Sleep Apnea Analysis Through Multimodal Time Series Data

Sophia Tong - Eman Wong - Rafael Garcia

UNIVERSITY OF MICHIGAN

Background

Sleep apnea is a sleep disorder characterized by recurring interruptions in breathing during sleep, which can be dangerous. Globally, an estimated 936 million adults are affected by sleep apnea. In this project, our objective is to analyze sleep apnea datasets to gain insights into the factors influencing the occurrence of sleep apnea.

Research Question

RQ 1: Which factors are important in predicting the next sleep apnea episode?
RQ 2: Which factors are more correlated with each other in sleep data?

Dataset

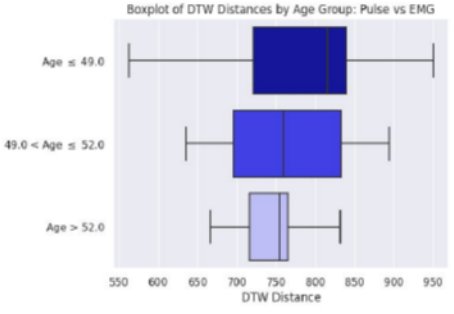
- St. Vincent's University Hospital/ University College Dublin Sleep Apnea database.
- 25 records for adult subjects with suspected sleep-disordered breathing
- Stationary data that shows each patient's health measurements.
- Time series data measured in their sleep, including EEG signals, airflow, movement, oxygen saturation, snoring and body position signals, in the polysomnograms, for each patient.
- Sleep stages, onset time, and duration of respiratory events, such as hypopnea or apnea, are annotated by experts.

E.D.A.

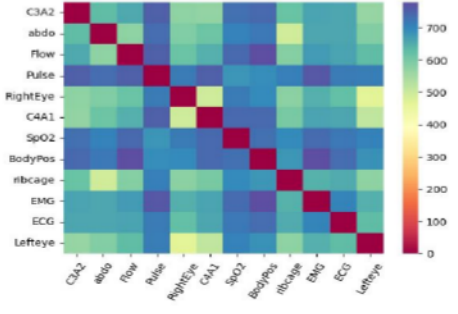
- Stationary data: all patients who suffer from sleep apnea have a BMI above the healthy range.
- Respiratory events time series: summarised duration for each type of abnormal breathing events→ hypopnea and apnea are positively correlated→ an episode of hypopnea might help predict the upcoming apnea.
- EEG signals time series: seasonal decomposition reflects different sleep stages, especially the REM stage→helpful to predict the apnea episodes.

Dynamic Time Warping

- Dynamic time warping was used to understand the similitude between patients' time series.
- Downsampling was required to make DTW computationally feasible.
- Factors such as BMI, Age, and Gender can affect the way a patient's ECG, Pulse, and EMG synchronize with each other.
- Some series, such as right-eye and left-eye wavelengths, showed to be strongly synchronized with each other.

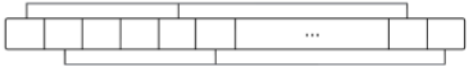


Feature Dissimilarity



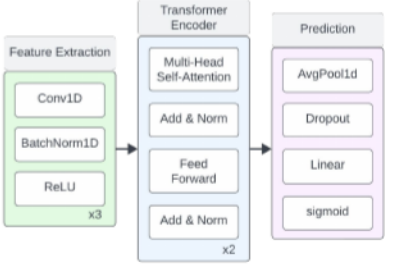
Data Pre-Processing & Augmentation

- Recordings divided into 40s segments with a siding step of 15s.
- If the next apnea event is within 30s, segment is considered as "prior".
- Segments are downsampled into 200 equally spaced points.
- There is a large class imbalance, with > 5 times the number of normal breathings segments vs "prior" segments.
- Data augmentation on "prior" class, by sliding downsample window.



Model Architecture

- 3 Convolutional Blocks
  Capture local & global features
- 2 Stacked Transformer Encoders
  Capture long-range dependencies
- 1 Prediction Block



Permutation Feature Importance

- Each feature in test set is permuted, and score is compared to original.
- Problematic features in PFI dropped: SpO2, Flow.
- Top 3 features: Pulse, Sleep Stage, Abdomen Movements.

Limitations & Future Work

- Limitations of permutation feature importance (correlated features, doesn't measure the intrinsic predictive value of a feature).
- Signal processing can be done to reduce the noise of inputs.
- Downsampling reduced the effectiveness of DTW.
- Only 25 individuals were studied.

Sleep Apnea Analysis Through Multimodal Time Series Data

Garcia, Rafael and **Tong, Xin** and **Wong, Eman**
University of Michigan

Abstract

Sleep apnea is a prevalent sleep disorder characterized by recurrent interruptions in breathing during sleep, posing significant health risks to affected individuals. In this study, we delve into the St. Vincent's University Hospital Sleep apnea database, which encompasses 25 records of adult subjects with suspected sleep-disordered breathing, featuring comprehensive time series measurements during sleep. We employ advanced time series analysis techniques as well as neural network to serve our primary objective: to identify the features strongly associated with sleep apnea and then to predict the onset time of the next sleep apnea episode. We discovered that the three features that contribute the most to predicting sleep apnea onset are pulse, sleep stage, and abdominal movements. There are limitations to the relatively small dataset, the noise in the input, as well as the permutation feature importance technique, but the study provides important guidance in predicting sleep apnea and even preventing it from happening. The code repository can be found here.

1 Introduction

As of 2023, 936 million adults around the world are estimated to suffer from sleep apnea. Research show that many people with obstructive sleep apnea develop high blood pressure, which can increase the risk of heart disease. The worse the obstructive sleep apnea, the greater the risk of coronary artery disease, heart attack, heart failure and stroke. The good news is that sleep apnea is treatable and preventable to a large extent. Hence, it is valuable if we identify the main factors contributing to sleep apnea with a data driven approach to assist the prevention and treatment of sleep apnea. In more extreme situations, when sleep apnea tends to stop a patient from breathing for an extended period of time, it is extremely helpful if we create an algorithm to predict the next onset of an episode of sleep apnea so that proper intervention may commence to save the life of a patient.

There are a multitude of studies conducting the detection of sleep apnea events, and more recently, Chen et al. proposed a CNN-transformer network showing promising improvements over classical RNN-based and CNN-RNN-based models. As such, we followed their data processing procedure and model architecture closely, in the hopes that it will be able to generalise well to our much wider dataset, comprising of 17 different signals as compared to the 5 used in their study. Furthermore, our dataset comprises of recordings of only 25 patients, in contrast to the 109 patients in their dataset. Thus, we note that model performance is not our main goal, but we instead want to use the proposed architecture to infer meaningful insights about the features present in our data.

2 Data

Our study uses the St.Vincent's University Hospital at University College Dublin Sleep Apnea database (McNicholas et al., 2004). This database contains 25 records for adult subjects with suspected sleep-disordered breathing. These subjects were selected randomly over a 6-month period, with 21 male and 4 female subjects, an age range of 28-68, a BMI range of 25.1-42.5 kg/m^2, and an AHI range of 1.7-90.9. This data is stored as a matrix, with each row representing the measurements for a single subject. The variables in the stationary data matrix are Height (cm), Weight (kg), Gender, PSG Start Time, PSG AHI, BMI, Age, Epworth Sleepiness Score, Study Duration (hr), and Sleep Efficiency (%).

Besides the combined stationary data, the database contains a rich amount of individual time series data for each subject, where each patient has their own matrices storing data measured in their sleep, which are EEG, EOG, EMG and ECG signals, airflow, movement, oxygen

saturation, snoring and body position signals in the polysomnograms. Each patient's EEG signal readings at each time stamp also come in three different channels. Specifically, the variable names in the dataset that we used for analysis and prediction are 'C3A2', 'abdo', 'Flow', 'Pulse', 'RightEye', 'C4A1', 'SpO2', 'Sound', 'BodyPos', 'ribcage', 'EMG', 'ECG', and 'Left-eye'. 'C3A2' and 'C4A1' are EEG readings, 'abdo' is abdomen movements measured by uncalibrated strain gauges, 'Flow' is oro-nasal airflow measured by thermistor, 'ribcage' is ribcage movements, 'SpO2' is oxygen saturation measured by finger pulse oximeter, and 'Sound' is snoring measured by tracheal microphone. We note that there was a 'Sum' variable present as well, which we suspect to be the sum of chest and abdomen movement signals. However, this variable is not documented well, and since the dataset includes signals for both chest and abdominal movement, we omit it from this study. In addition to the raw time series data, the sleep stages of each individual was annotated by a sleep technologist according to standard Rechtschaffen and Kales rules, where a 0 represents wakefulness, a 1 represents REM sleep, 2-5 represent sleep stages 1-4 respectively, and a 6 or a 7 represent artifacts and indeterminate recordings, which are essentially noisy labels. This data is represented as a sequence, with each value corresponding to a single sleep epoch, which is a 30 second period of time within this dataset. Lastly, the database contains annotations of onset time and duration of respiratory events, such as hypopnea or apnea, as well as whether they are central, obstructive, or mixed.

3 Exploratory Data Analysis

3.1 Initial Exploration

The time series data in our dataset are in .edf format. To read such a file format, we use the MNE package (Gramfort et al., 2013). However, we encountered issues while reading the data since the delimiters in our dataset for time stamps were incorrectly formatted, causing errors while reading in the time series data files. The EDF files not adhering to the correct date notation by using ':' instead of '.' as time separators. This formatting mistake appears to be common and has been reported to the MNE developers in this Github issue. We resolved the problem by modifying the source code of the MNE package (specifically, the edf.py file), implementing the solution proposed on the aforementioned Github issue page. After successfully reading the dataset, we chose to store the time series in a format other than EDF for ease of use. Initially, we tried storing the dataset in CSV format, but found that the average size of each patient's data was 933 MB. To save space, we decided to use a compressed file format that could be directly read by the Pandas software library, as it would be the primary tool for analysis (The pandas development team). We experimented with storing each patient's file as a pickle file (Pilgrim, 2009), which resulted in files averaging 400 MB, and as Parquet files (Vohra, 2016), which reduced the size to an average of 300 MB per patient. Given that the latter resulted in the least amount of memory usage after compression, we ultimately chose to store each patient's file in Parquet format.

As initial investigation of the stationary trends in the data revealed that all patients with sleep apnea have a BMI above the healthy range (with an upper bound of 24.9). This suggest that beyond the variables in time series, obesity might play an important role in causing sleep apnea, which aligns with the consensus in the field of sleep apnea research. The raw time series data for each patient's three-channel EEG signals does not show obvious trends, but when we plot the average EEG readings for each patient, as shown in the figure 1, we discover that the channel 1 and 2 have similar trends while channel 3 has the opposite trend compared to those two channels.

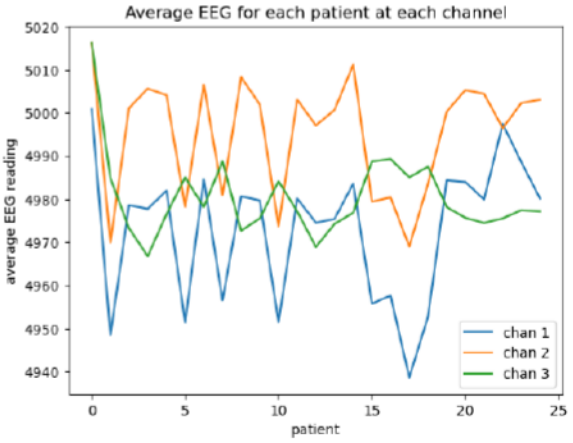


Figure 1: Average EEG of each patient from all three channels

This means that when we are trying to do dimension reduction in the future, we may one channel and discard the other two because they have indicative power of each other. From now on, we

conduct further exploratory data analysis on just channel 1. After doing seasonal-trend decomposition with LOESS on the EEG time series data for each patient, we discover that the REM sleep stage stands out on the Figure 2, because EEG readings are the highest during REM sleep stage among all sleep stages, almost as high as the readings from awake stage. Because during REM sleep there is an increase in the duration of apnea episodes according to research done by Oksenberg et al., this discovery could be helpful in future predictions of apnea, whenever REM happens or is about to happen.

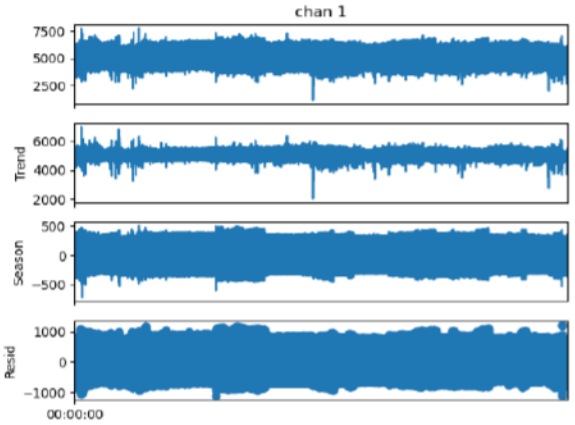


Figure 2: Seasonal Decomposition on Channel 1 EEG for a patient

To better analyze the trend across all patients for their respiratory events during over eight hours of sleep, we summarized the 25 time series data files for each patient into one large stationary data matrix. For example, each raw respiratory events file contains the time stamp of the onset of every single type of respiratory event during sleep, such as hypopnea-central, apnea-obstructive, etc. The data matrix also contains the duration of each recorded respiratory event, the oxygen desaturation quantity, snore arousal, indicator, and heart rate. We focus on the frequency and duration of each type of respiratory event for each patient, and discover that the Pearson correlation between total hypopnea count and total apnea count is about 0.53, which suggests a relatively high correlation between hypopnea and apnea. This may be helpful for future apnea onset prediction because hypopnea is a less severe respiratory event than apnea and it precedes apnea.

3.2 Dynamic Time Warping

Dynamic Time Warping (DTW) was used to quantify the relationship between each patient's time series. As described by Bankó and Abonyi, DTW can provide a reliable similarity measure for multivariate time series data and it can be complemented with techniques such as PCA. However, unlike Bankó and Abonyi's approach, this project aimed to use naive DTW to gain insights into the relationships among variables in the overall data set. Specifically, this was done to understand the strength of the relationship between a patient's series and BMI, age, and gender.

3.2.1 DTW Pre-processing

It was observed that, when examining each subject, the recording devices began measuring before being properly attached to the patients and continued recording for a couple of seconds after being removed from them. This was evident from unusual measurements at the beginning and end of each patient's time series, such as negative pulse readings. To rectify this, only recordings where the pulse exceeded 35 bpm were considered for DTW analysis, and it was assumed that the patients did not experience more extreme low heart rates (bradycardia) during the study. On average, each patient's time series comprised 3,195,381.76 observations, corresponding to signals recorded every 7.8ms over an average duration of 6.92 hours.

Constructing a synchronization matrix to calculate the DTW distance for two such time series would require approximately 82 TB of memory, which was infeasible due to resource constraints. Therefore, down-sampling was performed. Observations were grouped into 20-second windows, and the average of the observations in each window was chosen to represent the entire window. In cases where NaN values emerged in a window due to missing data, a backfill approach was used, where missing values were replaced with the value from the nearest preceding window without missing values. After down-sampling each of the series, Scikit-learn's StandardScaler was used to standardize each time series (Pedregosa et al., 2011). The purpose of this standardization was to obtain unitless DTW distances between each of the compared time series.

3.2.2 DTW Dissimilarity of Features

Once the series of each patient were processed, for each patient, their series were compared through

DTW producing a dissimilarity matrix. The following figure shows the average dissimilarity matrix of features across all patients.

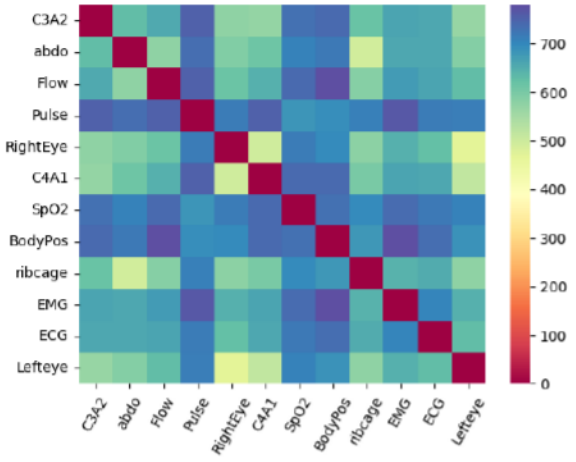


Figure 3: Average dissimilarity matrix of features

The matrix suggests that certain features exhibit similarities in terms of their DTW distances. For instance, abdominal movements (abdo) and ribcage movements (ribcage) demonstrate a relatively small DTW distance compared to other pairs in the matrix, indicating a potential connection between them. A similar observation is made for left-eye electrooculography (Lefteye) and right-eye electrooculography (RightEye), as well as for both electrooculography measures (left and right) when compared with the electroencephalogram measures (C3-A2, C4-A1).

3.2.3 DTW Variation Across Groups

The dissimilarity matrix from the previous section indicated that pulse was not as closely related to the other features, with the latter showing more overall similarity amongst themselves. To investigate pulse further, this section focuses on analyzing whether dividing patients into different BMI, age, and gender groups and then examining their group pulse DTW distances results in lower or higher median DTW distances across features.

Effect of Age in DTW distances

To investigate the impact of age on the DTW distances of patients' time series, subjects were categorized into three age groups. These groups were determined based on the $1/3$, $2/3$, and 1 quantiles of the subjects' ages. Subjects at or below the $1/3$ age quantile were placed in the first group. Those who were above the $1/3$ quantile but below the $2/3$ quantile were assigned to the second group, and the remaining subjects were assigned to the third group.

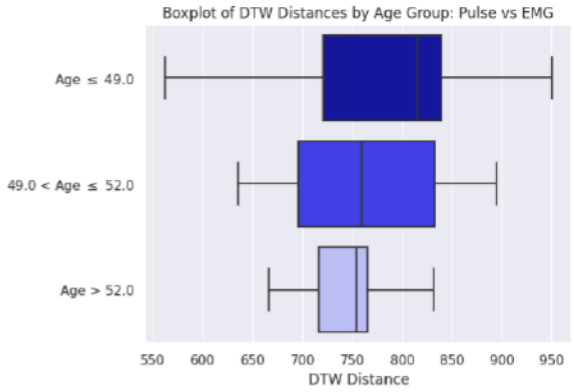


Figure 4: Pulse vs EMG Boxplots by Age

The results of this grouping showed that grouping by age led to a median decrease in both DTW distances and DTW variability as age increased, particularly in terms of EMG (Electromyography). This suggests that pulse tends to relate more with electromyography readings as age increases. Other variables showed similar trends in relation to pulse and age with minor variations.

Effect of BMI in DTW distances

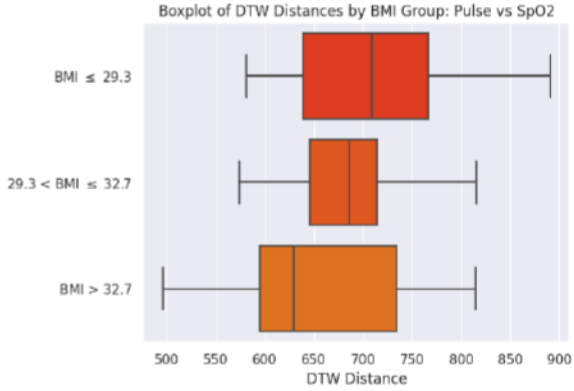


Figure 5: Pulse vs SpO2 Boxplots by BMI

To investigate the effect of BMI on the DTW (Dynamic Time Warping) distances of patients' time series, a similar approach to that described in the previous section was taken. Subjects were categorized into one of three BMI groups. These groups were based on the $1/3$, $2/3$, and 1 quantiles of the BMIs of all subjects. After grouping by BMI, it was found that SpO2 had a median decrease as BMI increased. This suggest a relationship between pulse and SpO2 as people are more overweight.

It is important to note and consider that all adults in the study had a BMI higher than the recommended healthy range and are considered overweight (BMI ≥ 25), as described by Gallagher et al..

The relationship between Pulse and SpO2 in terms of DTW distances for adults of normal weight remains unclear.

Effect of Gender in DTW distances

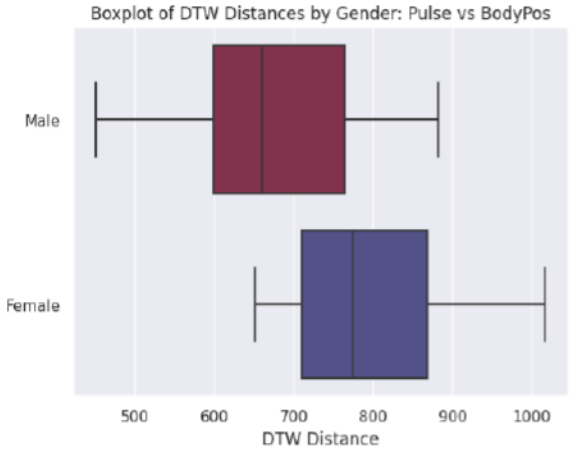


Figure 6: Pulse vs Body Position Boxplots by Gender

As described by MacWilliam, body position and various factors can affect a person's pulse. Following gender-based grouping, the results appeared consistent with MacWilliam's study, indicating that the DTW distances between body position signals and pulse showed slight variations by gender. Specifically, for males, pulse was found to be more closely related to body position compared to females, in terms of median DTW distance. However, when considering pulse, no significant dissimilarities in the other variables were observed after grouping by gender. This suggests that while body position has a differentiated impact on pulse between genders, other factors do not exhibit a noticeable gender-based disparity in their relationship with pulse.

4 Modelling

4.1 Data Pre-processing & Augmentation

The portions of recordings where the patient was still awake were removed, and the trimmed recordings were divided into 40 second segments with a sliding step of 15 seconds. The segments with a respiratory event occurring for more than 16 seconds are considered as disordered and are filtered out. Segments which have end less than 30 seconds from the onset of a respiratory event are considered "prior" segments, and serve as our positive class for the classification task. This preprocessing follows the procedure outlined by Chen et al., with some modifications to how we create labels. The proposed method labels the 3 windows leading up to a respiratory event as "prior", which we found to be inconsistent when experimenting with different segment lengths as well as sliding step size. A cut-off value of 30 seconds from the next event was chosen, as it coincides with the findings from Waxman et al., where it was found that predicting events up to 30 seconds into the future yielded the best results.

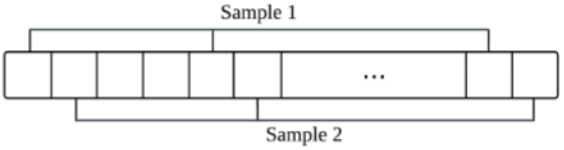


Figure 7: Combining downsampling with a rolling window

As compared to the data used by Chen et al., we have a much higher sampling rate of 128 Hz, and at a input segment length of 30 seconds, the resultant 3840 observations will cause extremely long training times. When attempting to train the model using all observations, we also encountered issues with memory limitations. Thus, the data was downsampled by selecting 200 equally spaced points within each window. Additionally, in order to combat the large class imbalance between the normal and "prior" segments, another rolling window was applied to "prior" segments for data augmentation. The combination of downsampling along with a rolling window allows us to generate multiple non-overlapping samples from the same segment, as illustrated in Figure 7. We find that for our dataset, there can be between 5 to 10 times the number of normal segments than "prior" segments, depending on our segment lengths, the time to the next event which we use for labelling, as well as the randomness when splitting the data into training and test sets by patient. Thus we use the lower bound and generate 5 samples per "prior" segment, and an equal number of normal breathing segments are randomly selected, allowing for an equal number of observations in both positive and negative classes.

4.2 Model Architecture

We follow the model architecture and parameters proposed by Chen et al., in which the model comprises of 3 modules: feature extraction, transformer encoder, and prediction. The feature extraction module is made up of 3 sequential convolutional blocks. Each convolutional block consists of a 1D-CNN layer, a batch normalisation layer,

and a ReLU activation layer. The first block's 1D-CNN layer uses a kernel size of 3 with a padding of 1, while the next two blocks use a kernel size of 29 with a padding of 14. This is in hopes that the initial convolutional block is able to capture local characteristics while the following blocks are able to capture characteristics on a macro scale. The transformer encoder module uses a stack of 2 transformer encoders, each consisting of a multi-head self-attention layer and a feed forward layer, with layer normalisation after each layer (Vaswani et al., 2017). Lastly, the outputs are fed into the prediction module, containing an average pooling layer with a kernel size of 200, followed by a dropout layer with a dropout rate of 0.5, and a linear output layer. These outputs are then mapped to output probabilities using a sigmoid function, and binary cross entropy is used as the loss function.

5 Model Evaluation

5.1 Model Fit

The model was initially fit with a batch size of 64, using the Adam optimiser with a learning rate of 1×10^{-4}, however, this caused the validation loss to explode from epoch 0. This was addressed by lowering the learning rate by several orders of magnitude, as well as by varying batch sizes. We observe the training and validation losses for some of these initial fits in figures 8 and 9 respectively, with "bad" runs being omitted. In these initial fits, we see that despite only altering the batch size and learning rate, the training and validation loss at epoch 0 vary greatly. This indicates that some form of weight initialisation would likely improve the stability of training as well as allow for the model to converge quicker. A rudimentary implementation of Xavier weight initialisation is applied to the convolutional and linear layers in the model; however we did not observe any meaningful effect, and a much more sophisticated implementation will be required, which is outside the scope of this project.

From our initial testing, we found that a batch size of 16, with a learning rate of 5×10^{-8}, trained for 25 epochs, provided the best results in terms of training stability, while maintaining a reasonable training time of under 30 minutes.

The model was implemented using PyTorch Lightning (Falcon et al., 2020), a lightweight PyTorch wrapper. The model fitting was performed on an Apple M2 Max Macbook Pro with 32GB of memory, utilising the MPS backend of PyTorch (Paszke et al., 2019), allowing for training to be done on GPU.

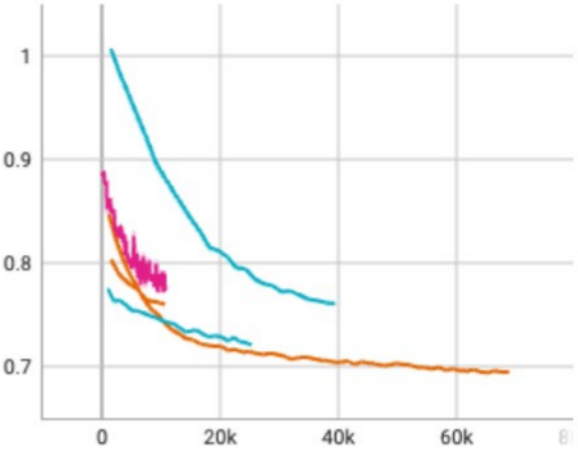


Figure 8: Training loss on initial fits

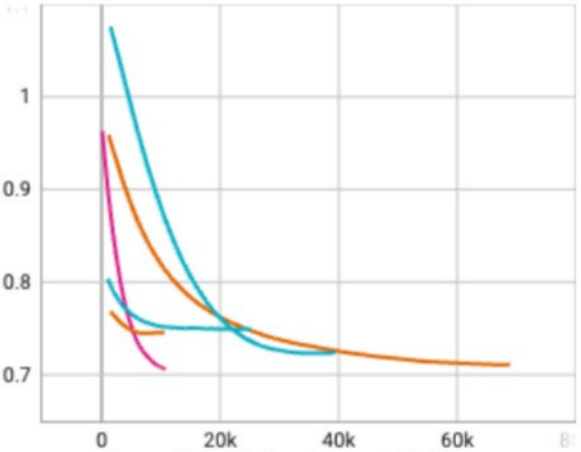


Figure 9: Validation loss on initial fits

5.2 Permutation Feature Importance

A main point of contention in deep neural networks is that they are black-box models, lacking explainability (von Eschenbach, 2021). In order to combat this, we use a global model-agnostic method in order to determine the key features which are driving our model—Permutation Feature Importance (PFI). PFI is calculated by training the model and obtaining its error score on the test set. Each feature in the test set is then randomly permuted, predictions are generated, and new error scores with the permuted feature are calculated. These scores can then be compared with our original score to give us a feature's importance; specifically, we calculate the quotient: $error_{permute}/error_{original}$ (Molnar, 2022). In our experiment, we repeat this random process 5 times for each feature, since PFI is susceptible to sampling variability.

The main advantage of PFI is that it allows us

to measure the importance of a feature without the need to retrain the model, unlike certain algorithms which iteratively remove low importance features and retrain the model. This is crucial for our use case, since both the data pre-processing as well as model training are very computationally intensive, with a total wall time of over an hour, even when using parallelisation.

The error metric that we choose for this process is 1-AUC (1 minus the area under the Receiver Operating Characteristic (ROC) curve). The metrics used for model evaluation by Chen et al. are accuracy, sensitivity and F1-score. We choose instead to use 1-AUC, as the model outputs probabilities for a positive class, and we are able to compute the 1-AUC score without explicitly setting an arbitrary threshold for the predicted probabilities.

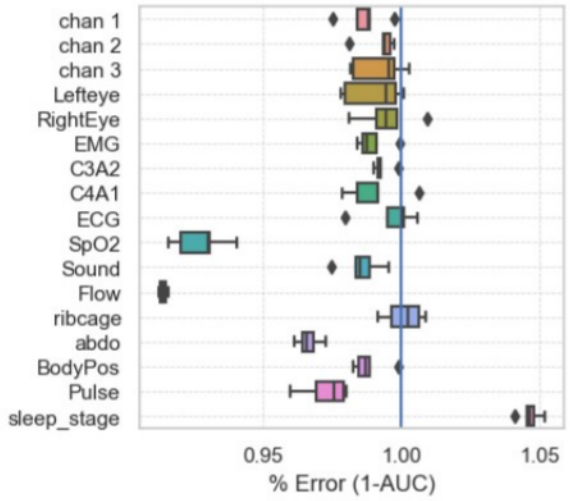


Figure 10: Permutation Feature Importance with All Features

We calculate the PFI for our model, as shown in Figure 10. From the plot, we see that when permuting the features SpO2 and Flow, we have a significant decrease in error as compared to the other features, with average error rates of 91.4% and 93.1% respectively. We also note that despite the majority of other features also showing a negative feature importance, they do not necessarily have poor intrinsic predictive value, as PFI merely indicates the importance of a feature within a specific model (Buitinck et al., 2013).

The large decrease in error from permuting SpO2 and Flow indicate that these variables are impacting the quality of the fit of our model, possibly causing the model to overfit, resulting in poor generalisability. These two features are removed, the model is retrained, and we calculate the PFI again, as shown in Figure 11. The retrained model shows much better performance, with an accuracy of 50.8% and an AUC of 50.4%, as compared to the previous model's accuracy of 39.2% and AUC of 37.6%.

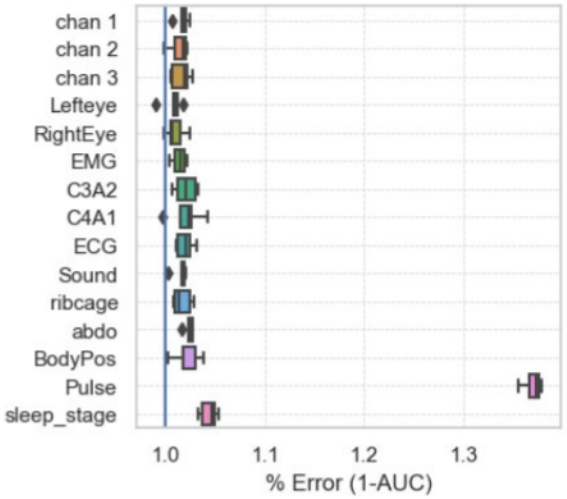


Figure 11: Permutation Feature Importance with SpO2 and Flow removed

We see that the importance of all features have improved drastically, with all mean percentage errors being greater than 1. We see that Pulse has an extremely high percentage error, with an average of 137%, indicating that this feature plays a very large role in driving the model's predictions. The next highest percentage error is obtained when permuting sleep stage (105%), followed by abdominal movements (103%).

6 Discussion

Dynamic Time Warping successfully identified relationships between abdominal and rib cage readings, as well as significant relationships between the electrooculography readings of each eye. Furthermore, DTW also showed a connection between electrooculography readings and the electroencephalogram measures provided. Notably, DTW distances in signals were found to vary based on features such as gender, BMI, and age. This finding opens avenues for future research, specifically investigating the potential of using BMI, gender, and age as predictors for the aforementioned patient signals. Any such study should aim to include a more diverse subject pool, encompassing various conditions. The current study's limitation, stemming from its focus on subjects with sleep apnea who were overweight, results in a sample that is not representative of the global population. Expanding the scope of future research to include a

broader range of subjects would enhance the applicability and relevance of the findings.

The metric of permutation feature importance has several drawbacks. As mentioned in section 5.2, PFI is unable to determine the intrinsic predictive value of a feature, thus it is difficult to ascertain any notion of causality using PFI. Additionally, PFI results are not always reliable when features are correlated. Permuting a feature which is highly correlated to another can cause instances in the data which are infeasible in real life, which we use to evaluate the model. Additionally, the inclusion of a correlated feature could decrease the importance of the associated feature by splitting the importance between them (Molnar, 2022). Hence, PFI is not perfect as a metric in our use case, where there exist highly correlated features, such as the 3 EEG channels present in our dataset. It is likely that the reliability and robustness of our chosen importance metric can be improved in the future using a more robust feature selection process, or by creating a composite feature.

As for our model, we believe that one contributing factor to the less than ideal performance is simply the size of our dataset. In this study, we closely follow the pre-processing procedure as well as the model architecture of Chen et al.. However, the dataset used in their study contained 109 patients, and only used 5 ECG signals as opposed to the 17 input signals in our dataset. With our much wider dataset, it is likely that we require a much larger volume of data to ensure that our model generalises well, and does not overfit to the noise. Ideally, we would also have been able to use a robust feature selection strategy as well as hyperparameter tuning using cross validation. However, this was infeasible within the time frame of the project simply due to the computational cost of the data pre-processing and augmentation, as well as model training.

In this study, we conducted a comprehensive analysis of time series data obtained from individuals with suspected sleep-disordered breathing, focusing on the prediction of sleep apnea onset. Our findings reveal compelling insights into the temporal dynamics of sleep-related respiratory events and provide a practical approach for real-time prediction. Leveraging transformer-based architectures, we successfully trained a predictive model that demonstrates promising potential in forecasting sleep apnea episodes. Our model shows great importance of pulse, sleep stage, and abdomen movements in predicting if the an episode of sleep apnea is about to happen. This finding corresponds with our exploratory data analysis quite decently, where the EDA unveils REM sleep stage in EEG readings, as well as when the pulse variable has a closer DTW with multiple other variables compared to some other single variables. The discovery of the strong association between abdomen movements with apnea aligns with state-of-the-art medical research, where breathing effort and airway blockage from apnea result in out-of-phase movement of the chest and abdomen Akbarian et al..

Although our model uncovers contributing features to effective apnea prediction, there are unexpected results in juxtaposition with current medical research. For example, body position during sleep is proved to influence the onset of apnea (supine position triggers onset much more than side position), our model did not rate body position as one of the top three features. Oxygen saturation is another factor that directly indicates the happening of apnea, yet this feature plays a destructive role in the prediction process of our model. Despite deserving more validation on robustness, our study contributes not only to the understanding of sleep apnea dynamics but also introduces a practical tool for early intervention and personalized treatment strategies. While the predictive model exhibits plausible efficacy, we acknowledge limitations in the dataset size and diversity, and we advocate for further validation on larger cohorts.

7 Conclusion

In conclusion, our research represents a significant step forward in the understanding and prediction of sleep apnea using in-depth time series analysis. The positive outcomes of our study demonstrate the potential of leveraging advanced computational techniques, particularly transformer-based models, for predicting the onset of sleep apnea episodes. By uncovering the intricate relationship between health measurements and apnea events, we have established a foundation for early intervention strategies. The successful training of our predictive model highlights the feasibility of integrating machine learning approaches into the realm of sleep medicine. Future research efforts could explore the integration of additional physiological signals and consider longitudinal studies

to enhance the robustness of predictive capabilities. Overall, our work opens avenues for advancements in sleep disorder management, emphasizing the potential for timely and targeted interventions based on predictive analytics derived from comprehensive time series analysis and neural network prediction models.

References

Sina Akbarian, Nasim Ghahjaverestan, Azadeh Yadollahi, and Babak Taati. 2019. Distinguishing obstructive versus central apneas in infrared video of sleep using deep learning: Validation study (preprint).

Zoltán Bankó and János Abonyi. 2012. Correlation based dynamic time warping of multivariate time series. *Expert Systems with Applications*, 39(17):12814–12823.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Yuhang Chen, Shuchen Yang, Huan Li, Lirong Wang, and Bidou Wang. 2023. Prediction of sleep apnea events using a cnn–transformer network and contactless breathing vibration signals. *Bioengineering*, 10(7):746.

William Falcon, Jirka Borovec, Adrian Wälchli, Nic Eggert, Justus Schock, Jeremy Jordan, Nicki Skafte, Ir1dXD, Vadim Bereznyuk, Ethan Harris, Tullie Murrell, Peter Yu, Sebastian Præsius, Travis Addair, Jacob Zhong, Dmitry Lipin, So Uchida, Shreyas Bapat, Hendrik Schröter, Boris Dayma, Alexey Karnachev, Akshay Kulkarni, Shunta Komatsu, Martin.B, Jean-Baptiste SCHIRATTI, Hadrien Mary, Donal Byrne, Cristobal Eyzaguirre, Cinjon, and Anton Bakhtin. 2020. Pytorchlightning/pytorch-lightning: 0.7.6 release.

Dympna Gallagher, Steven B Heymsfield, Moonseong Heo, Susan A Jebb, Peter R Murgatroyd, and Yoichi Sakamoto. 2000. Healthy percentage body fat ranges: an approach for developing guidelines based on body mass index. *The American journal of clinical nutrition*, 72(3):694–701.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti S. Hämäläinen. 2013. MEG and EEG Data Analysis with MNE-Python. *Frontiers in Neuroscience*, 7(267):1–13.

J. A. MacWilliam. 1933. Postural effects on heart-rate and blood-pressure. *Quarterly Journal of Experimental Physiology*, 23(1):1–33.

Walter McNicholas, Liam Doherty, Silke Ryan, John Garvey, Patricia Boyle, and Eric Chua. 2004. St. vincent's university hospital / university college dublin sleep apnea database.

Christoph Molnar. 2022. *Interpretable Machine Learning*, 2 edition.

Arie Oksenberg, Elena Arons, Khitam Nasser, Tatiana Vander, and Henryk Radwan. 2010. Rem-related obstructive sleep apnea: The effect of body position. *Journal of Clinical Sleep Medicine*, 06(04):343–348.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Mark Pilgrim. 2009. *Serializing Python Objects*, pages 205–223. Apress, Berkeley, CA.

The pandas development team. pandas-dev/pandas: Pandas.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Deepak Vohra. 2016. *Apache Parquet*, pages 325–335. Apress, Berkeley, CA.

Warren J. von Eschenbach. 2021. Transparency and the black box problem: Why we do not trust ai. *Philosophy &; Technology*, 34(4):1607–1622.

Jonathan A. Waxman, Daniel Graupe, and David W. Carley. 2010. Automated prediction of apnea and hypopnea, using a lamstar artificial neural network. *American Journal of Respiratory and Critical Care Medicine*, 181(7):727–733.