Developing a deep learning algorithm to improve diagnosis of otitis media

Phong Phu Nguyen

nguy1192@flinders.edu.au

Supervisors

Dr Trent Lewis, College of Science & Engineering

Dr Jacqueline Stephens, College of Medicine & Public Health

Submitted on October, 2021

Submitted to
the College of Science and Engineering
in partial fulfilment of the requirements for the degree of
Master of Science – Computer Science
at Flinders University – Adelaide Australia.

DECLARATION OF ORIGINALITY

I certify that this work does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief, it does not contain any material previously published or written by another person except where due reference is made in the text.

15/10/2021

Phong Phu Nguyen

ABSTRACT

Background: Otitis Media is a common childhood ailment. In clinical practice, the diagnosis of Otitis Media includes the visual examination of the tympanic membrane from otoscope (otoscopy images) and measuring the movement patterns of the eardrum for different pressure (tympanometry data). However, the diagnosis requires extensive tool usage training and result interpretation. Misdiagnosis is a common problem for community-based clinicians in remote areas, which could cause delays in the treatment of the disease. The report proposes machine learning systems that could give predictions on different types of otitis media by using multiple sources of diagnosis data available.

Methods: Support Vector Machine, Multi-Layer Perceptron and Convolutional Neural Network models are used to build the system on two different datasets. The first dataset is a public one, with 454 images of three different categories. The second dataset is the Swimming Pool database coming from a report in regional South Australia for child ear diagnosis. Data is available in terms of otoscopy videos and various diagnosis labelling. Input processing and normalisation, principal component analysis, and grid search have been utilised for training the model with available data in both multi-label and binary classification problems. The tympanometry data has been combined with the prediction results from extracted otoscopy images to boost the overall performance of the algorithm.

Results: The best model accuracy achieved on 3-label problem on the public dataset is 83%. The highest accuracies for a 5-label classification problem and binary classification problem on Swimming Pool data are 64% and 78%, respectively. By combining the tympanometry data with the probability prediction of otoscopy-image-based model, the accuracy of the system has been increased from 78% to 82% in the binary classification.

Conclusion: Machine learning models could be used to build a system that could support the otitis media diagnosis to support effective triage, timely patient referrals and effective treatment of the illness. The report also suggests the possibility of combining different diagnosis data types to increase the overall predictive performance of the system.

ACKNOWLEDGMENTS

I would like to offer my special thanks to my supervisor, Dr Trent Lewis, for his invaluable advice and continuous support during my project. His immense knowledge and experience in the field have steered me to the finish of the project. I would also like to thank Dr Jacqueline Stephens for the insightful support in understanding the structure and interpretation of the otitis media diagnosis data. Finally, I am also thankful to Associate Professor Linnett Sanchez and Professor Simon Carney for their guidance in collapsing the labels of otoscopy and tympanometry diagnosis data. Without their support, it would be impossible for me to complete this thesis.

CONTENTS

Table of Contents

List of figures

List of tables

INTRODUCTION

Otitis media (OM), also known as middle ear infection, is one of the most common childhood illnesses (Gaddey et al., 2019). OM is a group of inflammatory diseases, with the three most common forms are acute otitis media (AOM), which is caused by the inflammation of the middle ear and tympanic membrane, otitis media with effusion (OME) and chronic suppurative otitis media (CSOM). Some forms of otitis media cause pain, fever, otorrhea (AOM). However, other forms of otitis media could be asymptomatic and can go undiagnosed (Harmes et al., 2013). Approximately 80% of children will have one episode of Acute Otitis Media, and 80% to 90% of them will have one episode of Otitis Media with Effusion in school-age (Gaddey et al., 2019). Improper treatment of otitis media could result in hearing impairment, affecting cognitive development and educational outcomes (Williams and Jacobs, 2009). Currently, diagnosing otitis media includes a visual examination of the tympanic membrane from a standard otoscope. Other tools are available such as tympanometry, pneumatic otoscopy (Monroy et al., 2019). However, using those tools requires extensive training and is not very practical for the community-based worker. Human-based diagnosis is also subjected to reproducibility, as the conclusion is heavily dependent on the experience of the practitioners (Gaddey et al., 2019). There is a need to develop a system that is straightforward, which could support less confident technicians in diagnosing otitis media. Such a system could help to increase the efficiency of the patient's identification process.

In this report, three different machine learning models have been proposed and tested on two different datasets. Standard normalisation and feature extractions have been utilised. Traditional hyperparameter optimisation has been used to achieve the best models for a given set of parameter combinations. The report also discusses the difficulties in obtaining quality medical images from the videos and different ways to interpret the input label based on various needs. Finally, the report proposes the first system that utilises more than one source of data to boost the overall performance of the machine learning based otitis media diagnosis system.

PROBLEM STATEMENT

Neural network and machine learning, in general, have been widely utilised across multiple disciplines. Those approaches, in collaboration with computer vision techniques, have been used as a promising solution for various classification and image recognition tasks (Zeng et al., 2021). In the diagnosis of otitis media domain, otoscopic images have been analysed and used as input to develop machine learning algorithms that could classify multiple complex labels. The same tasks used to require complex training and tremendous experience from otolaryngologists. While the literature shows that the technique is still at its outset, multiple pieces of evidence suggest that the outcome accuracies of machine learning based systems could be comparable or even surpass those of traditional clinicians (Livingstone and Chau, 2020, Cai et al., 2021). However, to the best of my knowledge, there has been no research incorporating multiple data sources in developing such a classification system. For some diagnosis problems, using only otoscopy images is insufficient and could lead to incorrect results (Moberly et al., 2018). The aim of this project is to develop a diagnosis support system for otitis media related illnesses using machine learning. The system should give accurate predictions for various types of otitis media illnesses. Furthermore, the system should combine the input data from multiple sources to achieve better predictive performance. The input data include otoscopy images and other information acquired from the diagnostic tests.

LITERATE REVIEW

1. Machine Learning

One could find the term Artificial Intelligence, Machine Learning and their applications in many areas today. Although these two technologies are usually used interchangeably, each conveys different spectrums of concepts. According to John McCarthy, one of the first used the term "Artificial Intelligence", AI is the "the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable" (McCarthy, 2007). In short, Artificial Intelligence is the ability of a computer program to mimic human cognitive skills such as reasoning, generalising and behaving like a human being (Kersting, 2018). There is no limit on what an AI system could do. However, due to the complexity in constructing such a system and the limitation of contemporary hardware and algorithms, AI systems are primarily built for a specific problem to achieve a goal that is used to require human's logic and behaviour. On the other hand, Machine Learning is "concerned with the question of how to construct computer programs that automatically improve with experience" (Hierons, 1999). Experience refers to data such as observations or examples fed into the system in the learning phase. The system could then extract knowledge (patterns) from input data and take decisions or predictions for new observations without being explicitly programmed. So, both AI and ML are about building intelligent systems that can simulate human behaviour, and ML is considered a subset of AI. While AI encompasses problem-solving and reasoning in general, ML is more about learning from generally a large amount of data to generate accurate results or give predictions based on the historical data.

So how a computer could learn? There are three main components that are needed for a machine learning system. The first component is a dataset containing numbers, images, texts or any other forms of data related to the specific problem. It represents a collection of samples that are used for the training. Capturing and processing datasets is usually the component that takes most of the time and effort in building a Machine Learning system (Gudivada et al., 2017). The second component are features, which are the thing that the system needs to learn from the input data. Features are important variables and their correlations in the data that play salient roles in providing accurate results or predictions from the observation. Features could be represented in raw data, or they could be derived from the learning process. Finally, the third component of a machine learning system is an algorithm. The algorithm helps find the patterns in the data and adjust the processing and management of data to reflect the best outcomes. One task could be solved by using a different algorithm, and the choice of an algorithm also affects the accuracy and speed of the system. There have been many algorithms developed for various problems.

In practice, Machine Learning could be divided into three main categories (Kotsiantis, 2007):

- **Supervised Learning:** the training data is labelled with the correct outcome. The data is then fed into the algorithm for training, and the label is used to give feedback to the algorithm to tell if the algorithm has given the correct answer or not. Gradually, the algorithm could learn the correlations between the samples and their labels, and it could be able to predict a good label for a completely new observation. This type of learning is commonly utilised for classification and regression. Algorithms used for supervised learning include Decision Tree, K-Nearest Neighbours, Linear/Logistic Regression, Support Vector Machine, Neural Network. Supervised learning algorithm is the most common form of Machine Learning with applications in spam filtering, language detection or image classification system.

- **Unsupervised Learning:** the training data features no labels. Instead, the algorithm is fed with many data and with the algorithm, it will extract meaningful properties from the data to organise the data in a way that makes sense (Kotsiantis, 2007). Since the majority of data today is unlabelled (user logs, media, transactions, news), unsupervised learning offers a promising solution to very challenging problems of understanding the data and extracting useful insights analytics. Thus, unsupervised learning is considered data-driven learning. Algorithms developed for unsupervised learning problems are K-means clustering, Singular Value Decomposition, Principal Component Analysis and Latent Dirichlet allocation. Some areas have been advanced with unsupervised learning, such as user recommendation, analysing buying habits or grouping user logs.
- **Reinforcement Learning:** this type of learning is very similar to the way humans learn: trial and fail (Kaelbling et al., 1996). Unlike supervised learning, the algorithm does not have definite labels but only needs reinforcement signals in response to the outcome as negative or positive (Kotsiantis, 2007). The reinforcement signal usually comes from the environment as the agent (reinforcement learning algorithm) continuously interacts in the environment. By instructing the algorithm to associate good behaviour with positive signal and bad behaviour with negative signal, the agent learns to make fewer mistakes over time. Thus, reinforcement learning is behaviour driven. Common reinforcement learning algorithms are Q-Learning, Temporal Difference and Deep Adversarial Networks. Applications of such algorithms could be found in computer games, robotic machines and self-driving cars.

The categories of machine learning algorithms are not definite. In many cases, the line between different types of algorithms is vague, as a problem could be solved in various approaches depending on the way we state the problem in the first place. Moreover, an algorithm like Neural Network could be considered an architecture that could be applied in different machine learning paradigms (supervised or unsupervised learning).

In clinical areas, machine learning has been used to build systems to support illness diagnosis, prioritise optimal treatment and determine drug dosing (Scott et al., 2019). Some applications can perform faster and even more accurately than clinicians with better consistency on duplicated cases. With the availability of massive datasets, especially image related diagnosis data, machine learning and computer vision techniques have been used to support the diagnosis of various diseases such as diabetes, dermatology, cancer and ENT (Scott et al., 2019). Although the first use of neural network for the medical system could be found more than twenty years ago (Lo et al., 1995), in recent years, a key factor that contributes to the success of such system is the adoption of deep learning and convolutional neural network (Gibson et al., 2018). Deep learning is a subset of neural network, in which the structure of the network is inspired by the structure of a human brain. Each neuron in a network is a mathematical function that takes input from neurons in the previous layer, transforms the input and passes it to the next layer. A neuron is activated if the sum of all inputs received from the previous layers passes a predefined threshold. The activation means the original input contains a specific feature that could activate the neuron. The information from activated neurons is passed down to the following layers. This process continues to the second last layer, and the result is determined in the final layer. A neural network is constructed from multiple connected layers. Each layer contains many neurons. So, the neural network could be trained to detect very complex features that are hard to perceive from the raw input data. The term deep learning comes from the structure of the network and the number of hidden layers. Typically, more than three hidden layers are used in deep neural networks. However, in a simple problem, we could still see a single layer network.

When building a machine learning system with traditional methods, the data must be pre-processed manually through hand-crafted feature extraction, feature normalisation and selection. After that, an algorithm is selected (support vector machine, random forest or any other algorithm) is selected to perform the pattern recognition based on the extracted features (Castiglioni et al., 2021). The effort required for designing feature extraction methods and selecting the best fit classifiers are labour intensive and error prone (Wang et al., 2018). However, with deep learning approach, both feature extraction and pattern classification are performed automatically without manual intervention. Generally, in a deep neural network, lower layers (convolutional layer, pooling layer) are responsible for feature extraction, and higher layers (dense layer, recurrent layer) are responsible for pattern recognition. Building an accurate machine learning system in medical space also faces various challenges. First, the outcome of the algorithm depends greatly on the quality of input data. Collecting and labelling data for a large number of cases are tedious. Ununified quality, biased data collection, incorrect labelling or low volume of data could cause errors in system training (Scott et al., 2019). Second, a poorly designed algorithm could impede the outcome accuracy of the system. Finally, there is a lack of standards in assessing the safety and the utility of machine learning systems in reality, not to mention implementing and integrating such an algorithm in daily clinical workflows will require more liability clearance and privacy practices (Rajkomar et al., 2019).

2. Ear illnesses

The ear contains three main parts: outer ear, middle ear and inner ear (Mansour, 2019). The outer-ear consists of the pinna, which is the outside part of the ear, and the external auditory canal, which connects the outer ear and the middle ear. The eardrum (tympanic membrane - TM) is a thin cone-shaped membrane that separates the external ear from the middle ear (Isaacson, 2014). The middle ear anatomy includes three tiny bones (malleus, incus and stapes) that transmit the sound waves to the inner ear. There is also eustachian canal in the middle ear, which connects the middle ear with the nose. This tube helps to equalise the air pressure in the middle ear, which is essential for sound to be transferred properly. The inner ear consists of the cochlea, vestibule and semi-circular canals, which contain hearing nerve and balance receptors (Mansour, 2019).

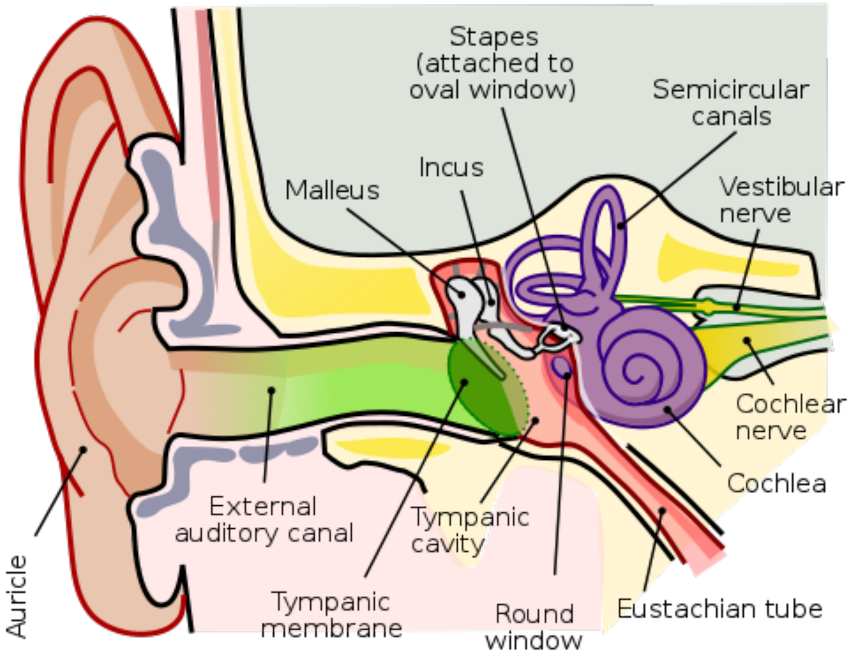


Figure 1: Anatomy of human ear

(photograph by Lars Chittka and Axel Brockmann, distributed under a CC-BY 2.5 license. (Wikipedia, 2009))

When we hear a sound, the sound waves travel from outside of the outer-ear through the external auditory canal and hit the eardrum (TM). The eardrum then vibrates and pass the vibrations to three ossicles bones. The ossicles amplify the sounds waves and send them to the inner ear, reaching the fluid-filled cochlea. In the organs on the cochlea, the vibrations are finally converted into electrical signals and sent to the brain for processing (Oxenham, 2018).

Middle ear infections have their root in the malfunction of the eustachian tube, which sits between the middle ear and the throat. In normal conditions, the tube helps to balance the pressure between the atmospheric pressure outside and the middle ear cavity. However, under illness conditions, the normal drainage of fluid is built up in the middle ear, behind the eardrum (Standford-Children's-Health, 2021). The circumstance leads to effusions and the growth of bacteria and viruses in the ear, which is the main reason that leads to Otitis Media. Middle ear infections could happen in three forms:

- **Acute otitis media (AOM):** the middle ear infection develops quickly, causing swelling and redness. Fluid and mucus are filled inside the middle ear, causing ear pain, fever or temporary hearing loss
- **Otitis media with effusion (OME):** Fluid and mucus continue to build up after the onset of AOM
- **Chronic otitis media with effusion (COME):** Fluids remain in the middle ear for a long time and happen again, without an actual ear infection. This condition does not cause severe illness by itself, but it could be a trouble if new ear infections come. (Standford-Children's-Health, 2021)

In order to diagnose the otitis media condition, the clinician could use otoscope and tympanometry test. An otoscope is a medical device that allows the health care provider to look inside the ear and view the outer ear and the tympanic membrane conditions. The output could be still images or videos. With the availability of low-cost handheld devices, this form of diagnosis is the most common form of assessment (Moberly et al., 2018). There is another form of otoscopy diagnosis, which is pneumatic otoscopy. With pneumatic otoscopy, a small suction of pressure is pushed inside the ear via an insufflator bulb and tube, which allows the device to record the eardrum mobility visually (Frumkin, 2018). Although pneumatic otoscopy provides higher accuracy in detecting the presence of middle ear effusion in comparison with simple otoscopy (Harmes et al., 2013), it is not frequently performed or taught by paediatric specialists (Frumkin, 2018). Another useful form of diagnosis is the tympanometry test. The test detects any changes in the vibration of the eardrum for different pressures pushed into the ear. The test is highly accurate in detecting the middle ear effusion; however, it is hard to perform on children as it requires the child to stay quiet and still. Other forms of tests are also performed to diagnose different forms of otitis media, such as computed tomography (Wang et al., 2020), acoustic reflectometry or tympanocentesis (Harmes et al., 2013). However, the main problem with all those tests is that they require extensive training to use the device and interpret the result. In a remote area, where access to tertiary ENT specialists is highly limited, the tasks of initial diagnosis need to be performed by community paediatricians and general practitioners, who are relatively inaccurate compared to the expert (Livingstone et al., 2019). A study shows that average diagnosis accuracies of paediatricians and general practitioners in differentiating Acute Otitis Media, Serous Otitis Media and normal ear conditions using video otoscopy are only 51% and 46%, respectively (Pichichero and Poole, 2005). Misdiagnosis of unconfident frontline workers makes the waiting lists in remote areas longer, which directly affects the timely access to proper treatment of the patients. So, there is a clear need to provide a system that could support the otitis media diagnosis process in those areas.

3. Machine learning in Otitis Media diagnosis

An important factor in determining various types of otitis media is locating and diagnosing the condition of the tympanic membrane (TM) in the middle ear (Lee et al., 2019). However, this process is highly subjected to clinician experience and the cooperation of the patients. In a study, Lee et al. (2019) proposed a method using Convolutional Neural Network (CNN) model for the tympanic membrane classification in detecting the TM side and the presence of perforation. The model was built with six simple layers (two convolutional layers with 32 filters, two max pooling layers and two fully connected layers). A class activation map (CAM) was also applied in the test phase of the training to visualise the feature variations of different parts of the input images under various lighting conditions. The results were 97.9% of accuracy in detecting the TM side, and 91.0% of accuracy in detecting the presence of perforation (Lee et al., 2019). It is also worth mentioning that the high accuracy could be achieved because the problem has been simplified to demonstrate the usefulness of ML algorithm. There were only two classes selected as output, and the data with postoperative conditions and lesions, which are generally more difficult to classify, were ignored. However, the results signify that a CNN architecture could be useful in classifying tympanic membrane abnormalities based on the otoscopic image.

Such tools could significantly reduce the requirement of clinician training and personal experience in the otoscopic examination.

In order to shorten training time and reuse a pre-trained model on a new problem, transfer learning could be used (Pan and Yang, 2010). Transfer learning, a design methodology, could be a great fit if the new problem does not have much data for efficient feature extraction and recognition during the training. It helps to achieve better performance in new problems by keeping the weights of the convolutional layer from pre-trained model and only adjusting the parameters of several last dense layers. It is common to utilise transfer learning with classification problems that use images as data input (Kaur and Gandhi, 2020). In otolaryngology field, a report that utilised transfer learning on Inception-V3 architecture could achieve an accuracy result of 76% on predicting the perforation on tympanic membrane (Habib et al., 2020). Other work reports the use of Xception model and MobileNet-V2 for training on a set of 10,703 high quality otoscopic images for three labels: normal, Acute Otitis Media and Otitis Media with Effusion. The accuracies for the two models were 97.45% and 95.72%, respectively. Those models were then used to give predictions on raw images taken from smartphone enabled wireless otoscope. They still achieved accuracies of 90.66% and 88.56% for Xception and MobileNet-V2 (Wu et al., 2021). These proof-of-concept models and their results suggest the promising application of machine learning and transfer learning in the field of otolaryngology. The novel system could be useful to support the screening process of ear disease in distance communities (Habib et al., 2020).

Advanced methods could be utilised with a CNN model to boost the accuracy of the classification system in otitis media diagnosis. In a recent report, Alhudhaif et al. (2021) proposed a novel multi-class system that used CNN with several cutting-edge techniques. The proposed model contains three main components: attention modules (a combination of channel and spatial model – CBAM), residual blocks and hyper-column techniques. The best model achieves an accuracy of 92.19% over five classes of otoscopic images. Sensitivity was 97.68%, and specificity was 99.3%. By utilising advanced methods, the model achieved shorter training time and superior results compared to existing pre-trained CNN models (AlexNet, GoogleNet and ResNets) and other machine learning models in other reports (Alhudhaif et al., 2021). The work not only suggests that CNN could be used to support the otitis media diagnosis process but also verifies that new advanced methods integrated into CNN model could enhance the efficiency of the model in the OM diagnosis area.

Not all approaches were utilising machine learning in diagnosing otitis media. Kuruvilla et al. (2013) suggested an interesting automated system built based on the validation of features set derived from the visual cues of the otoscopic image. The features could be understood by both otoscopists and engineers. They built a set of otitis media vocabularies, which described the colour, position and translucency of the tympanic membrane. Those features were then be used to construct a decision tree (otitis media grammar) based on the decision process used by real otoscopists. Such a system could be used to classify a tympanic membrane image into one of three categories: Acute Otitis Media, Otitis Media with effusion and No effusion. The proposed algorithm achieved 89.9% of classification accuracy, which was comparable with the diagnoses of expert otoscopists (Kuruvilla et al., 2013). The decision tree was also validated and compared with five automated classifiers with 5-fold cross validation setup, including correlation filter classification system, multiresolution classifier, SVM classifier, WND-CHARM classifier (Orlov et al., 2008) and Random Forest classifier. The results show that using decision tree on the input set provided the highest accuracy. The report provides valuable insights on how the otoscopic images could be understood and processed by an experienced clinician. Challenges with pre-processing the input image and possible solutions were discussed. The white-box approach proposed in this report can be of great benefit when building a fully automated system using machine learning for the classification of otitis media. In comparison with neural network approach, the decision tree is simpler and easier to interpret the result and the classification process (for most complex neural network models, it is not possible to interpret why the system gives a particular prediction at all). However, the tree structure is fixed and cannot adjust dynamically if more data with additional features become available later. As the number of feature increase, designing a decision tree also becomes more complicated as it is easy to get confused or derailed by the outliers ((Katz et al., 2014)). The neural network has fewer problems to overcome such complexities and could incorporate additional patterns effectively while still being able to perform classification correctly for previously unseen cases. However, neural network requires more training time and more expensive training resources as trade-offs.

CNN could also be used in classifying different forms of otologic abnormalities. Livingstone et al. (2019) proposed a CNN model to identify normal TM, TM with tympanostomy tubes and TM with cerumen impactions. The accuracies for each category were 93.3%, 86.7% and 84.4%, respectively. The input data was grey scaled, and some augmentation techniques were utilised (rotation, mirroring) before passing to the network. The augmentation has been used to increase the number of input data for the training and validation. The network had a small size architecture (three convolutional layers, with batch normalisation and a dropout layer of 40% fixed ratio). However, it could still achieve a reasonable accuracy in terms of target classification. The selection of a small size network could also prevent overfitting and increase the performance of the system, making it viable in integration with a low-cost image-capable otoscope that could be distributed to primary care providers (Livingstone et al., 2019). The report also discussed the difficulty of processing images from different sources (different resolutions, lighting levels or sizes of the otoscope tip). Those variants could greatly decrease the performance of the model, especially a simple model with fewer layers, which lack the capability to capture complicated feature correlations. Thus, the report only works with a single source dataset to demonstrate the feasibility of machine learning applications in the field.

Due to the development of low-cost handheld otoscopy devices, digital otoscopy images are usually the primary source of input for the diagnosis of different types of ear pathologies in remote areas (Moberly et al., 2018). However, only a limited number of studies had criticised the correctness of the diagnosis based on high resolution still images collected using handheld otoscope only. Moberly et al. (2018) had performed an experiment with 12 experts to review a large number of digital otoscopic images and provided diagnoses accordingly. The results were then compared with the ground truths based on clinical microscopic images with audiometry and/or tympanometry data. The results suggest that high-definition digital images of the eardrums could provide adequate information to make correct diagnoses for some pathologies. However, some diagnoses, such as middle ear effusion, are hard to decide based on the image only (Moberly et al., 2018). In such cases, the information conveyed in the image is not enough, and other input sources might be required.

Another report (Wang et al., 2015) suggests that tympanometry data was easier to use and could be accepted in GP diagnosis plan to review children with otitis media. However, tympanometry diagnosis has a higher cost, which could inhibit their wide usage. In general, both otoscopy image and tympanometry data should be used to increase the accuracy of the diagnosis. Another study also suggests that the combination of multiple diagnosis data could help in identifying otitis media conditions in children (combination of audiometry and tympanometry versus the use of pure tone audiometry alone) (Yockel, 2002).

Otoscopy image is not the only input for building a machine learning system that supports the diagnosis of Otitis media. In a recent report, Wang et al. (2020) proposed a deep learning tool that made use of temporal bone computed tomography scans to diagnose chronic otitis media. The framework contained two separate CNNs, one for extracting the regions of interest from 2-dimensional CT images, one for the classifications of chronic otitis media based on the extracted regions. The classification network was based on Inception-V3, a pre-trained CNN model. The validations results were compared with the decisions of a panel including six clinical experts. The accuracy in both binary classification task and multi-labels classification task of the model were higher than those of the panel (86% versus 81% for binary classification, and 76.7% versus 73.8% for three-class classification task). The system also provided superior consistency in duplicated cases (100% versus 81%) (Wang et al., 2020). The results imply a promising outlook in utilising machine learning in the ear disease diagnosis using CT images in comparison with a traditional approach of relying on clinical experts. The methodology also suggests the importance of proper feature extractions in developing a CNN model in the domain. In addition, the model only gave prediction on a single, two-dimensional CT image, while the clinical expert had access to all of the full-sized CT images of the cases when providing the diagnosis. This view signifies that if the model could incorporate additional information from adjacent CT slices, the accuracy of the model could be boosted to higher level (Wang et al., 2020). Another report that first utilised the characterised Wideband Absorbance Immittance (WAI) data to enable the automatic diagnosis of Otitis media with effusion using Machine Learning tools (Grais et al., 2021). After producing the 2D WAI image by interpolating the pressure axis, five different machine learning classifiers were built, including K-nearest neighbours classifier, Support Vector Machine, Random Forest, simple Feedforward neural network and a Convolutional Neural Network. The accuracies of all models were around 75% to 80%, with the CNN networks having slightly better results in comparison with other models. The report was not only providing clearer guidance to the practitioner in interpreting the WAI data, but also proving that Machine Learning could be a potential tool for building an automated diagnosis system from complex data in ENT clinics.

Besides machine learning, other methods also are used to automate the diagnoses of otitis media. (Myburgh et al., 2016) developed a computer-based image analysing system that utilised decision tree technique for the task. Image processing technique was used to extract features from input images for predefined diagnosis. A decision tree was then built to classify images into one of the five diagnosis groups. Experienced specialists would then re-evaluate the results. The authors also built a low-cost device to capture the image and run the trained decision tree on a personal notebook for evaluation. The total expense of low-cost device was around $84, which is at least five times less expensive than an entry-level commercial video-otoscope. The model output was comparable with a high degree of agreement (80.6%) to the two otologists' correspondences across five diagnostic categories (Myburgh et al., 2016). Interestingly, the result of low-cost device otoscopy running in low-end notebook could achieve a similar result of 78.7%. This achievement suggests that machine learning approach could be provided in automated OM diagnosis system in underserved communities in a fairly affordable manner. In general, the report signifies that techniques like computer vision and decision trees could be effectively utilised to support the process of otitis media diagnosis. However, unlike neural network approach, the major drawback of decision tree is that the structure of the system is fixed and need to be redesigned if more training images are available. In a later report, the authors compared the decision tree method with the neural network classifier approach (Myburgh et al., 2018). The accuracies of the neural network for five diagnosis groups were higher than those of the decision tree (86.84% overall accuracy for the neural network system). The feature extraction algorithm was also be redesigned. Several methods regarding image processing were applied, such as cropping to a fixed size and blur detection using a variation of the Laplacian. The system was built to run on an Android system using open-source software. The Android application allows users to process a video capturing the tympanic membrane and pre-process the image before sending the input to a server, where the feature extraction and classification are executed. The deployment model offers great flexibility and could be used to streamline the diagnosis of otitis media related illness in remote areas.

Chan et al. (2019) even go further in developing an effective screening tool to detect the presence of middle ear fluid using only the microphone and speaker of a smartphone. The smartphone speaker would be used to play an audio with a specific frequency. The reflected waves from the eardrum were then collected and analysed. If the ears were infected with otitis media with effusion or acute otitis media, the eardrum vibration would be restricted, which resulted in a narrower acoustic dip in the frequency domain. A simple paper funnel was used to ensure the transmission of sound to and from the eardrum was not interfered. The system used a logistic regression machine learning model, which is computationally inexpensive, to classify the waveforms. The leave-one-out validation and 10-fold cross-validation were performed to train the model. The system was tested on multiple smartphone platforms, and the results were comparable to established performance metrics of tympanometry and pneumatic otoscopy. The system even performed better than commercial acoustic reflectometry, which required expensive hardware (Chan et al., 2019). The result suggests that the classification of middle ear fluid is an important factor in determining both acute otitis media and otitis media with effusion, and a simple, low-cost screening tool could be possible with the help of machine learning model.

In addition to building a complete automated system that supports the diagnosis of otitis media, machine learning could also be used to develop tool that could provide similar looking images for a predefined diagnosis. Such utility could be helpful to increase the confidence of clinicians in the diagnosis process. Camalan et al. (2020) proposed a system that acted as a content-based image retrieval system using deep learning called OtoMatch. The system was developed from Inception-Resnet-V2 pre-trained model. To prevent overfitting, the first 820 layers were frozen, and the last three layers were retrained with the otoscope images. The output of the fully connected layers was used to calculate the distances to determine the similarity of input image with predefined labelled images. The system was trained on a 10-fold cross validation, and could achieve the accuracy of 80.58% and maximum F1 score of 0.90 for a three labels problem on a database of 454 eardrum images. The performance for this approach was higher and more stable in comparison with other traditional CBIR methods using handcrafted features. The system was a novel application of applying deep learning for image retrieval in the context of tympanic membrane diagnosis (Camalan et al., 2020). More importantly, this report proposes a common approach that could be applied to any deep learning architecture to convert a classification model to an image retrieval system. In later research, the researchers further experimented with the combination of both right and left ears examination to improve the accuracy (Camalan et al., 2021). The model's output in the previous report was used to generate a pair feature vectors of two eardrum images. The final classifier part was developed using Tree Bagger algorithm. A 3-fold validation was used to validate the model because the number of inputs was small. The proposed accuracy had been increased from 78.7% to 85.8%, which were quite promising. The work suggests that combining the data from both ears in the classification could result in better outcomes (Camalan et al., 2021). That process also mimics the way a clinician evaluates otoscopy images. In fact, otoscopy images are usually available in pair as the patient are expected to record both ears image in the screening procedure.

With the availability of cloud machine learning platforms, one could leverage the capability of machine learning model without a steep learning curve in the AI world. In a recent work, Livingstone and Chau (2020) obtained otoscopic images from open access repository with 14 labels (including acute otitis media and serous otitis media). After pre-processing, the images were uploaded to Google AutoML platform to train a multilabel image classification system. The result on the test set was compared to the diagnosis of paediatricians, otolaryngologists and family doctors. The accuracy of the algorithm was 86.1%, compared with 58.9% of human diagnoses (Livingstone and Chau, 2020). There were many misclassifications of classes with limited training data. The problem could be resolved partly by collecting sufficient data on all labels. In some cases, the information from a single image is not enough even for highly expert clinicians. The report suggested that different types of diagnosis data and clinical context should be given to reach an optimal decision.

There are many ways to build a classification model for a given problem. One could apply fusion techniques to enhance the accuracy of multiple models. Zeng et al. (2021) had developed nine different models from the data set of 20.542 otoscopy images for eight ear disease categories. Random X and Y flip horizontal and vertical were performed as the data augmentation method. Two best models built using transfer learning from DensNet-BC169 and DensNet-BC1615 were selected based on performance after the evaluations to ensemble a single classifier. The ensemble technique was selecting the maximum of the probability outputs for two classifiers as the final result. The average accuracy was 95.59% (Zeng et al., 2021). Given the promising results, the study suggests that ensemble techniques could be applied in the ear disease diagnosis by combining the outputs of multiple machine learning models training on the same data set.

In order to increase the accuracy of classification using tympanic membrane images, manual extraction of image is usually performed by skilled otolaryngologists before feeding the processed image to the training process. However, this procedure could be automated using CNN. In a recent report, a two-stage classification was proposed that used Class Activation Maps (CAMs) to identify important segments of the input image (Cai et al., 2021). The pipeline consisted of two stages. The first one, the main classifier, received the whole image and found important parts of the image by analysing the CAM and picking up discriminative segments from the result heat map. The second model, which was a focal one, only needed to classify based on the important segments. Classification results from the main and focal models were merged by averaging the two outputs. The result of the whole system was consistently higher than that of the main classifier only. The overall accuracy was 93.4%. This level of accuracy was equivalent to those of associate professors in otolaryngology that participated in the evaluation (Cai et al., 2021). Another report utilises a continuous improvement model by combining a normal classifier (offline) with feedback from real-time (online) classifier (Wang et al., 2015) to predict the condition of the tympanostomy tube for otitis media. Support Vector Machine algorithm was used on all models. The offline layer is a collection of three cascaded classifiers; each detects a different set of features of the original image. The online classifier allows the user to adjust the system by providing feedbacks on wrong predictions, which are then recorded by the system in a database and used to retrain the model. The offline model could be able to achieve a result of 90% in predicting the otoscopic image with and without a tube. The online model then improves the classification accuracy by 3-5% based on the feedbacks of additional images. The report proposed an interesting approach to continuously improve the effectiveness of machine learning based diagnosis support systems.

Besides the difficulty of feature extraction with image related problems, machine learning model also suffers from the lack of training labelled data. Some reports have tried to address both two problems using transfer learning (Kai et al., 2015). First, the ImageNet dataset of 15 million images is learnt through Alexnet (an image recognition CNN model). Then, for each image in the otitis media dataset, the image was fed into the model, and the features from three selected inner layers were extracted to produce a feature vector of the original image. The feature vectors, which contained high level abstraction of the original image, were trained using Support Vector Machine classifier with RBF kernel. Fusion technique to combine automated features with handcraft features was also applied. The accuracy of the system was 88.5%, which is considerably higher than previous trials which used feature extractions from domain experts (Kai et al., 2015). The study suggests that transfer learning could solve the issue of feature extraction and the scarcity of otitis media datasets in building an efficient model that supports the diagnosing process. The same method can be applied to solve the problem with medical image analysis in general.

METHOD

1. Data Sources

1.1 Test data

In the first phase of the project, a public dataset of the eardrum from Zenodo database was used to build proposed models of the system (Camalan et al., 2020). The same dataset has been utilised in the development of OtoMatch, a content-based eardrum image retrieval system using deep learning (Camalan et al., 2020). The data set contains 454 labelled images, collected from ENT and primary care clinics for both children and adult patients at Ohio State University and Nationwide Children's Hospital in Ohio, US. The images are colour images in JPEG format of different resolutions. There are three labels for the dataset: 179 images of middle ear effusion, 179 images of normal ear and 96 images of the ear with tympanostomy tube condition. The label data was provided in a separated CSV file which maps the file name with the name of the corresponding label.

1.2 Swimming data

The swimming dataset comes from a prospective 3-year cohort study, in which the researchers assessed whether access to saltwater swimming pools could reduce hearing loss and ear illness in children of aboriginal communities of South Australia (Sanchez et al., 2019). There are a total of 2107 children (ages 5 to 18) were assessed on multiple visits. Assessments and data recorded included audiometry screening, tympanometry diagnosis information and video otoscopy data for both left and right ears of each child. The data that is used in this report contains 2085 visits of 813 children. All the labelling data is stored in an excel file. There is a total of 4454 otoscopy videos of left and right ears. They are all in WMV format and have a frame rate per second of 30. The otoscopy diagnosis has a total of 8 labels, including Normal, Healed Chronic Otitis Media (Healed COM), Inactive mucosal COM, Inactive squamous COM, Active mucosal COM, Active squamous COM, Fluid (Otitis Media with Effusion) Tympanic Membrane intact and Pus (Acute Otitis Media) Tympanic Membrane intact. The tympanometry results have a total of 10 labels. There are many more diagnosis data and labelling. However, this report utilises video otoscopy and tympanometry as data sources, as they are the two most common forms of diagnosis available for otitis media. Besides the multi-labelling of otoscopy and tympanometry diagnosis, the output could also be collapsed to binary label for otoscopy (pass, fail) and three labels for tympanometry data (pass, fail or undetermined). The collapsing of output is crucial in combining multiple sources of diagnosis data in a single model pipeline. It is worth noting that the data was not collected for the purpose of building an automated analysis system using machine learning. Furthermore, the video dataset is huge (approximately 40.8 Gb) with varying qualities, which potentially introduces numerous challenges associated with data processing.

2. Models Selection

As suggested in the literature, Deep learning has been widely used in building otitis media diagnosis support systems with high accuracy. In this report, two of three models selected to train the data and validate the results belong to the deep learning umbrella: Multiplayer Perceptron and Convolutional Neural Network. Support Vector Machine model was also selected in order to comparatively evaluate the results given by deep learning models.

2.1 Support Vector Machine

Support Vector Machine (SVM) is a widely used model in classification tasks. The objective of support vector machine is to find a hyperplane in N-dimensional space (N is the number of features) that effectively separate different categories of data points (Noble, 2006). Effective separation is the one that yields the maximum margin or the maximum distance between data points of both classes. Maximising the margin distances could help to classify future data points with more confidence (or higher accuracy for unseen data). The input data will be transformed into a required form of processing data by a kernel function. Generally, the kernel functions do the transformation by adding additional dimensions to the original data, making it linearly separated in a higher dimensional space (Noble, 2006). The hyperparameters that could be tuned with SVM model are regularisation parameter (C value), kernel type to be used in the algorithm, the kernel coefficient (for rbf, poly and sigmoid), etc.

2.2 Multilayer Perceptron

Multilayer Perceptron (MLP) is a deep, artificial neural network, which composed of more than one perceptron. The network consists of one input layer, one output layer and an arbitrary number of connected hidden layers (Gardner and Dorling, 1998). While a single perceptron is a linear classifier, a layer of perceptron is capable of approximating any continuous function. Training an MLP model is performing the forward pass and backward pass multiple times to find the best parameter of the network. In the forward pass, the signals are transmitted from the input layers through hidden layers to output layers. After the forward pass, the current prediction is compared against the ground truth labels. In a backward pass, partial derivatives of cost function (errors) with respect to weight and biases values are back-propagated through the network. The weights and biases at each perceptron are then adjusted to minimise the error. The parameter that could be tuned with MLP model includes the number of hidden layers, the number of nodes for each layer or the number of propagation iteration.

2.3 Convolutional Neural Network

Convolutional Neural Network (CNN) is the standard of state-of-the-art machine learning model and has been used to solve many real-world problems. From the first work of the modern convolutional neural network in the 1990s (Lecun et al., 1998) to the massive success of AlexNet model in recognising a huge number of pictures in 2012 (Krizhevsky et al., 2017), CNNs has become the backbone of many computer vision and artificial intelligence applications. The main concern in building a CNN model for a particular problem is choosing the right structure of the network (number of convolutional layers and fully connected layers and the way they are interconnected) and balancing between the output metrics and the complexity of the model.

The model used in this report has been built from the inspiration of classic CNN architectures (AlexNet, VGG-16), which utilise a set of convolutional layers and max pooling layers at the start of the network, followed by fully connected layers at the end. The architecture starts with smaller filters to capture local information and continues with larger filters to represent more global features and also reduce the feature space. A typical kernel size of 3x3 has been used. The aim of the report is to produce a prototype of CNN based system that could efficiently support otitis media diagnosis using existing data, so a relatively simple architecture has been proposed. Also, due to the limitation of hardware, training time and scope of the work, other state of the art CNN architectures have not been fully explored. Such analysis should be subjected to future works.

3. Data processing

3.1 Test Data

The dataset taken from Zenodo database has an unequal number of data in one of the three labels (96 images with tympanometry tube condition, in comparison with 179 images in normal and effusion category). In order to balance the number of images in each group before training, data augmentation technique has been used to increase the number of images in the smaller group. The ImageDataGenerator utility from keras has been utilised for this task (keras). Due to the nature of using the otoscope, the tympanic membrane could be captured at an arbitrary angle or zoom level depending on how deep the device has been put into the ear. Thus, the data augmentation includes a random zoom in the range of 0.8 and 1.2, combined with random horizontal and vertical flips of the original image (Figure 2). The augmented images had then been saved with the original images. After the augmentation phase, each label now contains 179 images, which results in a total of 537 images that could be used for the training pipeline.

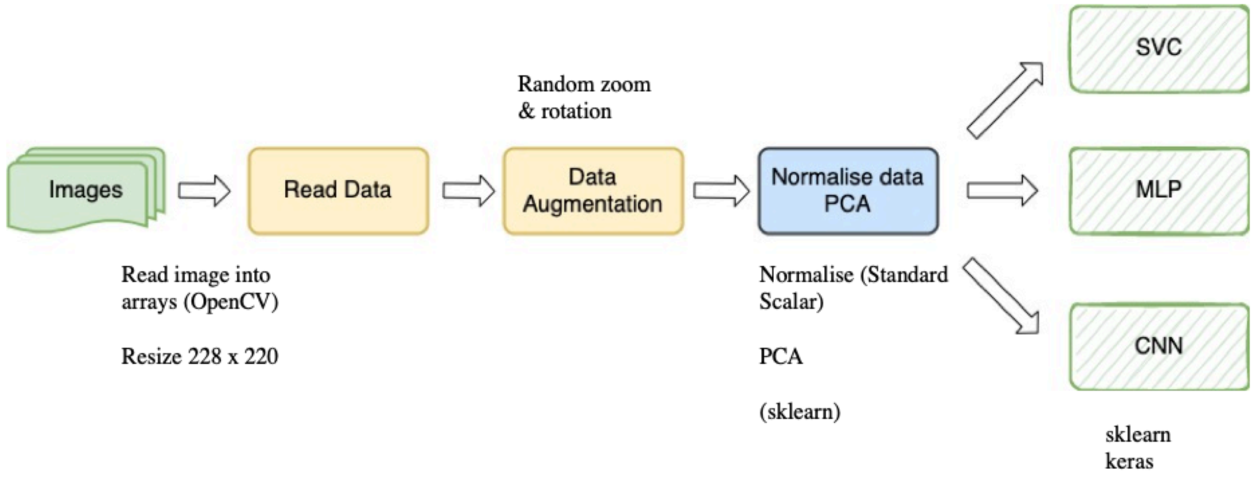


Figure 2: Input processing for public dataset

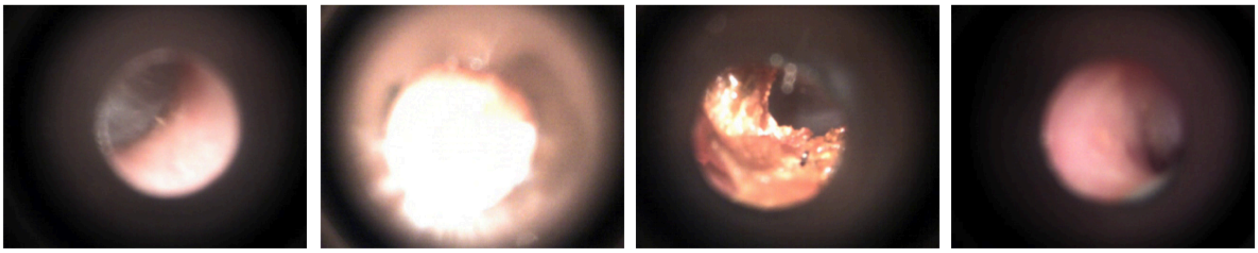
All the images are read into array using OpenCV library (OpenCV). As the training model requires all input to have the same shape, all images are resized to a fixed smaller size of 220x228. The scaling has been decided by calculating the average ratio of width and height from all the images in the dataset and a reasonable choice of the resized image, so the model does not have to train on very big features set.

To improve the overall performance of the classification and minimise the bias that the feature could introduce in the training (Singh and Singh, 2020), normalisation technique has been applied to the data. Mean and standard deviation of each feature is calculated and stored, then the original feature data has been scaled to have 0 means and unit variance in the output of the normalisation process. The normalisation technique has been applied before the training of all models (SVM, MLP and CNN). For SVM and MLP, Principal Component Analysis has been performed to extract features before feeding into the training. The aim is to reduce the dimensionality of the dataset, by transforming a large set of variables into a smaller one that still retains most of the information provided in the original data. The main idea behind PCA is to find a list of principal axes vectors that best describe the dataset. The direction of the vector represents the component of the data, while the length of the vector is determined by the variance of the data when projected onto the axis. The more scatter the projections are on the vector, the more important that axis is in capturing the distribution of the data (more information is captured in that direction of the vector). By keeping a number of most significant components, we could create a reduced dimensional dataset and make the training faster. However, we have to trade off the overall performance as some of the information is lost during the PCA process. In order to select the best number of features that fit the problem, Grid Search technique has been used, and the number of components that should be kept by PCA is included as one of the hyperparameters of the whole pipeline that needs to be tuned.

3.2 Swimming data

The swimming data comes in two parts, the video part and the metadata part (all the labelling) in excel format. Each record in the excel corresponds to one visit, defined by an EDID that is unique for each participant and a visit number. The record also contains the date and time of the visit and the date-time of the otoscopy diagnosis for left and right ears. The date-time of otoscopy diagnosis is also used as the name of the video file, so each video could be mapped to a record in the excel file. There is a total of 4454 distinct videos. After removing all duplicated video files and invalid labelling records (records with missing video mapping, invalid video file names or invalid results – contains more than one label for the diagnosis), there is a total of 1,776 visits and 3,552 videos left for processing.

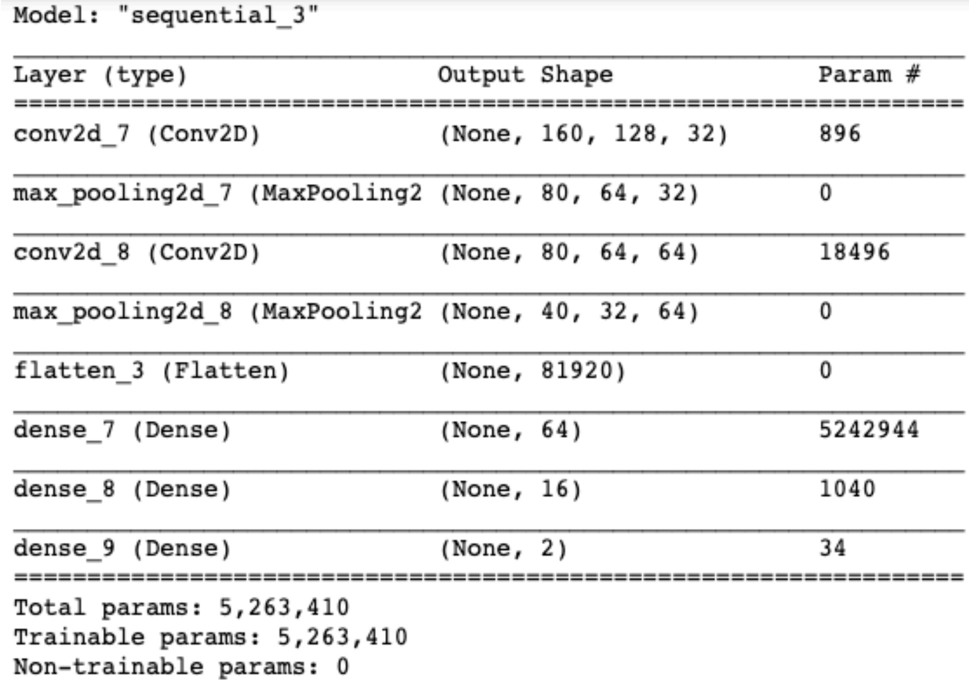
The model will need still images for training, so video frames will need to be extracted from the video. After manually analysing some of the videos, the first part and the last part of the video should be eliminated from the extraction, as they are usually capturing the devices putting in and pulling out from the ear. Most of the clear frames of the tympanic membrane are in the middle of the video. So, in the image extraction process, each video is loaded in the memory using OpenCV library (version 3.4.2). The first 18% and the last 12% of the video duration are skipped. For the rest of the video, an image is captured and saved at every 10% interval of the total video duration. This process results in approximately six to seven images extracted per video and a total of 22,458 still images. However, not all the images could be used for training. The quality of the videos is not consistent, and there is a substantial number of the extracted images that did not contain any useful information regarding the tympanic membrane and the middle ear condition (blur, too bright or too dark as the TM is totally obscured by cerumen or other obstacles (Figure 3))



(a) Outer ear (b) Too bright (c) Obscured by cerumen (d) Blurry

Figure 3: Extracted frames that should be filtered

In order to solve this problem, a subset of the total images is selected for manual filtering. The subset contains 3,368 images, which is 15% of the total images. In which, 1,301 images were removed (38,6%) and 2,607 images were kept (61,3%). A simple CNN model was built to solve the filtering problem. The model contains two pairs of convolutional layers, followed by a max pooling layer. After that, the data is flattened out and connected to 3 fully connected layers. The architecture of the filtering model is given in Figure 4.



```
Model: "sequential_3"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_7 (Conv2D)            (None, 160, 128, 32)      896
_________________________________________________________________
max_pooling2d_7 (MaxPooling2 (None, 80, 64, 32)        0
_________________________________________________________________
conv2d_8 (Conv2D)            (None, 80, 64, 64)        18496
_________________________________________________________________
max_pooling2d_8 (MaxPooling2 (None, 40, 32, 64)        0
_________________________________________________________________
flatten_3 (Flatten)          (None, 81920)             0
_________________________________________________________________
dense_7 (Dense)              (None, 64)                5242944
_________________________________________________________________
dense_8 (Dense)              (None, 16)                1040
_________________________________________________________________
dense_9 (Dense)              (None, 2)                 34
=================================================================
Total params: 5,263,410
Trainable params: 5,263,410
Non-trainable params: 0
```

Figure 4: CNN model architecture of filtering model

The accuracy of the filtering model was 84%, precision, recall and, f1-score was 83%. The filtering model then was used to classify the whole data set of 22,458 extracted images. A total of 7,227 images were removed by the classifier, and 15,231 images were left for further processing.

The numbers of images in each category are given in Table 1.

Table 1: Number of samples in each category after filtering - Swimming dataset

Label	Numbers
Normal	**2,415**
Healed COM	**5,952**
Inactive mucosal COM	**2,229**
Inactive squamous COM	**1,312**
Active mucosal COM	**1,872**
Active squamous COM	**4**
Fluid (OME) TM intact	**1,447**
Pus (AOM) TM intact	**0**
Total	**15,231**

There is no sample in Pus (AOM) TM intact group and only 4 in the Active squamous COM group. So, these two groups are removed from further processing. Another problematic category is Healed COM, which described a situation where the ear had some illness in the past but had been healed. However, without historical records and other types of diagnosis information, the sample in this category is tricky and could be easily misclassified as other types of illness. For this reason, Healed COM is also removed in the multi-label classification problem. After those considerations, there are five groups, with a total of 9,275 samples are selected for building models for multi-label classification problems (Table 2).

Table 2: Number of samples in five categories selected for training - Swimming dataset

Label	Numbers
Normal	**2,415**
Inactive mucosal COM	**2,229**
Inactive squamous COM	**1,312**
Active mucosal COM	**1,872**
Fluid (OME) TM intact	**1,447**
Total	**9,275**

Apart from the multi-labels classification, there is an attempt to build the models for binary classification problem of the otoscopy images given in the database. The two available labels are Pass and Fail for each video (and the relevant extract still images). Binary label also introduces the probability of combining otoscopy data with tympanometry data, which is already available as pass, fail and undetermined. After consulting with

ENT researchers, the translation from multi-label result to binary result are: Normal and Healed COM are grouped into pass, while all other labels are grouped into fail group. The numbers of samples in each group are given in Table 3:

Table 3: Number of samples in two categories for binary classification - Swimming dataset

Label	Numbers
Pass	**8,367**
Fail	**6,864**
Total	**15,231**

There are two types of label for tympanometry data in the original excel file. The first one is more accurate as it contains ten groups, which could be collapsed into one of three values: A-pass, B-fail and C-indeterminate. However, values for this label contains many records that are marked as not performed or otherwise blank (missing data). Filter the original otoscopy images by records with valid tympanometry data also leads to a very skew dataset. In order to simplify the problem, the binary label of tympanometry data has been used (Pass and Fail). Filtering the data by this label gives a total of 15,057 records with both otoscopy images and tympanometry diagnosis. There are 8,923 records in normal category and 6,764 records in abnormal category (Table 4).

Table 4: Number of samples in two categories for binary classification with tympanometry data - Swimming dataset

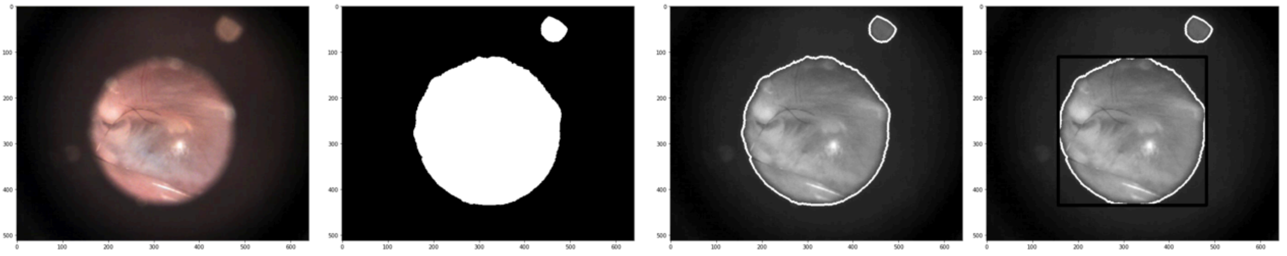
Label	Numbers
Normal	**8,293**
Abnormal	**6,764**
Total	**15,057**

A summary of all datasets used in different classification problems are given in below:

Table 5: Summary of experimental data sets

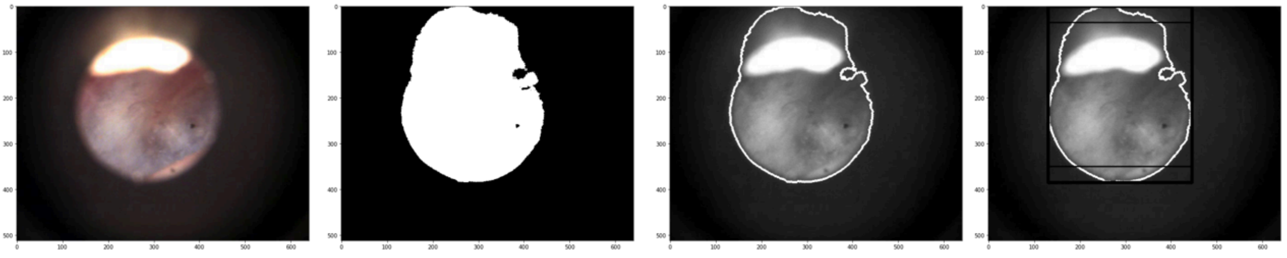
Data Source	Number of labels	Data Type	Data size
Public Test Data	3	Still, high quality images	454
Swimming Data Set	5	Extract frames from videos	9,275
Swimming Data Set	2	Extract frames from videos	15,231
Swimming Data Set	2	Extract frames from videos and tympanometry data	15,057

The next challenge is that the extracted still images, although share the same resolution of 640x512, have very different qualities. This comes from the fact that the videos have been captured using different devices by different clinicians over a long period. Normally the region of interest is a circle in the middle of the image, which captures the tympanic membrane. The region outside the circle is black. In order to extract the ROI from the image, binary thresholding has been applied in the grey scale version of the original image. Threshold value of 64 has been selected after several trials on random images. The result of this thresholding is a white circle (or some oval shape, depending on the light condition near the edge) on a black background. Then, using OpenCV, all contours of the thresholding image are identified, and a bounding box of the largest contour is selected. Applying the bounding box to the original image would give the ROI of the image. For high quality sample with clear shape of the circle, the process has extracted perfect ROI (Figure 5). However, for images with high intensity level pixels near the circle's edge (the lighting condition of the images is terrible, which also directly affects the information captured inside the ROI), the resulting ROI is a non-squared shape that covers unnecessary information (Figure 6). Non-squared shape could also affect the training process if, later on, the images need to be resized into the same shape, as the resized image might contain distorted elements. Other edge detections technique like Canny edge, Hough circles have been tried without success. In the end, a simple method has been used to capture a perfect square of the ROI. The square of ROI will have the size of the smaller edge of the bounding box while sharing the same centre as the bounding box. This selection will not affect the ROI capturing of high-quality images, and it also helps to extract a perfect square of ROI for problematic samples (Figure 6).



(a) Original (b) Thresholding (c) Draw contours (d) Bounding box

Figure 5: Cropping image - good example



(a) Original (b) Thresholding (c) Draw contours (d) Bounding box

Figure 6: Cropping image - bad example

4. Model training

All experiments were performed using a personal laptop, which had 2.8 GHz Quad-Core Intel Core i7 CPU, Intel Iris Plus 655 1536MB Graphic Card and 16 GB DDR3 RAM.

A split train/test set technique was used in model training and evaluation. The original dataset was divided into two disjoint subsets: a training set to fit the model and a test set to evaluate the model performance. The train set was bigger than the test set so that the model could learn the dataset effectively and generalise better. The test set was used to simulate the behaviour of the trained model on unseen data. In order to keep the variance rate low in the results, a testing set size of 30% was chosen for the validation of all the experiments.

4.1 Support Vector Machine

After normalisation and PCA has been performed on the input data, the training set (of 70%) has been put into the training process. Stratified sampling on the output has been performed to ensure the training set and the test set have the same percentage of each group sample in comparison with the original dataset. In order to select the best hyperparameters for SVC pipeline, GridSearch utility in sklearn package has been used. The following parameters have been tuned with several values:

- Number of components for PCA: 10, 50 and 100
- SVM kernel: rbf, linear and sigmoid
- C value of SVM: 0.1, 1 and 10

The score that GridSearch used to compare the performance of each model is the accuracy. In order to achieve better validation of the model, GridSearch was configured to run a five-fold cross validation on each of the hyperparameter combinations on the training set. The mean of five scores (accuracies) was then calculated and used to select the best model. Stratified technique is also applied as the number of samples on each category are different. After the best group of hyperparameters has been found, the model with those parameters has been trained again on the whole training set and then validated against the test set.

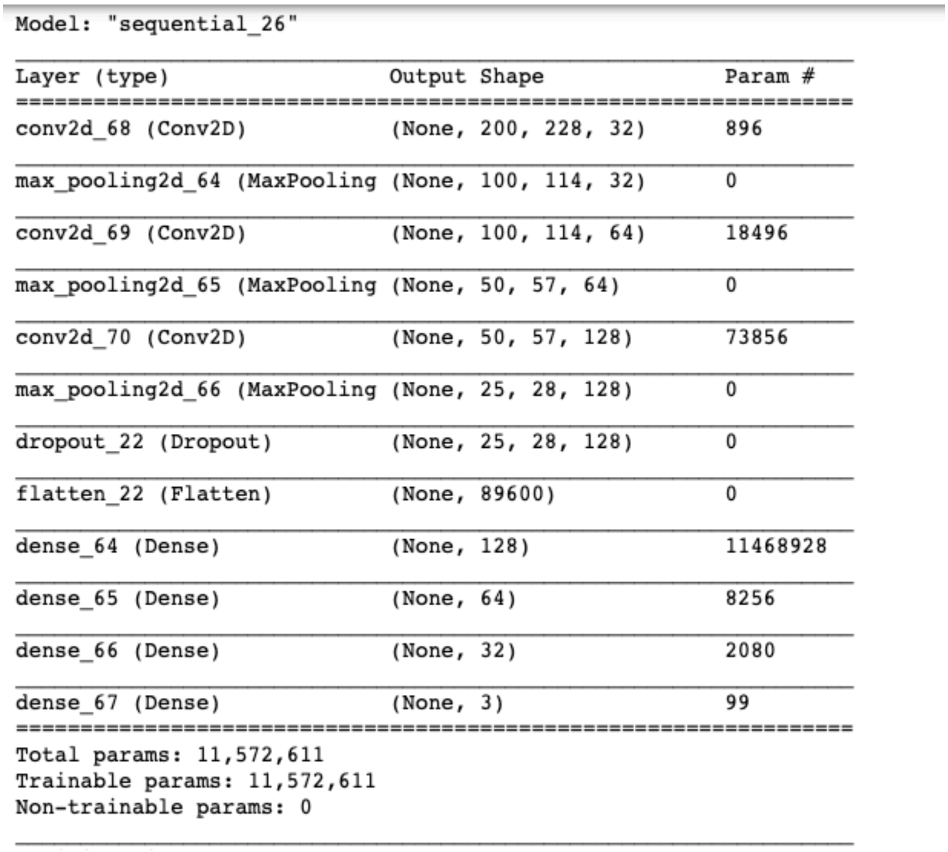
4.2 Multilayer Perceptron

Similar pipelines and approaches have been used to optimise the hyperparameters of MLP model. The parameter includes:

- Number of components for PCA: 10, 50 and 100
- Solver: sgd and adam
- Number of iteration that the solver iterates until convergence or this number has been reached: 500, 1000
- Hidden layer size: number of neurons in the single layer of MLP: 50, 100 and 250

4.3 Convolutional Neural Network

Input data has been normalised before feeding into CNN model. PCA is not needed as the features will be extracted using the convolutional layers in the network. The architecture of the CNN model consists of three convolutional layers with filter sizes of 32, 64 and 128. The filter size of convolutional layers (or the dimensionality of the output space after each convolutional layer) increase as the network tries to capture more complex patterns and combination of those patterns while trying to learn about the problem. Each convolutional layer is followed by a max pooling layer to reduce the size of output feature map while pertaining the most prominent features of it. After the convolutional steps, a Dropout layer is added to avoid overfitting of the problem. Finally, there are three fully connected dense layers at the end of the network to perform the classification based on the features extracted in previous layer. The number of neurons in the last layer represents the number of labels that model trying to predict (3 for the test data, 5 for swimming data with multi-label problem and 2 for the binary classification problem). The architecture of the CNN model is given in Figure 7.



```
Model: "sequential_26"
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
conv2d_68 (Conv2D)           (None, 200, 228, 32)      896       
_________________________________________________________________
max_pooling2d_64 (MaxPooling (None, 100, 114, 32)      0         
_________________________________________________________________
conv2d_69 (Conv2D)           (None, 100, 114, 64)      18496     
_________________________________________________________________
max_pooling2d_65 (MaxPooling (None, 50, 57, 64)        0         
_________________________________________________________________
conv2d_70 (Conv2D)           (None, 50, 57, 128)       73856     
_________________________________________________________________
max_pooling2d_66 (MaxPooling (None, 25, 28, 128)       0         
_________________________________________________________________
dropout_22 (Dropout)         (None, 25, 28, 128)       0         
_________________________________________________________________
flatten_22 (Flatten)         (None, 89600)             0         
_________________________________________________________________
dense_64 (Dense)             (None, 128)               11468928  
_________________________________________________________________
dense_65 (Dense)             (None, 64)                8256      
_________________________________________________________________
dense_66 (Dense)             (None, 32)                2080      
_________________________________________________________________
dense_67 (Dense)             (None, 3)                 99        
=================================================================
Total params: 11,572,611
Trainable params: 11,572,611
Non-trainable params: 0
_________________________________________________________________
```

Figure 7: CNN model architecture for training

4.4 Combining models

In an experiment to combine different types of diagnosis data in a single pipeline, SVM was chosen as the base model. The reason is that SVM is generally faster to train and could achieve higher performance metrics in most of the trials in this report (Results and analysis). Otoscopy images are normalised and go through a feature extraction process before being fed to SVM model. The results (categorical or probability values) are then merged with a binary-label tympanometry data of the same sample. This two-features input acts as the data for the second model, which could be a simple multilayer perceptron model or a decision tree (Figure 8). The performance on the validation set of the second model after training with combined data is then compared with the performance of SVM model on the otoscopy data alone. The results will show whether using more diagnosis data could boost the overall performance of the system.

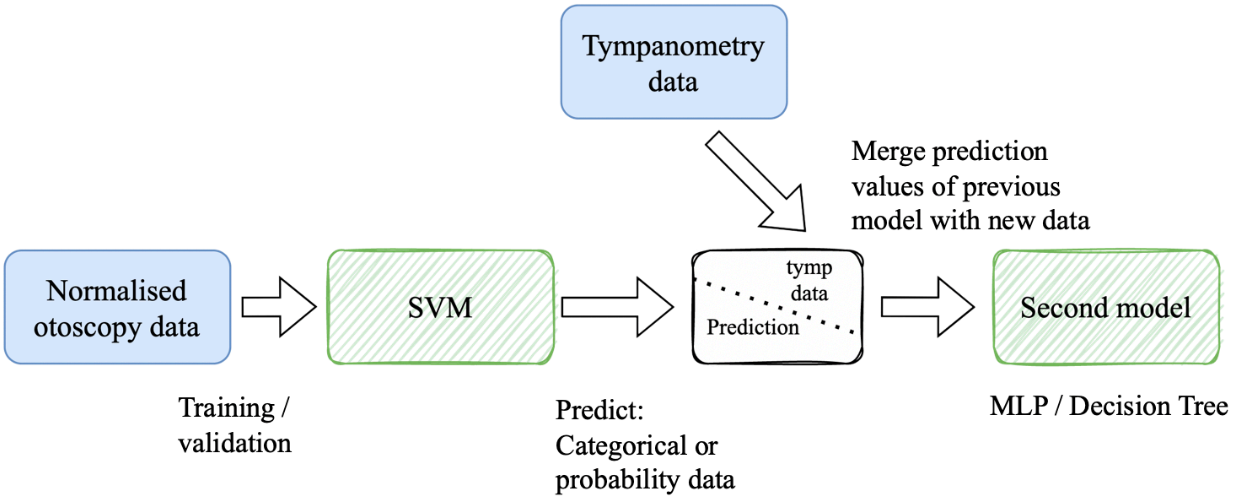


Figure 8: Combine models and different data sources

RESULTS AND ANALYSIS

1. Metrics

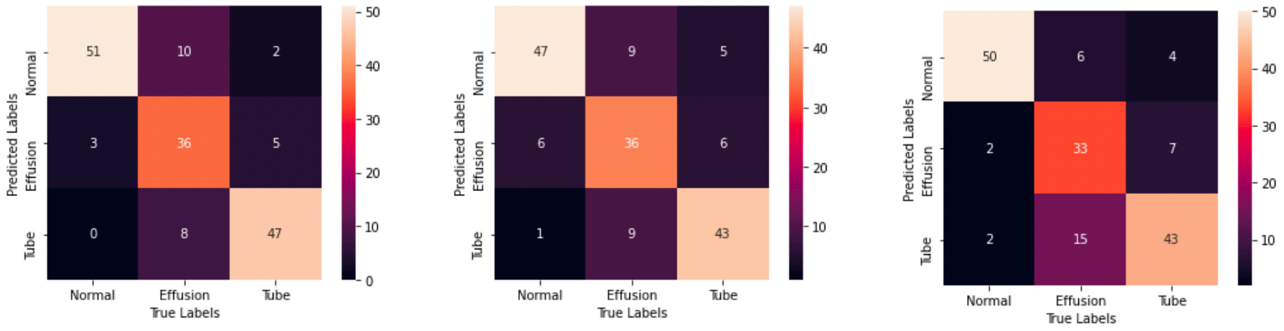
The primary performance metric used throughout the process is accuracy (in GridSearch hyperparameter tuning, model training on the training set and validation on the test set). However, in an otitis media diagnosis system, the costs of false positive and false negative cases are also considerable. False positive (case is normal, but has been diagnosed with illness) could cause more burden on the capacity of the ENT clinics, make the waiting list in rural areas longer and directly affect the time a real illness case could get any treatment. On the other hand, false negative (case has illness but has been diagnosed as normal) could cause serious problems as the patient is not going to receive any further assessment and treatment until the symptom gets worse. Thus, precision and recall metrics are also be monitored for each model. Another important metric is F1-score, as it seeks the balance between precision and recall on its calculation. F1-score is also a good indicator of a good model because, in the swimming dataset, most of the problems have an unequal number of samples in each category. However, the problem with both precision and recall (and F1-score as it is derived from precision and recall) is that they ignore the True Negative, or the ability of the model to correctly identify a negative case. Powers (2008) proposed informedness as an unbiased metric to address the problem. Informedness quantifies how informed the prediction is and indicates the probability that the prediction is informed versus random guessing (Powers, 2008). A value of 1 implies the model is fully informed about positives and negatives (every value of positive will be identified as positive, every value of negative will be identified as negative). A value of -1 implies the opposite, while a value of 0 is similar to random guessing, indicating a bad system. Informedness is the only metric that measures how similar a classifier is compared to random guessing (Chicco et al., 2021). However, in this report, a similar statistical metric called Matthews correlation coefficient (MCC) is used for the same purpose. MCC is generally more informative than other metrics (Chicco et al., 2021), and most importantly, it is readily available in sklearn, the library used to develop our models.

2. Test data

GridSearch on SVM and MLP results in the following best hyperparameters combination given the set of pre-defined parameters:

- SVM: number of components for PCA to retain: 100, rbf kernel with C value of 10
- MLP: number of components for PCA to retain: 100, sgd solver with hidden layer size of 500 and maximum iterations of 1000

After training on 70% of the dataset, each model is validated by giving prediction on the test set of 30% dataset (162 samples in total, 54 samples in each category). For SVM, the overall accuracy was 83%. The model misclassified 10 effusion samples as normal and 8 samples as tube conditions, giving the lowest value of recall for this class as 0.67. The model was able to give predictions correctly for the majority of sample with normal conditions and tube conditions. The average precision, recall and F1-score are 0.83, 0.83 and 0.82, respectively. The MCC value was 0.74, which indicates a strong informed prediction of the model. The confusion matrix for SVM model is given in Figure 9.



(a) SVM
Accuracy: 83%, Precision: 0.83,
Recall: 0.83, F1-score: 0.82, MCC 0.74

(b) MLP
Accuracy: 78%, Precision: 0.78,
Recall: 0.78, F1-score: 0.78, MCC 0.67

(c) CNN
Accuracy: 78%, Precision: 0.78,
Recall: 0.78, F1-score: 0.77, MCC 0.67

Figure 9: Confusion matrices for models trained with 3-label classification problem - public dataset

The MLP model obtained slightly worse performance on the same test set. The overall accuracy was 78%. Average precision, recall and F1-score are 0.78. The MLP model struggles to classify effusion samples correctly, given by lowest precision, recall and F1-score for this category. The MCC value for this model was 0.67, closer to the random guessing mechanism in comparison with SVM model (Figure 9).

CNN model was built with 12 layers and a total of 11,572,611 trainable params. The architecture of the model is given in Figure 7. Model was trained in 50 epochs. The accuracy on the test set was 78%, the MCC value was 0.67, which is equal to the value of MLP model. The average precision, recall and F1-score was 0.78, 0.78 and 0.77, respectively. From the confusion matrix (Figure 9), the CNN model was even more confusing with effusion samples (misclassified 15 of the total 54 effusion samples as tube condition). There are also more normal samples classified as tube (2, compared with 1 of MLP and 0 of SVM). The training loss and validation loss suggest that the model was struggling to overfit the problem in the training set (Figure 10). The reason could be that there has not enough data in the training set for the model to converge properly, and the model infrastructure was too complicated for the task.

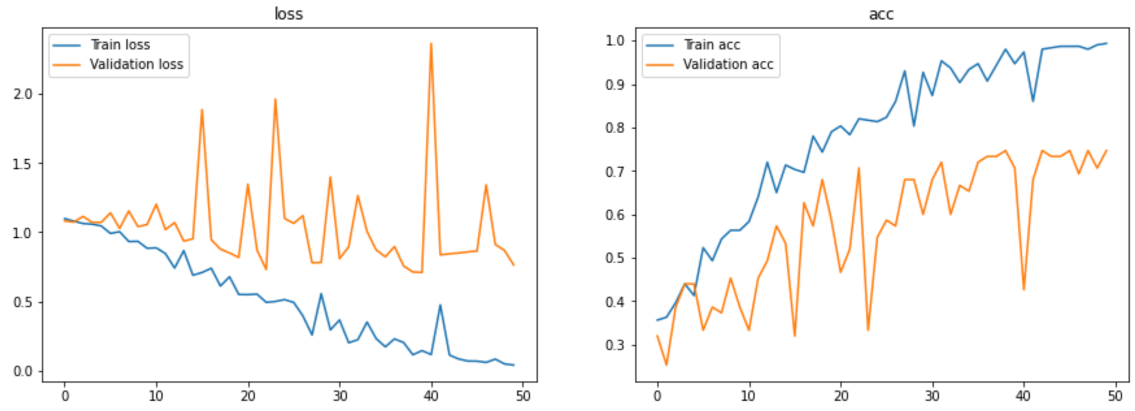
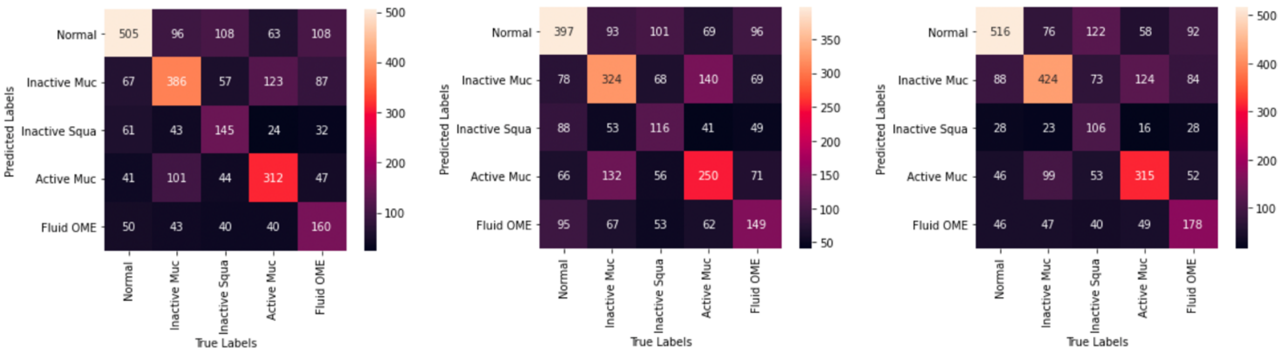


Figure 10: Training and validation loss / accuracy for CNN model on 3-label classification problem – public dataset

3. Swimming data 5 classes

After applying the auto filtering model and excluding all the problematic labels (Healed COM, Active squamous and Pus AOM), the dataset contains 9,275 samples in five categories. The same approaches and models have been built for the five labels classification problem. For SVM, the overall accuracy was 54%. The precision, recall and F1-score are 0.53, 0.51 and 0.52, respectively. The MCC value for SVM model was 0.42, indicating a fairly moderated informed of the prediction. The confusion matrix shows that the model was mostly confused between Active Mucosal and Inactive Mucosal samples (Figure 11). This could indicate that whether the model has not been capturing sufficient data to distinguish the two categories efficiently, or the information in the otoscopy image could be insufficient to recognise these cases, and additional data might be needed.



(a) SVM
Accuracy: 54%, Precision: 0.53,
Recall: 0.51, F1-score: 0.52, MCC 0.42

(b) MLP
Accuracy: 44%, Precision: 0.42,
Recall: 0.42, F1-score: 0.42, MCC 0.29

(c) CNN
Accuracy: 55%, Precision: 0.54,
Recall: 0.52, F1-score: 0.52, MCC 0.43

Figure 11: Confusion matrices for models trained with 5-label classification problem - Swimming dataset, cropped images

Similar to the classification problem on the Zenodo dataset, the MLP model performed worse than that of the SVM model. Overall accuracy was 44% for all classes. Average precision, recall and F1-score were 0.42. MCC value was 0.29, indicating that the model performed slightly better than random guessing (Figure 11). The accuracy and MCC of this model are the lowest in comparison with other models for this 5-class classification problem. The plotted training and validation accuracies suggest that while the model was becoming very good at the training set, it always struggled to pass the 55% threshold on the validation set (Figure 12). This is a clear sign of overfitting training.

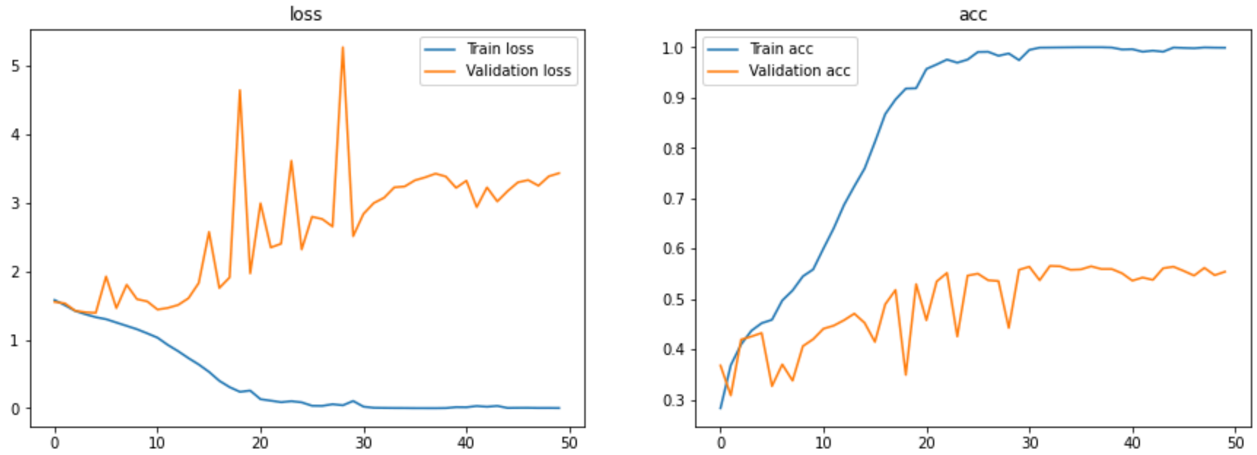
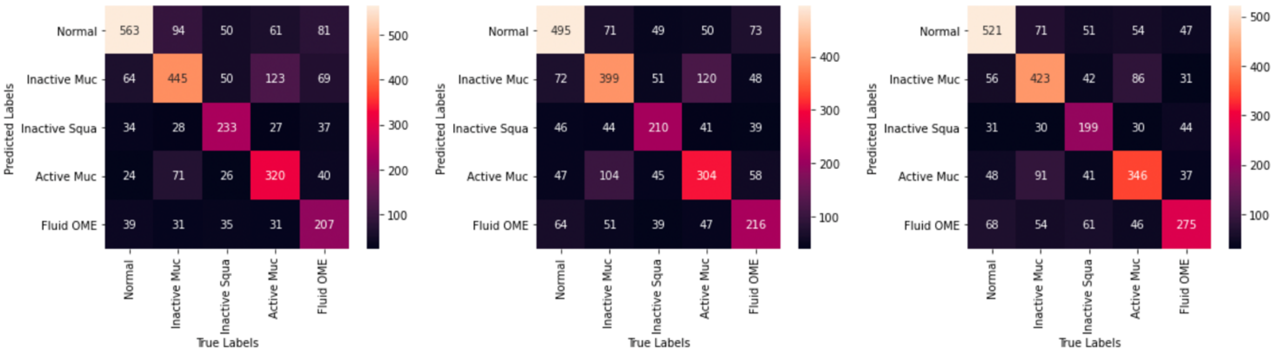


Figure 12: Training and validation loss / accuracy for CNN model on 5-label classification problem – Swimming dataset, cropped images

CNN model with the same architecture (except for the last dense layer to predict 5 classes) has an accuracy of 55%, which is comparable with SVM model. Average recall and F1-score were 0.52, while the average precision was 0.54. MCC value for CNN model was 0.43.

Those numbers are still comparable for the prediction of a community-based worker on similar research. In the diagnosis of 4 different labels using otoscopy videos, the paediatricians and general practitioners would only give an accuracy of 51% and 46%, respectively. Otolaryngologists are more accurate but still far from perfect, with an accuracy of around 74% (Pichichero and Poole, 2005).

Interestingly, when applying the same model to the input data before extracting the ROI using binary thresholding technique, all models could achieve higher performance. SVM model has the accuracy of 64% and MCC value of 0.54. MLP model has the accuracy of 58% and MCC value of 0.47. These metrics for CNN model were 63% and 0.54 (Figure 13, Figure 14). Most of the metrics for all separated labels from models with cropped data are lower than those from models with original data.



(a) SVM
Accuracy: 64%, Precision: 0.63,
Recall: 0.62, F1-score: 0.62, MCC 0.54

(b) MLP
Accuracy: 58%, Precision: 0.57,
Recall: 0.57, F1-score: 0.57, MCC 0.47

(c) CNN
Accuracy: 63%, Precision: 0.62,
Recall: 0.62, F1-score: 0.62, MCC 0.54

Figure 13: Confusion matrices for models trained with 5-label classification problem - Swimming dataset, original images

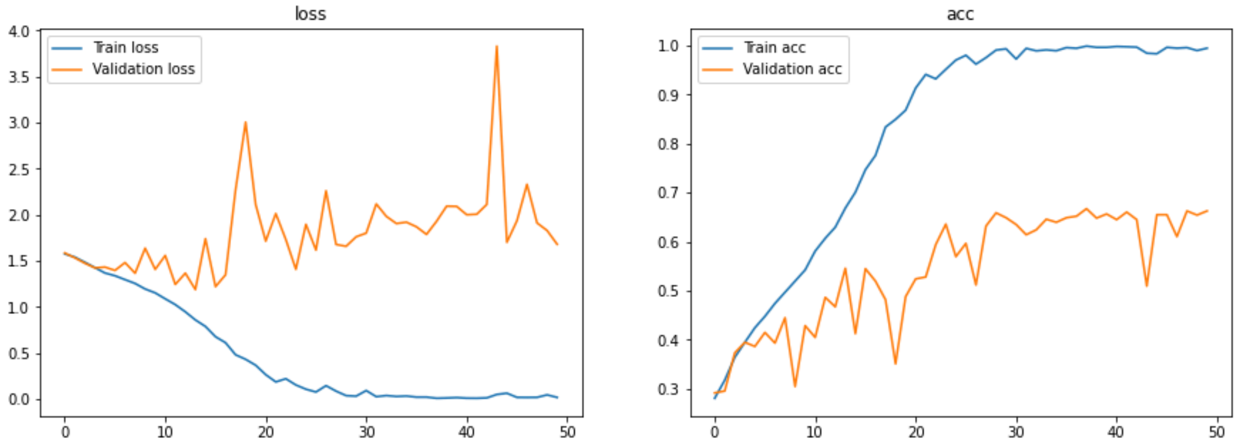
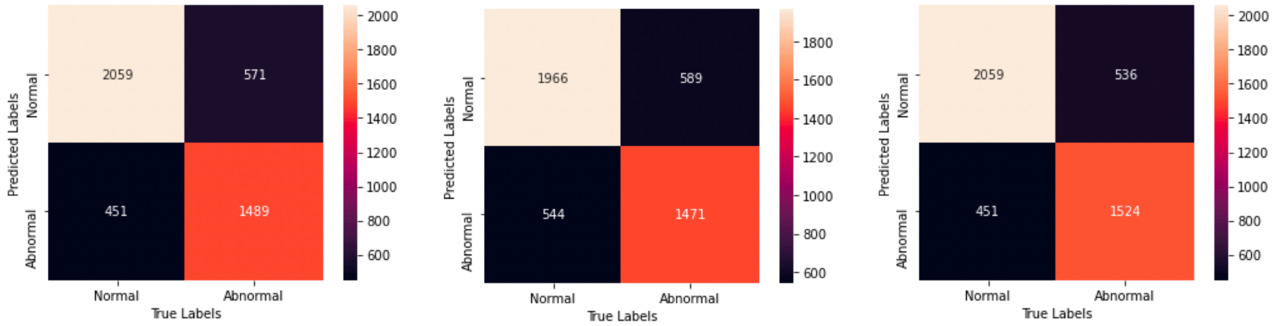


Figure 14: Training and validation loss / accuracy for CNN model on 5-label classification problem – Swimming dataset, original images

4. Swimming data 2 classes

Collapsing the label from eight to two (normal and abnormal) makes all filtered images become available for training. The size of the dataset is 15,231 samples. This time, the original images (un-cropped) version has been used. SVM model gives an overall accuracy on binary classification of 78%. The average precision is 0.78, while both average recall and F1-score are 0.77. The MCC value of SVM is 0.55, which indicates a pretty robust prediction. The model still makes errors by classifying 451 normal samples as abnormal and 571 abnormal samples as normal. F1 scores of normal and abnormal classes are 0.80 and 0.74, respectively (Figure 15).



(a) SVM
Accuracy: 78%, Precision: 0.78,
Recall: 0.77, F1-score: 0.77, MCC 0.55

(b) MLP
Accuracy: 75%, Precision: 0.75,
Recall: 0.75, F1-score: 0.75, MCC 0.50

(c) CNN
Accuracy: 78%, Precision: 0.78,
Recall: 0.78, F1-score: 0.78, MCC 0.56

Figure 15: Confusion matrices for models trained with 2-label classification problem - Swimming dataset, original images

MLP model achieves slightly lower performance, with an overall accuracy of 75%. 0.75 was also the average precision, recall and F1-score. MCC value for MLP model is 0.50 (Figure 15). CNN model achieves accuracy of 78%, which equals to the performance of SVM model. However, other metrics are slightly higher, with 0.78 of precision, recall and F1-score. MCC value for CNN model is 0.56. The plotted loss and accuracy of the training and validation processes of CNN model still show the signs of overfitting on the dataset (Figure 16). From epoch 10, the training loss continues to decrease while the validation loss starts to increase. The accuracy in validation also does not increase as much as in the training from epoch 10. These are signs of overfitting during the training. The spikes in the validation loss could be a result of noisy samples, which are hard to learn from the dataset. As the data is derived from raw video, many of them could contain irrelevant features regarding the target class. Therefore, the model could not converge properly.

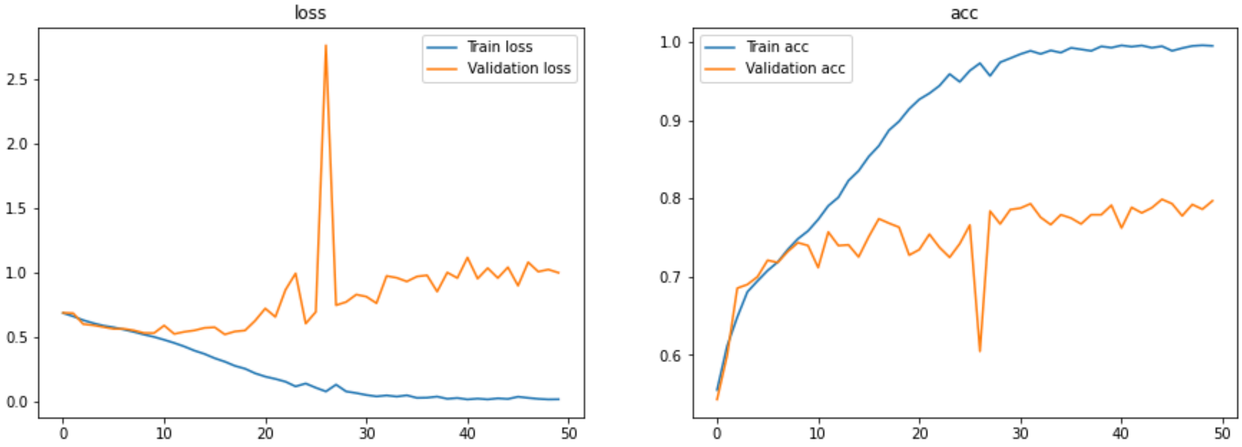
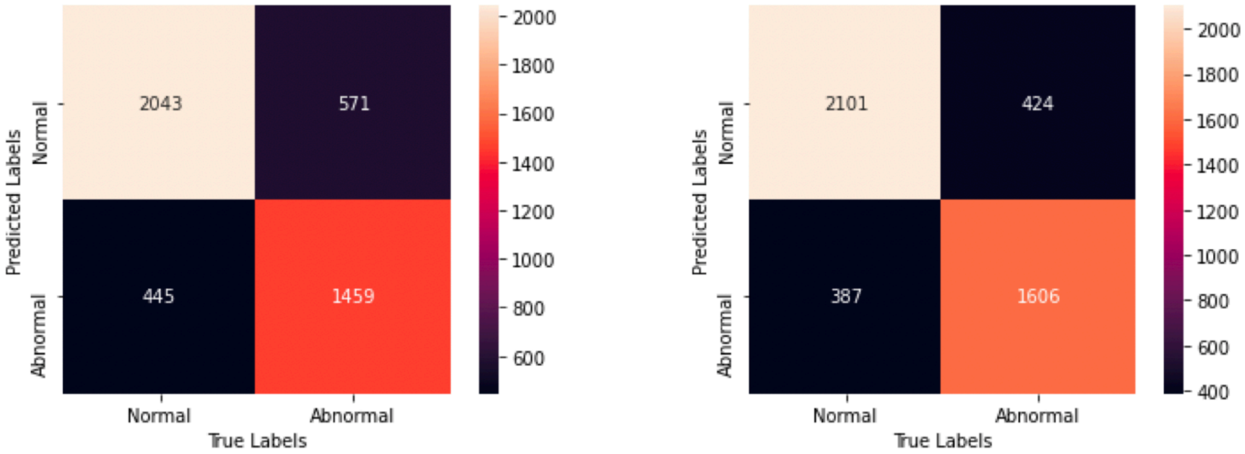


Figure 16: Training and validation loss / accuracy for CNN model on 2-label classification problem – Swimming dataset, original images

In order to boost the performance of the model, tympanometry data has been used. The idea was to combine the output of a model with the tympanometry data and train another model to give final prediction on the same training set. The performance of the intermediate and final model was then validated on the testing set. SVM model has been used to run the training again on 70% of the dataset. The model performance on the test set was similar to the one that trained in the previous binary classification problem, with overall accuracy of 78%. In order to boost the accuracy of single SVM model, another model has been introduced and trained on the combination of SVM prediction and the tympanometry data. In the first try, SVM model predicts an integer number (0 for normal, 1 for abnormal). The combined data (with only two features as the input, one for SVM prediction, other for tympanometry data) is then fed into two different models: a simple MLP model with 10 neurons in a single layer and a decision tree. The subsequent model then gives prediction on the test set. The results were that all the performance metrics had been the same compared with the prediction given by single SVM model alone. Adding another model did not help as the input data is too simple. In a second try, SVM model predicts a probability number for the two classes. The probability indicates how definite the model in predicting the sample belongs to a given category; thus, it contains more information than just binary labels of normal and abnormal. The probability number for the Normal class has been extracted and combined with the tympanometry data before feeding to final model again. Decision tree still fails to capture a useful decision procedure as it was confused by the number of probabilities given in the first SVM. However, the simple MLP, trained with only 50 iterations, did help to boost the accuracy of the pipeline from 78% with single SVM to 82%. The average precision, recall and F1-score was also 0.82, in comparison with 0.77 of SVM model alone.

The MCC value for the combined model was 0.63 (the value for SVM alone is 0.55), which is quite robust for this problem (Figure 17). This result signifies that additional types of data could help to increase the performance of a machine learning based otitis media. This experiment utilised a very simple mechanism of stacking multiple models in a single pipeline. However, more advanced ensemble techniques could be applied to get better performance.



(a) SVM
otoscopy image
Accuracy: 78%, Precision: 0.77,
Recall: 0.77, F1-score: 0.77, MCC 0.54

(b) SVM + MLP
otoscopy image + tympanometry
Accuracy: 82%, Precision: 0.82,
Recall: 0.82, F1-score: 0.82, MCC 0.64

Figure 17: Confusion matrices for pipelines with single SVM model and SVM + MLP models

All the performance metrics for different models are given in Table 6.

Table 6: All model's performance metrics

Data	# Labels	Model	Overall Accuracy	Average Precision	Average Recall	Average F1-Score	MCC value
Zenodo dataset	**3**	**SVM**	**83%**	**0.83**	**0.83**	**0.82**	**0.74**
Zenodo dataset	**3**	**MLP**	**78%**	**0.78**	**0.78**	**0.78**	**0.67**
Zenodo dataset	**3**	**CNN**	**78%**	**0.78**	**0.78**	**0.77**	**0.67**
Swimming dataset - original images	**5**	**SVM**	**64%**	**0.63**	**0.62**	**0.62**	**0.54**
Swimming dataset - original images	**5**	**MLP**	**58%**	**0.57**	**0.57**	**0.57**	**0.47**
Swimming dataset - original images	**5**	**CNN**	**63%**	**0.62**	**0.62**	**0.62**	**0.54**
Swimming dataset - cropped images	**5**	**SVM**	**54%**	**0.53**	**0.51**	**0.52**	**0.42**
Swimming dataset - cropped images	**5**	**MLP**	**44%**	**0.42**	**0.42**	**0.42**	**0.29**
Swimming dataset - cropped images	**5**	**CNN**	**55%**	**0.54**	**0.52**	**0.52**	**0.43**
Swimming dataset - original images	**2**	**SVM**	**78%**	**0.78**	**0.77**	**0.77**	**0.55**
Swimming dataset - original images	**2**	**MLP**	**75%**	**0.75**	**0.75**	**0.75**	**0.5**
Swimming dataset - original images	**2**	**CNN**	**78%**	**0.78**	**0.78**	**0.78**	**0.56**
Swimming dataset - cropped images	**2**	**SVM**	**72%**	**0.72**	**0.72**	**0.72**	**0.44**
Swimming dataset - cropped images	**2**	**MLP**	**67%**	**0.67**	**0.67**	**0.67**	**0.34**
Swimming dataset - cropped images	**2**	**CNN**	**73%**	**0.73**	**0.72**	**0.73**	**0.46**
SVM alone - otoscopy images	**2**	**SVM**	**78%**	**0.77**	**0.77**	**0.77**	**0.54**
SVM with MLP – otoscopy and tympanometry	**2**	**SVM + MLP**	**82%**	**0.82**	**0.82**	**0.82**	**0.64**

DISCUSSION

With three labels and a small set of data, the proposed models could achieve the accuracy of 83% (SVM) and 78% (MLP and CNN). The results were in the lower band in comparison with other models proposed in the literature for otitis media diagnosis. However, the results are still higher than those given by paediatricians and general practitioners in a similar problem, which are 51% and 46%, respectively (Pichichero and Poole, 2005). Even the worse performing models in five-label classification problem on cropped data are still comparable with the number given in the report. However, in order to build a usable otitis media system that could be useful in supporting the diagnosis, the proposed model needs to be enhanced to reach the high-level accuracy given by the expert. Problems and results that arise from the training could help to reveal possible enhancement paths.

First of all, from the training and validation accuracy and loss graph, we could see that overfitting is a common issue throughout all the training. Although GridSearch has been used to tune the hyperparameters for SVM and MLP model, there is only a limited number of parameter combinations has been tried. Further research on the algorithm and extending the set of possible hyperparameters could help to find a better solution. For CNN model, a fairly simple architecture has been used throughout all the problems. A better architecture could be suitable for this task. One could also try to apply transfer learning to leverage the power of the existing available model. Another approach could be using ensemble methods to merge multiple models for better predictive performance.

Second, the performances of all the models training on cropped images are worse than those trained on original images. One possible explanation is that the image processing technique applied to extract ROI did not really help to boost the feature extraction in the CNN model as well as the PCA procedure of SVM and MLP. It could also amplify the noise in the image as well as the intensity level of pixels around the circle edges (many images has been affected by this lighting condition). Note that no additional image processing efforts have been made (denoise, sharpen or histogram equalisation), so suitable image processing techniques could be applied in future works on the same dataset to increase the overall performance of the models. Another important aspect of the dataset is that they are filtered by the auto-filter model, which has been built based on a manual image selection of a non-ENT expert with no prior training. Although the writer has tried the best to understand the dataset, the selection could be problematic, especially for different ear conditions. This might directly affect the performance of the target model, in comparison with the other models in the literature, which has been built mostly by still-image of the ears that are well-selected and processed by an otolaryngologist. The results also suggest that more works are needed in the selection of the extracted images from video otoscopy. Otherwise, the model could be more accurate if it had been built using a dataset of still images captured from otoscope directly.

Finally, the report proposed an approach to combine multiple data sources to improve the diagnosis of otitis media. The system contains two models, SVM and MLP. SVM model gives predictions based on otoscopy images. The resulting probability is combined with tympanometry data and fed to MLP model. Although the setup is simple, the accuracy of the overall prediction has been increased from 78% to 82%. To the best of my knowledge, this is the first report that incorporates multiple diagnosis data in an otitis media diagnosis system based on machine learning. This approach is also similar to real world procedure, where the clinicians need to perform more than one diagnosis to identify the patient. It is worth noting that the problem has been simplified by scaling the input to a binary classification. In the future, more advanced techniques, like feature fusion and decision fusion (Boulahia et al., 2021), should be applied to build a more robust system that could utilise multiple data sources and effectively combine the results of different models to give better predictions on different types of otitis media.

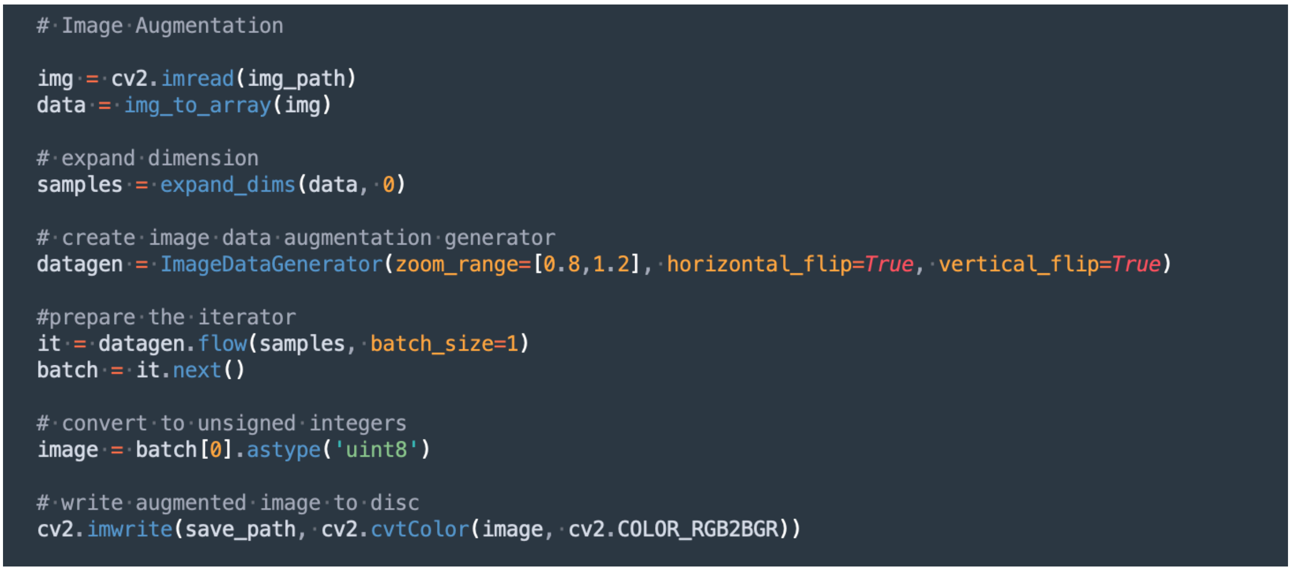
CONCLUSION

The report proposes three different machine learning models: SVM, MLP and CNN to solve various classification problems on otitis media based on otoscopy images and tympanometry data. The highest accuracy achieved on 3-label problem on a public dataset of 454 images is 83%. The highest accuracies for a 5-label classification problem and 2-label classification problems on Swimming Pool data are 64% and 78%, respectively. The report also confirms the possibility of combining multiple data types to boost the performance of an otitis media diagnosis system based on machine learning. The overall accuracy has been increased from 78% to 82% by combining tympanometry data in a later model with the prediction result from the previous model. In the future, multiple enhancements could be made to the system in image processing, hyperparameters tuning, CNN architecture exploration and other fusion approaches. A more accurate system with an effective deployment model could expose the technology capability in supporting the otitis media diagnosing procedure, especially in remote areas.

APPENDICES

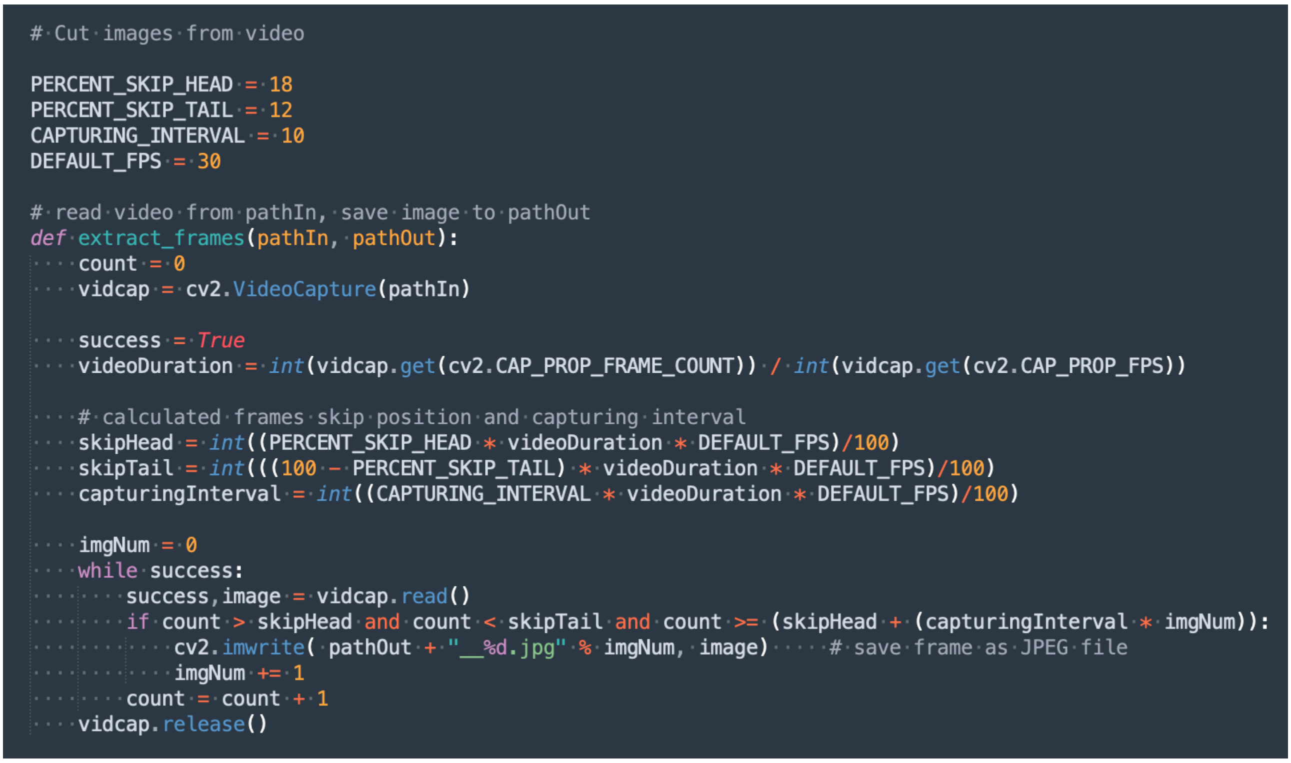
Following are a list of code fragments that have been used throughout the experiments in this report.

Image augmentation performed on Zenodo dataset:



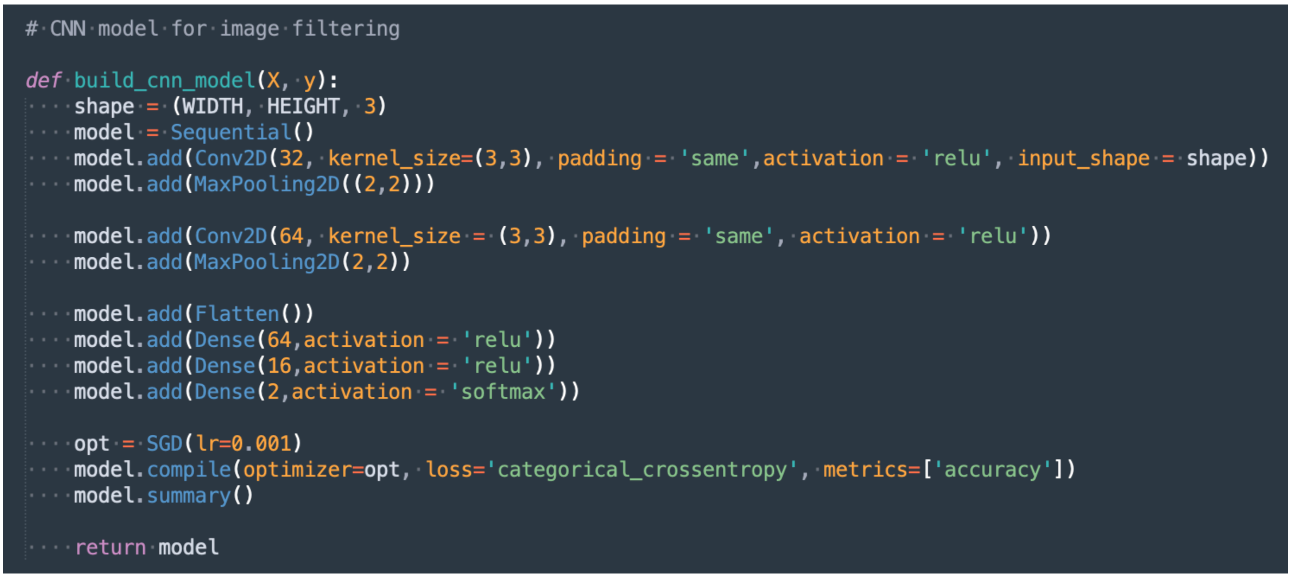
```python
# Image Augmentation

img = cv2.imread(img_path)
data = img_to_array(img)

# expand dimension
samples = expand_dims(data, 0)

# create image data augmentation generator
datagen = ImageDataGenerator(zoom_range=[0.8,1.2], horizontal_flip=True, vertical_flip=True)

#prepare the iterator
it = datagen.flow(samples, batch_size=1)
batch = it.next()

# convert to unsigned integers
image = batch[0].astype('uint8')

# write augmented image to disc
cv2.imwrite(save_path, cv2.cvtColor(image, cv2.COLOR_RGB2BGR))
```

Cut Image from video:



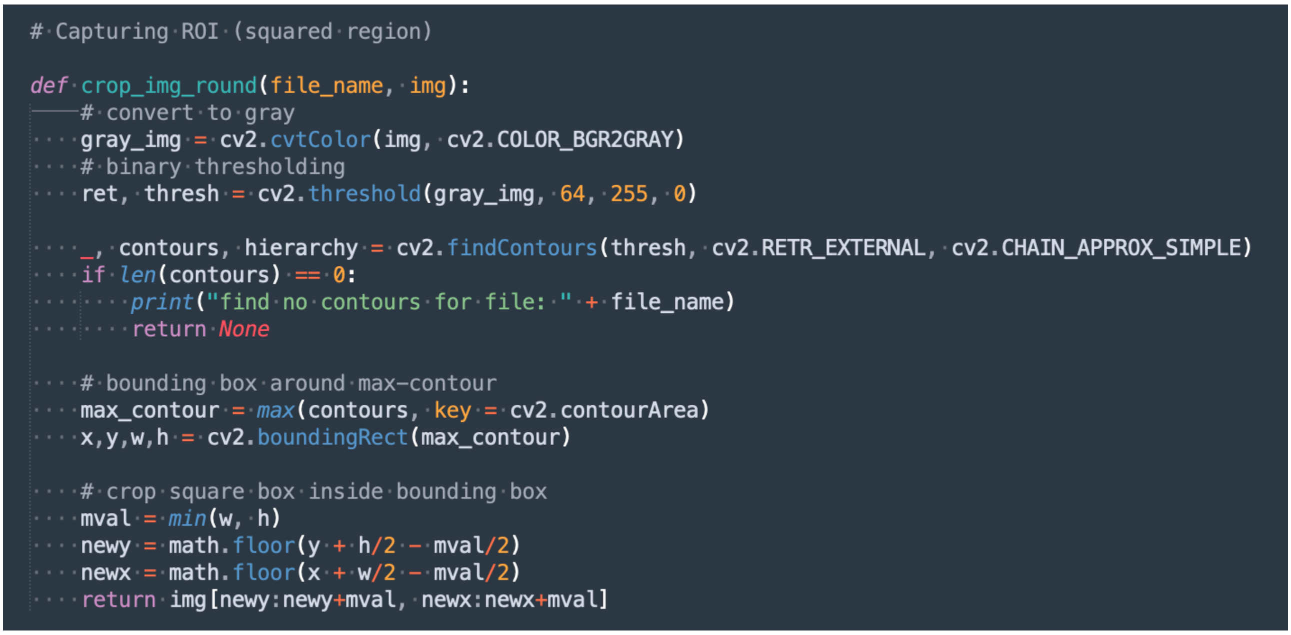
```python
# Cut images from video

PERCENT_SKIP_HEAD = 18
PERCENT_SKIP_TAIL = 12
CAPTURING_INTERVAL = 10
DEFAULT_FPS = 30

# read video from pathIn, save image to pathOut
def extract_frames(pathIn, pathOut):
    count = 0
    vidcap = cv2.VideoCapture(pathIn)

    success = True
    videoDuration = int(vidcap.get(cv2.CAP_PROP_FRAME_COUNT)) / int(vidcap.get(cv2.CAP_PROP_FPS))

    # calculated frames skip position and capturing interval
    skipHead = int((PERCENT_SKIP_HEAD * videoDuration * DEFAULT_FPS)/100)
    skipTail = int(((100 - PERCENT_SKIP_TAIL) * videoDuration * DEFAULT_FPS)/100)
    capturingInterval = int((CAPTURING_INTERVAL * videoDuration * DEFAULT_FPS)/100)

    imgNum = 0
    while success:
        success,image = vidcap.read()
        if count > skipHead and count < skipTail and count >= (skipHead + (capturingInterval * imgNum)):
            cv2.imwrite( pathOut + "__%d.jpg" % imgNum, image)     # save frame as JPEG file
            imgNum += 1
        count = count + 1
    vidcap.release()
```

CNN model for filtering image:



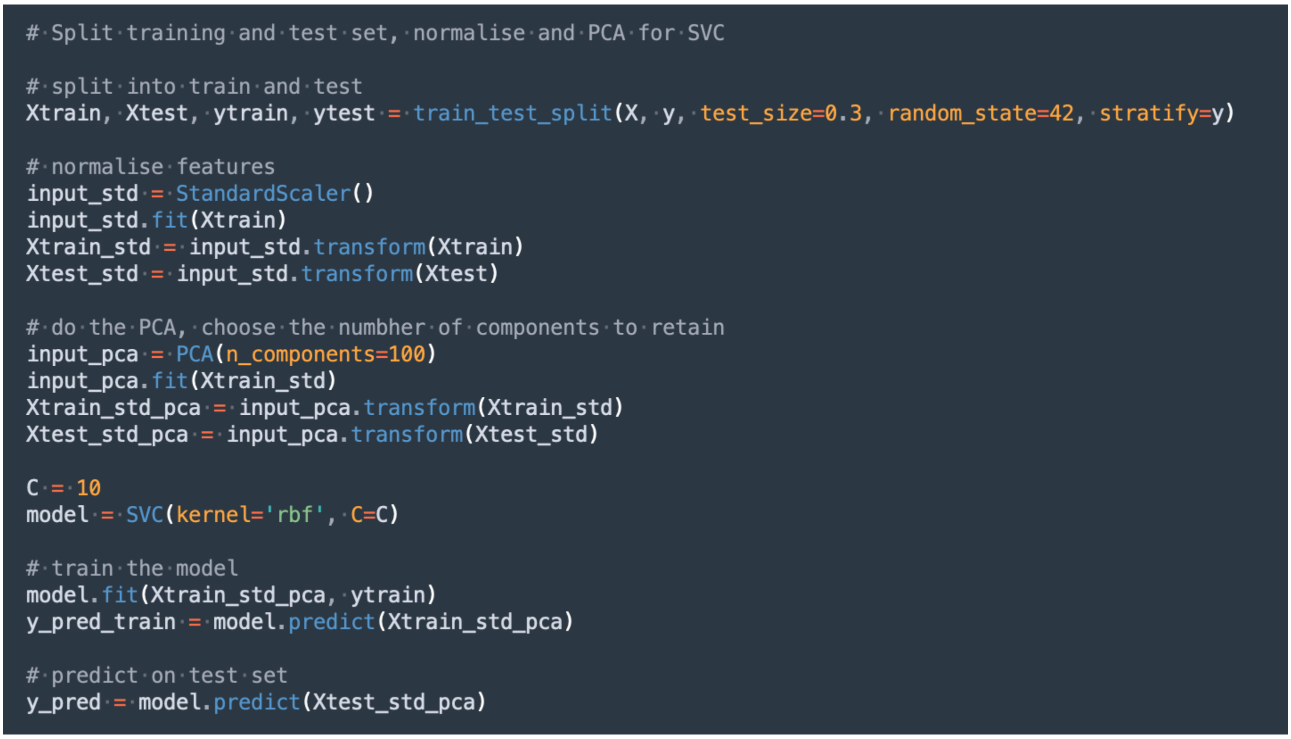
```
# CNN model for image filtering

def build_cnn_model(X, y):
    shape = (WIDTH, HEIGHT, 3)
    model = Sequential()
    model.add(Conv2D(32, kernel_size=(3,3), padding = 'same',activation = 'relu', input_shape = shape))
    model.add(MaxPooling2D((2,2)))

    model.add(Conv2D(64, kernel_size = (3,3), padding = 'same', activation = 'relu'))
    model.add(MaxPooling2D(2,2))

    model.add(Flatten())
    model.add(Dense(64,activation = 'relu'))
    model.add(Dense(16,activation = 'relu'))
    model.add(Dense(2,activation = 'softmax'))

    opt = SGD(lr=0.001)
    model.compile(optimizer=opt, loss='categorical_crossentropy', metrics=['accuracy'])
    model.summary()

    return model
```

Capturing ROI from image:



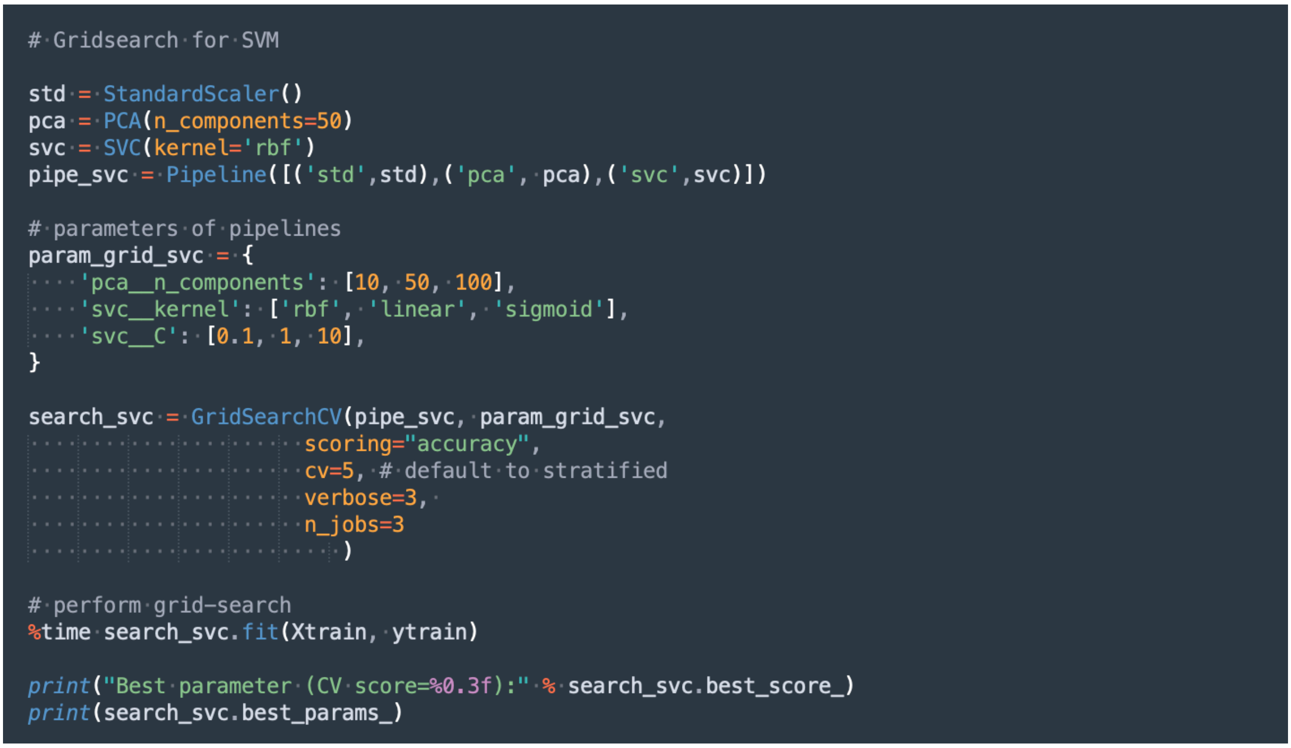
```
# Capturing ROI (squared region)

def crop_img_round(file_name, img):
    # convert to gray
    gray_img = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)
    # binary thresholding
    ret, thresh = cv2.threshold(gray_img, 64, 255, 0)

    _, contours, hierarchy = cv2.findContours(thresh, cv2.RETR_EXTERNAL, cv2.CHAIN_APPROX_SIMPLE)
    if len(contours) == 0:
        print("find no contours for file: " + file_name)
        return None

    # bounding box around max-contour
    max_contour = max(contours, key = cv2.contourArea)
    x,y,w,h = cv2.boundingRect(max_contour)

    # crop square box inside bounding box
    mval = min(w, h)
    newy = math.floor(y + h/2 - mval/2)
    newx = math.floor(x + w/2 - mval/2)
    return img[newy:newy+mval, newx:newx+mval]
```

SVM model training and validation:



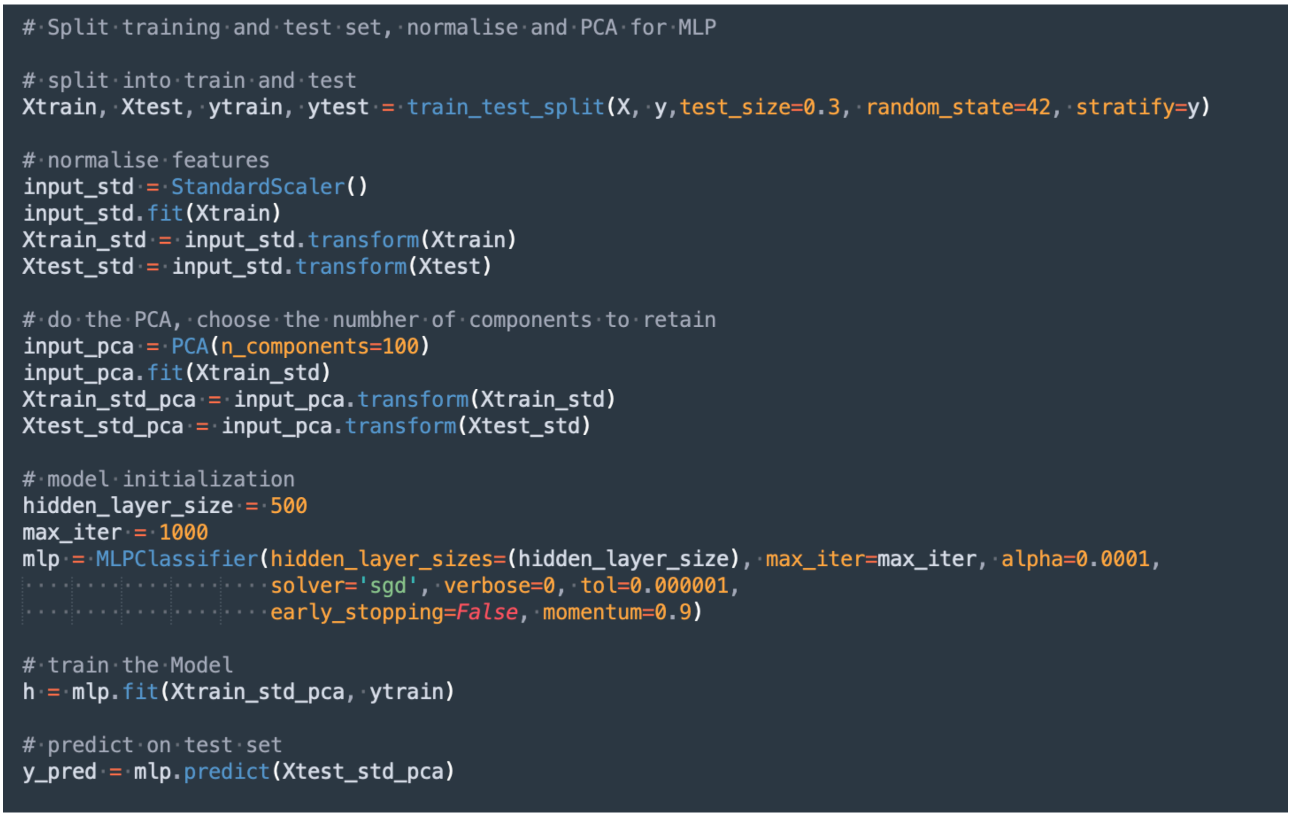
```
# Split training and test set, normalise and PCA for SVC

# split into train and test
Xtrain, Xtest, ytrain, ytest = train_test_split(X, y, test_size=0.3, random_state=42, stratify=y)

# normalise features
input_std = StandardScaler()
input_std.fit(Xtrain)
Xtrain_std = input_std.transform(Xtrain)
Xtest_std = input_std.transform(Xtest)

# do the PCA, choose the numbher of components to retain
input_pca = PCA(n_components=100)
input_pca.fit(Xtrain_std)
Xtrain_std_pca = input_pca.transform(Xtrain_std)
Xtest_std_pca = input_pca.transform(Xtest_std)

C = 10
model = SVC(kernel='rbf', C=C)

# train the model
model.fit(Xtrain_std_pca, ytrain)
y_pred_train = model.predict(Xtrain_std_pca)

# predict on test set
y_pred = model.predict(Xtest_std_pca)
```

GridSearch for SVM model hyperparameters tuning:



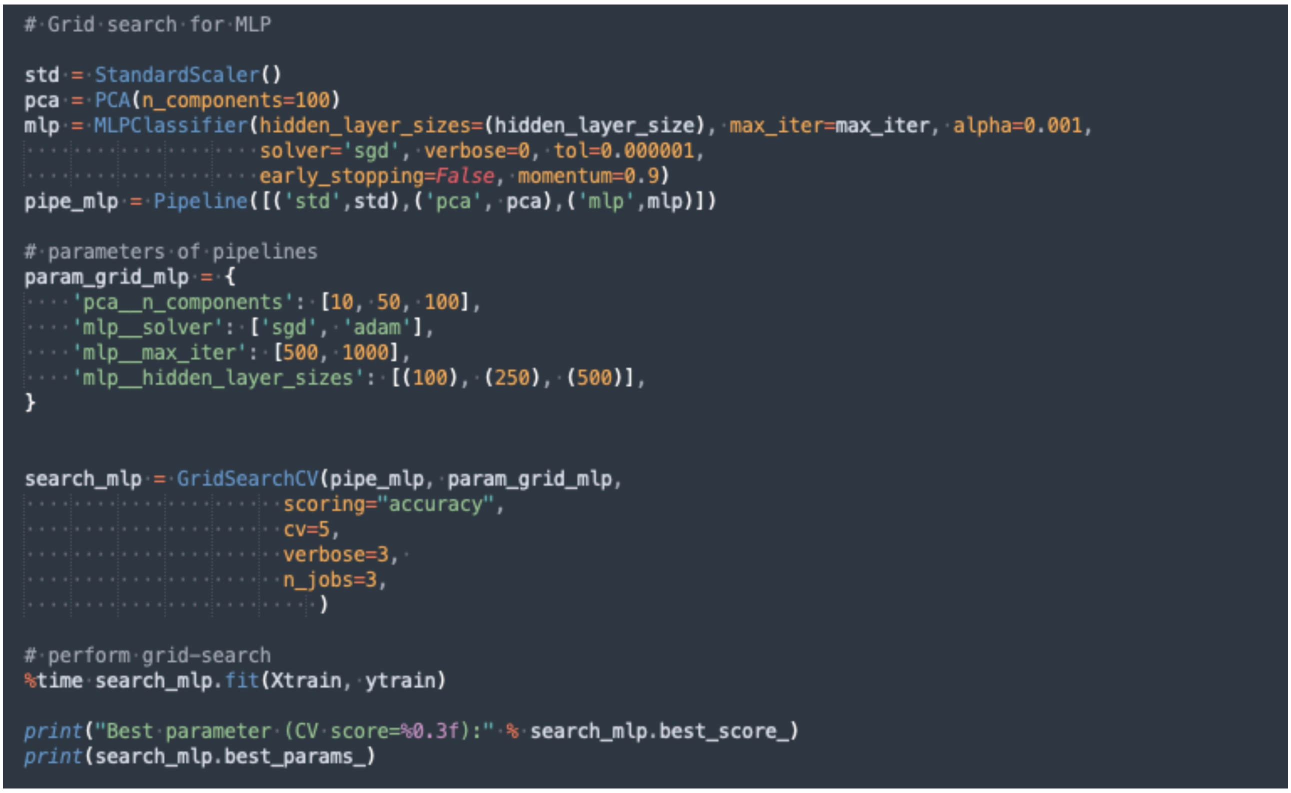
```
# Gridsearch for SVM

std = StandardScaler()
pca = PCA(n_components=50)
svc = SVC(kernel='rbf')
pipe_svc = Pipeline([('std',std),('pca', pca),('svc',svc)])

# parameters of pipelines
param_grid_svc = {
    'pca__n_components': [10, 50, 100],
    'svc__kernel': ['rbf', 'linear', 'sigmoid'],
    'svc__C': [0.1, 1, 10],
}

search_svc = GridSearchCV(pipe_svc, param_grid_svc,
                      scoring="accuracy",
                      cv=5, # default to stratified
                      verbose=3,
                      n_jobs=3
                         )

# perform grid-search
%time search_svc.fit(Xtrain, ytrain)

print("Best parameter (CV score=%0.3f):" % search_svc.best_score_)
print(search_svc.best_params_)
```

MLP model training and validation:



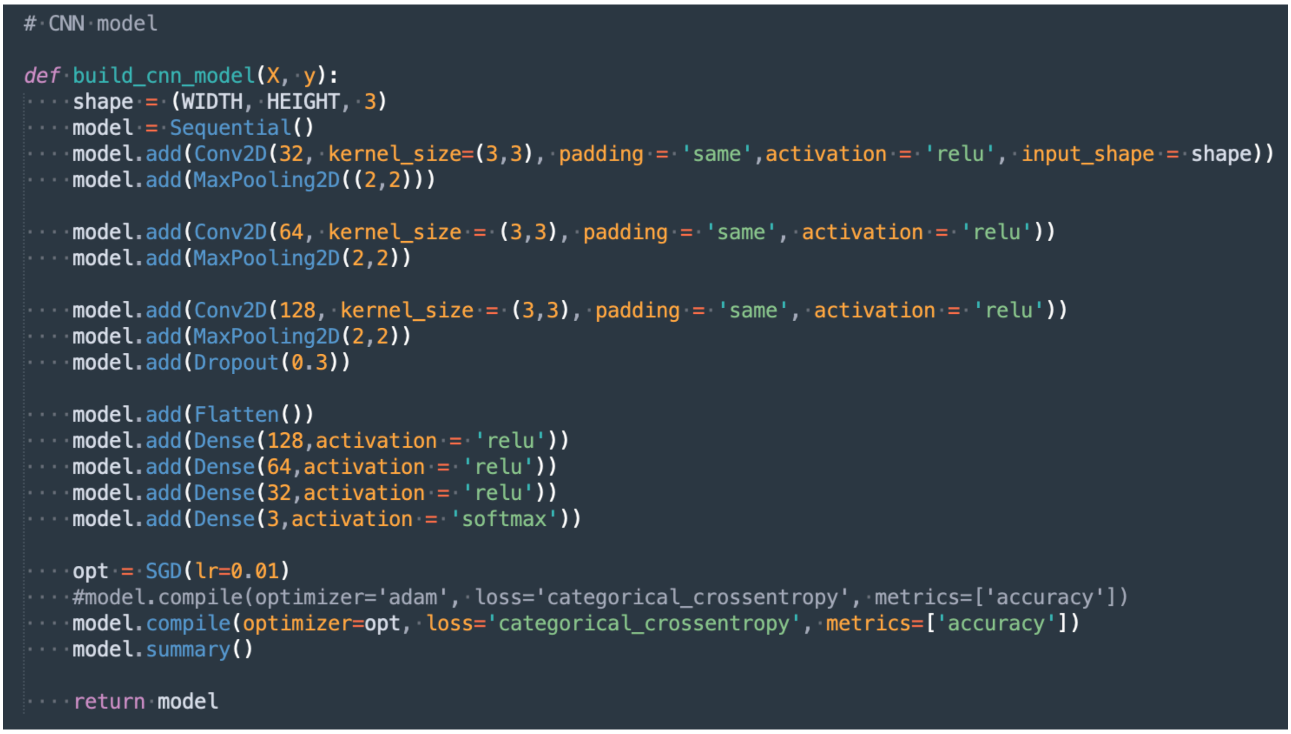
```python
# Split training and test set, normalise and PCA for MLP

# split into train and test
Xtrain, Xtest, ytrain, ytest = train_test_split(X, y,test_size=0.3, random_state=42, stratify=y)

# normalise features
input_std = StandardScaler()
input_std.fit(Xtrain)
Xtrain_std = input_std.transform(Xtrain)
Xtest_std = input_std.transform(Xtest)

# do the PCA, choose the numbher of components to retain
input_pca = PCA(n_components=100)
input_pca.fit(Xtrain_std)
Xtrain_std_pca = input_pca.transform(Xtrain_std)
Xtest_std_pca = input_pca.transform(Xtest_std)

# model initialization
hidden_layer_size = 500
max_iter = 1000
mlp = MLPClassifier(hidden_layer_sizes=(hidden_layer_size), max_iter=max_iter, alpha=0.0001,
                    solver='sgd', verbose=0, tol=0.000001,
                    early_stopping=False, momentum=0.9)

# train the Model
h = mlp.fit(Xtrain_std_pca, ytrain)

# predict on test set
y_pred = mlp.predict(Xtest_std_pca)
```

GridSearch for MLP model hyperparameters tuning:



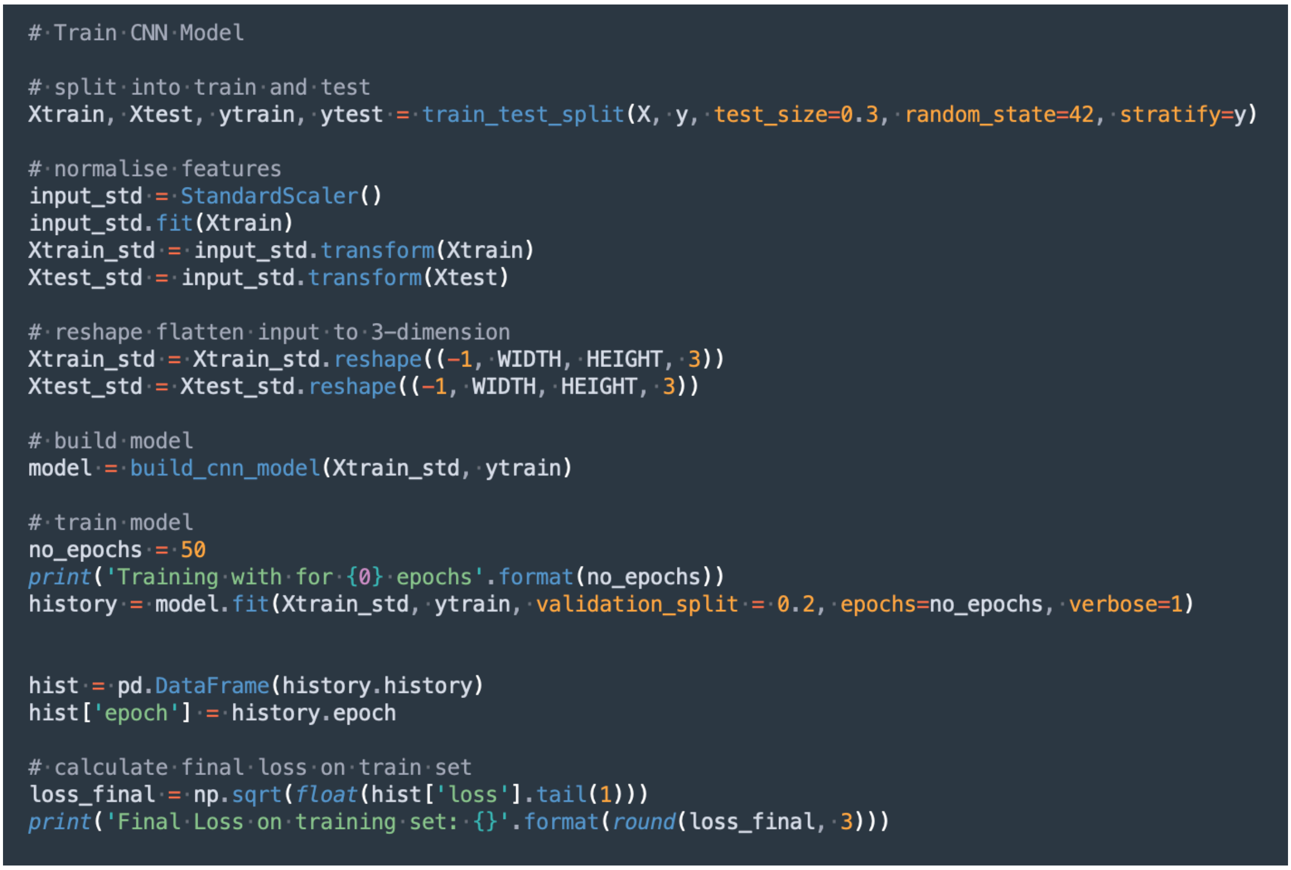
```python
# Grid search for MLP

std = StandardScaler()
pca = PCA(n_components=100)
mlp = MLPClassifier(hidden_layer_sizes=(hidden_layer_size), max_iter=max_iter, alpha=0.001,
                    solver='sgd', verbose=0, tol=0.000001,
                    early_stopping=False, momentum=0.9)
pipe_mlp = Pipeline([('std',std),('pca', pca),('mlp',mlp)])

# parameters of pipelines
param_grid_mlp = {
    'pca__n_components': [10, 50, 100],
    'mlp__solver': ['sgd', 'adam'],
    'mlp__max_iter': [500, 1000],
    'mlp__hidden_layer_sizes': [(100), (250), (500)],
}


search_mlp = GridSearchCV(pipe_mlp, param_grid_mlp,
                     scoring="accuracy",
                     cv=5,
                     verbose=3,
                     n_jobs=3,
                        )

# perform grid-search
%time search_mlp.fit(Xtrain, ytrain)

print("Best parameter (CV score=%0.3f):" % search_mlp.best_score_)
print(search_mlp.best_params_)
```

Constructing CNN model:



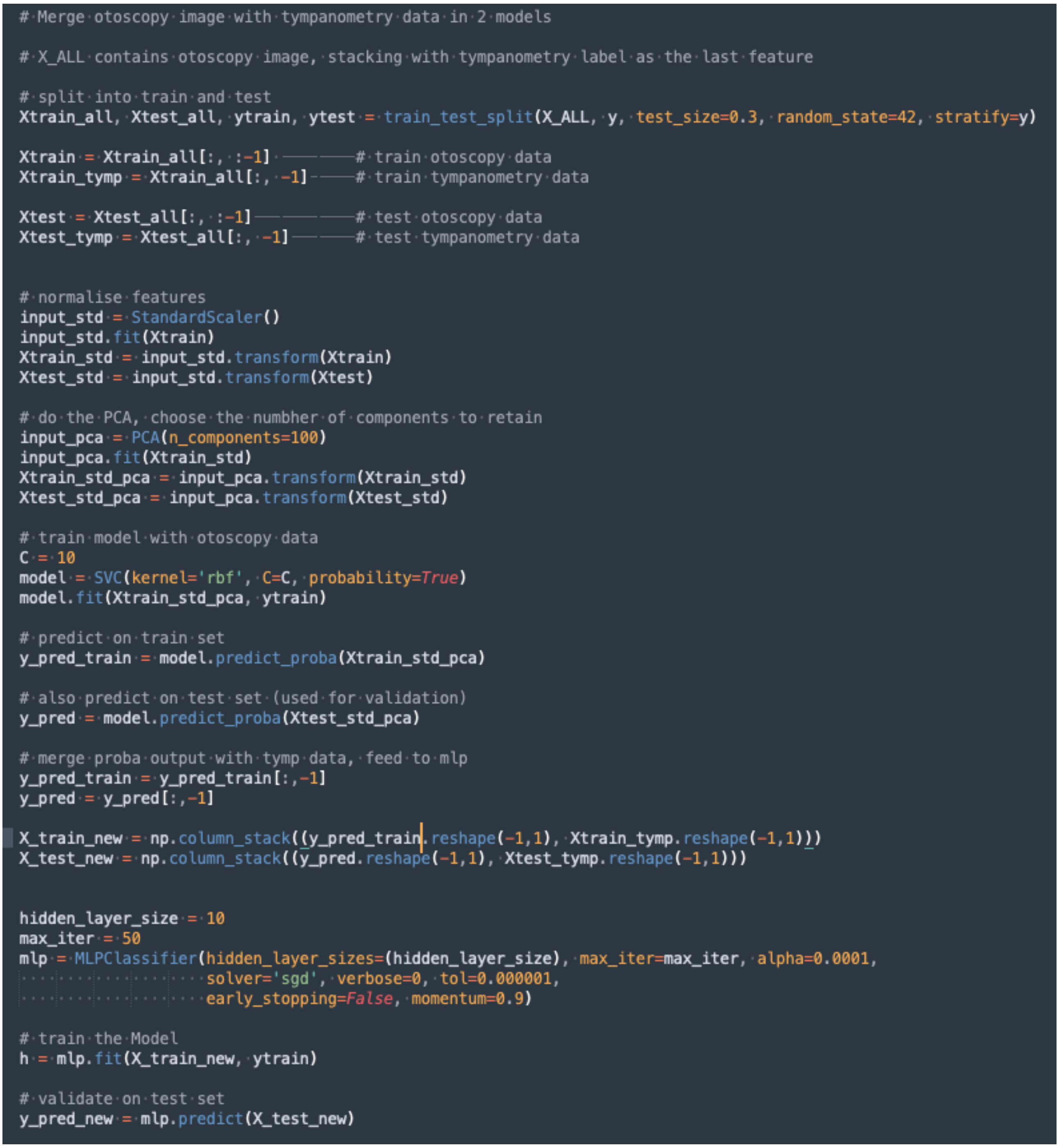
```
# CNN model

def build_cnn_model(X, y):
    shape = (WIDTH, HEIGHT, 3)
    model = Sequential()
    model.add(Conv2D(32, kernel_size=(3,3), padding = 'same',activation = 'relu', input_shape = shape))
    model.add(MaxPooling2D((2,2)))

    model.add(Conv2D(64, kernel_size = (3,3), padding = 'same', activation = 'relu'))
    model.add(MaxPooling2D(2,2))

    model.add(Conv2D(128, kernel_size = (3,3), padding = 'same', activation = 'relu'))
    model.add(MaxPooling2D(2,2))
    model.add(Dropout(0.3))

    model.add(Flatten())
    model.add(Dense(128,activation = 'relu'))
    model.add(Dense(64,activation = 'relu'))
    model.add(Dense(32,activation = 'relu'))
    model.add(Dense(3,activation = 'softmax'))

    opt = SGD(lr=0.01)
    #model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])
    model.compile(optimizer=opt, loss='categorical_crossentropy', metrics=['accuracy'])
    model.summary()

    return model
```

CNN model training and validation:



```
# Train CNN Model

# split into train and test
Xtrain, Xtest, ytrain, ytest = train_test_split(X, y, test_size=0.3, random_state=42, stratify=y)

# normalise features
input_std = StandardScaler()
input_std.fit(Xtrain)
Xtrain_std = input_std.transform(Xtrain)
Xtest_std = input_std.transform(Xtest)

# reshape flatten input to 3-dimension
Xtrain_std = Xtrain_std.reshape((-1, WIDTH, HEIGHT, 3))
Xtest_std = Xtest_std.reshape((-1, WIDTH, HEIGHT, 3))

# build model
model = build_cnn_model(Xtrain_std, ytrain)

# train model
no_epochs = 50
print('Training with for {0} epochs'.format(no_epochs))
history = model.fit(Xtrain_std, ytrain, validation_split = 0.2, epochs=no_epochs, verbose=1)


hist = pd.DataFrame(history.history)
hist['epoch'] = history.epoch

# calculate final loss on train set
loss_final = np.sqrt(float(hist['loss'].tail(1)))
print('Final Loss on training set: {}'.format(round(loss_final, 3)))
```

Merging otoscopy image with tympanometry data in two models:



```
# Merge otoscopy image with tympanometry data in 2 models

# X_ALL contains otoscopy image, stacking with tympanometry label as the last feature

# split into train and test
Xtrain_all, Xtest_all, ytrain, ytest = train_test_split(X_ALL, y, test_size=0.3, random_state=42, stratify=y)

Xtrain = Xtrain_all[:, :-1]         # train otoscopy data
Xtrain_tymp = Xtrain_all[:, -1]     # train tympanometry data

Xtest = Xtest_all[:, :-1]           # test otoscopy data
Xtest_tymp = Xtest_all[:, -1]       # test tympanometry data


# normalise features
input_std = StandardScaler()
input_std.fit(Xtrain)
Xtrain_std = input_std.transform(Xtrain)
Xtest_std = input_std.transform(Xtest)

# do the PCA, choose the numbher of components to retain
input_pca = PCA(n_components=100)
input_pca.fit(Xtrain_std)
Xtrain_std_pca = input_pca.transform(Xtrain_std)
Xtest_std_pca = input_pca.transform(Xtest_std)

# train model with otoscopy data
C = 10
model = SVC(kernel='rbf', C=C, probability=True)
model.fit(Xtrain_std_pca, ytrain)

# predict on train set
y_pred_train = model.predict_proba(Xtrain_std_pca)

# also predict on test set (used for validation)
y_pred = model.predict_proba(Xtest_std_pca)

# merge proba output with tymp data, feed to mlp
y_pred_train = y_pred_train[:,-1]
y_pred = y_pred[:,-1]

X_train_new = np.column_stack((y_pred_train.reshape(-1,1), Xtrain_tymp.reshape(-1,1)))
X_test_new = np.column_stack((y_pred.reshape(-1,1), Xtest_tymp.reshape(-1,1)))


hidden_layer_size = 10
max_iter = 50
mlp = MLPClassifier(hidden_layer_sizes=(hidden_layer_size), max_iter=max_iter, alpha=0.0001,
                    solver='sgd', verbose=0, tol=0.000001,
                    early_stopping=False, momentum=0.9)

# train the Model
h = mlp.fit(X_train_new, ytrain)

# validate on test set
y_pred_new = mlp.predict(X_test_new)
```

Complete source code used in the report could be found in the following repository:

https://github.com/phunp/otitis-media

BIBLIOGRAPHY

ALHUDHAIF, A., CÖMERT, Z. & POLAT, K. 2021. Otitis media detection using tympanic membrane images with a novel multi-class machine learning algorithm. *PeerJ Comput Sci,* 7**,** e405-e405.

BOULAHIA, S. Y., AMAMRA, A., MADI, M. R. & DAIKH, S. 2021. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine vision and applications,* 32.

CAI, Y., YU, J.-G., CHEN, Y., LIU, C., XIAO, L., M GRAIS, E., ZHAO, F., LAN, L., ZENG, S., ZENG, J., WU, M., SU, Y., LI, Y. & ZHENG, Y. 2021. Investigating the use of a two-stage attention-aware convolutional neural network for the automated diagnosis of otitis media from tympanic membrane images: a prediction model development and validation study. *BMJ Open,* 11**,** e041139-e041139.

CAMALAN, S., MOBERLY, A. C., TEKNOS, T., ESSIG, G., ELMARAGHY, C., TAJ-SCHAAL, N. & GURCAN, M. N. 2021. OtoPair: Combining Right and Left Eardrum Otoscopy Images to Improve the Accuracy of Automated Image Analysis. *Applied sciences,* 11**,** 1831.

CAMALAN, S., NIAZI, M. K. K., MOBERLY, A. C., TEKNOS, T., ESSIG, G., ELMARAGHY, C., TAJ-SCHAAL, N. & GURCAN, M. N. 2020. OtoMatch: Content-based eardrum image retrieval using deep learning. *PLoS One,* 15**,** e0232776-e0232776.

CASTIGLIONI, I., RUNDO, L., CODARI, M., DI LEO, G., SALVATORE, C., INTERLENGHI, M., GALLIVANONE, F., COZZI, A., D'AMICO, N. C. & SARDANELLI, F. 2021. AI applications to medical images: From machine learning to deep learning. *Phys Med,* 83**,** 9-24.

CHAN, J., RAJU, S., NANDAKUMAR, R., BLY, R. & GOLLAKOTA, S. 2019. Detecting middle ear fluid using smartphones. *Sci Transl Med,* 11**,** eaav1102.

CHICCO, D., TOTSCH, N. & JURMAN, G. 2021. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Min,* 14**,** 13-22.

FRUMKIN, K. 2018. News: Diagnosing Otitis Media? Pneumatic Otoscopy Simplified. *Emergency medicine news,* 40**,** 28-28.

GADDEY, H. L. M. D., WRIGHT, M. T. D. O. & NELSON, T. N. M. D. 2019. Otitis Media: Rapid Evidence Review. *Am Fam Physician,* 100**,** 350-356.

GARDNER, M. W. & DORLING, S. R. 1998. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment (1994),* 32**,** 2627-2636.

GIBSON, E., LI, W., SUDRE, C., FIDON, L., SHAKIR, D. I., WANG, G., EATON-ROSEN, Z., GRAY, R., DOEL, T., HU, Y., WHYNTIE, T., NACHEV, P., MODAT, M., BARRATT, D. C., OURSELIN, S., CARDOSO, M. J. & VERCAUTEREN, T. 2018. NiftyNet: a deep-learning platform for medical imaging. *Comput Methods Programs Biomed,* 158**,** 113-122.

GRAIS, E. M., WANG, X., WANG, J., ZHAO, F., JIANG, W., CAI, Y., ZHANG, L., LIN, Q. & YANG, H. 2021. Analysing wideband absorbance immittance in normal and ears with otitis media with effusion using machine learning. *Sci Rep,* 11**,** 10643-10643.

GUDIVADA, V., APON, A. & DING, J. 2017. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software,* 10**,** 1-20.

HABIB, A. R., WONG, E., SACKS, R. & SINGH, N. 2020. Artificial intelligence to detect tympanic membrane perforations. *J Laryngol Otol,* 134**,** 311-315.

HARMES, K., BLACKWOOD, R. A., BURROWS, H., COOKE, J., HARRISON, V. & PASSAMANI, P. 2013. Otitis Media: Diagnosis and Treatment. *Am Fam Physician,* 88**,** 435-440.

HIERONS, R. 1999. Machine learning. Chichester, UK: John Wiley & Sons, Ltd.

ISAACSON, G. 2014. Endoscopic Anatomy of the Pediatric Middle Ear. *Otolaryngol Head Neck Surg,* 150**,** 6-15.

KAELBLING, L. P., LITTMAN, M. L. & MOORE, A. W. 1996. Reinforcement learning: A survey. *The Journal of artificial intelligence research,* 4**,** 237-285.

KAI, C., HISANG, S. C., NAN, C. C., HSI, C. M., CHANG, W. & Y, E. 2015. Transfer representation learning for medical image analysis. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),* 2015-**,** 711-714.

KATZ, G., SHABTAI, A., ROKACH, L. & OFEK, N. 2014. ConfDTree： A Statistical Method for Improving Decision Trees. *Journal of Computer Science and Technology,* 29**,** 392-407.

KAUR, T. & GANDHI, T. K. 2020. Deep convolutional neural networks with transfer learning for automated brain image classification. *Machine vision and applications,* 31.

KERAS. Available: https://keras.io/ [Accessed 1-10 2021].

KERSTING, K. 2018. Machine learning and artificial intelligence: two fellow travelers on the quest for intelligent behavior in machines. *Frontiers in big Data,* 1**,** 6.

KOTSIANTIS, S. B. 2007. Supervised machine learning: A review of classification techniques. *Informatica (Ljubljana),* 31**,** 249-268.

KRIZHEVSKY, A., SUTSKEVER, I. & HINTON, G. 2017. ImageNet classification with deep convolutional neural networks. *Communications of the ACM,* 60**,** 84-90.

KURUVILLA, A., SHAIKH, N., HOBERMAN, A. & KOVAČEVIĆ, J. 2013. Automated Diagnosis of Otitis Media: Vocabulary and Grammar. *Int J Biomed Imaging,* 2013**,** 327515-15.

LECUN, Y., BOTTOU, L., BENGIO, Y. & HAFFNER, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE,* 86**,** 2278-2324.

LEE, J. Y., CHOI, S.-H. & CHUNG, J. W. 2019. Automated Classification of the Tympanic Membrane Using a Convolutional Neural Network. *Applied sciences,* 9**,** 1827.

LIVINGSTONE, D. & CHAU, J. 2020. Otoscopic diagnosis using computer vision: An automated machine learning approach. *Laryngoscope,* 130**,** 1408-1413.

LIVINGSTONE, D., TALAI, A. S., CHAU, J. & FORKERT, N. D. 2019. Building an Otoscopic screening prototype tool using deep learning. *Journal of otolaryngology-head and neck surgery,* 48**,** 66-66.

LO, S. C. B., LOU, S. L. A., JYH-SHYAN, L., FREEDMAN, M. T., CHIEN, M. V. & MUN, S. K. 1995. Artificial convolution neural network techniques and applications for lung nodule detection. *IEEE Trans Med Imaging,* 14**,** 711-718.

MANSOUR, S. 2019. *Comprehensive and Clinical Anatomy of the Middle Ear*, Springer International.

MCCARTHY, J. 2007. What is artificial intelligence?

MOBERLY, A. C., ZHANG, M., YU, L., GURCAN, M., SENARAS, C., TEKNOS, T. N., ELMARAGHY, C. A., TAJ-SCHAAL, N. & ESSIG, G. F. 2018. Digital otoscopy versus microscopy: How correct and confident are ear experts in their diagnoses? *J Telemed Telecare,* 24**,** 453-459.

MONROY, G. L., WON, J., DSOUZA, R., PANDE, P., HILL, M. C., PORTER, R. G., NOVAK, M. A., SPILLMAN, D. R. & BOPPART, S. A. 2019. Automated classification platform for the identification of otitis media using optical coherence tomography. *NPJ Digit Med,* 2**,** 22-22.

MYBURGH, H. C., JOSE, S., SWANEPOEL, D. W. & LAURENT, C. 2018. Towards low cost automated smartphone- and cloud-based otitis media diagnosis. *Biomedical signal processing and control,* 39**,** 34-52.

MYBURGH, H. C., VAN ZIJL, W. H., SWANEPOEL, D., HELLSTRÖM, S. & LAURENT, C. 2016. Otitis Media Diagnosis for Developing Countries Using Tympanic Membrane Image-Analysis. *EBioMedicine,* 5**,** 156-160.

NOBLE, W. S. 2006. What is a support vector machine? *Nat Biotechnol,* 24**,** 1565-1567.

OPENCV. Available: https://opencv.org/ [Accessed 1-10 2021].

ORLOV, N., SHAMIR, L., MACURA, T., JOHNSTON, J., ECKLEY, D. M. & GOLDBERG, I. G. 2008. WND-CHARM: Multi-purpose image classification using compound image transforms. *Pattern Recognit Lett,* 29**,** 1684-1693.

OXENHAM, A. J. 2018. How We Hear: The Perception and Neural Coding of Sound. *Annu Rev Psychol,* 69**,** 27-50.

PAN, S. J. & YANG, Q. 2010. A Survey on Transfer Learning. *IEEE transactions on knowledge and data engineering,* 22**,** 1345-1359.

PICHICHERO, M. E. & POOLE, M. D. 2005. Comparison of performance by otolaryngologists, pediatricians, and general practioners on an otoendoscopic diagnostic video examination. *Int J Pediatr Otorhinolaryngol,* 69**,** 361-366.

POWERS, D. 2008. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *Mach. Learn. Technol.,* 2.

RAJKOMAR, A., DEAN, J. & KOHANE, I. 2019. Machine Learning in Medicine. *N Engl J Med,* 380**,** 1347-1358.

SANCHEZ, L., CARNEY, A. S., ESTERMAN, A., SPARROW, K. & TURNER, D. 2019. Does access to saltwater swimming pools reduce ear pathology and hearing loss in school children of remote arid zone aboriginal communities? A prospective 3-year cohort study. *Clin Otolaryngol,* 44**,** 736-742.

SCOTT, I. A., COOK, D., COIERA, E. W. & RICHARDS, B. 2019. Machine learning in clinical practice: prospects and pitfalls. *Med J Aust,* 211**,** 203-205.e1.

SINGH, D. & SINGH, B. 2020. Investigating the impact of data normalization on classification performance. *Applied soft computing,* 97**,** 105524.

STANDFORD-CHILDREN'S-HEALTH. 2021. *Otitis Media (Middle Ear Infection)* [Online]. Available: https://www.stanfordchildrens.org/en/topic/default?id=otitis-media-middle-ear-infection-90-P02057 [Accessed 01-10 2021].

WANG, H., SHANG, S., LONG, L., HU, R., WU, Y., CHEN, N., ZHANG, S., CONG, F. & LIN, S. 2018. Biological image analysis using deep learning-based methods: Literature review. *Digital medicine,* 4**,** 157-165.

WANG, X., VALDEZ, T. A. & BI, J. 2015. Detecting tympanostomy tubes from otoscopic images via offline and online training. *Comput Biol Med,* 61**,** 107-118.

WANG, Y.-M., LI, Y., CHENG, Y.-S., HE, Z.-Y., YANG, J.-M., XU, J.-H., CHI, Z.-C., CHI, F.-L. & REN, D.-D. 2020. Deep Learning in Automated Region Proposal and Diagnosis of Chronic Otitis Media Based on Computed Tomography. *Ear Hear,* 41**,** 669-677.

WIKIPEDIA. 2009. *Middle ear* [Online]. Available: https://en.wikipedia.org/wiki/Middle_ear [Accessed 1-10 2021].

WILLIAMS, C. J. & JACOBS, A. M. 2009. The impact of otitis media on cognitive and educational outcomes. *Med J Aust,* 191**,** S69-S72.

WU, Z., LIN, Z., LI, L., PAN, H., CHEN, G., FU, Y. & QIU, Q. 2021. Deep Learning for Classification of Pediatric Otitis Media. *Laryngoscope,* 131**,** E2344-E2351.

YOCKEL, N. J. 2002. A Comparison of Audiometry and Audiometry With Tympanometry to Determine Middle Ear Status in School-Age Children. *J Sch Nurs,* 18**,** 287-292.

ZENG, X., JIANG, Z., LUO, W., LI, H., LI, H., LI, G., SHI, J., WU, K., LIU, T., LIN, X., WANG, F. & LI, Z. 2021. Efficient and accurate identification of ear diseases using an ensemble deep learning model. *Scientific reports,* 11**,** 10839-10839.