Exploration of deep learning algorithms for the use of 3D facial imaging for the detection of Obstructive Sleep Apnoea

Unit Name: Applied Project

Unit Code: CSG3101

Group: 16

Semester: S2, 2024

Supervisor: Dr Syed Mohammed Shamsul ISLAM

Group Members	Student #
Tristan Edward Gardiner	10042832
Mahathir Abdul Basher	10560535
Jon Sveinbjornsson	10396563
Steven Cobain	10475725

Adherence to project proposal

The project did not follow the expected timelines for a multitude of reasons. A large part of this was coordination of meeting times as students were in different time zones and countries. The project is being delivered late, with the use of an extension.

Project Outputs (deliverables)

The project has all provided all the deliverables stated in the project proposal.
Timeline was not accurate for all items.
The final predictive system is present.
The source code is present.
The data is present.
The project report is present.
The optional literature review is not present.
The project demo is present: https://youtu.be/1rYbhSP7-kY

Quality indicators

Yes, the provided outputs are a complete end to end user application with all project deliverables embedded in the user interface. The Model was developed, trained, and useable for inference within a single page application. The data was collected as requested at medical sites and at ECU Joondalup.

Teamwork (collaboration, communication and professionalism)

Despite living in different countries, we were able rally together and spend the time required to complete the project to a high standard we were all proud of.

Abstract

Purpose: The aim of this study is to propose an accurate approach to detect obstructive sleep apnoea with the use of a deep neural network for end-to-end classification of inputted 3D craniofacial scans to be potentially used for an online web application.

Method: The use of Multiview Landmark detection was adopted to for the automation in preprocessing the 3D craniofacial scans for input into a custom-made classifier utilising the PyTorch library. Allowing seamless results for user inputted 3D scans for classification.

Results: The final classification model achieved an accuracy of 100% over 10 folds, with a mean AUC of 77. Novel data tests gave an accuracy score of 60-70%.

Conclusion: This study shows that a classification tool for the detection of obstructive sleep apnoea can be achieved with 3D craniofacial scans and deep learning models.

Contents

1.0 Introduction

Obstructive Sleep Apnoea (OSA) is a major modern day health concern affecting 15% of middle-aged men and 5% of middle-aged women (Eastwood et al., 2020; Chen et al., 2024). It estimated that 75% of cases are unidentified, with a majority of cases being misdiagnosed as symptoms of other causes, with a majority of OSA sufferers unaware that they have the condition. OSA has serious implications on health as it lowers blood oxygen levels to vital organs causing serious problems such as hypertension and atherosclerosis, attributing to major cardiovascular problems such as heart attacks and strokes (Chen et al., 2024). As OSA can be attributed to a lower quality of sleep causing daytime drowsiness (Shetty & Jardin, 2023), which has been linked to road safety concerns from fatigued drivers (Chen et al, 2024). Thus, the need for a faster, cheaper, and accurate model for prediction, and potentially diagnosis, of sleep apnoea would be beneficial to the health and safety of the larger population.

Obstructive Sleep Apnoea (OSA) is a respiratory disorder which occurs during sleep, characterised by a partial or complete blockage of the airways resulting in a temporary pause in respiration. The cause of the obstruction is due to the closing of the upper airways, including the tissues at the base of the tongue, soft palate, and walls of the throat (Shetty & Jardin, 2023). These obstructions, also known as events, can occur multiple times within an hour. These events per hour are recorded as the apnoea-hypopnea index (AHI), creating a measure for the severity of the condition (Chen et al, 2024). An event is either an apnoea, complete obstruction of the airways, or a hypopnea, a partial blockage of the airways, lasting for more than 10 seconds. Mild cases of OSA in adults record an AHI of 5 to15, moderate cases occur between 15 and 30, and severe cases conclude with 30 or more events per hour. Although it should be considered that AHI records the frequency of events but may not be an accurate record the severity of the overall condition, as a patient with a low AHI could have only severe events as opposed to a patient with a high AHI suffering with only mild hypopnea events (Shetty & Jardin, 2023). Nevertheless, AHI is an important measurement as it indicates a decrease in blood oxygen supply levels with each event taken place, accumulating to the potential development of more severe future health issues.

The current benchmark for diagnosing OSA is nocturnal polysomnography (PSG), this is an overnight supervision of a patient's sleep within a sleep clinic setting via sleep technicians monitoring various sensor outputs, such as respiration rate, blood oxygen levels, rapid eye movements, body movements and twitches, as well as electroencephalograph (EEG), and electrocardiograms (ECG) readings (Hanif et al., 2021). Although these sessions are very thorough, there are certain disadvantages of PSG that make it impractical compared to a more statistical approach. The main issues are due to the expense of the sessions due to the need for trained sleep technicians within a clinical setting to observe the patient throughout the night, whom then need to review and evaluate the data which can be prone to human error. Data may also be misinterpreted as the clinical setting may not be comfortable for the patient presenting an inability to acquire a true representation of their typical night of sleep: as they are in a foreign setting with the need to wear specialty equipment wired to machines for reading all the required outputs. Consequently, this outlines the need for a more practical approach that artificial intelligence models can help to achieve.

To determine what input data is most adequate to train an artificial intelligence (AI) model, a look into the common causes of OSA should be looked at. The pathology of OSA is not completely determinate on a single cause as there are multiple contributing factors attributed to apnoea (Hanif et al., 2021). Such factors include genetics, gender, age, obesity, smoking, alcohol and other drug use. Although these are all contributing factors to OSA in different levels, a more independent variable can be determined via cranial facial features, or craniofacial morphology (Ozdemir et al, 2019; Eastwood et al, 2020). Such features are related to skeletal restriction, regional adiposity, and obesity. Craniofacial risk factors include a wider and flatter mid and lower face, a shorter and retracted mandible, a smaller enclosed area within the mandible, and more soft tissues or fat deposition on the anterior neck (Eastwood et al., 2020; Chen et al., 2024). It has also been linked to characteristics of a long face coupled with mandibular prognathism or retrognathism (Monna et al., 2022). Therefore, the use of 3D imaging of craniofacial anatomy would be beneficial to train an AI model as it can determine the shape and contour of a patient's facial surface structure, which 2D imaging would not typically supply detailed surface depth.

The purpose of this study is to use a dataset of 3D craniofacial scans to propose an accurate approach for the detection of OSA with a deep neural network for end-to-end classification of inputted 3D craniofacial scans to be potentially used for an online web application. There are two potential approaches considered: using pre-trained face recognition classifier to establish feature differences within the dataset, or by using automatic landmark detection to establish distances between features to feed into a deep learning classifier to determine OSA in the dataset. Potentially, both approaches combined could be a beneficial, but research, testing, and successful implementation will be factors in its reliability.

As the scope of the research only involved only 3D craniofacial scans, research results for 2D photography were omitted from the research unless it contained relevant information about using artificial intelligence learning models with medical images for OSA. Key words such as "obstructive sleep apnea", "obstructive sleep apnoea", "craniofacial scan", "craniofacial morphology", "morphology", "3D", "three-dimensional", "artificial intelligence", "deep learning", "machine learning", and "medical imaging" were used with Google Scholar to find research papers on the study of AI and OSA. The literature review starts with the influential parent research paper that was published 2018 (Islam et al., 2018), and then only papers from the last 6 years to present were selected within the research results. These searches also coincided with the selection of relevant articles from the reference lists on related papers. Overall, 9 papers were selected for a literature review.

2.0 Related Works

We start the journey of using deep learning models with 3D craniofacial scans for the detection of OSA, with a study done in 2018, from Islam et al, utilising the 3D mapping of facial structures to be converted into 2D mesh depth maps (Islam et al., 2018; TaghiBeyglou et al., 2024). Although the research consisted of a small dataset, 39 male and 30 female adults, it is the first to research depth map-based models for sleep apnoea detection pioneering a novel approach. The 3D scans were collected and converted to 2D depth maps of the frontal facial features. Three pretrained models were selected using VGG Face and Pose Aware CNN Models (PAMs) for facial recognition with PAMs-VGG-19, and PAMs-AlexNet. The outcome found VGG Face to be the most proficient with transfer learning for depth maps, with the other two models facing small issues with pose and rotation problems within the dataset. The model achieved validation accuracy of 68.75% and a test accuracy of 67.42%, with directly fed input images to the model achieving 62%. The network achieved end-to-end classification of OSA using depth facial data with a promising performance.

Ozdemir et al. conducted a study to investigate craniofacial differences within adults with and without OSA (Ozdemir et al., 2019) with sample size of 106 Turkish patients, 50 with OSA and 56 without. The participants were all involved in a PSG, with a 3D surface scan being taken at the sleep clinic. The group focused on 12 landmarks of the front of the face, mostly involving the nasal region. Geometric morphometric analysis was used to determine the difference between each participant, but results gave no statistical comparison between the landmarks selected. Although differences were found within some of the inter-landmark distances located with the nasal region, showing a 11% greater distance in OSA sufferers compared to 29% in non-OSA sufferers, giving the hypothesized conclusion that the found inter-landmark distances are a possible indicator of OSA causing a decrease in the flow of air within the oral cavity. Thus, the importance of this study showed that greater distances shown within landmarks may be indicative of the occurrence of OSA.

Eastwood et al. (2020) also uses 3D facial surface analysis of linear and geodesic (shortest line between points over a curved surface) distances to determine the best combinations to predict OSA severity. They used a larger sample size of 400, with 100 adults as a control group, 100 with mild AHI, 100 with a moderate AHI, and 100 with a severe AHI. The group selected 24 landmarks to measure using linear and geodesic distances within each 3D surface scan. To determine the accuracy of these measurements, linear discriminant analysis and receiver operating characteristic analysis was used on different combinations of measurements which is then repeated on each of the mild to severe AHI groups. They found that the Geodesic measurements improved the ability to identify individuals with OSA, with an accuracy of 86% with and 89% without OSA, $P < 0.1$. The implementation of linear and geodesic measurements was used with a single predictive algorithm to yield the best results, with a 91% accuracy rate. The conclusion of this study showed that the use of linear and geodesic measures gives the best results in the prediction of OSA with 3D surface scans.

The two primary aims of Hanif et al's 2021 study were to investigate the accuracy of a system predicting AHI and identifying which regions of the face and neck were most useful in predicting the severity of OSA (Hanif et al., 2021). The dataset for the study at the time was four times larger than any other study. Unfortunately, the dataset is not publicly available due to 3D facial scans being considered personally identifiable information which would result in confidentiality issues. The data was collected at 11 different sleep clinics where 1756 3D scans were performed using a Structure Sensor device from Occipital Inc. attached to an iPad Pro. The machine learning, or ML, model based on the craniofacial images had an overall accuracy of 67% in predicting AHI. The main benefit of the study was the bypassing of manual labour in using a landmark-based measured features approach that was commonplace with other studies at the time. Also, the ML model appeared to be just as accurate as three sleep specialists in predicting AHI. The most important features of the face in predicting high AHI values were those of neck, jaw, and midface (Hanif et al., 2021). Limitations within the study included quality of the 3D scans as it varied significantly due to the 11 different locations as well as lightning conditions. The study showed that it is possible to derive AHI values from 3D scans and deep learning techniques with accuracy on par or better than experienced sleep physicians.

Monna et al (2022) used 3D scans using geometric morphometrics focusing on the neck and the submandibular section for their association for OSA diagnosis from a selected dataset of 280 Caucasian men, who participated in a sleep clinic study with PSG, which also included a 3D scan at the clinic. The scans were cleaned and processed into PLY file format, giving a 3D polygon file type with 7 manually plotted landmarks on the 3D mesh. The submandibular landmarks were located on each ear lobe, the nasal bridge and the tip of the chin. The other 3 landmarks were identifiers of the lower neck boundary with a landmark on each side of the acromioclavicular joints, and the last one in-between, at the sternal fork. The outer landmark coordinates were used to align the 3D meshes within 3D space using Generalized Procrustes Analysis (GPA), with pose corrections being minimized at scanning with specially designed glasses with three spirit levels and slits in the arms to align to the participants eye level. 13 Machine Learning algorithms were selected to be trained and tested consisting of simple to advanced techniques, with 2 to 5 PCs. The best performing algorithms were LR, LDA, Adaboost, Extra Trees Classifier, and XGBoost. It was found that LR was the preferred classifier as its efficiency surpassed the rest, obtaining an auROC of 0.70, with a specificity of 60%, and a sensitivity of 74%. Furthermore, selecting the best performing algorithm for morphometric data, with an added layer for implementing the questionnaire and anthropometric data from ten selected variables and symptoms using the XGBoost algorithm for numerical and categorical inputs. The performance scores are then boosted to an auROC of 0.75, 56% specificity, and 80% sensitivity, giving it a 95% confidence score. As a result, this study shows that there is a correlation between OSA, and the shapes of the submandibular and neck areas within their participant sample. More importantly, this study shows an approach that machine learning can be quick and efficient, as well as an inexpensive way to screen for OSA.

He et al. (2022) conducted a study utilising deep learning (DL) for obstructive sleep apnoea (OSA) diagnosis based on craniofacial photographs. In this study, the researchers captured images from five facial angles: frontal view, 90° left and right profiles, and 45° left and right profiles. They employed a convolutional neural network (CNN) as their DL method. The model's performance was assessed using sensitivity, specificity, and area under the curve (AUC). To evaluate the prediction accuracy, the researchers compared the CNN's results against polysomnography (PSG), which is considered the gold standard for OSA diagnosis. The study involved a total of 393 participants. At an AHI limit of 5 events/h, the model achieved a sensitivity of 0.95, a specificity of 0.80, and an AUC of 0.916 (95% CI: 0.847–0.960). At a higher limit of 15 events/h, the sensitivity was 0.91, the specificity was 0.73, and the AUC was 0.812 (95% CI: 0.729–0.878). The study concluded that the CNN model demonstrated both effectiveness and efficiency in detecting OSA. The results highlight the potential of this deep learning technique to serve as a valuable tool in clinical settings for diagnosing OSA in patients.

A similar study conducted by Ohmura et al. (2022) explored the prediction and assessment of OSA severity in patients by analysing jaw measurements from facial profiles, independent of sex and weight. The study involved 37 adult patients with ages ranging from 30 years to 86 years. This study follows the same methodology as that of He et al. (2022), where PSG was used to establish a baseline for comparing the results of the DL model. The results revealed that the participants' OSA severity levels were classified as follows: “normal (AHI < 5, n = 5), mild (5 ≤ AHI < 15, n = 5), moderate (15 ≤ AHI < 30, n = 11), and severe (AHI ≥ 30, n = 16)” (Ohmura et al., 2022). The facial features were scanned using a device called the Shining 3D EinScan Pro. This non-contact camera uses a pattern of white LED light to scan the object. It collects data by capturing how the light pattern reflects off the surface from different angles, allowing it to recognize the shape of the face. The camera scanned each patient whilst moving around their face for approximately 30s, and the scans were then measured using Fusion 360, a 3D modelling program. The 3D photogrammetry captures various mandible features: Mandibular Width (Mw) is the distance between the right and left gonion. Mandibular Depth (Md) measures the distance from the mentum straight out to the width. Mandibular Length (Ml) is the span from the gonion to the mentum. Mandibular Width-Length Angle (Mwla) is the angle formed between the left and right gonion and the mentum. Lastly, Mandibular Area (Ma) refers to the area enclosed by the right gonion, mentum, left gonion, and the width. In statistics, the R value measures how strongly two variables are related, with values ranging from -1 to 1 (JMP, n.d.). A value close to 1 or -1 indicates a strong relationship, while a value near 0 suggests a weak or no relationship (JMP, n.d.). The p value indicates the likelihood that the observed relationship is due to chance. A p value less than 0.05 typically means the relationship is statistically significant (JMP, n.d.). Ohmura et al. (2022) found that Mwla, Mw, and Md were significantly linked to the severity of OSA. Mwla showed a strong correlation ($R = 0.73$, $p < 0.01$), while Mw ($R = 0.39$, $p < 0.05$) and Md ($R = -0.34$, $p < 0.05$) also had significant associations. Even after accounting for sex, age, BMI, and neck circumference, Mwla and Md continued to be significant predictors of the AHI. Furthermore, the diagnostic analysis revealed that Mwla was useful for detecting OSA ($AHI \geq 5$). With a cutoff value of 78.6, it had a sensitivity of 0.938 and a specificity of 0.800. The AUC was 0.931, showing excellent performance. The study indicates that Mwla, measured with 3D photogrammetry, is a dependable predictor of the sleeping disorder and relates its severity in patients, regardless of their obesity or sex (Ohmura et al., 2022).

A study conducted by Chen et al. (2023) followed a different approach for OSA detection. This paper focused on using two-dimensional craniofacial photographs captured by a Pad Mi 2s, a Huawei brand mobile device. This study focused on using 2D craniofacial imagery to detect OSA with machine learning techniques, as it is more time-efficient and can be easily conducted by anyone with a mobile device, allowing for repeated assessments. The study uses a dataset involving 653 participants, the majority of whom were male, with just over half diagnosed with OSA. In addition to the participants in the dataset, they had 19 clinical variables that were linked to OSA, some of which include facial characteristics, medical history and demographic details. Each participant was placed in a controlled environment setting (predetermined lighting and location) and a 2D photograph captured the front and right side of the face of each participant. The method evaluated 18 machine learning algorithms, and after further analysis and performance comparison, the top five models were identified: CATBOOST, LIGHTGBM, RBFSVM, ET, and LR. From the top 5, CATBOOST demonstrated the best performance for OSA detection, achieving a sensitivity of 0.75, specificity of 0.66, accuracy of 0.71, and an AUC of 0.76. Chen et al. (2023) concluded that 2D craniofacial images have shown a promising ability to predict OSA in patients by extracting facial features, potentially enabling individuals to engage in self-diagnosis for OSA where they can conduct repeated assessments in a timely manner.

Collier et al (2023) used 3D-sterophotogrammetry with landmarks placed on 3D images to measure distances using linear, angular and volumetric measurements to compare facial and neck characteristics, adjusted for BMI and sex, to determine OSA severity by predicting AHI. The study consisted of 91 participants, 61 male and 30 female. The participants all were involved in a PSG where age, BMI, as well as neck, abdominal and hip circumferences where recorded. The following day, a stereophotogrammetry scan was obtained using 3dMD head motion system, with orientation correction and adjustments was done sing image software 3dMDVultus. Landmarks where then added based on previous works by Lee et al (2009). This study showed another example of using 3D scans to identify craniofacial morphology, in conjunction with commonly related variables and symptoms, could be used to predict OSA severity of 51%. Thus, also predicting the presence of OSA within the participants.

Year	Study	Dataset size	Deep Learning Model (if used)	Major Outcome of Results
2018	Islam et al	69	VGG-Face, PAMs with VGG-19 and AlexNet.	The network achieved end-to-end classification of OSA using depth facial data with a promising performance with test accuracy of 67.42%.
2019	Ozdemir et al	106	n/a	Inter-landmark distances are possible indicator of OSA but no indication of a statistical correlation between landmarks were found.
2020	Eastwood et al	400	n/a	91% accuracy was achieved by combining geodesic and linear measurements.
2021	Hanif et al	1366	Deep-MVLM (for preprocessing), ResNet18.	Using a combination of craniofacial scans, demographics and questionnaire data, the model achieved 67% accuracy to predict AHI values. When the output was compared with sleep specialist ranking the data, a similar accuracy was observed.
2022	Monna et al	280	13 different supervised algorithms tested. Best results were using LR, LDA, Adaboost, Extra Trees Classifier, and XGBoost.	With a combination of questionnaires and anthropometric data with morphometric data, a 95% confidence score was achieved.
2022	He et al	393	EfficientNet-B4 (for preprocessing), LightGBM.	Efficiency in detecting OSA in photographs from the input of raw 2D photos without landmark detection with an Accuracy of 90%.
2022	Ohmura et al	37	Statistical Model: LinearR	Found that a strong correlation with Mwla measurements and AHI severity exists, regardless of gender or BMI.
2023	Chen et al	653	18 Common ML models used,	Accuracy of 71%, a machine learning approach using 2D mobile phone

			CATBOOST performed the best.	photography can be effective and efficient for OSA detection.
2023	Collier et al	91	Statistical Model: LinearR	Width of the face and soft tissue at the neck measurement were found to be indicative of OSA.

Table 1. Overview of research papers investigated, including their deep learning models (if used) and their main findings.

2.1 Dataset Collection

The datasets used for OSA detection are not available to the public due to privacy of the participants (TaghiBeygou et al., 2024). Due to this reason, there are not many studies that have had extensive datasets from a large demographic, as typically the data is collected in conjunction with local sleep clinics (Islam et al., 2018; Ozdemir et al., 2019; Eastwood et al., 2020; Hanif et al., 2021; Monna et al., 2022; He et al., 2022; Ohmura et al., 2022; Chen et al., 2023; Collier et al., 2023) and, thus, commonly came in conjunction with data from questionnaires and medical data from the subsequent PSG results. As a results of this, quite a few of the studies selected have a small dataset, with 4 having 106 or less participants (Islam et al., 2018; Ozdemir., 2019; Ohmura et al., 2022; Collier et al., 2023) although it could be argued that the results varied based on the approach that was taken, as the results did not necessarily reflect the dataset size (see table 1 for a breakdown on each study). Therefore, it seems that a larger cooperative dataset, from multiple sleep clinics, would be beneficial for future development of a trustworthy OSA detection model.

2.2 Feature Selection

Feature selection is dependent on the areas that each study had predicted to have significance to an association with OSA and AHI. A few of the studies did not rely on landmarks for statistical analysis as they used a pretrained deep neural network to establish their feature extraction, which included VGG-Face and PAMs with VGG-19 and AlexNet (Islam et al, 2018), as well as ResNet-18 (He et al., 2021) and EfficientNet-B4 (He et al., 2022). In such models, an overall craniofacial scan focusing on the frontal face, submandibular and neck regions are proven to be sufficient for promising results (Islam et al., 2018; Hanif et al., 2021; Monna et al., 2022; Ohmura et al., 2022). Adopting automation for defining landmarks has been found to be more reliable than manually placing the landmarks on a 3D mesh or 2D photography (TaghiBeyglou et al., 2024). Landmark placement is important as they are the points for measuring distances between two or three points of reference, via either Euclidean or Geodesic measurements, to obtain differences in facial structures (Eastwood et al., 2020; Monna et al., 2022). Overall, both methods have found that the focus on the jaw line and neck are important features to include for the prediction of OSA.

2.4 Deep Learning, Machine Learning and Statistical Models

The deep learning methods only appeared a few studies (Islam et al., 2018; He et al., 2022; Hanif et al., 2021), others used statistical models (Ozdemir et al., 2019; Ohmura et al., 2022; Collier et al., 2023), or machine learning methods were used (Eastwood et al., 2020; Monna et al., 2022; Chen et al., 2023). The best results found was with a combination of questionnaires and anthropometric data with morphometric data, a 95% confidence score was achieved using machine learning techniques (Monna et al., 2022). This was done using LR as a predictive model and classifying the additional collected data together with XGBoost. Although promising results came from deep learning models and automated landmark placement, achieving 90% accuracy from 2D imagery with EfficientNet-B4 and LightGBM (He et al., 2021). It could be hypothesized that deep learning models with automated landmark detection is a superior approach for 2D and 3D facial photography for the detection of the presence of OSA.

3.0 Methodology

3.1 Dataset

The dataset obtained for this model was cleaned and pre-processed for the reduction of noise and missing gaps in the surface data beforehand and thus was ready for input into the model in the format of ply files. The dataset was obtained from ECU science staff and coincidently was the same dataset for the research from Islam et al. (2018), which was obtained from patients of Genesis SleepCare undergoing home and lab-based sleep studies. The dataset consists of 39 males and 30 females, including 37 with OSA and 32 without, making a total of 69 craniofacial surface scans. Although the dataset is small, and some problems may arise from the lower training dataset, adequate results displayed from previous studies still pertains plausibility to acquire an adequate outcome in the model's overall results.

3.1 MVLM Automatic Detection approach

The MVLM Automatic Detection approach was a tool called 'Deep-MVLM' (Paulsen et al., 2019). This pre-trained model utilised the DTU3D-depth configuration for placing landmarks on three dimensional scans of the face and neck to create a dataset for prediction of OSA. The dependencies required for the Deep-MVLM tool were:

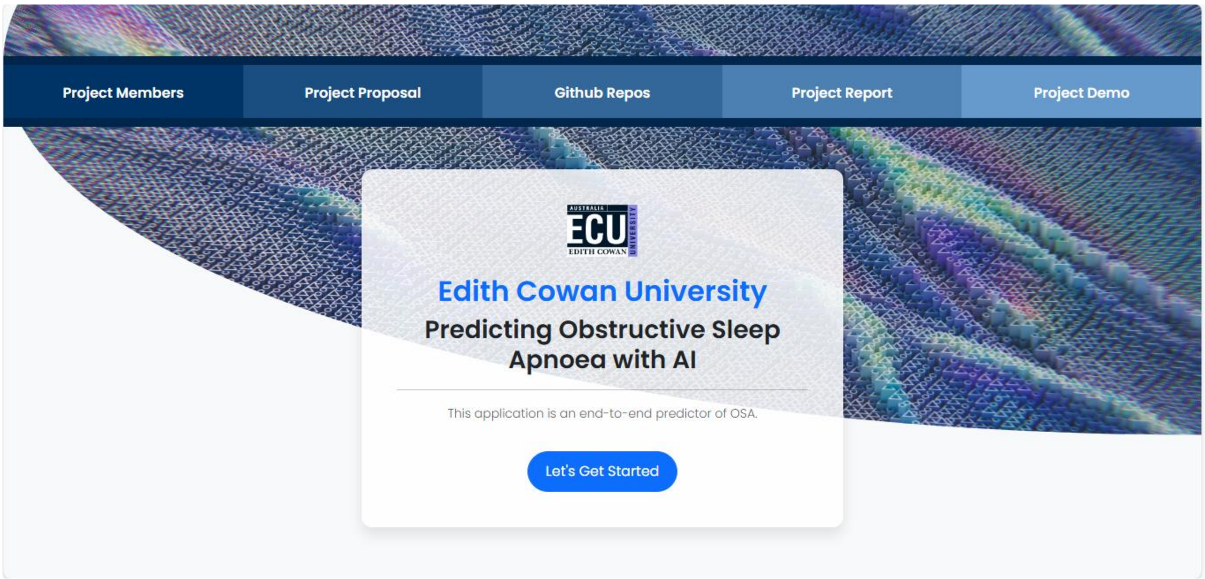
- Pytorch 1.2 (framework for building machine learning models)
- Vtk 8.2 (image processing)
- Libnetcdf 4.7.1 (array-oriented data access)
- Imageio 2.6 (interface for reading and writing image data)
- Matplotlib 3.1.1 (data visualisation)
- Scipy 1.3.1 (solving mathematical problems)
- Scikit-image 0.15 (image processing)
- Tensorboard 1.14 (tracking and visualising metrics such as accuracy and loss)
- Absl-py 0.8 (building Python applications)
- 3.2 Classifier

A classifier was coded from the ground up by the team's technical lead using Pytorch (Paszke et al., 2017). K-fold cross validation was utilised to prevent overfitting, which occurs when machine learning models are trained too well on training data and then perform poorly on unseen novel data. Cross validation involved evaluating the model on multiple validation sets and calculating an average to ensure a more robust predictive performance. The ideal scenario would be to have a dataset which enables a 'goldilocks' number of folds – not too little and not too much; too little folds would result in not enough training for adequate predictive ability, and too many folds would cause meaningful metrics and noise to be harder to differentiate if novel data isn't as clean as training data. Therefore, the limiting factor with cross validation was the small size of the dataset. The hyperparameters for the classifier algorithm were learning rate, batch size, number of epochs, and number of folds.

The body of the classifier code contains separate for loops for the folds and epochs. For the folds, the for loop creates data loaders for the current fold by creating training and validation subsets. The average accuracy across the folds is calculate with the following line of code:

```
average_accuracy = sum(fold_results) / len(fold_results)
```

In addition to the classifier, a user interface application was also designed by the technical leads for the purposes of uploading facial images for OSA prediction. An integral part of UI development was providing an easy end-to-end user experience. The top banner of the UI contains links to the four deliverables, and the centre of the screen enables the user to upload a facial image for analysis.



Edith Cowan University

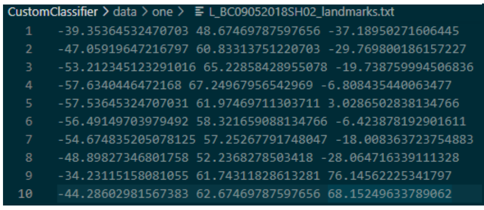
Predicting Obstructive Sleep Apnoea with AI

This application is an end-to-end predictor of OSA.

Figure 1. Web based user interface for the classification of OSA from 3D scans.

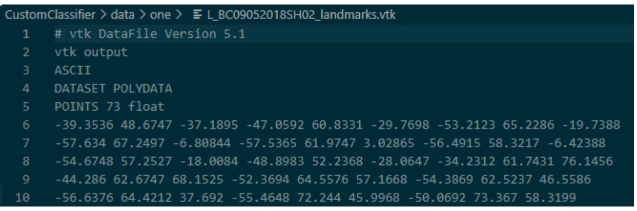
3.2 Image Preprocessing

To acquire the dataset that was used to train the classifier, 3D scans from the DTU3D-depth configuration were run through the MVLM tool using a script. The data was comprised of three-dimensional facial scans that were cleaned to remove holes, eliminate noise that would confound meaningful metrics, and resolve any posing issues. The script called **predict.py,** developed by the technical leads, generated **.txt** and **.vtk** files from the DTU3D-depth configuration which were assigned to directories labelled '0' and '1' representing non-OSA and OSA respectively.



CustomClassifier > data > one > L_BC09052018SH02_landmarks.txt

```
-39.35364532470703 48.67469787597656 -37.18950271606445
-47.05919647216797 60.83313751220703 -29.769800186157227
-53.212345123291016 65.22858428955078 -19.738759994506836
-57.6340446472168 67.24967956542969 -6.808435440063477
-57.53645324707031 61.97469711303711 3.0286502838134766
-56.49149703979492 58.321659088134766 -6.423878192901611
-54.674835205078125 57.25267791748047 -18.008363723754883
-48.89827346801758 52.2368278503418 -28.064716339111328
-34.23115158081055 61.74311828613281 76.14562225341797
-44.28602981567383 62.67469787597656 68.15249633789062
```



CustomClassifier > data > one > L_BC09052018SH02_landmarks.vtk

```
# vtk DataFile Version 5.1
vtk output
ASCII
DATASET POLYDATA
POINTS 73 float
-39.3536 48.6747 -37.1895 -47.0592 60.8331 -29.7698 -53.2123 65.2286 -19.7388
-57.634 67.2497 -6.80844 -57.5365 61.9747 3.02865 -56.4915 58.3217 -6.42388
-54.6748 57.2527 -18.0084 -48.8983 52.2368 -28.0647 -34.2312 61.7431 76.1456
-44.286 62.6747 68.1525 -52.3694 64.5576 57.1668 -54.3869 62.5237 46.5586
-56.6376 64.4212 37.692 -55.4648 72.244 45.9968 -50.0692 73.367 58.3199
```

The classifier was then trained on these files. The line of code that ran the script to create the file types can be seen below:

```
python predict.py --c configs/DTU3D-RGB.json --n <txt file> <vtk file>
```

Within the script, the display command was commented out so it could be automated over the whole dataset. The line of code commented out was:

```
#dm.visualise_mesh_and_landmarks(file_name, landmarks)
```

The ReLU activation function was utilised in the script for binary classification of 0 and 1 values to the dataset. The benefit of this function in machine learning models is its efficient computation as the only operations utilised are addition and multiplication in addition to comparison. The only potential limiting factor of the ReLU function is neurons within the network can become inactive for all inputs which may reduce the capacity of the model for prediction of OSA the longer it is used.

4.0 Results

The model achieved an AUROC score of 66% across all folds. The model is consistently capable of classifying 6 or 7 out of 10 novel and unseen data files correctly, which aligns with the training statistics. We were able to train multiple versions of the classifier to confirm its accuracy using the same dataset, after manually altering the dataset with each newly trained model. The manual alterations were to take out 10 data items, randomly, to use as manual validation sets, which was completed with python scripts running the inference of the trained model over the novel data. The overall accuracy across these tests ranged between 60% and 70% consistently. These results were also verified using the absolute script located in the CustomClassifier directory, ensuring consistency in performance evaluation, which is an automation script used for displaying the model's performance on novel data.

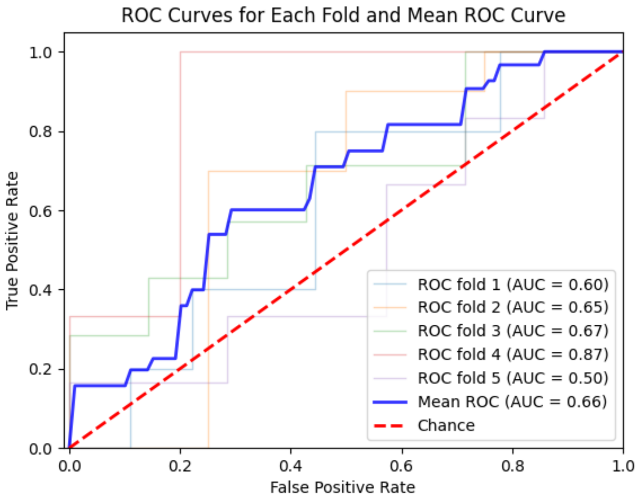


Figure 2. Results showing the ROC curve and AUC over 5 folds.

5.0 Discussion

The MVLM automatic detection approach combined with the custom classifier algorithm aligned with the most promising research in the field of OSA prediction. The review of relevant literature showed that machine learning models have predictive ability surpassing sleep technicians with the added benefit of increased speed of analysis and decreased overheads. Accuracy of the classifier for OSA prediction was 79.33%; this score improved upon the scores of reviewed studies using similar models such as Islam et al.'s (2018) study and Hanif et al.'s (2021) study with 67.42% accuracy and 67% accuracy respectively. Confidence scores of over 90% were recorded in some studies, though this denoted the probability of an event perceived by the algorithms rather than the ability of the algorithms to make precise predictions (accuracy).

Manipulation of the hyperparameters of the classifier enabled the technical lead to achieve an accuracy score of 66.6% average across all folds combined. The learning rate of the classifier algorithm was set to 0.001 which was found to be equal or superior for OSA prediction compared to slower learning rates such as 0.0001 or 0.00001. Just as the number of folds in cross validation is important to not be too high or too low, the rate of learning should exhibit a goldilocks attribute; if the learning rate is too low, training the model may be too slow and it may become stuck. Likewise, if the learning rate is too high, the model may overshoot. It also assists with preventing overfitting as well as underfitting, which is where models fail to recognise patterns in the dataset. Due to the smaller size of the dataset, having a learning rate faster than 0.0001 would result in inferior performance.

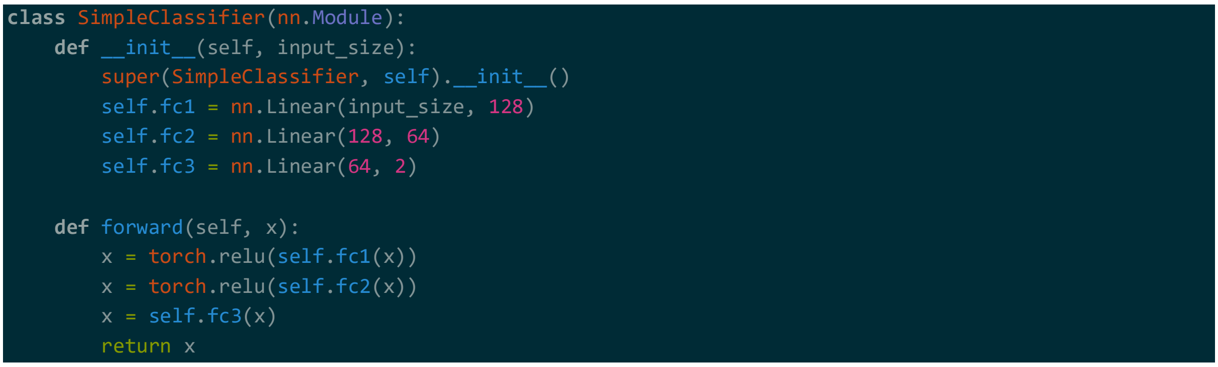
The batch size of the classifier algorithm defined the number of samples to work through within the dataset before updating the internal parameters of the model. As batch size increases, model performance decreases in terms of generalisation, or the ability to adapt and perform well on new unseen data despite high performance on training data (Devansh, 2022). The most effective batch size for the classifier was found to be 30.

The number of folds hyperparameter concerned the k-fold cross-validation and was set to 5. A higher number of folds would train the model on a larger training set and test it on a smaller test fold, with the opposite being true for a smaller number of folds. It is expected that a higher prediction error on average would occur with a smaller number of folds compared to a higher number of folds. Performance was repeated with cross-validation by manipulating the number of folds and observing the difference in prediction error, as an effective method for creating a robust predictive model (Olsen, 2024).

The number of epochs used in the classifier algorithm was set to 1000 for early stoppage, though the longest run across the many training tests performed was 31 epochs. An epoch refers to one complete pass of the training dataset by the algorithm. While a greater number of epochs does not guarantee better results, the smaller size of the dataset necessitated a greater number of passes to produce acceptable predictive ability for OSA. Just as with the learning rate hyperparameter, too many epochs may result in overfitting which reduces generalisation, and too few epochs may result in underfitting or inability to recognise meaningful patterns relevant to detecting OSA.

The technical lead achieved an AUC of 0.77 by further modifying the classifier, tuning hyperparameters, and incorporating a sigmoid activation. Changes included using the ADAM optimizer, a learning rate scheduler, model reloading, and smaller batch sizes to prevent overfitting. The new classifier version also introduced batch normalization, dropout layers, and increased hidden layer sizes, improving regularization and network capacity. These additions, along with the use of a sigmoid function for binary classification, helped stabilize training and enhance performance.

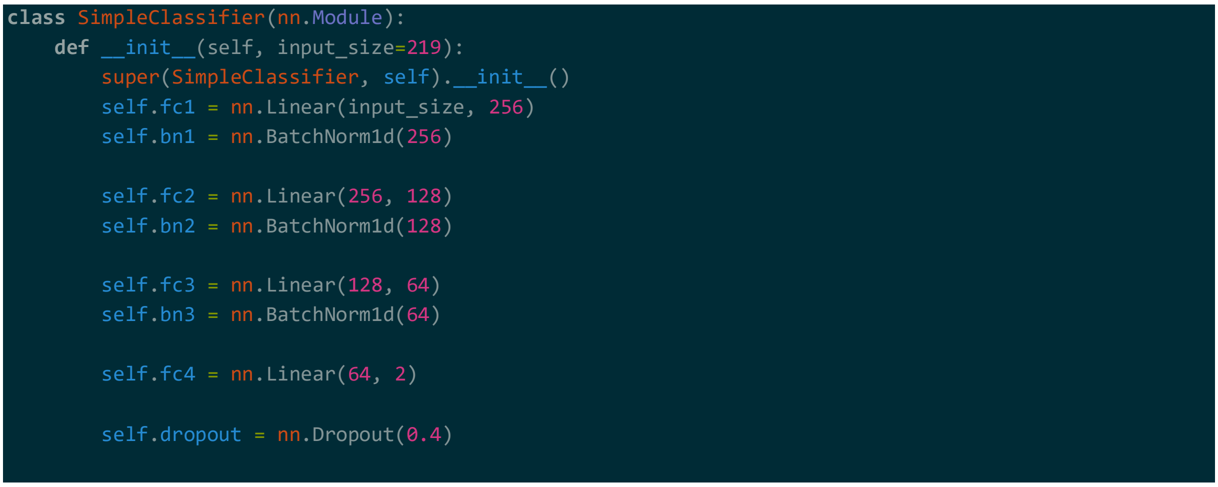
The original classifier architecture:



```
class SimpleClassifier(nn.Module):
    def __init__(self, input_size):
        super(SimpleClassifier, self).__init__()
        self.fc1 = nn.Linear(input_size, 128)
        self.fc2 = nn.Linear(128, 64)
        self.fc3 = nn.Linear(64, 2)

    def forward(self, x):
        x = torch.relu(self.fc1(x))
        x = torch.relu(self.fc2(x))
        x = self.fc3(x)
        return x
```

The new classifier architecture:



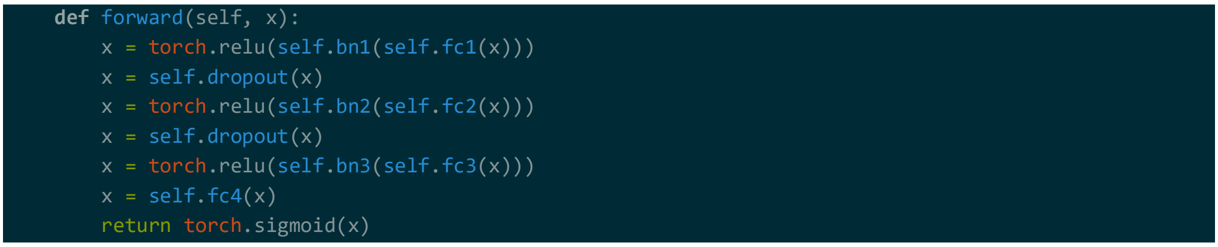
```
class SimpleClassifier(nn.Module):
    def __init__(self, input_size=219):
        super(SimpleClassifier, self).__init__()
        self.fc1 = nn.Linear(input_size, 256)
        self.bn1 = nn.BatchNorm1d(256)

        self.fc2 = nn.Linear(256, 128)
        self.bn2 = nn.BatchNorm1d(128)

        self.fc3 = nn.Linear(128, 64)
        self.bn3 = nn.BatchNorm1d(64)

        self.fc4 = nn.Linear(64, 2)

        self.dropout = nn.Dropout(0.4)
```



```
    def forward(self, x):
        x = torch.relu(self.bn1(self.fc1(x)))
        x = self.dropout(x)
        x = torch.relu(self.bn2(self.fc2(x)))
        x = self.dropout(x)
        x = torch.relu(self.bn3(self.fc3(x)))
        x = self.fc4(x)
        return torch.sigmoid(x)
```

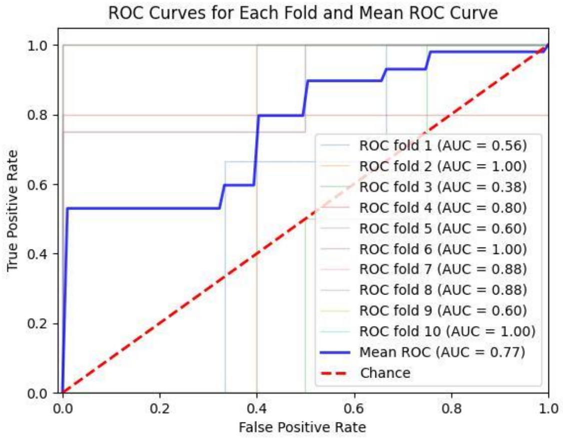


Figure 3. Updated results showing the ROC curve and AUC over 10 folds.

5.1 Challenges and future works (lessons learnt)

As previous discussed, the primary challenge that limited the predictive ability of the deep learning model was working with a small dataset. While hyperparameter manipulation achieved higher accuracy levels when finetuned, the smaller size of the dataset affects the generalisability and transferability of the model.

Challenges may arise with data quality when users upload facial images using the UI. While data was cleaned when training the model to reduce noise and resolve posing issues, the presence of such confounding factors in novel data may prevent the model from predicting OSA effectively. Further issues that may hinder predictive performance may include inadequate lighting and poor image resolution.

Due to time constraints the team was unable to train the model extensively. For future works, updating the 3Dutils code in the deep-MVLM tool to produce snapshots of full frontal, left and right profile, as well as 45-degree under chin shots of the above poses would be beneficial. Giving the ability for automated processing of depth map snapshots to be implemented as identified by Islam et al.'s (2018) approach. As it was hypothesised that a multi-faceted approach could be beneficial for a positive outcome. Additionally, acquiring a larger dataset with a larger demographic would help to resolve the overfitting and reduce generalisation issues to create a more robust predictive model for OSA.

6.0 Conclusion

The aim of this project was to find an accurate approach to detect OSA with the use of a deep neural network model to be able to give end to end classification from inputted 3D craniofacial scans. This model was supported with a web application to successfully upload a 3D image file that takes the unknown craniofacial scan, automates landmark placement, and classifies the craniofacial scan to predict the presence of OSA, or not, based on its trained model of craniofacial morphology. The final classification model achieved an accuracy of 100% over 10 folds, with a mean AUC of 77, after decreasing the batch size. Novel data tests gave an accuracy score of 60-70%, which performed surprisingly well as we had used a small dataset for training. Thus, this project shows a great example of a potential screening tool for OSA detection using 3D scans of OSA patients. Future improvements to the model would be beneficial, such as a larger and more diverse dataset, as well as implementing depth maps to work in conjunction with the landmark detection approach to allow more geodesic surface shape comparisons.

7.0 Acknowledgments

We would like to acknowledge the help and support of ECU’s School of Sciences, Dr Syed Mohammed Shamsul Islam and Hassan Mahmood. I (Jon Sveinbjornsson) acknowledge the use of ChatGPT (version 4o) as a coding assistant.

References

Chen, B., Cao, R., Song, D., Qiu, P., Liao, C., & Li, Y. (2024). Predicting obstructive sleep apnea hypopnea syndrome using three-dimensional optical devices: A systematic review. DIGITAL HEALTH. doi:10.1177/20552076241271749

Chen, Q., Liang, Z., Wang, Q., Ma, C., Lei, Y., Sanderson, J. E., Hu, X., Lin, W., Liu, H., Xie, F., Jiang, H., & Fang, F. (2023). Self-helped detection of obstructive sleep apnea based on automated facial recognition and machine learning. *Sleep & Breathing*, *27*(6), 2379–2388. https://doi.org/10.1007/s11325-023-02846-9

Collier, E., Nadjmi, N., Verbraecken, J. & Van de Casteele, E. (2023). Anthropometric 3D evaluation of the face in patients with sleep related breathing disorders. *Sleep Breathing Physiology and Disorders*. 27, 2209-2221. Retrieved from https://link.springer.com/article/10.1007/s11325-023-02827-y

Devansh. (2022). How does Batch Size impact your model learning. *Medium*. Retrieved from https://medium.com/geekculture/how-does-batch-size-impact-your-model-learning-2dd34d9fb1fa

Eastwood, P., Gilani, S. Z., McArdle, N., Hillman, D., Walsh, J., Maddison, K., Goonewardene, M. & Mian, A. (2020). Predicting sleep apnea from three-dimensional face photography. *Journal of Clinical Sleep Medicine.* 16(4), 493-502. https://doi.org/10.5664/jcsm.8246

Hanif, U., Leary, E. B., Schneider, L. D., Paulsen, R. R., Morse, A. M., Blackman, A., Schweitzer, P. K., Kushida, C. A., Liu, S. Y., Jennum, P., Sorensen, H. B. D., & Mignot, E. (2021). Estimation of Apnea-Hypopnea Index using Deep Learning on 3D Craniofacial Scans. *IEEE Journal of Biomedical and Health Informatics, 25*(11), 4185-4194. https://doi.org/10.1109/JBHI.2021.3078127

He, S., Su, H., Li, Y., Xu, W., Wang, X., & Han, D. (2022). Detecting obstructive sleep apnea by craniofacial image–based deep learning. *Sleep and Breathing*, *26*(4), 1885–1895. https://doi.org/10.1007/s11325-022-02571-9

Islam, S. M. S., Mahmood, H., Al-Jumaily A. A., & Claxton, S. (2018). Deep Learning of Facial Depth Maps for Obstructive Sleep Apnea Prediction. *2018 International Conference on Machine Learning and Data Engineering (iCMLDE)* (pp. 154-157). IEEE. doi:10.1109/iCMLDE.2018.00036.

JMP. (n.d.). *Correlation Coefficient*. JMP. https://www.jmp.com/en_au/statistics-knowledge-portal/what-is-correlation/correlation-coefficient.html#404f1893-ae56-43ed-b84c-f6c99f313eca

Lee, R., Chan, A., Grunstein, Cistulli, P. (2009). Craniofacial Phenotyping in Obstructive Sleep Apnea — A Novel Quantitative Photographic Approach. Sleep, 32(1), 37–45, https://doi.org/10.5665/sleep/32.1.37

Monna, F., Messaoud, R., Navarro, N., Baillieul, S., Snachez, L., Loiodice, C., Tamisier, R., Joyeux-Faure, M., Pepin, J. (2022). Machine learning and geometric morphometrics to predict obstructive sleep apnea from 3D craniofacial scans. Sleep Medicine, 95, 76-83. Retrieved from https://www.sciencedirect.com/science/article/pii/S1389945722001538

Ohmura, K., Suzuki, M., Soma, M., Yamazaki, S., Uchida, Y., Komiyama, K., Shirahata, T., Miyashita, T., Nagata, M., & Nakamura, H. (2022). Predicting the presence and severity of obstructive sleep apnea based on mandibular measurements using quantitative analysis of facial profiles via three-dimensional photogrammetry. *Respiratory Investigation*, *60*(2), 300–308. https://doi.org/10.1016/j.resinv.2021.10.002

Olsen, L. R. (2024). Multiple-k: Picking the number of folds for cross-validation. *Comprehensive R Archive Network*. Retrieved from https://cranr.project.org/web/packages/cvms/vignettes/picking_the_number_of_folds_for_cross_validation.html

Ozdemir, S. T., Ercan, I., Can, F. E., Ocakoglu, G., Cetinoglu, E. D. & Ursavas, A. (2019). Three-Dimensional Analysis of Craniofacial Shape in Obstructive Sleep Apnea Syndrome Using Geometric Morphometrics. *International Journal of Morphometrics, 37*(1), 338-343. Retrieved from https://pdfs.semanticscholar.org/bc1f.bc0bbfb814bf317f11a1c27e1255e36162b4.pdf

Paszke, P., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in PyTorch. *31st Conference on Neural Information Processing Systems*. NIPS 2017. Retrieved from https://openreview.net/pdf?id=BJJsrmfCZ

Paulsen, R.R., Juhl, K.A., Haspang, T.M., Hansen, T., Ganz, M., Einarsson, G. (2019). Multi-view Consensus CNN for 3D Facial Landmark Placement. In: Jawahar, C., Li, H., Mori, G., Schindler, K. (eds) Computer Vision – ACCV 2018. ACCV 2018. Lecture Notes in Computer Science, vol 11361. Springer, Cham. https://doi.org/10.1007/978-3-030-20887-5_44

Shetty, A., & Baptista Jardín, P. M. (2023). A patient's guide to obstructive sleep apnea syndrome. Springer. https://doi.org/10.1007/978-3-031-38264-2

TaghiBeyglou, B., Ng, B., Bagheri, F., Yadollahi, A. (2024). Estimating the risk of obstructive sleep apnea during wakefulness using facial images: A review. Biomedical Signal Processing and Control, 96(a), Retrieved from https://www.sciencedirect.com/science/article/pii/S1746809424005615