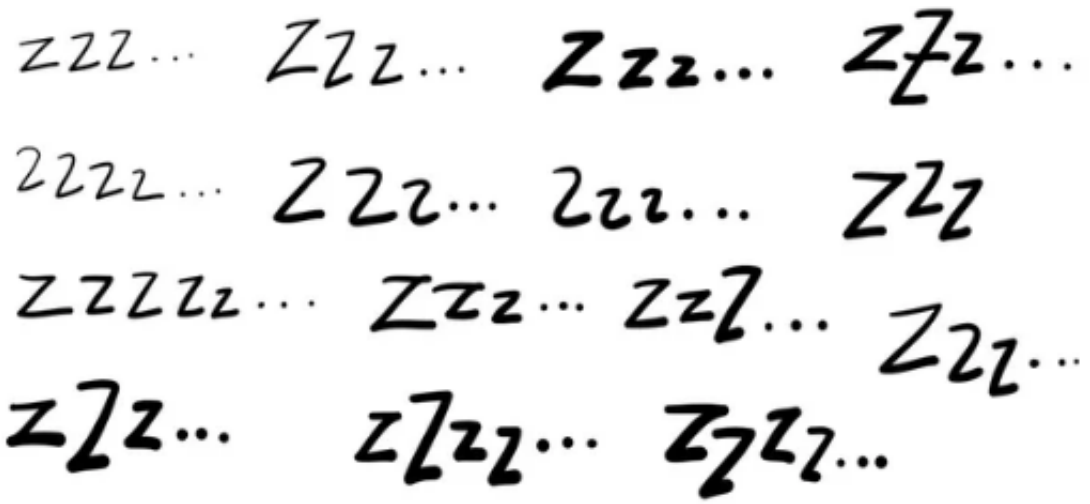
Deep Learning for Automatic Classification of Sleep Stages

Group 3

`CM50265 Machine Learning 2 (ML2), Coursework 2`



ID	Contribution
jg852	100%
sdlw20	100%
tvhf20	100%

Do all the members agree with the above contributions: yes

Shared files for code & model: please clone the repo from: https://github.com/scottwellington/sleepEEG/

`.py` and `.ipynb` scripts are supplied with the submission of this report. The two scripts contain the same code (`yaml` is employed in the `.ipynb` to handle `argv` inputs to `argparse` for the `.py` script equivalent.)

Please read the `README` in the repo for more info re: our pre-supplied test data.

Link to dataset: https://www.physionet.org/content/sleep-edfx/1.0.0/

Contents

List of Figures

List of Tables

Abbreviations

AASM American Academy of Sleep Medicine. 1

CNN convolutional neural network. iii, 4, 5, 7, 9, 13

EEG electroencephalogram. ii–iv, 1, 3–7, 11, 15–17
EMG electromyogram. ii, iii, 1, 3, 6, 11, 16, 17
EOG electro-oculogram. ii, iii, 1, 3, 6, 11, 16

FBCSP filter bank common spatial patterns. 4
FCN Fully Convolutional Network. 4, 13

GAN Generative Adverserial Network. 15

LSTM Long Short-Term Network. 4, 13

MFCCs Mel-frequency cepstral coefficients. 15

OSA obstructive sleep apnea. 1

PSD power spectral density. iii, 4, 5, 13
PSG polysomnogram. iii, 1, 4, 6, 11

REM rapid eye movement. iii, 1, 3, 8–11, 18, 19
RF Random Forest. iii, 5, 13
ROC Receiver Operating Characteristic. iii, 13

SWS slow-wave sleep. iii, 1, 8–11, 18, 19

1 Problem

Lack of sleep has serious consequences for health and productivity. Someone who regularly sleeps less than 7 hours a night shortens their age-related mortality (Heslop et al., 2002) and impairs their immune system (Irwin et al., 1996). They quadruple their risk of heart attack (Ohtsu et al., 2013); and double their chance of cancers like lung and ovarian cancer (Walker, 2017). Sleep deprivation causes 20% of motorway accidents (Horne and Reyner, 1995) and increases the risk of workplace accidents by 176% (Åkerstedt et al., 2002). Lack of sleep has a negative impact on productivity – for example, in one study, medical interns who did not get a good night's sleep made 36% more serious errors than those who did (Landrigan et al., 2004).

Polysomnography is the gold standard for classifying sleep stages to diagnose and therefore treat sleep disorders. A polysomnogram (PSG) is a multi-parametric assessment based on biophysiological changes while the participant is asleep, typically measuring brain activity with an electroencephalogram (EEG); eye movement with electro-oculogram (EOG); and muscle tension with a chin electromyogram (EMG) in 30-second 'epochs'. Experts visually categorise sleep into stages using standards set by American Academy of Sleep Medicine (Iber, 2007): wake (W); light sleep (S1/S2); slow-wave sleep (SWS) or deep sleep (S3/S4); and rapid eye movement (REM). This classification differs from the original classification of Rechtschaffen (1968) in that it now merges S3 and S4 to denote SWS.

There is growing demand for PSG sleep stage analysis, which is highly specialist and time consuming. Demand is growing, given growing awareness of the impact of sleep disorders on health – e.g. obstructive sleep apnea (OSA) now impacts on up to a third of the population (Tufik et al., 2010) and is linked to cardiovascular issues such as strokes (Redline et al., 2010). Table 1 summarises the specialists' methodology; see Figures 1, 2, 3, 4, 5 and 6 for visual examples; and Appendices A and B for measurement sites and EEG frequency bands. Visual classification takes 2-4 hours for one night of sleep (Ronzhina et al., 2012) – and there is the risk of inconsistencies between diagnoses.

Stage	Description	% of sleep	EEG (brain)	EOG (eyes)	EMG (muscle)
W	Awake	N/A	Low voltage, alpha >50%	-	High (>30 Hz)
S1	Light sleep: dozing off, twitching	5%	Theta waves 4-7 Hz114; alpha <50% – Figure 1	Slow eye movements – Figure 2	Elevated; <W.
S2	Light sleep	50%	Theta waves 4-7 Hz with 'sleep spindles' (≥0.5 sec 12-14 Hz waves – Figure 3) and/or 'K-complexes' (≥0.5 sec sharp negative then smooth positive – Figure 4).	-	Elevated; < W.
SWS	Deep sleep. Muscle tone, pulse and breathing slow	20%	Delta waves 0.5-3 Hz 20-50% of total; amplitude >75 mV from peak to peak - Figure 5; K-complexes and spindles.	-	<S2 to as low as REM
REM	Eyes move quickly; body paralysed; high brain activity	20-25%	Low-amplitude, mixed-frequency. Saw-tooth 2-6 Hz waves. No K-complexes or spindles.	Episodic REMs (<500 ms) - Figure 6	Lowest.

Table 1: Sleep stages – see Appendix C.2 for more details; source: Iber (2007), Vaughn and Giallanza (2008)

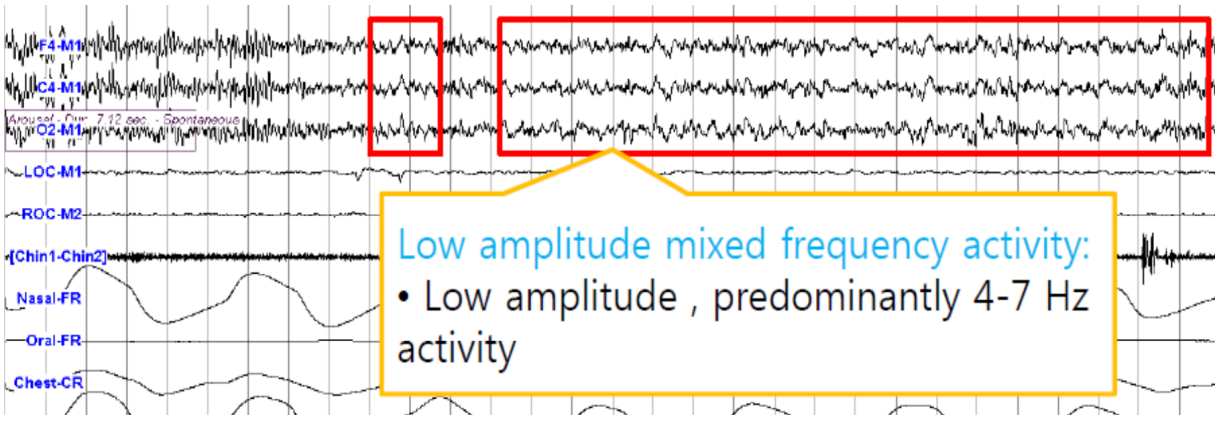


Figure 1: S1 EEG activity from F4, C4 and O2 – see Figure 18 for sites00, source: Iber (2007)

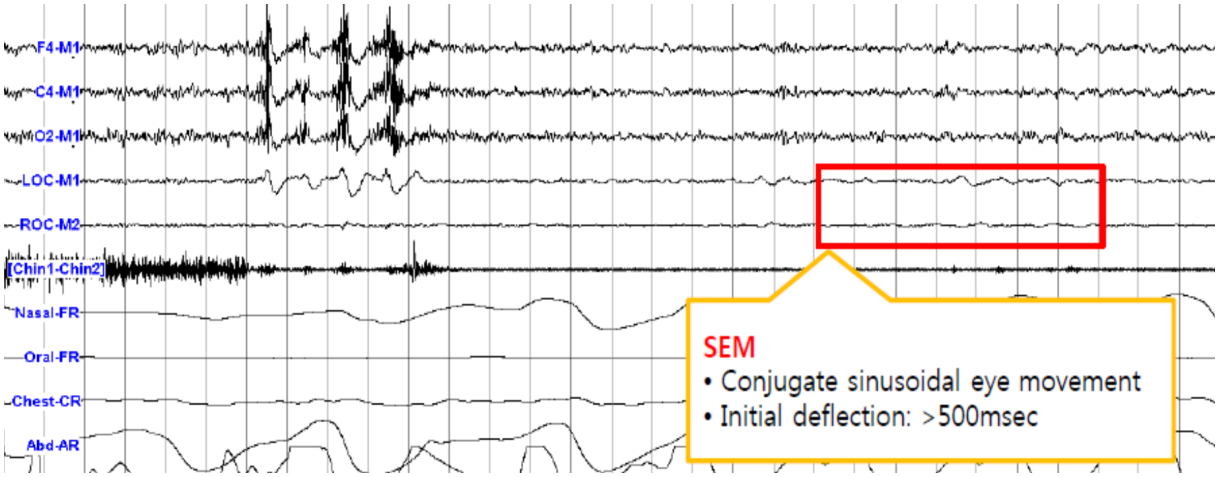


Figure 2: S1 EOG activity: slow eye movement, source: Iber (2007)

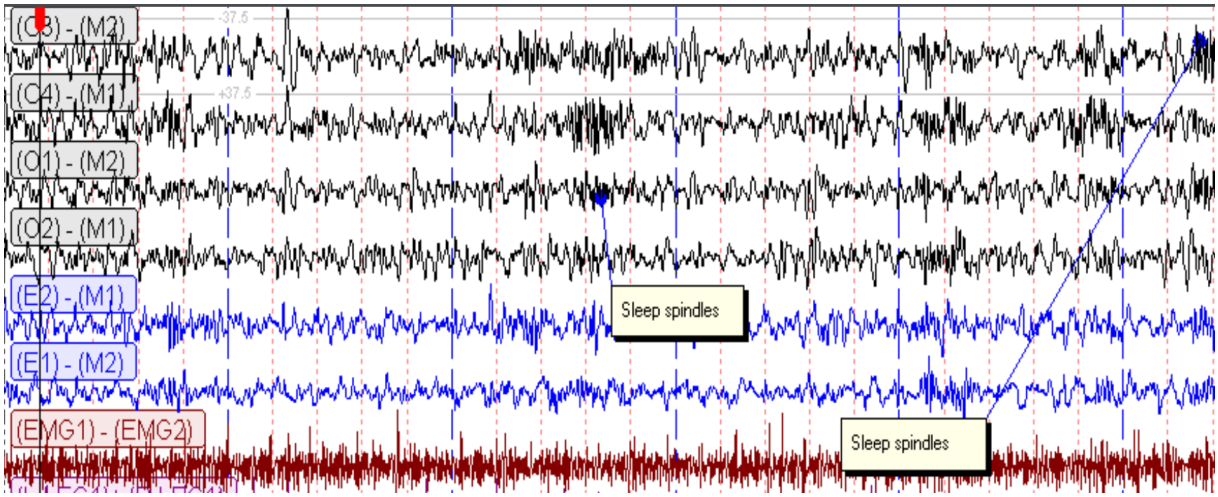


Figure 3: Sleep spindles, EEG in black, source: Iber (2007)

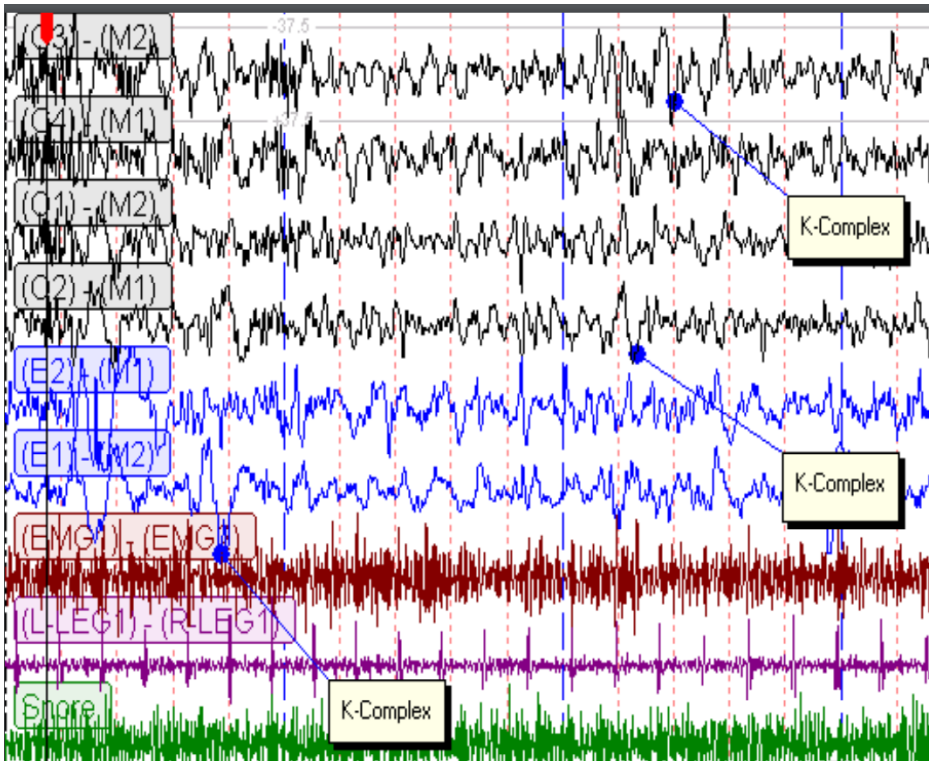


Figure 4: K-complexes, EEG in black, source: Iber (2007)

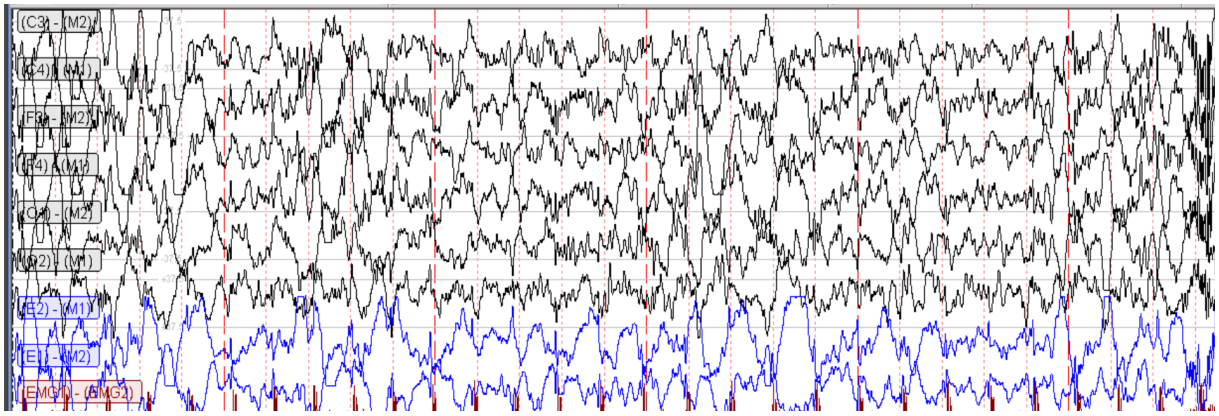


Figure 5: Slow delta waves (EEG in black), source: Iber (2007)

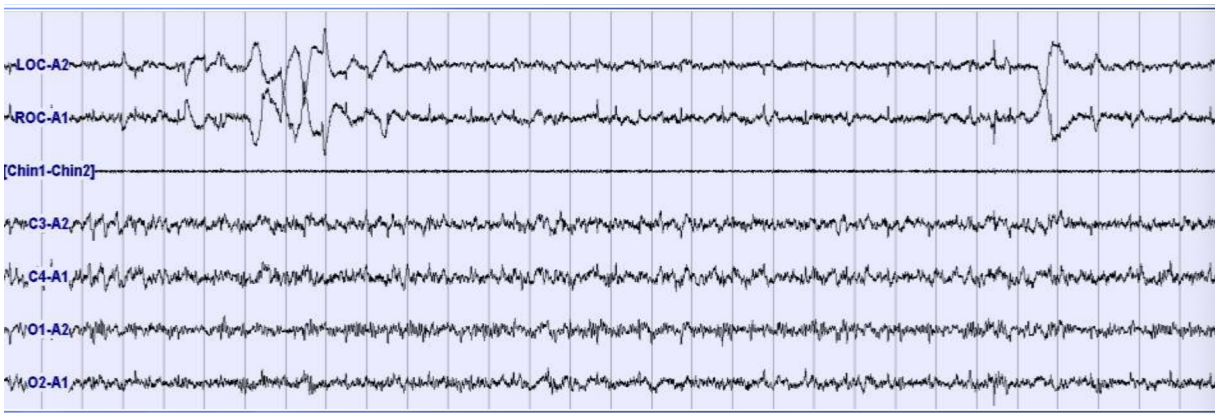


Figure 6: Episodic REM in EOG at the top; relatively flat low level activity at EMG at chin; EEG from C3, C3, O1 and O2 at the bottom with some saw tooth waves early on, source: Iber (2007)

2 Solution

Classification of PSG data could be done by a convolutional neural network (CNN) and validated by experts prior to use, as they build trust and learn how best to use the support.

CNNs have been used on EEG data. Lun et al. (2020) used 4 seconds of EEG data from the Physionet database to categorise motor imaginations tasks: left fist, right fist, both fists, and both feet. The model was a 5-block CNN network with 4 max pooling layers to reduce dimensionality, and a fully-connected layer for classification. Schirrmeister et al. (2017) used CNNs in combination with a FBCSP algorithm to decode EEG.

CNNs are in theory well-suited to classifying PSG time series data. CNNs automatically and adaptively learn spatial hierarchies of features through back propagation by using building blocks like convolution layers. While CNNs were originally developed for 2D image data, they are used to classify time series data – see Figure 7. Convolutions apply and slide a 'filter' over the data, performing a generic non-linear transformation. Images have 2 dimensions (width, height); time series have 1 (time).

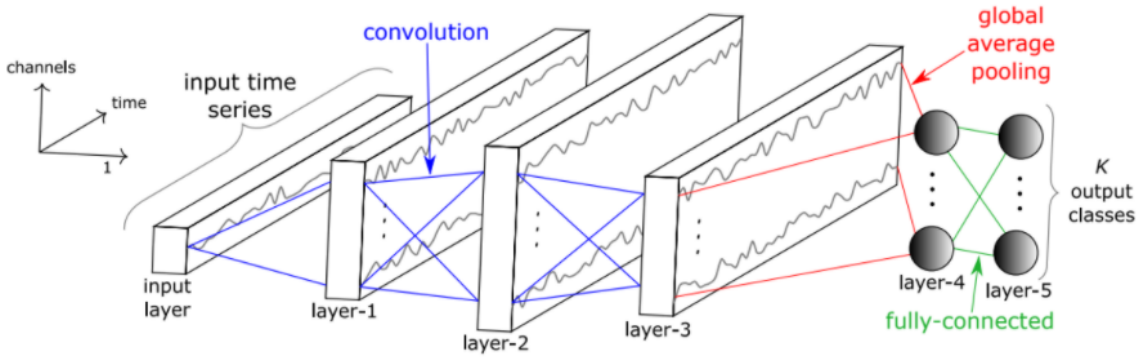


Figure 7: CNN applied to time series data; source Stylianou (2021)

A CNN will be applied to time series data and 3 additional CNN-variant models also will be trained. Filters will be applied to the PSG data after sliding windowed batching and sub-band frequency decomposition. The output will be a softmax layer with a set of probabilities that the input window is part of a particular class. The softmax function is a generalisation of the logistic function to multiple dimensions, where the probabilities of the classes in the output sum to one. The three additional models include: Fully Convolutional Network (FCN), 3-Dimensional CNN and Long Short-Term Network (LSTM) CNN.

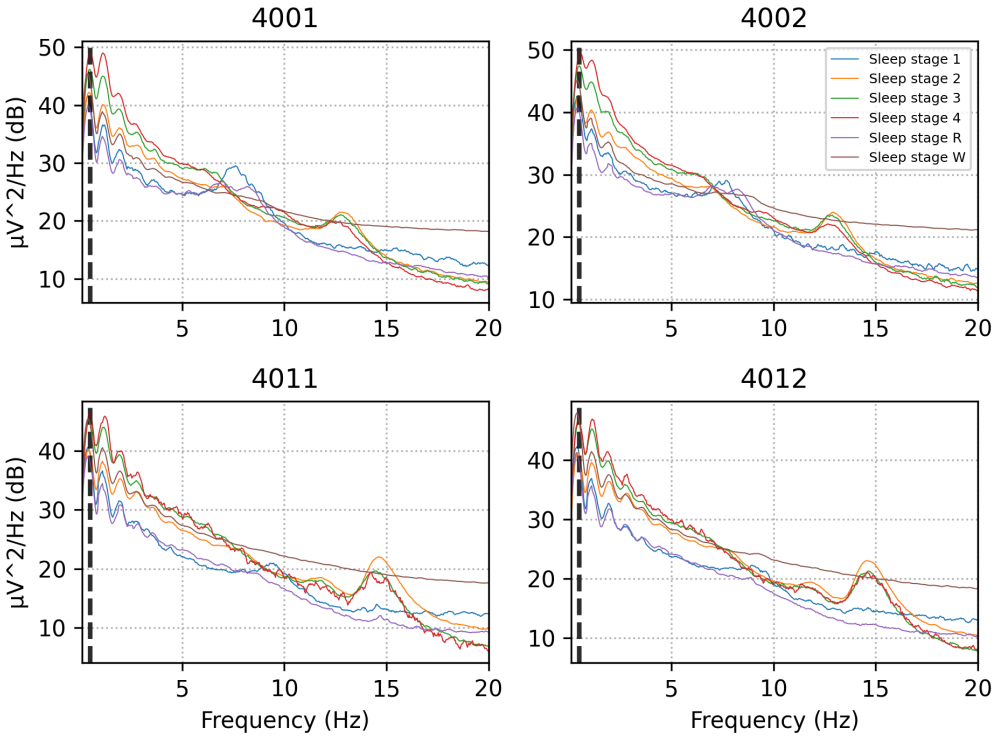


Figure 8: power spectral density (PSD) plots for participants 4001, 4002, 4011 and 4012

Figure 8 illustrates our proof-of-concept: following a power spectral density (PSD) transform, it is possible to determine the unique frequency (and harmonics) properties of the different sleep stages from the EEG signal. Since a PSD transform decomposes one of our 30-second epochs into only 2 values per channel, we decompose our EEG signal into frequency subbands only (using a band-pass Butterworth filter). This is a common preprocessing procedure for EEG data, e.g Gao et al. (2020), and allows the CNN model to extract predictive features from latent feature maps of data that have the same number of samples as the raw EEG signal.

We compare the CNN to a baseline Random Forest (RF) classifier, an ensemble of decision tree models. Each tree is created from a different bootstrap sample of the training dataset – see Figure 9. For classification, the prediction is the majority vote class label across all decision trees. RF can be used for time series forecasting, although it requires the dataset to first be transformed into a supervised learning problem.

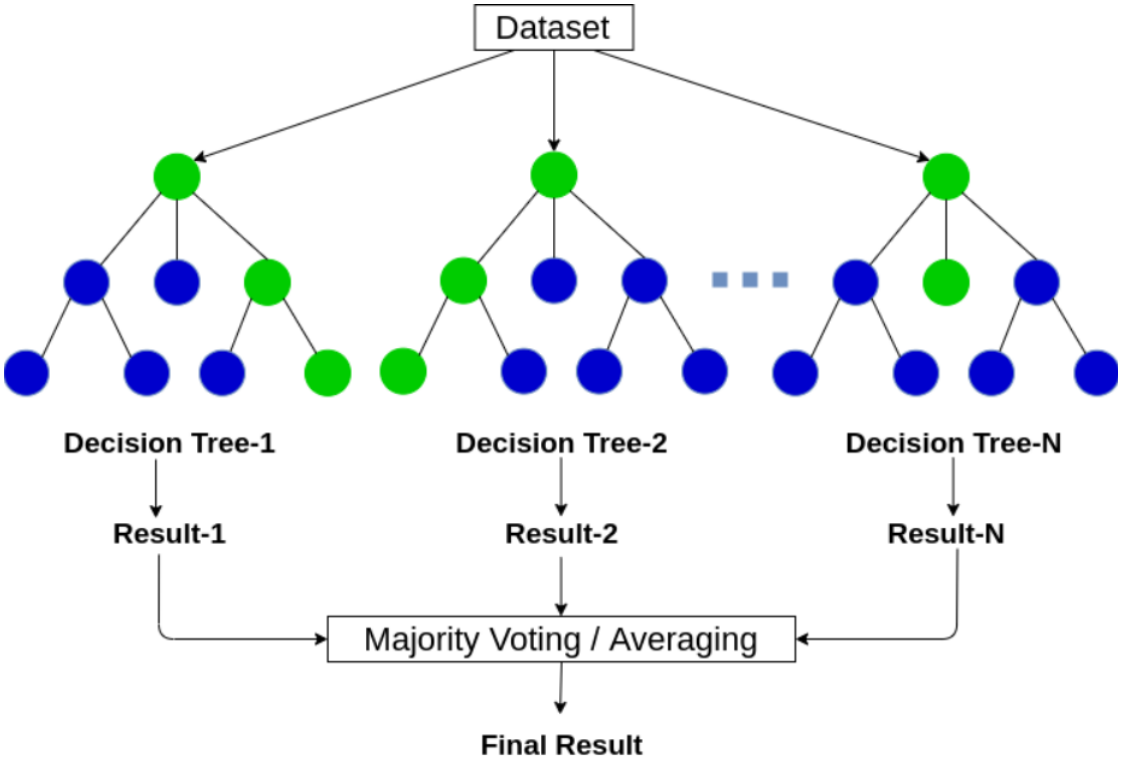


Figure 9: Visualisation of an RF model showing how the voting procedure of each tree leads to the model output

3 Dataset

3.1 Description

The 'sleep cassette' study from the sleep-EDF database of Kemp et al. (2000) was used, which contains 153 PSG whole night recordings. The database has standard EEG, EOG, EMG readings, as well as oro-nasal respiration and rectal body temperature and patient gender and age. The data was originally used to assess the impact of age and gender on sleep (Mourtazaev et al., 1995). Between 1987 and 1991, they recorded 2 nights of sleep at home for 78 healthy Caucasians (aged 25-101) without any sleep-related medication. They used modified Walkman-like cassette-tape recorders (Kemp, 1987), which only yielded 153 recordings, due to hardware mishaps with 3 recordings (82 female and 71 male nights).

The data was labelled by expert technicians, using the classification approach of Rechtschaffen (1968), which separates S3 and S4. See Figure 10 for an example of the labelling and Appendix E for the raw data.

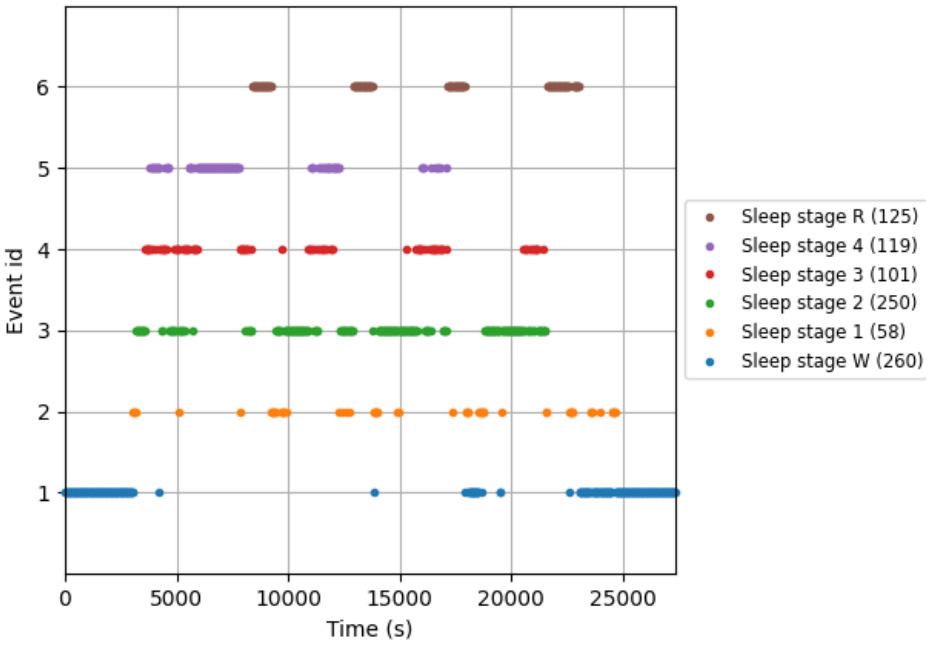


Figure 10: Example of expert labelling for one individual (participant 4001)

3.2 Preprocessing

Raw data comprising 6 channels (EEG ($\times$2), EOG, EMG, oro-nasal respiration and rectal temperature), originally recorded at 100 Hz, were decomposed into the 6 frequency subbands using a band-pass Butterworth filter, as per Table 2. The low- and high-pass values of these subbands are not codified in the literature (see Appendix B for examples); however, we employ typical values adapted from Memar and Faradji (2017) whose notable exception is γ (gamma), often taken as frequencies up to and beyond 100 Hz; however, our γ-band is curtailed due to Nyquist frequency constraints. The resulting data were epoched into 30-second non-overlapping windows, specifying no more than a 5-second overlap between sleep stages (as expert-labelled). The resulting data form a $6\times6\times3000$ tensor.

Subband	high-pass (Hz)	low-pass (Hz)
δ (delta)	0.5	4.5
θ (theta)	4.5	8.5
α (alpha)	8.5	11.5
σ (sigma)	11.5	15.5
β (beta)	15.5	32.5
γ (gamma)	32.5	49.5

Table 2: Our 6 frequency subbands with their low- and high-pass frequencies

4 Model

4.1 Training

Models were written to use PyTorch, as developed by Paszke et al. (2019). Models were trained in batches of 256 employing an 80/10/10 train/test/validation split, on the University of Bath's Hex compute cluster (using a GeForce RTX 3090 GPU). Models were set by default to train for 1000 epochs; but we implement early stopping with a patience of 30 epochs (training ends if there is no improvement to the validation loss after 30 epochs). The loss function for our multiclass classification task is cross-entropy loss. Our optimisation function implements the AdamW algorithm, developed by Loshchilov and Hutter (2017), with a learning rate of 0.001 and a weight decay of 0.0005. These are notably lower than typical classification tasks employing deep learning; but preliminary investigations revealed this learning rate and weight decay yield the best results (the models struggled to converge with higher learning rates), and such values were inspired by recommendations of Schirrmeister et al. (2017). Finally, we employ a cosine annealing scheduler with a restart period of 5 epochs; due to the high noise and variance of EEG signal, cosine annealing is an invaluable addition to deep learning models to mitigate the tendency of models using such data to converge on local minima during gradient descent (particularly due to the low hyperparameters we use for our optimisation function).

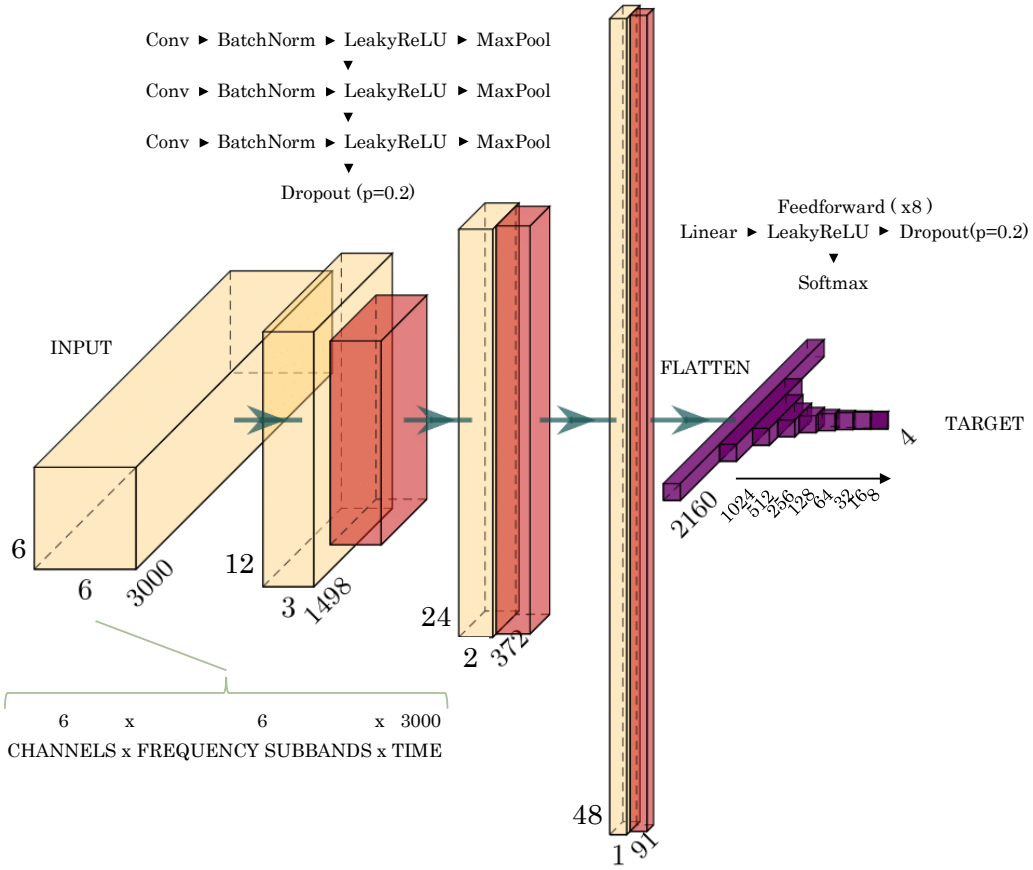


Figure 11: A high-level illustration of our chosen CNN architecture, as settled upon following experimentation of different architectures

Figure 11 is an illustration of our chosen model architecture, settled upon following inspiration from the literature (Section 2) and preliminary investigations with non-CNN and CNN-extended model architectures (Section 4.3). The hyperparameters for the best CNN model are provided in Table 3. These values were chosen following a comprehensive hyperparameter sweep. Table 4 provides the model accuracy scores for one participant across a sweep for the optimal kernel sizes for max-pooling and convolution operations. Our preliminary investigations revealed that convolutions across the channels (i.e. height-wise convolutions) lowered model performance considerably. We therefore limited our convolution (and max-pooling) operations to be across the samples (i.e. width-wise convolutions); Table 4 reveals the result of our sweep for kernel sizes, with an optimal 'sweet spot' of 6 for convolutions, and 2 for max-pooling.

CNN Hyperparameter	Value
in/out channels conv layer 1	$6 \rightarrow 12$
in/out channels conv layer 2	$12 \rightarrow 24$
in/out channels conv layer 3	$24 \rightarrow 48$
convolutions kernel shape	(1, 6)
max-pooling kernel shape	(1, 2)
dilation shape	(1, 1)
padding shape	(0, 0)
dropout probability	0.2
learning rate	1e-4
weight decay	5e-4
cosine annealing warm restart	5
early stopping patience	30

Table 3: Chosen hyperparameters for our best-performing CNN architecture

Classes	Max-Pooling Kernel (down) Convolutions Kernel (across)	(1,3)	(1,4)	(1,5)	(1,6)	(1,7)	(1,8)
6 (all sleep stages)	(1,1)	41.87	52.22	39.14	42.74	55.76	32.1
	(1,2)	52.82	42.91	59.97	67.16	52.11	43.66
	(1,3)	62.48	47.21	65.03	62.47	49.59	52.13
	(1,4)	51.37	40.84	53.41	69.67	68.44	38.22
5 (sleep stages ¾ merged)	(1,1)	41.23	53.44	49.25	61.19	40.52	37.76
	(1,2)	57.91	53.14	73.62	74.63	71.11	68.65
	(1,3)	55.9	49.8	70.12	64.17	56.55	41.32
	(1,4)	41.1	55.36	72.41	64.17	68.86	44.73
4 (sleep stages ½ and ¾ merged)	(1,1)	33.91	51.25	61.34	70.72	77.79	49.51
	(1,2)	49.25	68.34	81.19	85.07	56.55	68.65
	(1,3)	59.25	63.39	75.4	79.91	71.55	61.37
	(1,4)	40.75	68.34	68.82	79.05	59.66	58.7

Table 4: CNN model accuracies (with 5-fold cross-validation, and architecture as per 11) for a comprehensive sweep of kernel sizes for convolution and max-pooling operations, for a single participant's data (participant data 4001 and 4002)

4.2 Evaluation

4.2.1 Number of classes: 4, 5, 6

4 classes yield top accuracy: 74%. Tables 5, 6 and 7 compare the accuracy of 4, 5 and 6 classes for one subject's night of sleep. Accuracy declines from 0.74 for four to 0.63 for five and 0.57 for six classes, as would be expected from the literature (Table 14). This subject is relatively distinctive: S2 represents 35% of support (50% expected – Table 1) and SWS 37% (20% expected).

Wake and sleep spindles/ k-complex identification could be improved. Four classes predict wakefulness best (0.56 vs 0.05 and 0.04), with less confusion with S1 and REM (Figure 12). The model struggles to predict S1 (0.15, 0.05 accuracy), confusing it with SWS, which probably means it is not identifying 'sleep spindles' or 'k-complex' well, potentially because of the pre-processing. Combining S1 with S2 to give light sleep improves performance above S2 alone: 0.77, probably because they both have theta waves, and the model may not be picking up 'sleep spindles' or 'k-complexes' well. S3 prediction is poor compared to S4, and combining them improves performance above S4 alone. Since Iber (2007) published the updated criteria, it is no longer considered good practice to split them apart. This 0.73 accuracy on SWS is not observed in many other subjects (e.g. 20 subjects in Table 10 have 0.28 accuracy). It may be that the distinctiveness of the subject's SWS is helping improve classification. REM accuracy is relatively consistent and well predicted, rising slightly when S1 and S2 are separate, probably because slow eye movements occur in S1. REM is sometimes confused with wakefulness (Figure 12), which is consistent with expected eye movement.

	Precision	Recall	f1-score	Support
Sleep stage W	0.43	0.83	0.56	100
Sleep stage 1/2	0.81	0.74	0.77	675
Sleep stage 3/4	0.76	0.71	0.73	575
Sleep stage R	0.76	0.76	0.76	325
Accuracy	-	-	0.74	1675
Macro avg	0.69	0.76	0.71	1675
Weighted avg	0.76	0.74	0.74	1675

Table 5: 4 classes: results when sleep stages 1 & 2 (light sleep) and 3 & 4 (SWS) are combined

	Precision	Recall	f1-score	Support
Sleep stage W	0.05	0.05	0.05	100
Sleep stage 1	0.12	0.19	0.15	125
Sleep stage 2	0.72	0.69	0.70	550
Sleep stage 3/4	0.67	0.70	0.68	575
Sleep stage R	1.0	0.74	0.85	325
Accuracy	-	-	0.63	1675
Macro avg	0.51	0.47	0.49	1675
Weighted avg	0.67	0.63	0.64	1675

Table 6: 5 classes: results when sleep stages 3 & 4 are combined (modern definition of SWS)

	Precision	Recall	f1-score	Support
Sleep stage W	0.04	0.04	0.04	100
Sleep stage 1	0.07	0.04	0.05	125
Sleep stage 2	0.71	0.78	0.74	550
Sleep stage 3	0.32	0.13	0.18	275
Sleep stage 4	0.52	0.80	0.63	300
Sleep stage R	0.71	0.74	0.72	325
Accuracy	-	-	0.57	1675
Macro avg	0.39	0.42	0.39	1675
Weighted avg	0.52	0.57	0.53	1675

Table 7: 6 classes: results when all sleep classes are kept separate (former definition of SWS)

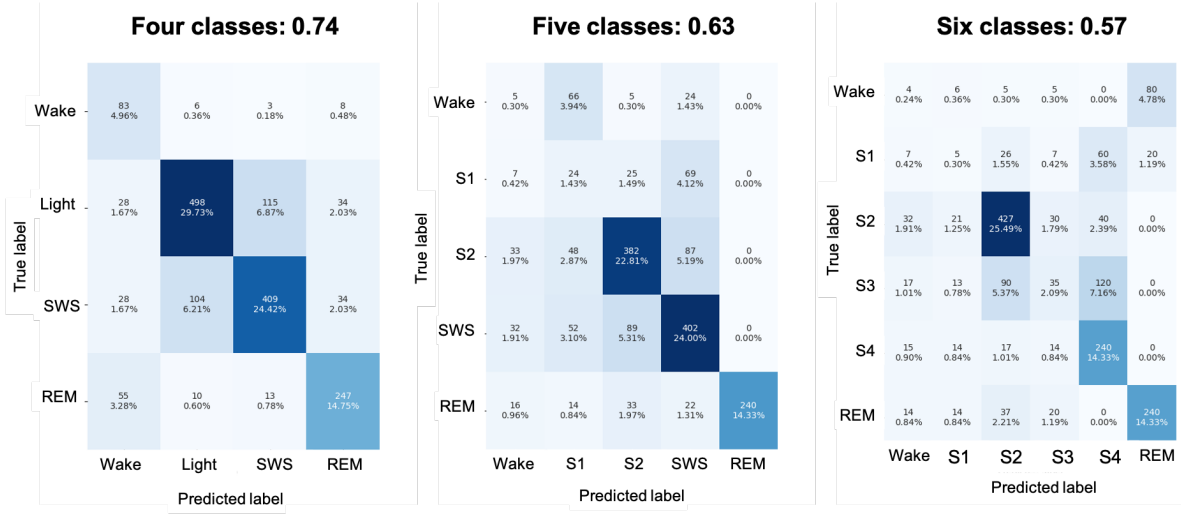


Figure 12: Confusion matrix for four, five and six classes. Depth of blue indicates the proportion of the data set in the box. If the CNN was predicting perfectly, one would expect the strongest blue boxes in a diagonal line from top left to bottom right – i.e. the predicted label was the true label most of the time. Four classes are W, light (1 and 2 combined), SWS (3 and 4 combined), REM. Five splits out 1 and 2; six splits out SWS into 3 and 4 (old definition)

4.2.2 Number of subjects: one, ten, twenty

Twenty subjects yield top accuracy of the 3 groups tested: 0.77. Tables 8, 9 and 10 compare the accuracy for one, ten and twenty subjects. While the '1' subject is different from the subject in Section 4.2.1, accuracy is similar at 0.72. Accuracy declines to 0.68 for 10 subjects and up again to 0.77 for twenty subjects. The sleep distribution of the 10 sample is most similar to what would be expected, as shown in Table 1: 59/21/20% for light/SWS/REM support respectively (expect 55/20/25%). The one and twenty samples are far higher on light sleep (74-78%) and lower on SWS and REM. All these subjects score well on W and light sleep prediction (0.75-0.86), poorly on SWS (0.01-0.28), and moderately well on REM (0.53-0.61). Given ten has a higher concentration of poorly predicted sleep stages, its overall accuracy suffers – had it the mix of one, accuracy would not have gone down for ten.

Confusion tends to decrease as subject numbers increase. There is limited confusion in predicting wakefulness and light sleep (Figure 13), especially once subjects get up to twenty. SWS is more often taken for light sleep and REM, even for twenty subjects, again indicating a potential issue with detecting spindles and k-complex. REM is taken for light sleep and SWS, with confusion decreasing as subjects increase.

	Precision	Recall	f1-score	Support
Sleep stage W	0.86	0.73	0.79	2100
Sleep Stage 1/2	0.75	0.76	0.75	1800
Sleep Stage 3/4	0.00	0.01	0.01	100
Sleep Stage R	0.45	0.65	0.53	400
Accuracy	-	-	0.72	4400
Macro avg	0.52	0.54	0.52	4400
Weighted avg	0.76	0.72	0.73	4400

Table 8: Model results for one subject

	Precision	Recall	f1-score	Support
Sleep stage W	0.86	0.81	0.83	1975
Sleep Stage 1/2	0.84	0.75	0.79	2600
Sleep Stage 3/4	0.15	0.05	0.08	925
Sleep Stage R	0.39	0.83	0.53	900
Accuracy	-	-	0.68	6400
Macro avg	0.56	0.61	0.56	6400
Weighted avg	0.68	0.68	0.66	6400

Table 9: Model results for ten subjects

	Precision	Recall	f1-score	Support
Sleep stage W	0.82	0.81	0.81	1625
Sleep Stage 1/2	0.91	0.82	0.86	3550
Sleep Stage 3/4	0.27	0.29	0.28	500
Sleep Stage R	0.51	0.75	0.61	725
Accuracy	-	-	0.77	6400
Macro avg	0.63	0.67	0.64	6400
Weighted avg	0.79	0.77	0.78	6400

Table 10: Model results for twenty subjects

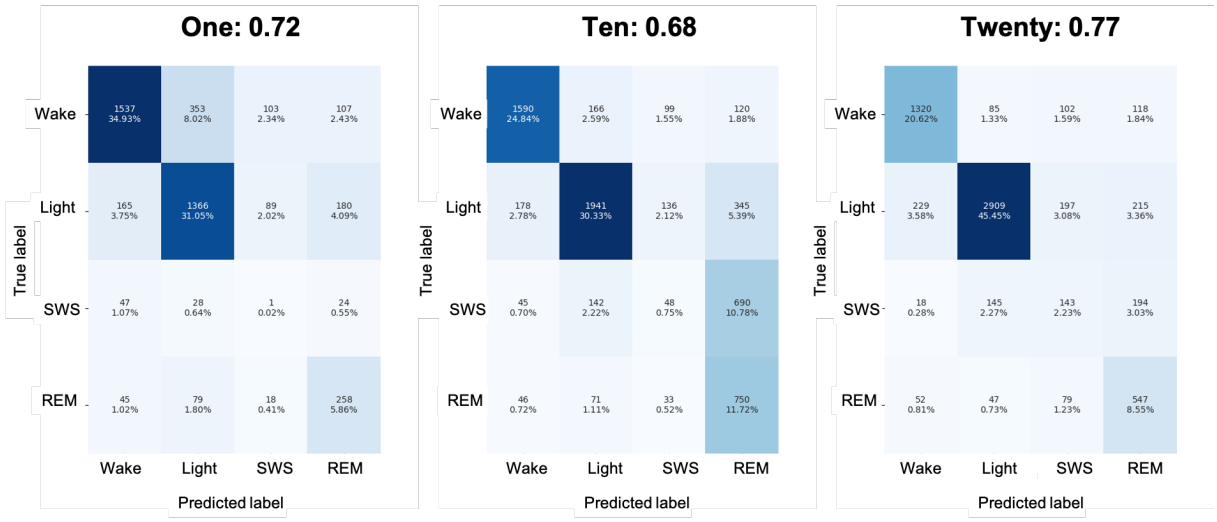


Figure 13: Confusion matrix for one, ten, twenty subjects. Depth of blue indicates the proportion of the data set in the box

4.2.3 Channel

Use of all channels improved performance for light and REM sleep prediction (Table 11). The improvement is most marked for light sleep and PSG with an 8 percentage point improvement – EMG recordings are probably helping to distinguish S1 from REM, because REM also improves (paralysis). SWS prediction decreases when EOG and EMG data are added, and improves again when breath and temperature are included, reflecting the slow breath nature of SWS (Table 1 and spindle/ k-complex detection issues).

	EEG only	PSG only	All
Light (1/2)	0.60	0.68	0.70
SWS (3/4)	0.78	0.73	0.78
REM	0.73	0.75	0.77

Table 11: Model accuracy by channel (4 classes). PSG only: EEG plus EOG and EMG. All: PSG plus rectal temperature and breath

4.2.4 Gender and age

Table 12 shows model accuracy for male and female subjects in 10-year age buckets from under 30 to over 90. There were no subjects aged 40-50 and it was not possible for the model to classify males over 90. Accuracy is relatively consistent by age and gender, between 0.61 and 0.78.

As can be seen in the confusion matrices in Figure 14, the model is failing to predict SWS well with these subjects. The male data set suffers particularly from this SWS prediction issue. This may be linked to older men experiencing declining amounts of SWS as a proportion of total sleep, as shown on the left-hand side of Figure 15. This group of males appear to reduce their proportion of SWS earlier in life than shown in the work of Ohayon et al. (2004), where the decline is most marked after 65 – see Appendix C.1. The females exhibit more typical sleep behaviour, with total amount of sleep declining with age, with less decline in SWS than males.

	<30	30-40	50-60	60-70	70-80	80-90	>90
Female	0.73	0.73	0.72	0.71	0.61	0.78	0.76
Male	0.72	0.68	0.74	0.62	0.68	0.77	N/A

Table 12: Model accuracy by age and gender. Each bucket includes a random selection of six subjects from all the subjects of that age/ gender, as that was the lowest observed bucket size. Note that no subjects were aged 40-50

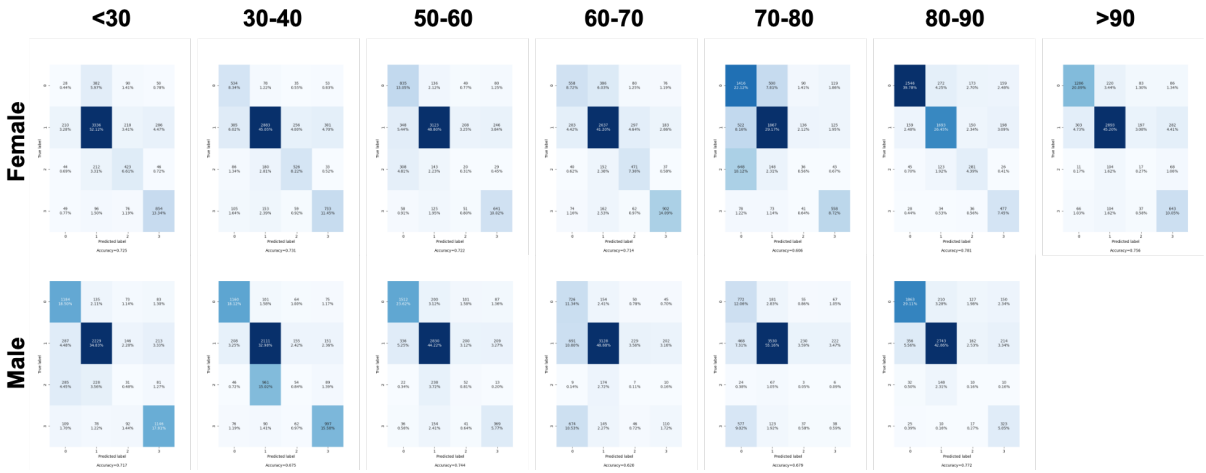


Figure 14: Confusion matrix for male and female subjects by age group. Note that no subjects were aged 40-50

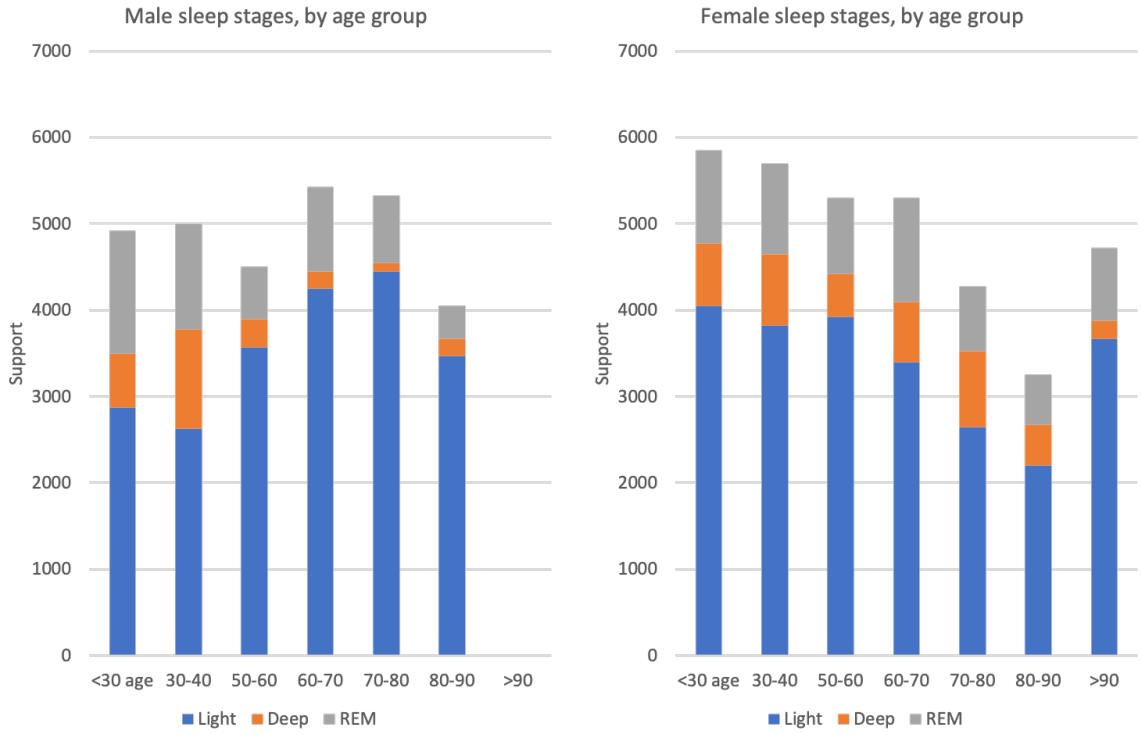


Figure 15: Support by age and gender. Note that no subjects were aged 40-50

4.3 Performance

The final model for the 6-class model had a validation and training loss of 0.60 and 0.68 respectively. This model also had a accuracy score on the validation set of 0.72. The performance of this model was significantly better than the other CNN variants trained on the data (see Table 13). The other CNN variants were significantly lower for the same pre-processed data to the point that the models performed only slightly better than random label allocation.

Model	Training Epochs	Training Loss	Validation Loss	Training Time	Test Accuracy
CNN	187	0.684	0.607	21.25 secs	0.71
FCN	147	2.905	3.07	34.14 secs	0.34
CNN-3D	203	1.32	1.313	45.98 secs	0.34
CNN-LSTM	629	1.3	1.273	10 mins, 55 secs	0.40

Table 13: Number of epochs and time to train, training and validation loss and final test set accuracy of our chosen CNN model, alongside the other models that formed part of our investigations during this research

Model architectures for the Fully Convolutional Network (FCN), the 3D-CNN (which employed 3D kernels upon the data), and the CNN-LSTM network (which employed a bi-directional LSTM between the CNN and linear network layers) can be found in Appendix D. It should be noted that the the flattened data for the FCN would contain 108,000 input features if we were to use the full 30-second epochs of our chosen model. The Hex compute cluster was unable to allocate enough resources to support such large-scale processing. We therefore took sub-crops of each window instead. Specifically, we employed a sliding window of width 32 (our 30 seconds per epoch is width 3,000), and a step size of 16. These sub-crops reduced the number of input features to the FCN from 108,000 to 1,152. We do not anticipate that the additional processing affected the integrity of the data.

Receiver Operating Characteristic (ROC) values for 4, 5, 6 class models are 0.78, 0.83, 0.78 respectively (see Figure 16). The higher true positive rate for 5 classes indicates that this model had the highest allocation of true positives across all classes for each of the class models.

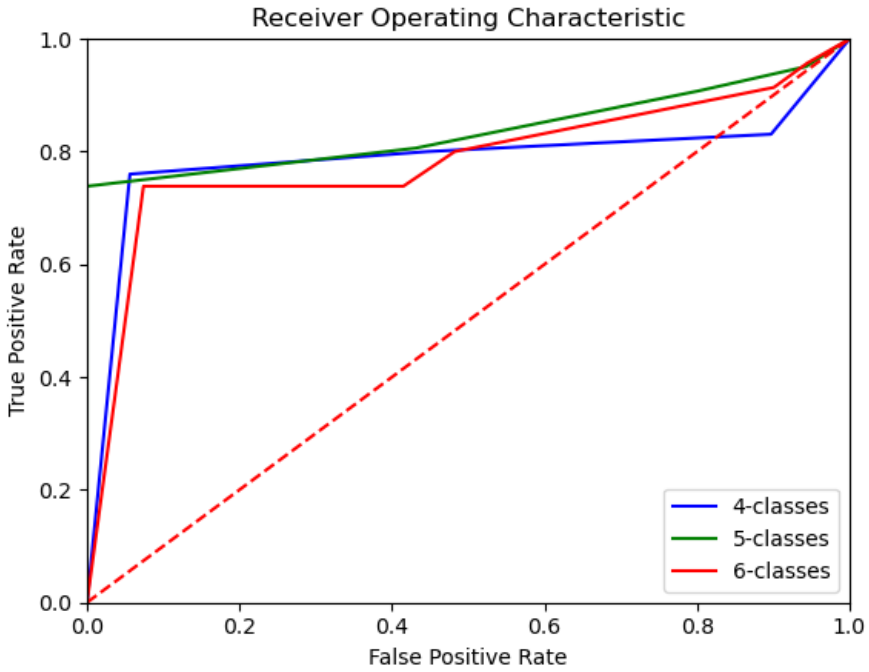


Figure 16: A receiver operated characteristics (ROC) plot for 4, 5, 6 classes model

The results for the baseline Random Forest classifiers were 0.48 for a model trained on the same frequency band decomposition data as the CNN model. A model was also trained on data processed with PSD transform (see Figure 8). This model achieved an accuracy of 0.78 for classification, the highest of the any of the models trained. It is important to note here that this type of preprocessing cannot be used for CNN models. This is because the output is two numbers that are highly correlated to the sleep state. An input of two features is not an appropriate input for a convolutional network so was not explored.

Saliency mapping was also used to assess what the model had learned regarding which features are important to a class classification, to improve the explainability of the model. The procedure, first proposed in Zeiler and Fergus (2014), uses a gradient-based approach to assess what features in the input are most important to the output classification given by the model. This can be seen in Figure 17, where the salience as a function of time is plotted for each of the input channels. The output of the saliency process is a normalised

histogram which is plotted as a heat-map for visual aid for the time series data. It is important to note here that gradient based saliency mapping is not just about what features are indicative of a given class being ascribed – it also extracts what input features change the result by the largest amount. This means that the result also carries information about how this output class relates to other classes the model has learned, and about which feature needs to be changed to lead the model to ascribe data to another class.

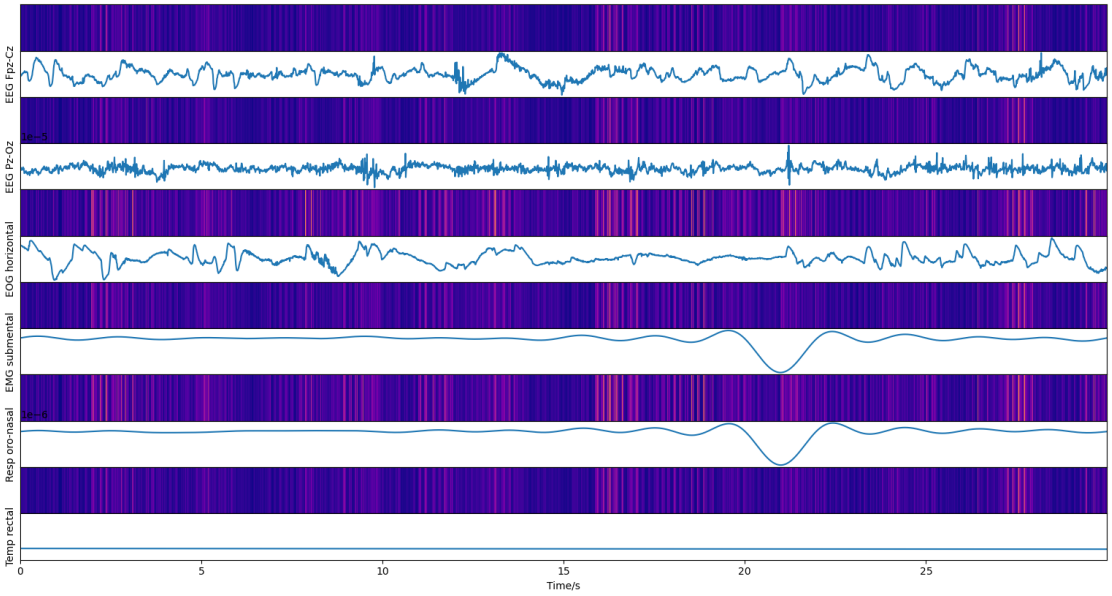


Figure 17: A plot of the saliency heatmaps for each of the raw input channels for a sleep window, which the model classified as Sleep Stage 1/2. Each heatmap is associated with the raw data channel directly below it

5 Comparison

Berthomier et al. (2007) quote an accuracy of 0.75 on 4-state data over fifteen subjects. This was achieved with their automatic sleep scoring software ASEEGA (see Table 14). Our model achieved an accuracy of 0.67 and 0.77 with ten and twenty participants respectively. Interpolation of these results would estimate an accuracy of 0.72 for fifteen participants, noting that participant sleep mix is an important determinant of accuracy (Section 4.2.2). This result indicates that while the model may not significantly outperform the other models in the literature, it is comparable. Although Table 6 presents the accuracy for 5 classes, this is only the results from one person; so the accuracy for the Berthomier et al. (2007) paper from Table 14 is not directly comparable. This is also the case for the Tagluk, Sezgin and Akin (2010) and Shimada, Shiina and Saito (2000) for 5 classes because their results are also the result of multi-participant studies.

Study	Model	Data Type	Accuracy
Lajnef et al. (2015)	DSVM	-	0.88
	DSVM	Restricted	0.92
Berthomier et al. (2007)	ASEEGA	2 states(wake/sleep)	0.82
	ASEEGA	3 states(wake/sleep/REM)	0.81
	ASEEGA	4 states (wake/REM/stages 1-2/SWS),	0.75
	ASEEGA	5 states (wake/REM/ stage 1/stage 2/SWS)	0.72
Lajnef et al. (2015)	ANN	6 states	0.74
Ronzhina et al. (2012)	ANN	-	0.70 - 0.97 (state dependent)
Tagluk et al. (2010)	ANN	5 states (REM, S1, S2, S3, S4)	0.74
Shimanda et al. (2010)	ANN	5 states (W, S1, S2, S3, S4)	0.72

Table 14: Results of different classification approaches from different sleep states from different publications. Note that the accuracy between different results is not always comparable, because the number of state classes varies between publications

6 Further work

We could look at **class balancing the EEG data** for under-represented classes with synthetic data. Class balancing is important for sleep EEG data as some classes are far more over-represented than others across different demographics i.e. lack of deep sleep in older people. This synthetic data could be generated with a Generative Adverserial Network (GAN) that is trained on the under-represented class and would allow for data augmentation as necessary in a given subject.

We may also wish to investigate **feature engineering during the data preprocessing stage**; specifically, we may wish to investigate the benefits of deep learning using feature-based representations of the data, as opposed to training on 'raw' data. This is a process commonly used in speech recognition: the speech signal may first be preprocessed as Mel-frequency cepstral coefficients (MFCCs), which decomposes the raw signal into 13 features (plus delta and delta-delta features) which 'describe' the speech envelope. A similar process for the EEG signal data would be a worthwhile investigation. This could include, for example, using a GAN-like architecture to model a Gaussian prior to inform a universal background model (UBM) for improved inter-subject EEG sleep stage classification. In this work, special focus should be given to enabling the model to **recognise spindles and k-complexes**, as this could greatly improve model forecast accuracy.

Appendices

A EEG, EOG, and EMG measurement sites

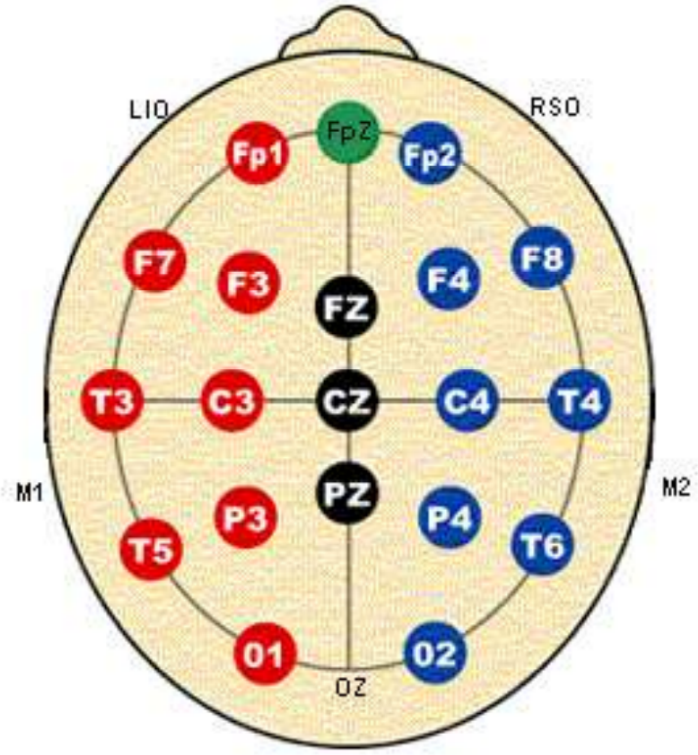


Figure 18: EEG brain activity sites: F4, C4, O2; back-up F3, C3, O1, source: Iber (2007)

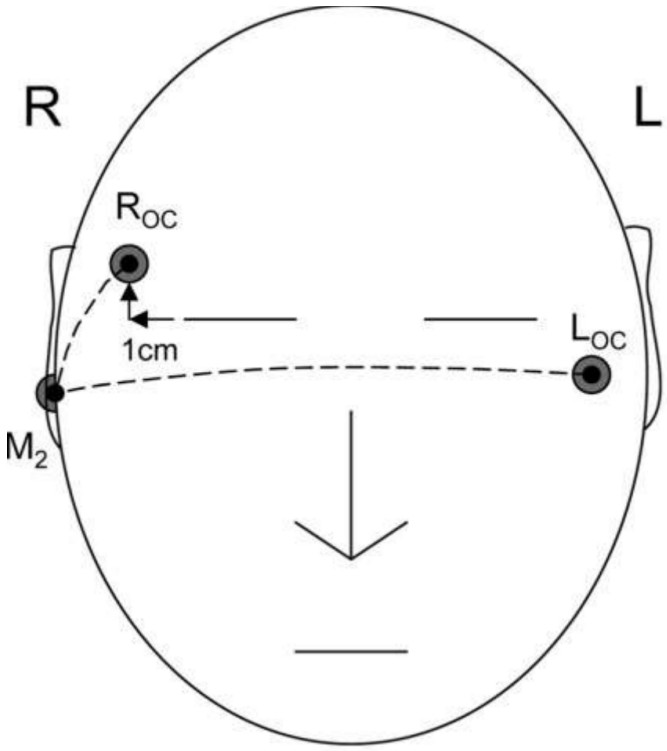


Figure 19: EOG eye movement sites: E1-M2 and E2-M2; E1 is placed 1cm below the left outer canthus (LOC); E2 1cm above the right OC, source: Iber (2007)

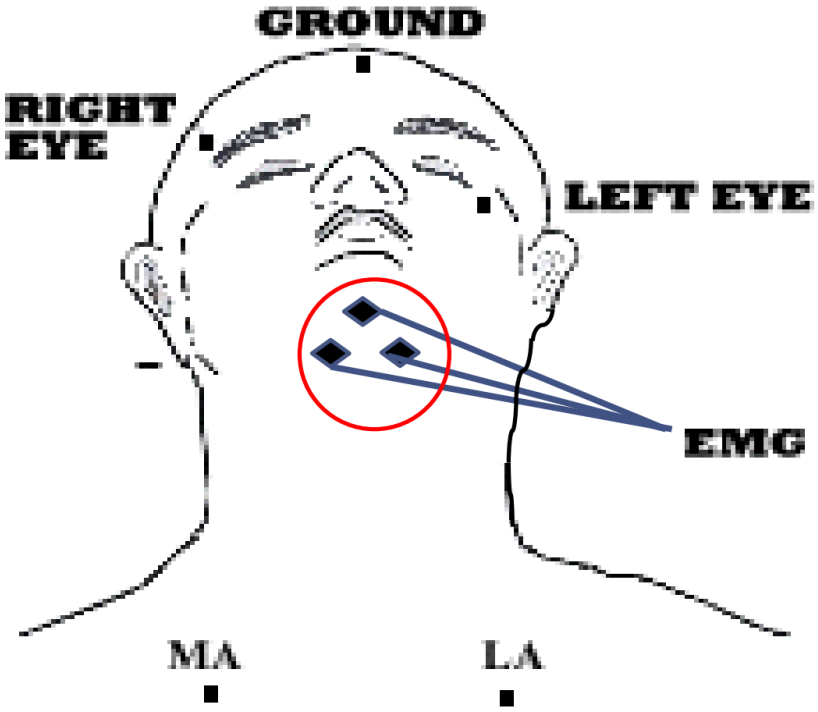


Figure 20: EMG muscle tone sites, source: Iber (2007)

B EEG frequency bands

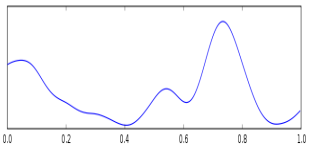
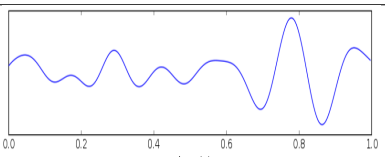
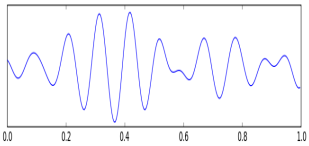
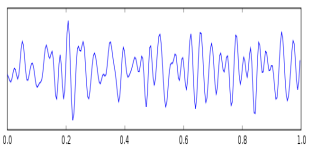
Wave	Frequency	Location on head	One second sample
Delta	<4 Hz (slow waves)	Frontally in adults	
Theta	4-8 Hz	Unrelated to task at hand	
Alpha	8-14 Hz	Posterior, both sides; higher on dominant side. Central sites at rest	
Beta	14-30	Symmetrical, both sides; most evident frontally	

Table 15: EEG frequency bands, source: Deuschl (1999) and Hugo Gamboa

C Sleep quantity and quality

C.1 Sleep and aging

Our ability to sleep diminishes as we age. People over the age of 70 sleep 36% less than 5-year-old children. In the study of Ohayon et al. (2004) on people 65 and older, only 12% reported no sleep issues; while over 50% had chronic sleep issues most of the time. Adults over the age of 65 wake on average 1.3 hours earlier than adults aged 20-30.

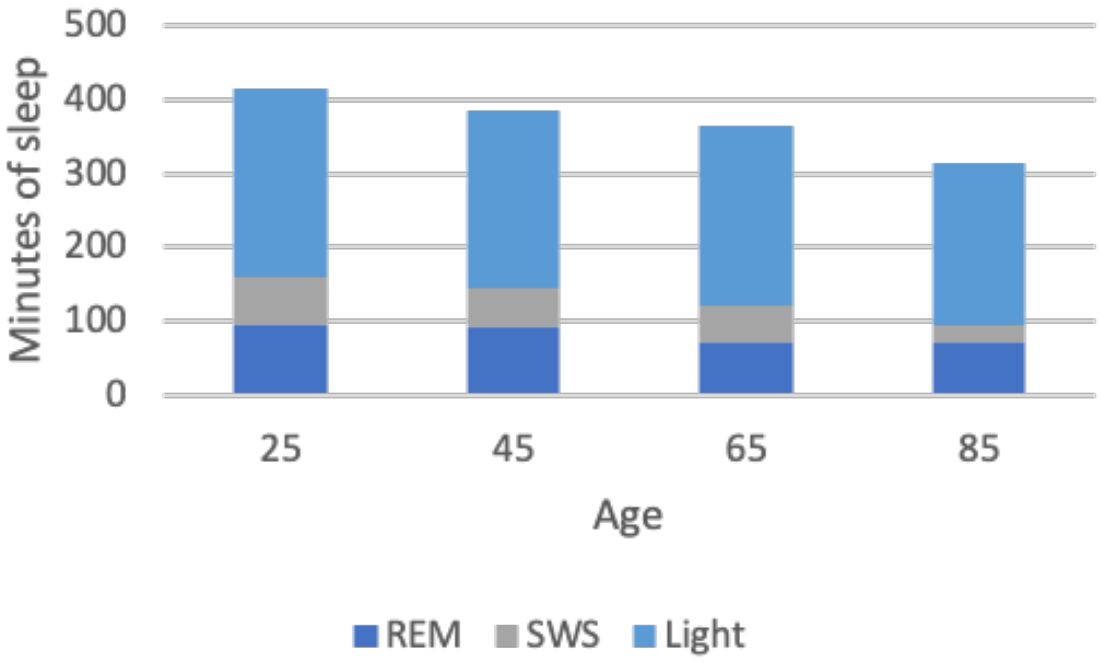


Figure 21: Aging and sleep, source: Ohayon et al. (2004)

Our need for sleep does not diminish with age. Not sleeping well at night increases the need to take naps during the day, slows response times, and makes it harder to concentrate and remember things. These symptoms can have more impact on older people, because they can be wrongly interpreted as early signs of dementia. Slower response times makes it harder to drive safely and increases the chances of falling. Older people who don't sleep well are twice as likely to die from a heart attack, stroke or cancer than an older person who sleeps well.

Some medications make it harder for us to sleep as we age, as do mental-health issues like depression, anxiety, and primary insomnia, as well as disruptions to our circadian rhythm. Mental-health issues and primary insomnia are best dealt with by a GP, who will have many good solutions including cognitive behavioural therapy. Sorting out disruptions to circadian rhythm is where sleep hygiene helps.

C.2 Sleep quality

Sleep is a cyclical process – we go through repeated cycles of REM and non-REM sleep (light and deep sleep). Sleep quality can be measured in terms of the amount of REM and deep sleep that we get. A healthy adult gets 20-25% REM sleep, and 20% deep sleep (or slow-wave sleep (SWS)). However, this proportion declines as we age. As shown in the chart above, comparing the sleep of 25-, 45-, 65- and 85-year olds, the total amount of sleep we get declines as we age. REM sleep declines a little faster than total sleep, while SWS sleep falls precipitously as a proportion of our sleep. This decline is particularly marked in elderly men – men aged 70 and older get 25-33% of the SWS of females of the same age.

C.2.1 REM

REM sleep is associated with dreaming and learning, memory, creativity and mood. It is negatively impacted by alcohol and insufficient sleep.

If we remember a vivid dream, it is most likely that we woke out of REM sleep – around 80% of vivid-dream recall awoke from REM. Scientists can't yet explain the purpose of dreaming but believe it may help us process emotions. People who suffer from stress and anxiety report especially frightening dreams.

REM stimulates the areas of our brain involved in learning and memory consolidation. People who are deprived of REM sleep in experiments have difficulty forming (or expressing) spatial and emotional memories. They also experience difficulty concentrating and processing social interactions. REM appears to prune, strengthen and maintain new synapses associated with motor learning. It is believed to support neuroplasticity and creativity.

Alcohol reduces the amount of REM we get, especially in high doses, as does too few hours of sleep, because more REM occurs towards the end of the night. REM decreases for people with brain diseases like Alzheimer's and Parkinson's, especially in the later stages of these diseases.

During REM, our eyes dart back and forth behind closed eyelids. Our brains become very active – especially in motor and sensory areas – flooded by theta and slow alpha waves. Our heart rate and breathing speed up. Blood flow to the brain increases 50-200%. Our body becomes temporarily paralysed; scientists believe this is happens so that we do not act out our dreams.

REM makes up to 50% of a newborn's sleep, where it is believed to support brain development. By the age of 2, REM reaches 20-25% of sleep, where it stays for the rest of our lives, declining marginally as we age.

The amount of REM we get increases throughout the night. The first cycle comes 90-110 minutes after we fall asleep, and may only last 1-10 minutes. The cycles then come roughly every 90 minutes, with periods getting longer the later they are in the night. We get most REM in the last third of the night, when each REM cycle can be up to 60 minutes in length/ duration.

C.2.2 Deep sleep

Deep sleep supports brain and body healing and repair. The level of glucose in the brain increases, supporting memory processing and learning. Secretion of human growth hormone supports cell regeneration in adults, regulating body composition and metabolism, and supporting muscle and bone growth. Muscle blood supply increases, carrying essential nutrients to help our muscles recover, which is especially important if we exercise regularly. Deep sleep also boosts the immune system; without it, we are more likely to get sick, gain weight and feel down.

We cycle through periods of deep sleep throughout the night. These periods are longest at the start of the night, lasting 45-90 minutes, and shorten with each sleep cycle. If we go to sleep later than usual, we can get less deep sleep because our circadian rhythm pulls us forward into REM and light sleep. You can increase the amount of deep sleep you get by getting to bed on time, doing vigorous exercise early in the day, or having a hot bath before you go to bed.

Deep sleep is also known as slow-wave sleep (SWS) or delta sleep, as our brains are flooded with delta waves. Our muscles relax, our heart rate and breathing slows. We are hardest to rouse during this period of sleep, even with loud noises. We can be decidedly groggy if we are woken from deep sleep, and experiments show that mental performance can be impaired for up to 30 minutes.

The older we are, the less deep sleep we get. Someone under 30 might get 2 hours; but someone who's over 70 might get only 30 minutes, or none at all. This is entirely normal and does not indicate a sleep disorder. Scientists don't know why it happens, but it is linked to body and brain growth and development.

D Print-out architectures for models used in this investigation

D.1 CNN parameters and hyperparameters

Arguments for model:

```
'in_channels': (6, 12, 24),
'out_channels': (12, 24, 48),
'input_size': (6, 6, 256),
'kernel_size': (1, 6),
'pool_size': (1, 2),
'stride': 2,
'dilation': (1, 1),
'padding': (0, 0),
'dropout': 0.2
```

Model architecture:

```
CNN(
  (cnn1): Conv2d(6, 12, kernel_size=(1, 6), stride=(2, 2))
  (bn1): BatchNorm2d(12, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
  (leakyrelu1): LeakyReLU(negative_slope=0.01)
  (maxpool1): MaxPool2d(kernel_size=(1, 2), stride=(1, 2), padding=0, dilation=1, ceil_mode=False)
```

```
  (cnn2): Conv2d(12, 24, kernel_size=(1, 6), stride=(2, 2))
  (bn2): BatchNorm2d(24, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
  (leakyrelu2): LeakyReLU(negative_slope=0.01)
  (maxpool2): MaxPool2d(kernel_size=(1, 2), stride=(1, 2), padding=0, dilation=1, ceil_mode=False)
  (cnn3): Conv2d(24, 48, kernel_size=(1, 6), stride=(2, 2))
  (bn3): BatchNorm2d(48, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
  (leakyrelu3): LeakyReLU(negative_slope=0.01)
  (maxpool3): MaxPool2d(kernel_size=(1, 2), stride=(1, 2), padding=0, dilation=1, ceil_mode=False)
  (dropout): Dropout(p=0.2, inplace=False)
  (feedforward): Sequential(
    (0): Linear(in_features=2160, out_features=1024, bias=True)
    (1): LeakyReLU(negative_slope=0.01)
    (2): Dropout(p=0.2, inplace=False)
    (3): Linear(in_features=1024, out_features=512, bias=True)
    (4): LeakyReLU(negative_slope=0.01)
    (5): Dropout(p=0.2, inplace=False)
    (6): Linear(in_features=512, out_features=256, bias=True)
    (7): LeakyReLU(negative_slope=0.01)
    (8): Dropout(p=0.2, inplace=False)
    (9): Linear(in_features=256, out_features=128, bias=True)
    (10): LeakyReLU(negative_slope=0.01)
    (11): Dropout(p=0.2, inplace=False)
    (12): Linear(in_features=128, out_features=64, bias=True)
    (13): LeakyReLU(negative_slope=0.01)
    (14): Dropout(p=0.2, inplace=False)
    (15): Linear(in_features=64, out_features=32, bias=True)
    (16): LeakyReLU(negative_slope=0.01)
    (17): Dropout(p=0.2, inplace=False)
    (18): Linear(in_features=32, out_features=16, bias=True)
    (19): LeakyReLU(negative_slope=0.01)
    (20): Dropout(p=0.2, inplace=False)
    (21): Linear(in_features=16, out_features=8, bias=True)
    (22): LeakyReLU(negative_slope=0.01)
    (23): Dropout(p=0.2, inplace=False)
    (24): Linear(in_features=8, out_features=4, bias=True)
  )
)
```

D.2 FCN parameters and hyperparameters

Arguments for model:

```
'layers': 5,
'input_size': (32, 36),
'output_size': (32,),
'dropout': 0.2
```

Model architecture:

```
FCN(
  (feedforward): Sequential(
    (0): Linear(in_features=1152, out_features=592, bias=True)
    (1): BatchNorm1d(592, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (2): LeakyReLU(negative_slope=0.01)
    (3): Dropout(p=0.2, inplace=False)
    (4): Linear(in_features=592, out_features=312, bias=True)
    (5): BatchNorm1d(312, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (6): LeakyReLU(negative_slope=0.01)
    (7): Dropout(p=0.2, inplace=False)
    (8): Linear(in_features=312, out_features=172, bias=True)
    (9): BatchNorm1d(172, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (10): LeakyReLU(negative_slope=0.01)
```

```
    (11): Dropout(p=0.2, inplace=False)
    (12): Linear(in_features=172, out_features=102, bias=True)
    (13): BatchNorm1d(102, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (14): LeakyReLU(negative_slope=0.01)
    (15): Dropout(p=0.2, inplace=False)
    (16): Linear(in_features=102, out_features=32, bias=True)
    (17): Softmax(dim=None)
  )
)
```

D.3 3D-CNN parameters and hyperparameters

Arguments for model:

```
'in_channels': (1, 2, 4),
'out_channels': (2, 4, 8),
'input_size': (6, 6, 256, 1),
'kernel_size': (1, 1, 6),
'pool_size': (1, 1, 2),
'stride': 2,
'dilation': (1, 1, 1),
'padding': (1, 1, 1),
'dropout': 0.2
```

Model architecture:

```
CNN3d(
  (cnn1): Conv3d(1, 2, kernel_size=(1, 1, 6), stride=(2, 2, 2), padding=(1, 1, 1))
  (bn1): BatchNorm3d(2, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
  (leakyrelu1): LeakyReLU(negative_slope=0.01)
  (maxpool1): MaxPool3d(kernel_size=(1, 1, 2), stride=(1, 1, 2), padding=0, dilation=1, ceil_mode=False)
  (cnn2): Conv3d(2, 4, kernel_size=(1, 1, 6), stride=(2, 2, 2), padding=(1, 1, 1))
  (bn2): BatchNorm3d(4, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
  (leakyrelu2): LeakyReLU(negative_slope=0.01)
  (maxpool2): MaxPool3d(kernel_size=(1, 1, 2), stride=(1, 1, 2), padding=0, dilation=1, ceil_mode=False)
  (cnn3): Conv3d(4, 8, kernel_size=(1, 1, 6), stride=(2, 2, 2), padding=(1, 1, 1))
  (bn3): BatchNorm3d(8, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
  (leakyrelu3): LeakyReLU(negative_slope=0.01)
  (maxpool3): MaxPool3d(kernel_size=(1, 1, 2), stride=(1, 1, 2), padding=0, dilation=1, ceil_mode=False)
  (dropout): Dropout(p=0.2, inplace=False)
  (feedforward): Sequential(
    (0): Linear(in_features=3312, out_features=1024, bias=True)
    (1): LeakyReLU(negative_slope=0.01)
    (2): Dropout(p=0.2, inplace=False)
    (3): Linear(in_features=1024, out_features=512, bias=True)
    (4): LeakyReLU(negative_slope=0.01)
    (5): Dropout(p=0.2, inplace=False)
    (6): Linear(in_features=512, out_features=256, bias=True)
    (7): LeakyReLU(negative_slope=0.01)
    (8): Dropout(p=0.2, inplace=False)
    (9): Linear(in_features=256, out_features=128, bias=True)
    (10): LeakyReLU(negative_slope=0.01)
    (11): Dropout(p=0.2, inplace=False)
    (12): Linear(in_features=128, out_features=64, bias=True)
    (13): LeakyReLU(negative_slope=0.01)
    (14): Dropout(p=0.2, inplace=False)
    (15): Linear(in_features=64, out_features=32, bias=True)
    (16): LeakyReLU(negative_slope=0.01)
    (17): Dropout(p=0.2, inplace=False)
    (18): Linear(in_features=32, out_features=16, bias=True)
    (19): LeakyReLU(negative_slope=0.01)
    (20): Dropout(p=0.2, inplace=False)
```

```
    (21): Linear(in_features=16, out_features=8, bias=True)
    (22): LeakyReLU(negative_slope=0.01)
    (23): Dropout(p=0.2, inplace=False)
    (24): Linear(in_features=8, out_features=4, bias=True)
  )
)
```

D.4 CNN+LSTM parameters and hyperparameters

Arguments for model:

```
'in_channels': (6, 12),
'out_channels': (12, 24),
'input_size': (256, 6, 8, 16),
'kernel_size': (1, 6),
'padding': (1, 3),
'layers': 3,
'dropout': 0.2
```

Model architecture:

```
CNNLSTM(
  (conv1): Conv2d(6, 12, kernel_size=(1, 6), stride=(1, 1), padding=(1, 3))
  (conv2): Conv2d(12, 24, kernel_size=(1, 6), stride=(1, 1), padding=(1, 3))
  (conv2_drop): Dropout2d(p=0.2, inplace=False)
  (rnn): LSTM(72, 48, num_layers=3, batch_first=True)
  (feedforward): Sequential(
    (0): Linear(in_features=48, out_features=32, bias=True)
    (1): LeakyReLU(negative_slope=0.01)
    (2): Dropout(p=0.2, inplace=False)
    (3): Linear(in_features=32, out_features=16, bias=True)
    (4): LeakyReLU(negative_slope=0.01)
    (5): Dropout(p=0.2, inplace=False)
    (6): Linear(in_features=16, out_features=8, bias=True)
    (7): LeakyReLU(negative_slope=0.01)
    (8): Dropout(p=0.2, inplace=False)
    (9): Linear(in_features=8, out_features=4, bias=True)
  )
)
```

E Visualisation of raw signal data for each channel

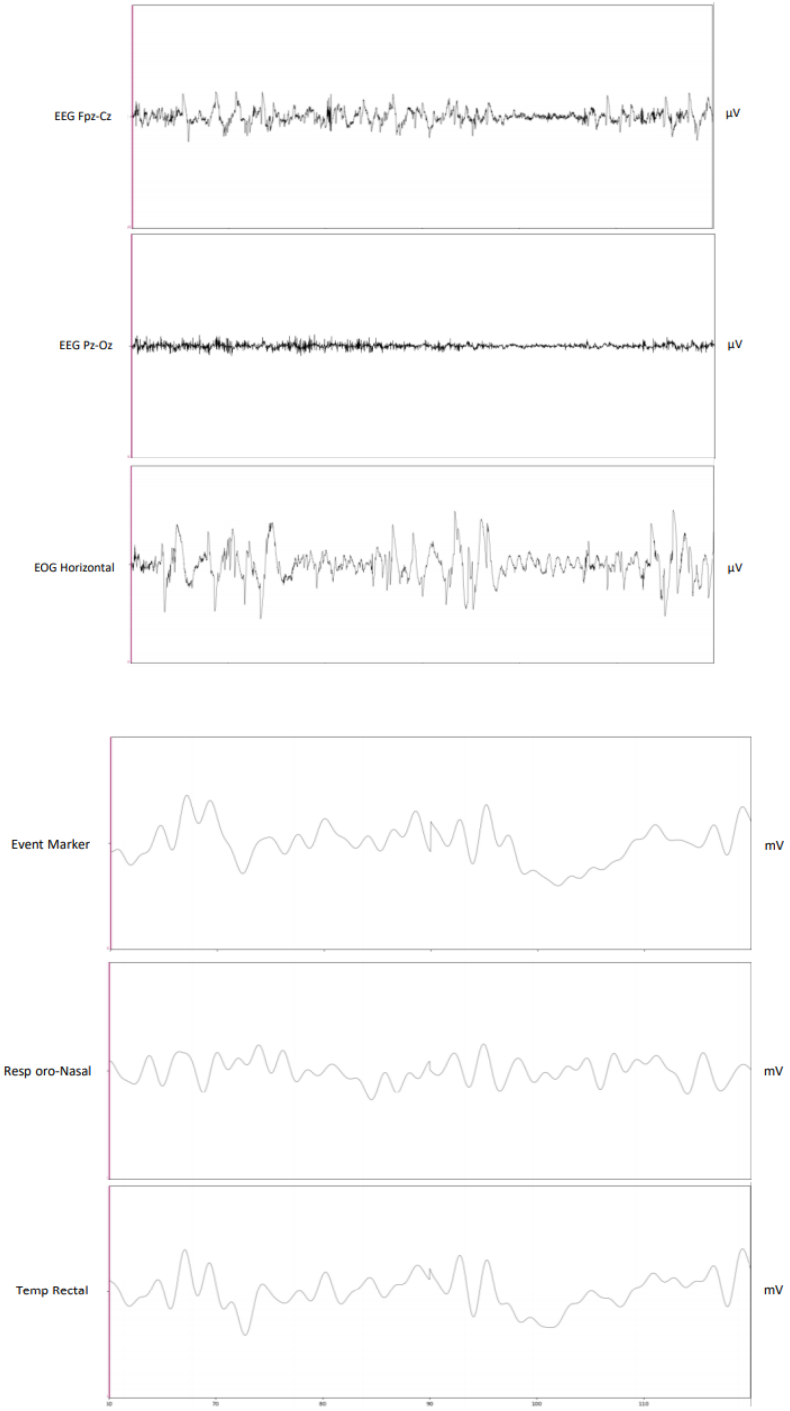


Figure 22: A diagram of the 6 raw input data channels labelled. Note the difference in scale between the top three (μV) and the bottom three (mV).

References

Åkerstedt, T., Fredlund, P., Gillberg, M. and Jansson, B., 2002. A prospective study of fatal occupational accidents–relationship to sleeping difficulties and occupational factors. *Journal of sleep research*, 11(1), pp.69–71.

Berthomier, C., Drouot, X., Herman-Stoïca, M., Berthomier, P., Prado, J., Bokar-Thire, D., Benoit, O., Mattout, J. and d'Ortho, M.P., 2007. Automatic analysis of single-channel sleep eeg: validation in healthy individuals. *Sleep*, 30(11), pp.1587–1595.

Deuschl, G., 1999. Recommendations for the practice of clinical neurophysiology. *Guidelines of the International Federation of Clinical Neurophysiology*.

Gao, Y., Gao, B., Chen, Q., Liu, J. and Zhang, Y., 2020. Deep convolutional neural network-based epileptic electroencephalogram (eeg) signal classification. *Frontiers in neurology*, 11.

Heslop, P., Smith, G.D., Metcalfe, C., Macleod, J. and Hart, C., 2002. Sleep duration and mortality: the effect of short or long sleep duration on cardiovascular and all-cause mortality in working men and women. *Sleep medicine*, 3(4), pp.305–314.

Horne, J.A. and Reyner, L.A., 1995. Sleep related vehicle accidents. *Bmj*, 310(6979), pp.565–567.

Iber, C., 2007. The AASM manual for the scoring of sleep and associated events: Rules. *Terminology and technical specification*.

Irwin, M., McClintick, J., Costlow, C., Fortner, M., White, J. and Gillin, J.C., 1996. Partial night sleep deprivation reduces natural killer and celhdar immune responses in humans. *The FASEB journal*, 10(5), pp.643–653.

Kemp, B., 1987. Model-based monitoring of human sleep stages. *Thesis from University of Twente*.

Kemp, B., Zwinderman, A., Tuk, B., Kamphuisen, H. and Oberye, J., 2000. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *Ieee transactions on biomedical engineering*, 47(9), pp.1185–1194. Available from: https://doi.org/10.1109/10.867928.

Landrigan, C.P., Rothschild, J.M., Cronin, J.W., Kaushal, R., Burdick, E., Katz, J.T., Lilly, C.M., Stone, P.H., Lockley, S.W., Bates, D.W. et al., 2004. Effect of reducing interns' work hours on serious medical errors in intensive care units. *New england journal of medicine*, 351(18), pp.1838–1848.

Loshchilov, I. and Hutter, F., 2017. Decoupled weight decay regularization. *arxiv preprint arxiv:1711.05101*.

Lun, X., Yu, Z., Chen, T., Wang, F. and Hou, Y., 2020. A simplified cnn classification method for mi-eeg via the electrode pairs signals. *Frontiers in human neuroscience*, 14.

Memar, P. and Faradji, F., 2017. A novel multi-class eeg-based sleep stage classification system. *Ieee transactions on neural systems and rehabilitation engineering*, 26(1), pp.84–95.

Mourtazaev, M., Kemp, B., Zwinderman, A. and Kamphuisen, H., 1995. Age and gender affect different characteristics of slow waves in the sleep eeg. *Sleep*, 18(7), pp.557–564.

Ohayon, M.M., Carskadon, M.A., Guilleminault, C. and Vitiello, M.V., 2004. Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: developing normative sleep values across the human lifespan. *Sleep*, 27(7), pp.1255–1273.

Ohtsu, T., Kaneita, Y., Aritake, S., Mishima, K., Uchiyama, M., Akashiba, T., Uchimura, N., Nakaji, S., Munezawa, T., Kokaze, A. et al., 2013. A cross-sectional study of the association between working hours and sleep duration among the Japanese working population. *Journal of occupational health*, 55(4), pp.307–311.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. and Chintala, S., 2019. Pytorch: An imperative style, high-performance deep learning library. In: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox and R. Garnett, eds. *Advances in neural information processing systems 32*. Curran Associates, Inc., pp.8024–8035. Available from: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Rechtschaffen, A., 1968. A manual for standardized terminology, techniques and scoring system for sleep stages in human subjects. *Brain information service*.

Redline, S., Yenokyan, G., Gottlieb, D.J., Shahar, E., O'Connor, G.T., Resnick, H.E., Diener-West, M., Sanders, M.H., Wolf, P.A., Geraghty, E.M. et al., 2010. Obstructive sleep apnea–hypopnea and incident stroke: the sleep heart health study. *American journal of respiratory and critical care medicine*, 182(2), pp.269–277.

Ronzhina, M., Janoušek, O., Kolářová, J., Nováková, M., Honzík, P. and Provazník, I., 2012. Sleep scoring using artificial neural networks. *Sleep medicine reviews*, 16(3), pp.251–263.

Schirrmeister, R.T., Springenberg, J.T., Fiederer, L.D.J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W. and Ball, T., 2017. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11), pp.5391–5420.

Shimada, T., Shiina, T. and Saito, Y., 2000. Detection of characteristic waves of sleep eeg by neural network analysis. *Ieee transactions on biomedical engineering*, 47(3), pp.369–379.

Stylianou, V., 2021. *Deep learning for time series classification (inception time)*. Available from: https://towardsdatascience.com/deep-learning-for-time-series-classification-inceptiontime-245703f422db [Accessed 2021-05-07].

Tagluk, M.E., Sezgin, N. and Akin, M., 2010. Estimation of sleep stages by an artificial neural network employing eeg, emg and eog. *Journal of medical systems*, 34(4), pp.717–725.

Tufik, S., Santos-Silva, R., Taddei, J.A. and Bittencourt, L.R.A., 2010. Obstructive sleep apnea syndrome in the Sao Paulo epidemiologic sleep study. *Sleep medicine*, 11(5), pp.441–446.

Vaughn, B.V. and Giallanza, P., 2008. Technical review of polysomnography. *Chest*, 134(6), pp.1310–1319.

Walker, M., 2017. *Why we sleep: Unlocking the power of sleep and dreams*. Simon and Schuster.

Zeiler, M.D. and Fergus, R., 2014. Visualizing and understanding convolutional networks. *European conference on computer vision*. Springer, pp.818–833.