E, we can define the diagnostic difficulty. If the four additional ENTs all agree with the original ENT on the diagnosis, the case is easy, but if the ENTs are split on the decision, the case is more difficult. In this paper, the ground truth difficulties are used to develop methods for estimating how hard it is for a medical professional to diagnose a case, both when ground-truth difficulties are available for the model and when they are not. The methods are based on embeddings generated by neural networks trained using deep metric learning. The methods are evaluated on two medical datasets: the otoscopy dataset from this PhD project, and a skin lesion dataset.

### *Combining it all*

#### **Paper G: Multi-modal deep learning for diagnosing otitis media and estimating diagnostic difficulty**

This paper presents the final classification model, which ties together most of the previous contributions from this PhD project. The classification pipeline is based on both otoscopy and WBT data, and the model is used to predict both diagnoses and estimate the diagnostic difficulty of each case. In the paper, the use of a multi-task neural network for the classification and difficulty estimation tasks is compared with an deep metric learning approach, showing the strengths of deep metric learning. Furthermore, the further challenges of diagnosing otitis media are discussed, such as detecting mild cases of otitis media.

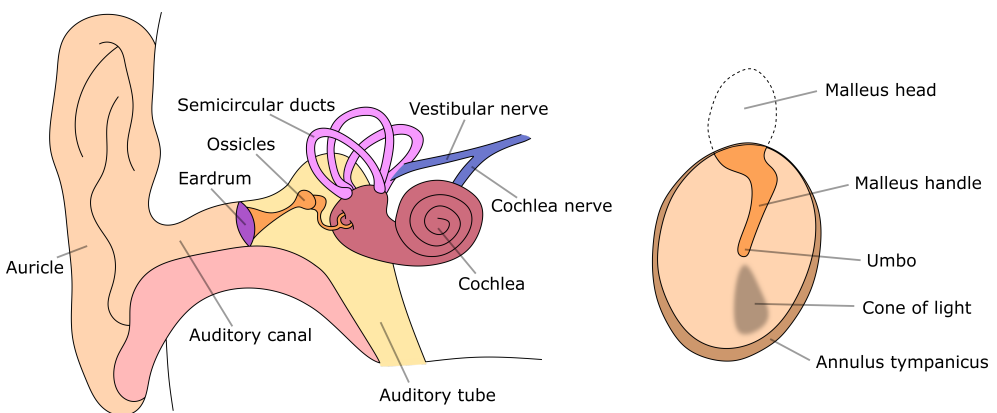
# CHAPTER 2

## Background

This chapter provides the required background knowledge needed to understand the contributions of this thesis. The first part describes the clinical background including the anatomy of the ear, the two modalities used in this project: otoscopy images and wideband tympanometry measurements, and the medical condition otitis media. Second, a technical background is provided, focussing on the field of deep metric learning and generative models.

### 2.1 Clinical background

The human ear, seen in Figure 2.1(Left), is typically described as consisting of three parts: the outer ear, the middle ear cavity, and the inner ear. The outer ear is the external part of the ear, including the auricle and the auditory canal. The middle ear consists of an air-filled cavity with the three ossicles (stapes, incus, malleus), which are the smallest bones in the human body. The middle ear connects to the upper throat through the opening of the auditory tube (also called the Eustachian tube). The inner ear is where the cochlea lies, which is a spiral shell-shaped organ responsible



**Figure 2.1:** Left: anatomy of the human ear. Right: anatomy of the eardrum.

for the transduction of sound to a neural code that propagates through the auditory pathway to the auditory cortex in the brain.

The eardrum, also called the tympanic membrane, is a thin layer of tissue that separates the outer and middle ear cavity. The eardrum receives sound vibrations through the auditory canal and transmits these to the ossicles, causing movement of the fluid in the cochlea. In the cochlea, mechanical vibration is converted to electrical activation in the auditory nerve, which propagates through the brainstem and into the hearing centres of the brain. The ossicles also provide a degree of impedance matching between the air-propagated sound in the ear canal and the fluid propagated sound transmission in the cochlea. Middle ear pathologies that result in pressure offsets relative to ambient and/or fluid that interferes with the motion of these bones will negatively impact their performance and reduce the hearing threshold.

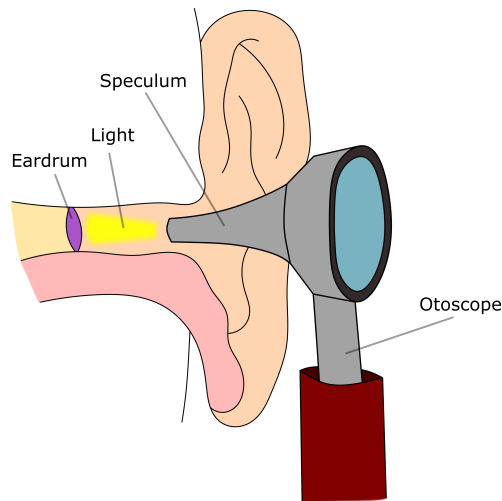
The shape of the tympanic membrane can vary from round to oval, and the edges of the membrane are attached to a bone ring (annulus tympanicus). Figure 2.1(Right) shows the anatomical structure of the eardrum. The most important features are labelled on the schematic presentation, and include the malleus bone, umbo, and cone of light, which is the reflection of light from the otoscope. These characteristics are important when examining an eardrum for middle ear diseases, as those diseases are typically diagnosed depending on the appearance and mobility of the membrane.

## 2.1.1 Diagnostic tools for ear examination

Changes in physical appearance and function of the tympanic membrane, or the ear canal, can indicate various diseases, and several diagnostic tools for examining this exist. Otoscopy is a widely used diagnostic tool for visualisation of the tympanic membrane, while wideband tympanometry assesses the acoustic function of the middle ear.

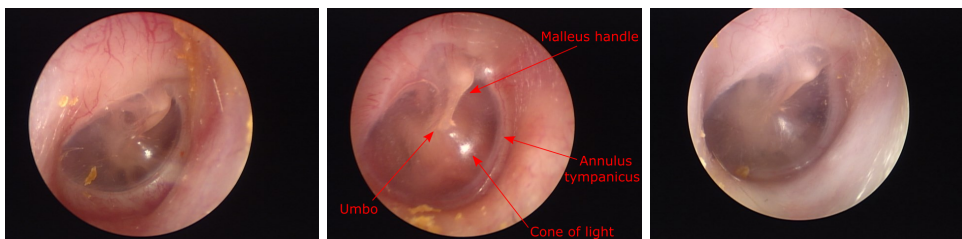
### 2.1.1.1 Otoscopy

An otoscope is a medical device used to obtain a visual impression of the tympanic membrane through the ear canal. The most commonly used otoscope is a handheld version consisting of a handle and a head, containing a light source, a magnifying lens, and a disposable ear speculum. The ear speculum is inserted into the ear canal, as seen in Figure 2.2, while the doctor straightens the ear canal by pulling the auricle or earlobe. The doctor looks through the lens to get a view of the tympanic membrane highlighted by the light source. Modern otoscopes are digital and can capture videos and still images of the eardrum. Examination can be challenged by the presence of foreign bodies, pus, cerumen (earwax), or canal skin edema, since these obscure the view of the tympanic membrane. Some otoscopes are designed with tools to remove such obstructions so that it does not compromise the examination. Furthermore, children with narrow or curved ear canals are challenging to examine, as it is not straightforward to insert the otoscope.



**Figure 2.2:** Otoscopy examination with a handheld otoscope.

Another type of commonly used otoscope is a pneumatic otoscope. This kind of otoscope has a rubber bulb and tubing, and the speculum is designed to fit tightly into the ear canal in order to create an airtight seal. The doctor can then squeeze and release the bulb to change the pressure in the ear canal to examine the mobility of the tympanic membrane in response to positive and negative pressure. This gives additional information about the characteristics of the tympanic membrane beside the visual appearance. As ENTs examine both ear, nose, and throat, it can be more convenient to use the same device for all three examinations. This can be solved by utilising an endoscope. An endoscope is a long, flexible instrument used to look into body openings. An endoscope will usually have a light at the end, to light up the examination area. Figure 2.3, 2.6 and 2.7 show examples of images of the tympanic



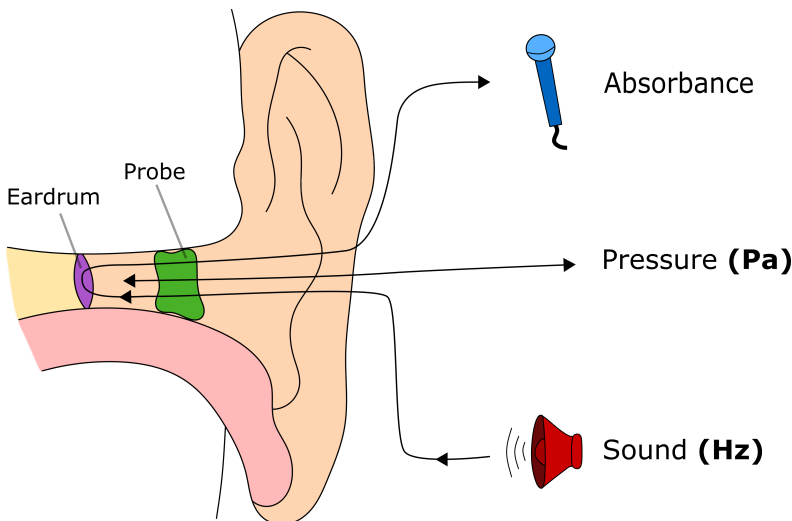
**Figure 2.3:** Otoscopy images of a normal tympanic membrane with no effusion. Photo credit: Kamide ENT clinic, Shizouka, Japan.

membrane captured with an endoscope. These images show a part of the ear canal and the tympanic membrane. The orientation of the membrane varies slightly, and both left and right ears are shown.

### 2.1.1.2 Wideband tympanometry

Tympanometry is an examination of the condition of the middle ear and the mobility of the tympanic membrane. It is an objective test of middle ear function, evaluating the energy transmission through the middle ear, but it does not as such assess the sensitivity of hearing. A standard tympanometry measurement is performed by presenting a 226 Hz tone into the ear canal, where the sound strikes the tympanic membrane, causing vibration of the middle ear. Some of this sound is reflected back and picked up by the instrument, as indicated in Figure 2.4. The test is performed by inserting an airtight tympanometer probe into the ear canal. The instrument changes the pressure in the ear, typically from +200 to -400 daPa, generates a pure tone, and measures the eardrum responses to the sound at different pressures. This produces a series of data showing how the proportion of acoustic energy absorbed by the middle ear varies with pressure, which is plotted as a tympanogram.

Tympanometry provides quantitative information, which can indicate the presence of fluid in the middle ear and assess the mobility of the tympanic-ossicular



**Figure 2.4:** Tympanometry measurement setup. The components include a microphone, a pressure regulation system, and a loudspeaker. Figure inspired by [26].

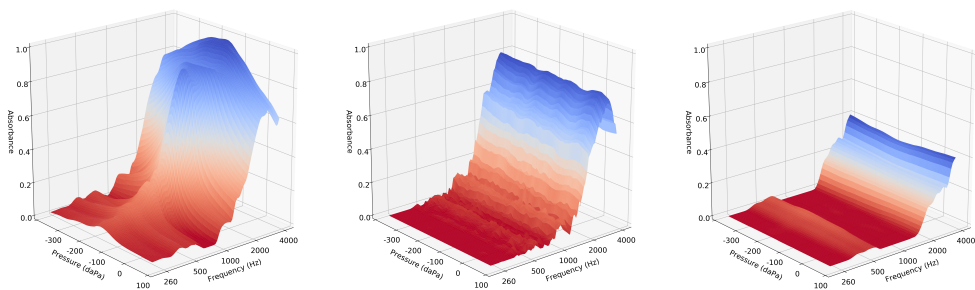
system. The method does have limitations, including a lack of specific norms for different populations (children, infants, adults), since the eardrum and the external ear are anatomically different in children compared to adults [26]. For infants of approximately 6 months of age, measurements with a high-frequency tone (1000 Hz) can be more sensitive in identifying middle-ear changes than those conducted with a 226 Hz probe tone. However, the 1000 Hz tympanometry trace is different from the traditional trace at 226 Hz. For many subjects, the 1000 Hz trace presents a double peak, and its clinical interpretation can be quite complicated.

The use of a wideband stimulus (e.g., an acoustic click or chirp) has been shown to be more efficient and precise for the evaluation of the middle ear by providing more detailed information on the mechanical and acoustic status of the middle ear than the standard 226 Hz tympanogram [51]. This measurement is called a wideband tympanogram (WBT), and evaluates the middle ear function with a transient stimulus and displays the results in the frequency range of 226 to 8000 Hz. Assessment of middle-ear function over such a broad bandwidth provides detailed information on the middle-ear status and can assist considerably in any needed diagnosis. Furthermore, it simultaneously provides the tympanometric information over a wide frequency range in an equivalent test time to standard single frequency tympanometry.

Examples of WBT measurements are shown in Figure 2.5. Higher absorbance values suggest a more efficient middle ear, in the sense that more acoustics energy potentially is transferred, as seen in Figure 2.5(Left). Lower values mean that the tympanic membrane cannot move properly, suggesting otitis media, as seen in the middle and right plot in Figure 2.5.

### 2.1.2 Otitis media

Otitis media is a group of diseases in the middle ear. Otitis media can manifest itself in different ways, leading to two of the major diagnostic groups: acute otitis media



**Figure 2.5:** Examples of wideband tympanometry measurements from each of the three diagnostic groups. Left: no effusion (NOE), middle: otitis media with effusion (OME), right: acute otitis media (AOM).



(AOM) and otitis media with effusion (OME). AOM is an acute infection in the middle ear with a rapid onset of signs and symptoms, whereas otitis media with effusion (OME) is an inflammation in the middle ear with a collection of fluid in the middle ear cavity. OME patients have no signs of infection or perforation of the tympanic membrane. Table 2.1 gives an overview of the most important characteristics of AOM, OME, and no effusion (NOE).

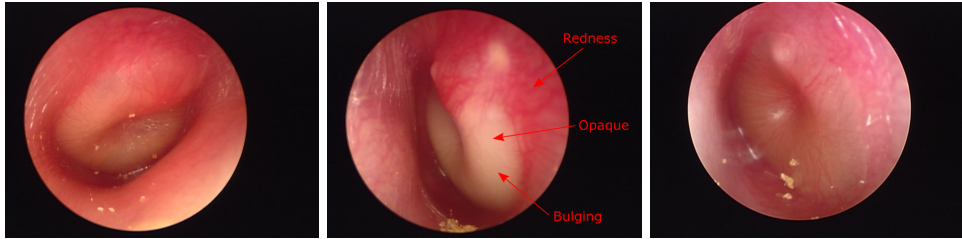
### 2.1.2.1 Acute otitis media

AOM is the second most common reason for a visit to the doctor and accounts for 10-15% of all childhood visits to the doctor [82]. AOM has a very high prevalence in young children, with around 80% of children having an episode during the first year of life [15]. It is most common in very young children, and the incidence rate decreases rapidly after 5 years of age. The aetiology of AOM includes infectious, allergic, and environmental factors. Infection has been found to be associated with genetic predisposition, anatomic abnormalities, cochlear implants, vitamin A deficiency, immunodeficiencies, bacterial pathogens, viral pathogens, allergies, lack of breastfeeding, passive exposure to smoke, attendance at the daycare, and lower socioeconomic status [15]. AOM starts as an inflammatory process, which obstructs the narrowest part of the Eustachian tube. This leads to a decrease in ventilation, causing a cascade of increase in negative pressure, accumulation of mucosal secretions, and colonisation of bacterial and viral organisms in the middle ear [15]. This can cause a visible bulging of the tympanic membrane. The condition is clinically diagnosed considering otoscopy findings and other symptoms. Symptoms appear with rapid onset and include ear pain, fever, ear discharge, vomiting, and diarrhoea. Otoscopy findings include red eardrum, clear inflammation, bulging, potentially not visible malleus, lack of airspace, or pus-filled middle ear cavity, as also seen in Table 2.1, and shown in Figure 2.6.

AOM is typically a self-limiting condition and will recover by itself within 3-7 days. Diagnosis is difficult because no standard diagnostic criterion exists, nor any specific medical or laboratory test to confirm the diagnosis. The severe ear pain makes it a challenging task to examine a small child using an otoscope, especially a pneumatic

Characteristics	AOM	OME	NOE
Visible malleus	no	yes	yes
TM shape	bulging	retracted	neutral
Colour	pale yellow, red	amber, grey, opaque	pearly white/grey
Fluid	can be present	yes	no
Cone of light	can be visible	can be visible	visible
Translucency	opaque	opaque	translucent

**Table 2.1:** Key characteristics of the tympanic membrane for the three diagnostic groups [38, 49].



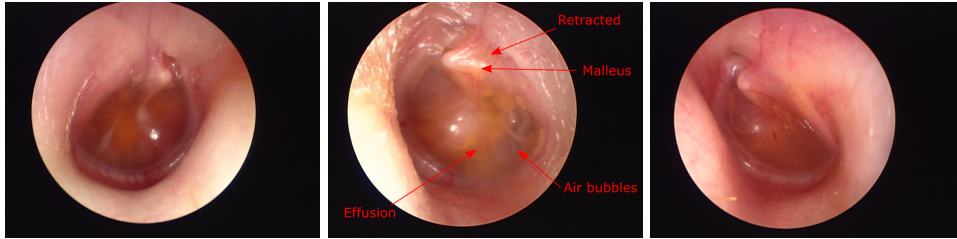
**Figure 2.6:** Otoscopy images of tympanic membrane with AOM. Photo credit: Kamide ENT clinic, Shizouka, Japan.

otoscope. Treatment is highly debated in the medical literature, as AOM is the single diagnosis responsible for most antibiotic prescriptions [82, 9]. There are controversies about prescribing antibiotics in the early stages of AOM. Watchful waiting is the best practise in most of Europe, and this approach does not show an increased incidence of complications. However, watchful waiting has not gained wide acceptance in the United States, where antibiotics are still the most common treatment [15].

#### 2.1.2.2 Otitis media with effusion

OME is the most common cause of acquired hearing loss in childhood [58]. OME accounts for 25% to 35% of all cases of otitis media, and about 80% of all children younger than 4 years have had at least one episode of OME. The prevalence declines drastically beyond 8 years of age. OME occurs more frequently in the fall and winter months. Contributing factors to OME include upper respiratory tract infection and narrow upper respiratory airways. Risk factors also include a large number of siblings, having a cold, attending daycare, passive exposure to smoke, and bottle feeding. The development of OME is related to the position of the Eustachian tube, which is more horizontal in younger children. As the child grows, the tube elongates, and the angle changes. The Eustachian tube helps to equalise pressure between the external ear and the middle ear. A clogged Eustachian tube prevents normal drainage of fluid from the middle ear, causing a build-up of fluid in the middle ear cavity. Conditions that change the development of the Eustachian tube, such as Down syndrome and cleft palate, increase the risk of developing OME [72].

Signs and symptoms of OME vary greatly and can vary in intensity over time, but often include hearing difficulties, loss of balance, or delayed speech development. Unlike AOM, patients with OME will not experience pain, fever, or malaise, making the diagnosis of OME difficult. The child is often brought to the doctor due to parents' concern regarding the child's behavior, performance in school, or language development [12]. OME is diagnosed based on an otoscopy of the tympanic membrane. Examples of OME cases are shown in Figure 2.7, and common characteristics include a retracted membrane with visible fluid build-up. The color of the membrane can



**Figure 2.7:** Otoscopy images of the tympanic membrane with OME. Photo credit: Kamide ENT clinic, Shizouka, Japan.

vary from amber to grey, and it will usually be opaque. These characteristics are also summarised in Table 2.1.

OME is typically a self-limiting condition, and 50% of the cases will be resolved within 3 months, 95% within 1 year. In these prolonged cases, complications such as tympanic membrane perforation, tympanosclerosis, ear discharge, and hearing deficits can occur. Treatment for cases that persists for longer than 3 months will usually be an insertion of a tympanostomy tube to drain the fluid, which will solve the issue until the Eustachian tube is developed enough to allow for natural drainage [12].

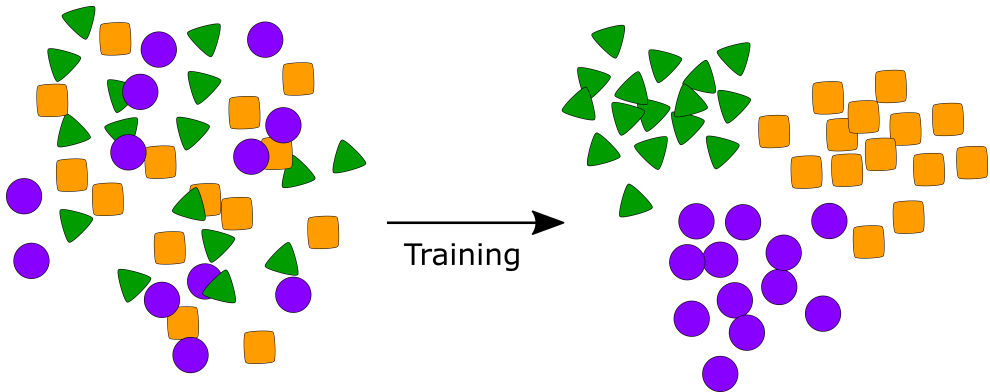
## 2.2 Technical background

This PhD project works with various models, but the common theme of the proposed methods is deep learning. A general introduction to deep learning will, however, not be given in this thesis. The focus of most of the work has been on the field of deep metric learning, and this section will therefore give an introduction to this topic. Additionally, a brief introduction to generative models is provided, as these are employed in Paper C and D.

### 2.2.1 Metric learning

Metric learning is based on a distance metric evaluating the similarity of the data examples, which have been mapped to an embedding, or feature, space. The goal is to learn an embedding space in which similar data examples are moved together and the dissimilar examples are moved further apart. In deep metric learning, the embeddings are acquired using a neural network as a feature extractor. The network architecture is thus different from a standard classification network, as the output is a feature vector of, e.g., dimensionality 32, representing the input sample. The training process of a deep metric learning embedding space is shown in Figure 2.8.

Deep metric learning have been used for various tasks, such as face recognition [59], person re-identification [86], and retrieval tasks [85]. These tasks usually have a



**Figure 2.8:** Schematic representation of the training process of the embedding space in a deep metric neural network. The colours indicate three different classes.

high number of classes and few samples, or images, in each class. Deep metric learning increases the size of the data for training by employing pairs or triplets [32], which makes these methods well suited for these tasks. This also makes deep metric learning well-suited for small datasets with high class imbalance, which is very common in medical image analysis.

Using deep metric learning, a higher level of representation of the data is provided, compared to a standard classification network. A feature vector describing the model input is acquired which can be interpreted or used for further analysis, such as classification. Training the network based on similarities, instead of with a specific classification goal, grants a better ability to represent the data, and results in a more meaningful and interpretable model.

It is, for example, possible to plot the feature vectors of the full dataset and evaluate the placement of each sample in the feature space in relation to each other. In, e.g., a medical image classification task, the disease patterns should be similar in the different diagnostic groups. By visualising the feature vectors of the dataset, the distribution can be investigated. There may be a diagnostic group that is slightly divided into two subgroups in this feature space, because the condition can appear with different signs or other underlying features of the dataset that can be inferred from the feature space.

Deep metric neural networks are trained with loss functions designed to manipulate the embedding space to move examples of the same class closer together, and to push examples of different classes away from each other, as seen in Figure 2.8. There are various loss functions designed to learn this transformation. The most simple loss function is contrastive loss [24]. This loss function is computed based on the Euclidean distance between two embedding vectors. If the two input examples are

from the same class (positive pairs), the embedding vectors will be moved closer to each other, and further apart if they are from two different classes (negative pairs). The loss function is a measure of the distance between the two embedding vectors, which ideally should be  $y_i = 0$  for positive pairs and  $y_i = 1$  for negative pairs. The loss function can be written as [24]:

$$L_c = \sum_{i=1}^N (1 - y_i) (\|f_{1,i} - f_{2,i}\|_2)^2 + y_i \{\max(0, m - \|f_{1,i} - f_{2,i}\|_2)\}^2, \quad (2.1)$$

where  $f_{1,i}, f_{2,i}$  represents the embedding vectors generated by the network for each input,  $N$  is the number of pairs, and  $m$  is the margin, usually set to 1.0. The margin is the desired distance between the clusters, and it is used both for the loss function and for sampling. It is shown in Figure 2.9.

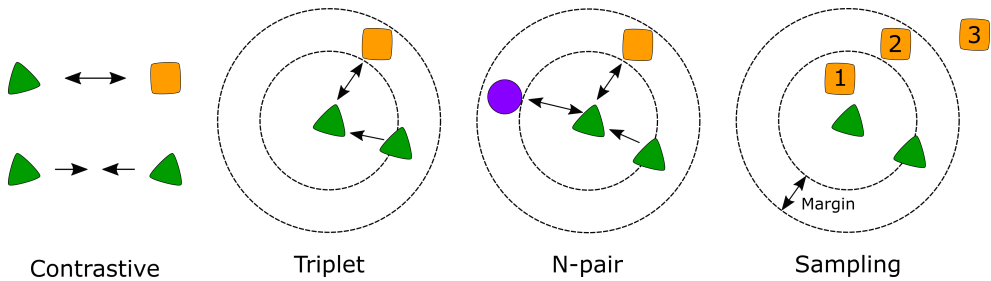
A further development of contrastive loss is the triplet loss [59], which optimises both positive and negative samples at the same time. To calculate this loss, three embedding samples, called a triplet, are used. A triplet consists of an anchor  $f^a$ , from which similarities are computed, a positive sample  $f^p$ , and a negative example  $f^n$ . The triplet loss function is computed based on distances between the samples, given as [59]:

$$L_{\text{triplet}} = \sum_{i=1}^N \max(0, m + \|f_{a,i} - f_{p,i}\|_2^2 - \|f_{a,i} - f_{n,i}\|_2^2). \quad (2.2)$$

Multi-class N-pair loss [64] is a generalisation of triplet loss, which takes into account  $j = N - 1$  negative samples in each iteration, instead of only one.  $N$  is the number of available pairs in a mini-batch. For  $N = 2$ , the loss highly resembles triplet loss. The loss function reduces the computational cost by optimising over the distance, computed as the cosine similarity, against all classes in one iteration, and it is given as [64]:

$$L_{\text{m-c}} = \sum_{i=1}^N \log\left(1 + \sum_{j \neq i}^{N-1} \exp(f_{a,i} f_{n,j} - f_{a,i} f_{p,i})\right). \quad (2.3)$$

A schematic representation of these three loss functions is shown in Figure 2.9. When training with deep metric loss functions, the sample selection of the pairs is crucial [59]. If the neural network is trained with all possible pairs in the dataset for every epoch, the training time would be very long. Efficient sampling, also called mining, is, therefore, a key part of the training process to ensure that all pairs are informative to the network. Figure 2.9 on the right shows three different situations for the sampling of negative samples. Negative sample 1 is a hard negative because the distance from the anchor to the negative sample is smaller than to the positive sample. Semi-hard sampling was proposed by Schroff et al. [59], and selects the negative samples within the margin defined as the distance between the anchor and the positive sample plus a margin  $m$ , shown with the negative sample 2 in the figure. Sample 3 in



**Figure 2.9:** Schematic representation of deep metric loss functions and sampling. The points represent three different classes (green, orange, and purple), and the arrows indicate the desired transformation for each of the loss functions. For sampling, the green point in the middle is the anchor, the other green point is the positive example, and the orange points are negative samples. Orange point 1 is a hard negative, orange point 2 is a semi-hard negative, and orange point 3 is an easy negative. Modified from [68] and [32].

the figure is an easy sample, further away from the anchor than the positive sample plus the margin  $m$ . This would therefore not contribute any information if used for training. The semi-hard sampling strategy is widely used for triplet loss training.

As stated by Kaya et al. [32], the benefits of N-pair loss are lost when the number of classes decreases. Since we have a limited number of classes, three, in our dataset, we might not benefit greatly from N-pair loss. Another loss function is the multi-similarity loss [79], which considers both self-similarity and relative similarities, allowing for more efficient and accurate sampling of pairs during training. With the proposal of multi-class loss, the authors also introduced a new mining strategy claimed superior to previous strategies, and a general pair weighting framework for analysing pair-wise loss function. Like the N-pair loss function, the multi-class loss uses the cosine similarity as the distance function. I recommend reading the paper by Wang et al. [79] for a detailed description of the multi-similarity loss.

There are many other deep metric loss functions, but the ones mentioned here, are employed in this PhD project. Kaya et al. [32] provide a good overview of deep metric learning in general and some of the many other deep metric loss functions.

## 2.2.2 Generative models

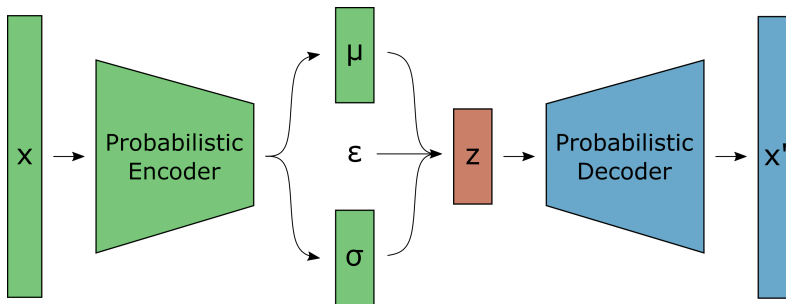
Deep generative models are deep neural networks trained to synthesise artificial data from the distribution of the training data. There are many different types of generative models, but the most widely used are Variational Autoencoders (VAEs) [35] and Generative Adversarial Networks (GANs) [21].

### 2.2.2.1 Variational autoencoder

This is a brief introduction to VAEs from a neural network perspective. This section will not go into the mathematical background or motivation for these models. A VAE is a neural network designed to reduce the high-dimensional input data to a latent representation, and then reconstruct the input data from this latent representation. An illustration of a standard VAE is shown in Figure 2.10. In the VAE, the input data is mapped to a certain distribution through the encoder. The training of the network is regularised to ensure that this latent space follows a standard normal distribution, enabling easy generation of new samples. Instead of a standard autoencoder, where the encoder output is the latent representation of the input, the output of the variational encoder is a distribution in the latent space. During training, a point from this distribution is then sampled at each iteration, which is reconstructed through the decoder.

The loss function of a standard VAE consists of two parts: a reconstruction term and a regularisation term. The reconstruction term penalises the reconstruction by comparing the input data with the reconstructed data, thus improving the generative performance of the model. The reconstruction loss could be e.g. mean squared error, structural similarity [80] or based on perceptual similarity [14]. The regularisation term penalises the predicted distributions and enforces a standard normal distribution. It is expressed as the Kullback-Leibler divergence [37] between the latent distribution and a standard Gaussian.

Once the model is trained, new latent vectors can be sampled from the standard Gaussian distribution and run through the decoder for the generation of new data. VAEs are theoretically appealing, as it is a well-defined statistical model. A major challenge of VAEs is, however, that the generated images tend to be blurry. This is partially due to the bottleneck of a small latent dimension in the model, and the use of loss functions based on mean squared error, causing imperfect reconstruction. Several works have focused on solving this problem, such as the VQ-VAE-2 model [55], which can generate high-resolution images. Other applications of VAEs include semi-supervised classification [84] and denoising [28].



**Figure 2.10:** Basic structure of a VAE model.

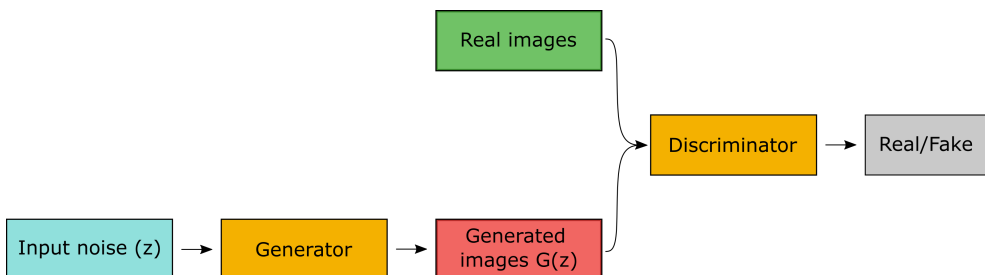
### 2.2.2.2 Generative adversarial network

A GAN model consists of two neural networks: a generator and a discriminator, as seen in Figure 2.11. The networks are trained simultaneously, and the goal is to learn to transform random noise into realistic artificial data examples from the training data distribution. The generator learns to capture the data distribution and transforms the input noise into realistic data examples, and the discriminator estimates the probability that the input data is a generated image or from the training dataset. Training these two networks is a delicate balance, as the discriminator needs to be able to distinguish between real and fake images, and the generator needs to generate realistic data to fool the discriminator [20].

After training, the generator model can be used without the discriminator to generate new examples from the training distributions. GANs are very capable generative models and produce highly realistic fake images. The GAN training process is, however, one of the main drawbacks of these models. GANs suffer from several issues such as mode collapse, imbalance between the generator and discriminator causing overfitting, vanishing gradients, and they are highly sensitive to hyperparameter selections [20].

Just like the VAE, many variations of GANs have been developed: the conditional GAN, which generates data with some desired feature, such as images from certain classes [47]; the text-2-image network, which generates images from text descriptions [56]; the SRGAN, generating photo-realistic super-resolution (SR) images [39]; and the cycleGAN, which is an unpaired image-to-image translation model [87].

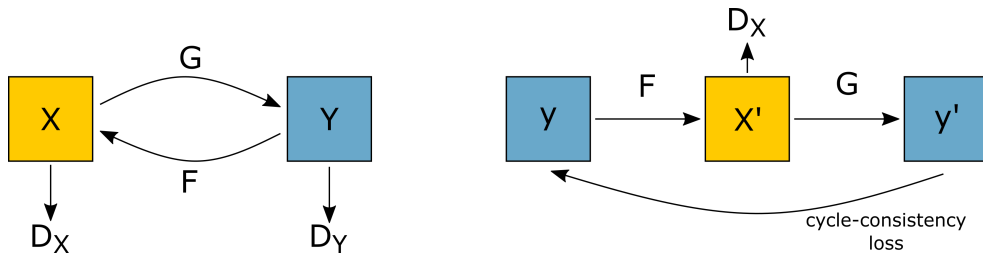
The cycleGAN learns a mapping between two image distributions, for example, domain  $X$  and  $Y$ , without the use of paired examples. The network follows the basic structure of a GAN model, but with two generators,  $F$  and  $G$ , and two discriminators,  $D_X$  and  $D_Y$ , as seen in Figure 2.12. The networks are trained on two datasets from domain  $X$  and domain  $Y$ , and as it is an unpaired model, image correspondence between the two domains is not needed. This makes this model much more accessible and easier to train. During training, the two generators learn to transfer images between the two domains, while the discriminators evaluate the quality of the mapped



**Figure 2.11:** Basic structure of a GAN model.



images. Figure 2.12 also show the cycle-consistency loss that enforces the intuition that an image mapped from, e.g., domain  $Y$  to  $X$ , and back to  $Y$ , should look the same as the original image. Once the networks are trained, the two generators can thus be used to transfer images from one domain to another. More details on the cycleGAN model are found in the paper by Zhu et al. [87]. As shown in the paper, the cycleGAN model can be used for style transfer, such as turning Monet paintings into naturally realistic photos, object transfiguration, where, e.g., images of zebras are turned into images of horses, or photo enhancement. In medical image analysis, cycleGANs have been used for, e.g., unpaired translation between T1 and T2 weighted MRI scans [5], and for generating synthetic CT scans from PET scans [17]. Another application is to rectify domain shifts when the medical data used for a deep learning model are acquired at different hospitals or with different equipment. This was addressed by, e.g., Wollmann et al. [81], and will be the focus of Paper C of this thesis.



**Figure 2.12:** Structure of the cycleGAN model. Left: the two generators  $F$  and  $G$  between domain  $X$  and  $Y$ , and the two discriminators. Right: mapping between the domains. Figure inspired by Zhu et al. [87].

# CHAPTER 3

## Related works

---

This chapter presents the current state of the art of otitis media diagnosis. Each contribution included in this thesis contains presentations of the related works for the specific research area of the individual contributions. Thus, this chapter is an introduction to the overall theme of this PhD project focussing on the clinical practise for the diagnosis of otitis media and other works on automated diagnosis based on otoscopy images or WBT measurements.

Treatment and diagnosis of otitis media are highly debated in the medical literature, due to the large increase in recent years of drug-resistant infections in acute otitis media. Historically, there has been a global tendency to over-prescribe antibiotics in cases where middle ear effusion is present, but it is not clear if there is an infection. The usual high prevalence of not fully adhering to a full course of antibiotics and the very high general prevalence of OME and AOM in young children has led to a rise in drug-resistant bacteria and the desire to reduce general antibiotic use. The differentiation of AOM and OME has therefore become more critical in an era of increased resistance to antibiotics among bacterial pathogens that cause AOM.

Several studies have investigated the diagnostic process and accuracy of the diagnosis of otitis media by different medical professionals. Jensen et al. [30] found that general practitioners (GP) were certain of their diagnosis of AOM in 67% of new cases in children younger than 2 years of age. For children over 2 years of age, the diagnostic certainty increased to 75%. Pichichero et al. [53] discovered that paediatricians correctly distinguished between NOE, OME, and AOM 50% of the time, while the accuracy of the ENTs was 75%. Blomgren et al. [8] found that four medical professionals agreed on the diagnosis in 64% of the AOM diagnosis and that the ENT was less likely to diagnose AOM, compared to the GP (44% compared to 64% of the cases).

These studies all show how challenging the diagnosis of otitis media is, no matter the clinical background of the doctors. The studies report specific challenges in the diagnostic process, such as familiarisation with the pneumatic otoscope, which increases diagnostic performance [53], or access to tympanometry equipment [8]. Thus, diagnostic accuracy depends on specific training, experience, and available diagnostic equipment. But even with the correct diagnostic tools, diagnosis is still highly subjective and diagnostic performance differs greatly between medical professions.

## 3.1 Clinical practise

In order to streamline the diagnosis and treatment of diseases, clinical practise guidelines are published in many countries. These guidelines include recommendations intended to optimise patient care for a specific disease. They are based on a systematic review of evidence and assess the benefits and harms of different care options. For otitis media, guidelines are often focused on OME or AOM. These guidelines show the worldwide differences in clinical practise, and many reviews and comparisons of these guidelines have been conducted.

### 3.1.1 Diagnosis of AOM

Tamir et al. [73] compared national guidelines, consensus papers, and position documents from 62 countries. All guidelines discussed diagnostic criteria for AOM, the pillars of diagnosis being the characteristics of the tympanic membrane (bulging, red, opaque, or position) and the coexistence of acute onset of symptoms such as fever and ear pain. Developed countries generally recommend using pneumatic otoscopy for further examination, while in developing countries, this is rarely mentioned. Similarly, many developed countries recommend the use of tympanometry as a supplementary tool for the diagnosis of AOM when the tympanic membrane is not visible, while this tool is barely mentioned in developing countries. In developing countries and remote regions of developed countries, cases of AOM are often diagnosed and treated by GPs or paediatricians rather than ENTs, and the diagnosis is based on symptoms rather than otoscopic evaluation. Differences in clinical practise therefore often originate from variations in accessibility to local healthcare services.

The AOM clinical guideline published by the American Academy of Pediatrics (AAP) and the American Academy of Family Physicians [41] is widely used and is referenced in many other guidelines. Baumer et al. [6] compared the AAP guideline and the Scottish guideline [60]. Both guidelines are based on current best practices but set in two different healthcare systems. There is strong agreement between the two guidelines, indicating that most developed countries in the world follow a similar clinical practise, as also shown by Tamir et al. [73]. The Scottish guideline does, however, make a stronger statement against antibiotics in children under 2 years of age, compared to the American guideline.

Despite the development of national guidelines to enforce a streamlined clinical practise for the diagnosis of AOM, many cases of overdiagnosis of AOM, and thus overprescription of antibiotics, still occur. This could be due to a lack of adherence to the guidelines. Céline et al. [9] raise concerns about the adherence to treatment guidelines and the need for an effective strategy of guideline implementation. The study focused on Sweden, where watchful waiting is recommended for most children over 2 years of age with AOM, while antibiotics are recommended for children younger than 2 years. The way in which the AOM guideline is followed was evaluated in a busy paediatric emergency department. Adherence to the guidelines was greater in patients under 2 years of age, compared to those older than 2 years. This shows that when

the recommendation is to give antibiotics, the guidelines were followed. An intense information campaign was carried out, due to the assumption that this deviation from the recommendations was caused by a lack of knowledge. The campaign did not, however, produce any significant improvement. The fear of complications could outweigh the guilt over bad adherence to guidelines, especially when the treatment guideline is to do nothing, as when the child is older than 2 years. This study raises questions about how these guidelines should be implemented since an intense information campaign is not enough to enforce these recommendations.

Similarly, Marchisio et al. [44] conducted a study concerning Italian paediatricians and ENTs and their attitude and adherence to the Italian AOM guideline. Only 9% of the 2012 included physicians had received any specific AOM medical education during medical school, but in post-residency, the percentage increased to 53%. 40% reported a positive attitude towards the AOM guideline, but only 21% reported an appropriate diagnostic method for AOM (pneumatic otoscopy). This survey shows how these guidelines have not been extensively integrated into the practices of Italian pediatricians or ENTs. Very similar results are seen in equal studies performed in the US [77, 78] and Israel [57], where most physicians are familiar with the AOM guideline, but fail to follow the diagnosis and treatment recommendations. Chandler et al. [11] examined how well clinical trials from 1994 to 2005 met the AOM guidelines from AAP [41]. They found that only 17% of the clinical trials met the three criteria of AAP for a certain diagnosis; a history of acute onset of symptoms, the presence of effusion in the middle ear, and signs of inflammation in the middle ear indicated by redness or ear pain. However, 80% of the clinical trials required at least one AAP criteria for diagnosis, and the most commonly used symptom was middle ear effusion.

### 3.1.2 Diagnosis of OME

Most clinical practise guidelines are focusing on AOM, since AOM is treated with antibiotics. In order to prevent over-prescription of antibiotics, a correct diagnosis of AOM is needed. OME is usually not treated with antibiotics but with a tympanostomy tube to drain the effusion. But even without the social problem of drug-resistant bacteria, it is still important to properly diagnose and treat OME to avoid complications and challenges related to persistent episodes of OME.

Stewart et al. [65] conducted a study on practise patterns of physicians when diagnosing OME. The overall performance in answering questions about the guideline was very poor, with only around 50% correct answers on average. This study was performed five years after the publication of a national guideline and shows the lack of integration of these guidelines in the clinical practice.

The AAP OME guideline [3] addresses the need for adequate documentation when children with OME visit the doctor. Kalu et al. [31] examined clinical compliance with these recommendations related to documentation of presence, laterality, resolution, persistence, and surveillance for hearing loss or speech decay. It was found that the initial diagnosis of OME had a high documentation rate, but continuity of care

and follow-up visits were poorly documented. This study also found that contrary to recommendations, 43% chose to prescribe antibiotics for children without risk of speech or language decays, where watchful waiting is recommended.

Cullas Ilarslan et al. [13] investigated how Turkish ENTs and paediatricians adhere to the clinical practise guideline. Turkey does not have a national guideline for OME, but most medical professionals follow the AAP guideline [3]. A high level of self-confidence was recognised when surveying the medical professionals, even though low adherence to the clinical practise guidelines was found. Only half of the physicians could correctly define pneumatic otoscopy findings for OME, while 17% employed a pneumatic otoscope in their practice. They also found that older and more experienced physicians are less likely to prescribe antibiotics. This can be related to insecurity in younger physicians, but it also shows how important proper training is, in order for younger physicians to know how to handle OME cases.

With the publication of these AOM and OME guidelines, the attempt is to provide evidence-based recommendations to assist medical professionals in making diagnostic and treatment decisions. However, as seen with these reviews of implementation and integration of the guidelines, this is a challenging task. One study found that 21% of paediatricians never used guidelines, 44% sometimes used them, and only 35% routinely consulted guidelines [19]. With statistics like this, it is very difficult to change the practice of pediatricians or ENTs, in order to limit the use of antibiotics and get the children the best treatment.

As mentioned, many countries publish national guidelines and since data for this project are collected in Japan, the Japanese AOM and OME guidelines [36, 29] will be adhered to during the diagnostic process of the patients included in this project.

## 3.2 Automatic otitis media detection

As discussed above, the diagnosis of otitis media is still highly subjective, in spite of the publication of clinical practise guidelines. Key problems in the diagnostic process have been shown to be lack of specific training, limited availability of necessary diagnostic tools [30, 53], lack of experience in handling cases of otitis media, and lack of adherence to clinical guidelines, sometimes due to the attitude and behaviour of physicians about guidelines [9, 19]. A diagnosis support system would thus be of great value in order to improve the diagnosis and treatment of otitis media and ensure adherence to the diagnostic guidelines.

With machine learning growing rapidly in many fields, including medical image analysis, it has also taken an impact on the field of automatic otitis media detection. This task is in many cases performed by classifying an otoscopy image of the tympanic membrane into a diagnostic group, but there has also been some work in classification based on tympanometry measurements. Some studies have focussed on AOM and OME, as we are focusing on in this PhD project, but other diagnostic groups such as chronic suppurative otitis media, tympanic perforation, or attic retraction are also included in some studies.

### 3.2.1 Otoscopy analysis

Some of the first published approaches for the detection of otitis media are based on hierarchical rule-based decision trees to determine if an otoscopy image is from an OME, AOM, or NOE patient [38, 49]. The decision trees are of varying complexity but are based on hand-made, manually selected features. The features include color, bulging, translucency, light, bubbles, the presence of malleus, and concavity. The decision trees were manually constructed, which means that the decision is purely based on the decision process of the person designing the decision tree. This method mimics the decision process of an ENT and achieves performance ranging from 80 to 85%. Other works are also based on manually selected features [63, 48], but they use other classification methods such as neural networks or Adaboost, which outperform decision trees with a performance of around 86-88%.

Mironica et al. [46] evaluated different machine learning methods for classification of otitis media, including k-nearest neighbour, decision tree, linear discriminant analysis, Bayes, multi-layer neural network, and support vector machines. The methods were only designed to distinguish between normal and abnormal tympanic membranes. The results showed that neural networks and support vector machine have the highest performance on this classification task. Similarly, Myburgh et al. [48] found that a neural network outperformed their previous method using a decision tree [49]. This is similar to the overall trend in the field of machine learning, where deep neural networks outperform previous methods in many fields. This has also influenced the methods used for otitis media detection and analysis of otoscopy images. The first publications employing neural networks employ simple networks, such as fully connected neural networks, while the newer publications employ convolutional neural networks. Senaras et al. have published two papers on the classification of normal and abnormal tympanic membranes, one using deep neural networks [62] and the other employing a fuzzy stacked generalisation algorithm [61]. Tran et al. [75] employed a multi-task joint sparse representation-based algorithm to classify based on a tympanic membrane segmentation and manually selected, but automatically extracted, features. During the extent of this PhD project, many studies on deep learning-based detection of ear diseases have been published. This includes Cha et al. [10], who collected a large otoscopy database of over 10,000 images, covering six different middle ear diseases - although not AOM. They employ pre-trained convolutional neural networks fine-tuned for the classification task and show impressive results on these diagnostic groups. Khan et al. [34] showed a similar approach for the classification of chronic suppurative otitis media, OME, and normal tympanic membrane, and Alhudhaif [2] classified AOM, chronic suppurative otitis media, earwax, and normal tympanic membrane. Only Wu et al. [83] have focussed on the same subclassification of otitis media, as we have (AOM, OME, and normal) and employed deep neural networks for the classification task. They demonstrated their approach on a large dataset of over 10,000 otoscopy images, achieving an accuracy of 97%. They show an impressive classification performance on all three classes. It is important to notice how the dataset for the different studies are created. For the work presented by Wu

et al. [83], several inclusion criteria were enforced such as consensus diagnosis by two otologists, clear image details, no out-of-focus images, and images had to cover the full tympanic membrane. An extensive review of artificial intelligence-based methods for the classification of ear diseases based on otoscopy images can be found in the paper by Habib et al. [23].

### 3.2.2 WBT analysis

Automatic classification of tympanometry measurements is not as extensively investigated as otoscopy images. Terzi et al. [74] employed a receiver operating characteristic (ROC) test to distinguish between NOE and OME cases based on WBT measurements from pediatric patients. Ellison et al. [18] analysed the measurements only at ambient pressure using a likelihood ratio classifier and found that absorbance is sensitive to stiffness of the middle ear and middle ear effusion. Aithal et al. [1] showed that wideband absorbance at ambient pressure and tympanometry peak pressure can be used successfully to detect OME, although not significantly better than a 226 Hz tympanogram. These studies lay the ground for the development of an automatic classification model, as there is clearly relevant diagnostic information available in the WBT measurements.

Recent studies have thus shown an interest in the automatic classification of WBT measurements. Merchant et al. [45] created a multivariate prediction model based on the three first principal components using logistic regression, showing good results for otitis media and NOE classification. Binol et al. [7] automatically detected NOE or OME based on a combination of otoscopy imaging and tympanograms. Their analysis used a random forest classifier on hand-selected features (peak admittance, peak pressure, tympanogram width, and ear canal volume) of a standard 226 Hz tympanogram, which was combined using majority voting with the output of a convolutional neural network predicting diagnosis based on the patient's otoscopy image. Grais et al. [22] showed that convolutional neural networks can be used to analyse and classify WBT measurements into OME and NOE classes. No studies have attempted to classify AOM, OME, and NOE based on WBT measurements, except ours. Helenius et al. [27] investigated the discrimination of diagnosis into the subgroups AOM and OME based on standard 226 Hz tympanometry, and found that this measurement can be used to distinguish between NOE and otitis media cases, but not to diagnose specific types of otitis media.

# CHAPTER 4

## Contributions

---

This chapter will discuss the contributions included in this thesis, relate the papers to each other, and put the work into a broader perspective. Details on the work in each contribution will therefore not be presented but can be found in the publications in appendices A to G.

The work presented in this thesis is based on a clinical dataset collected at Kamide ENT clinic in Shizuoka, Japan. The dataset consists of otoscopy images captured using an endoscope and WBT measurements performed using the Titan system (Interacoustics, Denmark) from patients aged between 2 months and 12 years. Each case was diagnosed by Dr. Kamide when the patient was examined in the clinic based on symptoms, patient history, otoscopy, and WBT measurements. The diagnostic groups included in the dataset are acute otitis media (AOM), otitis media with effusion (OME), and no effusion (NOE). In addition to the diagnoses, Dr. Kamide also noted whether AOM or OME was mild or severe, based on the appearance of the symptoms. The dataset has evolved during the PhD project, as data has continuously been collected. The number of cases has therefore increased over the duration of the project.

### 4.1 Computer-aided diagnosis of otitis media

Otoscopy and WBT measurements are the most commonly used diagnostic tools for the diagnosis of otitis media in the clinic. Both tools require specialised training for operation and interpretation of the results, and the diagnosis of otitis media can be challenging, even for trained specialists. The two papers in this section, Paper A and Paper B, thus focus on computer-aided diagnosis of otitis media based on otoscopy images and WBT measurements, respectively.

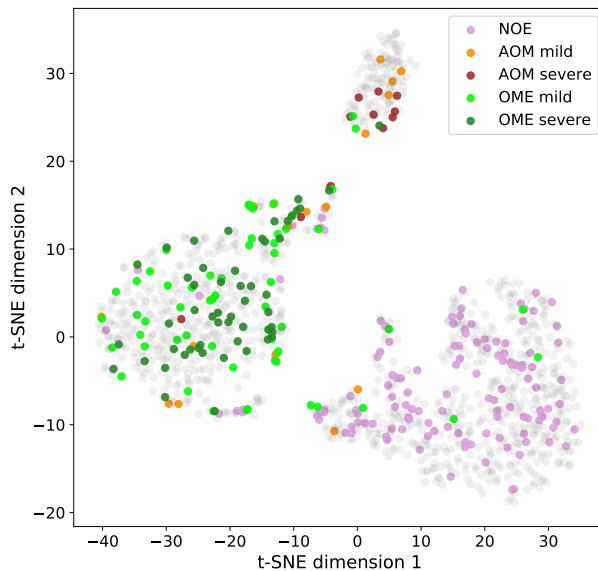
#### 4.1.1 Otoscopy classification (Paper A)

Paper A addresses the problem of diagnosing otitis media based on otoscopy images of the tympanic membrane. This work considers several challenges of this clinical problem: The images in the dataset have a large imbalance among the three classes: AOM, OME, and NOE; the images are of varying diagnostic difficulty and image quality; and there is bound to be some label noise, as the ground truth labels are based on the diagnosis made by a single doctor.



We proposed employing deep metric learning for this classification task and showed that a triplet-based neural network handled the class imbalance better than, e.g., one based on class weighted cross-entropy loss. Another benefit of the deep metric learning approach is the output embedding features of all image examples. A plot of the embedding space is shown in Figure 4.1 using t-SNE dimensionality reduction [76]. A plot like this allows interpretation of the distribution of the data. The AOM and OME cases were further labelled mild and severe. This graduation of the severity is also plotted in the figure and shows how the most severe test cases are generally in the center of the cluster, while the milder cases can be found in other class clusters or on the border between the clusters. This shows that the model is more certain of the diagnosis of severe cases, most likely due to the clear diagnostic signs in the otoscopy images. This relationship between the distribution in the embedding space and the diagnostic difficulty was further examined in Paper F.

The embeddings were used to classify the test cases into the three diagnostic groups, and the overall accuracy of the approach was 86%. This is a similar performance as similar earlier studies. Senaras et al. [62] diagnosed healthy and otitis



**Figure 4.1:** Visualizations of train and test embeddings from Paper A. The colored points show the test dataset, while the grey points show the training dataset.

media ears, without the subclassification of AOM and OME, with an accuracy of 84.4%, Kuruvilla et al. [38] predicted AOM, OME, and NOE with an accuracy of 85.6%, and Wu et al. [83] achieved a 97.8% accuracy on the same three classes. It is, however, impossible to directly compare the performance of these studies, as no benchmark dataset is available. The performance relies heavily on the quality of the input images and ground truth labels. The ground truth labels for the study by Kuruvilla et al. [38] was provided by a panel of three clinical experts, while it was a consensus diagnosis evaluated by two clinical experts for the study by Wu et al. [83]. Furthermore, Wu et al. [83] excluded blurry images and images where the full tympanic membrane was not visible. This ensures high quality images in the dataset, but it also removes the challenging cases. The performance might therefore not reflect the realistic performance of the trained model on standard clinical data. We decided to include all images in the dataset to include a natural variance in image quality, diagnostic difficulty, etc. This means that there probably is a limit for the achievable accuracy on our dataset, but it also allows us to develop methods to identify challenging cases, and examine the diagnostic challenges further.

We still want to find ways to improve this diagnostic model and identify the challenges of this classification task. Table 3 in Paper A shows the confusion matrix of the full dataset. The correctly classified images in the green cells show textbook examples of the three conditions of the middle ear, whereas the diagnostic signs and image quality of the misclassified images in the red cells are much more varied. The misclassified images show common otoscopy challenges, such as earwax, narrow ear canals, and blurry images. These image examples raised questions such as whether there was an upper limit for the achievable performance of this task. Since the ground truth diagnosis is based on manual annotations from only one ENT, there are bound to be some annotation errors. We also do not know whether crucial diagnostic information is missing when only the image is assessed. The ground truth diagnosis is based on a full examination of the patient, and the ENT will therefore have additional information that the model does not have access to. These results and considerations motivated the human inter-rater study presented in Paper E and will be discussed further in Section 4.3.1. Paper A thus shows that classification of otitis media based on otoscopy images is possible, but this work also raised many other questions and relevant research areas related to the development of a deep learning model for this task.

#### 4.1.2 WBT classification (Paper B)

Besides the otoscopy images, the dataset also contains WBT measurements from the patients. Paper B presents our approach for automatic classification of otitis media based solely on these WBT measurements. As mentioned in Section 3.2.2, Helenius et al. [27] showed that standard 226 Hz tympanometry measurements do not allow discrimination of specific types of otitis media, but are useful in distinguishing between no effusion (NOE) and otitis media (OM). Based on this knowledge, the first

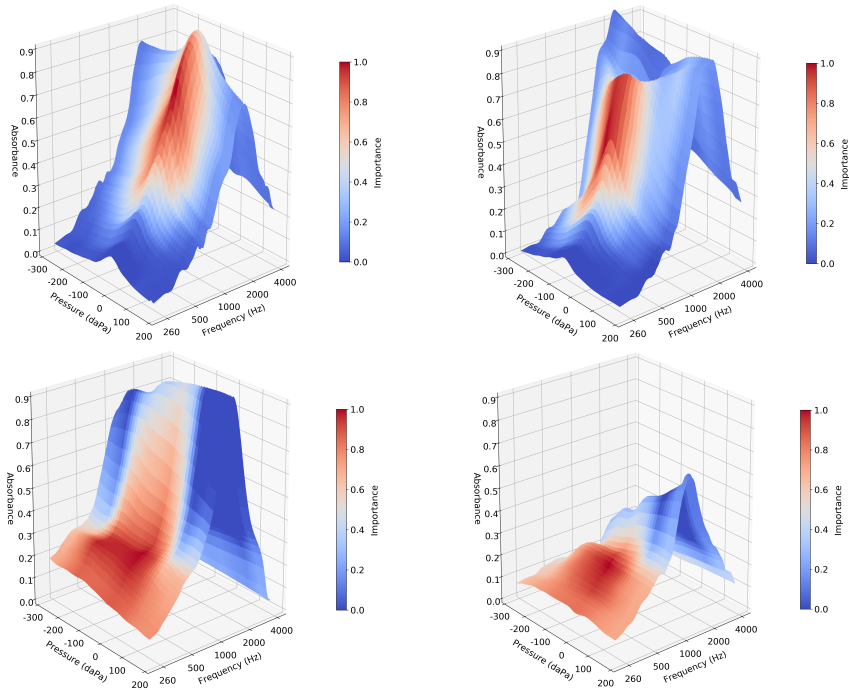
approach presented in our paper is the classification of NOE and OM based on the WBT measurements. Some of the main contributions of this paper include the pre-processing procedure for WBT measurements and experiments with various types of data augmentation for this specific type of data. WBT measurements differ from the standard image input for convolutional neural networks. We are analysing physical data, where the measurement grid is the same for all input examples. Standard data augmentation using geometric transformations is thus not feasible, as it would change the nature of the data. The paper presents experiments with a variety of noise distortions and intensity manipulations for data augmentation in both 2D and 1D for the proposed approaches, showing how this improves the performance of the model.

Our approach for binary classification of NOE and OM shows a very good classification accuracy of 92.6%, and we show that the network trained on full WBT measurements achieves higher performance than networks trained on either tympanogram or ambient absorbance alone. For the second approach, where we attempt to distinguish between AOM, OME, and NOE, the accuracy drops drastically to 70.9%. Recall and precision are still high for the NOE class, but the AOM and OME classifications show low performance. This shows that the WBT does not contain diagnostic information on the specific type of otitis media and that more patient information is needed for the subclassification. However, the high performance of the binary classification task suggests that WBT would add great value to an automatic diagnostic model in combination with other patient data. These results thus inspired the work presented in Paper G, where we combine otoscopy images and WBT measurement into a single classification model.

Another important part of this work is the utilisation of GradCAM for the interpretation of the model output. These saliency maps show the key feature areas in the input WBT, leading to the final diagnostic decision made by the model, and can be generated for each input example. Examples of saliency maps are found in Figure 4.2. This adds great value to a diagnostic system, as it guides the user of the model and instils trust in the model. At the same time, it can be used as a training tool for medical professionals without specific WBT training, as they will be presented with the most important diagnostic features of each WBT, as the model analyses them. As seen in the top row of Figure 4.2, the most important features of the NOE cases are around the tympanic peak pressure and in frequencies of 500 to 1000 Hz, while the OM cases in the bottom row have key features in the low-frequency area and on the full pressure axis. Thus, it is important for diagnosing the OM cases to evaluate whether or not there is an alteration in absorbance along the pressure axis. These observations correspond well to the known features of WBT measurements, as also discussed in Paper B.

## 4.2 Generative models

Paper C and D investigate two different aspects of generative models. Paper C shows the use of generated data examples for neural network training and presents a



**Figure 4.2:** Examples of WBT measurements with the saliency map plotted as heatmaps. Top row: NOE, bottom row: OM.

solution to a specific use case for this type of model. The work focusses on domain shifts, which are commonly found in medical image analysis, when models trained on a specific training dataset have to be used on data from, e.g., a different hospital or equipment. Paper D presents a model for the generation of the specific data used in this project.

#### 4.2.1 Domain shifts (Paper C)

The main challenge addressed in Paper C is related to the issue of having two different data distributions in the training and test set, i.e., domain shifts. When collecting medical data, it is very common for data to be acquired with different systems, such as different scanners or cameras. The modality is the same, but the specific systems and/or settings are different, which leads to poor generalization and models that do not work in practice. This contribution was part of the Retinal Fundus Glaucoma Challenge [50, 40], held as part of the MICCAI 2020 conference, in which the objective was to evaluate and compare automated algorithms for glaucoma detection and optic disc, and cup segmentation on a common dataset of retinal fundus images. The

challenge had three tasks: glaucoma classification, optic disc and cup segmentation, and fovea detection. The challenge dataset was collected at different sources: the training data was collected using two different cameras, while the test data was collected using a third camera. This is a very likely scenario for a medical image analysis case. Normally, it would be very challenging to train a neural network, or any other prediction model, on a training set and then apply it to data from another camera at test time. A neural network is a data-driven model and thus learns the descriptive features of the training data set, and these features cannot easily be transferred to data from another distribution.

CycleGAN is an unpaired image-to-image translation model that learns a mapping between two image domains, as described in Section 2.2.2.2. The model does not require pairs of images from each domain as it learns the special characteristics of the original domains from a set of examples and figures out how to translate these characteristics into the other domain. As shown in Paper C, this generative model is used to translate image examples between the three domains and generate fake images from the test domain for training. Table 1 in the paper shows examples of both real and fake images, and these clearly indicate the impressive generative performance of the cycleGAN model. This approach allows the neural networks for the three different tasks to be trained on fake images from the test domain, while at the same time increasing the size of the training dataset drastically, whereby the networks learn descriptive features and characteristics of the test domain.

This work demonstrates how a cycleGAN can be used for a clinical problem, and even though it is shown in the specific task of retinal fundus images, it is applicable in many fields. The artificial extension of training data is a key issue to keep in mind when employing a neural network in a new test dataset. This is especially important in medical image analysis, where it is time-consuming and expensive to acquire data, and where data likely will stem from different sources, for example, various hospitals. It is also crucial for this PhD project, as the data used for training is from a single clinic with annotations from one ENT. Should this model be employed in another clinic or with different equipment, domain adaptation would probably be needed.

## 4.2.2 Data augmentation (Paper D)

Another widely used application of generative models is data augmentation. Data augmentation is especially important in medical image analysis since the available datasets are usually quite small. Neural networks typically have millions of parameters that are tuned during training, but this process relies on a large amount of training data. It can be challenging to acquire data for certain medical conditions, for example, if the group of patients is small for rare diseases, while it is easier to acquire normal data. It is therefore valuable to be able to generate data in specific classes to equalize class imbalance in order to improve the training of classification models.

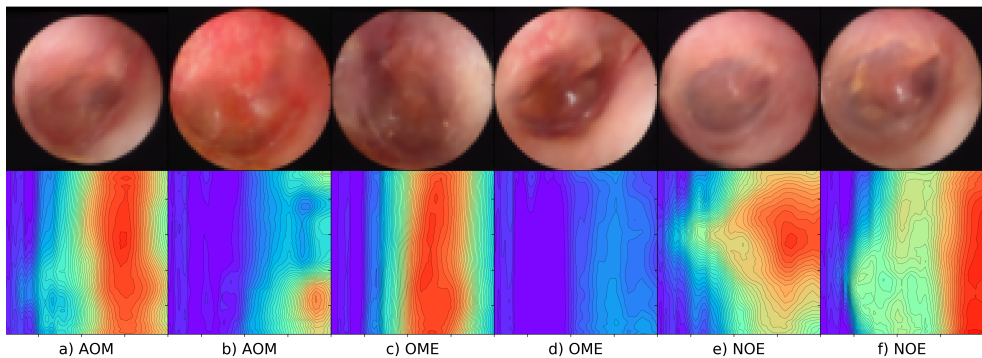
The specific dataset used in this project also has a class imbalance problem, as

the dataset contains fewer AOM cases compared to NOE and OME. In Paper A, we included simple data augmentation during training, such as geometric transformations (flipping, rotation) and colour alterations of the input images. In Paper B, geometric transformations were not fitting for the input WBT data, and instead we experimented with various types of noise as data augmentation. In both papers, it was shown that even simple data augmentation increased the performance of the trained networks. Therefore, we wanted to explore more advanced methods for data augmentation, more specifically generative models.

Paper D presents our work on developing a generative model which can generate pairs of otoscopy images and WBT measurements in each of the three diagnostic groups: AOM, OME, and NOE. This is achieved by employing a variational autoencoder with a metric learning loss function in the embedding space. During training, the embedding space will be enforced to generate class clusters, which can be used for sampling new embedding vectors for generating new data. The implementation details of this model are explained in the paper.

There is, of course, great variability in the appearance of both otoscopy images and WBT measurements, depending on the severity of the symptoms. The generated data show the same range from mild to severe symptoms, and the generated pairs show that the two modalities are representing the same severity. This is shown in Figure 4.3. Figure 4.3 a) and c) show mild cases of AOM and OME, where the respective otoscopy image shows no severe signs of otitis media, and the absorbance in the WBT is also high. On the other hand in Figure 4.3 b) and d) the otoscopy images show a severe infection in the AOM case and effusion in the OME case, accompanied by very low absorption values in both WBT measurements.

This paper indicates that it is possible to generate correlated pairs of data. The



**Figure 4.3:** Generated pairs of otoscopy images and WBT measurements. a) mild AOM case, b) severe AOM case, c) mild OME case, d) severe OME case, e) NOE case, and f) NOE case.

model is currently restricted to generating  $64 \times 64$  pixel images, which is much lower than the resolution required to train the classification models in Paper A ( $299 \times 299$ ). Furthermore, there is still a challenge with blurry generated images for the proposed model. The practical use of the generated data is thus left for future work due to research priorities. For the time being, this is a proof-of-concept of the triplet-based variational autoencoder for pairwise data generation.

## 4.3 Diagnostic difficulty

Otitis media is a challenging disease to diagnose, even for trained specialists. This has been well established by previous studies, as described in Chapter 3. However, it has not been investigated how the availability of a WBT affects the diagnostic difficulty, nor do we know the difficulty of cases in the dataset. This is investigated in the human inter-rater study presented in Paper E. Furthermore, we developed a method to predict this diagnostic difficulty from embeddings obtained using deep metric learning, which is presented in Paper F.

### 4.3.1 Human inter-rater study (Paper E)

The main goals of Paper E were to investigate which cases in the otitis media dataset are more difficult than others and to determine a reinforced ground truth diagnosis of each case in the dataset by eliminating the human annotator errors. The diagnoses of the original dataset were determined by one expert ENT based on otoscopy examination, WBT measurements, symptoms, and patient history. This study thus allows an investigation of how other expert ENTs would diagnose each of the cases if only presented with the otoscopy image and the WBT.

The "true" ground truth diagnosis of otitis media can only be determined by performing a myringotomy, which is an incision in the eardrum, and analysing the content of the middle ear. Since this information is not available, we have to rely on the diagnosis made non-invasively by doctors. There are bound to be some human errors in the annotation of a large dataset of cases, and the idea was to attempt to eliminate these errors by obtaining several diagnoses for each case from a panel of expert ENTs. As seen in Table 2 of Paper E, there is, however, large discrepancies between the diagnostic assessment made by the original ENT, Dr. Kamide, who examined the patients, and the annotations by the four ENTs in the study. The majority voting among the four ENTs shows that several cases that were diagnosed as NOE by Dr. Kamide, were now diagnosed as OME. As discussed in the paper, ENTs who only examine the otoscopy images and WBT measurements seem to be more prone to detect disease than no effusion. The four ENTs agreed with the Dr. Kamide's diagnosis in 65.1% of the cases on average (min: 60.5%, max: 72.3%) on the test dataset. This low agreement, and the predisposition to diagnose OM, indicate that these annotations cannot be used to establish a better ground truth diagnosis.

There are clearly some challenges for the ENTs in the diagnostic process when they are only presented with the image and WBT and do not have the possibility to examine the patient. Instead of using the annotations to estimate a new ground truth, we can instead use them to determine the diagnostic difficulty of each case. If the ENTs all agree with the original diagnosis, it is expected that the case is easy to diagnose, whereas if some of the ENTs disagree with the original diagnosis, the case is probably more challenging. This human inter-rater study also shows that self-evaluated diagnostic certainty correlates well with the agreement between the ENTs, and that certainty and agreement increased when presented with both the otoscopy image and WBT, compared to only evaluating the image. These observations validate the use of agreement and certainty as a measurement of diagnostic difficulty, which is further explored in Paper F.

This human inter-rater study evaluated the specific clinical dataset employed in this PhD project, and these additional annotations create a state-of-the-art clinical dataset. The annotations establish a performance benchmark for our deep learning models since the expert ENTs only evaluated the otoscopy image and WBT measurement. This allows one to determine a realistic performance based on this limited patient information.

### 4.3.2 Estimating human annotation difficulty (Paper F)

The annotations collected in the human inter-rater study in Paper E were used to estimate the difficulty of each otoscopy image in the dataset, which was further explored in Paper F. The difficulty is computed based on the annotations of diagnosis and self-assessed certainty from the ENTs. If all four ENTs agree with the original diagnosis made by Dr. Kamide and they rated high certainty, the diagnostic difficulty is low. On the other hand, if the four ENTs do not all agree with Dr. Kamide, and rated a lower certainty, the diagnostic difficulty is higher. It is computed as

$$D = 1 - \mu_{correct} \cdot \mu_{certainty} , \quad (4.1)$$

where  $\mu_{correct}$  is the fraction of correct ENT answers and  $\mu_{certainty}$  is the average self-evaluated certainty.

The difficulties were evaluated using a “leave-one-annotator-out” analysis, where the difficulty estimated from one annotator and the estimate from the other three were compared. The Kendall’s  $\tau$  of the difficulty from one annotator against the difficulty estimated from the rest shows how consistent the annotations are among the annotators. Table 4.1 shows the results of this evaluation of the test dataset used in the paper, which consists of 204 images annotated by all four ENTs. In previous studies, only the fraction of incorrect diagnoses from the raters, or  $\mu_{correct}$ , has been used to define difficulty [4]. Table 4.1 therefore includes the evaluation of difficulty computed with and without certainty. The results show that including certainty in the definition of difficulty increases the consistency of the estimated difficulty. Kendall rank correlation coefficient, also called Kendall’s  $\tau$  [33], is a non-parametric measurement of the correlation between two ranked variables. It evaluates



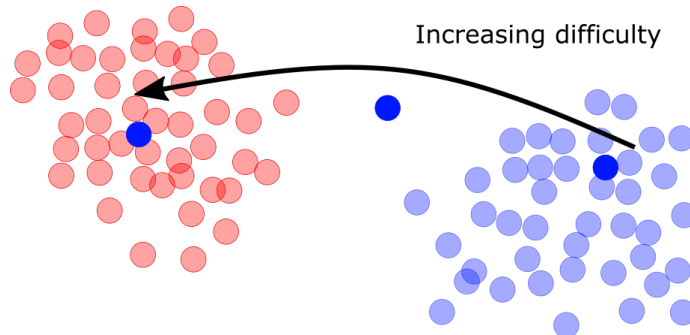
	Mean	Min	Max
<b>Difficulty consistency (Kendall’s <math>\tau</math>)</b>			
Difficulty with certainty	$0.570 \pm 0.036$	0.513	0.611
Difficulty without certainty	$0.548 \pm 0.057$	0.456	0.607
<b>Prediction accuracy</b>	$69.5 \pm 2.5\%$	67.2%	73.5%

**Table 4.1:** Top: Kendall’s  $\tau$  for the comparisons of responses of one annotator to the difficulties estimated from the remaining annotators. The number after  $\pm$  indicates the standard deviation. Bottom: The accuracy of the annotators.

the relationship between the two variables based on the ordering, or ranking, of the samples. It is thus not important to achieve the same specific difficulty value as the ground truth, but the ordering has to be the same in order to achieve a high Kendall’s  $\tau$ . A Kendall’s  $\tau$  of 0.570 corresponds to having ranked 78.5% of the images in the correct order. The small standard deviation also shows that no annotator stands out from the group with a much lower agreement than the rest. Furthermore, the table shows that the four ENTs only agree with the original diagnosis made by Dr. Kamide in 69.5% of the cases.

In Paper F, we present methods for estimating the diagnostic difficulties of medical images based on the embedding of the image obtained using deep metric learning. We evaluate our methods on both the otoscopy data from this PhD project, and on a skin lesion dataset. In this paper, we show that there is a correlation between the placement of an input image in the embedding space and the diagnostic difficulty. This concept is shown in Figure 4.4. The intuition is that the three dark blue points increase in difficulty as they move out of the blue cluster and into the red cluster. The difficulty can thus be estimated on the basis of the position in the embedding space and the ground truth class. We show that we can estimate the difficulty unsupervised based on this principle of distance in the embedding space, and we also present supervised methods for difficulty estimation.

This relationship between the diagnostic difficulty of a case and the embedding space was briefly discussed in Paper A. At the time of writing Paper A, we did not have the additional annotations, but Dr. Kamide evaluated whether the otitis media cases were mild or severe. We expect that the severe cases are easier to diagnose, as the diagnostic signs and symptoms are clearer compared to the mild cases, where the diagnostic signs are not as pronounced and the diagnosis could be made based on symptoms like ear pain, which is not included in our recorded information about the patients. The average difficulty of the mild AOM and OME cases in the dataset is 0.60 and 0.34, respectively, and for the severe AOM and OME cases, it is 0.23 and 0.25, respectively, so our expectations were correct. In Figure 4.1, the severity is included in the plot of the embeddings, and the pattern is similar to the one seen in Figure 2 of Paper F: the mild cases, which are also more difficult to diagnose, are



**Figure 4.4:** Samples from two classes (red and blue) are visualized in a two-dimensional space. The arrow shows how the difficulty increases, as the blue point moves into the red cluster. Classification difficulty thus depends not only on the position in the space but also on the ground truth class.

generally placed furthest away from the cluster center or even within other clusters, while the severe and easier cases are primarily placed within the class cluster.

The diagnostic difficulty is another valuable output of a future diagnostic tool. Being able to estimate the difficulty of every analysed case allows the operator to better assess the patient and the model output. If the estimated difficulty is high, it can be recommended to refer the patient to an expert ENT for further examination. It could also mean that the quality of the input data is too low for the model to infer a diagnosis. As seen in Figure 1 in Paper F, the diagnostic difficulty is clearly affected by the quality of the images. In such cases, the operator could remove ear wax or other obstructions from the ear canal, or simply redo the otoscopy and WBT measurements and run the analysis again.

## 4.4 Combining it all (Paper G)

Paper G ties together several previous contributions for otitis media classification and difficulty estimation based on both otoscopy images and WBT measurements. The network architecture employed in this paper is a combination of the network architecture for otoscopy image classification from Paper A and the WBT network from Paper B. Paper G also concerns itself with the estimation of diagnostic difficulty, defined for each case in the dataset based on the annotations from the human inter-rater study presented in Paper E. Deep metric learning is employed for the training of a multi-modal embedding network, and from the embedding space, the methods from Paper F are utilised for classification and difficulty estimation, respectively. The deep metric learning approach is compared with the performance of a multi-task neural network, which also predicts both diagnostic group and difficulty. As seen in Paper

E, the diagnostic certainty and agreement of the ENTs is increased when presented with both otoscopy image and WBT, compared to otoscopy alone. The assumption was that this will also be the case for the prediction models.

The results presented in Paper G show that both classification and difficulty estimation performance is increased when including both otoscopy images and WBT measurements in the model, compared to a model trained solely on the individual modalities. The improvement from combining the modalities is not drastic, but still worth pursuing. The performance table and confusion matrices in Paper G show that detection of AOM is improved by adding WBT measurements to the model, which is crucial to ensure proper treatment.

The annotations by Dr. Kamide are used as the ground truth labels for the evaluation of the models. The best performing network achieves a classification performance of 86.5%, while the average accuracy of the four ENTs who also examined only the otoscopy images and WBT measurements from Paper E was 64.0%. Despite the fact that our deep learning models have the same limited information as the ENTs, the models are able to achieve a much better performance. This is very promising for a future diagnostic tool and shows the strength of deep learning models.

A further investigation of the performance of the model compared to the four ENTs is shown in Table 4.2. The table show how the true positive rate (TPR) of the network predictions and ENT annotations differ between the diagnostic groups, and both ground truth and estimated difficulty are also shown. As discussed in Paper G, the TPR and difficulty are very different among diagnostic groups. Both network and ENT TPR is lower for the mild cases than for the severe, and the diagnostic difficulty is thus higher for the mild cases. It is seen in the table that the average predicted difficulty for each of the groups is close to the ground truth difficulties computed from the ENT annotations. The lower performance in cases of mild otitis media is a challenge if this model is to be used as a diagnostic tool. It is expected that this limitation arises from the limited patient information available for both the models and ENTs in Paper E. Additional patient information, such as a list of symptoms or

	OME		AOM		NOE
	Mild	Severe	Mild	Severe	
<b>Network predictions</b>					
True positive rate [%]	73.2	92.0	62.3	85.1	91.5
Average predicted difficulty	0.35	0.24	0.48	0.32	0.66
<b>ENT annotations</b>					
True positive rate [%]	82.2	88.1	52.1	88.7	49.3
Average difficulty	0.34	0.25	0.60	0.23	0.67

**Table 4.2:** True positive rate and average difficulty for each diagnostic group for network predictions and ENT annotations on the full dataset.

patient temperature, is clearly necessary for proper identification of mild cases.

Paper G demonstrate some of the strengths of deep metric learning over the standard classification, or multi-task, networks. As also shown in Paper A, the deep metric learning approach handles class imbalance better than the standard classification network, which means that the prediction of AOM is improved. Another strength is the fact that a full dataset of ground truth difficulties is not needed for the training of the neural network, as the embeddings are only based on diagnostic class. This allows the training dataset to include cases where ground truth difficulties are not known, thus reducing the cost of acquiring the training dataset. This makes it possible to add new cases to the dataset, possibly improving the model even further, without the need for four additional ENTs to evaluate each case.

## 4.5 Further challenges

This project is a proof-of-concept for the diagnosis of otitis media based on this high-quality dataset from the Kamide ENT clinic. While the methods presented in the contributions show impressive results and performance, models like these are not ready to be deployed directly in a diagnostic tool to be used in the clinic. There are many remaining challenges regarding the implementation of deep learning methods in the clinic, as also discussed in the review paper by Habib et al. [23]. The challenges highlighted by Habib et al. include acquiring properly labelled datasets based on consensus among multiple independent experts, and evaluation of real-life test performance and applicability in daily clinical practice. Pichichero also raised several concerns regarding automatic diagnosis based on otoscopy in a recent commentary paper [54]. Thus questions were raised related to the practicalities of a digital diagnostic tool in the clinic, such as whether these models would still be better than doctors to diagnose otitis media even if ear wax obscures the view of the tympanic membrane, or whether the doctor would want to use a digital diagnostic tool if the examination time is increased. These are all very valid points of concern and it is important to ensure a good and time-efficient workflow for doctors with a future digital diagnostic tool.

For the deployment of deep learning models, the most important thing to ensure is a good training dataset. The dataset used for this PhD project works very well for a proof-of-concept but is limited by being collected in only one specific clinic, annotated by only one ENT, collected with the same equipment for all patients, from a very specific patient group of Japanese children. Furthermore, the otoscopy equipment used captures high-quality images, beyond what is common in a standard clinic. Since deep learning models are data-driven models, they do not generalise well to unseen data from other distributions, even though the task is the same. Therefore, models trained in this data set are likely limited to work on data from this clinic with specific equipment, and the model has learnt to mimic the decision process of the ENT, who annotated all cases. For the use of these models in other clinics, the

cycleGAN-based domain adaptation used in Paper C could be used to fine-tune the model on images from the new domain.

Another challenge, raised by both Habib et al. and Pichichero, is to ensure proper data quality for the models when a model is being used in the clinic. A digital diagnostic system is not expected to be able to analyse all otoscopy images, and a certain image quality must be ensured for a proper analysis result. This could be done with a screening tool, which could prompt the operator to retake the image if the quality is not good enough, or if ear wax obscures the view of the tympanic membrane. A preliminary investigation of this idea was undertaken in terms of the BSc project "Image-based quality evaluation of otoscopy images", where the student developed an algorithm for detecting blurry images, together with an application prompting the user to retake the image, if the image is too blurry [52]. There is, however, still much more work to be done on this topic.

# CHAPTER 5

## Conclusion

---

In this thesis, we have introduced and discussed the research that was carried out as a part of this PhD project. The work resulted in several scientific publications related to the automatic diagnosis of otitis media. This overall theme was investigated in various ways, from both a technical and clinical point of view, and with various methods for classification, data generation, and diagnostic difficulty estimation.

The thesis contributes with new methods for automatic diagnosis of otitis media, which can be used to improve the diagnostic process in the clinic. Paper A, B, and G present the three models constituting the originally intended outcome of this PhD project. Paper A contributes to a growing field of interest in deep learning-based approaches for otoscopy image analysis, while Paper B was one of the first to present a deep learning model for WBT classification. Similarly, Paper G is the first publication to present a multi-modal classification approach based on otoscopy images and WBT measurements. These publications show potential for the development of a diagnostic tool based on these two types of data.

During the work on the classification models, several challenges were encountered. The first issue to be addressed was the class imbalance in the dataset. This was addressed in the design of the classification model in Paper A, but this also sparked an interest in developing a generative model for multi-modal data generation. Paper D delivers just that, by utilizing deep metric learning to develop a conditional multi-modal generative model. Secondly, during the work on Paper A, it quickly became obvious that there was a limit to the achievable performance based on the current diagnostic labels. This started a discussion on the best way to conduct an evaluation of the dataset and possibly reinforce the ground truth labels. This resulted in the human inter-rater study presented in Paper E. This study shed light on the clinical aspects of this PhD project, as it became evident that even ENTs find it difficult to diagnose patients only based on otoscopy images and WBT measurements; not having the patient in front of them is a serious disadvantage. Instead of using the annotations to improve the ground truth diagnosis, they were utilised to determine the diagnostic difficulty for each of the cases in the dataset. These difficulty ratings were then used for the work in Paper F, which presents a difficulty estimation approach based on image embeddings.

In the summer of 2020, I participated in the Retinal Fundus Glaucoma challenge together with two fellow PhD students, which resulted in Paper C. This was a side project not directly related to the PhD project, but the theme of domain shifts in medical image datasets fits well into the common thread of this PhD.

The main contributions from the work in this PhD project can be summarized as follows:

- Presented a deep learning model for otitis media classification based on otoscopy images (Paper A).
- Showed that deep metric learning handles class imbalance better than the standard deep learning loss functions (Paper A and G).
- Demonstrated the use of various types of data augmentation for biomedical non-image data based on noise and intensity manipulation (Paper B).
- Presented a deep learning model, including GradCAM for generation of saliency maps, for the classification of otitis media based on WBT measurements (Paper B).
- Showed that cycleGANs can be applied for domain shifts in order to handle the classic medical image analysis challenge of datasets from various sources (Paper C).
- Developed a triplet-based variational autoencoder for conditional generation of multi-modal data (Paper D).
- Conducted a human inter-rater study on the diagnosis of otitis media and presented a statistical analysis of the responses from four ENTs (Paper E).
- Developed methods for both unsupervised and supervised estimation of diagnostic difficulty based on image embeddings (Paper F).
- Presented the first multi-modal classification model for the diagnosis of otitis media based on otoscopy images and WBT measurements (Paper G).

This PhD project is an example of applied research conducted in close collaboration with industry. The Interacoustics supervisors have expert domain knowledge on otoscopy and WBT, and the applications of these diagnostic tools, which has been a huge help during the project, while the DTU supervisors have provided guidance in the technical part of the project. Furthermore, we have collaborated with clinical experts; Dr. Kamide and the ENTs from Lund University Hospital. These collaborations have allowed us to provide new insights into the automatic diagnosis of otitis media, both from a clinical and technical point of view.

# Bibliography

---

- [1] Venkatesh Aithal, Sreedevi Aithal, Joseph Kei, Shane Anderson, and David Wright. “Predictive Accuracy of Wideband Absorbance at Ambient and Tympanometric Peak Pressure Conditions in Identifying Children with Surgically Confirmed Otitis Media with Effusion.” In: *Journal of the American Academy of Audiology* 31.7 (2020), pages 471–484. ISSN: 21573107. DOI: 10.3766/jaaa.19012.
- [2] Adi Alhudhaif, Zafer Cömert, and Kemal Polat. “Otitis media detection using tympanic membrane images with a novel multi-class machine learning algorithm.” In: *PeerJ Computer Science* 7 (2021), pages 1–22. ISSN: 23765992. DOI: 10.7717/PEERJ-CS.405.
- [3] American Academy of Pediatrics. “Clinical Practice Guideline: Otitis Media With Effusion.” In: *Pediatric Research* 113.5 (2004). ISSN: 0031-3998. DOI: 10.1203/00006450-198504000-00002.
- [4] Barbara André, Tom Vercauteren, Anna Buchner, Muhammad Waseem Shahid, Michael Wallace, and Nicholas Ayache. “An image retrieval approach to setup difficulty levels in training systems for endomicroscopy diagnosis.” In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2010, pages 480–487. DOI: 10.1007/978-3-642-15745-5\_59.
- [5] Karim Armanious, Chenming Jiang, Sherif Abdulatif, Thomas Küstner, Sergios Gatidis, and Bin Yang. “Unsupervised medical image translation using Cycle-MeDGAN.” In: *European Signal Processing Conference*. 2019. ISBN: 9789082797039. DOI: 10.23919/EUSIPCO.2019.8902799.
- [6] J. Harry Baumer. “Comparison of two otitis media guidelines.” In: *Archives of Disease in Childhood: Education and Practice Edition* 89.3 (2004). ISSN: 17430585. DOI: 10.1136/adc.2004.065490.
- [7] Hamidullah Binol, Aaron C. Moberly, M. Khalid Khan Niazi, Garth Essig, Jay Shah, Charles Elmaraghy, Theodoros Teknos, Nazhat Taj-Schaal, Lianbo Yu, and Metin N. Gurcan. “Decision fusion on image analysis and tympanometry to detect eardrum abnormalities.” In: *Medical Imaging 2020: Computer-Aided Diagnosis*. Volume 11314. 2020, pages 375–382. ISBN: 9781510633957. DOI: 10.1117/12.2549394.



- [8] Karin Blomgren and Anne Pitkäranta. “Is it possible to diagnose acute otitis media accurately in primary health care?” In: *Family Practice* 20.5 (2003), pages 524–527. ISSN: 02632136. DOI: 10.1093/fampra/cm9505.
- [9] Jimmy Cé Lind, Liv Södermark, and Ola Hjalmarson. “Adherence to treatment guidelines for acute otitis media in children. The necessity of an effective strategy of guideline implementation.” In: *International Journal of Pediatric Otorhinolaryngology* 78.7 (2014), pages 1128–1132. ISSN: 18728464. DOI: 10.1016/j.ijporl.2014.04.029.
- [10] Dongchul Cha, Chongwon Pae, Si Baek Seong, Jae Young Choi, and Hae Jeong Park. “Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database.” In: *EBioMedicine* 45 (2019), pages 606–614. ISSN: 23523964. DOI: 10.1016/j.ebiom.2019.06.050.
- [11] Stephanie M. Chandler, Shawn M.S. Garcia, and David P. McCormick. “Consistency of diagnostic criteria for acute otitis media: A review of the recent literature.” In: *Clinical Pediatrics* 46.2 (2007), pages 99–108. ISSN: 00099228. DOI: 10.1177/0009922806297163.
- [12] Children’s Hospital of Philadelphia. *Otitis Media with Effusion (OME)*. URL: <https://www.chop.edu/conditions-diseases/otitis-media-effusion-ome> (visited on October 1, 2019).
- [13] Nisa Eda Cullas Ilarslan, Fatih Gunay, Seda Topcu, and Ergin Ciftci. “Evaluation of clinical approaches and physician adherence to guidelines for otitis media with effusion.” In: *International Journal of Pediatric Otorhinolaryngology* 112 (2018), pages 97–103. ISSN: 18728464. DOI: 10.1016/j.ijporl.2018.06.040.
- [14] Steffen Czolbe, Oswin Krause, Ingemar Cox, and Christian Igel. “A loss function for generative neural networks based on Watson’s perceptual model.” In: *Advances in Neural Information Processing Systems*. Volume 33. 2020, pages 2051–2061. arXiv: 2006.15057.
- [15] Amina Danishyar and John V. Ashurst. *Acute otitis media*. StatPearls Publishing LLC, 2019.
- [16] Paula Lopez Diez, Kristine Aavild Juhl, Josefine Vilsbøll Sundgaard, Hassan Diab, and Jan Margeta. “Deep Reinforcement Learning for Detection of Abnormal Anatomies.” In: *Proceedings of the Northern Lights Deep Learning Workshop*. Volume 3. 2022. DOI: 10.7557/18.6280.
- [17] Xue Dong, Tonghe Wang, Yang Lei, Kristin Higgins, Tian Liu, Walter J. Curran, Hui Mao, Jonathon A. Nye, and Xiaofeng Yang. “Synthetic CT generation from non-attenuation corrected PET images for whole-body PET imaging.” In: *Physics in Medicine and Biology* 64.21 (2019). ISSN: 13616560. DOI: 10.1088/1361-6560/ab4eb7.

- [18] John C. Ellison, Michael Gorga, Edward Cohn, Denis Fitzpatrick, Chris A. Sanford, and Douglas H. Keefe. “Wideband acoustic transfer functions predict middle-ear effusion.” In: *Laryngoscope* 122.4 (2012), pages 887–894. ISSN: 0023852X. DOI: 10.1002/lary.23182.
- [19] Glenn Flores, Mina Lee, Howard Bauchner, and Beth Kastner. “Pediatricians’ attitudes, beliefs, and practices regarding clinical practice guidelines: A national survey.” In: *Pediatrics* 105.3 (2000), pages 496–501. ISSN: 00314005. DOI: 10.1542/peds.105.3.496.
- [20] Ian Goodfellow. “NIPS 2016 tutorial: Generative adversarial networks.” In: *arXiv preprint arXiv:1701.00160* (2016). arXiv: 1701.00160.
- [21] Ian J Goodfellow, Jean Pouget-abadie, Mehdi Mirza, Bing Xu, and David Warde-farley. “Generative Adversarial Nets.” In: *Advances in Neural Information Processing Systems*. Volume 27. 2014.
- [22] Emad M Grais, Xiaoya Wang, Jie Wang, Fei Zhao, Wen Jiang, and Yuexin Cai. “Analysing wideband absorbance immittance in normal and ears with otitis media with effusion using machine learning.” In: *Scientific Reports* 11.1 (2021), pages 1–12. ISSN: 2045-2322. DOI: 10.1038/s41598-021-89588-4.
- [23] Al-Rahim Habib, Majid Kajbafzadeh, Zubair Hasan, Eugene Wong, Hasantha Gunasekera, Chris Perry, Raymond Sacks, Ashnil Kumar, and Narinder Singh. “Artificial intelligence to classify ear disease from otoscopy: A systematic review and meta-analysis.” In: *Clinical Otolaryngology* (2022). DOI: 10.1111/coa.13925.
- [24] Raia Hadsell, Sumit Chopra, and Yann LeCun. “Dimensionality reduction by learning an invariant mapping.” In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2006, pages 1735–1742. ISBN: 0769525970. DOI: 10.1109/CVPR.2006.100.
- [25] Morten Rieger Hannemose, Josefine Vilsbøll Sundgaard, Niels Kvorning Ternov, Rasmus R. Paulsen, and Anders Nymark Christensen. “Was that so hard? Estimating human classification difficulty.” In: *arXiv preprint arXiv:2203.11824* (2022), pages 1–10. arXiv: 2203.11824.
- [26] Thais Antonelli Diniz Hein, Stavros Hatzopoulos, Piotr Henryk Skarzynski, and Maria Francisca Colella-Santos. “Wideband Tympanometry.” In: *Advances in Clinical Audiology*. IntechOpen, 2017. DOI: 10.5772/67155.
- [27] Kjell K. Helenius, Miia K. Laine, Paula A. Tähtinen, Elina Lahti, and Aino Ruohola. “Tympanometry in discrimination of otoscopic diagnoses in young ambulatory children.” In: *Pediatric Infectious Disease Journal* 31.10 (2012), pages 1003–1006. ISSN: 08913668. DOI: 10.1097/INF.0b013e31825cac94.
- [28] Daniel Jiwoong Im, Sungjin Ahn, Roland Memisevic, and Yoshua Bengio. “Denosing criterion for variational auto-encoding framework.” In: *AAAI Conference on Artificial Intelligence*. Volume 31. 2017. arXiv: 1511.06406.

- [29] Makoto Ito, Haruo Takahashi, Yukiko Iino, Hiromi Kojima, Sho Hashimoto, Yosuke Kamide, Fumiyo Kudo, Hitome Kobayashi, Haruo Kuroki, Atsuko Nakano, Hiroshi Hidaka, Goro Takahashi, Haruo Yoshida, and Takeo Nakayama. “Clinical practice guidelines for the diagnosis and management of otitis media with effusion (OME) in children in Japan, 2015.” In: *Auris Nasus Larynx* 44.5 (2017), pages 501–508. ISSN: 18791476. DOI: 10.1016/j.anl.2017.03.018.
- [30] Peter M. Jensen and Jørgen Lous. “Criteria, performance and diagnostic problems in diagnosing acute otitis media.” In: *Family Practice* 16.3 (1999), pages 262–268. ISSN: 02632136. DOI: 10.1093/fampra/16.3.262.
- [31] Stella U. Kalu and Matthew C. Hall. “A study of clinician adherence to treatment guidelines for otitis media with effusion.” In: *Wisconsin Medical Journal* 109.1 (2010), pages 15–20. ISSN: 10981861.
- [32] Mahmut Kaya and Hasan Sakir Bilge. “Deep metric learning: A survey.” In: *Symmetry* 11.9 (2019). ISSN: 2073-8994. DOI: 10.3390/sym11091066.
- [33] Maurice George Kendall. “Rank correlation methods.” In: (1948).
- [34] Mohammad Azam Khan, Soonwook Kwon, Jaegul Choo, Seok Min Hong, Sung Hun Kang, Il Ho Park, Sung Kyun Kim, and Seok Jin Hong. “Automatic detection of tympanic membrane and middle ear infection from oto-endoscopic images via convolutional neural networks.” In: *Neural Networks* 126 (2020), pages 384–394. ISSN: 18792782. DOI: 10.1016/j.neunet.2020.03.023.
- [35] Diederik P. Kingma and Max Welling. “Auto-encoding variational bayes.” In: *arXiv preprint arXiv:1312.6114* (2013). arXiv: 1312.6114.
- [36] Ken Kitamura, Yukiko Iino, Yosuke Kamide, Fumiyo Kudo, Takeo Nakayama, Kenji Suzuki, Hidenobu Taiji, Haruo Takahashi, Noboru Yamanaka, and Yoshifumi Uno. “Clinical Practice Guidelines for the diagnosis and management of acute otitis media (AOM) in children in Japan - 2013 update.” In: *Auris Nasus Larynx* 42.2 (2015), pages 99–106. ISSN: 18791476. DOI: 10.1016/j.anl.2014.09.006.
- [37] Solomon Kullback and Richard A. Leibler. “On Information and Sufficiency.” In: *The Annals of Mathematical Statistics* 22.1 (1951), pages 79–86.
- [38] Anupama Kuruvilla, Nader Shaikh, Alejandro Hoberman, and Jelena Kovačević. “Automated diagnosis of otitis media: Vocabulary and grammar.” In: *International Journal of Biomedical Imaging* (2013). ISSN: 16874188. DOI: 10.1155/2013/327515.
- [39] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. “Photo-realistic single image super-resolution using a generative adversarial network.” In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2017, pages 105–114. ISBN: 9781538604571. DOI: 10.1109/CVPR.2017.19.

- [40] Fei Li, Diping Song, Han Chen, Jian Xiong, Xingyi Li, Hua Zhong, Guangxian Tang, Sujie Fan, Dennis S.C. Lam, Weihua Pan, Yajuan Zheng, Ying Li, Guoxiang Qu, Junjun He, Zhe Wang, Ling Jin, Rouxi Zhou, Yunhe Song, Yi Sun, Weijing Cheng, Chunman Yang, Yazhi Fan, Yingjie Li, Hengli Zhang, Ye Yuan, Yang Xu, Yunfan Xiong, Lingfei Jin, Aiguo Lv, Lingzhi Niu, Yuhong Liu, Shaoli Li, Jiani Zhang, Linda M. Zangwill, Alejandro F. Frangi, Tin Aung, Ching yu Cheng, Yu Qiao, Xiulan Zhang, and Daniel S.W. Ting. “Development and clinical deployment of a smartphone-based visual field deep learning system for glaucoma detection.” In: *NPJ Digital Medicine* 3.1 (2020), pages 1–8. ISSN: 23986352. DOI: 10.1038/s41746-020-00329-9.
- [41] Allan S. Lieberthal, Aaron E. Carroll, Tasnee Chonmaitree, Theodore G. Ganiats, Alejandro Hoberman, Mary Anne Jackson, Mark D. Joffe, Donald T. Miller, Richard M. Rosenfeld, Xavier D. Sevilla, Richard H. Schwartz, Pauline A. Thomas, and David E. Tunkel. “The diagnosis and management of acute otitis media.” In: *Pediatrics* 131.3 (2013), pages 964–999. ISSN: 00314005. DOI: 10.1542/peds.2012-3488.
- [42] Paula Lopez Diez. *Deep learning for landmark detection and segmentation with geodesic path finding of facial and cochlear nerves*. 2021.
- [43] Paula López Diez, Josefine Vilsbøll Sundgaard, François Patou, Jan Margeta, and Rasmus Reinhold Paulsen. “Facial and Cochlear Nerves Characterization Using Deep Reinforcement Learning for Landmark Detection.” In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2021, pages 519–528. ISBN: 9783030872014. DOI: 10.1007/978-3-030-87202-1\_50.
- [44] Paola Marchisio, Eugenio Mira, Catherine Klersy, Fabio Pagella, Susanna Esposito, Sonia Bianchini, Giuseppe Di Mauro, Michela Fusi, Erica Nazzari, Marta Tagliabue, Luisa Bellussi, and Nicola Principi. “Medical education and attitudes about acute otitis media guidelines: A survey of italian pediatricians and otolaryngologists.” In: *Pediatric Infectious Disease Journal* 28.1 (2009), pages 1–4. ISSN: 15320987. DOI: 10.1097/INF.0b013e318184ef02.
- [45] Gabrielle R. Merchant, Sarah Al-Salim, Richard M. Tempero, Denis Fitzpatrick, and Stephen T. Neely. “Improving the Differential Diagnosis of Otitis Media With Effusion Using Wideband Acoustic Immittance.” In: *Ear & Hearing* 42.5 (2021), pages 1183–1194. ISSN: 0196-0202. DOI: 10.1097/aud.0000000000001037.
- [46] Ionuț Mironica, Constantin Vertan, and Dan Cristian Gheorghe. “Automatic pediatric otitis detection by classification of global image features.” In: *2011 E-Health and Bioengineering Conference (EHB)*. IEEE, 2011, pages 1–4. ISBN: 9781457702921.
- [47] Mehdi Mirza and Simon Osindero. “Conditional Generative Adversarial Nets.” In: *arXiv preprint arXiv:1411.1784* (2014). arXiv: 1411.1784.

- [48] Hermanus C. Myburgh, Stacy Jose, De Wet Swanepoel, and Claude Laurent. “Towards low cost automated smartphone- and cloud-based otitis media diagnosis.” In: *Biomedical Signal Processing and Control* 39 (2018), pages 34–52. ISSN: 17468108. DOI: 10.1016/j.bspc.2017.07.015.
- [49] Hermanus C. Myburgh, Willemien H. van Zijl, De Wet Swanepoel, Sten Hellström, and Claude Laurent. “Otitis Media Diagnosis for Developing Countries Using Tympanic Membrane Image-Analysis.” In: *EBioMedicine* 5 (2016), pages 156–160. ISSN: 23523964. DOI: 10.1016/j.ebiom.2016.02.017.
- [50] José Ignacio Orlando, Huazhu Fu, João Barbossa Breda, Karel van Keer, Deepthi R. Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng Ann Heng, Jeyoung Kim, Joon Ho Lee, Joonseok Lee, Xiaoxiao Li, Peng Liu, Shuai Lu, Balamurali Murgesan, Valery Naranjo, Sai Samarth R. Phaye, Sharath M. Shankaranarayana, Apoorva Sikka, Jaemin Son, Anton van den Hengel, Shujun Wang, Junyan Wu, Zifeng Wu, Guanghui Xu, Yongli Xu, Pengshuai Yin, Fei Li, Xiulan Zhang, Yanwu Xu, and Hrvoje Bogunović. “REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs.” In: *Medical Image Analysis* 59 (2020). ISSN: 13618423. DOI: 10.1016/j.media.2019.101570.
- [51] Arto Palmu, Heikki Puhakka, Tapani Rahko, and Aino K. Takala. “Diagnostic value of tympanometry in infants in clinical practice.” In: *International Journal of Pediatric Otorhinolaryngology* 49.3 (1999), pages 207–213. ISSN: 01655876. DOI: 10.1016/S0165-5876(99)00207-4.
- [52] Freja Rindel Peulicke. *Image-based quality evaluation of otoscopy images*. 2021.
- [53] M. E. Pichichero and M. D. Poole. “Assessing diagnostic accuracy and tympanocentesis skills in the management of otitis media.” In: *Archives of Pediatrics and Adolescent Medicine* 155.10 (2001), pages 1137–1142. ISSN: 10724710. DOI: 10.1001/archpedi.155.10.1137.
- [54] Michael E. Pichichero. “Can machine learning and AI replace otoscopy for diagnosis of otitis media?” In: *Pediatrics* 147.4 (2021). ISSN: 10984275. DOI: 10.1542/peds.2020-049584.
- [55] Ali Razavi, Aäron Van Den Oord, and Oriol Vinyals. “Generating diverse high-fidelity images with VQ-VAE-2.” In: *Advances in Neural Information Processing Systems*. Volume 32. 2019. arXiv: 1906.00446.
- [56] Scott Reed, Zeynep Akata, Xinchun Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. “Generative adversarial text to image synthesis.” In: *International Conference on Machine Learning*. 2016, pages 1060–1069. ISBN: 9781510829008. arXiv: 1605.05396.

- [57] Haim Reuveni, Elad Asher, David Greenberg, Joseph Press, Natalya Bilenko, and Eugene Leibovitz. “Adherence to therapeutic guidelines for acute otitis media in children younger than 2 years.” In: *International Journal of Pediatric Otorhinolaryngology* 70.2 (2006), pages 267–273. ISSN: 01655876. DOI: 10.1016/j.ijporl.2005.06.016.
- [58] Peter J. Robb and Ian Williamson. “Otitis media with effusion in children: Current management.” In: *Paediatrics and Child Health* 26.1 (2016), pages 9–14. ISSN: 1878206X. DOI: 10.1016/j.paed.2015.09.002.
- [59] Florian Schroff, Dmitry Kalenichenko, and James Philbin. “FaceNet: A unified embedding for face recognition and clustering.” In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2015, pages 815–823. ISBN: 9781467369640. arXiv: 1503.03832.
- [60] Scottish Intercollegiate Guidelines Network. “Diagnosis and management of childhood asthma in primary care.” In: *Royal College of Physicians of Edinburgh* (2014).
- [61] Caglar Senaras, Aaron C. Moberly, Theodoros Teknos, Garth Essig, Charles Elmaraghy, Nazhat Taj-Schaal, Lianbo Yu, and Metin Gurcan. “Autoscope: automated otoscopy image analysis to diagnose ear pathology and use of clinically motivated eardrum features.” In: *Medical Imaging 2017: Computer-Aided Diagnosis*. Volume 10134. 2017, pages 500–507. DOI: 10.1117/12.2250592.
- [62] Caglar Senaras, Aaron C. Moberly, Theodoros Teknos, Garth Essig, Charles Elmaraghy, Nazhat Taj-Schaal, Lianbo Yua, and Metin N. Gurcan. “Detection of eardrum abnormalities using ensemble deep learning approaches.” In: *Medical Imaging 2018: Computer-Aided Diagnosis*. Volume 10575. 2018, pages 295–300. DOI: 10.1117/12.2293297.
- [63] Chuen Kai Shie, Hao Ting Chang, Fu Cheng Fan, Chung Jung Chen, Te Yung Fang, and Pa Chun Wang. “A hybrid feature-based segmentation and classification system for the computer aided self-diagnosis of otitis media.” In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2014, pages 4655–4658. DOI: 10.1109/EMBC.2014.6944662.
- [64] Kihyuk Sohn. “Improved deep metric learning with multi-class N-pair loss objective.” In: *Advances in Neural Information Processing Systems*. 2016.
- [65] Michael G. Stewart, Spiros Manolidis, Rhoda Wynn, and Marilyn Bautista. “Practice patterns versus practice guidelines in pediatric otitis media.” In: *Otolaryngology - Head and Neck Surgery* 124.5 (2001), pages 489–495. ISSN: 01945998. DOI: 10.1067/mhn.2001.115497.
- [66] Josefine Vilsbøll Sundgaard, Peter Bray, Søren Laugesen, James Harte, Yosuke Kamide, Chiemi Tanaka, Anders Nymark Christensen, and Rasmus R. Paulsen. “A deep learning approach for detecting otitis media from wideband tympanometry measurements.” In: *IEEE Journal of Biomedical and Health Informatics* (2022). DOI: 10.1109/JBHI.2022.3159263.

- [67] Josefine Vilsbøll Sundgaard, Morten Rieger Hannemose, Søren Laugesen, Peter Bray, James Harte, Yosuke Kamide, Chiemi Tanaka, Rasmus R. Paulsen, and Anders Nymark Christensen. “Multi-modal data generation with a deep metric variational autoencoder.” In: *arXiv preprint arXiv:2202.03434* (2022). arXiv: 2202.03434.
- [68] Josefine Vilsbøll Sundgaard, James Harte, Peter Bray, Søren Laugesen, Yosuke Kamide, Chiemi Tanaka, Rasmus R. Paulsen, and Anders Nymark Christensen. “Deep metric learning for otitis media classification.” In: *Medical Image Analysis* 71 (2021). ISSN: 13618423. DOI: 10.1016/j.media.2021.102034.
- [69] Josefine Vilsbøll Sundgaard, Kristine Aavild Juhl, Klaus Fuglsang Kofoed, and Rasmus Reinhold Paulsen. “Multi-planar whole heart segmentation of 3D CT images using 2D spatial propagation CNN.” In: *Medical Imaging 2020: Image Processing*. 2020, pages 477–484. ISBN: 9781510633933. DOI: 10.1117/12.2548015.
- [70] Josefine Vilsbøll Sundgaard, Kristine Aavild Juhl, and Jakob Mølkjær Slipsager. “EyeLoveGAN: Exploiting domain-shifts to boost network learning with cycleGANs.” In: *arXiv preprint arXiv:2203.05344* (2022). arXiv: 2203.05344.
- [71] Josefine Vilsbøll Sundgaard, Maria Värendh, Franziska Nordström, Yosuke Kamide, Chiemi Tanaka, James Harte, Rasmus R. Paulsen, Anders Nymark Christensen, Peter Bray, and Søren Laugesen. “Inter-rater reliability of the diagnosis of otitis media based on otoscopic images and wideband tympanometry measurements.” In: *International Journal of Pediatric Otorhinolaryngology* 153 (2022), page 111034. ISSN: 18728464. DOI: 10.1016/j.ijporl.2021.111034.
- [72] Frederick T. Searight, Rahul Kumar Singh, and Diana C. Peterson. *Otitis Media With Effusion*. StatPearls Publishing LLC, 2019.
- [73] Sharon Ovnat Tamir, Andres Sibbald, Vedantam Rupa, Paola Marchisio, Preben Homøe, Sam J. Daniel, Frida Enoksson, and Tal Marom. “Guidelines for the Treatment of Acute Otitis Media: Why Are There Worldwide Differences?” In: *Current Otorhinolaryngology Reports* 5.2 (2017), pages 101–107. DOI: 10.1007/s40136-017-0149-1.
- [74] S. Terzi, A. Özgür, Erdivanli, Z. Coşkun, M. Ogurlu, M. Demirci, and E. Durşun. “Diagnostic value of the wideband acoustic absorbance test in middle-ear effusion.” In: *Journal of Laryngology and Otology* 129.11 (2015), pages 1078–1084. ISSN: 17485460. DOI: 10.1017/S0022215115002339.
- [75] Thi Thao Tran, Te Yung Fang, Van Truong Pham, Chen Lin, Pa Chun Wang, and Men Tzung Lo. “Development of an automatic diagnostic algorithm for pediatric otitis media.” In: *Otology and Neurotology* 39.8 (2018), pages 1060–1065. ISSN: 15374505. DOI: 10.1097/MAO.0000000000001897.
- [76] Laurens Van Der Maaten and Geoffrey Hinton. “Visualizing data using t-SNE.” In: *Journal of Machine Learning Research* 9.11 (2008). ISSN: 15324435.

- [77] Louis Vernacchio, Richard M. Vezina, and Allen A. Mitchell. “Knowledge and practices relating to the 2004 acute otitis media clinical practice guideline: A survey of practicing physicians.” In: *Pediatric Infectious Disease Journal* 25.5 (2006), pages 385–389. ISSN: 08913668. DOI: 10.1097/01.inf.0000214961.90326.d0.
- [78] Louis Vernacchio, Richard M. Vezina, and Allen A. Mitchell. “Management of acute otitis media by primary care physicians: Trends since the release of the 2004 American Academy of Pediatrics/American Academy of Family Physicians clinical practice guideline.” In: *Pediatrics* 120.2 (2007), pages 281–287. ISSN: 00314005. DOI: 10.1542/peds.2006-3601.
- [79] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. “Multi-similarity loss with general pair weighting for deep metric learning.” In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2019, pages 5017–5025. ISBN: 9781728132938. DOI: 10.1109/CVPR.2019.00516.
- [80] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. “Image quality assessment: from error visibility to structural similarity.” In: *IEEE Transactions on Image Processing* 13.4 (2004), pages 600–612. DOI: 10.1109/TIP.2003.819861.
- [81] T. Wollmann, C. S. Eijkman, and K. Rohr. “Adversarial domain adaptation to improve automatic breast cancer grading in lymph nodes.” In: *International Symposium on Biomedical Imaging*. IEEE, 2018, pages 582–585. ISBN: 9781538636367. DOI: 10.1109/ISBI.2018.8363643.
- [82] Graham Worrall. “ARI Series Acute otitis media.” In: *Canadian Family Physician* (2007).
- [83] Zebin Wu, Zheqi Lin, Lan Li, Hongguang Pan, Guowei Chen, Yuqing Fu, and Qianhui Qiu. “Deep Learning for Classification of Pediatric Otitis Media.” In: *Laryngoscope* 131.7 (2021), E2344–E2351. ISSN: 15314995. DOI: 10.1002/lary.29302.
- [84] Weidi Xu, Haoze Sun, Chao Deng, and Ying Tan. “Variational autoencoder for semi-supervised text classification.” In: *AAAI Conference on Artificial Intelligence*. 2017. arXiv: 1603.02514.
- [85] Liu Yang, Rong Jin, Lily Mummert, Rahul Sukthankar, Adam Goode, Bin Zheng, Steven C.H. Hoi, and Mahadev Satyanarayanan. “A boosting framework for visuality-preserving distance metric learning and its application to medical image retrieval.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.1 (2010), pages 30–44. ISSN: 01628828. DOI: 10.1109/TPAMI.2008.273.
- [86] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. “Deep metric learning for person re-identification.” In: *International Conference on Pattern Recognition*. IEEE, 2014, pages 34–39. ISBN: 9781479952083. DOI: 10.1109/ICPR.2014.16.



- 
- [87] Jun Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks.” In: *IEEE International Conference on Computer Vision*. 2017, pages 242–251. ISBN: 9781538610329. DOI: 10.1109/ICCV.2017.244.

CONTRIBUTION **A**

# Deep metric learning for otitis media classification

---

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

**Journal** Medical Image Analysis, vol 71, 102034, 2021

**Status** Published

**DOI** [10.1016/j.media.2021.102034](https://doi.org/10.1016/j.media.2021.102034)



Contents lists available at ScienceDirect

# Medical Image Analysis

journal homepage: [www.elsevier.com/locate/media](http://www.elsevier.com/locate/media)

## Deep metric learning for otitis media classification



Josefine Vilsbøll Sundgaard<sup>a,\*</sup>, James Harte<sup>b</sup>, Peter Bray<sup>c</sup>, Søren Laugesen<sup>b</sup>,  
Yosuke Kamide<sup>d</sup>, Chiemi Tanaka<sup>e</sup>, Rasmus R. Paulsen<sup>a,1</sup>, Anders Nymark Christensen<sup>a,1</sup>

<sup>a</sup> Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby, Denmark<sup>b</sup> Interacoustics Research Unit, c/o Technical University of Denmark, Lyngby, Denmark<sup>c</sup> DGS Diagnostics, Snørum, Denmark<sup>d</sup> Kamide ENT clinic, Shizuoka, Japan<sup>e</sup> Demant Japan K.K., Kanagawa, Japan

### ARTICLE INFO

#### Article history:

Received 9 June 2020

Revised 22 February 2021

Accepted 8 March 2021

Available online 14 March 2021

#### MSC:

68T07

68U10

92C55

94A08

#### Keywords:

Otitis media

Deep metric learning

Convolutional neural network

Image classification

### ABSTRACT

In this study, we propose an automatic diagnostic algorithm for detecting otitis media based on otoscopy images of the tympanic membrane. A total of 1336 images were assessed by a medical specialist into three diagnostic groups: acute otitis media, otitis media with effusion, and no effusion. To provide proper treatment and care and limit the use of unnecessary antibiotics, it is crucial to correctly detect tympanic membrane abnormalities, and to distinguish between acute otitis media and otitis media with effusion. The proposed approach for this classification task is based on deep metric learning, and this study compares the performance of different distance-based metric loss functions. Contrastive loss, triplet loss and multi-class N-pair loss are employed, and compared with the performance of standard cross-entropy and class-weighted cross-entropy classification networks. Triplet loss achieves high precision on a highly imbalanced data set, and the deep metric methods provide useful insight into the decision making of a neural network. The results are comparable to the best clinical experts and paves the way for more accurate and operator-independent diagnosis of otitis media.

© 2021 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)

### 1. Introduction

Otitis media is a group of diseases in the middle ear, which can be divided into two major diagnostic groups: acute otitis media (AOM) and otitis media with effusion (OME). Each year, around 11% of the world's population suffer from AOM (Monasta et al., 2012), and it is the second most common reason for a visit to the doctor (Worrall, 2007). Acute otitis media is an acute middle-ear infection with a rapid onset, characterized by a bulging and red eardrum, due to a pus-filled middle-ear cavity, with a clear indication of inflammation, as shown in Fig. 1(a). Symptoms include fever, otalgia, otorrhea, vomiting, and diarrhea. The disease is usually treated with antibiotics, and it is the single diagnosis responsible for most prescriptions of antibiotics (Worrall, 2007), even though 'watch-and-wait' is advised by many clinical guidelines to limit the overuse of antibiotics.

Otitis media with effusion is the most common cause of acquired hearing loss in childhood (Robb and Williamson, 2016) and

80% of all children younger than 4 years old have had at least one episode of the disease. An example of an eardrum with OME is shown in Fig. 1(b), which shows a build-up of fluid in the middle ear and a retracted and opaque tympanic membrane. Signs and symptoms of OME vary greatly and change in intensity, but often include hearing difficulties, loss of balance, and delayed speech development. Otitis media with effusion does not cause pain, fever or malaise, and is therefore more difficult to detect and diagnose. The effusion is not an infection, and should therefore not be treated with antibiotics. The condition is self-limiting, and in persistent cases a tube can be inserted to drain the fluid. For comparison, Fig. 1(c) shows a healthy eardrum with no effusion (NOE).

Otitis media is mostly diagnosed with the use of an otoscope, which is a small handheld medical device with a light source and a magnifying lens, allowing the general practitioner (GP) to get a visual impression of the tympanic membrane. Otolaryngologists/Ear-Nose-Throat specialists (ENTs) usually use an endoscope or microscope to diagnose otitis media, as they are trained to use more advanced and specialized tools. Modern otoscopes and endoscopes are equipped with digital cameras, as shown in Fig. 2, making the images suited for automated enhancements and computer-aided diagnostics. Examples of images captured with an otoscope are

\* Corresponding author.

E-mail address: [josh@dtu.dk](mailto:josh@dtu.dk) (J.V. Sundgaard).<sup>1</sup> Shared senior authorship



Fig. 1. Otoscopy images of tympanic membrane with acute otitis media (a), otitis media with effusion (b), and no effusion (c).

shown in Fig. 1, Table 3, and Fig. 6. The diagnosis is decided by the ENT based on the appearance of the tympanic membrane, medical history, and other signs and symptoms, such as fever or ear pain. To provide proper care and treatment, doctors must be able to distinguish between AOM and OME, but it can be challenging for them to do so. In addition, differentiation of AOM from OME has become more critical in the current era that sees rising antibiotic resistance among bacterial pathogens that cause AOM, and therefore a desire to reduce general use of antibiotic drugs (Pichichero, 2000). The rise in drug-resistant bacteria is related to many patients not adhering to a full course of antibiotics and to the high general prevalence of OME and AOM in young children.

Treatment and diagnosis of otitis media is highly debated in the medical literature. Historically, there has been a global tendency to over-prescribe antibiotics in cases where middle-ear effusion is present, even when it is not clear if there is infection (Cullasllarslan et al., 2018). The diagnosis of otitis media is still highly subjective, in spite of the publication of clinical practice guidelines in many countries around the world. Key problems in the diagnostic process include lack of specific training, lack of experience in handling otitis media cases, limited availability of necessary diagnostic tools (Jensen and Lous, 1999; Pichichero and Poole, 2001), and lack of adherence to clinical guidelines, which can be due to physicians' attitude and behaviour concerning guidelines (Céлинд et al., 2014; Flores et al., 2000). Studies have compared diagnostic accuracy across different medical professionals. Pichichero and Poole (2001) compared the diagnostic accuracy of paediatricians with that of ENTs. Paediatricians correctly distinguished between the NOE, OME, and AOM 50% of the time, while the accuracy of the ENTs was 75%. The biggest issue for paediatricians was the fact that they were usually not familiar with the pneumatic otoscope, which is known to increase the diagnostic performance.



Fig. 2. Sketch of an otoscopic examination with a modern otoscope. The image of the tympanic membrane is shown on an external monitor. Image from Intercoustics A/S.

These results indicate the need for ENTs or properly trained primary care physicians to better diagnose otitis media. Jensen and Lous (1999) studied the performance of GPs and found that they were certain about their diagnosis in 67% of new AOM cases regarding children younger than 2 years old. For children over 2 years old, the self-evaluated diagnostic certainty increased to 75%.

A diagnostic support system would be of great value for a GP or pediatrician with limited training in otitis media, in order to streamline the diagnosis and treatment, to ensure adherence to clinical practice guidelines, and to limit the prescription of unnecessary antibiotics. This requires an automatic system that is able to distinguish between AOM, OME, and NOE. Image-based diagnostics based on digital otoscopy images has shown to be a promising approach. Previous approaches have primarily focused on hierarchical rule-based decision trees (Kuruvilla et al., 2013; Myburgh et al., 2016). The features for the decision trees were manually selected, and included colour, bulging, translucency, light, bubbles, presence of malleus, and concavity of the membrane. The decision trees were then manually constructed, mimicking the decision process of an ENT. Other studies are also based on manually selected features, but employ more advanced classification methods, such as neural networks or Adaboost, which outperform decision trees (Shie et al., 2014; Myburgh et al., 2018).

In more recent studies, deep neural networks or other advanced machine learning algorithms have been employed to detect eardrum abnormalities. Tran et al. (2018) performed segmentation of the tympanic membrane, from which relevant features such as colour and shape were extracted. These features were used to classify AOM, OME, and NOE by employing multitask joint sparse representation-based classification. Shie et al. (2015) performed classification of otitis media using hand-crafted features and automatically extracted features from a convolutional neural network. Mironica et al. (2011) evaluated many different machine learning methods for classification of normal and abnormal tympanic membranes, including k-nearest neighbour, decision tree, linear discriminant analysis, naïve Bayes classifier, multi-layer neural network, and support vector machine. Neural network and support vector machine were found to be superior, as also seen in the general trend in the field of machine learning, where deep neural networks are gaining ground in medical image analysis and computer vision in general (Litjens et al., 2017).

Most previous attempts at classification of tympanic membrane diseases have been based either solely on manually extracted features, or a combination of learned and manual features, but in recent years more studies have focused on using deep neural networks for classification. Senaras et al. (2018) employed deep neural networks for both feature extraction and classification, as they utilized an ensemble model of a pre-trained Inception V3 network and a convolutional auto-encoder for the classification of normal

or abnormal eardrum. Similarly, [Binol et al. \(2020\)](#) employed a pre-trained Inception-ResNet-v2 network for otoscopy image classification combined with analysis of tympanometric measurements for the classification of normal or abnormal eardrum. Other studies have focused on other diseases of the tympanic membrane, including [Cha et al. \(2019\)](#), who used an ensemble of convolutional neural networks to classify eardrums into six categories of ear diseases: NOE, OME, perforation, attic retraction, myringitis and EAC tumour. [Xiao et al. \(2019\)](#) employed fine-grained visual classification to classify NOE, secretory otitis media, active chronic suppurative otitis media and static chronic suppurative otitis media. These studies detail the applicability of a broad range of deep neural networks in the analysis of otoscopy images of the tympanic membrane.

The present paper focuses on deep neural networks, as they have not yet been employed for the classification of AOM, OME, and NOE, and since deep neural networks may help distinguish between OME and AOM, which would in turn help ensure proper treatment of patients. This distinction between OME and AOM is, as mentioned earlier, clinically very challenging, since the signs and symptoms vary greatly within each diagnostic group, and no clear diagnostic guidelines are available. Furthermore, the current methods for this classification task employing manual features are time consuming and less effective than newer automatic feature extraction approaches, for example the approaches that use deep neural networks. In this paper, we present a deep neural network approach that aims to eliminate manually selected features and perform the classification of NOE, OME, and AOM automatically by employing advanced deep metric learning methods that have not been utilised before in this field.

Metric learning, or similarity learning, is the overall expression for machine learning approaches based directly on similarities between samples. An example is large margin nearest neighbor, which learns a pseudometric for k-nearest neighbor classification ([Weinberger and Saul, 2009](#)), increasing the distance between samples from different classes and creating dense clusters of same-class samples. As mentioned, deep learning is making an impact in many areas of image analysis and machine learning in general, and metric learning is no exception, with the introduction of deep metric learning. The first attempts at deep metric learning were used for face recognition and person re-identification, as these similarity-based methods hold many advantages when working with only few image examples of each target. This resulted in the presentation of siamese and triplet networks ([Chopra et al., 2005](#); [Schroff et al., 2015](#)). Deep metric learning has also gained ground over the last few years in analysis of images, videos, speech, and text ([Kaya and Bilge, 2019](#)). In deep metric learning, an embedding representation of the input image is computed using a convolutional neural network, and the similarity of different images can be evaluated using these embedding representations. With deep metric learning for medical image analysis and, more specifically, diagnosis detection, it is possible to get an insight into the decision-making of the neural network, and thus get a sense of how widely spread each diagnostic group is. The clusters of the embedding representations provide insight into each diagnostic group, since the centre of the cluster will be the textbook examples of a certain disease, while the examples surrounding the cluster will be variations of this diagnostic group. This can be used to determine clear signs and symptoms for each diagnostic group. The embedding representations can also be used for outlier detection, and sanity checks of the diagnostic decision for each example.

As these methods were developed for face recognition, we believe that deep metric learning is a well-suited approach for our classification task, as the data set is highly unbalanced. Thus, the goal is to capture the variation of the under-represented class as well as the larger classes. In this work, we propose employing deep

metric learning for automatic detection of otitis media in otoscopy images.

The main contribution of this paper is the application of state-of-the-art deep metric learning methods for otitis media classification on a state-of-the-art data set of otoscopy images. Three different distance-based loss functions are evaluated for the task, and compared with the widely used cross-entropy loss and class-weighted cross-entropy loss. This paper investigates the use of deep metric learning (developed for one-shot learning) for the classification task, and shows the advantages of these methods when working with a highly imbalanced data set for disease detection.

## 2. Material and methods

In deep metric learning, the output of the neural network is an embedding representation of the input, instead of a one-hot encoded vector or a soft-max output, as with standard classification networks. These embedding representations are learnt by the network to keep inputs from the same class close together in embedding space, and create a margin between the different classes, thus creating clusters of examples from each class.

A key element in metric learning is the definition of an appropriate loss function, in order to ensure fast convergence and optimise the global minimum search. There are many different suggestions for loss functions, including contrastive loss, triplet loss, and multi-class N-pair loss, which are all based on the Euclidean distance between the training inputs in embedding space. A schematic representation of the loss functions is shown in [Fig. 3](#).

### 2.1. Loss functions

Contrastive loss focuses on either negative or positive pairs for each training iteration. Positive pairs of same-class examples are penalized to move closer together, while negative pairs of two different classes are pushed away from each other, as shown in [Fig. 3a](#). The loss function is a measure of the distance between two embedding vectors, which ideally should be  $y_i = 0$  for positive pairs and  $y_i = 1$  for negative pairs. The loss function is defined as ([Hadsell et al., 2006](#)):

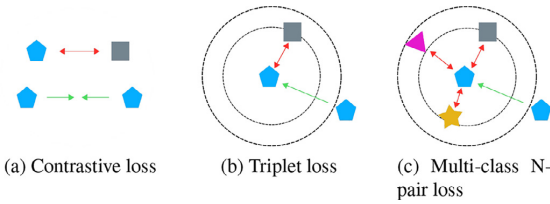
$$L_c(x_{1,i}, x_{2,i}) = \sum_{i=1}^N [(1 - y_i) \|f_{1,i} - f_{2,i}\|_2 + (y_i) \{\max(0, m - \|f_{1,i} - f_{2,i}\|_2)\}^2], \quad (1)$$

where  $x_{1,i}, x_{2,i}$  is the training input from two classes,  $f_{1,i}, f_{2,i}$  represents the embedding vectors generated by the network to each training input,  $N$  is the number of samples, and  $m$  is the margin, usually set to 1.0.

Triplet loss employs three training examples for each iteration. A triplet contains an anchor,  $x^a$ , from which the distances are computed, and a positive sample,  $x^p$  and a negative example,  $x^n$ . This loss function simultaneously penalizes a short distance between an anchor and a negative sample and a long distance between an anchor and a positive sample, and is given as ([Schroff et al., 2015](#)):

$$L_{\text{triplet}}(x_i^a, x_i^p, x_i^n) = \sum_{i=1}^N \max(0, m + \|f_i^a - f_i^p\|_2^2 - \|f_i^a - f_i^n\|_2^2) \quad (2)$$

For triplet loss, the selection of triplets is crucial to improve convergence. Therefore, semi-hard or hard triplets, where the negative sample is closer to the anchor than the positive, are selected, which enforces the network to handle challenging triplet constellations.



**Fig. 3.** Illustration of each loss function. Arrows indicate direction of successful optimization, with red indicating increasing distance between differently labelled samples and green indicating a decreasing distance between same-class samples.

Multi-class N-pair loss is a generalization of triplet loss, which takes into account negative samples from  $j = N - 1$  negative classes in each iteration, instead of only one, as shown in Fig. 3. The loss function reduces the computational cost by optimizing over the distance against all classes in one iteration, and it is given as (Sohn, 2016):

$$L_{m-c}(x_i^a, x_i^p, x_j^p) = \sum_{i=1}^N \log(1 + \sum_{j \neq i}^{N-1} \exp(f_i^a f_j^p - f_i^p f_j^p)) . \quad (3)$$

Besides these three loss functions, classification is also performed using a standard cross-entropy and a class-weighted cross-entropy loss function for comparison.

### 2.2. Network architecture and training details

The network architecture employed for this work is the Inception V3 network (Szegedy et al., 2016) initialized with weights pre-trained on the ImageNet dataset. Other network structures (ResNet and VGG) were also evaluated, but Inception V3 was found superior on this task, and this architecture was also used by both Senaras et al. (2018) and Cha et al. (2019) for otitis media classification. When fine-tuning a pre-trained neural network on a smaller data set, a standard approach is to freeze some of the weights. Experiments with various amounts of frozen weights were conducted, and an optimal setting was found by freezing the first half of the network (first four inception modules and the first grid size reduction). A final linear layer was added to the network, where the output dimensions were set to the desired dimensions of the embedding representation, in this case 32. Classification of test examples was performed using k-nearest neighbor with  $k = 25$  in the embedding space based on the ground truth labels of the training examples. The size of the embedding vector and  $k$  were empirically chosen, and variations of these parameters are explored in Table 2.

The input size for this network architecture is 299x299x3, as the images are RGB images. All networks were trained using the Adam optimizer (Kingma and Ba, 2014), with decreasing learning rate with a factor of 0.1 every eighth epoch. The initial learning rate was set to 0.001 for cross-entropy and contrastive loss, and 0.0001 for triplet and multi-class N-pair loss. The networks were trained using early stopping, and the average number of epochs was 66.0 epochs for cross-entropy loss, 90.8 epochs for contrastive loss, 21.2 epochs for triplet loss, and 79.4 epochs for multi-class N-pair loss. All trained networks had an average training time per epoch around 17 seconds, when trained on an NVIDIA Quadro P5000 16GB GPU.

For each training epoch, balanced mini-batches were created with 30 training examples from each class in each batch. For each iteration in an epoch, the training pairs/triplets were generated for each mini-batch and used for training. For contrastive loss, negative pairs were randomly generated to match the number of positive pairs in the batch. For triplet loss and multi-class N-pair

loss, the pair/triplet generation scheme from the original papers (Sohn, 2016; Schroff et al., 2015) was followed to ensure optimal pair/triplet selection. The approaches were implemented in Pytorch using libraries from Bielski (2018) and Musgrave et al. (2019).

### 2.3. Data

The data used for this study include otoscopy images of the tympanic membrane collected at Kamide ENT clinic, Shizouka, Japan, from patients aged between 2 months and 12 years. The images were captured with an endoscope. The data set consists of 1336 images of both left and right ear from 519 patients, shared between the three diagnostic groups: NOE (658 images), OME (533 images), and AOM (145 images). Diagnosis was decided by an experienced ENT specialist based on signs and symptoms, patient history, otoscopy examination, and, when applicable, wide-band tympanometric measurements (Hein et al., 2017). Furthermore, the ENT graded the severity of OME and AOM as either mild or severe, with the following frequencies: AOM - 76 mild, 69 severe, OME - 274 mild, 259 severe. This grading was not used for classification, but for validation of the results. The data were collected during visits to the clinic, and for 27% of the patients, data were collected for more than one visit (up to five visits). For 74% of individual visits, two images, one of each ear side, were captured. In 20% of visits, only one image was captured, usually because the other ear side was healthy, and for the final 6% of the visits, three to six images were captured, usually to capture different angles of the tympanic membrane if the view was obstructed, for example by earwax. It was ensured that data from one patient was only used for either training or testing, as images captured of the same ear at different times will undoubtedly be very similar.

The original image size was 640x480 pixels, which was cropped to a square to limit the amount of background. Cropping was performed by detecting the outline of the circular image using the circular Hough transform (Yuen et al., 1990), and cropping a square around the detected circles. The images were then downsampled to 299x299, to fit the Inception V3 network structure. Data augmentation was employed in a manner imitating the natural variance of the data set with a certainty of  $p = 0.5$  for each epoch. Horizontal flipping was performed to ensure ear side invariance, together with random erasing (Zhong et al., 2017). This data augmentation method randomly erases one region in the input image with a proportion from 0.02 to 0.33 of the erased area against the input image. The erased region also has various aspect ratios from 0.3 to 3.3. This augmentation method was utilised to force the network to learn features in all areas of the input image.

Due to the limited number of images in the data set, a stratified five-fold cross validation scheme was employed to evaluate each method. The train-test splits were created on a patient level, to ensure that images from one patient were only present in either a training or testing fold. The same train-test splits were used for all methods, which makes the performances directly comparable.

### 3. Results

We evaluate the classification performance of each loss function by computing the accuracy for all classified images, and the recall and precision of each class (AOM, OME, and NOE). The performance measures are computed as the average across the five validation folds, and the standard deviation represents the variation across the five folds, and is shown in Table 1.

The test accuracy is not significantly different for the five loss functions as determined by one-way ANOVA ( $F(4, 20) = 0.94, p = .46$ ), neither is the AOM precision ( $F(4, 20) = 1.56, p = .22$ ). Normal distribution of the residuals were ensured by evaluating the

**Table 1**  
Five-fold cross-validated classification performance (mean  $\pm$  standard deviation) of the neural networks trained with five different loss functions.

	AOM		OME		NOE		Acc. [%]	
	Recall	Precision	Recall	Precision	Recall	Precision	Training	Test
<b>CE</b>	74 $\pm$ 9	67 $\pm$ 10	82 $\pm$ 10	85 $\pm$ 3	89 $\pm$ 3	90 $\pm$ 4	89 $\pm$ 2	85 $\pm$ 2
<b>Class-weighted CE</b>	72 $\pm$ 8	72 $\pm$ 9	84 $\pm$ 4	82 $\pm$ 4	87 $\pm$ 3	89 $\pm$ 2	94 $\pm$ 2	84 $\pm$ 2
<b>Contrastive</b>	50 $\pm$ 9	76 $\pm$ 13	78 $\pm$ 4	84 $\pm$ 5	94 $\pm$ 3	84 $\pm$ 3	99 $\pm$ 0	84 $\pm$ 3
<b>Triplet</b>	61 $\pm$ 8	82 $\pm$ 6	86 $\pm$ 5	84 $\pm$ 3	92 $\pm$ 3	89 $\pm$ 3	98 $\pm$ 1	86 $\pm$ 1
<b>Multi-class</b>	58 $\pm$ 8	74 $\pm$ 8	87 $\pm$ 5	79 $\pm$ 5	87 $\pm$ 3	89 $\pm$ 4	91 $\pm$ 2	84 $\pm$ 3

QQ-plots, thus fulfilling the requisites of the ANOVA test. A one-way ANOVA on the recall of AOM reveals that one or more loss functions are significantly different from the others at a 0.05 significance level ( $F(4, 20) = 6.7651, p = .0013$ ), and a Tukey's post-hoc test shows that contrastive loss recall is significantly lower than that of both cross-entropy ( $p = .003$ ) and class-weighted cross-entropy ( $p = .006$ ), while multi-class loss recall is significantly lower than that of cross-entropy ( $p = .006$ ). This shows, that contrastive and multi-class loss functions perform worse than the standard cross-entropy on this task. There is, however, no significant difference between the performance of triplet loss and either cross-entropy or class-weighted cross entropy. In spite of the fact that the differences among these three loss functions are not statistically significant, Table 1 shows that the precision of AOM does increase from  $67 \pm 10$  by the cross-entropy loss, to  $72 \pm 9$  by the class-weighted cross-entropy loss and then again to  $82 \pm 6$  by the triplet loss. The results show that utilising class-weighted cross-entropy has increased the precision on the under-represented class by 5%, at the expense of a lower AOM recall compared to standard cross-entropy loss, which was expected when introducing class-weights in the loss function, while the rest of the performance measures are very similar to those of the standard cross-entropy measure. Precision and recall are linked, and it is thus often a trade-off between one or the other, as recall usually decreases as precision increases and vice versa, which is also seen in the case of AOM recall and precision for these three loss functions. We will return to this trade-off in the discussion. As triplet loss is the best performing metric for learning loss function, although not significantly better than cross-entropy measures, the rest of the results will be presented for the network trained with the triplet loss function.

The method described by Kuruvilla et al. (2013) was tested on the images from our study. Unfortunately, it was not possible to achieve comparable results to the results reported in the Kuruvilla et al. (2013) paper. The method is based on manual feature selection and a careful selection of hyperparameters, for example, splits in a decision tree. Apparently, the nature of the images in this publication and the images used by Kuruvilla et al. (2013) are of such different quality and nature that the hyperparameter settings found in Kuruvilla et al. (2013) made the approach fail in a majority of the images in our data set.

Test accuracy of variations of the proposed method using triplet loss is shown in Table 2. The proposed method is the neural network trained with triplet loss function, with  $k = 25$ , embedding dimensions 32, with data augmentation and trained with five-fold cross validation, and this table shows the results with variations of these parameters. The classification accuracy is very stable for various values of  $k$  in the range 10–50, and decreases at higher and very low  $k$ -values. The classification accuracy decreases for both halved and doubled embedding dimensions, and Table 2 shows how data augmentation increases the accuracy. The pipeline was also evaluated with 10-fold cross validation, which showed very similar results to those of five-fold cross validation, although the standard deviation increased.

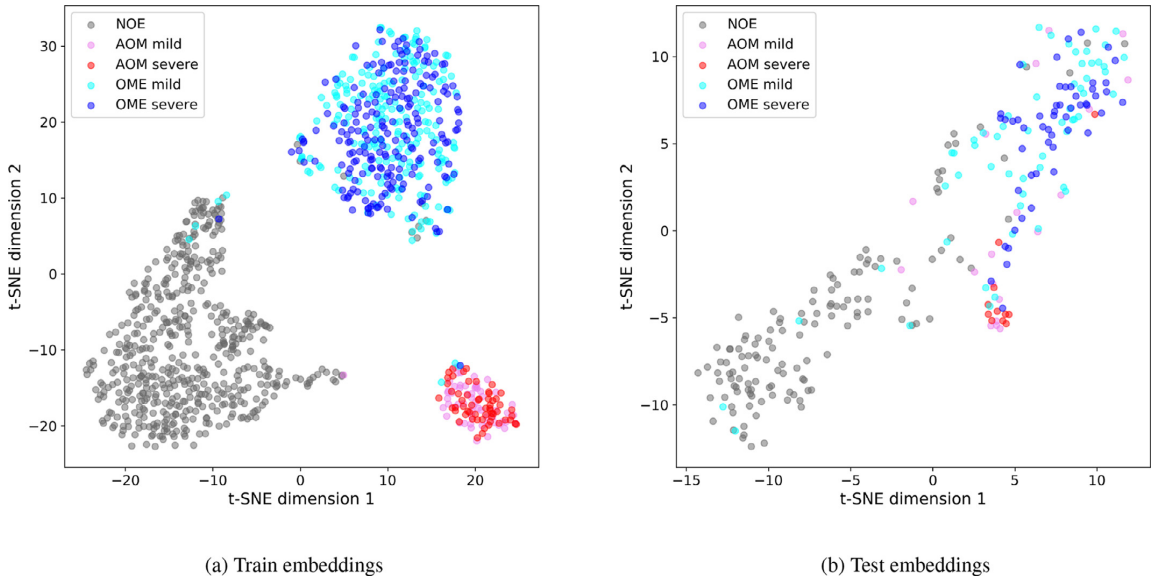
**Table 2**  
Test accuracy of setting and hyperparameter variations of the triplet loss neural network. Proposed approach is the neural network trained with triplet loss function, with  $k = 25$ , embedding dimensions 32, with data augmentation, and trained with five-fold cross validation (CV).

Variations of settings and hyperparameters	Acc. [%]
Proposed approach	86 $\pm$ 1
$k = 10$	86 $\pm$ 1
$k = 55$	85 $\pm$ 1
Embedding dim. = 16	83 $\pm$ 2
Embedding dim. = 64	83 $\pm$ 1
No augmentation	84 $\pm$ 3
10 fold CV	85 $\pm$ 4

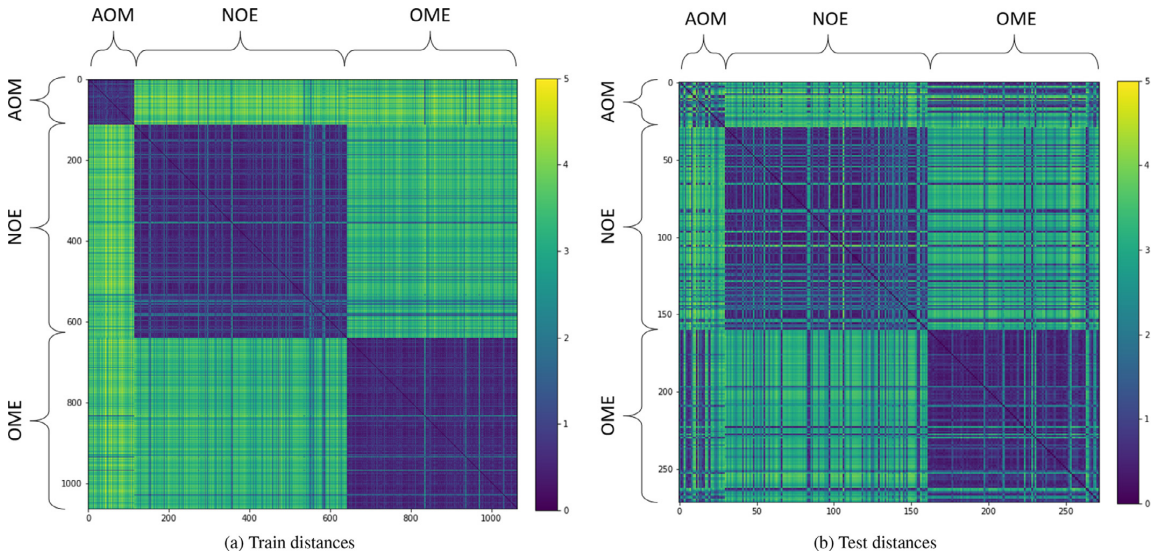
Fig. 4 shows the embeddings created with the triplet loss function for both training data and test data for the fold with precision closest to the overall average precision. As the embeddings are of dimension 32, a t-SNE dimensionality reduction (Van Der Maaten and Hinton, 2008) was performed to obtain this visualization. Each point in the plots represents an otoscopy image of a tympanic membrane. The clusters are not positioned exactly similarly for the train and test 2D plot, due to the nature of the t-SNE reduction. The t-SNE dimensionality reduction is generated separately for the two sets of embeddings, which will create two different mappings from the high dimensional space to 2D. The clusters will therefore be placed similarly in the high dimensional space, but not in completely the same position in this plot. The grading of OME and AOM into mild or severe is plotted as well, to show which cases are most commonly misclassified.

The train embeddings in Fig. 4(a) show clear clustering of the images into the three diagnostic groups, but there are a few outliers of OME images around the NOE and AOM clusters. The test embeddings also show a clear clustering pattern, but with considerably more misclassifications. Here, the clusters blend together in the middle, with no clear boundary between them. In this area, mostly mild AOM and OME mixed with the NOE cases are found, while the severe cases of AOM and OME are primarily kept in the separate clusters. This indicates again that when strong cues are present in the otoscopy image in the severe cases, they are more easily classified.

Fig. 5 shows the pairwise standard Euclidean distance of the 32 dimensional embedding vectors for the same train/test split, as in Fig. 4. Fig. 5(a) shows three clear training-set groups, while there are still images with smaller distance to images in another diagnostic group, especially between NOE and OME. It is clear from this plot that NOE and OME are closest to each other in embedding space, compared to AOM. Fig. 5(b) shows a similar image of the three test-set groups, where specially AOM looks very different. From this figure, it appears that AOM has a few different sub-groups, where one of them appears more like OME. Furthermore, a sub-group of OME images has smaller distance to NOE than any other class.



**Fig. 4.** t-SNE visualizations of train (a) and test (b) embeddings created with triplet loss function. Grey is NOE, pink is mild AOM, red is severe AOM, light blue is mild OME, and dark blue is severe OME.



**Fig. 5.** Pair-wise distance matrix between images in embedding space. Images are grouped by their ground truth label.



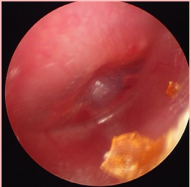
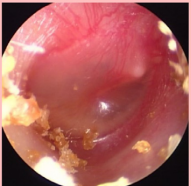


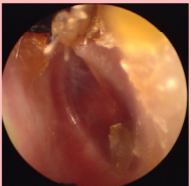
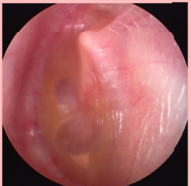

To further investigate how triplet loss manages to classify the otoscopy images, a confusion matrix is shown in Table 3 of the test set from each fold, thus including the full data set. The main errors are false negatives, where AOM- or OME-labelled images are classified as NOE. Furthermore, the neural network does not detect all AOM cases (88 out of 145 are detected), as also seen in the recall performance of AOM, but it does not have a tendency to over-diagnose AOM, as only 3 NOE and 17 OME cases were classified as AOM, as also seen in the high AOM precision. Of the 57 AOM cases that were misclassified as OME or NOE, 44 were diagnosed as mild AOM. Similarly, for the 56 OME cases mis-

classified as NOE, 44 of them were diagnosed as mild OME. This shows, as in Fig. 4, that the severe cases of AOM and OME were classified correctly to a higher degree, and mainly initial stages of mild AOM or OME were misclassified. The table also shows typical image examples for each type of error or correctly classified images. The three correctly classified images are classic examples of: AOM, with a bulging and red membrane; OME, with retracted membrane and visible fluid; and NOE, with translucent membrane with no signs of inflammation. The misclassified images show signs in between these three conditions, and have thus been challenging to diagnose with the algorithm. The example images of



**Table 3**

Confusion matrix for the neural network trained with triplet loss. The first number in each cell shows the number of images for each type of result, and the numbers in the parentheses represent the number of mild and severe cases for each cell (mild/severe). An image example of each type of result is furthermore shown.

Prediction \ Target	AOM	OME	NOE	Total
AOM	88 	17 (10/7) 	3 	108
OME	38 (29/9) 	460 	54 	552
NOE	19 (15/4) 	56 (44/12) 	601 	676
Total	145	533	658	1336

OME and NOE misclassified as AOM are both very red with clear blood vessels around the membrane, and in the OME image, the effusion is visible behind the tympanic membrane. For the AOM image misclassified as OME, the blood vessels are clearly seen, but the inflammation is not as clear. Furthermore, ear wax can challenge the diagnosis, as seen in the NOE image misclassified as OME, where the membrane is not fully visible due to ear wax. The inherent challenges of otoscopy images will be further discussed below.

**4. Discussion**

The results show that otitis media can be classified with a high accuracy with all five loss functions. The data set is highly unbalanced, and some of the loss functions struggle to capture the variance of the under-represented class AOM. The Tukey's pairwise comparison test showed that contrastive and multi-class achieved significantly lower recall on the AOM class. Triplet loss, however, achieves the highest recall among the deep metric methods, and the highest precision over all loss functions on AOM images. As mentioned, precision and recall are interlinked, and it is a trade-off when training a model, as precision will decrease, as recall increases. It is therefore important to optimize the metric most important for the specific application, while keeping a balance between the two. Precision is important, when false positives are expensive, whereas recall is important in cases where false negatives are expensive. In this case, where otitis media diagnosis is considered, the premium is on over-diagnosing AOM, since the problem we want to solve is the over-prescription of antibiotics. Therefore, we want to be very sure that the patients that are diagnosed with

AOM actually have AOM, which is why precision is crucial. It does not cost much to have a false negative, because AOM usually resolves itself after 3–7 days. In persistent cases, the patient will return to their doctor to be checked, probably presenting clearer signs and symptoms that would make AOM detectable. The mistakes made by the neural network are primarily false negatives of mild cases of OME and AOM. The biggest issue in the clinic today is that AOM is over-diagnosed in up to 30% of children, as shown by [Blomgren and Pitkäranta \(2003\)](#), which increases the unnecessary use of antibiotics. In the present study, using deep metric learning with triplet loss had a high precision, and is thus less-likely to over-diagnose AOM compared to the standard cross-entropy loss functions. The higher standard deviation of the AOM class seen in [Table 1](#) is somewhat related to the class imbalance, since this metric is highly susceptible to the specific split of the five cross validation folds. Since the dataset only contains 145 AOM images, and some of the images are very challenging to classify, the standard deviation is dependent on the kind of images in each test set. This is not as big a concern for the larger classes of OME and NOE, where the test set in each fold is much bigger.

Deep metric learning was originally created for face detection, and is therefore designed to classify from only a few images per class. This is very beneficial for the present case, where AOM is under-represented. Triplet loss performs well in this task, and manages to classify each class with above 80% precision, and with the highest test accuracy. Unbalanced data is a very common issue in medical diagnosis classification, as data from one disease class can be challenging to acquire. This is therefore a relevant aspect of the application of deep metric learning in classification tasks. The overall accuracy of otitis media classification with the triplet loss

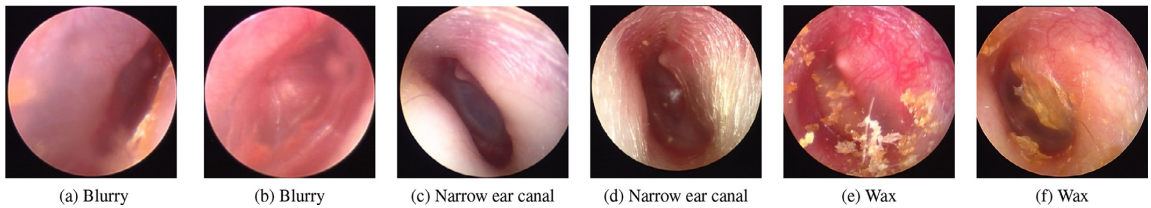


Fig. 6. Examples of quality variations of clinical otoscopy images.

implementation was 86%, which is a satisfying result, when compared to the reported performance of GPs and ENTs, which ranges from 50 to 75% (Pichichero and Poole, 2001). This suggests that an automatic diagnostic support system can improve the performance of otitis media diagnosis in the clinic.

When employing deep metric learning for classification, the pipeline has two steps. First the clusters are generated by the neural network, and then a clustering algorithm classifies based on the generated clusters, as opposed to standard classification networks where the network directly classifies each image. Generally, the precision is increased for the under-represented class when using deep metric learning loss functions, compared to cross-entropy loss functions. This increase in precision is due to the fact that only images located at a certain cluster are classified as belonging to that cluster. Then the model will miss some cases, as seen in the lower recall, because the class is limited to the images located at the cluster centre. Which method would be best will therefore depend on the application. If recall is important, then these results indicate that the cross-entropy loss functions would be a better choice.

It is very challenging to examine otitis media patients, as they are primarily children or babies in pain. Thus, capturing a focused images of the tympanic membrane when a child is moving, screaming, and crying is almost impossible. There are other inherent challenges to acquiring high quality otoscopy images of the tympanic membrane, and some examples are shown in Fig. 6. Fig. 6(a) and (b) show blurry images of the tympanic membrane, where only a few features can be distinguished. Fig. 6(c) and (d) show examples of narrow ear canals, which can make it challenging, and sometimes impossible, to insert the endoscope deep enough into the ear canal, or to get the proper angle, in order to get a high-quality image of the tympanic membrane. Another common problem during ear examinations is ear wax, as shown in Fig. 6(e) and (f). Ear wax can either be found around the ear canal, as in (e), where the ENT sometimes can navigate around it or remove it during the examination, or it can cover the tympanic membrane, as in (f). A high-quality image of the tympanic membrane, with the membrane in focus and with no obstructions or other disruptive elements, is very important to ensure a proper analysis of the image. The images seen in Fig. 6 are, however, realistic images of what would be found in ENT clinics, and they need to be included in the pipeline alongside the high-quality images. The quality variation in otoscopy images currently constitutes a major and unsolved clinical challenge. It can be more challenging to examine children with AOM than OME or NOE patients, as they are generally in more pain. This makes it difficult to get a high quality image of the tympanic membrane, as the child is screaming, crying and moving around. This is clearly visible in our dataset, with more blurry images of AOM cases, and would also account for some of the variation seen in the performance in Table 1.

The data used for this study were assessed and classified by an experienced ENT. Using only one expert opinion in the diagnosis creates a potential bias, since the diagnosis of otitis media

is highly subjective and no objective examination exists. Blomgren and Pitkäranta (2003) found that four medical professionals (a GP, an ENT, and two experienced clinicians) agreed on the diagnosis in 64% of the AOM cases. This uncertainty and lack of objective measurements is a major challenge when working with automatic otitis media diagnosis, and many other medical conditions. It is important to note that the diagnostic decisions for this data set were made by an ENT with many years of experience with otitis media cases, but despite this, we cannot be fully confident in all cases. There might therefore be misdiagnosis in the ground truth data set, which is a common issue in medical image analysis. An improvement of this pipeline would be to perform a human inter-operator study to have a second opinion on each diagnosis from other experienced ENTs, and to be able to evaluate the certainty of the diagnosis of each case. It is a future goal of this research to perform such a study.

## 5. Conclusion

In this work, we demonstrate that it is possible to do automated classification of otitis media, and thus develop a diagnostic tool for detecting acute otitis media, otitis media with effusion, or no effusion. This study compares the performance of five loss functions: cross-entropy, class-weighted cross-entropy, contrastive, triplet and multi-class loss. The results show that the deep metric loss functions achieve a high precision on the under-represented class at the expense of a lower recall. Triplet loss achieved the highest precision on the AOM class without a significant drop in recall, compared to class-weighted cross-entropy loss. Triplet loss has therefore shown good results on this classification task, where the ultimate goal is to reduce the over-prescription of antibiotics by achieving a high precision on the diagnostic predictions. The developed approach shows a high classification accuracy of 85%, thus paving the way for more accurate and operator-independent diagnosis of otitis media.

## Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Søren Laugesen, Pete Bray, James Harte and Chiemi Tanaka works for the Demant Group that develop and manufacture otoscopy equipment.

## CRedit authorship contribution statement

**Josefine Vilsbøll Sundgaard:** Conceptualization, Methodology, Software, Visualization, Writing - original draft. **James Harte:** Conceptualization, Supervision. **Peter Bray:** Conceptualization, Supervision. **Søren Laugesen:** Supervision, Writing - review & editing. **Yosuke Kamide:** Data curation. **Chiemi Tanaka:** Data curation. **Rasmus R. Paulsen:** Conceptualization, Supervision, Validation, Writing - review & editing. **Anders Nymark Christensen:**

Conceptualization, Supervision, Validation, Writing - review & editing.

**Acknowledgments**

We would like to thank William Demant Fonden (Denmark) for financially supporting this study.

**References**

Bielski, A., 2018. Siamese and triplet networks.

Binol, H., Moberly, A.C., Niazi, M.K.K., Essig, G., Shah, J., Elmaraghy, C., Teknos, T., Taj-Schaal, N., Yu, L., Gurcan, M.N., 2020. Decision fusion on image analysis and tympanometry to detect eardrum abnormalities. *Proc. SPIE 11314, Medical Imaging 2020: Computer-Aided Diagnosis (March)* doi:10.1117/12.2549394.

Blomgren, K., Pitkäranta, A., 2003. Is it possible to diagnose acute otitis media accurately in primary health care? *Fam. Pract.* 20 (5), 524–527. doi:10.1093/fampra/cm505.

Célinde, J., Södermark, L., Hjalmarson, O., 2014. Adherence to treatment guidelines for acute otitis media in children: the necessity of an effective strategy of guideline implementation. *Int. J. Pediatr. Otorhinolaryngol.* 78 (7), 1128–1132.

Cha, D., Pae, C., Seong, S.B., Choi, J.Y., Park, H.J., 2019. Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database. *EBioMedicine* 45, 606–614. doi:10.1016/j.ebiom.2019.06.050.

Chopra, S., Hadsell, R., LeCun, Y., 2005. Learning a similarity metric discriminatively, with application to face verification. In: *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005* doi:10.1109/CVPR.2005.202.

Cullas Ilarslan, N.E., Gunay, F., Topcu, S., Ciftci, E., 2018. Evaluation of clinical approaches and physician adherence to guidelines for otitis media with effusion. *Int. J. Pediatr. Otorhinolaryngol.* 112, 97–103.

Flores, G., Lee, M., Bauchner, H., Kastner, B., 2000. Pediatricians' Attitudes, beliefs, and practices regarding clinical practice guidelines: a national survey. *Pediatrics* 105 (3), 496–501.

Hadsell, R., Chopra, S., LeCun, Y., 2006. Dimensionality reduction by learning an invariant mapping. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition doi:10.1109/CVPR.2006.100*.

Hein, T.A.D., Hatzopoulos, S., Skarzynski, P.H., Colella-Santos, M.F., 2017. Wideband Tympanometry. In: *Advances in Clinical Audiology* doi:10.5772/67155.

Jensen, P.M., Lous, J., 1999. Criteria, performance and diagnostic problems in diagnosing acute otitis media. *Fam. Pract.* 16 (3), 262–268.

Kaya, M., Bilge, H.S., 2019. Deep metric learning: a survey. *Symmetry (Basel)* 11 (9).

Kingma, D.P., Ba, J.L., 2014. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kuruviilla, A., Shaikh, N., Hoberman, A., Kovačević, J., 2013. Automated diagnosis of otitis media: vocabulary and grammar. *Int. J. Biomed. Imaging*.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. *10.1016/j.media.2017.07.005*

Mironica, I., Vertan, C., Gheorghe, D.C., 2011. Automatic pediatric otitis detection by classification of global image features. *2011 E-Health and Bioengineering Conference, EHB 2011* 1–4.

Monasta, L., Ronfani, L., Marchetti, F., Montico, M., Brumatti, L., Bavcar, A., Grasso, D., Barbiero, C., Tamburlini, G., 2012. Burden of disease caused by otitis media: systematic review and global estimates. *PLoS ONE* 7 (4).

Musgrave, K., Lim, S.-N., Belongie, S., 2019. *PyTorch Metric Learning*.

Myburgh, H.C., Jose, S., Swanepoel, D.W., Laurent, C., 2018. Towards low cost automated smartphone- and cloud-based otitis media diagnosis. *Biomed. Signal Process. Control* 39, 34–52.

Myburgh, H.C., van Zijl, W.H., Swanepoel, D.W., Hellström, S., Laurent, C., 2016. Otitis media diagnosis for developing countries using tympanic membrane image-Analysis. *EBioMedicine* 5, 156–160.

Pichichero, M.E., 2000. Acute otitis media: part II, treatment in an era of increasing antibiotic resistance. *Am. Fam. Physician* 61 (8), 2410.

Pichichero, M.E., Poole, M.D., 2001. Assessing diagnostic accuracy and tympanocentesis skills in the management of otitis media. *Archives of Pediatrics and Adolescent Medicine* 155 (10), 1137–1142.

Robb, P.J., Williamson, I., 2016. Otitis media with effusion in children: current management. *Paediatr. Child Health (Oxford)* 26 (1), 9–14.

Schroff, F., Kalenichenko, D., Philbin, J., 2015. FaceNet: A unified embedding for face recognition and clustering. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 815–823.

Senaras, C., Moberly, A.C., Teknos, T., Essig, G., Elmaraghy, C., Taj-Schaal, N., Yua, L., Gurcan, M.N., 2018. Detection of eardrum abnormalities using ensemble deep learning approaches. *Proceedings SPIE, Medical Imaging 2018: Computer-Aided Diagnosis* 10575.

Shie, C.K., Chang, H.T., Fan, F.C., Chen, C.J., Fang, T.Y., Wang, P.C., 2014. A hybrid feature-based segmentation and classification system for the computer aided self-diagnosis of otitis media. *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 4655–4658.

Shie, C.K., Chuang, C.H., Chou, C.N., Wu, M.H., Chang, E.Y., 2015. Transfer representation learning for medical image analysis. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 711–714.

Sohn, K., 2016. Improved deep metric learning with multi-class N-pair loss objective. *Adv. Neural Inf. Process. Syst.* 1857–1865.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the Inception Architecture for Computer Vision. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826.

Tran, T.T., Fang, T.Y., Pham, V.T., Lin, C., Wang, P.C., Lo, M.T., 2018. Development of an automatic diagnostic algorithm for pediatric otitis media. *Otology and Neurotology* 39 (8), 1060–1065.

Van Der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*.

Weinberger, K.Q., Saul, L.K., 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research* doi:10.1145/1577069.1577078.

Worrall, G., 2007. ARI Series acute otitis media. *Canadian Family Physician*.

Xiao, L., Yu, J.G., Ou, J., Liu, Z., 2019. Fine-Grained Classification of Endoscopic Tympanic Membrane Images. In: *Proceedings - International Conference on Image Processing, ICIP* doi:10.1109/ICIP.2019.8802995.

Yuen, H., Princen, J., Illingworth, J., Kittler, J., 1990. Comparative study of hough transform methods for circle finding. *Image Vis. Comput.* doi:10.1016/0262-8856(90)90059-E.

Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y., 2017. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*.

CONTRIBUTION **B**

# A deep learning approach for detecting otitis media in wideband tympanometry measurements

---

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

**Journal** IEEE Journal of Biomedical and Health Informatics, 2022

**Status** Published early access. The included paper below is the postprint.

**DOI** [10.1109/JBHI.2022.3159263](https://doi.org/10.1109/JBHI.2022.3159263)

# 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: Our 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].

\* shared senior authorship

This work was supported by William Demant Foundation.

Josefine Vilsbøll Sundgaard (e-mail: joshi@dtu.dk), Anders Nymark Christensen, and Rasmus R. Paulsen are with Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark.

Peter Bray is with Interacoustics A/S, Middelfart, Denmark.

Søren Laugesen and James Harte are with Interacoustics Research Unit, c/o Technical University of Denmark, Denmark.

Yosuke Kamide is with Kamide ENT clinic, Shizuoka, Japan.

Chiemi Tanaka is with Diatec Japan K.K., Kanagawa, Japan.

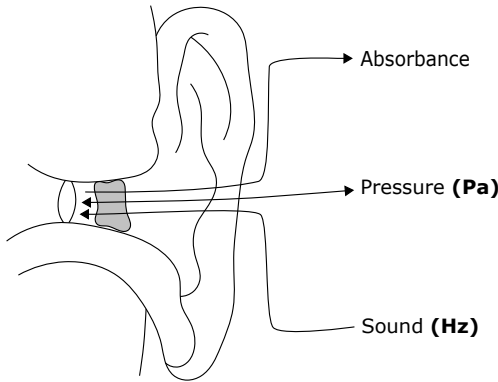
Even though AOM and OME are similar, their clinical classification is important because antibiotics are only recommended for the treatment of AOM, which is caused by infections. Antibiotics are not used to treat OME as it is self-limiting and is not an infection. Diagnosing which type of OM a patient 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.

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 226Hz to 8000Hz) 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].



**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.

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.

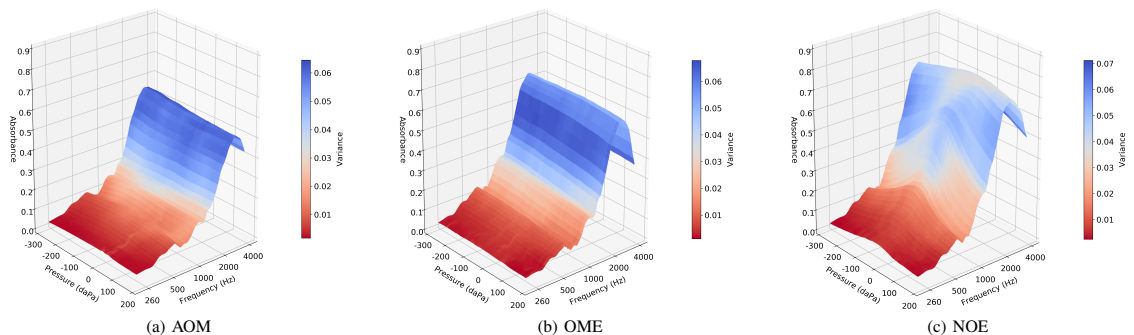
Clinical assessment of OM using WBT could benefit from an automatic diagnostic system designed to assist medical experts when diagnosing patients. As described above, WBT is an objective measurement, and it has been established that it can be successfully used to diagnose OM. Further, its traditional 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



**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.

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 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 (Intera-

coustics, 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.

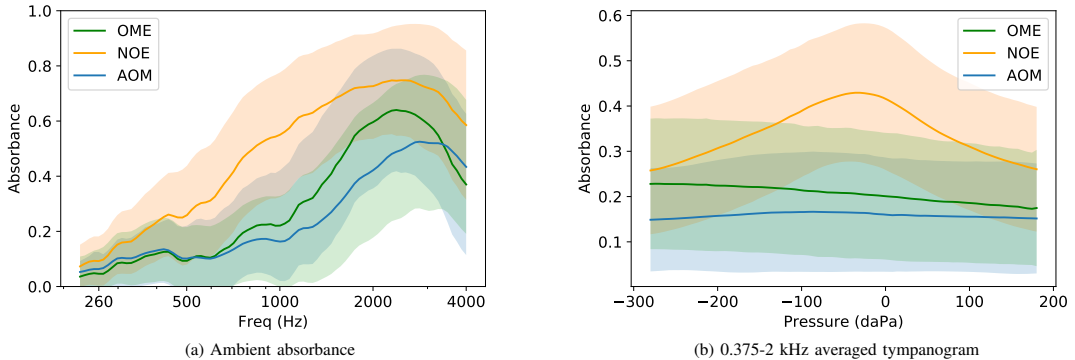


Fig. 3: Average ambient absorbance measurements (a) and 0.375-2 kHz averaged tympanogram (b) of each diagnostic group: OME (green), NOE (orange), and AOM (blue). The faded background curves show the standard deviation of each group.

### 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 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  $84 \times 84$  WBT. Through repeated 2D convolution and max pooling, features are extracted from the WBT, and finally the output of the network 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  $224 \times 224$ , or  $299 \times 299$ , 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

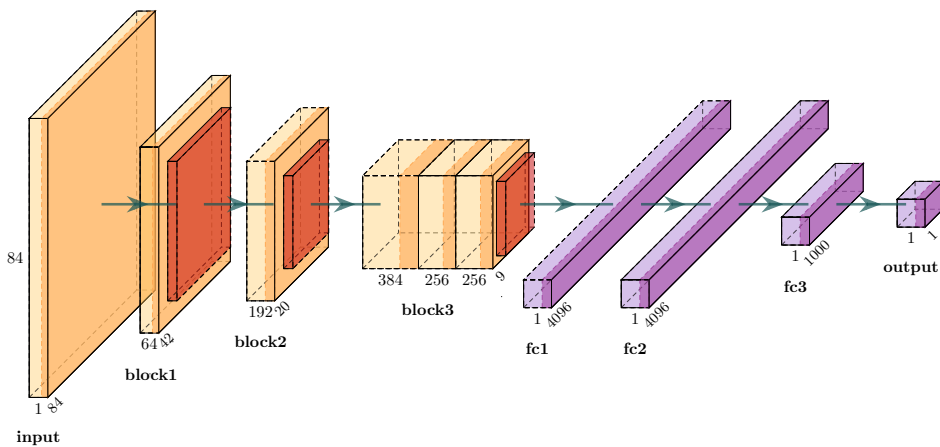
do not require a large-scale network for classification. The input dimensions are much lower ( $84 \times 84$ ), the input only consists of one channel, and the measurements consist of fewer details compared to images, as seen in Fig. 2. It is therefore not feasible, nor necessary, to employ a pre-trained 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<sup>th</sup> 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.

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





**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.

### 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  $9 \times 9$ ). The coarse map is upsampled using bilinear interpolation to obtain a full input size heat map of  $84 \times 84$ .

### 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. 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.

	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	<b>92.2%</b>	87.6%	90.3%
WBT w/o aug.	90.0%	<b>0.97</b>	88.8%	91.2%	90.0%
WBT	<b>92.6%</b>	<b>0.97</b>	<b>92.2%</b>	<b>92.9%</b>	<b>92.6%</b>

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 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

**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%

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.

**TABLE IV:** Performance of multi-class classification (NOE, AOM, and OME). The table shows recall and precision for each class and the overall accuracy.

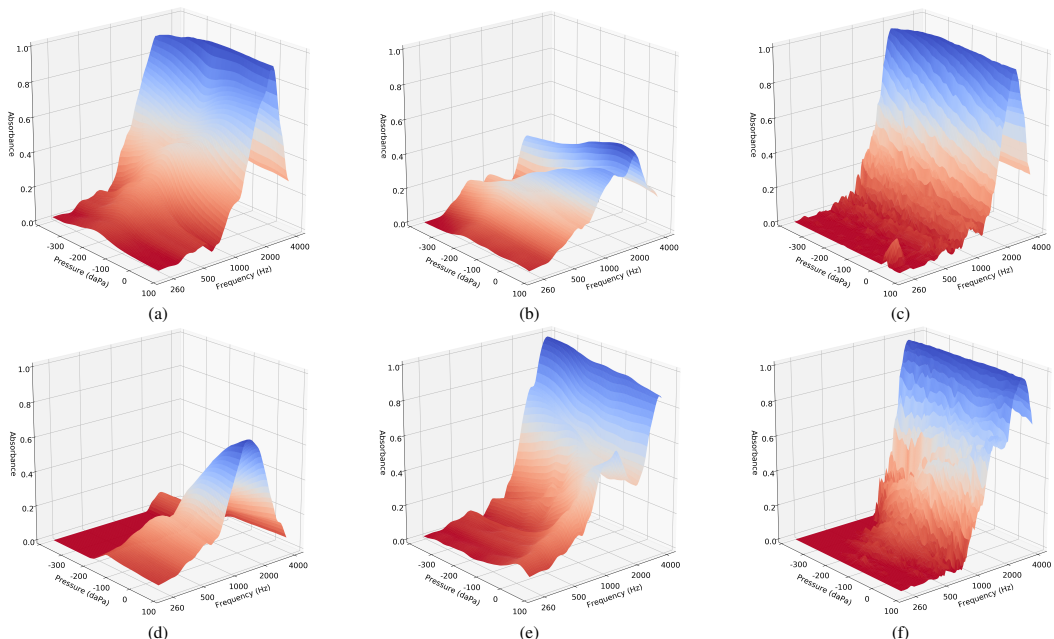
NOE		AOM		OME		Acc.
Recall	Precision	Recall	Precision	Recall	Precision	
90.3%	90.3%	36.4%	52.2%	80.7%	72.0%	79.0%

## V. DISCUSSION

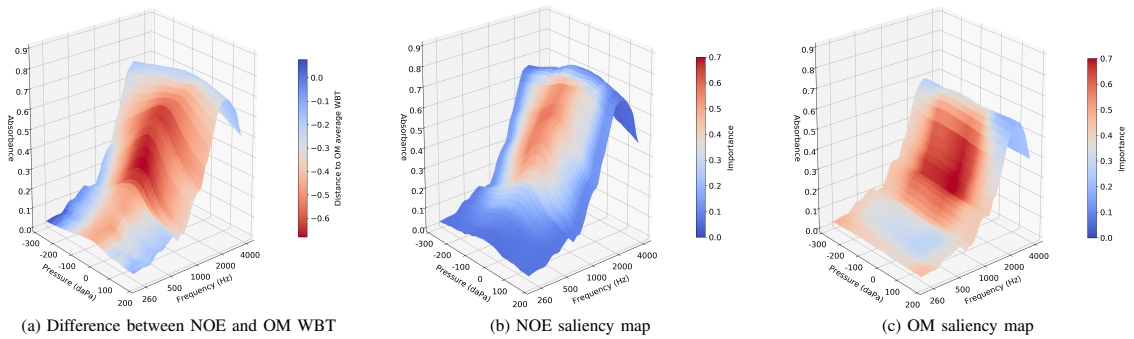
The classification results in Table II show very high performance in all performance metrics for the WBT approach to classifying NOE from OM cases. The averaged tympanogram and ambient absorbance approaches are inferior to WBT, except for sensitivity, where the WBT and averaged tympanogram approaches 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



**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.



**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.

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, "ARI Series Acute otitis media," *Canadian Family Physician*, 2007.
- [2] M. E. Pichichero and M. D. Poole, "Assessing diagnostic accuracy and tympanocentesis skills in the management of otitis media," *Archives of Pediatrics and Adolescent Medicine*, vol. 155, no. 10, pp. 1137–1142, 2001.
- [3] J. V. Sundgaard, J. Harte, P. Bray, S. Laugesen, Y. Kamide, C. Tanaka, R. R. Paulsen, and A. N. Christensen, "Deep metric learning for otitis media classification," *Medical Image Analysis*, vol. 71, 2021.
- [4] T. A. D. Hein, S. Hatzopoulos, P. H. Skarzynski, and M. F. Colella-Santos, "Wideband Tympanometry," in *Advances in Clinical Audiology*. BoD – Books on Demand, 2017.
- [5] A. Biswas and N. Dutta, "Wideband Tympanometry," *Annals of Otolaryngology and Neurology*, vol. 01, no. 02, pp. 126–132, 2018.
- [6] C. A. Sanford, L. L. Hunter, M. Patrick Feeney, and H. H. Nakajima, "Wideband acoustic immittance: Tympanometric measures," *Ear and Hearing*, vol. 34, no. SUPPL. 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 and Hearing*, 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*, 2012.
- [9] D. H. Keefe and J. L. Simmons, "Energy transmittance predicts conductive hearing loss in older children and adults," *The Journal of the Acoustical Society of America*, 2003.
- [10] S. Terzi, A. Özgür, Erdivanli, Z. Coşkun, M. Ogurlu, M. Demirci, and E. Dursun, "Diagnostic value of the wideband acoustic absorbance test in middle-ear effusion," in *Journal of Laryngology and Otolaryngology*, 2015.
- [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," *European Archives of Oto-Rhino-Laryngology*, 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," *International Journal of Pediatric Otorhinolaryngology*, 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," *American Journal of Audiology*, 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," *Journal of the American Academy of Audiology*, 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 & Hearing*, vol. Publish Ah, pp. 1–12, 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 Generation Computer Systems*, 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," *Annals of Translational Medicine*, 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, A. C. Moberly, M. K. K. Niazi, G. Essig, J. Shah, C. Elmaraghy, T. Teknos, N. Taj-Schaal, L. Yu, and M. N. Gurcan, "Decision fusion on image analysis and tympanometry to detect eardrum abnormalities," *Proc. SPIE 11314, Medical Imaging 2020: Computer-Aided Diagnosis*, vol. 11314, no. March, 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," *Scientific Reports*, 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," *The Journal of the Acoustical Society of America*, vol. 146, no. 2, pp. 1350–1361, 2019.
- [22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances In Neural Information Processing Systems*, 2012.
- [23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 770–778, 2016.
- [24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp. 1–14, 2015.
- [25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.
- [26] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," pp. 248–255, 2009.
- [27] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [28] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding Data Augmentation for Classification: When to Warp?" *2016 International Conference on Digital Image Computing: Techniques and Applications, DICTA 2016*, 2016.
- [29] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *arXiv preprint arXiv:1708.04896*, 2017.
- [30] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *Journal of Imaging*, 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," *Proceedings of the IEEE international conference on computer vision*, 2017.
- [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," *Pediatric Infectious Disease Journal*, vol. 31, no. 10, pp. 1003–1006, 2012.
- [33] J. N. Mandrekar, "Receiver operating characteristic curve in diagnostic test assessment," *Journal of Thoracic Oncology*, vol. 5, no. 9, pp. 1315–1316, 2010.
- [34] G. Volandri, F. Di Puccio, P. Forte, and C. Carmignani, "Biomechanics of the tympanic membrane," *Journal of Biomechanics*, vol. 44, no. 7, pp. 1219–1236, 2011.

CONTRIBUTION 

# EyeLoveGAN: Exploiting domain-shifts to boost network learning with cycleGANs

---

**Authors** Josefine Vilsbøll Sundgaard\*, Kristine Aavild Juhl\*, and Jakob Mølkjær Slipsager.

**Journal** arXiv.org

**Status** Published

**Link** <https://arxiv.org/abs/2203.05344>

---

\*Authors contributed equally

# EyeLoveGAN: Exploiting domain-shifts to boost network learning with cycleGANs

Josefine Vilsbøll Sundgaard\*, Kristine Aavild Juhl\*, and Jakob Mølkjær Slipsager

Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby, Denmark

**Abstract.** This paper presents our contribution to the REFUGE challenge 2020. The challenge consisted of three tasks based on a dataset of retinal images: Segmentation of optic disc and cup, classification of glaucoma, and localization of fovea. We propose employing convolutional neural networks for all three tasks. Segmentation is performed using a U-Net, classification is performed by a pre-trained InceptionV3 network, and fovea detection is performed by employing stacked hour-glass for heatmap prediction. The challenge dataset contains images from three different data sources. To enhance performance, cycleGANs were utilized to create a domain-shift between the data sources. These cycleGANs move images across domains, thus creating artificial images which can be used for training.

**Keywords:** Glaucoma detection · cycleGAN · Convolutional neural network

## 1 Introduction

Glaucoma is a group of eye conditions that damage the optic nerve. It is one of the leading causes of irreversible, but preventable, blindness [6], and the incidence is expected to increase. The condition is typically caused by a high pressure in the eye, which damages the optic nerve with no warning signs. The condition of the retina, and thus the optical nerve, is examined using color fundus photography. This imaging technique is both economical and non-invasive. The Retinal Fundus Glaucoma Challenge (REFUGE2) [3] is a competition held as part of the Ophthalmic Medical Image Analysis (OMIA) workshop at the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2020. The goal of this challenge is to provide key tools for diagnosing glaucoma by releasing a large scale database. The challenge consists of three tasks: optic disc and cup segmentation, glaucoma classification, and fovea localization. The problem is challenged by the fact that the data is acquired from three different data sources. The training dataset is acquired with two different cameras, and the test dataset is acquired using a third camera. This

---

\* These authors contributed equally to this work.

paper presents our contribution to this challenge. Segmentation is performed using a U-Net, classification is performed by a pre-trained InceptionV3 network, and fovea detection is performed by employing stacked hour-glass for heatmap prediction of fovea location. To enhance performance and cope with the challenges of different data sources, cycleGANs are utilized to create a domain-shift between the data sources. These cycleGANs move images across domains, thus creating artificial images which can be used for training.

## 2 Data

The approach is trained on the REFUGE challenge data consisting of 1200 annotated images from two different cameras (400 from one, 800 from another). The test dataset consists of 400 images from a third camera. The camera used for the first set of training images (later called domain 1) have the image dimensions 2124x2056, while the dimensions of the other part of the training dataset (later called domain 2) have the dimensions 1634x1634, and the test dataset (later called domain 3) has dimensions 1940x1940. Image examples from all three domains are seen in Figure 1.

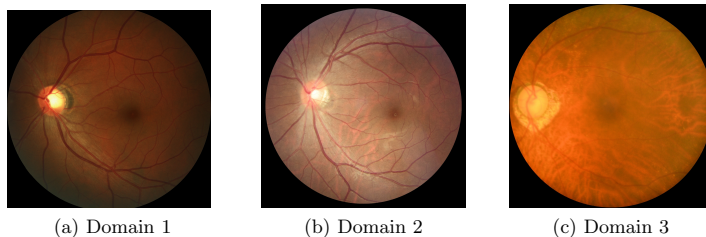


Fig. 1: Image examples from the three different domains in the challenge dataset

For training of the various neural networks for this proposed method, the training data is split into training and validation using a 90/10 split. As the dataset is highly unbalanced with only 10% of glaucoma cases, a class-wise stratified train-validation split was employed. The split was also generated to ensure that 10% of images in each of the two training domains were used for validation.

For classification of glaucoma and optic disc and cup segmentation, a region of interest is cropped out of the original images. The region is detected using a stacked hour-glass neural network for heatmap prediction of the center of the optic disc. This methods is explained in detail in Section 3.1. A region of 500x500 pixels is cropped of the area of the optic nerve head, and this was used as training inputs.



### 3 Methods

As the dataset consist of images from three different domains, we decided to incorporate domain-shift into our approach. The goal is to train an unpaired domain-shift network, in order to create training examples from the domain of the otherwise unlabelled test domain, to generalize the classification and segmentation networks by learning robust features across domains.

The full pipeline is shown in Figure 2. For the first step of ROI detection and fovea localization, we employ the stacked hourglass neural network trained on the original annotated training data from domain 1 and 2. The cropped input images in all three domains are then used to train three cycleGAN’s for domain transfer across all three domains. An example of each image is artificially created in each domain, resulting in an increase in training and test data by a factor of 3. A combination of artificial domain-transferred images and original images are used to train an Inception V3 network for glaucoma classification and a U-net for optic disc and cup segmentation. All the steps of the proposed method will be described in details in the following sections.

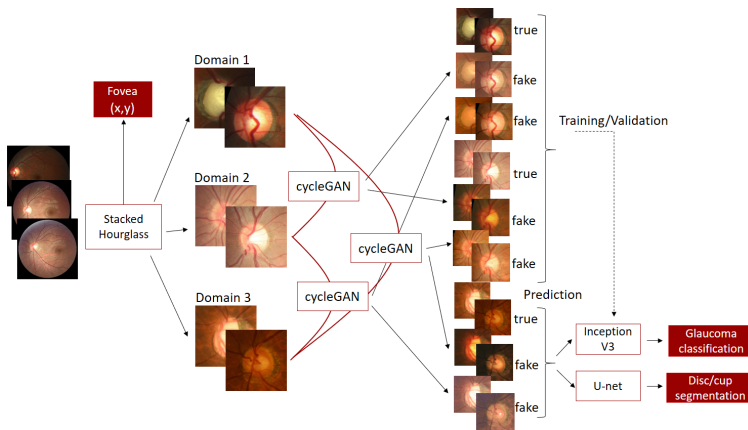


Fig. 2: Schematic representation of proposed method

The proposed method was implemented in Python using the deep learning framework Pytorch, and trained using a GeForce GTX 1070 8GB RAM graphics card.

#### 3.1 ROI detection and fovea localization

Detecting the region-of-interest around the optic nerve head and locating fovea is carried out together using a stacked hourglass network [2] with two stacks.

The network was originally proposed for human pose estimation and is chosen due to its ability to incorporate the interrelationship between the position of the optic nerve head and fovea. This setup allows the network to not only use image features for predicting the position of fovea and the optic nerve head, but also their relative position. In the training images the center of the optic cup is found from the segmentation maps and a heatmap is created as a 2D gaussian placed at the optic cup center with a variance of 100 pixels. Similarly, a heatmap is created with a gaussian at the fovea location. The input image and heatmaps are resized to 256x256 and normalized to the range  $[0, 1]$ . During training the images are augmented using the following transforms: random affine transformations with up to 20 degrees rotation, 100 pixels translation both vertical and horizontal and scaling with up to 20%, color jitter which randomly changes hue ( $\pm 10$ ), saturation ( $[-0.2, 0.5]$ ), and value ( $\pm 0.3$ ), and horizontal and vertical flip. The transformation were applied with a probability of 0.5.

The network was trained using the Adam optimizer [1] with decreasing learning rate with a factor of 0.1 every 50th epoch. The initial learning rate was set to 0.001, and trained with early stopping with a patience of 10 epochs. The network makes use of spatial dropout with a dropout rate of 0.2. The batch size was set to 8 and all input images were shuffled during training.

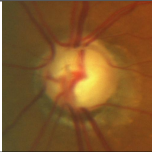
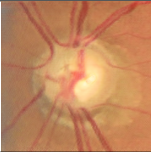
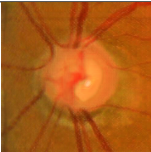
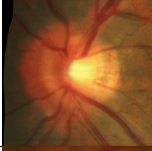

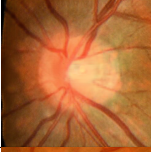



Location of fovea and center of optic cup are determined as pixel with the maximum intensity in each of the predicted heatmaps. The region of interest is constructed as 500x500 pixels cropped around the optic cup center.

### 3.2 Domain shift using cycleGAN

To create the domain shifts between the image domains, several cycleGAN's were trained. The cycleGAN is a image-to-image translation model, where a mapping between an input and an output image is learned without the use of paired examples [7]. The cycleGAN models are trained using a group of images from the source domain and from the target domain, and the model learns to transfer images between these two domains. As seen in Figure 2, three cycleGAN's were needed to transfer between all combinations of the three domains. The standard implementation of cycleGAN was employed, with no modifications from the original implementation from [7]. No data augmentation was used, only the cropped region-of-interest from each of the three domains. The advantage of using cycleGAN is the fact that no labels are needed, as the domain shift is unpaired. The neural network will learn the representation of each domain without other information than the images. This is a big advantage for this application, where we have an unlabelled test dataset from a different domain than the training dataset. The results of domain transfer of three image examples can be seen in Figure 1. The images in the diagonal shows the original images from each of the domains, while the off-diagonal images show artificial images created by the cycleGAN. The figure shows that the cycleGAN creates artificial images that preserves the features of the original image (ie. the vessels and cup/disc size and shape) while incorporating features from the new domain (ie. colors).

The neural networks for classification and segmentation were trained on the fixed training data, consisting of images from both domain 1 and 2. These images were transferred into each of the other domains, e.g. an image in domain 1 is transferred into both domain 2 and 3. The neural networks were therefore trained on image examples from all three domains, although many of them are artificially generated by the cycleGAN. This increases the size of the training dataset by a factor of 3, and ensures that the neural network learns the characteristics of the unlabelled test dataset.

Table 1: Image examples from each of the three domains, and the domain-shifted images

	Domain 1	Domain 2	Domain 3
Image from domain 1			
Image from domain 2			
Image from domain 3			

### 3.3 Glaucoma classification

Classification of glaucoma was performed using the Inception V3 network [5] for binary classification initialized with weights pre-trained on the ImageNet dataset. The weighted cross-entropy loss function was employed to enforce the network to learn to classify the under-represented class, glaucoma. The last layer of the neural network was changed to have the output dimensions two, to match the number of classes. Furthermore, the first half of the network (first four inception modules and the first grid size reduction) was frozen during training, and only the last half of the network was fine-tuned for the glaucoma classification task.

The input size for this network architecture is 299x299x3, as the images are RGB images. Thus the 500x500 input regions were resized to fit the network input. The network was trained using the Adam optimizer [1] with decreasing learning rate with a factor of 0.1 every eighth epoch. The initial learning rate was set to  $10^{-4}$ , and trained with early stopping with a patience of 10 epochs. The batch size was set to 60 and all input images were shuffled during training.

Data augmentation was employed with a variation of transformations: random affine transformations with up to 20 degrees rotation, 60 pixels translation both vertical and horizontal, and scaling with up to 20%, color jitter which randomly changes brightness ( $\pm 10$ ), contrast ( $\pm 10$ ), saturation ( $\pm 10$ ), and hue ( $\pm 10$ ), grey scale transformations, random perspective, and horizontal and vertical flips. The transformations were applied in a random order with a probability of 0.5. Before training, the images were also normalized with the standard parameters for the pre-trained Inception V3 network for Pytorch.

For prediction on the 400 test images, test-time augmentation was employed. In test-time augmentation, the test image is evaluated by the neural network several times with different transformations applied. The same transformations were used as during training, though only one transformation at a time, and each input image was evaluated with 10 different transformations. Besides the test-time augmentation, each image was also evaluated in each of the three domains. The final prediction of glaucoma risk is computed by averaging the output of all 30 prediction (10 transformations in each of the three domains), and applying a softmax activation to obtain the probability for the two output classes.

### 3.4 Optic disc and cup segmentation

Segmentation of the optic disc and cup is carried out using a standard U-net [4]. The input image and heatmaps are resized to 256x256 and normalized to the range [0, 1]. During training the images are augmented using the following transforms: random affine transformations with up to 20 degrees rotation, 60 pixels translation both vertical and horizontal, and scaling with up to 20%, color jitter which randomly changes hue ( $\pm 10$ ), saturation ( $[-0.2, 0.5]$ ), and value ( $\pm 0.3$ ) and horizontal and vertical flip. The transformations were applied with a probability of 0.5.

The network was trained using the Adam optimizer [1] with decreasing learning rate with a factor of 0.1 every 50th epoch. The initial learning rate was set to 0.001, and trained with early stopping with a patience of 10 epochs. The batch size was set to 8 and all input images were shuffled during training.

At prediction time, the image is evaluated 10 times in each of the three domains with different transformations similar to the augmentations used during training. In total the 30 different segmentation proposals are combined by averaging the network outputs before applying a softmax activation to obtain the final segmentation.

## 4 Challenge evaluation

Our contribution was evaluated in the semi-final leaderboard with a overall ranking of 13 out of 22 contributions. For the classification task, our contribution ranked 13th with an AUC of 0.95. Optic disc and cup segmentation resulted in a mean cup dice of 0.86, a mean disc dice of 0.96 and cup-to-disc ratio relative mean error (CDR RME) of 0.04. On the segmentation task, our contribution ranked 11th out of 23 contributions. For fovea localization our contribution ranked 18th with a average euclidian distance of 29.7.

## References

1. Kingma, D.P., Ba, J.L.: Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings (2015)
2. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2016). [https://doi.org/10.1007/978-3-319-46484-82\\_9](https://doi.org/10.1007/978-3-319-46484-82_9)
3. Orlando, J.I., Fu, H., Breda, J.B., van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P.A., Kim, J., Lee, J., et al.: Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis* **59**, 101570 (2020)
4. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2015). [https://doi.org/10.1007/978-3-319-24574-42\\_8](https://doi.org/10.1007/978-3-319-24574-42_8)
5. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the Inception Architecture for Computer Vision. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2016). <https://doi.org/10.1109/CVPR.2016.308>
6. Tham, Y.C., Li, X., Wong, T.Y., Quigley, H.A., Aung, T., Cheng, C.Y.: Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology* **121**(11), 2081–2090 (2014)
7. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In: Proceedings of the IEEE International Conference on Computer Vision (2017). <https://doi.org/10.1109/ICCV.2017.244>

CONTRIBUTION **D**

# Multi-modal data generation with a deep metric variational autoencoder

---

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

**Journal** arXiv.org

**Status** Published

**Link** <https://arxiv.org/abs/2202.03434>

# Multi-modal data generation with a deep metric variational autoencoder

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

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

<sup>2</sup>*Interacoustics Research Unit, c/o Technical University of Denmark, Denmark*

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

<sup>4</sup>*Kamide ENT clinic, Shizuoka, Japan*

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

**We present a deep metric variational autoencoder for multi-modal data generation. The variational autoencoder employs triplet loss in the latent space, which allows for conditional data generation by sampling in the latent space within each class cluster. The approach is evaluated on a multi-modal dataset consisting of otoscopy images of the tympanic membrane with corresponding wideband tympanometry measurements. The modalities in this dataset are correlated, as they represent different aspects of the state of the middle ear, but they do not present a direct pixel-to-pixel correlation. The approach shows promising results for the conditional generation of pairs of images and tympanograms, and will allow for efficient data augmentation of data from multi-modal sources.**

## I. Introduction

Deep generative models can generate new data within the distribution of the training dataset, and can be used for advanced data augmentation in cases where data are costly to annotate or difficult to acquire [1]. A widely used model is the variational autoencoder (VAE). The VAE is a probabilistic model, consisting of an encoder that learns an approximation of the posterior distribution of the data, and a decoder that learns to reconstruct the original input from a latent representation. An advantage of VAEs over generative adversarial networks (GANs) is that the VAE learns a smooth latent representation of the input data. The latent space can therefore be used for sampling new latent representations and thus be used to generate new examples from the distribution of the training dataset.

Conditional data generation, e.g., the conditional VAE [2], allows us to specify which class in the dataset to generate data from. Here, both the latent representations and the input data are conditioned by, e.g., class label. Instead of conditioning the model

for class specific data generation, Karaletsos et al. [3] proposed the triplet-loss based VAE for generation of interpretable latent representations that separate the classes in the latent space with deep metric learning. Karaletsos et al. [3] put their main focus on learning the latent representations, whereas we are interested in using the triplet-loss based VAE for data generation.

We propose a generative approach using a triplet-loss based VAE, and we expand the network architecture and training process to allow for multi-modal data generation. The multi-modal dataset consists of pairs of otoscopy images of the tympanic membrane and wideband tympanometry (WBT) measurements, examples of which are presented in Figure 1. The two types of data are very different, as the first is an image from a camera, and the other is the results of an acoustic measurement. Furthermore, they reflect different aspects of the state of the middle ear. The otoscopy image shows the visual impression of the tympanic membrane, which can show signs of e.g. infection or effusion, while the WBT measurement provides quantitative indications about the presence of fluid in

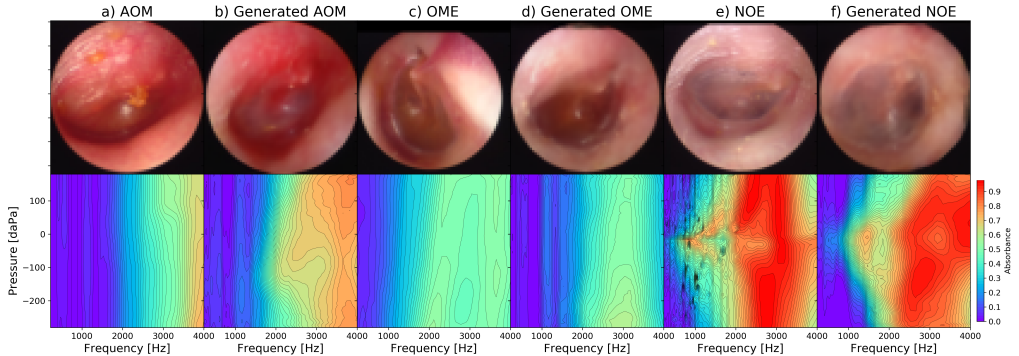


Figure 1. Examples from the dataset and generated examples: otoscopy images (top) and WBT measurements (bottom). Acute otitis media (left two images), otitis media with effusion (middle two images), no effusion (right two images).

the middle ear, the mobility of the tympanic-ossicular system, and the volume of the external auditory canal. The two types of data are therefore correlated but do not have a direct pixel-to-pixel relation, and they reflect two different aspects of the state of the middle ear.

Otitis media can be separated into two main diagnostic groups: acute otitis media (AOM) and otitis media with effusion (OME). Figure 1 shows the difference between these two groups, where AOM is an acute infection with redness and a bulging eardrum, and OME is a build-up of fluid in the middle ear. An example of a normal eardrum with no effusion (NOE) is also shown. The WBT measurements in Figure 1 show how the absorbance across the pressure axis does not change in AOM or OME measurements, whereas the NOE measurements typically show a general increase in absorbance around 0 daPa, compared to negative or positive relative pressures. Furthermore, the general absorbance level at lower frequencies is lower for AOM and OME, than for NOE measurements. These two types of data can both be used for the diagnosis of otitis media. Several studies have developed different approaches for otitis media classification based on either otoscopy images [4–6] or WBT measurements [7, 8]. A combined deep learning classification approach based on standard single-frequency tympanograms and otoscopy images was proposed by Binol et al. [9].

The aim of this paper is to generate new pairs of

otoscopy images and WBTs from each of the three diagnostic groups: AOM, OME, and NOE, and for this task, we propose the multi-modal triplet VAE. The generated otoscopy image and WBT pairs can be used as advanced data augmentation for a multi-modal classification pipeline. Our multi-modal generative model can also be used in other domains such as pairs of cardiac images and electrocardiograms, or brain scans and electroencephalograms. These modalities have a correlation, while reflecting different aspects - visual and functional - of the condition of the examined organ. This work can also be used for the training of doctors and models while preserving patient privacy. Generated data ensures anonymity and allows for data to be shared without regulations such as EU’s GDPR, and some studies have already shown the usability of variational autoencoders in this field [1, 10].

## II. Methods

The multi-modal triplet VAE consists of two encoders and two decoders - one for each modality, and the structure is shown in Figure 2 together with the structure of the upsampling and downsampling blocks used to construct the encoders and decoders. The encoders consist of five residual downsampling blocks using 2D average pooling, and take the  $64 \times 64 \times 3$  otoscopy images and the  $64 \times 64 \times 1$  WBT measurements as input. They start with 64 features in the first block, and double the number of features in each consecutive block. The output feature maps from each encoder



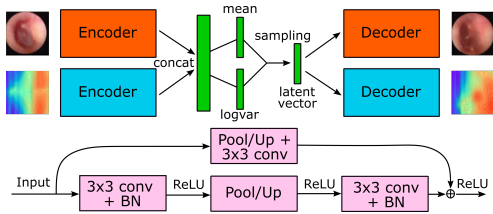


Figure 2. Structure of the multi-modal triplet VAE. Top figure shows the overall structure with two encoders, concatenation of the outputs, sampling, and two decoders. Bottom figure shows the residual blocks used in both encoders and decoders. BN refers to batch normalization.

( $2 \times 2 \times 512$ ) are concatenated, and two  $2 \times 2$  convolutional layers are used to obtain the mean and variance in the 128-dimensional latent space. Using the reparameterization trick [2], a latent vector is sampled, which is passed to both decoders. The decoders consist of six residual upsampling blocks using nearest neighbour upsampling, and the number of features is halved for each block starting at 512. The final layer is a single  $3 \times 3$  convolutional layer going from 32 feature maps to the desired number of channels of the output - one channel for WBTs and three for the otoscopy images. Because the encoder outputs are concatenated, we achieve a common latent space for both modalities, which allows for sampling in the latent space to generate new pairs from each class. The decoders will thus receive information from both image and WBT for the reconstruction of each modality.

The training loss function consists of several parts. The difference between reconstructed WBT and input WBT is penalized using binary cross entropy (BCE) loss. The reconstruction of the image is evaluated using structured similarity index (SSIM) loss [11], which is a local measurement comparing the reconstruction and original image based on luminance, contrast, and structural information. In the latent space, both Kullback–Leibler (KL) divergence and triplet loss [12] are computed. The KL divergence forces the latent embeddings close to a standard normal distribution, while the triplet loss forces examples from the same class to cluster together and pushes examples from different classes further apart [12]. The loss function

terms related to the embedding space are weighted lower than the rest of the terms, and the value 0.1 was experimentally chosen, leading to a loss function defined as:

$$Loss = L_{SSIM} + L_{BCE} + 0.1 \cdot (L_{KL} + L_{triplet}) \quad (1)$$

Balanced sampling is performed during training, with a batch size of 60 (20 pairs from each class) to ensure a balanced representation of every class in each training batch and to cope with the class imbalance in the dataset. The triplets are sampled in each batch using semi-hard mining [12] based on the encoder-generated mean vector from each input pair. The VAE is trained for 5000 epochs using the Adam optimizer [13] with a learning rate of 0.0004. Data augmentation is performed using random erasing [14] on both image and WBT measurement, while horizontal flipping and rotation with  $\pm 20$  degrees is also performed on the images.

Once the network is trained, the test set is passed through the encoders, obtaining the latent representation of each image and WBT pair in the test set. In order to sample new latent vectors for generation of data pairs in each class, the distribution of each class in the latent space is approximated using kernel density estimation for each class. Kernel density estimation estimates the probability density function in the latent space by placing a Gaussian kernel on each sample. The bandwidth of the kernel is fine-tuned using five-fold cross validation. The kernel density estimation is performed only on the test set. When the distribution of each class is estimated, new samples can be generated. The sampled latent vectors are then run through both decoders, to generate new pairs of images and WBTs.

## A. Data

The dataset consists of 1420 pairs of images and WBT measurements collected at Kamide ENT clinic, Shizuoka, Japan, from patients aged between 2 months and 12 years. Each pair was assigned one of the three classes: NOE (537 pairs), OME (419 pairs), and AOM (211 pairs) by an experienced ENT specialist based on signs, symptoms, patient history, otoscopy examination, and WBT measurements. The data was 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 all participants or their parent or guardian.

An otoscopy image is captured using an endoscope (dedicated video otoscope) inserted into the ear canal, allowing a visual inspection of the tympanic membrane. The original image size was  $640 \times 480$  pixels, which was cropped to a square to limit the amount of black background and then downsampled to  $64 \times 64$  to fit the proposed architecture. 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 relative to the ambient pressure from 200 to -300 daPa [15]. The measurements were performed using the Titan system (Interacoustics, Denmark). From the WBT measurement, it is possible to derive conclusions about both tympanic membrane mobility and middle ear condition, and thus additional diagnostic power can be gained over visual inspection alone. WBT measurements were bilinearly resampled to a common grid from 180 daPa to -280 daPa in 64 steps on a linear scale for the pressure axis, and from 226 Hz to 4 kHz in 64 steps for the frequency axis. Examples of both images WBT measurements are shown in Figure 1.

The dataset is split into a train (80%) and test (20%) set. It was ensured that data from one patient was only used for either training or testing, to prevent data leakage.

### III. Results

The test embeddings are shown in Figure 3. The 128-dimensional latent representation of each image has been reduced to two dimensions using t-SNE dimensionality reduction [16] in order to visualize the latent space. The test embeddings clearly show three clusters, but they do blend in the transition areas between the classes, as the images and WBTs can look quite similar across the diagnostic groups. Some of the overlap could also arise from the drastic dimensionality reduction from 128 to two dimensions. The clusters will likely be more separable in the high-dimensional space.

New latent representations are sampled in the full

128-dimensional space within the three class distributions estimated with kernel density estimation, and examples of generated otoscopy images and WBTs are plotted in Figures 4 and 5.

Figure 4 shows examples of generated images in the three diagnostic groups. The images look realistic, as they all contain a tympanic membrane, clear diagnostic markers, and the malleus bone is seen in several examples. The top row of AOM images shows signs of redness and bulging eardrum, and the OME cases clearly have effusion behind the eardrum. The NOE cases appear pale and translucent, as expected.

Other examples of generated pairs of otoscopy images and WBTs are shown side by side with original examples from the dataset in Figure 1. These are not reconstructions, but new generated images. In this figure, it is possible to compare the diagnostic markers of the conditions across modalities, while also comparing the generated examples with original examples. Figure 1(a-b) show similar signs of AOM redness and infection and reduced absorbance in the WBT, which is relatively flat across the pressure axis. The two OME cases in Figure 1(c-d) show very similar diagnostic signs on both the original and generated data with yellow effusion behind the tympanic membrane. Likewise, the absorbance is much lower with very little variation across pressures. The NOE cases in Figure 1(e-f) show normal tympanic membranes and high absorbance in the WBT with a

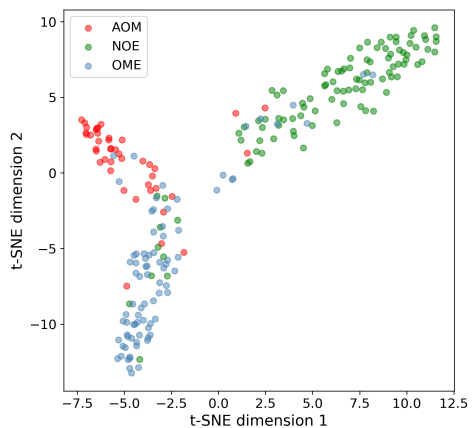


Figure 3. t-SNE visualization of test data latent embeddings.

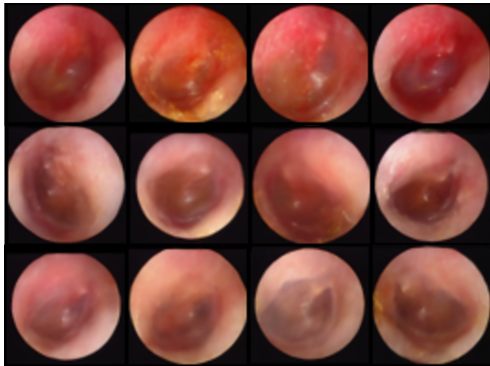


Figure 4. Examples of generated otoscopy images. Top row: AOM, middle row: OME, bottom row: NOE. Best viewed with zoom.

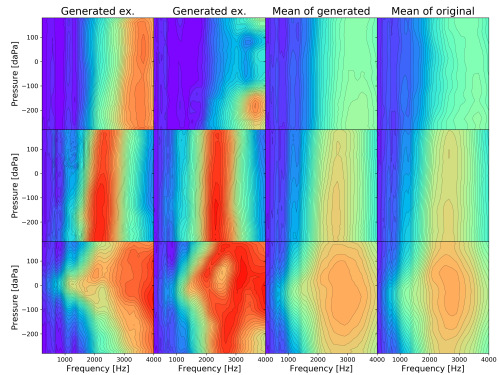


Figure 5. Overview of generated WBT measurements. Top row: AOM, middle row: OME, bottom row: NOE. Best viewed with zoom.

change across pressure.

The generation of WBT measurements is summarized in Figure 5, where generated examples are shown together with the average WBT of the generated samples as well as the original dataset for each of the three diagnostic groups. The average of the generated samples is computed from 500 samples in each diagnostic group. The two average WBT measurements look very similar. This shows that the generated WBT measurements within each diagnostic group follow the same pattern as the mean of the original dataset, thus the distribution of the classes has been captured quite well. The generated examples also indicate great variation within each class.

#### IV. Discussion and Conclusion

The proposed multi-modal triplet-loss based VAE is able to generate highly realistic conditional pairs of otoscopy images and WBT measurements. The generated images examples in Figures 1 and 4 show that the proposed triplet-loss based VAE generates images with a large variation in appearance, and with clear diagnostic markers. The generated images does appear a bit blurry, which is a common VAE problem [17]. The use of SSIM loss [11] has improved the quality of the generated images drastically, compared to employing BCE loss. Other studies have found ways to improve the quality even further, and have thus synthesized high resolution images using

VAEs [18, 19]. However, incorporating this into our approach remains future work. The WBT is a simpler type of data to generate, as it does not contain the same level of detail as an image. BCE loss is therefore sufficient for this modality, and the results in Figures 1 and 5 show that the generated WBTs corresponds very well to the appearance and structure of the original WBTs.

In this study, we propose a VAE structure for conditional multi-modal data generation, even when no direct pixel-to-pixel correlation is present in the two modalities. This multi-modal VAE structure is very flexible, as the encoder and decoder for each modality are completely de-coupled from the other modality. This allows different architectures to be used for each modality depending on the specific needs of the modalities. The employed network architecture for the otoscopy images could be changed to allow for generation of larger and more high-quality images. Likewise, the architecture could be altered to fit temporal data, such as electrocardiograms or electroencephalograms, if this method was to be employed in other domains.

Furthermore, the results show how conditional data generation can be accomplished when employing triplet loss in the latent space of the VAE. This way, conditioning the input or latent space is not needed, as one can simply sample within a certain class cluster.

## V. Acknowledgements

This study was financially supported by William Demant Foundation.

## References

- [1] Shin, H., Tenenholtz, N. A., Rogers, J. K., Schwarz, C. G., Senjem, M. L., Gunter, J. L., Andriole, K. P., and Michalski, M., "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," *International workshop on simulation and synthesis in medical imaging*, Springer, 2018.
- [2] Kingma, D. P., Rezende, D. J., Mohamed, S., and Welling, M., "Semi-supervised learning with deep generative models," *Advances in Neural Information Processing Systems*, Vol. 4, 2014.
- [3] Karaletsos, T., Belongie, S., and Rätsch, G., "Bayesian representation learning with oracle constraints," *4th International Conference on Learning Representations, ICLR*, 2016.
- [4] Senaras, C., Moberly, A. C., Teknos, T., Essig, G., Elmaraghy, C., Taj-Schaal, N., Yua, L., and Gurcan, M. N., "Detection of eardrum abnormalities using ensemble deep learning approaches," *Proceedings SPIE, Medical Imaging 2018: Computer-Aided Diagnosis*, Vol. 10575, 2018.
- [5] Sundgaard, J. V., Harte, J., Bray, P., Laugesen, S., Kamide, Y., Tanaka, C., Paulsen, R. R., and Christensen, A. N., "Deep metric learning for otitis media classification," *Medical Image Analysis*, Vol. 71, 2021. doi: 10.1016/j.media.2021.102034.
- [6] Myburgh, H. C., Jose, S., Swanepoel, D., and Laurent, C., "Towards low cost automated smartphone- and cloud-based otitis media diagnosis," *Biomedical Signal Processing and Control*, Vol. 39, 2018.
- [7] Grais, E. M., Wang, X., Wang, J., Zhao, F., Jiang, W., and Cai, Y., "Analysing wideband absorbance immittance in normal and ears with otitis media with effusion using machine learning," *Scientific Reports*, 2021. doi: 10.1038/s41598-021-89588-4, URL <https://doi.org/10.1038/s41598-021-89588-4>.
- [8] Terzi, S., Özgür, A., Erdivanli, Coşkun, Z., Ogurlu, M., Demirci, M., and Dursun, E., "Diagnostic value of the wideband acoustic absorbance test in middle-ear effusion," *Journal of Laryngology and Otology*, 2015. doi: 10.1017/S0022215115002339.
- [9] Binol, H., Moberly, A. C., Niazi, M. K. K., Essig, G., Shah, J., Elmaraghy, C., Teknos, T., Taj-Schaal, N., Yu, L., and Gurcan, M. N., "Decision fusion on image analysis and tympanometry to detect eardrum abnormalities," *Proc. SPIE 11314, Medical Imaging 2020: Computer-Aided Diagnosis*, Vol. 11314, 2020. doi: 10.1117/12.2549394.
- [10] Li, S., Tai, B., and Huang, Y., "Evaluating variational autoencoder as a private data release mechanism for tabular data," *2019 IEEE 24th Pacific Rim International Symposium on Dependable Computing (PRDC)*, IEEE, 2019.
- [11] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P., "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, Vol. 13, 2004.
- [12] Schroff, F., Kalenichenko, D., and Philbin, J., "FaceNet: A unified embedding for face recognition and clustering," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015.
- [13] Kingma, D. P., and Ba, J. L., "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [14] Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y., "Random Erasing Data Augmentation," *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, 2020. doi: 10.1609/aaai.v34i07.7000.
- [15] Hein, T. A. D., Hatzopoulos, S., Skarzynski, P. H., and Colella-Santos, M. F., "Wideband Tympanometry," *Advances in Clinical Audiology*, BoD – Books on Demand, 2017. doi: 10.5772/671155.
- [16] Van Der Maaten, L., and Hinton, G., "Visualizing data using t-SNE," *Journal of Machine Learning Research*, 2008.
- [17] Dosovitskiy, A., and Brox, T., "Generating images with perceptual similarity metrics based on deep networks," *Advances in Neural Information Processing Systems*, 2016.
- [18] Huang, H., Li, Z., He, R., Sun, Z., and Tan, T., "IntroVAE: Introspective variational autoencoders for photographic image synthesis," *Advances in Neural Information Processing Systems*, , No. NeurIPS, 2018.
- [19] Zhao, S., Song, J., and Ermon, S., "Towards Deeper Understanding of Variational Autoencoding Models," *arXiv preprint arXiv:1702.08658*, 2017. URL <http://arxiv.org/abs/1702.08658>.



CONTRIBUTION **E**

# Inter-rater reliability of the diagnosis of otitis media based on otoscopic images and wideband tympanometry measurements

---

**Authors** Josefine Vilsbøll Sundgaard, Maria Värendh, Franziska Nordström, Yosuke Kamide, Chiemi Tanaka, James Harte, Rasmus R. Paulsen, Anders Nymark Christensen, Peter Bray, and Søren Laugesen.

**Journal** International Journal of Pediatric Otorhinolaryngology, vol 153, 111034, 2022

**Status** Published

**DOI** [10.1016/j.ijporl.2021.111034](https://doi.org/10.1016/j.ijporl.2021.111034)



Contents lists available at ScienceDirect

## International Journal of Pediatric Otorhinolaryngology

journal homepage: [www.elsevier.com/locate/ijporl](http://www.elsevier.com/locate/ijporl)

## Inter-rater reliability of the diagnosis of otitis media based on otoscopic images and wideband tympanometry measurements

Josefine Vilsbøll Sundgaard<sup>a,\*</sup>, Maria Värendh<sup>b</sup>, Franziska Nordström<sup>b</sup>, Yosuke Kamide<sup>c</sup>, Chiemi Tanaka<sup>d</sup>, James Harte<sup>e</sup>, Rasmus R. Paulsen<sup>a</sup>, Anders Nymark Christensen<sup>a</sup>, Peter Bray<sup>f,1</sup>, Søren Laugesen<sup>e,1</sup>

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

<sup>b</sup> Lund University, Skåne University Hospital, Department of Clinical Sciences Lund, Otorhinolaryngology, Head and Neck Surgery, Lund, Sweden

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

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

<sup>e</sup> Interacoustics Research Unit, C/o Technical University of Denmark, Denmark

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

## ARTICLE INFO

## Keywords:

Otitis media  
Inter-rater reliability  
Agreement  
Diagnosis  
Otoscopy  
Wideband tympanometry

## ABSTRACT

**Objectives:** This study aims to investigate the inter-rater reliability and agreement of the diagnosis of otitis media with effusion, acute otitis media, and no effusion cases based on an otoscopy image and in some cases an additional wideband tympanometry measurement of the patient.

**Methods:** 1409 cases were examined and diagnosed by an otolaryngologist in the clinic, and otoscopy examination and wideband tympanometry (WBT) measurement were conducted. Afterwards, four otolaryngologists (Ear, Nose, and Throat doctors, ENTs), who did not perform the acute examination of the patients, evaluated the otoscopy images and WBT measurements results for diagnosis (acute otitis media, otitis media with effusion, or no effusion). They also specified their diagnostic certainty for each case, and reported whether they used the image, wideband tympanometry, or both, for diagnosis.

**Results:** All four ENTs agreed on the diagnosis in 57% of the cases, with a pairwise agreement of 74%, and a Light's Kappa of 0.58. There are, however, large differences in agreement and certainty between the three diagnoses. Acute otitis media yields the highest agreement (77% between all four ENTs) and certainty (0.90), while no effusion shows much lower agreement and certainty (34% and 0.58, respectively). There is a positive correlation between certainty and agreement between the ENTs across all cases, and both certainty and agreement increase for cases where a WBT measurement is shown in addition to the otoscopy image.

**Conclusions:** The inter-rater reliability between four ENTs was high when diagnosing acute otitis media and lower when diagnosing otitis media with effusion. However, WBT can add valuable information to get closer to the ground-truth diagnosis without myringotomy. Furthermore, the diagnostic certainty increases when the WBT is examined together with the otoscopy image.

### 1. Introduction

Otitis media is very common in children, with around 80% of children having at least one episode during their first years of life [1]. The diagnosis of otitis media is challenging because the two main conditions, otitis media with effusion (OME) and acute otitis media (AOM), can appear with various signs and symptoms. Furthermore, performing

specialized examinations of patients requires specific training and tools such as an endoscopic examination of the tympanic membrane, pneumatic otoscope, or tympanometry equipment. It is, however, crucial to diagnose the two conditions correctly, since clinical guidelines only recommend antibiotics for AOM, whereas OME will resolve on its own. AOM is the single diagnosis responsible for most prescriptions of antibiotics [2,3], and there are controversies about prescribing antibiotics in

\* Corresponding author.

E-mail address: [josh@dtu.dk](mailto:josh@dtu.dk) (J.V. Sundgaard).

<sup>1</sup> shared senior authorship.

<https://doi.org/10.1016/j.ijporl.2021.111034>

Received 2 December 2021; Accepted 31 December 2021

Available online 7 January 2022

0165-5876/© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

early acute otitis media. Watchful waiting is considered best practice in most of Europe, and this approach shows no increased incidence of complications. However, watchful waiting has not gained wide acceptance in the United States, where antibiotics is still the most common treatment [1].

Historically, there has been a global tendency to overprescribe antibiotics in cases where middle ear effusion is present, even if it is not clear that there is an infection [4]. The diagnosis of otitis media is still highly subjective, despite the publication of clinical practice guidelines in many countries around the world. Key problems in the diagnostic process include lack of specific training, lack of experience in handling otitis media, limited availability of necessary diagnostic tools [5,6], and lack of adherence to clinical guidelines, which can be due to the attitude and behavior of the physicians concerning guidelines [3,7].

Several studies have investigated the diagnostic process and accuracy of the diagnosis of otitis media by various medical professionals. Jensen et al. [5] assessed the performance of AOM diagnosis of Danish general practitioners (GPs) based on surveys. The study included 368 children with AOM and 151 GPs, and the study found that the GPs' certainty was 67% of new AOM cases regarding children younger than 2 years old. For children over 2 years old, the diagnostic certainty increased to 75%. Key criteria for the diagnosis included symptoms of earache, fever, elevated audiometric threshold, and findings of bulging or red eardrum and purulent otorrhea. These signs and symptoms were present in more than 80% of the AOM cases. This suggests that diagnostic certainty is highly correlated with the visibility of the eardrum and with the use of pneumatic otoscopy.

Pichichero et al. [6] compared diagnoses by 14 pediatricians with those of 188 ENTs by viewing nine videotaped pneumatic otoscopic examinations. This study focused on the distinction between AOM and OME, and found that pediatricians correctly distinguished between normal, OME, and AOM 50% of the time, while the accuracy of the ENTs was 75%. Lack of experience with pneumatic otoscopy was the biggest issue for pediatricians, and more experience would have increased the diagnostic performance. These results indicate the need for ENTs or properly trained primary care physicians to distinguish between AOM or OME.

A similar study was performed by Blomgren *et al.* [8], who compared the diagnosis of 50 children examined by a GP, an expert ENT, and two experienced clinicians. The GP and ENT doctor performed an examination of the patient individually, while the two experts examined images of the tympanic membrane and a tympanogram and made their diagnosis from this information without examining the patient themselves. The four medical professionals agreed on the diagnosis in 64% of the AOM cases, and the ENT was less likely to diagnose AOM compared with the GP (44% compared to 64% of the cases). The two experts agreed on the diagnosis more often when both image and tympanogram were available, compared to only examining the image of the tympanic membrane. This study concluded that the diagnostic accuracy of AOM could possibly be increased if primary care clinics had access to appropriate equipment, such as tympanometry and pneumatic otoscopes, and that proper education is crucial when using these diagnostic tools.

Pichichero [9] further compared the diagnostic accuracy across pediatricians in different countries, including Italy, Greece, South Africa, and the USA. Each pediatrician assessed nine videos of otoscopic examinations of the tympanic membrane, and their ability to distinguish between OME, AOM, and no effusion (NOE) was then evaluated. The correct diagnosis was found by each group of pediatricians with the following frequencies: Italy 54%, Greece 36%, South Africa 53%, and USA 51%, and the frequency of over-diagnosed AOM was: Italy 18%, Greece 34%, South Africa 23%, and USA 26%. These results show how OME is frequently misdiagnosed as AOM.

The great inter-variability in the presented studies shows how challenging it can be to establish the correct diagnosis based only on the opinions of doctors. The ground-truth diagnosis can only be found by

performing myringotomy, where an incision is created in the tympanic membrane to relieve pressure or drain effusion, and then analyzing the content of the middle ear. Since this is not desirable, or ethical, in many cases, it is necessary to rely on the diagnosis of doctors. Since even specialized ENTs identify the condition correctly in only 75% of the cases [6], this is a challenging task.

The presented work is part of a larger study aiming to provide an automated pipeline for otitis media diagnosis using deep learning. However, a ground-truth diagnosis is needed in order to employ supervised learning. As predictive models require a large amount of training data, it is not feasible to use only myringotomy-confirmed cases. The ground-truth diagnosis can therefore only be established based on annotations by ENTs, where it is assumed that using annotations from several ENTs will provide the best possible estimated ground-truth. The automated analysis can be based on otoscopic images [10] and/or wideband tympanometry (WBT) [11]. WBT is a fairly new method of measuring the middle-ear absorbance as a function of both frequency and pressurization of the ear canal [12,13], where the normal tympanogram only measures the absorbance at 226 or 1000 Hz.

This study evaluates agreement among four experienced ENTs when diagnosing NOE, OME, and AOM cases. The aim is to evaluate how well the ENT doctors agree on the diagnosis of otitis media cases, and whether a WBT can add additional information valuable in the diagnosing process. Furthermore, we aim to establish an estimated ground-truth diagnosis for a large number of cases based on annotations by several ENTs.

## 2. Materials and methods

### 2.1. Study design

The study includes 1409 cases collected during the clinical routine at Kamide ENT clinic, Shizouka, Japan, from patients aged between 2 months and 12 years. The data was collected under ethical approval from the non-profit organization MINS Institutional Review Board (reference number 190221), and with either opt-out consent, or informed consent from all participants or their parent or guardian. The otoscopic images were captured with a digital endoscope. Fig. 1 shows examples from each diagnostic group. WBT measurements were performed using the Titan system (Interacoustics, Denmark) in the range from +200 to -300 daPa pressure and from 226 Hz to 4 kHz frequency. WBTs were not measured in patients that reported pain in their ears. The grand averages of the WBT measurements for each of the three diagnostic classes are shown in Fig. 2.

During the clinical routine, an experienced ENT diagnosed each case based on otoscopic image, WBT, signs and symptoms, and patient history. These diagnostic labels were used to create the experimental study design. To decrease the amount of work for the four additional ENTs who annotated the cases, the study design was split into two parts. The first part consisted of 204 cases annotated by all four ENTs in a balanced and complete study design outlined in Table 1. The second part consisted of 1205 cases, each annotated by two ENTs in a balanced incomplete block design. The second part is included as supplementary material in order to support the analysis made on the balanced and complete study of the 204 cases, but it is not a part of the main analysis presented in the paper.

The 204 cases were equally divided among AOM, OME, and NOE, in which half of the cases only included an otoscopic image and the other half included both image and WBT. This even distribution among classification categories was recommended by Mitani et al. [14] in order to employ Kappa coefficients for inter-rater reliability estimation (as discussed later). The 204 cases were randomly selected from a larger pool of patient data from the normal clinical routine. The cases were not selected on a patient basis, rather each ear was selected individually. This was done to ensure the correct distribution across diagnostic groups and with/without WBT. Therefore, the cases were not necessarily from



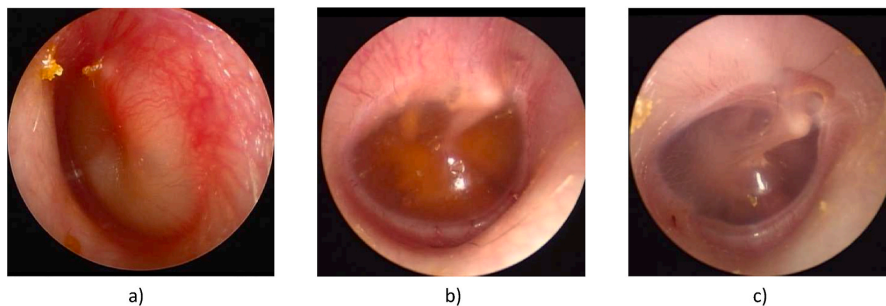


Fig. 1. Otoscopy images of tympanic membrane with acute otitis media (a), otitis media with effusion (b), and no effusion (c).

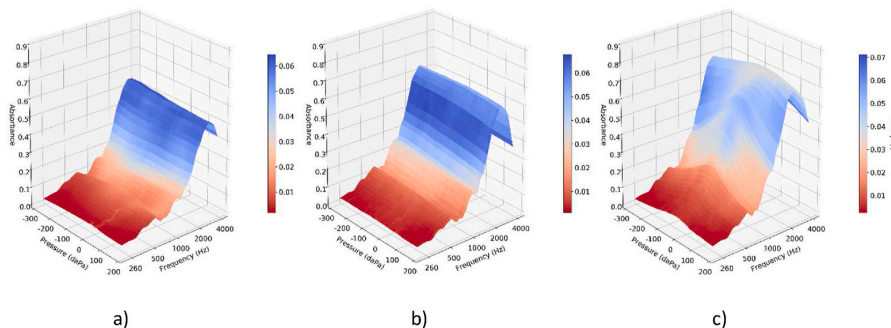


Fig. 2. Grand average WBT of acute otitis media (a), otitis media with effusion (b), and no effusion (c) cases. Color scale shows the variance across the measurements. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

both ears of a patient and can be from both left and right ears.

The four ENTs never performed the physical examination of the patients in the present study, and are not related to the clinic where the data was collected, although they are all experienced with otoscopy and tympanometry for the diagnosis of otitis media. The ENTs have completed training for interpretation of WBTs, and they use the standard tympanogram regularly. The ENTs evaluated all 204 cases, resulting in a fully crossed study. The ENTs were presented with one case at a time and were shown the otoscopy image, the full WBT, the absorbance curve at ambient pressure, and the more familiar standard 226-Hz tympanogram, as shown in Fig. 3. They were asked to determine the diagnosis (AOM, OME, NOE, or Unknown), what data they used to decide the diagnosis of the current case (image, WBT, or both), and finally to perform a self-evaluation of the certainty of the diagnosis (very low, low, medium, moderate, or high), similar to the approach regarding certainty evaluation found in Ref. [5].

2.2. Statistical methods

One of the goals of this study is to determine the most correct diagnosis of each of the cases in the dataset, under the understanding that the ground-truth (determined from myringotomy) is not available.

Table 1

Overview of the balanced and complete study design consisting of 204 cases, divided equally between the three diagnostic groups, and between cases with and without a WBT.

	AOM	OME	NOE	Total
Image and WBT	34	34	34	102
Only image	34	34	34	102
Total	68	68	68	204

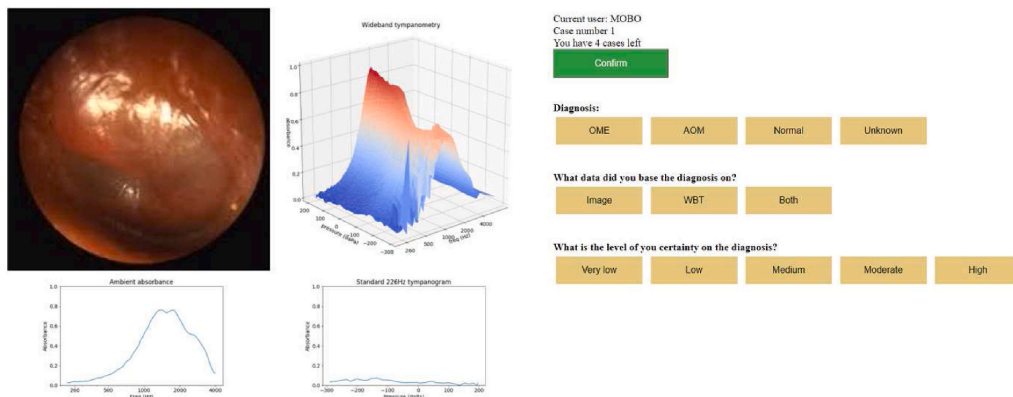
The final diagnosis is determined based on majority voting weighted by the certainty reported by each ENT. Thus, votes with higher reported certainty counted more towards the final diagnosis than votes with low reported certainty. The ENTs were allowed to answer Unknown diagnosis if they could not determine the diagnosis from the otoscopic image and WBT measurement. If a majority of the ENTs reported Unknown diagnosis on a case, the case was removed from the rest of the analysis.

The statistical inter-rater reliability was computed to evaluate agreement across the four ENTs. Light’s Kappa [15] was employed, as the annotation data is categorical and the agreement between several raters was evaluated [16]. Light’s Kappa is computed as the arithmetic mean of the Cohen’s Kappa [17,18] for all rater pairs, which provides an overall metric of agreement. The Kappa coefficient ranges between 0 and 1, and the scale of interpretation is [19]: slight agreement (0–0.2), fair agreement (0.21–0.4), moderate agreement (0.41–0.6), substantial agreement (0.61–0.8), and perfect agreement (0.81–1). The advantage of Kappa over percentage agreement is the ability to account for chance agreement, but it can be more challenging to interpret. As suggested by McHugh [20], both percent agreement and Kappa is reported for this study, as they both have advantages and limitations.

The Mann Whitney U test was used to examine whether the changes in certainty and time spent on evaluation between different groups are statistically significant. The Mann Whitney U test is a non-parametric statistical significance test comparing two independent samples from a population with the same distribution.

3. Results

The weighted majority voting results for all cases are presented in Table 2 as a confusion matrix relative to the original annotations used to set up the study. This table shows that the OME group is now much



**Fig. 3.** Interface of the annotation system, which filled the entirety of a standard 17-inch computer screen monitor. To the left, the otoscopy image, WBT, ambient absorbance, and standard 226-Hz tympanogram are shown. To the right, the questions for the ENT are presented with response buttons.

larger, as 12 AOM and 31 NOE cases were moved to the OME group. Furthermore, six cases were removed from the study as they have been labelled as Unknown diagnosis and will not be used in the rest of the analysis. These cases are of such low data quality, that it is not possible to diagnose based on the data. This could be because the tympanic membrane is not visible, the image is too blurry, or similar.

The overall agreement was evaluated and shows that all four ENTs agree on 57% of the cases, at least three of them agree on 81% of the cases, and they are split two and two on 19% of the cases. Table 3 shows Cohen’s Kappa together with the percentage-wise agreement on the diagnosis for each ENT pair. Based on the Kappa interpretation scale [19], the agreement ranges from moderate to substantial, with the pairs ENT2/ENT3 and ENT1/ENT4 achieving substantial agreement. Light’s Kappa (arithmetic mean of all pairwise Cohen’s Kappa) is 0.58, which is at the high end of the moderate range of agreement between the four ENTs on the otitis media diagnosis. The average pairwise agreement is 74%.

The relationships among the three variables, agreement, average certainty, and average time spent per annotation is evaluated using Pearson’s correlation coefficient. A moderately strong positive correlation of 0.62 ( $p < 0.001$ ) is found between average certainty and percentage agreement. The violin plot of the two variables in Fig. 4 (left) also shows that certainty increases as agreement increases. Similarly, the correlation with the time spent on each case was investigated. This resulted in a correlation between agreement and average time of  $-0.27$  ( $p < 0.001$ ) shown in Fig. 4 (middle), and between average time and average certainty of  $-0.46$  ( $p < 0.001$ ) shown in Fig. 4 (right). All correlations are thus statistically significant, but only agreement and average certainty show a strong correlation, whereas time shows weak relationship with agreement and certainty.

The certainty reported by each ENT can reflect several challenging aspects of the image or WBT. In Fig. 5, the five images with the lowest average certainty are shown. Some of the images such as a, b, d, and e

**Table 2**  
Confusion matrix between the original annotations used to set up the study and the majority voting results for all cases in the dataset.

	Original diagnosis					
	OME	AOM	NOE	Unknown	Total	
<b>Majority voting</b>	<b>OME</b>	63	12	31	0	106
	<b>AOM</b>	2	55	0	0	57
	<b>NOE</b>	3	0	32	0	35
	<b>Unknown</b>	0	1	5	0	6
	<b>Total</b>	68	68	68	0	204

are challenging since the eardrum is not clearly visible, and in some cases earwax is blocking the view, while other cases such as c show diagnostic signs that are not specific for a certain diagnosis.

As presented in Table 1, the dataset was split into two groups: 102 cases with both otoscopy image and WBT, and 102 cases only with an otoscopy image. Table 4 shows the agreement, average certainty, and average time spent per annotation for the cases in each of these two groups from each of the diagnostic groups. The table shows that in all three diagnostic groups, the agreement, certainty, and time increases when a WBT is presented together with the image. Statistical tests were run to examine whether the changes are significant. The  $\chi^2$ -contingency test was employed to determine whether agreement is different between the groups. The tests show that agreement is not significantly different for either of the diagnostic groups between with or without WBT. Thus, agreement does not significantly increase, as more information (the WBT) is presented for the cases. The Mann Whitney U tests indicate that both certainty and time is significantly increased when presenting the WBT compared with presenting only the image, except for the time in NOE cases, and certainty in AOM cases.

Table 4 also presents clear differences among the diagnostic groups. Agreement and certainty are higher for the AOM and OME cases, while NOE shows much lower agreement and certainty. The time spent also varies a lot between AOM and NOE, where AOM is much faster to diagnose than NOE, and OME is in the middle between the two other groups.

For the 102 cases with both image and WBT, the ENTs answered which data they used to determine their diagnosis. The ENTs reported that they used both image and WBT in 57% of the annotated cases, only otoscopy image in 38% of the annotated cases, and only WBT in 5% of the annotated cases.

The supplementary material includes the analysis presented in this section performed on the balanced incomplete block study of 1205 cases. The analysis in the supplementary materials shows the same tendencies as presented in this section.

**Table 3**  
Pairwise Cohen’s Kappa and percentage-wise agreement for each ENT pair.

	ENT1	ENT2	ENT3	ENT4
<b>ENT1</b>	–	0.56/72%	0.55/71%	0.71/83%
<b>ENT2</b>	–	–	0.65/77%	0.51/69%
<b>ENT3</b>	–	–	–	0.52/70%
<b>ENT4</b>	–	–	–	–

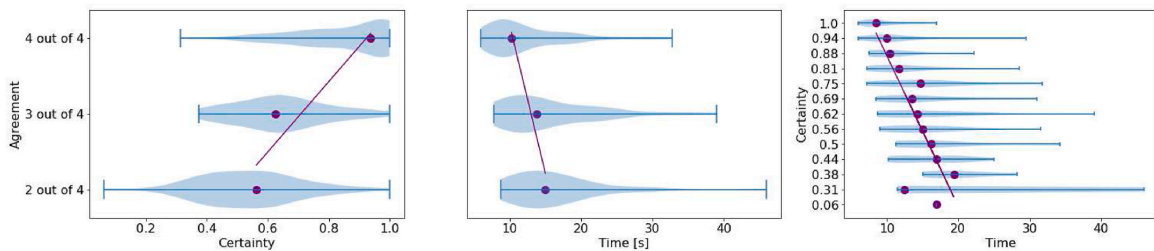


Fig. 4. Violin plots between agreement and certainty, agreement and time, and certainty and time. Purple dot marks the medians, and the regression lines are computed based on the median values, as the data does not follow a normal distribution. The y-axis on the middle graph is the same as that of the left graph. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

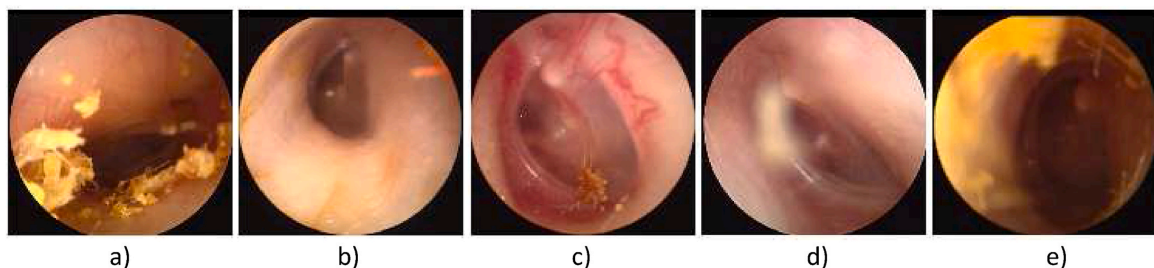


Fig. 5. Low certainty images. Average certainty from left: 0.063, 0.31, 0.31, 0.31, and 0.38. Earwax and restricted visibility of the tympanic membrane due to anatomical constraints affects certainty of the diagnosis.

Table 4

Agreement, certainty, and time for annotations in each of the three diagnostic groups with either only image or image and WBT.

		Image and WBT	Only image	p-value
OME	All 4 ENTs agree	56%	51%	0.8
	Min. 3 ENTs agree	80%	81%	–
	Average certainty	0.72	0.67	0.03
	Average time	16.1 s	12.1 s	<0.001
AOM	All 4 ENTs agree	77%	77%	0.6
	Min. 3 ENTs agree	96%	90%	–
	Average certainty	0.91	0.88	0.1
	Average time	13.2 s	9.0 s	0.003
NOE	All 4 ENTs agree	47%	20%	0.24
	Min. 3 ENTs agree	73%	40%	–
	Average certainty	0.65	0.51	0.001
	Average time	17.6 s	15.9 s	0.08

4. Discussion

The average pairwise agreement between the four ENTs of 74% compares quite well with the 75% accuracy of ENTs reported by Pichichero et al. [6]. The two measurements do not directly correspond, as one is agreement, and the other is accuracy compared to a ground-truth, but both show that ENTs disagree on the diagnosis of around 25% of cases. Our results also correspond well with the agreement reported by Blomgren et al. [8], who reported that all four medical professionals agreed on the diagnosis in 64% of the cases, whereas the four ENTs in our study agreed in 57% of the cases. However, their study only included AOM, for which all four of our ENTs agreed on 77% of all cases. The responses of this study show a higher certainty regarding the diagnosis of AOM compared to the 75% certainty found by Jensen et al. [5]. In our study, the reported certainty of AOM cases was 90%.

When comparing with previous work, it is important to note that this study only includes images and WBTs from the normal clinical routine. Thus, the quality of the data will vary greatly, as shown in Fig. 5. The

image examples with the lowest certainties show some of the common issues with otoscopy images, such as earwax, blurry images, or lack of appropriate illumination to obtain a clear visual impression of the tympanic membrane. This is a very different dataset than the data used by Pichichero et al. [6], for example, where all cerumen was removed before examination.

The reported certainties, computed agreement, and time for annotations were further examined, and the correlations indicate relationships between all three variables. These correlations are as expected, as time increases with lower certainty and lower agreement, indicating that more time is spent on annotating challenging cases. Furthermore, a strong positive correlation was found between agreement and certainty, which shows that self-reported certainty is a good indicator of how difficult a specific case is to diagnose.

This study examines how ENTs diagnose when only presented with an image and a WBT, without examining the actual patient, similar to the study design of Blomgren et al. [8]. There is, of course, a huge difference in the diagnostic process if the face-to-face patient examination with a detailed medical history is not included. Thus, the results from the *a posteriori* examination of the four ENTs are more representative of what can be expected in a remote-care scenario, and moreover they correspond exactly with the conditions for the deep learning diagnostic system described in the introduction. Table 2 shows how the weighted majority voting diagnosis and the original diagnosis by the ENT who examined the patients differ. The largest differences are found in the top row, which shows that 31 NOE cases and 12 AOM cases have moved to OME when diagnosed only by studying the image and WBT. This observation, together with the low agreement and certainty for NOE cases, suggests that the ENTs struggled to differentiate between otitis media and NOE. It is possible that the ENTs were biased towards choosing a diagnosis, due to the design of the study. It is also possible that it is easier to spot symptoms than the lack thereof. It is therefore hypothesized that the ENTs were reluctant to diagnose NOE, and inclined to diagnose either AOM, or in most cases, OME. Furthermore, the

most common sign of OME is lack of movement of the tympanic membrane during pneumatic otoscopy examination. Since the ENTs were only presented with a static image, even the slightest signs of effusion would count towards an OME diagnosis, instead of NOE. Contrary to the belief that OME is frequently misdiagnosed as AOM, and that AOM is usually over-diagnosed [9], the results of our study show that the most common error is that NOE is misdiagnosed as OME. This is not as serious an issue, however, as long as doctors follow clinical guidelines and do not prescribe antibiotics for the treatment of OME.

Blomgren et al. [8] showed that agreement increased when presenting a WBT with the otoscopy image. Our results show that the time increased as expected, as the ENTs had to study and analyze more data when presented with both image and WBT. The results generally show that the ENTs are more confident in their diagnosis when the WBT is presented, which is likely to lead to higher confidence when advising the following treatment. However, agreement across examiners did not increase significantly when the WBT was added. The ENTs responded that the WBT was the sole basis for the diagnosis in 5% of the cases, and contributed to the diagnostic decision in 57% of the cases, which suggests that the WBT does add diagnostic value, especially in cases where the image is not useful, or in challenging cases. The WBT measurements are mostly used for OME cases, and rarely for AOM, since the children are in too much pain to perform the measurement. The signs for AOM are also much clearer in the otoscopy image, which explains the higher agreement and certainty for the AOM cases, as well as why the difference in certainty between with or without WBT is not significant for this diagnostic group.

The results show that it is challenging to correctly identify otitis media in children from a static image and a WBT alone, and that ENTs do not always agree on the diagnosis. This makes it challenging to develop an automatic diagnostic tool, due to the lack of a consistent ground-truth definition based only on non-invasive examination. There are different ways of defining an approximate ground-truth when myringotomy-confirmed cases are not available. Some studies, such as Myburgh et al. [21], only included the cases where two specialists agree on the diagnosis. This ensures the most correct diagnosis of each case, but also removes a lot of challenging cases. This means that the performance of the diagnostic system is boosted but might not properly represent the performance of such a tool in real life cases. It also does not allow the diagnostic system to learn how to handle challenging cases. Based on the results from this study, that would mean removing 25–40% of the cases obtained from the normal clinical routine. On the other hand, if the ground-truth is based only on one ENT's opinion, it will be biased, and possibly include 25% incorrect diagnoses. For this study, we decided to do a majority voting between the ENTs weighted with their self-evaluated certainties. A normal majority voting could also be employed without the weighting, but as we have an equal number of raters, the vote will in some cases be tied. The weighting thus allows for a diagnosis to be determined even when the raw voting is tied.

#### 4.1. Strengths and limitations

This study has several major strengths. First, a large number of cases are included compared to the references studies, and the statistical analysis is more extensive than seen in previous studies. Furthermore, the study is blinded, as the ENTs did not have any prior knowledge about the patients. The quality of the data is generally good from both otoscopy images and WBT measurements. Finally, some cases include both otoscopic image and wideband tympanometry measurements, while others only include otoscopic image, potentially affecting the inter-rater agreement, as seen in Blomgren et al. [8], although they used a standard tympanometry, not a WBT. The main limitation of this study is that no myringotomy was performed, and thus no ground-truth diagnosis is available for comparison. Furthermore, despite the high quality of the data, the ENTs were only presented with static images. Thus, the ENTs in this study did not benefit from observing the full otoscopy examination,

nor from gaining additional diagnostic information by interacting with the patient.

## 5. Conclusion

This study illustrates that, under these conditions, it is challenging to diagnose otitis media with effusion when the ENT is only provided with static images and WBT data. The inter-rater reliability between the four ENTs was high when diagnosing acute otitis media and lower when diagnosing otitis media with effusion. However, WBT can add valuable information to get closer to the ground-truth diagnosis without myringotomy. It was furthermore shown that diagnostic certainty increased significantly when showing both image and WBT, compared to when only presenting the image. This study provides a useful comparison benchmark for future work on an automated deep learning approach using the same diagnostic inputs, as well as an estimate of the ground-truth diagnosis for each case.

## Acknowledgements

We would like to thank the William Demant Foundation (Denmark) for financially supporting this study. We would furthermore like to thank Fredrik Tjernström, Frida Enoksson, and Anders Åkerlund from the Department of Clinical Sciences at Lund University in Sweden for participating in this study. Thank you to Morten Bo Svendsen from Copenhagen Academy of Medical Education and Simulation (CAMES) for his help on developing the web interface for rating the images.

## Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ijporl.2021.111034>.

## References

- [1] A. Danishyar, J.V. Ashurst, Acute Otitis Media, StatPearls Publishing LLC, 2019.
- [2] G. Worrall, ARI Series Acute otitis media, Can. Fam. Physician 53 (12) (2007) 2147–2148.
- [3] J. Céldind, L. Södermark, O. Hjalmarson, Adherence to treatment guidelines for acute otitis media in children. The necessity of an effective strategy of guideline implementation, Int. J. Pediatr. Otorhinolaryngol. 78 (7) (2014) 1128–1132.
- [4] N.E. Cullas Ilarslan, F. Gunay, S. Topcu, E. Ciftci, Evaluation of clinical approaches and physician adherence to guidelines for otitis media with effusion, Int. J. Pediatr. Otorhinolaryngol. 112 (2018) 97–103.
- [5] P.M. Jensen, J. Lous, Criteria, performance and diagnostic problems in diagnosing acute otitis media, Fam. Pract. 16 (3) (1999) 262–268.
- [6] M.E. Pichichero, M.D. Poole, Assessing diagnostic accuracy and tympanocentesis skills in the management of otitis media, Arch. Pediatr. Adolesc. Med. 155 (10) (2001) 1137–1142.
- [7] G. Flores, M. Lee, H. Bauchner, B. Kastner, Pediatricians' attitudes, beliefs, and practices regarding clinical practice guidelines: a national survey, Pediatrics 105 (3) (2000) 496–501.
- [8] K. Blomgren, A. Pitkäranta, Is it possible to diagnose acute otitis media accurately in primary health care? Fam. Pract. 20 (5) (2003) 524–527.
- [9] M.E. Pichichero, Diagnostic accuracy of otitis media and tympanocentesis skills assessment among pediatricians, Eur. J. Clin. Microbiol. Infect. Dis. 22 (9) (2003) 519–524.
- [10] J.V. Sundgaard, et al., Deep metric learning for otitis media classification, Med. Image Anal. 71 (2021).
- [11] J.V. Sundgaard, et al., A deep learning approach for detecting otitis media from wideband tympanometry measurements, In submission (2022).
- [12] J.C. Ellison, M. Gorga, E. Cohn, D. Fitzpatrick, C.A. Sanford, D.H. Keefe, Wideband acoustic transfer functions predict middle-ear effusion, Laryngoscope 122 (4) (2012) 887–894.
- [13] T.A.D. Hein, S. Hatzopoulos, P.H. Skarzynski, M.F. Colella-Santos, "Wideband Tympanometry," in *Advances in Clinical Audiology*, 2017.
- [14] A.A. Mitani, P.E. Freer, K.P. Nelson, Summary measures of agreement and association between many raters' ordinal classifications, Ann. Epidemiol. 27 (10) (2017) 677–685.
- [15] R.J. Light, Measures of response agreement for qualitative data: some generalizations and alternatives, Psychol. Bull. 76 (5) (1971) 365–377.
- [16] K.A. Hallgren, Computing inter-rater reliability for observational data: an overview and tutorial, Tutor. Quant. Methods Psychol. 8 (1) (2012) 23.
- [17] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1) (1960) 37–46.

- [18] L. Cyr, K. Francis, Measures of clinical agreement for nominal and categorical data: the kappa coefficient, *Comput. Biol. Med.* 22 (4) (1992) 239–246.
- [19] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, *Biometrics* (1977) 159–174.
- [20] M.L. McHugh, Interrater reliability: the kappa statistic, *Biochem. Med.* 22 (3) (2012) 276–282.
- [21] H.C. Myburgh, S. Jose, D.W. Swanepoel, C. Laurent, Towards low cost automated smartphone- and cloud-based otitis media diagnosis, *Biomed. Signal Process Control* 39 (2018) 34–52.

CONTRIBUTION **F**

# Was that so hard? Estimating human classification difficulty

---

**Authors** Morten Rieger Hannemose\*, Josefine Vilsbøll Sundgaard\*, Niels Kvorning, Rasmus R. Paulsen, and Anders Nymark Christensen.

**Status** In submission

**Link** <https://arxiv.org/abs/2203.11824>

---

\*Authors contributed equally

# Was that so hard? Estimating human classification difficulty

Morten Rieger Hannemose<sup>\*1</sup>, Josefine Vilsbøll Sundgaard<sup>\*1</sup>, Niels Kvorning Ternov<sup>2</sup>, Rasmus R. Paulsen<sup>1</sup>, and Anders Nymark Christensen<sup>1</sup>

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

<sup>2</sup> Department of Plastic Surgery, Copenhagen University, Herlev and Gentofte Hospital, Copenhagen, Denmark

**Abstract.** When doctors are trained to diagnose a specific disease, they learn faster when presented with cases in order of increasing difficulty. This creates the need for automatically estimating how difficult it is for doctors to classify a given case. In this paper, we introduce methods for estimating how hard it is for a doctor to diagnose a case represented by a medical image, both when ground truth difficulties are available for training, and when they are not. Our methods are based on embeddings obtained with deep metric learning. Additionally, we introduce a practical method for obtaining ground truth human difficulty for each image case in a dataset using self-assessed certainty. We apply our methods to two different medical datasets, achieving high Kendall rank correlation coefficients, showing that we outperform existing methods by a large margin on our problem and data.

**Keywords:** Difficulty estimation · Deep metric learning · Human classification.

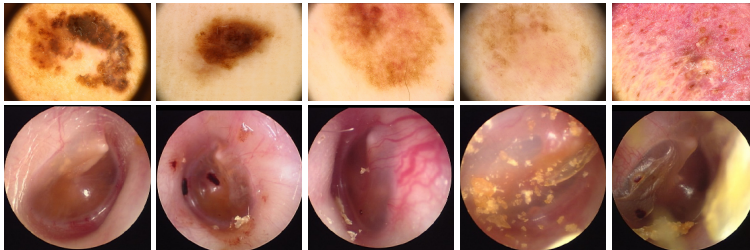
## 1 Introduction

When doctors are diagnosing patients, not all cases have the same difficulty. A case can be very easy if there are clear diagnostic signs. However, if the typical signs are missing or give conflicting information, a doctor will be more likely to assign an incorrect diagnosis. When doctors are trained to diagnose certain diseases, they learn faster when starting with easy cases and then gradually progressing to harder cases [23]. Knowing how hard each case is to classify is thus useful in an educational context. This concept is well-known in pedagogy [8] and applies to many other areas such as language training, mathematics, etc.

In this paper, we present a novel approach for estimating human difficulty in image classification using deep metric learning. In deep metric learning, high-dimensional data (in our case, images) are mapped to a lower-dimensional embedding that captures similarities between the training examples: Similar images

---

\* These authors contributed equally



**Fig. 1.** Image examples from the skin lesion (top row) and eardrum (bottom row) datasets. The difficulty increases from left to right from 0 to 1 in steps of 0.25 for each image. For the skin lesion dataset, only images from the melanoma class are shown, while the eardrum images are from all three diagnostic classes, see Section 3.

cluster together, and dissimilar images are pushed apart. In our paper, we define metrics in the embedding space that capture human classification difficulty. We evaluate our methods on two different medical datasets, one containing images of skin lesions and the other of eardrums, see Fig. 1.

The term *difficulty* is used in various ways in image analysis. Difficulty can be defined as how hard it is for machine learning to reach high accuracy on a given dataset [26], how challenging it is to automatically segment an image [16], visual complexity and clutter in the image [20], the time needed for a human to segment an image [34], or the human response time for a visual search task [32]. The latter definition was employed by Ionescu *et al.* [32], who proposed a method based on a pretrained neural network for feature extraction, followed by support vector regression to estimate the difficulty score. They presented a dataset with difficulty scores on the PASCAL VOC2012 dataset evaluated by 736 raters. Ma *et al.* [17] presented an approach on the same dataset using an end-to-end multi-loss network trained to optimize Kendall’s  $\tau$  coefficient to predict the difficulty scores. Both approaches achieved high Kendall’s  $\tau$  coefficients of 0.472 and 0.476, respectively. In contrast to our approach, neither of these use any knowledge about the ground truth class of the image but instead estimated the difficulty directly from image features. By using both the ground truth class and an embedding space our approach becomes more interpretable [25].

In some clinical problems, a specific difficulty scale is already defined. Yoo *et al.* [37] predict the difficulty of extracting a mandibular molar from a panoramic radiographic image using a pretrained convolutional neural network. In their study, the Pederson difficulty score is used, which is a pre-defined difficulty scale for extracting mandibular molars. However, this score is purely related to the difficulty of performing the required procedure and not the difficulty of diagnosis. André *et al.* [1] propose a method to estimate interpretation difficulty in endomicroscopy videos. Their approach is based on the content-based video retrieval method known as bag-of-visual-words, and the difficulty is given by the percentage of false diagnoses among annotators compared to a ground truth di-



agnosis from biopsies. We define human difficulty, similarly to André *et al.* [1], as the fraction of incorrect classifications from people familiar with the classification task. For one of our datasets, we use a self-evaluated certainty of all raters to obtain a less noisy estimated difficulty ground truth with few annotators.

We compare our work to methods in active learning and curriculum learning. Active learning accelerates labeling efficiency by selecting the most useful samples from an unlabeled dataset for labeling, thus reducing the labeling cost [22]. The intuition behind the most commonly used approach, the uncertainty-based approach, is that with a lower certainty on a specific example, a higher amount of informativeness will be added to the classifier when utilizing the example for training [36]. Curriculum learning is inspired by the learning process of humans, where examples are presented with increasing order of difficulty. This concept is transferred to neural networks to increase training speed and performance by introducing easy examples at the beginning of training, and to gradually increase the difficulty of the training examples [2].

In this paper, we present a new procedure for obtaining ground truth human difficulty from several annotators by including a self-evaluated certainty. We also propose a new method for estimation of human difficulty based on embeddings of images learned using deep metric learning, which outperforms existing methods by a large margin. We propose methods that both utilize ground truth difficulties and methods that do not. Finally, we are the first to utilize the ground truth class label for human difficulty estimation, which increases the performance of our methods even further.

## 2 Estimating image difficulty

Our difficulty estimation models are all based on the embedding space learned using a deep neural network, trained using metric learning. By training a model this way, instead of as a classification network, we learn the similarities in the training dataset. The output from the network is an embedding vector, mapping each individual image to the embedding space. The idea is that easy cases will be placed far from decision boundaries in the embedding space, while difficult cases will be further away from the class cluster center, and possibly closer to other cluster classes. We separate our proposed methods into two categories depending on whether or not they utilize ground truth difficulties during training. An overview of these methods is in Table 1.

**Methods without ground truth difficulties** are all based on embeddings of samples, extracted using a trained neural network. As our neural networks are trained using cosine similarity, our methods for estimating difficulties are thus also based on cosine similarities. As difficulties should be high for points far from their cluster, we refer to inverse similarity which is one minus the similarity. The methods still apply to neural networks trained using Euclidean distances, and in that case, one would use the Euclidean distance in the embedding space instead.

*Inverse similarity* is a naïve approach to estimating the difficulty, found by taking the similarity between the sample and the cluster center of its ground

truth class. This is intuitive, as samples less similar to the cluster center are typically more similar to other class clusters, and thus harder to classify. To find the difficulty, and not the easiness, we report the inverse of the similarity.

*Inverse softmax of similarity* is an improvement of inverse similarity. Samples can have low similarity to their cluster center without being close to other classes. To handle this, we compute the similarity between the sample and all cluster centers and normalize these with softmax. The difficulty is the inverse of the softmax output corresponding to the ground truth class. This method is related to decision margin sampling in active learning [33], except we can go on both sides of the decision boundary since the ground truth label is known.

*Sample classification power* is an alternative way of obtaining an estimate of image difficulty. Here, we evaluate how many of the neighboring points in the embedding space belong to the ground truth class of a certain sample. To do that for a single sample  $s$  from class  $c$ , we imagine classifying the closest  $k$  samples as  $c$ , and classifying the rest as not  $c$ . By varying  $k$  from one to the number of samples, we can draw a receiver operating characteristic (ROC) curve. We then use the area under the curve (AUC) of this ROC curve as our estimate of the difficulty of  $s$ . To handle class imbalance, we use the weighted ROC curve, with the weights being the inverses of the class frequencies.

*Normalization* is carried out on the estimated difficulties, by introducing the assumption that each class has the same average difficulty. To enforce this assumption, we propose normalizing the difficulty on a per-class basis by dividing it by the average estimated difficulty of that class. We refer to this as “norm”.

**Methods with ground truth difficulties** are methods, where the ground truth difficulties of a training set are employed. We set up a regression problem to predict the difficulty scores directly from the pre-trained image embeddings. We employ the tree-based ensemble model extra trees [9] for the regression problem. In addition to only predicting from the embeddings, we also fit a model using the ground truth label as additional input. The ground truth label will allow the model to learn that samples placed close to incorrect class clusters should have a higher difficulty, than samples within their correct class cluster.

### 3 Datasets

To validate our method, we have performed experiments on two medical image datasets, examples of which are shown in Fig. 1. We have obtained estimates of the human difficulty for a number of images from both datasets, which we use as our test-sets for evaluating our proposed approaches.

**The skin lesion dataset** consists of dermoscopic images of skin lesions divided into eight diagnoses, which include benign (nevus [NV], keratoses [BKL], vascular lesions [VASC], dermatofibromas [DF]), pre-malignant (actinic keratoses), and malignant (melanoma [MEL], squamous cell carcinoma [SCC], basal cell carcinoma [BCC]). The diagnoses were determined by histopathology or as the consensus between two to three domain experts. We have a dataset of 52 292

images from the 2019 ISIC Challenge training set [31,4,5]<sup>3</sup> and our own dataset (Permission to access and handle the patients’ data was granted by the Danish Patient Safety Authority (Jr.# 3-3013-2553/1) and the Data Protection Agency of Southern Denmark (Jr.# 18/53664)).

Skin lesion difficulties are obtained for 1723 images from our own dataset, based on diagnoses from 81 medical students with an interest in dermatology (Ethical waiver: Jr.#: H-20066667, data handling agreement case #: P-2019-556). On average, each student diagnosed 609 randomly sampled images. It was ensured that at least eight students diagnosed each case.

The images were diagnosed into seven different categories, as we expected actinic keratoses would be too difficult for the medical students. We estimate the difficulty of a case as the fraction of students answering incorrectly.

**The eardrum dataset** contains 1409 images collected during the standard clinical routine at an Ear-Nose-and-Throat (ENT) clinic. The data was collected under the ethical approval from the Non-Profit Organization MINS Institutional Review Board (ref.# 190221). The images show the patients’ eardrum captured using an endoscope and are diagnosed into three different diagnoses: acute otitis media, otitis media with effusion, and no effusion by an experienced ENT specialist. The dataset is split into a training and test set of 1209 and 204 images.

Eardrum difficulties were estimated by getting the test set of 204 equally class sampled eardrum images analyzed and diagnosed by four additional experienced ENTs. The ENTs diagnosed each case as one of the three diagnoses or “unknown”, counting as an incorrect diagnosis. Furthermore, each ENT rated their certainty of each diagnosis on the scale: very low, low, medium, moderate, or high, which is converted to a scale from 0 to 1. More details on this dataset are in Sundgaard *et al.* [29]. For a case,  $\mu_{correct}$  is the fraction of correct ENT answers and  $\mu_{certainty}$  is the average self-evaluated certainty. The difficulty of each case is then

$$1 - \mu_{correct} \cdot \mu_{certainty}. \quad (1)$$

We evaluate the difficulties with “leave-one-annotator-out”. This gave an average Kendall’s  $\tau$  of 0.548 based only on the fraction of correct ENT answers, which increased to 0.570 when including the self-evaluated certainty, showing that this improves the estimated difficulties.

## 4 Experiments

The embeddings of the images in our proposed methods are computed using neural networks trained with a metric loss function. All experiments are conducted in PyTorch (v. 1.10) using the PyTorch metric learning library [19]. The neural networks are trained using the multi-similarity loss function [35] ( $\alpha = 2$ ,  $\beta = 50$ , base = 1) and a multi similarity miner ( $\epsilon = 0.1$ ) using cosine similarity to optimize the selection of training pairs. Our models are pretrained on the ImageNet database [7] and trained using the Adam optimizer [13]. The fully connected

<sup>3</sup> License: CC-BY-NC

layer before the final softmax of the model is replaced by a fully connected layer without an activation function, which returns the embedding space. The output embeddings are L2 normalized.

The skin lesion network is based on a ResNet-50 model [11], with a 64-dimensional embedding space. The model is trained for 350 epochs with a learning rate of  $10^{-5}$ . The input images ( $256 \times 256$ ) are color normalized using the Minkowski norm ( $p = 6$ ). Data augmentation consists of flips, rotations, scaling, and color jitter. We do inference with the same augmentations, and compute each prediction as the average of 64 random augmentations.

The eardrum network is based on the Inception V3 network [30], with a 32-dimensional embedding space. The Inception V3 network has been used by several others for similar images [3,28]. The parameters of the first half of the network (until first grid size reduction) were frozen to avoid over-fitting. The initial learning rate ( $10^{-3}$ ) is decreased by a factor of 0.1 every 50<sup>th</sup> epoch. Training is continued until the training loss has not decreased for 20 epochs, resulting in 111 training epochs. Data augmentation consists of horizontal flips, rotations, color jitter, and random erasing. Images are resized to  $299 \times 299$ .

We use Kendall’s  $\tau$  [12] to evaluate how well our methods can predict the ground truth difficulties. This is a non-parametric measurement of the correlation between two ranked variables. As it only compares how the images are ranked, it is not important to achieve the exact same difficulty as the ground truth estimate, as long as the ordering of samples is correct.

We use Extra trees [9] for supervised difficulty estimation, with five-fold cross-validation. This allows us to obtain predictions for all samples in the test set, and thus compute a single Kendall’s  $\tau$  for the entire test set. All our experiments with extra trees use 500 trees, with 10 as the minimum number of samples required to split an internal node.<sup>4</sup>

**Comparisons** are made between our methods and methods from both active and curriculum learning using a standard trained classification network, and with the approach proposed by Ionescu *et al.* [32]. The classification networks employ the same architecture as our embedding networks, but the dimension of the output is the number of classes in each dataset. The networks are trained with cross-entropy loss weighted by the inverse frequency of each class, but otherwise using the same setup as described for the embedding networks.

Visual search difficulty proposed by Ionescu *et al.* [32] is used for comparison. We replicate their method by passing each image ( $299 \times 299$ ) through VGG-16 [27] once and using the penultimate features to fit a  $\nu$ -support vector regression.<sup>4</sup>

We compare to the following approaches from active learning, all based on the softmax output of a classification network: classification uncertainty, which is one minus the maximum value of the softmax [14]; entropy of the softmax probabilities [6]; and classification margin found by computing the difference between the second-highest and highest probabilities of the softmax [15].

We also compare to three approaches from curriculum learning: standard deviation of the images [24]; transfer scores obtained by running all images through

<sup>4</sup> Unspecified parameters are the defaults in Scikit-Learn v. 0.24.2 [21].

**Table 1.** Kendall’s  $\tau$  for all methods on both datasets. It is indicated which methods utilize: the ground truth class label for prediction (L) or a training set of ground truth difficulties (D). Bold indicates the significantly best performance, and bold with a star indicates the methods without D performing significantly better than the rest. We used bootstrap based hypothesis testing with 50 000 replicates and  $\alpha = 5\%$ .

Method	Uses	Skin lesion	Eardrum
Visual search difficulty [32]	D	0.142	0.117
<b>Curriculum learning</b>			
Std. of image [24]		-0.070	0.011
Transfer scoring [10]	L	0.115	0.213
Self-taught scoring [10]	L	0.176	0.261
<b>Active learning</b>			
Classification uncertainty [14]		0.094	0.217
Entropy of probabilities [6]		0.118	0.216
Classification margin [15]		0.068	0.215
<b>Ours</b>			
Inverse similarity	L	0.137	-0.140
Inverse softmax of similarity	L	<b>0.239*</b>	0.354
Inverse softmax of similarity norm.	L	<b>0.239*</b>	0.380
Sample classification power	L	0.201	0.143
Sample classification power norm.	L	<b>0.247*</b>	<b>0.440*</b>
Extra trees: embeddings	D	0.322	0.465
Extra trees: embeddings + label	L D	<b>0.398</b>	<b>0.517</b>

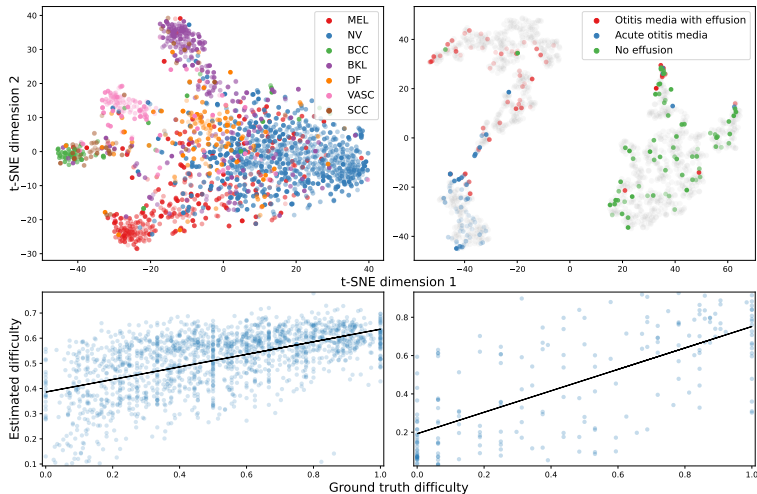
a pretrained Inception V3 network using the penultimate features to train a support vector classifier to obtain the confidence of the model [10]; and one minus the softmax output of the ground truth class from our classification network [10].

## 5 Results

The Kendall’s  $\tau$  for all experiments is reported in Table 1. The table also gives an overview of whether the ground truth label is used for prediction, and whether a training set of ground truth difficulties has been used. The embeddings for the two datasets are shown in the top of Figure 2. For the eardrum data, we see how most easy examples are located within the class clusters, while the difficult examples are the ones located in another class cluster, or at the edge of the clusters. The same tendencies are visible in some classes of the skin lesion embeddings. The bottom of Figure 2 shows scatter plots of the ground truth difficulties versus predicted difficulties for both datasets.

## 6 Discussion and conclusion

We have shown that neural networks trained using metric learning can be used to estimate diagnostic difficulty. Our methods for difficulty estimation outperform



**Fig. 2.** **Left:** skin lesion dataset. **Right:** eardrum dataset. **Top:** visualization of the embeddings in two dimensions with t-SNE [18]. The transparency of each point indicates the ground truth difficulty with very transparent being the easiest. Grey points are the training samples for the eardrum data. **Bottom:** scatter plots of ground truth difficulties and difficulties estimated with the *embeddings + label* approach, together with the least squares regression lines.

all existing methods in both active and curriculum learning. Ionescu *et al.* [32] report a Kendall’s  $\tau$  of 0.472, while their method achieves 0.142 and 0.117 on our datasets. Our methods are significantly better, with our best achieving a Kendall’s  $\tau$  of 0.398 and 0.517. This corresponds to 69.9% and 75.8% of pairs being ordered correctly, which is an improvement of 12.8 and 11.1 percentage points from the best performing existing method (self-taught scoring).

Table 1 shows that our contribution of incorporating the ground truth class greatly increases performance. A similar tendency is seen in the higher performance of self-taught scoring compared to classification uncertainty, as the only difference between these two methods is the knowledge about the ground truth class. This intuition is also visible in Fig. 2, especially for the eardrum dataset, where the most difficult examples are often placed in the extremities of the clusters, or placed inside other clusters. This indicates that the embedding has a relation to difficulty, and shows the relevance of including the ground truth class label when estimating difficulty. Our methods have demonstrated great potential in the estimation of human classification difficulty of medical images, which can be used to optimize and improve the training of medical professionals.

## References

1. André, B., Vercauteren, T., Buchner, A.M., Shahid, M.W., Wallace, M.B., Ayache, N.: An image retrieval approach to setup difficulty levels in training systems for endomicroscopy diagnosis. In: MICCAI. pp. 480–487 (2010)
2. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: ICML. pp. 41–48 (2009)
3. Cha, D., Pae, C., Seong, S.B., Choi, J.Y., Park, H.J.: Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database. *EBioMedicine* **45**, 606–614 (2019)
4. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 168–172. IEEE (2018)
5. Combalia, M., Codella, N.C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., et al.: Bcn20000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288 (2019)
6. Dagan, I., Engelson, S.P.: Committee-based sampling for training probabilistic classifiers. In: In Proceedings of the Twelfth International Conference on Machine Learning. pp. 150–157. Morgan Kaufmann (1995)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
8. Elio, R., Anderson, J.R.: The effects of information order and learning mode on schema abstraction. *Memory & cognition* **12**(1), 20–30 (1984)
9. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. *Machine learning* **63**(1), 3–42 (2006)
10. Hacothen, G., Weinshall, D.: On the power of curriculum learning in training deep networks. In: ICML. pp. 2535–2544 (2019)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Kendall, M.G.: Rank correlation methods. (1948)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: SIGIR’94. pp. 3–12. Springer (1994)
15. Li, X., Guo, Y.: Active learning with multi-label svm classification. In: Twenty-Third International Joint Conference on Artificial Intelligence (2013)
16. Liu, D., Xiong, Y., Pulli, K., Shapiro, L.: Estimating image segmentation difficulty. In: International Workshop on Machine Learning and Data Mining in Pattern Recognition. pp. 484–495. Springer (2011)
17. Ma, D., Zhang, H., Wu, H., Zhang, T., Sun, J.: Estimating difficulty score of visual search in images for semi-supervised object detection. In: Pacific Rim Knowledge Acquisition Workshop. pp. 1–9. Springer (2019)
18. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. *Journal of machine learning research* **9**(11) (2008)
19. Musgrave, K., Belongie, S., Lim, S.N.: Pytorch metric learning (2020)

20. Nagle, F., Lavie, N.: Predicting human complexity perception of real-world scenes. *Royal Society open science* **7**(5), 191487 (2020)
21. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011)
22. Ren, P., Xiao, Y., Chang, X., Huang, P.Y., Li, Z., Chen, X., Wang, X.: A survey of deep active learning. *arXiv preprint arXiv:2009.00236* (2020)
23. Roads, B.D., Xu, B., Robinson, J.K., Tanaka, J.W.: The easy-to-hard training advantage with real-world medical images. *Cognitive Research: Principles and Implications* **3**(1), 1–13 (2018)
24. Sadasivan, V.S., Dasgupta, A.: Statistical measures for defining curriculum scoring function. *arXiv preprint arXiv:2103.00147* (2021)
25. Sanakoyeu, A., Tschernezki, V., Buchler, U., Ommer, B.: Divide and conquer the embedding space for metric learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 471–480 (2019)
26. Scheidegger, F., Istrate, R., Mariani, G., Benini, L., Bekas, C., Malossi, C.: Efficient image dataset classification difficulty estimation for predicting deep-learning accuracy. *The Visual Computer* **37**(6), 1593–1610 (2021)
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014)
28. Sundgaard, J.V., Harte, J., Bray, P., Laugesen, S., Kamide, Y., Tanaka, C., Paulsen, R.R., Christensen, A.N.: Deep metric learning for otitis media classification. *Medical Image Analysis* **71**, 102034 (2021)
29. Sundgaard, J.V., Vårendh, M., Nordström, F., Kamide, Y., Tanaka, C., Harte, J., Paulsen, R.R., Christensen, A.N., Bray, P., Laugesen, S.: Inter-rater reliability of the diagnosis of otitis media based on otoscopic images and wideband tympanometry measurements. *International Journal of Pediatric Otorhinolaryngology* p. 111034 (2022)
30. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: *CVPR*. pp. 2818–2826 (2016)
31. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data* **5**(1), 1–9 (2018)
32. Tudor Ionescu, R., Alexe, B., Leordeanu, M., Popescu, M., Papadopoulos, D.P., Ferrari, V.: How hard can it be? estimating the difficulty of visual search in an image. In: *CVPR*. pp. 2157–2166 (2016)
33. Tuia, D., Volpi, M., Copa, L., Kanevski, M., Munoz-Mari, J.: A survey of active learning algorithms for supervised remote sensing image classification. *IEEE Journal of Selected Topics in Signal Processing* **5**(3), 606–617 (2011)
34. Vijayanarasimhan, S., Grauman, K.: What’s it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In: *CVPR*. pp. 2262–2269 (2009)
35. Wang, X., Han, X., Huang, W., Dong, D., Scott, M.R.: Multi-similarity loss with general pair weighting for deep metric learning. In: *CVPR*. pp. 5022–5030 (2019)
36. Wu, J., Sheng, V.S., Zhang, J., Li, H., Dadakova, T., Swisher, C.L., Cui, Z., Zhao, P.: Multi-label active learning algorithms for image classification: Overview and future promise. *ACM Computing Surveys (CSUR)* **53**(2), 1–35 (2020)
37. Yoo, J.H., Yeom, H.G., Shin, W., Yun, J.P., Lee, J.H., Jeong, S.H., Lim, H.J., Lee, J., Kim, B.C.: Deep learning based prediction of extraction difficulty for mandibular third molars. *Scientific Reports* **11**(1), 1–9 (2021)





CONTRIBUTION 

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

---

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

**Status** In preparation

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

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

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

<sup>2</sup>*Interacoustics Research Unit, c/o Technical University of Denmark, Denmark*

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

<sup>4</sup>*Kamide ENT clinic, Shizuoka, Japan*

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

**In this study, we propose a diagnostic model for automatic detection of otitis media based on a combined input of otoscopy images and wideband tympanometry measurements. We present a neural network-based model for the joint prediction of otitis media, in the subclassifications acute otitis media and otitis media with effusion, and diagnostic difficulty. The proposed approach is based on deep metric learning, and we compare this with the performance of a standard multi-task network. The proposed deep metric approach shows good performance on both tasks, and it is shown that the multi-modal input increases the performance for both classification and difficulty estimation, compared to the models trained on the modalities individually. 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. This study shows that deep metric learning enables detection of otitis media with high performance, and demonstrates the strengths of a multi-modal diagnostic tool using both otoscopy images and wideband tympanometry measurements.**

## I. Introduction

Automatic diagnosis of otitis media has been tackled in various ways. Previous studies have employed datasets of otoscopy images [1–4], tympanometry measurements [5–7], optical coherence tomography [8], or computed tomography [9]. The approaches have focussed on a single modality, and utilised a variety of machine learning algorithms for the data analysis and classification task, progressing from simpler methods such as random forest [10] and support vector machines [11], to deep neural networks [1, 5, 7, 12, 13]. However, when a doctor examines a patient, the diagnostic decision is rarely based solely on one modality of the clinical examination. Binol et al. [14] was the first to combine otoscopy images and standard tympanometry measurements for the classification of normal

or abnormal tympanic membrane. 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. 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

---

\*Shared senior authorship

mechanical and acoustic status of the middle ear than the standard 226 Hz tympanogram [15]. Furthermore, a higher classification accuracy can be achieved using WBT measurements for detection of otitis media, compared to both ambient absorbance and standard tympanograms [7]. In the present paper, 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, acute otitis media, and no effusion. Examples of images and WBT measurements from the three groups are shown in Figure 1.

We are the first to propose a purely neural network-based model for the analysis of otoscopy images and WBT measurements combined in a single model. Furthermore, our models are developed for joint prediction of otitis media, in the subclassifications acute otitis media and otitis media with effusion, and diagnostic difficulty. This subclassification is important to ensure proper treatment.

There has been an increasing interest in neural network-based diagnosis of otitis media, and other

middle ear conditions, based on otoscopy images. Habib et al. [16] recently published a review on this topic including 39 papers published over the past 10 years. They conclude that these classification models have been shown to be 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 to allow 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 [7, 17]. Another valuable output would be an estimate of the diagnostic difficulty of the input case. This 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. The estimation of diagnostic difficulty was investigated by Hannemose et al. [18] based on image embeddings from a metric learning-based neural network. In the paper, several supervised

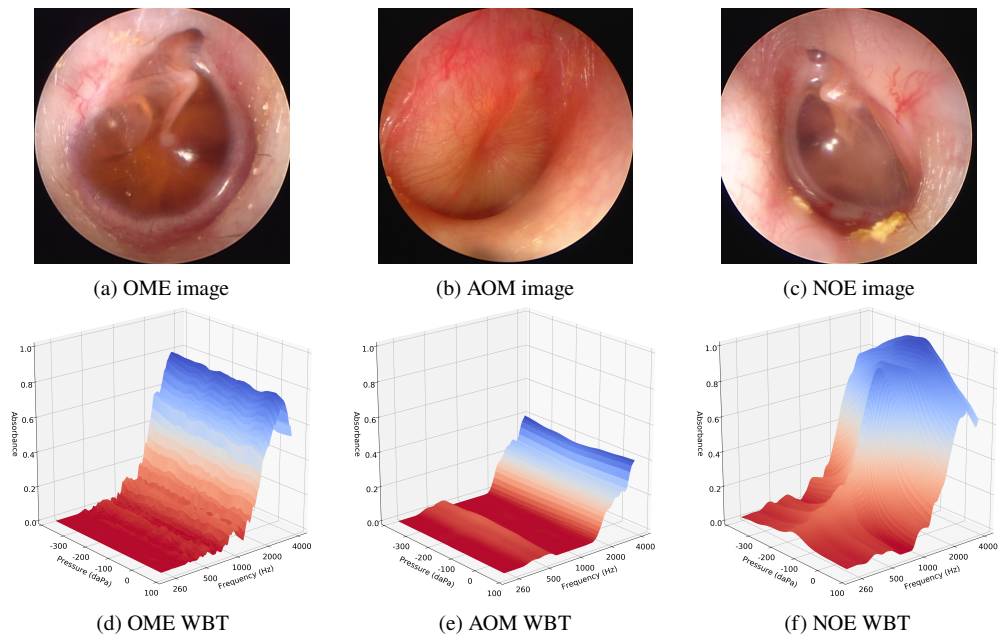


Figure 1. Otoscopy images and WBT measurements from patients otitis media with effusion (a and d), acute otitis media (b and e), and no effusion (c and f).

and unsupervised methods are presented to estimate the difficulty from the distribution of the dataset in the embedding space.

The goal of the present work is to predict both diagnostic class and difficulty for each case, and we will evaluate two methods for this task. One of the proposed methods is a deep metric learning approach, where the diagnostic class and difficulty will be predicted from the embedding space, using the supervised method presented by Hannemose et al. [18], and the other is a multi-task network for the joint prediction.

## II. Methods

We propose a single network for the combined analysis of otoscopy images and WBT measurements. The network architecture, seen in Figure 2, consists of a pre-trained Inception V3 [19] 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. [7]. 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 of the two different inputs into a single decision. The size of the layers gradually decreases through the network until the final 32-dimensional vector.

In this paper, we will compare the use of a multi-task neural network for simultaneous prediction of otitis media and diagnostic difficulty with a deep metric learning model, where the output embeddings of the test set are used to predict the otitis media diagnostic and to estimate the diagnostic difficulty. In deep metric learning, 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 2. In deep metric learning, the goal is to learn similarities in the dataset by measuring the distances between different samples in the embedding space. During training, the network learns to move similar cases together and move 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. This allows us to use this embedding space for either classification or derivation of other metrics, like diagnostic difficulty.

The deep metric learning network is trained using the multi-similarity loss function [20] ( $\alpha = 2$ ,  $\beta = 50$ , base = 1) and a multi-similarity miner ( $\epsilon = 0.1$ ) using cosine similarity to optimise the selection of training pairs. Classification is performed in the embedding space by assigning the class with the closest training data cluster centre to the current test

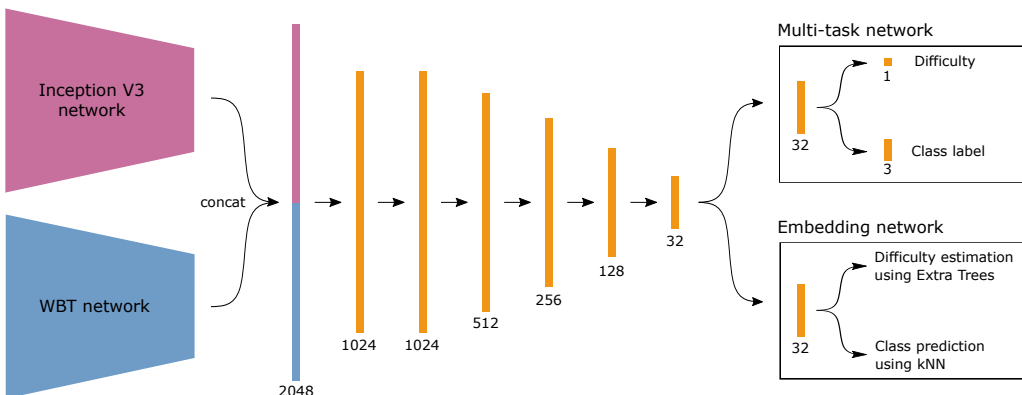


Figure 2. Network architecture of the multi-modal otoscopy image and WBT network. Numbers below the boxes indicate the size of the layer. The boxes to the right show the final layers of the multi-task approach and the embedding network approach.

example. Difficulty estimation is performed with the supervised method employing extra trees [21] with both embeddings and ground truth labels as input [18].

The other approach is the multi-task network. The output of this network consists of two fully connected layers, one with a single output for the difficulty, and another with a softmax output with size 3 for prediction of 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 class prediction, using the inverse frequency of each class as weights.

All networks were trained with a learning rate of 0.0001, decreased by a factor of 0.1 every 50<sup>th</sup> epoch. The training is stopped using early stopping with a patience of 50 epochs. During training, data augmentation of both input modalities was performed. For input images, transformations include horizontal flips, random rotation, colour jitter, and random erasing. For the WBT measurements, we employ the transformations shown to improve training of the WBT network [7]: 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 noise affecting a larger area in the input than the random noise.

### A. Data

The dataset consists of 1014 pairs of otoscopy images and WBT measurements collected at Kamide ENT clinic, Shizuoka, Japan, from patients between 2 months and 12 years of age. Otoscopy images were captured with an endoscope, and 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 based on signs, symptoms, patient history, otoscopy examination, and WBT measurements. The data was 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 all participants or their parent or guardian.

The otoscopy images are of size  $640 \times 480$  pixels but are cropped to a square, as the sides are mostly black, and downsampled to  $299 \times 299$ . The WBT measurements are not necessarily uniformly sampled in regard to 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 computation 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 [18]

$$D = 1 - \mu_{\text{correct}} \cdot \mu_{\text{certainty}}. \quad (1)$$

More details on the human inter-rater study with the four ENTs can be found in Sundgaard et al. [22].

Due to the limited number of cases in the dataset, all experiments were performed with 5-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.

### III. Results

Figure 3 shows the embeddings of the train and test data generated for one of the cross-validation folds for the image + WBT model in two dimensions using t-SNE dimensionality reduction [23]. From these embeddings, classification and difficulty estimation were performed. Table 1 shows the performance in both tasks: otitis media classification and estimation of the diagnostic difficulty for all proposed models.

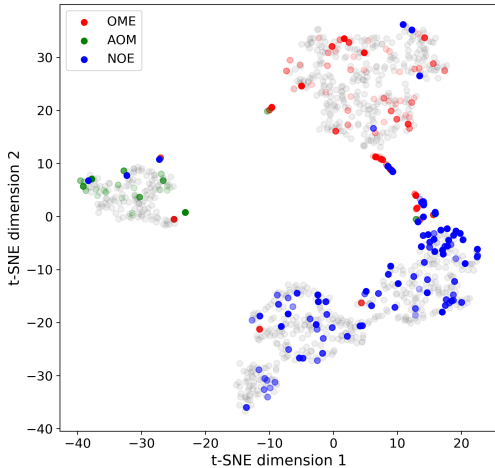


Figure 3. Visualization of embeddings. The transparency of each point indicates the ground truth difficulty with very transparent being the easiest. Grey points are the training samples.

For classification, both accuracy and class-wise F1-scores are computed. The F1-score is the harmonic mean of the precision and recall. Kendall rank correlation coefficient [24], also called Kendall’s  $\tau$ , was used to evaluate the estimation of difficulty. 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. The stated performance values in the table are the

average performances across the five cross-validation folds.

Table 2 shows 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 image + WBT model, as both accuracy and each class-wise F1-score is superior to the other methods. It is also clear 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 72.6% of the cases correctly. The following results are generated using the image + WBT embedding model and the supervised prediction model. Figure 4 shows a scatter plot of ground truth difficulties versus estimated difficulties.

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 0.480, corresponding to 74.0% correctly ranked cases, while for misclassified cases it is 0.163, corresponding to only 58.2% 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.

Table 1. Performance of the proposed models: Accuracy (Acc.) and F-1 score of the classification task and Kendall’s  $\tau$  of the estimated diagnostic difficulty. Each performance metric is the average across all five cross-validation folds. Bold font marks the highest performance in each column.

Method	Acc. [%]	F1-score			Difficulty $\tau$
		OME	AOM	NOE	
Image multi-task	85 $\pm$ 4	0.82 $\pm$ 0.05	0.78 $\pm$ 0.05	0.88 $\pm$ 0.03	0.39 $\pm$ 0.03
Image embedding	85 $\pm$ 3	0.83 $\pm$ 0.04	0.77 $\pm$ 0.02	<b>0.90 <math>\pm</math> 0.02</b>	0.43 $\pm$ 0.01
WBT multi-task	74 $\pm$ 2	0.69 $\pm$ 0.04	0.53 $\pm$ 0.05	0.87 $\pm$ 0.03	0.42 $\pm$ 0.03
WBT embedding	68 $\pm$ 3	0.51 $\pm$ 0.10	0.51 $\pm$ 0.03	0.87 $\pm$ 0.03	0.36 $\pm$ 0.07
Image + WBT multi-task	85 $\pm$ 4	0.83 $\pm$ 0.05	0.77 $\pm$ 0.05	<b>0.90 <math>\pm</math> 0.03</b>	0.40 $\pm$ 0.02
Image + WBT embedding	<b>86 <math>\pm</math> 2</b>	<b>0.84 <math>\pm</math> 0.04</b>	<b>0.82 <math>\pm</math> 0.04</b>	<b>0.90 <math>\pm</math> 0.02</b>	<b>0.45 <math>\pm</math> 0.02</b>

Table 2. Confusion matrices for the three embedding models: image, WBT, and image + WBT.

Target \ Pred.	Image			WBT			Image + WBT		
	OME	AOM	NOE	OME	AOM	NOE	OME	AOM	NOE
OME	<b>306</b>	24	45	<b>162</b>	175	38	<b>307</b>	18	50
AOM	23	<b>117</b>	15	25	<b>121</b>	9	21	<b>124</b>	10
NOE	34	7	<b>443</b>	56	20	<b>408</b>	31	7	<b>446</b>
Total	363	148	503	243	316	455	359	149	506

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 versus 0.23 for AOM, respectively, and 0.36 versus 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 62.3% and 85.1% for mild and severe cases, respectively, while for OME it is 73.2% and 92.0%, respectively.

#### IV. 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 1 shows that, in addition to the overall superior perfor-

mance, the combined embedding network manages to improve the classification of AOM, from a 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 was also previously shown by Sundgaard et al. [1].

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 [7, 25]. 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 as distinguishing between OME and AOM is crucial in deciding whether or not 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 ENTs, as indicated by the higher ground truth diagnostic difficulty. It shows that the mild symptoms are not 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 have to reach a certain severity or intensity before the model can detect otitis media.

These results show that the multi-modal model performs better for both classification and difficulty

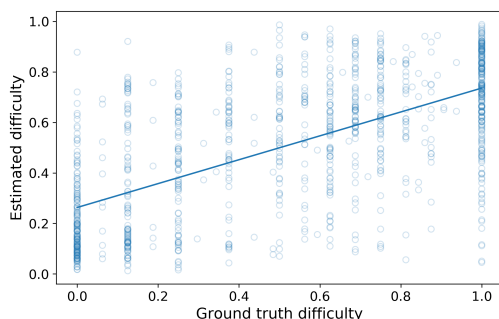


Figure 4. Scatter plot of ground truth difficulties and difficulties estimated with the supervised approach, together with the least-squares regression line.



estimation compared to the models trained on the modalities separately. The four ENTs in the human inter-rater study [22] achieved a 64.0% accuracy on this dataset based on the same amount of patient information used in the multi-modal embedding model, which achieved 86.5%. 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.

## V. Acknowledgements

This study was financially supported by the William Demant Foundation.

## References

- [1] Sundgaard, J. V., Harte, J., Bray, P., Laugesen, S., Kamide, Y., Tanaka, C., Paulsen, R. R., and Christensen, A. N., "Deep metric learning for otitis media classification," *Medical Image Analysis*, Vol. 71, 2021. doi: 10.1016/j.media.2021.102034.
- [2] Senaras, C., Moberly, A. C., Teknos, T., Essig, G., Elmaraghy, C., Taj-Schaal, N., Yua, L., and Gurcan, M. N., "Detection of eardrum abnormalities using ensemble deep learning approaches," *Medical Imaging 2018: Computer-Aided Diagnosis*, Vol. 10575, 2018, pp. 295–300. doi: 10.1117/12.2293297.
- [3] Shie, C. K., Chang, H. T., Fan, F. C., Chen, C. J., Fang, T. Y., and Wang, P. C., "A hybrid feature-based segmentation and classification system for the computer aided self-diagnosis of otitis media," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2014, pp. 4655–4658. doi: 10.1109/EMBC.2014.6944662.
- [4] Wu, Z., Lin, Z., Li, L., Pan, H., Chen, G., Fu, Y., and Qiu, Q., "Deep Learning for Classification of Pediatric Otitis Media," *Laryngoscope*, Vol. 131, No. 7, 2021, pp. E2344–E2351. doi: 10.1002/lary.29302.
- [5] Grais, E. M., Wang, X., Wang, J., Zhao, F., Jiang, W., and Cai, Y., "Analysing wideband absorbance immittance in normal and ears with otitis media with effusion using machine learning," *Scientific Reports*, Vol. 11, No. 1, 2021, pp. 1–12. doi: 10.1038/s41598-021-89588-4.
- [6] Terzi, S., Özgür, A., Erdivanlı, Coşkun, Z., Ogurlu, M., Demirci, M., and Dursun, E., "Diagnostic value of the wideband acoustic absorbance test in middle-ear effusion," *Journal of Laryngology and Otolaryngology*, Vol. 129, No. 11, 2015, pp. 1078–1084. doi: 10.1017/S0022215115002339.
- [7] Sundgaard, J. V., Bray, P., Laugesen, S., Harte, J., Kamide, Y., Tanaka, C., Christensen, A. N., and Paulsen, R. R., "A deep learning approach for detecting otitis media from wideband tympanometry measurements," *IEEE Journal of Biomedical and Health Informatics*, 2022. doi: 10.1109/JBHI.2022.3159263.
- [8] Monroy, G. L., Won, J., Dsouza, R., Pande, P., Hill, M. C., Porter, R. G., Novak, M. A., Spillman, D. R., and Boppart, S. A., "Automated classification platform for the identification of otitis media using optical coherence tomography," *NPJ Digital Medicine*, Vol. 2, No. 1, 2019, pp. 1–11. doi: 10.1038/s41746-019-0094-0.
- [9] Wang, Y. M., Li, Y., Cheng, Y. S., He, Z. Y., Yang, J. M., Xu, J. H., Chi, Z. C., Chi, F. L., and Ren, D. D., "Deep Learning in Automated Region Proposal and Diagnosis of Chronic Otitis Media Based on Computed Tomography," *Ear and Hearing*, Vol. 0, 2020, pp. 669–677. doi: 10.1097/AUD.0000000000000794.
- [10] Kuruvilla, A., Shaikh, N., Hoberman, A., and Kovačević, J., "Automated diagnosis of otitis media: Vocabulary and grammar," *International Journal of Biomedical Imaging*, 2013. doi: 10.1155/2013/327515.
- [11] Mironica, I., Vertan, C., and Gheorghe, D. C., "Automatic pediatric otitis detection by classification of global image features," *2011 E-Health and Bioengineering Conference (EHB)*, IEEE, 2011, pp. 1–4.
- [12] Khan, M. A., Kwon, S., Choo, J., Hong, S. M., Kang, S. H., Park, I. H., Kim, S. K., and Hong, S. J., "Automatic detection of tympanic membrane and middle ear infection from oto-endoscopic images via convolutional neural networks," *Neural Networks*, Vol. 126, 2020, pp. 384–394. doi: 10.1016/j.neunet.2020.03.023.
- [13] Cha, D., Pae, C., Seong, S. B., Choi, J. Y., and Park, H. J., "Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database," *EBioMedicine*, Vol. 45, 2019, pp. 606–614. doi: 10.1016/j.ebiom.2019.06.050.
- [14] Binol, H., Moberly, A. C., Niazi, M. K. K., Essig, G., Shah, J., Elmaraghy, C., Teknos, T., Taj-Schaal, N., Yu, L., and Gurcan, M. N., "Decision fusion on image analysis and tympanometry to detect eardrum abnormalities," *Medical Imaging 2020: Computer-Aided Diagnosis*, Vol. 11314, 2020, pp. 375–382. doi: 10.1117/12.2549394.
- [15] Harris, P. K., Hutchinson, K. M., and Moravec, J., "The use of tympanometry and pneumatic otoscopy for predicting middle ear disease," *American Journal of Audiology*, Vol. 14, No. 1, 2005, pp. 3–13. doi: 10.1044/1059-0889(2005/002).
- [16] Habib, A., Kajbafzadeh, M., Hasan, Z., Wong, E., Gunasekera, H., Perry, C., Sacks, R., Kumar, A.,

- and Singh, N., “Artificial intelligence to classify ear disease from otoscopy: A systematic review and meta-analysis,” *Clinical Otolaryngology*, 2022. doi: 10.1111/coa.13925.
- [17] Lee, J. Y., Choi, S. H., and Chung, J. W., “Automated classification of the tympanic membrane using a convolutional neural network,” *Applied Sciences*, Vol. 9, No. 9, 2019. doi: 10.3390/app9091827.
- [18] Hannemose, M. R., Sundgaard, J. V., Ternov, N. K., Paulsen, R. R., and Christensen, A. N., “Was that so hard? Estimating human classification difficulty,” *arXiv preprint arXiv:2203.11824*, 2022, pp. 1–10.
- [19] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z., “Rethinking the Inception Architecture for Computer Vision,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.
- [20] Wang, X., Han, X., Huang, W., Dong, D., and Scott, M. R., “Multi-similarity loss with general pair weighting for deep metric learning,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5017–5025. doi: 10.1109/CVPR.2019.00516.
- [21] Geurts, P., Ernst, D., and Wehenkel, L., “Extremely randomized trees,” *Machine learning*, Vol. 63, 2006, pp. 3–42. doi: 10.1007/s10994-006-6226-1.
- [22] Sundgaard, J. V., Värendh, M., Nordström, F., Kamide, Y., Tanaka, C., Harte, J., Paulsen, R. R., Christensen, A. N., Bray, P., and Laugesen, S., “Inter-rater reliability of the diagnosis of otitis media based on otoscopic images and wideband tympanometry measurements,” *International Journal of Pediatric Otorhinolaryngology*, Vol. 153, 2022, p. 111034. doi: 10.1016/j.ijporl.2021.111034.
- [23] Van Der Maaten, L., and Hinton, G., “Visualizing data using t-SNE,” *Journal of Machine Learning Research*, Vol. 9, No. 11, 2008.
- [24] Kendall, M. G., “Rank correlation methods,” 1948.
- [25] Helenius, K. K., Laine, M. K., Tähtinen, P. A., Lahti, E., and Ruohola, A., “Tympanometry in discrimination of otoscopic diagnoses in young ambulatory children,” *Pediatric Infectious Disease Journal*, Vol. 31, No. 10, 2012, pp. 1003–1006. doi: 10.1097/INF.0b013e31825cac94.





