UNIVERSITY OF OULU

ARTIFICIAL INTELLIGENCE IN STUDYING AND EVALUATION OF OTITIS MEDIA BY ACOUSTIC REFLECTOMETRY

Iikka Nissi

Master's Thesis

Medical and Wellness Technology

Faculty of Medicine

University of Oulu

2022

Nissi Iikka (2022), Faculty of Medicine, University of Oulu, Master's thesis, 52 pages.

Abstract

Objective:

Acute otitis media (AOM) is usually associated with upper respiratory tract infections and common colds, but many times it can last longer than the initial symptoms. An acoustic reflectometry device can be used to objectify the diagnostic process. The purpose of the study was to train the neural network to identify ears with symptoms of AOM using the acoustic response of the device.

Methods:

An acoustic reflectometry sample of 53 ears from 39 patients was collected during laryngoscopy operation from patients with recurrent ear infections. In addition to the acoustic samples, the doctor determined whether ear had visual signs of otitis media (OM) and whether there was effusion in it. These three parameters were used in the construction of feedforward neural network. Two neural network layouts were selected, one with samples of effusion-only sick ears and the other with sick ears based on other visual indications of OM, independent of effusion.

Results:

The sensitivity and specificity of the trained networks were about 90%. Two different groupings of samples clearly showed that diseased ears without effusion could be identified as sick with sensitivity of 80-90%, when similar ears were included in the category of sick ears. Network with sick ears with effusion as training material had a sensitivity of 20-30% identifying sick ears without effusion. The inclusion of both types of sick ears in single network caused slight drop in sensitivity and specificity compared to just one type.

Conclusion:

Acoustic reflectometry can detect more than just standard cases of acute otitis media, in which effusion typically occurs. An accurate neural network for identifying sick ears without effusion can be achieved with a relatively small sample size. This indicates a possibility of conducting an in-depth analysis of other diseases within the OM group or the transition between these diseases.

Keywords: Artificial intelligence; Acoustic reflectometry; Otitis media; Otoscopy

Nissi Iikka (2022), Lääketieteellinen tiedekunta, Oulun yliopisto, Pro gradu –tutkielma, 52 sivua.

Tiivistelmä

Työtarkoitus:

Akuutti otitis media (AOM) yhdistetään yleensä ylempien hengitysteiden tulehduksiin ja nuhaan, mutta monesti otitis media (OM) oireet voivat kestää tulehdustilaa tai nuhaa pidempään. Akustista reflektometriaa käyttävän laitteen avulla diagnosointi prosessia voidaan tarkastella objektiivisesti. Työn tarkoitus oli opettaa neuroverkko, mikä tunnistaa AOM-oireisen korvan akustisen reflektometrin akustisesta mittauksesta.

Menetelmät:

Laryngoskopiaoperaation aikana kerättiin akustisella reflektometrialla otos 53 korvasta. Operaatio suoritettiin 39 potilaalle, joilla oli uusiutuvia korvatulehduksia. Akustisten näytteiden lisäksi operaation aikana lääkäri määritti visuaaliset OM-merkit ja eritteen määrän. Näitä kolmea tietoa käytettiin eteenpäin kytkeytyvän neuroverkon rakentamiseen. Kaksi neuroverkkoa rakennettiin, joissa ensimmäisessä oli pelkästään eritettä sisältävät korvat, ja toisessa kaikki visuaalisesti OM-merkit täyttävät korvat, mukaan lukien eritettä sisältävät korvat.

Tulokset:

Opetettujen neuroverkkojen sensitiivisyys ja spesifisyys olivat 90 % luokkaa. Kahteen ryhmään jaettu aineisto osoitti, että sairaat eritteettömät korvat voidaan tunnistaa sairaiksi 80–90 % sensitiivisyydellä, kun neuroverkolle opetetaan sekä eritteiset että eritteettömät sairaat korvat. Pelkästään eritteisiä korvia sisältävä neuroverkko tunnisti eritteettömät sairaat korvat 20–30 % sensitiivisyydellä. Eritteisten ja eritteettömien korvien käyttö samassa neuroverkossa laski sensitiivisyyttä ja spesifisyyttä verrattuna pelkkien eritteisten käyttöön.

Johtopäätökset:

Akustinen reflektometria voi tunnistaa muitakin tiloja kuin tyypillisen eritteisen akuutin otitis median. Pienellä näytemäärällä voidaan saavuttaa tarkka neuroverkko, mikä tunnistaa otitis median ilman eritteen läsnäoloa. Tämä viittaa mahdollisuuteen, että syvällisellä analyysillä voidaan saada lisää tietoa taudin etenemisestä tai taudin muista tiloista.

Avainsanat: Tekoäly; Akustinen reflektometria; Otitis media; Otoskopia

Foreword

Thanks to Otometri founder and the original inventor of the product Manne Hannula, whose work with acoustic reflectometry and neural networks this study is based on.

Thanks also to

Tuomas Holma for providing his expertise on ear related physiology and otology.

Esko Alasaarela and Timo Jämsä for the instructions and guidance during this master’s thesis.

Oulu, 12.1.2022

Iikka Nissi

Abbreviations used

AI	Artificial Intelligence
ANN	Artificial neural network
AOM	Acute otitis media
AR	Acoustic reflectometry
BP	Back-propagation
CC	Correlation coefficient
CNN	Convolutional neural network
FFT	Fast Fourier transform
FNN	Feed-forward neural network
LM	Levenberg-Marquardt
LPC	Linear Prediction Coefficients
MEE	Midlle ear effusion
MFCC	Mel Frequency Cepstrum Coefficients
NN	Neural network
OM	Otitis media
OME	Otitis media with effusion
PCA	Principal component analysis
purelin	Linear transfer function
RNN	Recurrent neural network
SG-AR	Spectral gradient acoustic reflectometry
tansig	Hyperbolic tangent sigmoid transfer function
TM	Tympanic membrane

Table of contents

1 Introduction

Otitis media (OM) means generally an inflammatory condition in the middle ear having various background factors as well as symptoms and courses of the disease.

Acute otitis media (AOM) is a common infectious disease particularly among young children all over the world (Teele, D. W., Klein, & Rosner, 1989). It is a bacterial or viral ear infection that appears usually in connection with an upper respiratory tract infection (Granath, 2017; Wiertsema & Leach, 2009). Its typical symptoms, ear pain and hearing impairment, cause the child suffering and the child’s illness strains the family in many ways.

The inflammatory process in the upper respiratory tract and the middle ear leads in swelling of mucous membranes and increased mucus secretion. As a result, the auditory tube (Eustachian tube) connecting the middle ear cavity and the nasopharynx becomes constricted or even completely blocked. Further, the middle ear ventilation and the mucus transporting function of the tube becomes insufficient. Resulting condition is called otitis media with effusion (OME). The accumulated mucus in the middle ear and the swollen eardrum causes so called conductive hearing impairment because the structures of the middle ear cannot normally conduct an ambient sound vibration to the inner ear. (Parmar et al., 2019; Wiertsema & Leach, 2009)

The spontaneous healing of AOM occurs usually in the period of 1-2 weeks. Anyhow, prolonged ear infections are not rare. Besides that, some children suffer from recurrent otitis media that cause need for frequent doctor visits.

After the acute period of otitis media there may be individually varying symptoms and signs of OM. The effusion may still stay in the middle ear albeit the ear pain is not existing anymore. During that kind of “silent” form of OM the child is prone to get a new AOM and he or she may have a hearing impairment which the parents cannot notice (Parmar et al., 2019). Older children can communicate and describe if they suffer from ear locks or pain, but younger children just become restless, touch a painful ear with their hands and cry. It is difficult to detect of otitis media at home. It is also challenging for parents to decide when a doctor’s appointment is needed.

At the doctor's office, the primary method to examine an otitis media is to use an otoscope, which is an optical tubular device to help to see and evaluate the condition of the tympanic membrane (TM) under well-lit conditions. The freedom of movement of the TM can be tested by applying an air pressure pulse to the ear canal with a pneumatic otoscope (Abbott, Rosenkranz, Hu, Gunasekera, & Reath, 2014; Lous, Ryborg, Damsgaard, & Munck, 2012; Won et al., 2018). This method does not work when the TM cannot be seen due to wax or a narrow ear canal (Pichichero, 2000). Moreover, the examination may be uncomfortable or painful for the child, that may reduce the willingness to cooperate. So, the diagnostic of OAM is in many cases challenging and the result of examination remains uncertain.

Another method for detecting the presence of fluid is acoustic reflectometry (AR) (Kimball, 1998; Pichichero, 2000; Puhakka, 2014; Teele, David W. & Teele, 1984), which does not require as much cooperation with struggling children as using otoscope (Barnett et al., 1998; Chianese et al., 2007; Teppo & Revonta, 2009). Acoustic reflectometry, in simple terms, is a recording of reflection, which is produced by projecting sound towards TM. From that recorded reflection, it is possible to figure out many different parameters, e.g., spectral changes or delays in time of reflection. The use of various parameters can be utilized for different physical conditions that occur in otitis media cases. AR has been used mainly for the detection of effusion in previous solutions and studies.

This study is based on the method developed by Hannula et al., where acoustic reflectometry data were collected using a PC internal sound card via a USB port and analyzed using a generalized regression-based neural network (NN). Acoustic data was converted to spectral data by fast Fourier transform (FFT) and the neural network was trained using this information in combination with the amount of middle ear effusion (MEE) taken from the ear, scaled between 0 and 1. Previous studies focused on the incidence and amount of effusion observed in the ear (Hannula, Hinkula, Holma, Löfgren, & Sorri, 2009; Hannula, Holma, Löfgren, Hinkula, & Sorri, 2011). This study used new smartphone device, remade spectral conversion and focused on the detection of AOM, regardless of effusion.

2 Background

2.1 Otitis media

Otitis media is a general term to address middle ear inflammation, which is distinguished between different states using additional terms. In most OM cases mucus is present in middle ear after membranes swell and the extraction of mucus is limited. Acute OM is used when bacterial infection is present. OME is a non-infected state where the mucus is blocking the middle ear e.g., after virulent upper respiratory infection. Cases of AOM or OME are typically short term, but prolonged episodes can lead to more severe complications. (Mandel et al., 2016; Schilder et al., 2016)

The environmental factors increasing the risk of getting OM include age, genetical differences and socioeconomic status. The probability of inheriting OM susceptibility from parents is relatively high. (Schilder et al., 2016)

2.1.1 Eustachian tubes

One explaining factor for prolonged OM cases is the changes in the eustachian tube (ET) development. ET act as drainage and pressure equalizing for the middle ear. This enables the tympanic membrane to work freely without being forced inwards or outwards by the pressure difference in outer ear. When the efficiency of ET function is decreased, mucus is formed faster than it is drained out. This symptom can be caused natural physiological difference where the opening of tubes is not functioning. One significant difference in ET structure is between young children and adults. The position of ET for children is horizontal from middle ear to nasopharynx, whereas for adults ET is more vertical, close to 45 degrees (Schilder et al., 2016). Also, the ability to manipulate the opening of ET by jaw and muscle movements is much more effective for adults. (Bylander, 1980; Mandel et al., 2016)

2.2 Diagnostic methods

Hearing problems are many times present in OME cases, but it can’t be utilized when child is too young to communicate (Parmar et al., 2019; Teele, D. W., Stewart, Teele, Smith, & Tregonning, 1990). Definitive way to diagnose OME is myringotomy operation, but it is not feasible in check-ups at primary health care or general clinics. Clinical uses include pneumatic otoscope and tympanometer (Barnett et al., 1998; Laine, 2015). Less common method is acoustic reflectometer, which has not reached popularity in clinical use.

2.2.1 Otoscopes and tympanometer

AOM diagnostic in general practices can be limited to pneumatic otoscopes and tympanometers as non-pneumatic otoscopes don't provide enough information on MEE (Abbott et al., 2014; Laine, 2015). Microscopy is used in specialized environments. Pneumatic otoscope uses positive and negative air pressure to visually check TM for irregular mobility, which can indicate presence of different pathologies. Method is limited by the high level of competence required to provide diagnosis, which makes it hard to generalize between studies (Fagan, 2014).

Tympanometer can be described as pneumatic acoustic reflectometer, as the mechanism has similarities between both solutions. A negative to positive air pressure is varied inside ear canal and measuring sound is recorded during this change in pressure. Tympanogram is printed for every pressure value, resulting in graph with pressure on horizontal axis and compliance on vertical axis. Plotted curve should have distinguishable peak near 0 pressure for healthy ear. Different peak location and undisguisable peak indicate abnormal states. (Lous et al., 2012)

The most challenging part in standard otoscopic diagnostic method is identifying MEE; both pneumatic otoscope and tympanometer specialize in this (Abbott et al., 2014). Tympanometer is more likely to give standardized results as it is not as dependent on the expertise of users but is expensive compared to pneumatic otoscope. Pneumatic otoscope reliant diagnosis is more likely to mistake healthy ear as OME. Tympanometer sensitivity and specificity is in most studies between 70 and 95% but there was high cap as the lower value was in most cases around 50 and 60%. (Laine, 2015).

2.2.2 Acoustic reflectometry

AR utilizes the reflectance of soundwaves from surfaces to detect changes in the structure of point of interest. In the case of OM, TM is many times rigid and possibly unmovable from the mucus or pressure difference (Laine, 2015). The density change inside middle ear can also have effect. Reflectance is measured by recording the summarized value from both loudspeaker output sound and reflected sound from ear. Basic methodology of AR is to display the reflected energy profile of target in some form e.g., intensity of sound, time difference measurements or as power of frequency. Then analysis is based on the thresholds or profiles of known healthy and sick values. These thresholds and profiles are investigated in clinical trials and confirmed by other state-of-the-art methods.

2.2.2.1 Methods

Acoustic reflectometry was popularized by studies performed by Teele et. al. The method utilizes the nature of reflected and incident sound being out of phase when the distance from microphone to TM is quarter of sound wavelength. This creates a low point in measured sound intensity, which is then used to detect middle ear effusion. (Teele, David W. & Teele, 1984)

In spectral gradient acoustic reflectometry (SG-AR), a spectral slope, also called spectral gradient, is used to measure the physiological changes. A Fourier transform is used to show spectral information, which is then used to detect the slope location and angle. The steepness of this slope is the defining parameter in SG-AR. Commercial adaption of this method was made by EarCheck and earliest studies also mention EarCheck (Block, S. L. et al., 1998). Other mentions of method used are scarce and use different terms e.g., angle of the curve (Kemaloğlu, Sener, Beder, Bayazit, & Göksu, 1999). Newer studies are very much synonymous with EarCheck and SG-AR. Device measures the angle and lower value indicates higher risk of middle-ear effusion, which in many cases indicate OME (Barnett et al., 1998)

One more comparable method is by Hannula et al. (Hannula et al., 2009; Hannula et al., 2011), where frequencies of acoustic reflectometry are used as such to input to NN, without using any specific singular point to determine the result. Method is similar when compared to SG-AR, but the determining of result value is left to NN.

More complex audio processing and analysis is nonexistent in AR studies. This can be seen how already common methods such as Mel-Frequency Cepstral Coefficients (MFCC) have been in use by different areas of medicine (Bahoura & Ezzaidi, 2013; Balamurali et al., 2021). When considering the wide range of OM related symptoms and individual pathological categories and only diagnostic focus in recorded sound is the presence of MEE (Coleman, 2021; Schilder et al., 2016; Teele, D. W. et al., 1989).

Processing and audio analysis methods can be used to increase the number of features extracted from the reflected audio, aside from the physical dimension changes that can be made to recording device.

2.2.2.2 Diagnostic relevance

Many studies compare AR with tympanometry as it is the widely accepted standard for OM diagnosing. Study results are divided between tympanometry or AR being superior solution. Differences are produced by used methods and selected test criteria. Study comparison can be divided into sound intensity-based studies, SG-AR based studies and rest into different acoustic analysis methods.

Sound intensity-based results were the first studies utilizing AR. Teele's study showed 94.4% sensitivity and 79.2% specificity, after confirming results with tympanometer and acoustic admittance (Teele, David W. & Teele, 1984).

SG-AR method by EarCheck Ltd seems to be equal method compared to tympanometry in many studies, where definitive diagnosis was made with either pneumatic otoscopy, microscopy or surgery (Barnett et al., 1998; Block, S. L. et al., 1998; Block, Stan L., 1999; Kimball, 1998; Laine, 2015). This is a common comparison as they are considered clinically easy to use compared to pneumatic otoscopy which requires experienced operators (Fagan, 2014). Two studies slightly favoured EarCheck over tympanometer (Babb, Hilsinger, Korol, & Wilcox, 2004; Puhakka, 2014) and two studies vice versa (Chianese et al., 2007; Muderris et al., 2013). Aside from EarCheck, in older studies acoustic otoscopes have showed more promising results with SG-AR being equal or slightly better than tympanometer (Douniadakis, Nikolopoulos, Tsakanikos, Vassiliadis, & Apostolopoulos, 1993; Kemaloğlu et al., 1999). The dependent variable in each study seems to be the selected cut-off for the measured angle. SG-AR studies show that measurements between different angles can produce high accuracy while sacrificing the sensitivity or specificity (Chianese et al., 2007; Teppo, Revonta, Lindén, & Palmu, 2006; Teppo & Revonta, 2009). Angle values with balanced sensitivity and specificity produce in most cases between 70% and 80% (Chianese et al., 2007; Muderris et al., 2013). The best cut-off point varied between different studies and showed that EarCheck can have very high accuracy in single gradient level, but significantly biased towards negative accuracy at other levels. The high negative predictive values were proposed to be a good screening tool for clinic or home use as negative result was seldom false (Babb et al., 2004; Teppo et al., 2006; Teppo & Revonta, 2009).

Neural network results from Hannula were mostly from preliminary testing of fluid level detection (Hannula et al., 2009; Hannula, Hirvikoski, Hinkula, & Holma, 2009; Hannula et al., 2011). No clinical OME results were published. Results state that a correlation between neural network output and fluid level exists (Hannula et al., 2011).

2.3 Artificial Intelligence and machine learning

Artificial intelligence is machine using technology to operate like intellectual biological being while solving problems. Machine learning is this machine using premade algorithms to develop these problem-solving functions from existing data. Algorithm will make assumption of the data structure and then build the functions to try and match the input and output variables with minimal error. First step is the selection of algorithm that is suitable for the data, which can be binary, continuous numerical values or multiple category classification. When learning is done, these functions can be used to predict outcomes or operate different functions.

2.3.1 Input or feature selection

Point is to use most relevant values for building the AI system. This way there are no redundant or irrelevant data making the processing heavier with increased number of calculations. Correct feature selection can also lead to more accurate system.

2.3.1.1 Principal component analysis (PCA)

PCA is a dimensionality reduction method when handling large datasets. It is used to create new variables which have no correlation to the original data, while preserving the information the dataset has. Method principle is finding new linear functions that maximize the variance of dataset. This is done by solving eigenvector problems. (Jolliffe & Cadima, 2016)

2.3.1.2 Refief

Relief utilizes a filter-type feature selection. It is flexible method suitable for many different data characteristics. It calculates a score for each feature to determine the relevance. Scores for each input parameter is dependent to nearest neighbor relations. Relief cycles through datasets and calculates values for differences between neighbor datapoints. Scores are then updated based on these values. (Urbanowicz, Meeker, La Cava, Olson, & Moore, 2018)

2.3.1.3 Filtering data

After selecting NN model and functions suitable for the data, many times the data needs to be preprocessed. The NN in most cases is not flexible enough to remove datapoints, which have undesired impact on the NN. This problem is the scope in deep learning (Goodfellow, Bengio, & Courville, 2016). For machine learning either filtering or manual selection is needed for datapoints depending on the size of the data. Detecting and removing influential observations from the data will increase accuracy and reliability. Datapoints causing large deviations to the calculated network are called outliers. Outliers are random occurrences, which don't describe the observed phenomenon. These are most likely caused by data collection errors, or in the worst-case scenario, phenomena the testing equipment cannot understand, but still relevant physiological measurements outside the intended scope. (Singh & Upadhyaya, 2012)

2.3.2 System

Selecting the system components depends on the type of input values and desired output. Machine learning system is constructed by the algorithms selected and creating a structure matching that algorithm. Typically, machine learning algorithms can be divided into regression and classification methods. These two are not mutually exclusive and there are algorithms that can be used for both (Torgo & Gama, 1997).

2.3.2.1 Regression

Regression is used when making models to predict values. When a linear relationship exists between dependent and independent variable, dependent variable values can be predicted even when there is no existing data representing the exact independent variable value. One of the regression algorithm models is linear regression, where continuous numerical values are summarized and the linear relationship between dependent and independent variable is calculated. The relationship is represented by equation that leaves minimal error when these two variables are used in the equation as input and output. Simple linear regression model can be built for predicting weight when e.g., height is known. System with more than one independent variable is called multiple linear regression. (Osborne, 2021)

2.3.2.2 Classification

Classification methods can be simplified to finding borders from dataset to either find labels, which can be used to define limits for a class or divide dataset into wanted number of classes using known labels. Understanding and predicting the labels is important. This can again be expanded to binary or multi-class systems. In case of OM binary classifier could be OME and healthy, and multi-class could be OME, AOM and healthy. The selection of input parameters and the feasibility of distinguishing the difference between very similar OM conditions is what must be understood when making such classification system. (Kotsiantis, Zaharakis, & Pintelas, 2006)

Example of classification algorithm would be decision tree, where each node has defined number of possible outcomes depending on the feature that is observed in that node. Decision trees are mainly classification algorithms but can be adapted easily to different tasks, e.g., as regression trees, when the node functions are defined as numerical thresholds (Yang, Liu, Tsoka, & Papageorgiou, 2017).

2.3.2.3 Artificial neural network

The difference in structure of nodes and connections are very different in each of these systems. Simple linear regression is defined as formulas depicting the correlation between dependent and independent variable. Decision trees are flowcharts with different feature comparisons and thresholds in each step of chart, which leads to classification in the bottom of the chart. In more complex cases the number of variables and features increase, which can lead to redundancies. When handling many features, artificial neural network (ANN) is more feasible.

ANN is a collection of many neurons that can be layered to form multiple levels for different computational functions. The main parts are the input layer, hidden layers, and output layer. Example of neural network model in Fig. 1. Most common neural network types are feed-forward neural network (FFN), convolutional neural network (CNN) and recurrent neural network (RNN). ANN is the basic term for every NN, and it is many times used in place of the FFN, because it was the first NN model and most common. FFN operates in linear manner from input layer towards output layer. CNN can skip neurons in layers by focusing on distinct properties in the input data, e.g., filtering images.

RNN stores data in layers and uses that stored data to make predictions of sequential or time series data.

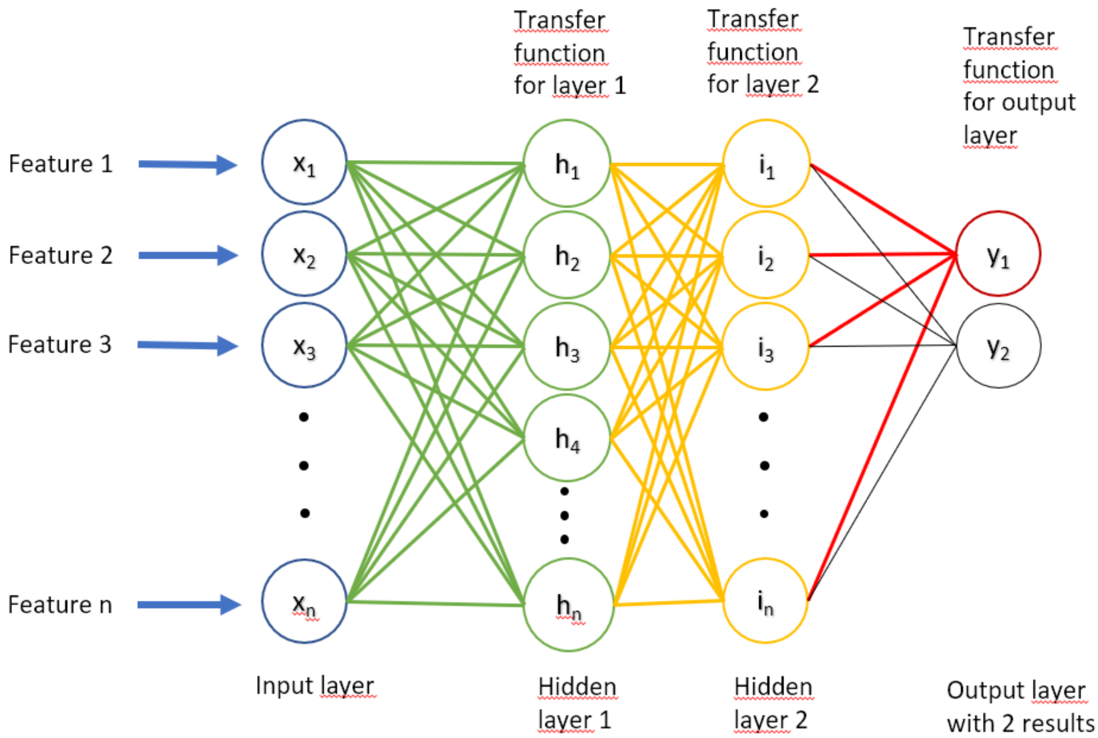


Fig. 1. ANN diagram

Figure shows the network structure with layers where the lines between represent the calculated weights of the transfer function inputs. Every neuron in layer is connected to all neurons on the next layer.

Training a FNN is done by finding suitable weights between each neuron. The weights in case of FNNs can be found with back-propagation (BP). BP means calculating the gradients of network layers to minimize the loss occurring with current weights set for FNN (Goodfellow et al., 2016). This calculation is performed by checking the output error from current network and propagating the error back to the network layers (Gavin, 2013). Transfer functions are not typically modified during BP in NNs (Tutunji, 2009). Transfer function selection and layer sizing are crucial parts, since they are static during BP. Once the NN is trained, hidden layers combined with weights between neurons should produce close to desired results with minimal loss.

Levenberg-Marquardt (LM) algorithm is the tool utilized in processing the gradients, which are calculated during BP. LM algorithm is used to find solutions to non-linear problems. The algorithm operates by minimizing the square error between the matching function and the input. This is the so-called least squares method. LM is a combination of two different algorithms, gradient descent algorithm and Gauss-Newton algorithm. LM balances between these two methods by changing the damping factor after checking whether approximation is getting better or worse after each iteration. (Gavin, 2013)

2.3.3 Validation

Last step in AI model is the validation. When a NN is constructed, training algorithms can lead to different answers caused by the random initial weight selection (Cao, Wang, Ming, & Gao, 2017). Results from validation are more reliable when error is minimal in subsequent trainings with random initial weights. Validation is many times included in the training function or it can be performed manually by leaving part of the data out of the training dataset. The validation data is input as data to the constructed model and the model output is compared with the actual values. This provides error, which can be used in the evaluation of the model. The selection of validation data can be done by rotating the data randomly or applying different cross-validation methods (Vabalas, Gowen, Poliakoff, & Casson, 2019). It is better to have as many iterations as possible. In large datasets this might not be as clear, but in smaller sample sizes a single abnormal datapoint can cause large deviations. The biggest problem in uneven distribution is overfitting of model (Ying, 2019). Overfitted model will only give good results for the phenomenon that is most represented in the training data and will perform poorly with new data. With smaller dataset underfitting can occur, which has bad performance with both training and new data.

2.3.3.1 Validity of data for generalization

In machine learning models the generalization is one of the main points. It is expected that model can solve unseen cases from the same problem category, not just the data input during training. This is reached by having large dataset with enough diversity. Large size means that it is more probable that both training and validation datasets are diverse. Cross validation methods help with the diversity and distribution. Over- and underfitting can also happen with selected machine learning algorithms when number of features is too large. This leads back to dimensionality reduction. (Ying, 2019)

The validity of model can be expressed in the amount of error or correlation. In medicine the standard way is to describe the results with true and false basis. The range is binary, and the model gives either positive or negative result, which is either true or false. This is given as sensitivity, correct positive readings divided by all positive cases, and specificity, correct negative readings divided by all negative cases.

3 Objective of the study

The objective of the study is to collect clinical data using a next-generation mobile device to verify and investigate the accuracy provided by AR data for use in diagnosing otitis media, and not just to detect effusion behind the tympanic membrane (Hannula et al., 2011). The study consisted of data collection in the Department of Otorhinolaryngology, Oulu University Hospital, and modelling a system for detecting sick ears.

4 Material and methods

The data was gathered from the patients who were scheduled for a myringotomy operation providing an accurate assessment of ear status by professionals with state-of-the-art devices. A total of 39 patients were studied, completing 75 ears altogether. Data of 22 ears were abandoned, and 53 accepted. The abandoned ears were classified as 9 already tubed TMs, 2 perforated TMs, 3 too much environmental noise and 8 device malfunctions. The accepted ears were classified as 28 healthy and 25 sick, of which there were 19 cases with effusion and 6 cases with minimal effusion. Those 6 cases had clear visual signs of AOM but could not be clearly classified as AOM like the other 19 cases. Unclear cases were handled as own group and taken into consideration in the study.

Data collected from each ear included a binary diagnosis of otitis media (healthy - sick), the amount of effusion as rough estimates using the number of + signs, three AR measurements, and three additional measurements if there was wax in the ear and had to be cleaned. The data also included notes from the measurement staff describing the original condition of the ears, for example, whether the TM was already tubed, or the outer ear canal was completely filled with wax.

Acoustic reflectometry data were collected using a prototype measuring device developed by Otometri Ltd which was connected to the Huawei smartphone model Y5 II as shown in Fig. 2. The system consists of end-user software running on the smartphone, determined by a special calibration protocol for collecting AR data in a clinical setting. The device is connected to the smartphone via a 3.5 mm aux connector. The device has a speaker and a microphone that are controlled by the wired headset software of the Android smartphone.

Fig. 2. a) An Otometri measuring setup, including a measuring device connected to a smartphone and b) operational block diagram of the setup.

A measurement sound (chirp) is played and recorded simultaneously for about one second. Chirp used to measure acoustic reflectivity was a series of different frequencies in ascending order from 1.5 kHz to 2.5 kHz at 20 Hz transitions. This pattern was repeated twice to improve error handling and improve the time fitting. Envelope of the chirp is seen in Fig. 3.

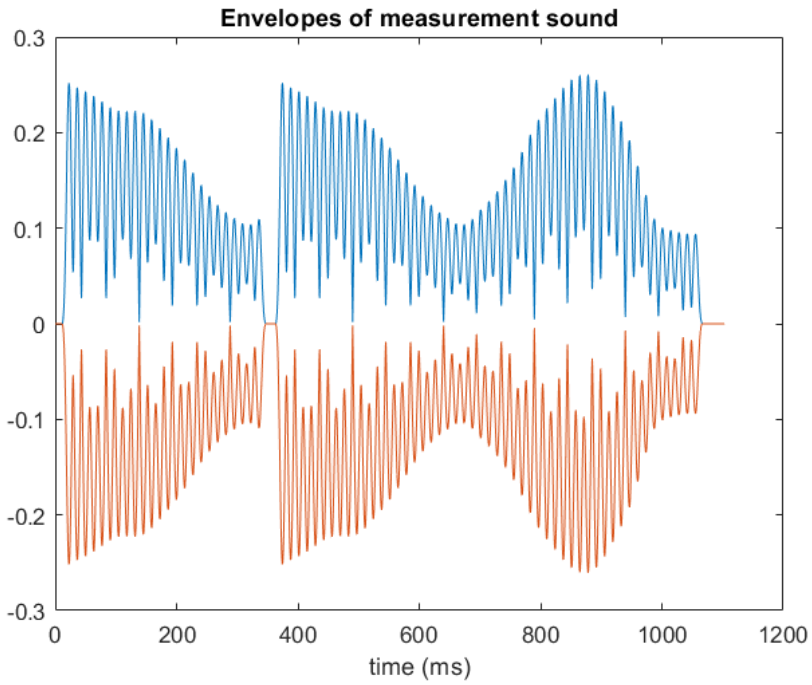


Fig. 3. Chirp envelope in time domain

The device was calibrated for the smartphone using the measurement against free air as a reference. The measurement result is used to adjust the sound amplitude levels at each frequency to match the desired profile. The profile was defined as an input signal whose amplitudes were half at each frequency compared to the transmitted sound. The corrected audio file is downloaded to the smartphone, and this process is repeated until the sum of the errors in each frequency band is at an acceptable level. This meant that if the total sum of frequency deviation or single frequency deviation was more than 1%, the calibration was not acceptable. This final air measurement result is part of the input parameters used for the neural network. The approval process is performed each time an application is launched for new measurements.

In the smartphone application interface, there are two selections: # 15 = sick and # 85 = healthy ear. Once the mode is selected, measurements are performed simply by placing the tip of the device at the entrance of the ear canal and pointing it towards TM. The Start button starts the measurements. The application has a customized mode of operation for the collection of clinical data, where 3 identical recordings are made at the start of the recording and sent to the device server. After the measurement, the researchers fill in additional information in the prepared documents.

The acoustic data is decomposed over the frequency range with the original input frequencies between 1.5 and 2.5 kHz in 20 Hz steps. This can be seen in the time domain envelope of the chirp in Fig. 3, which shows different amplitudes for the different frequencies.

4.1 Models

Study used MATLAB's (MathWorks Inc) back-propagation network training function trainlm (Demuth & Beale, 2001), which utilizes the LM algorithm to find the best fit for neural network weights and bias values. Algorithm utilizes least squared errors in nonlinear curve fitting. Each training loop was an individual process, and no data was stored for subsequent loops. A preliminary test between classification and regression models was conducted to select the method to be used in this study. The original regression-based model was compared to a new approach using classification with more advanced input parameters.

4.1.1 Classification method

In classification approach a Wiener filter was used to remove background noise (Plapous, Marro, & Scalart, 2006). 68 acoustic features and channel features were extracted from the recordings, namely MFCCs and differential MFCCs (Mel Frequency Cepstrum Coefficients), energy, Linear Prediction Coefficients (LPC), pitch, formant, maximum of frequency domain channel response, mean of frequency domain channel response, cumulative sum of edge of frequency domain channel response, cumulative sum of center of frequency domain channel response, maximum of histogram of frequency domain channel response, and length of histogram of frequency domain channel response. Frequency domain channel was calculated as 2048-point FFT. The cumulative sum of edge of frequency domain channel response is the sum of $H(f)$ from f=0 till f=2048/3 whereas the cumulative sum of center of frequency domain channel response is the sum of $H(f)$ from f=2048/3 till f=2048*2/3. The length of histogram of frequency domain channel response is the number of histograms. Other features extracted are widely used audio processing features. Example of healthy is shown ear in Fig. 4, where sum borders are displayed as red lines.

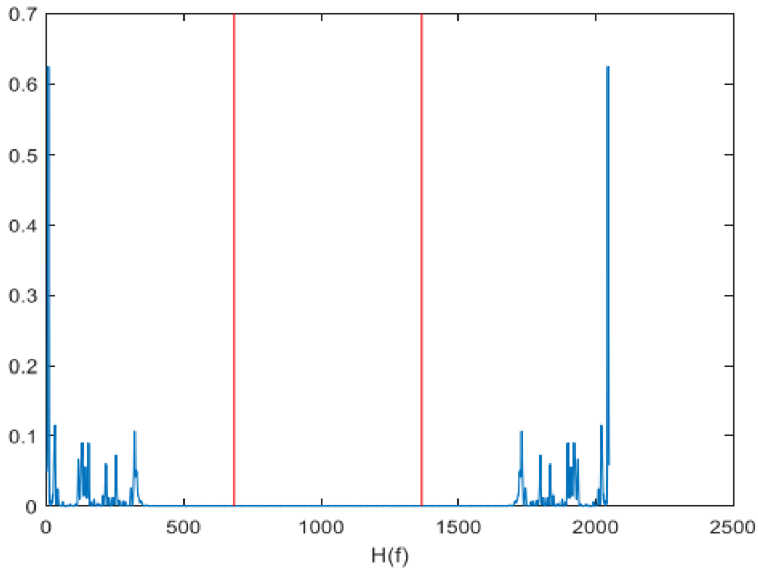


Fig. 4. Example frequency domain from healthy ear

Selection of features was performed by Relief-F (k=9). Relief-F is a filter-type feature selecting algorithm which calculates the weights of features according to intra-class distances and inter-class distances. During Relief-F process, features with negative weights were discarded, and finally 41 features were retained. These features contained maximum of frequency domain channel response, mean of frequency domain channel response, cumulative sum of edge of frequency domain channel response, cumulative sum of center of frequency domain channel response, length of histogram of frequency domain channel response, 5 MFCCs, 8 first order differential MFCCs, 11 second order differential MFCCs, 11 LPCs, and pitch.

Class labels were the doctor defined study indexes of #15 or #85, which were changed to 1 and 0. A neural network was used as the classifier, where fitnet function was used with trainlm. Two thirds of the samples were used as the training set, and one third were used as the testing set. The classification was between-subject, i.e., the training set and the testing set have no overlap.

4.1.2 Regression method

The frequency feature extraction in case of regression was a bit different from the one in classification. Spectral samples are taken from the chirp in a known time interval order. This meant that there was no need to use Fourier transform when by knowing the time slots of each frequency the changes could be compared by overlapping the original and recorded sound. Previous studies indicated problems in the use of FFT, which led to this method. The number of spectral samples is doubled from 51 to 102 by doubling each sample. The profile of frequencies from calibrated chirp before doubling is shown in Fig. 5. Calibration tunes the amplitudes of each frequency to match the profile against air as seen in the figure. The profile of the chirp was selected through experimenting with the typical ear acoustic responses.

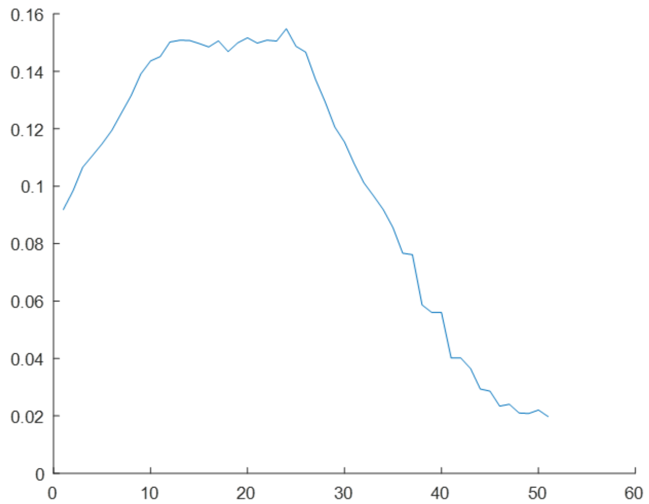


Fig. 5. Profile of calibrated chirp according to spectral samples before doubling

The resulting jagged vector is smoothed with a Blackman window filter. The final focused frequency spectrum is 102 samples between 1.5 and 2.5 kHz in 10 Hz steps. This vector is constructed from both air and ear measurements. The final input is the ratio of the two vectors. Each frequency was weighted to 1 by function:

$$Spectrum = 1 + \frac{Spectrum_{ear} - Spectrum_{air}}{\overline{Spectrum_{air}}} \tag{1}$$

and normalized between 0 and 2 by changing the values outside this range to a maximum of 2 and a minimum of 0. The constructed vector is used as input for the neural network and as a visual aid for filtering of unsuitable data. The filtering is done by looking at the plotted frequencies and listening to the recordings by ear. Loud artifacts and spikes in the frequency are reasons to remove the measurement from the data set.

5 Results

5.1 Pre-study

5.1.1 Input parameters

The acoustic data were processed by converting the recordings into a frequency domain and taking the relevant values from selected frequencies using the PCA method. PCA compares all frequency vectors and calculates a principal component coefficient matrix that identifies relevant values across the group. It was chosen as preferred method to avoid overfitting in case of effusion being a dominant factor in the selected input range. The components obtained are then used as input to an FNN. The network was built with back-propagation using the LM algorithm.

5.1.2 Method selection

The problem solving was compared as classification and regression. In terms of effusion levels and the correlation to OM a regression model is naturally better, but in binary terms a classification method could show better results. Regression was selected as it was showing better initial test results, which is described in Material and methods. For regression modelling the inputs needed to be edited according to the used frequency range. The range and sampling selection rationale consisted of previous unpublished tests performed by Otometri Ltd and the measurement device limitations. Device was intended to be used in varied environments, which meant that manufacturing focus was on robustness and simplicity. Microphones with higher response range were expensive.

The selected input is the normalized combination vector that is reduced using PCA to reduce the computational burden on the neural network as well as to prevent overfitting. Previous experiments had shown using plain frequency inputs would not give good results, especially with large dataset. This was found to be the case even for smaller dataset in this study. The final input is a vector of 16 values. The inputs for the back-propagation algorithm to construct the FNN are the study index and the error index. Study index refers to the doctor defined #15 or #85 values, which were input as 0.15 and 0.85 to better suit the NN functions. The error index was not a focus of this study. The selected NN structure is input layer, two hidden layers and output layer, which is shown in Fig. 6. The number of neurons in the input layer correspond to the 16 values of PCA output. The sizes of the hidden layers are tested in the preliminary trainings. The selected sizes for the two hidden layers are 20 and 11. There are 2 neurons in the output layer for study and error indices. The two hidden layers use a hyperbolic tangent sigmoid transfer function (tansig) and the output layer use a linear transfer function (purelin) to approximate the FNN result. The linear output is found to be the right choice to simulate the linearity of changing ear status from healthy to sick.

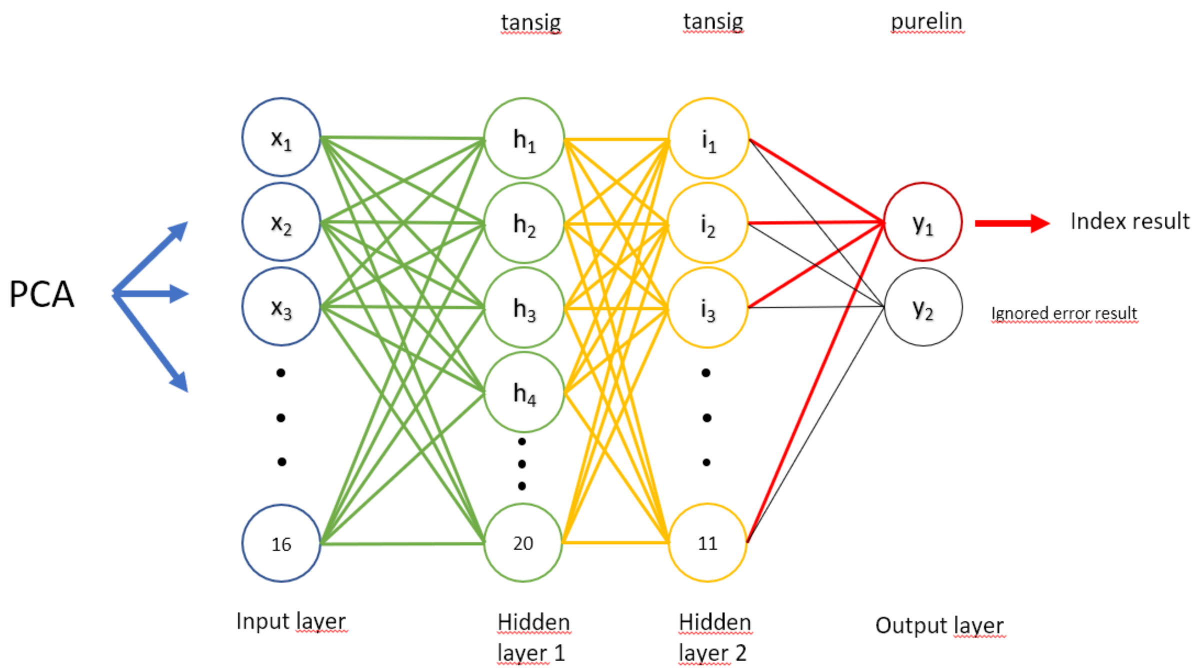


Fig. 6. Feedforward neural network diagram

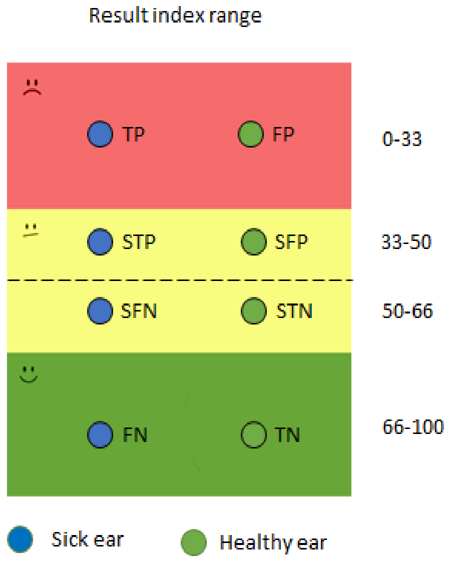
Tansig fits the input into -1 to 1 range, and purelin gives linear results from negative to positive without minimum or maximum limitation (Namin, Leboeuf, Wu, & Ahmadi, 2009).

The training takes place by dividing the ears into train, test and validate groups. The appropriate data distribution is tested with short preliminary training sets. According to these preliminary tests, using 66 % of the ears as training material instead of 50 %, gives better results. The risk of bias is a bit higher with this selection, since the test and validate groups have less ears. The groups are defined so that the ears do not overlap in any of the groups. Both the cleaned ear and the unclean ear are placed in the same group.

A backpropagation FNN was created with newff function from Neural Network Toolbox. Training was performed with trainlm function. (Demuth & Beale, 2001)

5.2 Main study

The preliminary comparison between classification and regression ended in favour of regression. The binary classification method produced best accuracy of 80%. In this best network, the number of neurons in single hidden layer was 48. Model reached binary sensitivity of 73.7% and specificity of 87.5%. Binary sensitivity and specificity explained in Fig. 7. The regression model preliminary tests had accuracy over 85% without neural network size tuning. Classification model had promise and it was deduced it would be good to continue the experimentation in the future.



Description of terms used in calculation of sensitivity, specificity and accuracy:

- True positive (TP): Sick people correctly identified as sick (0–33)
- False positive (FP): Healthy people incorrectly identified as sick (0–33)
- Slight true positive (STP): Sick people with value 33–50
- Slight false positive (SFP): Healthy people with value 33–50
- Slight false negative (SFN): Sick people with value 50–66
- Slight true negative (STN): Healthy people with value 50–66
- False negative (FN): Sick people incorrectly identified as healthy (66–100)
- True negative (TN): Healthy people correctly identified as healthy (66–100)

Equations

Sensitivity:
- Standard: TP / (TP+FN)
- Binary: (TP+STP) / (TP+STP+FN+SFN)
- Strict: TP / (TP+STP+FN+SFN)

Specificity:
- Standard: TN / (TN+FP)
- Binary: (TN+STN) / (TN+STN+FP+SFP)
- Strict: TN / (TN+STN+FP+SFP)

Accuracy:
- Standard/Strict: (TP+TN) / All
- Binary: (TP+STP+TN+STN) / All

Fig. 7. Classification of index groups and calculation equations for sensitivity, specificity and accuracy.

For the study, multi class approach was adopted for both methods. In this 3-class model there were healthy ears, sick ears with symptoms of OME and sick ears with OM visual signs with minimal effusion. The classification method was ruled out as it showed significant class imbalance problems with the small dataset and the accuracy was worse than with binary class. The more complex feature collection method of classification experiments was tested for the regression model, but the accuracy was lower than previously, which led to using the original method.

In case of regression model, the 3-class approach was done by dividing the data into two datasets. First dataset included the ears with symptoms of OME. The second dataset was the same as previous but also included ears with OM visual signs with minimal effusion. The choice to use these two datasets was to differentiate the ability of NN to diagnose OM with different states. AR method, which is mainly known for its fluid sensitivity, was put to the test. These two datasets are called later as OME ONLY and AOM ALL. Technically these cases were not acute otitis media, but prolonged cases needing treatment. Dataset groups were rearranged randomly between train, test and validate to produce to Random 1 and Random 2. The distribution of the data is shown in Table 1.

Table 1. Data distribution and randomization into two random versions. All unique ears were divided into six equal sized groups (first row). These groups had unique ears stated before the parenthesis and number of measurements from those unique ears inside the parenthesis. Each unique ear had at least one measurement. Each of six groups were randomly categorized as train, test or validate (Random 1 and Random 2), where there were four train groups, one test group and one validate group.

Groups for OME ONLY

Group number	1	2	3	4	5	6
Healthy	5 (16)	5 (9)	5 (13)	4 (14)	6 (21)	3 (13)
Sick	3 (15)	3 (12)	3 (15)	4 (16)	2 (9)	4 (17)
Random 1	Train	Test	Train	Validate	Train	Train
Random 2	Train	Validate	Train	Train	Train	Test

Groups for AOM ALL

Group number	1	2	3	4	5	6
Healthy	6 (12)	5 (17)	5 (22)	3 (12)	5 (11)	4 (12)
Sick	3 (13)	4 (18)	4 (19)	6 (25)	4 (15)	4 (21)
Random 1	Train	Train	Train	Train	Test	Validate
Random 2	Train	Test	Validate	Train	Train	Train

Table 1 shows the 6 different groups and the distribution between them. The 3 main groups were train, test and validate. First value shows the unique ear count and the value inside parenthesis is the total data packet count recorded from these ears. OME ONLY

Random 1 dataset had unique ear ratios of 19:12, 5:3 and 4:4, where the value before colon is number of healthy and after colon is number of sick. Complete number of data packets were 63:56, 9:12 and 14:16. Second randomization had ratios of 20:12 (64:55), 3:4 (13:17) and 5:3 (9:12). Total packet divisions were 119:21:30 and 119:30:21 (train, test, validate). AOM ALL Random 1 had ratios of 19:17 (63:75), 5:4 (11:15) and 4:4 (12:21). Random 2 had unique ratio of 18:17 (47:74), 5:4 (17:18), and 5:4 (22:19). Total packet divisions were 138:26:33 and 121:35:41.

The correlation coefficient (CC) between inputted indexes and results of trained neural network was calculated by a corrcoef function of MATLAB and used as an initial specifier to identify good training results. Sensitivity and specificity were calculated on multiple levels by limiting the area of results to 4 distinct groups, as shown in Fig. 7. A choice to use 4 groupings was made to further inspect the functionality of linear transfer function instead of simple binary classification.

Values between 33 and 66 represent transition values, which can occur when state of ear is between healthy and sick definition. This is intended as follow-up functionality to show the progress of recovery. For study reasons the results were split into three following categories, a standard result, which ignores the values between 33 and 66, a binary result with 50-50 distribution, and strict result with 33 cut-offs in both directions. Strict result counts values between 33 and 66 to be false in both positive and negative.

When both OME ONLY and AOM ALL datasets had two randomized groups, there was a total of 4 neural networks to be trained. Results of the networks are displayed in Table 2. Training was done by selecting the groups for MATLAB code and running the training process 30 times. The highest CC values were picked out of 30 trained networks for every case. Every CC value given by corrcoef function had p-value better than 0.0001.

Table 2. NN training results.

OME ONLY, Random 1, CC = 0.856, Average index error = 12.83				
Confusion matrix	Positive	Negative	Slight positive	Slight negative
True	65	66	14	15
False	2	2	3	3
Calculations	Sensitivity	Specificity	Accuracy	
Standard	97.01	97.06	77.06	
Binary	94.05	94.19	94.12	
Strict	77.38	76.74	77.06	

AOM ALL, Random 1, CC = 0.787, Average index error = 13.81				
Confusion matrix	Positive	Negative	Slight positive	Slight negative
True	73	73	22	6
False	3	5	4	11
Calculations	Sensitivity	Specificity	Accuracy	
Standard	93.59	96.05	74.11	
Binary	85.59	91.86	88.32	
Strict	65.77	84.88	74.11	

OME ONLY, Random 2, CC = 0.879, Average index error = 10.74				
Confusion matrix	Positive	Negative	Slight positive	Slight negative
True	70	72	7	11
False	0	4	3	3
Calculations	Sensitivity	Specificity	Accuracy	
Standard	94.59	100	83.53	
Binary	91.67	96.51	94.12	
Strict	83.33	83.72	83.53	

AOM ALL, Random 2, CC = 0.846, Average index error = 12.47				
Confusion matrix	Positive	Negative	Slight positive	Slight negative
True	88	56	16	20
False	2	0	8	7
Calculations	Sensitivity	Specificity	Accuracy	
Standard	100	96.55	73.10	
Binary	93.69	88.37	91.37	
Strict	79.28	65.12	73.10	

The standard sensitivity and specificity values give a reference to how in a normal diagnostic situation there is always a degree of trust in the diagnostic result. For the classification in this study, the standard value is the result with high degree of trust. This means that the unsure cases are not displayed. The standard/strict accuracy here displays the number of ears, which have the high degree of trust. The fairest way to assess the results would be the binary, which includes every measurement, but on the expense of degree of trust.

The results for the OME ONLY cases have high accuracy and very small number of results are outside the intended index area. From Table 2, we can see that the accuracy of the measurements is slightly higher in the OME ONLY dataset. The accuracy is below 80 % for AOM ALL, but the binary result is around 90 %. Compared to the OME ONLY dataset, there is a difference of around 5 % units. The Random 2 network in OME ONLY is the best version with good accuracy of over 80 % for every calculation option. The only problem with it is the biased weight on the specificity. This weighted result raises the strict accuracy but is not necessarily a good thing. The other versions were also more tipped to either good sensitivity or good specificity, but the accuracy was still pretty good overall. The networks with AOM ALL as training data had a bit lower accuracy compared to OME ONLY. To demonstrate the difference between these two networks, the data of ears with minimal effusion was inputted into the trained FNNs in Table 2. The results are shown on Table 3.

Table 3. Results from trained FNNs using AOM ears without effusion.

OME ONLY, Random 1, CC = 0.856, Average index error = 43.66				
Confusion matrix	Positive	Negative	Slight positive	Slight negative
True	5	-	4	-
False	-	11	-	7
Calculations	Sensitivity /Accuracy			
Standard	31.25			
Binary	33.33			
Strict	18.52			

AOM ALL, Random 1, CC = 0.787, Average index error = 12.89				
Confusion matrix	Positive	Negative	Slight positive	Slight negative
True	10	-	12	-
False	-	3	-	2
Calculations	Sensitivity /Accuracy			
Standard	76.92			
Binary	81.48			
Strict	37.04			

OME ONLY, Random 2, CC = 0.879, Average index error = 48.08				
Confusion matrix	Positive	Negative	Slight positive	Slight negative
True	3	-	2	-
False	-	12	-	10
Calculations	Sensitivity /Accuracy			
Standard	20			
Binary	18.52			
Strict	11.11			

AOM ALL, Random 2, CC = 0.846, Average index error = 9.66				
Confusion matrix	Positive	Negative	Slight positive	Slight negative
True	18	-	7	-
False	-	0	-	2
Calculations	Sensitivity /Accuracy			
Standard	100			
Binary	92.59			
Strict	66.67			

It was clear that the AOM ALL networks, which had the ears with minimal effusion mixed into the data, had much better detection results. The OME ONLY networks classified these ears as healthy since they did not have any effusion. This proves the fact; how dominant the presence of effusion is in the input values. This also proves that the healthy and OM ears without effusion do have other differences, which the NN is able to differentiate from the AR measurement.

The network model was able to identify the ears with no effusion, but it was clearly more challenging for the algorithm to find the right network weights. CC and accuracy were lower overall. The differences are visible in the histograms for the results in Fig. 8. Histograms were divided into 25 bins using histcounts function to detect edges and the counts for the bins. OME ONLY had around equal amount of healthy and sick ear measurements, whereas AOM ALL had the extra 27 sick measurements.

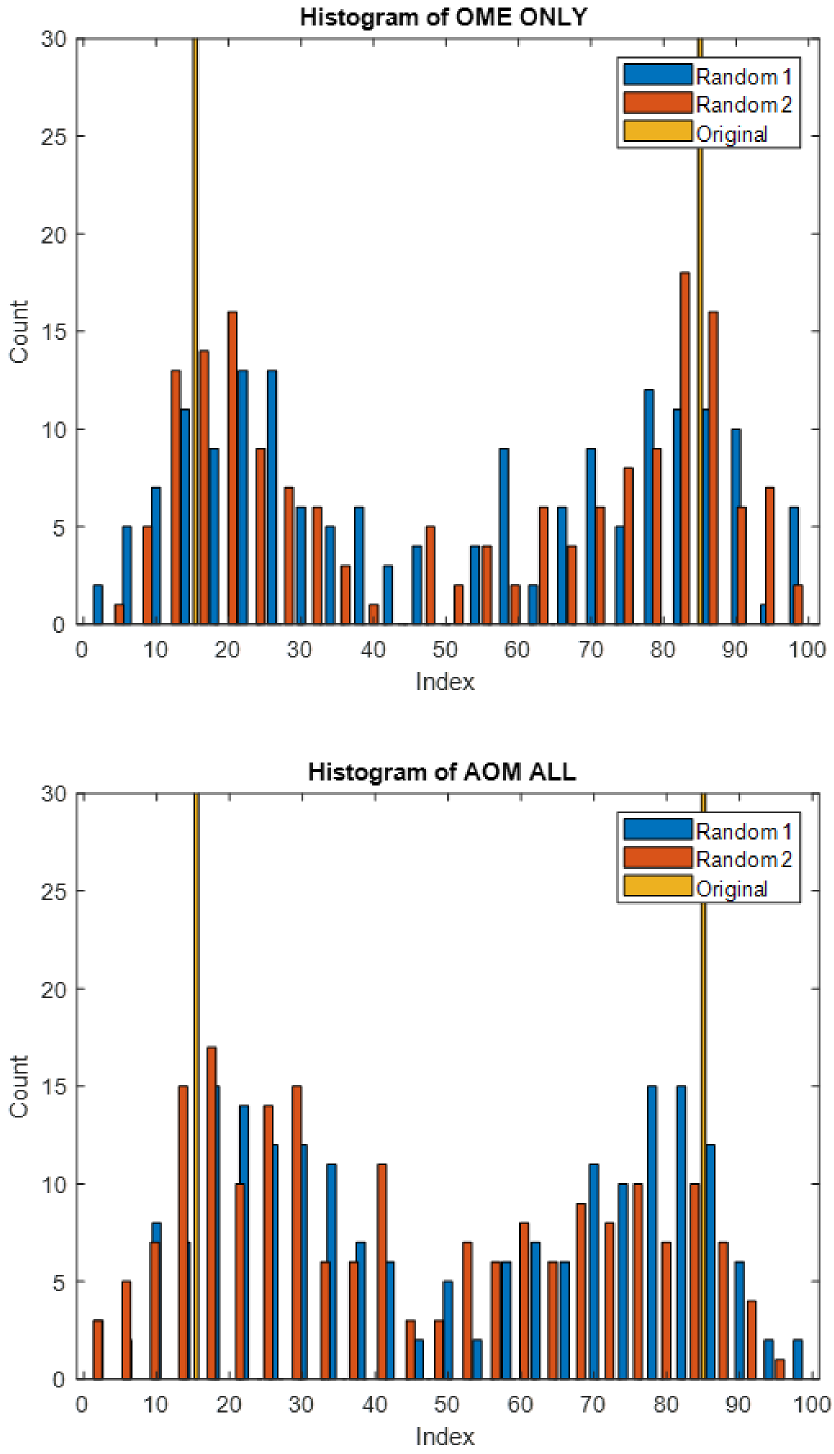


Fig. 8. a) Histogram of OME ONLY and b) histogram of AOM ALL datasets.

Rough comparison from the Fig. 8a and 8b shows that the localization near the index values is better with OME ONLY network. Comparing these histograms, it can be said that the OME ONLY network gives better results when effusion is present. On the other hand, that same network cannot identify ears, which have OM with minimal effusion. Distinct features for OME ONLY is the ability to give better binary results, whereas the AOM ALL can detect two types of sick ears, but the results are not as clear.

The difference between these two types is very subtle. It is nearly impossible to differentiate using just the frequency. The converted frequency spectrum of whole data given in the window of 1500–2500 Hz shows how similar even the sick and healthy ears are. Spectrum is shown in Fig. 9a, where the y-axis represents the relation to the free-air calibrated measurement. These spectrums were weighted around 1 as stated in the Eq.1. When signal is straight line over value 1, the ear recording matches the free-air calibrated measurement. When value is lower, the ear measurement has lower amplitude on those frequencies. Means of frequencies are shown in Fig. 9b and variances are shown in Fig. 9c. Figures provide information about AOM without effusion, OME and healthy ears.

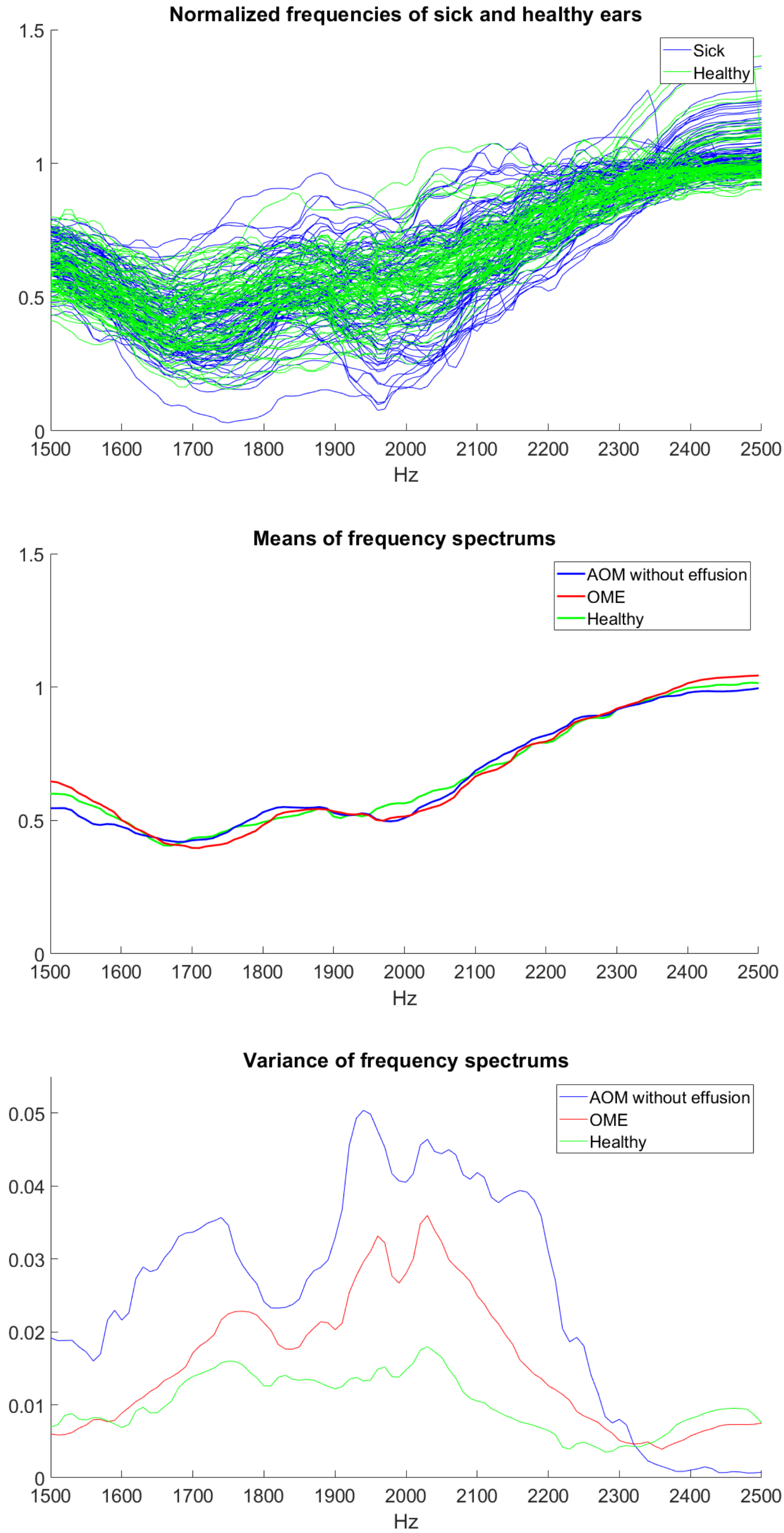


Fig. 9. a) Normalized frequencies of sick and healthy ears for all measurements, b) Means of healthy, OME ONLY and OAM ALL frequencies, and c) variances of healthy, OME ONLY and OAM ALL frequencies.

The colors in figures represent the original indexes, not the neural network results. Blue curves represent individual sick ears and green curves healthy ears in Fig. 9a. The most significant visual information is that the sick ears tend to have much higher variation in all frequencies. This means that healthy ears will have more predictable reflection and attenuation, which will make any bigger changes likely to be associated with OM. Variance change is clear between healthy and sick ears. AOM without effusion shows higher variance compared to OME, but this could be caused by smaller subject group size. The variances are more meaningful between 1650–2100 Hz for sick and healthy ears. It is also notable that healthy ears reflectance falls in between the two sick ear groups on the ends of the frequency window in Fig. 9b. This indicates the possibility of discarded information at frequencies below 1500 Hz and above 2500 Hz.

6 Discussion

The used AR device combined with FNN frequency analysis can handle linear regression detecting of AOM at least on same level as tympanometry or SG-AR. Results showed comparable accuracy levels to Teele studies (Teele, David W. & Teele, 1984), but as the data used in training is relatively small for a NN, a definitive result cannot be concluded. Still notable finding is that correlation with MEE is not necessarily needed for AOM detection. The slight accuracy difference between the two constructed NNs does indicate that effusion has big part in the overall variability of PCA outputted values. In terms of this variance, healthy ears and sick ears without effusion are very closely related. The result from this comparison indicates that there is a variance aspect, which can differentiate these sick ears without effusion from healthy ears.

What we can deduct from the used methods is that the frequency variance between different ears is the most important variable for this system. PCA uses each frequency as a separate feature and calculates the most significant values by considering the entire data set, including the variance. PCA might have problems with sustaining some information, but in this case that does not seem to produce problems. In the future, several different combinations will have the possibility to adapt different curve profiles for better classification of the ears. The current binary indexing seems simple, but the fact that sick ears include more than one type of OM pathology makes this identification sensitive to small trends. NN cannot determine the output based on higher values showed by OME cases e.g., at the 1500 Hz frequency. This is emphasized since the amount of effusion was not reported in any way to NN. There is also the possibility of using multiple smaller frequency areas or even combinations of time and frequency domain data when a large and well-documented data has been gathered.

The nature of neural network analysis is very accurate and rigorous, as evidenced by the fact that OM cases with no effusion cannot be identified when they are not included in the data. The most significant finding of these results is that the neural network can detect a binary result reasonably well with two different OM types, without other indicators in the network input. Even if the accuracy is slightly lower with both types than with OME ONLY, the level of accuracy displayed is well within the requirements of a home-use or even a clinical-use device. It is also worth mentioning that there was no input data for the area of uncertainty between sick and healthy ears, making strict computation questionable and binary accuracy thus more credible. The binary accuracy was between 88–95%. In the long run, the goal is to use linear regression to determine the range of uncertainty between diseased and healthy ears, which is why the linear transfer function is used in the output layer. Direct binary classification using the sigmoid function in the output layer would probably have given slightly better binary results.

A typical training outcome was that networks placed more emphasis on either sensitivity or specificity. This can be explained in some cases by the similarity of the acoustic responses. This was also observed in the classification method, where increasing the classification from binary to 3 classes caused significant drops in accuracy. The best way to combat these problems is to either gather more data to reduce the problems caused by imbalance in datasets or use figure out better features to separate the classes. When data pool is larger, more specific classification can be utilized, such as differences in a child's age or stages of the ear healing process, or the process of OM in an ear that looks relatively healthy. The age of the child affects the volume of the ear canal and middle ear. The change in volume is inversely proportional to the frequency of the system formed by the ear and the device. This is significant when the target group of patients is young children aged a few months or older, where the size and tissue structure become a factor (Teele, D. W. et al., 1989). Due to the patient group, it is possible that many of the healthy ears were already healing ears with only residual symptoms of OM because the selected group were patients screened for myringotomy surgery. The surgery is only intended for cases of diseased ears that do not heal normally and where recurrent infections are suspected unless surgery is performed. A healthy group would also need more information from other sources.

There are several other methods, which could be used instead of simple frequency domain data. The classification method pursued in this study used a much more complicated feature extraction, which could easily have much promise with a more profound dataset. Pre-processing the time domain information into a visual information and using image related AIs, which have already been developed to great lengths in other solutions (Khamparia et al., 2019). Decision tree methods could also be a potential branch of AI technology to utilize with a classification approach.

During the study, there were some problems that should be solved in the future. The number of patients was less than 10 patients per week, which limited the number of ears eligible for the study and slowed down data collection. It would be better to take only one measurement from each ear or to use the averages of several measurements, but due to the small group size all the collected measurements had to be used separately. This was not only a negative thing, as the included environmental noise makes the network more flexible. The effects of errors were minimized by omitting the most erroneous measurements, which led to rejection of several measurements. The rejection also included few unique ears caused by device malfunction. Small group size was highlighted as the main negative point.

The use of a smartphone and built-in logic for sound processing is a large unknown variable, which creates uncertainty about the repeatability of measurements. This uncertainty is due in part to the calibration process because the smartphone responds strongly to the surrounding environment. The only way to ensure that the smartphone does not make automatic changes to the recording sensitivity would be to tamper with the device's hardware, which will void the device's warranty. An independent probe for performing acoustic operations would fix most of the negatives found during this study.

7 Conclusions

The accuracy of the presented neural network adapted AR solution is higher than expected with very minimal dataset. Still a clear advantage in comparison to other solutions is hard to prove objectively, as there is no clear baseline for acoustic reflectometry solutions in terms of accuracy, because of varied methods and cut-off values. It is only an approximation to say any solution to be better than other. A clearly defined test settings between other solutions is needed for this.

The presented solution showed the ability to detect typical OM cases with effusion in high sensitivity. Also, the sensitivity was good for abnormal OM cases, where effusion was minimal, but ears showed clear visual signs of OM. Despite having high sensitivity there was no major degradation in the specificity or vice-versa.

The versatility of the method has not been studied in-depth to this day, when considering the vast development in NN and AI technology. A simple frequency domain-based analysis gives more than adequate detection accuracy for AOM in home or healthcare environment. There is a widespread and growing need for diagnostic devices that can help physicians or patients to monitor the status of their health. Solution is a prime example of old viable technology being included into new innovative methods to process data and form results. The AR is a solution with a relatively long history in technological diagnostic methods, but it is still very much relevant when enhanced with computer-assisted ways. The results are applicable to current medicine, and solution has several possible ways for future development.

The visible changes to ear have been the prevalent way to diagnose ears, which has already been adapted by AIs specializing in image analysis. The combined outputs of image analysis and acoustic properties would be new and a comprehensive method for avoiding the weaknesses of single diagnostic solutions. Moreover, would also give more depth and reliability to diagnosis and follow-up of OM.

References

Abbott, P., Rosenkranz, S., Hu, W., Gunasekera, H., & Reath, J. (2014). The effect and acceptability of tympanometry and pneumatic otoscopy in general practitioner diagnosis and management of childhood ear disease. *BMC Family Practice, 15*(1), 181. doi:10.1186/s12875-014-0181-x

Babb, M. J., Hilsinger, R. L.,Jr, Korol, H. W., & Wilcox, R. D. (2004). Modern acoustic reflectometry: Accuracy in diagnosing otitis media with effusion. *Ear, Nose, & Throat Journal, 83*(9), 622-624.

Bahoura, M., & Ezzaidi, H. (2013). *Hardware implementation of MFCC feature extraction for respiratory sounds analysis* doi:10.1109/WoSSPA.2013.6602366

Balamurali, B. T., Hee, H. I., Kapoor, S., Teoh, O. H., Teng, S. S., Lee, K. P., . . . Chen, J. M. (2021). Deep neural network-based respiratory pathology classification using cough sounds. *Sensors, 21*(16), 5555. doi:10.3390/s21165555

Barnett, E. D., Klein, J. O., Hawkins, K. A., Cabral, H. J., Kenna, M., & Healy, G. (1998). Comparison of spectral gradient acoustic reflectometry and other diagnostic techniques for detection of middle ear effusion in children with middle ear disease. *The Pediatric Infectious Disease Journal, 17*(6), 556-9; discussion 580. doi:10.1097/00006454-199806000-00035 [doi]

Block, S. L., Mandel, E., McLinn, S., Pichichero, M. E., Bernstein, S., Kimball, S., & Kozikowski, J. (1998). Spectral gradient acoustic reflectometry for the detection of middle ear effusion by pediatricians and parents. *The Pediatric Infectious Disease*

Journal, 17(6), 560-4; discussion 580. doi:10.1097/00006454-199806000-00036 [doi]

Block, S. L. (1999). Spectral gradient acoustic reflectometry: Detection of middle ear effusion in suppurative acute otitis media. *The Pediatric Infectious Disease Journal, 18*(8), 741-744. doi:10.1097/00006454-199908000-00024

Bylander, A. (1980). Comparison of eustachian tube function in children and adults with normal ears. *The Annals of Otology, Rhinology & Laryngology.Supplement, 89*(3 Pt 2), 20-24. doi:10.1177/00034894800890s308 [doi]

Cao, W., Wang, X., Ming, Z., & Gao, J. (2017). A review on neural networks with random weights. *Neurocomputing, 275* doi:10.1016/j.neucom.2017.08.040

Chianese, J., Hoberman, A., Paradise, J. L., Colborn, D. K., Kearney, D., Rockette, H. E., & Kurs-Lasky, M. (2007). Spectral gradient acoustic reflectometry compared with tympanometry in diagnosing middle ear effusion in children aged 6 to 24 months. *Archives of Pediatrics & Adolescent Medicine, 161*(9), 884-888. doi:10.1001/archpedi.161.9.884

Coleman, A. (2021). The microbiome of otitis media and development of a probiotic to prevent otitis media in indigenous australian children.

Demuth, H. B., & Beale, M. H. (2001). *Neural network toolbox for use with MATLAB: User's guide*. Natick, Mass: MathWorks.

Douniadakis, D. E., Nikolopoulos, T. P., Tsakanikos, M. D., Vassiliadis, S. V., & Apostolopoulos, N. J. (1993). Evaluation of acoustic reflectometry in detecting otitis media in children. *British Journal of Audiology, 27*(6), 409-414. doi:10.3109/03005369309076717 [doi]

Fagan, J. (2014). (2014). Open access guide to audiology and hearing aids for otolaryngologists. Paper presented at the

Gavin, H. P. (2013). (2013). The levenberg-marquardt method for nonlinear least squares curve-fitting problems c \textcopyright. Paper presented at the

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep feedforward networks* MIT Press.

Granath, A. (2017). Recurrent acute otitis media: What are the options for treatment and prevention? *Current Otorhinolaryngology Reports, 5*(2), 93-100. doi:10.1007/s40136-017-0151-7

Hannula, M., Hinkula, H., Holma, T., Löfgren, E., & Sorri, M. (2009). Artificial neural network analysis in the evaluation of ear canal and tympanic membrane properties from acoustic reflectometry data. *IFMBE Proceedings, 25* doi:10.1007/978-3-642-03882-2_622

Hannula, M., Hirvikoski, A., Hinkula, H., & Holma, T. (2009). Application of artificial neural network in analysis of acoustic reflectometry data. *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications, AIA 2009,* , 175-178.

Hannula, M., Holma, T., Löfgren, E., Hinkula, H., & Sorri, M. (2011). Preliminary results of clinical tests of a new neural-network-based otitis media analysis system.

NCTA 2011 - Proceedings of the International Conference on Neural Computation Theory and Applications, , 365-368.

Jolliffe, I., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*, 20150202. doi:10.1098/rsta.2015.0202

Kemaloğlu, Y. K., Sener, T., Beder, L., Bayazit, Y., & Göksu, N. (1999). Predictive value of acoustic reflectometry (angle and reflectivity) and tympanometry. *International Journal of Pediatric Otorhinolaryngology, 48*(2), 137-142. doi:S0165-5876(99)00017-8 [pii]

Khamparia, A., Gupta, D., Nguyen, N. G., Khanna, A., Pandey, B., & Tiwari, P. (2019). Sound classification using convolutional neural network and tensor deep stacking network. *IEEE Access, 7*, 7717-7727. doi:10.1109/ACCESS.2018.2888882

Kimball, S. (1998). Acoustic reflectometry: Spectral gradient analysis for improved detection of middle ear effusion in children. *The Pediatric Infectious Disease Journal, 17*(6), 552-5; discussion 580. doi:10.1097/00006454-199806000-00034 [doi]

Kotsiantis, S., Zaharakis, I., & Pintelas, P. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review, 26*, 159-190. doi:10.1007/s10462-007-9052-3

Laine, M. (2015). *Tympanometry and spectral gradient acoustic reflectometry in the diagnosis of otitis media in young children* Retrieved from http://urn.fi/URN:ISBN:978-951-29-6143-6

Lous, J., Ryborg, C. T., Damsgaard, J. J., & Munck, A. P. (2012). Tympanometry in general practice: Use, problems and solutions. *Family Practice, 29*(6), 726-732. doi:10.1093/fampra/cms045

Mandel, E. M., Casselbrant, M. L., Richert, B. C., Teixeira, M. S., Swarts, J. D., & Doyle, W. J. (2016). Eustachian tube function in 6-year-old children with and without a history of middle ear disease. *Otolaryngology--Head and Neck Surgery : Official Journal of American Academy of Otolaryngology-Head and Neck Surgery, 154*(3), 502-507. doi:10.1177/0194599815620149

Muderris, T., Yazıcı, A., Bercin, S., Yalçıner, G., Sevil, E., & Kırıs, M. (2013). Consumer acoustic reflectometry: Accuracy in diagnosis of otitis media with effusion in children. *International Journal of Pediatric Otorhinolaryngology, 77*(10), 1771-1774. doi:https://doi.org/10.1016/j.ijporl.2013.08.019

Namin, A. H., Leboeuf, K., Wu, H., & Ahmadi, M.Artificial neural networks activation function HDL coder. Paper presented at the - *2009 IEEE International Conference on Electro/Information Technology,* 389-392. doi:10.1109/EIT.2009.5189648

Osborne, J. W. (2021). *Regression & linear modeling: Best practices and modern methods*. Thousand Oaks, California: doi:10.4135/9781071802724

Parmar, S., Davessar, J. L., Singh, G., Arora, N., Kansal, L., & Singh, J. (2019). Prevalence of otitis media with effusion in children with hearing loss. *Indian Journal of Otolaryngology and Head and Neck Surgery : Official Publication of the Association of Otolaryngologists of India, 71*, 1276-1281. doi:10.1007/s12070-018-1310-y

Pichichero, M. E. (2000). Acute otitis media: Part I. improving diagnostic accuracy. *American Family Physician, 61*(7), 2051-2056.

Plapous, C., Marro, C., & Scalart, P. (2006). Improved signal-to-noise ratio estimation for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(6), 2098-2108. doi:10.1109/TASL.2006.872621

Puhakka, T. (2014). Comparison of spectral gradient acoustic reflectometry and tympanometry for detection of middle-ear effusion in children. *The Pediatric Infectious Disease Journal, 33*(8), e183-e186. doi:10.1097/INF.0000000000000273

Schilder, A. G. M., Chonmaitree, T., Cripps, A. W., Rosenfeld, R. M., Casselbrant, M. L., Haggard, M. P., & Venekamp, R. P. (2016). Otitis media. *Nature Reviews.Disease Primers, 2*(1), 16063. doi:10.1038/nrdp.2016.63

Singh, K., & Upadhyaya, S. (2012). (2012). Outlier detection: Applications and techniques. Paper presented at the

Teele, D. W., Klein, J. O., & Rosner, B. (1989). Epidemiology of otitis media during the first seven years of life in children in greater boston: A prospective, cohort study. *The Journal of Infectious Diseases, 160*(1), 83-94. doi:10.1093/infdis/160.1.83 [doi]

Teele, D. W., Stewart, I. A., Teele, J. H., Smith, D. K., & Tregonning, S. J. (1990). Acoustic reflectometry for assessment of hearing loss in children with middle ear effusion. *The Pediatric Infectious Disease Journal, 9*(12), 870-872. doi:10.1097/00006454-199012000-00002 [doi]

Teele, D. W., & Teele, J. (1984). Detection of middle ear effusion by acoustic reflectometry. *The Journal of Pediatrics, 104*(6), 832-838. doi:https://doi.org/10.1016/S0022-3476(84)80476-X

Teppo, H., & Revonta, M. (2009). Consumer acoustic reflectometry by parents in detecting middle-ear fluid among children undergoing tympanostomy. *Scandinavian Journal of Primary Health Care, 27*(3), 167-171. doi:10.1080/02813430903072165

Teppo, H., Revonta, M., Lindén, H., & Palmu, A. (2006). Detection of middle-ear fluid in children with spectral gradient acoustic reflectometry: A screening tool for nurses? *Scandinavian Journal of Primary Health Care, 24*, 88-92. doi:10.1080/02813430600699997

Torgo, L., & Gama, J. (1997). Regression using classification algorithms. *Intelligent Data Analysis, 1*(1), 275-292. doi:https://doi.org/10.1016/S1088-467X(97)00013-9

Tutunji, T. (2009). *Approximating transfer functions using neural network weights* doi:10.1109/NER.2009.5109378

Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., & Moore, J. H. (2018). Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics, 85*, 189-203. doi:https://doi.org/10.1016/j.jbi.2018.07.014

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *Plos One, 14*(11), e0224365. Retrieved from https://doi.org/10.1371/journal.pone.0224365

Wiertsema, S. P., & Leach, A. J. (2009). Theories of otitis media pathogenesis, with a focus on indigenous children. *Medical Journal of Australia, 191*, S50-S54. doi:https://doi.org/10.5694/j.1326-5377.2009.tb02927.x

Won, J., Monroy, G. L., Huang, P. C., Dsouza, R., Hill, M. C., Novak, M. A., . . . Boppart, S. A. (2018). Pneumatic low-coherence interferometry otoscope to quantify tympanic membrane mobility and middle ear pressure. *Biomedical Optics Express, 9*(2), 397-409. doi:10.1364/BOE.9.000397 [doi]

Yang, L., Liu, S., Tsoka, S., & Papageorgiou, L. G. (2017). A regression tree approach using mathematical programming. *Expert Systems with Applications, 78*, 347-357. doi:https://doi.org/10.1016/j.eswa.2017.02.013

Ying, X. (2019). An overview of overfitting and its solutions. *Journal of Physics: Conference Series, 1168*, 022022. doi:10.1088/1742-6596/1168/2/022022