Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Deep metric learning for otitis media classification

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

[a] *Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby, Denmark*
[b] *Interacoustics Research Unit, c/o Technical University of Denmark, Lyngby, Denmark*
[c] *DGS Diagnostics, Smørum, Denmark*
[d] *Kamide ENT clinic, Shizuoka, Japan*
[e] *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.

1. Introduction

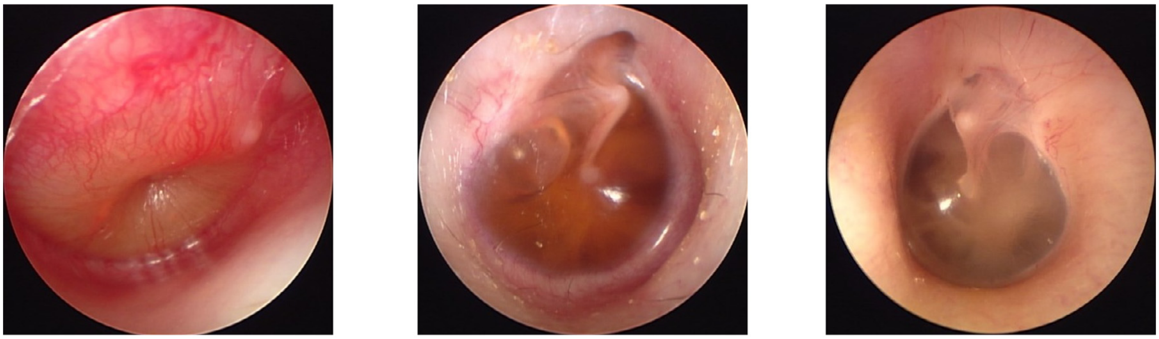
Otitis media is a group of diseases in the middle ear, which can be divided into two major diagnostic groups: acute otitis media (AOM) and otitis media with effusion (OME). Each year, around 11% of the world's population suffer from AOM (Monasta et al., 2012), and it is the second most common reason 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 (J.V. Sundgaard).
[1] Shared senior authorship

https://doi.org/10.1016/j.media.2021.102034
1361-8415/



(a) AOM

(b) OME

(c) NOE

Fig. 1. Otoscopy images of tympanic membrane with acute otisis 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 (Cullas Ilarslan 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 (Célind 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 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. 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%.

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 Interacoustics A/S.

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, naòve 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 Euclidian 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):

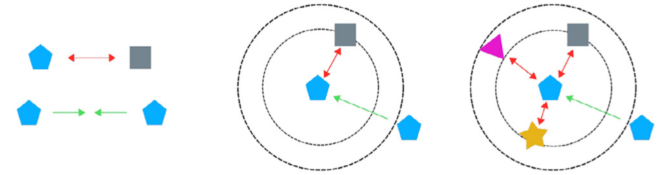
$$\begin{aligned} L_{\mathrm{c}}(x_{1,i}, x_{2,i}) = \sum_{i=1}^{N} [&(1 - y_i)||f_{1,i} - f_{2,i}||_2 \\ &+ (y_i)\{\max(0, m - ||f_{1,i} - f_{2,i}||_2)\}^2], \end{aligned} \tag{1}$$

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):

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

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.



(a) Contrastive loss (b) Triplet loss (c) Multi-class N-pair loss

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):

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

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 pretrained 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 ± standard deviation) of the neural networks trained with five different loss functions.

	AOM		OME		NOE		Acc. [%]	
	Recall	Precision	Recall	Precision	Recall	Precision	Training	Test
CE	74 ± 9	67 ± 10	82 ± 10	85 ± 3	89 ± 3	90 ± 4	89 ± 2	85 ± 2
Class-weighted CE	72 ± 8	72 ± 9	84 ± 4	82 ± 4	87 ± 3	89 ± 2	94 ± 2	84 ± 2
Contrastive	50 ± 9	76 ± 13	78 ± 4	84 ± 5	94 ± 3	84 ± 3	99 ± 0	84 ± 3
Triplet	61 ± 8	82 ± 6	86 ± 5	84 ± 3	92 ± 3	89 ± 3	98 ± 1	86 ± 1
Multi-class	58 ± 8	74 ± 8	87 ± 5	79 ± 5	87 ± 3	89 ± 4	91 ± 2	84 ± 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 ± 10 by the cross-entropy loss, to 72 ± 9 by the class-weighted cross-entropy loss and then again to 82 ± 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.

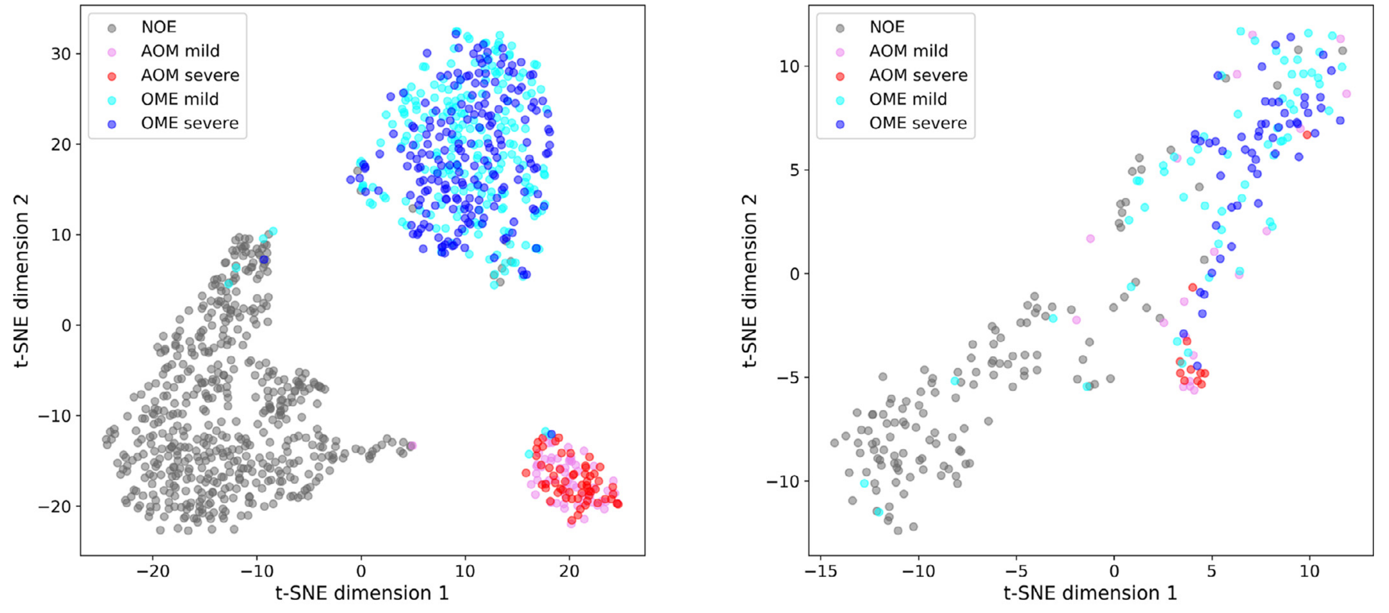
Table 2
Test accuracy of setting and hyperparameter variations of the triplet loss neural network. Proposed approach is the neural network trained with triplet loss function, with $k = 25$, embedding dimensions 32, with data augmentation, and trained with five-fold cross validation (CV).

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

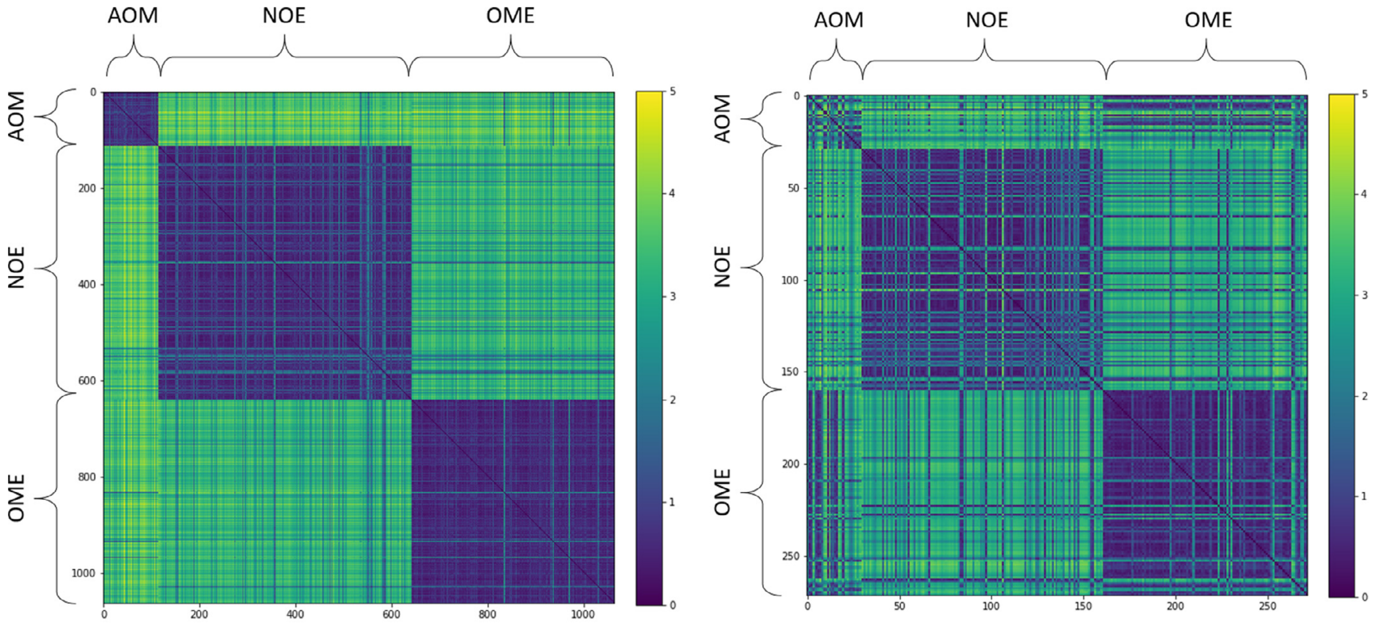
Fig. 5 shows the pairwise standard Euclidean distance of the 32 dimensional embedding vectors for the same train/test split, as in Fig. 4. Fig. 5(a) shows three clear training-set groups, while there are still images with smaller distance to images in another diagnostic group, especially between NOE 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 subgroups, where one of them appears more like OME. Furthermore, a sub-group of OME images has smaller distance to NOE than any other class.



(a) Train embeddings

(b) Test embeddings

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.



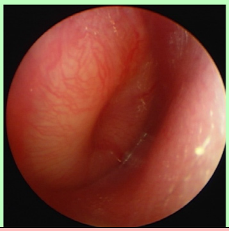
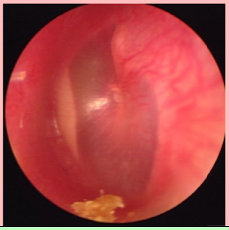
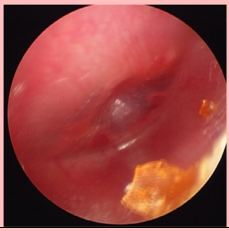
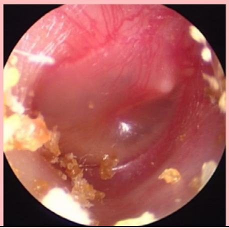
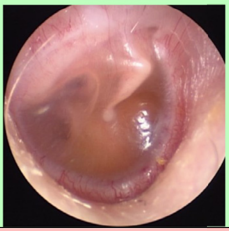
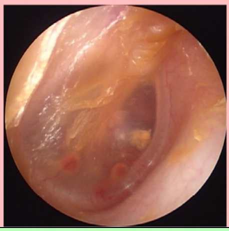
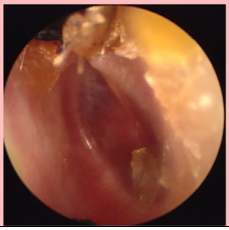
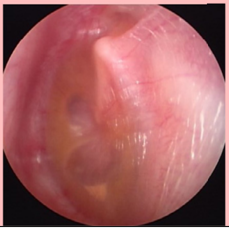
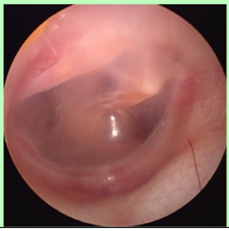
(a) Train distances

(b) Test distances

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 misclassified 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 tradeoff 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

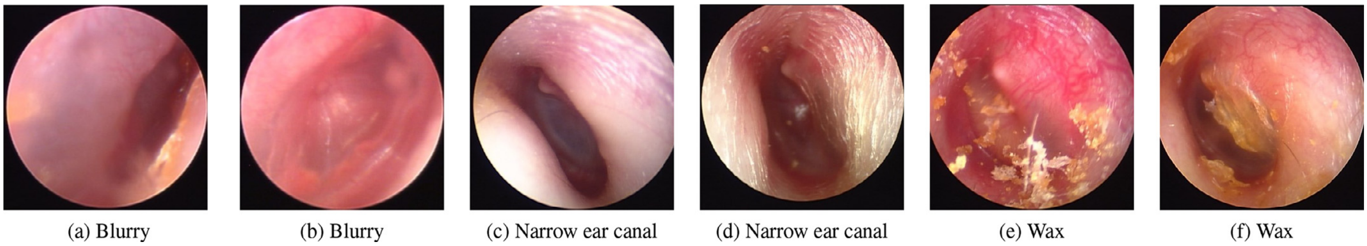


Fig. 6. Examples of quality variations of clinical otoscopy images.

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 that 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/cmg505.

Célind, 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.

Kuruvilla, 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.