北京郵電大學 Beijing University of Posts and Telecommunications

Queen Mary University of London

UNIVERSIDAD POLITÉCNICA DE VALENCIA

Undergraduate Project Report
2020/21

Dimension Reduction Methods Applied to Sleep Stage Analysis

Name:	Zekuan Liu
School:	International School
Class:	2017215120
QM Student No.:	171043570
BUPT Student No.:	2017213176
Programme:	Internet of Things Engineering

Date: 03-05-2021

Table of Contents

Abstract

Sleep is critical to the human body, but sleep disorders bother people frequently in modern daily life. To diagnose sleep disorders, the proportion between sleep stage duration as well as the content of physiological signals are used. A sleep cycle consists of several stages: wake, rapid eye movement (REM), stage 1, stage 2, stage 3, and stage 4. The sleep cycle is repeated several times overnight. Recently, Polysomnography (PSG) is widely used in sleep stage analysis and further address sleep disorders. It records a wide range of physiological signals at the same time, e.g., Electroencephalography (EEG), Electrocardiography (ECG), blood oxygenation, airflow, etc. However, multiply measurements in PSG add data redundancy, and features extracted from physiological data can be very high dimensional compared to the number of data records. In order to tackle this problem, this project has applied dimension reduction methods to high dimensional data, thereby reducing data redundancy, saving computational resource and preventing overfitting. In addition, the complete procedure for sleep stage analysis, including data pre-processing, feature extraction, and downstream classification are implemented as well. Experiments demonstrated the effectiveness of the sleep staging procedure and verified the necessity and superiority of dimension reduction under the sleep stage analysis scenario.

摘要

睡眠对人体非常重要，但是在现代生活中人们经常被睡眠障所困扰，而不同睡眠阶段持续时间的比例与生理信号本身可被用于诊断睡眠障碍。一个睡眠周期包括以下几个阶段：清醒期，快速眼动期（REM），睡眠 1 期（S1），睡眠 2 期（S2），睡眠 3 期（S3）和睡眠 4 期（S4）。睡眠周期会在一晚的睡眠中循环多次。近年来，多导睡眠监测仪被广泛用于睡眠分期并进一步诊断睡眠障碍。多导睡眠监测仪可以同时记录多种生理信号，例如：脑电图（EEG），心电图（ECG），血氧饱和度，呼吸气流等。但多导睡眠监测仪中测量的多种信号会增加数据冗余，且相比于生理信号本身从信号中提取的特征可能有非常高的维度。该项目为解决此问题，将降维方法应用于高维数据，以减少数据冗余，节省计算资源并防止过拟合。此外，该项目还实现了完整的睡眠分期程序，包括数据预处理，特征提取和下游分类。实验证明了睡眠分期程序的有效性，并验证了在睡眠分期场景中降维的必要性和优越性。

Chapter 1: Introduction

1.1 Motivation

Sleep is important for the human body. It is a prerequisite for both physical and mental health. During sleep, the body will protect the metabolizable energy, mature the neuronal connections, and consolidate learning and memory (Faust et al., 2019). However, due to the quickened life rhythm and unhealthy lifestyle, sleep disorders become an increasingly frequent problem in modern daily life. Demographic research reveals that up to 24% of people are faced with regular sleep problems (Willemen et al., 2014), due to sleep apnea, insomnia, hypersomnia, etc. Sleep disorders associated with not only disturbances in cognitive and psychological function, but also increased morbidity and mortality (Hillman and Lack, 2013). For instance, obstructive sleep apnea syndrome (OSAS) has a significant side effect of increased risk of cardiovascular diseases (Boostani et al., 2016). In addition, adverse effects of sleep disorders threaten people's safety, productivity, and quality of life. A study shows that drowsy driving is a key factor in about 100,000 traffic accidents occurring each year in the United States, resulting in thousands of deaths and injuries (Stutts et al., 2003)A French research found that employees with sleep problems missed as twice many as workdays in a year compared to normal sleepers (Faust et al., 2019). Baldwin et al. found that subjective sleep symptoms are comprehensively associated with poorer quality of life (Baldwin et al., 2001).

Nowadays, overnight Polysomnography (PSG) is considered the gold standard in sleep research and allows accurate assessment of sleep. It is widely adopted in sleep stage analysis and further sleep disorders diagnosis (Agarwal and Gotman, 2001; Sulistyo et al., 2017). PSG is a multi-parametric measurement equipment that records a wide range of physiological signals at the same time, including Electroencephalography (EEG), Electrocardiography (ECG), Electrooculography (EOG), Electromyography (EMG), airflow, blood oxygenation, respiratory effort, etc. In conventional systems, PSG sleep stage analysis is carried out manually by experts. However, more and more computer automatic sleep stage analysis systems have replaced humans to perform this tiring and time-consuming task in recent years. In addition to saving the cost of labour, computer systems can help to reduce intra- and inter-observer variability, and further improve the quality of the analysis (Faust et al., 2019). Despite the superiorities, problems still exist. One of the problems is that multiple measurements add data redundancy, i.e., the ECG merely confirms the information already extracted from an EEG signal. Similarly,

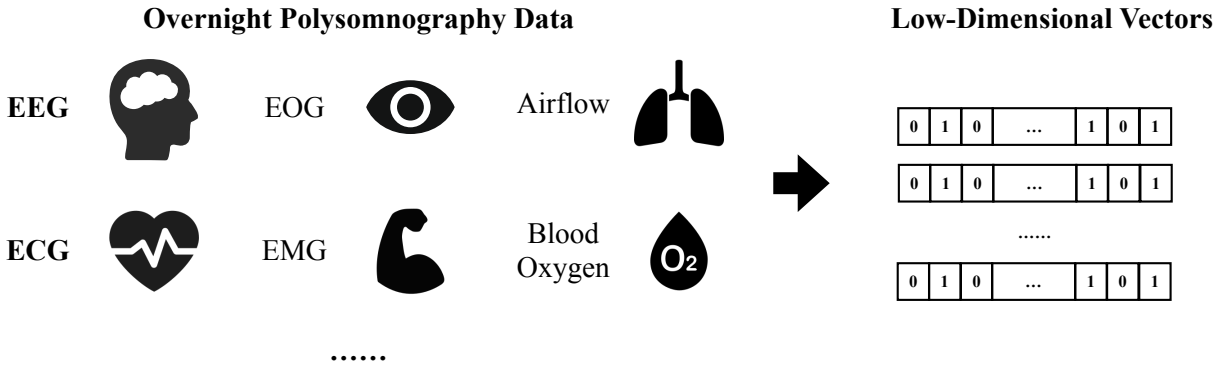


Figure 1 Dimension Reduction on PSG Data

features extracted from physiological data are redundant as well. Therefore, as Figure 1 shows, this project aims to apply dimension reduction methods in computer-based automatic sleep stage analysis, in order to reduce data redundancy, save computational resource, and avoids the issue of over-fitting. Specifically, this project only focused on EEG and ECG of polysomnography data. Finally, obtained data will be used in the downstream classification of different sleep stages.

1.2 Overview

Figure 2 shows the overview of this project. The project is divided into four parts: pre-processing, feature extraction, dimension reduction, and classification.

Data pre-processing is required because a lot of interferences and artefacts could affect the record of the raw data. On top of these, signals have to be segmented into regular 30s epochs as units of further classification.

After pre-processing, various features can be extracted for each epoch of obtained cleaned data. For instance, powers in the frequency band of different waves and Hjorth parameters are estimated from EEG data, so that high dimensional feature vectors are obtained.

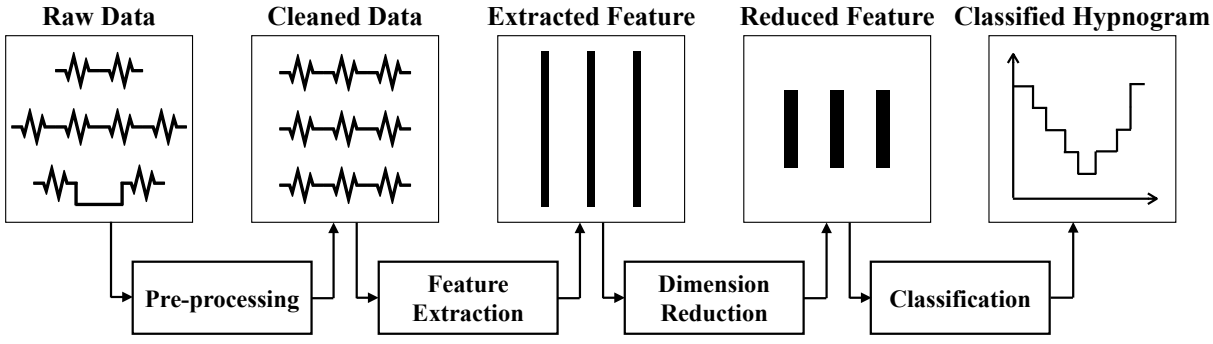


Figure 2 Project Overview

Regarding dimension reduction, the key component of this project, high dimensional vectors are compressed, reducing the size of vectors but preserving most of the information required.

The final step of this project adopted machine learning models to classify overnight sleep into different stages based on the dimension reduced vectors. According to the classification results, overnight hypnograms can be plotted.

1.3 Achievement

In general, the achievements of this project are as follow.

1. Established a pre-processed dataset for sleep stage analysis from raw physiological data.
2. Implemented the feature extraction and dimension reduction processing step of EEG and ECG data.
3. Implemented the classification processing step for various type of classification with multiple classifiers and evaluation matrices.
4. Verified the necessity and superiority of dimension reduction under the sleep stage analysis scenario.
5. Designed a variable weighted PCA based on mutual information, which outweighs conventional PCA in most cases.
6. Proposed an improved representation of hypnogram which is easier to read compared to conventional hypnogram.

Chapter 2: Background

2.1 Polysomnography

Polysomnography (PSG) data is adopted to analyse the sleep stages of patients in this project. PSG is a multi-parametric measurement apparatus that records a wide range of physiological signals simultaneously, Electroencephalogram (EEG), Electrocardiogram (ECG), Electromyogram (EMG), Electrooculogram (EOG), blood oxygenation, respiratory effort, airflow and so on. Despite the variety of data, most of the studies focus on EEG and/or ECG, because sleep is caused by significant changes in brain activities (Faust et al., 2019). Therefore, this project only focused on sleep stage analysis with EEG and ECG signals.

2.1.1 Electroencephalography

Electroencephalography (EEG) signals are recordings of the electrical activity of the human brain. The special device called Electroencephalogram records the signals through electrodes or leads placed on the scalp. Generally, several electrodes are arranged according to standards, but in this project, to simplify the problem, only one channel (C3-A2) of EEG signal is analysed. EEG patterns show different characteristics during sleep stages. EEG waveforms are classified based on their frequency.

Delta wave (0.1 – 4 Hz) is the slowest wave but has the highest amplitude. It is related to the grey matter in the brain. It can be observed in all sleep stages especially in stage 3 and 4, but not in the stages of awake.

Theta wave (5 – 7 Hz) is associated with subconscious activities. It is found in deep relation and meditation. It can increase in sleep stage 1.

Alpha wave (8 – 12 Hz) is observed in awake but relaxed adults. It represents white matter activities of the brain and serves as a bridge between the subconscious and conscious mind.

Sigma wave ranges from 13 to 15 Hz. Sleep stage 2 is characterized by sleep spindles, i.e., transient runs of rhythmic activity in sigma wave.

Beta wave (16 – 30 Hz) is regarding action and behaviour. It is commonly observed during conscious behaviours (e.g., speaking, thinking, and decision making).

Gamma wave (30 – 100 Hz) occurs during the integration of sensory inputs and hyper-alertness. It seldom occurs in normal sleep (Kumar and Bhuvaneswari, 2012).

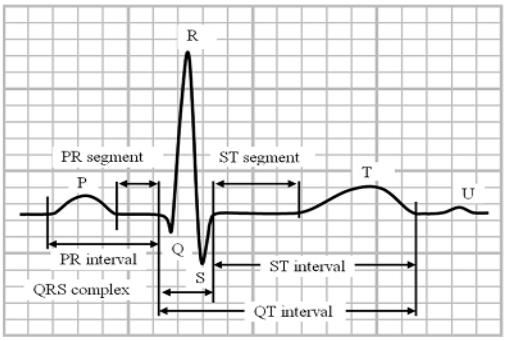


Figure 3 ECG Waveform

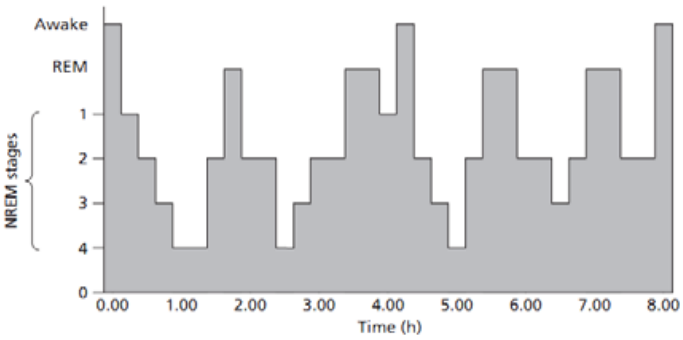


Figure 4 Hypnogram

2.1.2 Electrocardiography

Electrocardiography (ECG) is a recording of the electrical activity of the heart. Similar to EEG signal, single-lead ECG is applied. ECG signals of healthy human are highly structured, and each signal component can be identified via visual inspection (Noviyanto et al., 2011). ECG waveform and its attributes are shown in Figure 3. Sleep stages are also indicated in subtle changes in ECG signals. For instance, the absolute voltage for the P, Q, R, S and T points of the ECG signal in the stage of awake are greater than the stage of sleep. Additionally, the number of samples in the P-QRST interval in the stage of awake is lesser than the stage of sleep. Furthermore, the total amount of R peaks in the constant time interval is also distinguishing between the two stages (Yücelbaş et al., 2018). These difference in the physiological signal provides the foundation of sleep stage analysis in this project.

2.2 Sleep Stage Analysis

Sleep specialists analyse sleep stages following well-established guideline (Berry et al., 2015). Overnight sleep is scored in 30-second epochs. Each epoch is labelled as either wakefulness, rapid eye movement (REM) sleep or one of four stages during non-rapid eye movement (NREM) sleep, including stage 1 (S1), stage 2 (S2), stage 3 (S3), and stage (S4). In some scenarios, stage 1 and stage 2 were merged to Light Sleep (LS), while stage 3 and stage 4 were merged to stage 3, also known as Slow Wave Sleep (SWS) (Faust et al., 2019). The scoring result plotted in temporal consequence is a hypnogram. Figure 4 provides an example of hypnogram (Noviyanto et al., 2011). In this project, sleep stage analysis is completed automatically with computer programs.

Table 1 Signal length of subjects in 30-second epochs

Subject	Epoch	Subject	Epoch	Subject	Epoch	Subject	Epoch	Subject	Epoch
1	748	6	768	11	811	16	852	21	908
2	882	7	925	12	774	17	752	22	711
3	826	8	907	13	916	18	913	23	838
4	808	9	900	14	789	19	787	24	893
5	813	10	864	15	822	20	861	25	721

2.3 Dataset

This project adopted St. Vincent's University Hospital / University College Dublin Sleep Apnea Dataset, which is publicly accessible in PhysioNet (Goldberger et al., 2000). The dataset contains twenty-five overnight polysomnograms with synchronised three-channel Holter ECG, from adult subjects with suspected sleep-disordered breathing.

Subjects were arbitrarily selected over 6 months (from Sept. 2nd to Feb. 3rd) from patients referred to the Sleep Disorders Clinic at St Vincent's University Hospital, Dublin, for potential diagnosis of central sleep apnea, obstructive sleep apnea, or primary snoring. Subjects had to be over 18 years of age, without known autonomic dysfunction and cardiac disease, and not on medication known to interfere with heart rate. Twenty-five subjects (twenty-one males and four females) were selected (age: 50 ± 10 years, range 28 - 68 years; AHI: 24.1 ± 20.3, range 1.7 - 90.9; BMI: 31.6 ± 4.0 kg/m², range 25.1 - 42.5 kg/m²). The signal length of each subject in 30-

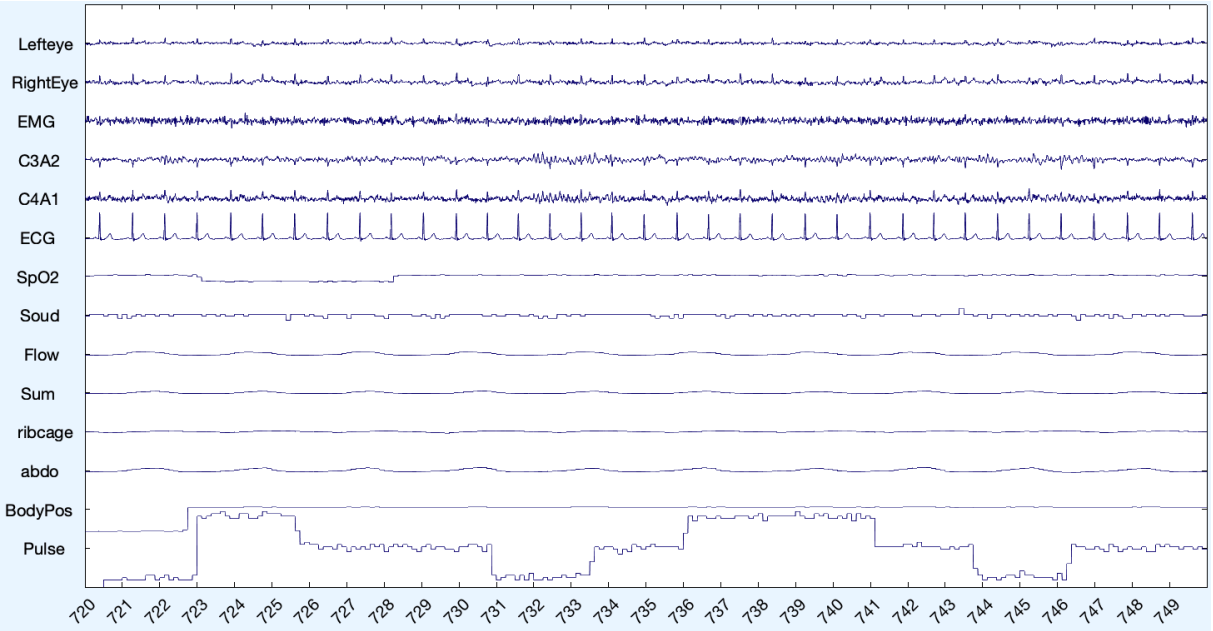


Figure 5 A 30-second sample of the PSG dataset

second epochs is shown in Table 1.

Polysomnograms were acquired with the Jaeger-Toennies system (Erich Jaeger GmbH, Germany). Figure 5 shows a piece of the dataset. Signals recorded were EEG (C4-A1), EEG (C3-A2), left and right EOG, submental EMG, ECG (modified lead V2), oronasal airflow (thermistor), ribcage movements, abdomen movements (uncalibrated strain gauges), oxygen saturation (finger pulse oximeter), snoring (tracheal microphone) and body position. The files are stored in EDF format. In this project, only EEG (C3-A2) and ECG signal are used, and the labels were scored by a skilled sleep technologist according to standard Rechtschaffen and Kales rules. The original dataset is open access at https://doi.org/10.13026/C26C7D.

2.4 Related Work

This project aims at computer-based sleep stage analysis, which stages the sleep with physiological data automatically. In recent years, numerous researches and experiments have published on this task. For instance, Hassan and Bhuiyan have decomposed EEG signals and built a sleep classification system using the RUSBoost classifier (Hassan and Bhuiyan, 2017). The accuracy has been improved in (Bajaj and Pachori, 2013) with an alternative SVM classifier, and Mousavi et al. proposed a model to classify sleep stages further enhanced the performance with the introduction of deep learning (Michielli et al., 2019). Although most of the works were concerning with EEG signals for the reason that changes in brain activities are the origin of sleep and sleep stages, ECG signals are also widely adopted by many researchers because they pick up sleep-related changes in the automatic nervous system and relatively easy to collect compared to EEG signals (Faust et al., 2019). Yücelbaş et al. applied ECG signals to score the sleep (Yücelbaş et al., 2018). Mendez et al. have implemented sleep stage analysis based on heart rate variability (HRV), which is even simpler and cheaper to access (Mendez et al., 2010). In addition to EEG and ECG signals, other signals, e.g., EOG and respiratory effort, are adopted in sleep stage analysis as well (Long et al., 2014; Rahman et al., 2018). Some researchers use multiple types of data at the same time. Phan et al. used multi-modal learning on both EEG and EOG data for sleep staging (Phan et al., 2019). However, most of these works have not focused on dimension reduction.

Chapter 3: Design and Implementation

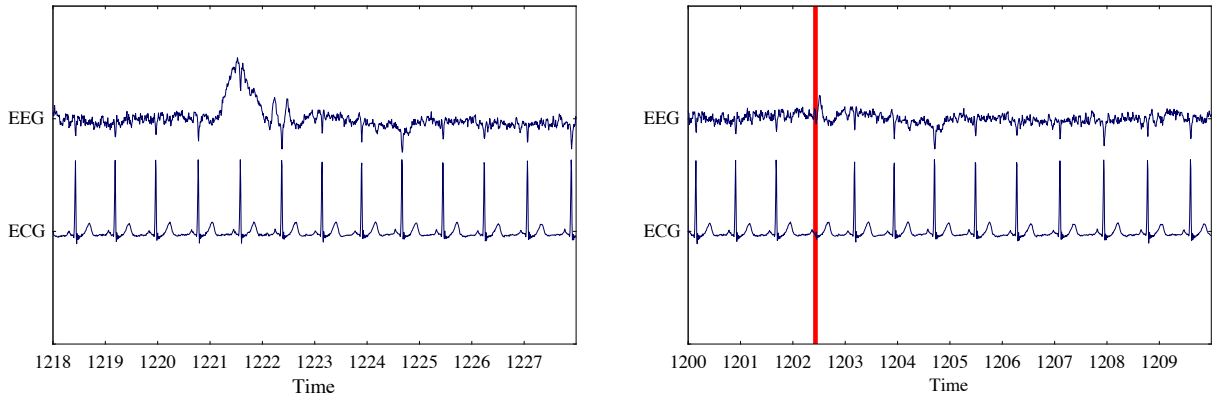
The program in this project is implemented in MATLAB R2021a under macOS Big Sur, with Statistics and Machine Learning Toolbox, Bioinformatics Toolbox, Signal Processing Toolbox, Wavelet Toolbox and DSP System Toolbox. The code is available at https://github.com/kayzliu/DRSleep.

3.1 Pre-processing

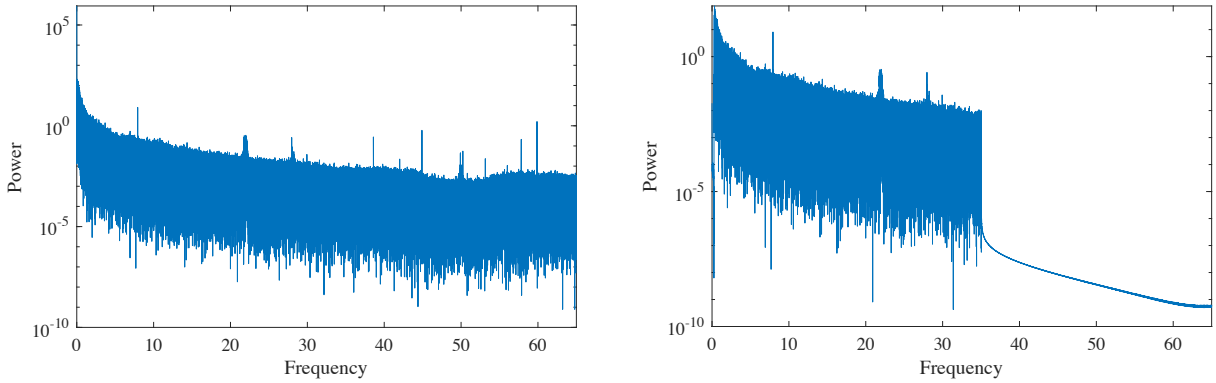
Electroencephalogram (EEG) and electrocardiogram (ECG) are recorded by sensitive electrodes. Due to the complication of the sensing environment, artefacts are inevitable. Basically, there are two types of artefacts, biological artefacts, and environmental artefacts. As the EEG electrodes are placed on the scalp, it will not only record brain neuron electrical activities but also those from eyes (ocular artefacts), muscles (muscle artefacts), even heart (ECG artefacts). These are called biological artefacts. On the other hand, environmental artefacts refer to the artefacts that originated outside of the body, e.g., body movement, 50 or 60 Hz artefacts from the power supply, etc. These artefacts may lead to significant variation to the signal (Motamedi-Fakhr et al., 2014). To reduce the influence of artefacts, three methods can be used in pre-processing of the raw data, rejecting bad data, filtering, and Independent Component Analysis (ICA). In addition to the pre-processing stated above, other commonly used pre-processing steps are also performed, including normalisation, calibration, etc. Furthermore, the signals are required to be divided into uniform 30-second epochs for the later procedure. In this project, EEGLAB (a toolbox developed by UCSD) is applied to implemented part of pre-processing, including selecting channels, removing DC offset, filtering, rejecting bad segment, etc (Delorme and Makeig, 2004).

3.1.1 Removing Bad Data

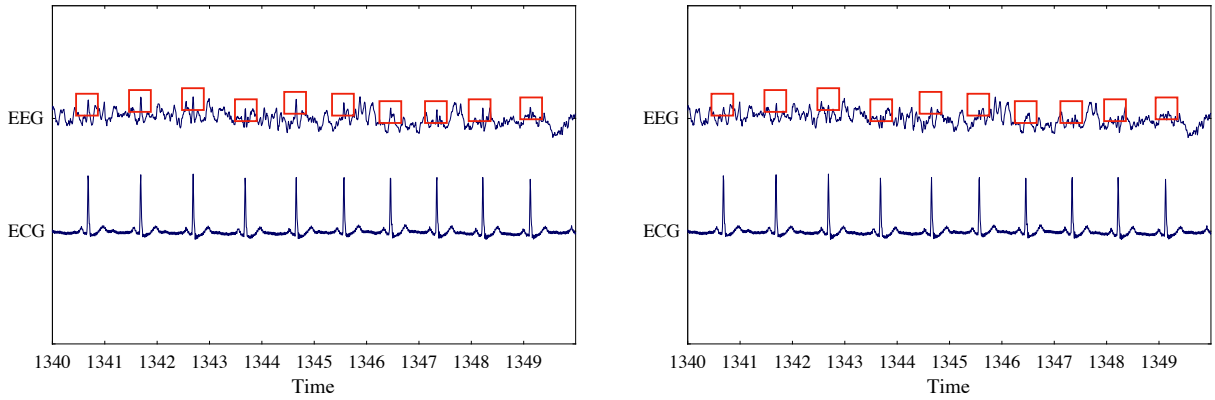
Due to various artefacts, there could be short-time bursts in the recorded signals. For instance, as Figure 6 (a) Left shows, there is an abnormally large value between 1221 second and 1222 second in the EEG channel. To prevent the negative effect on the later procedure, the standard deviation within a 0.5-second window is estimated. When the estimated result exceeds the threshold (20 in this case), the corresponding signal segment will be considered as a bad segment and removed from all channels. As Figure 6 (a) Right shows, the segment is removed



(a) Removing Bad Data



(b) Filtering



(c) Independent Component Analysis

Figure 6 Signal comparison between before and after pre-processing

and replaced with a red vertical line in its original position. Similarly, flatlines, noise and low-frequency drifts can also be removed by this mean.

3.1.2 Filtering

Although the method used in the last step can remove most of the artefacts, some artefacts remain, e.g., alternating current power supply, which has a small volume but continuously exists

during the record of the signal. Hence, filtering is used to further reduce the artefacts. As for EEG signal, a band-pass filter (0.3 – 35 Hz) is used. As Figure 6 (b) Right shows, the high-frequency components and direct current component in Figure 6 (b) Left have been removed after filtering. Similarly, regarding ECG data, a high-pass filter of 0.3 Hz is applied.

3.1.3 Independent Component Analysis

Another approach that can be conducted is Independent Component Analysis (ICA) which uses the multi-channel property of signals to decompose the dataset into several components which are maximally independent. The electromagnetic property of EEG and ECG signals indicates that these signals satisfy the assumptions of the ICA algorithm, i.e., ones based on an instantaneous mixture model (Motamedi-Fakhr et al., 2014). In addition, a variant sequential independent component analysis, which takes temporal dependence into account can be used to improve the performance (Safont et al., 2014). The crucial application of ICA is the suppression of other electromagnetic characters (e.g., ECG) from EEG signal. As Figure 6 (c) shows, the impulses in red boxes of EEG signal in Figure 6 (c) Left, which are result from ECG signal, are diminished in Figure 6 (c) Right after ICA. The red boxes are in the exact same position in the two figures. However, the main problem of ICA is the lack of automatic and reliable identification of different components, i.e., which component is related to EEG and which component is related to ECG. The incorrect identification can result in an inconsistency problem of the data.

3.1.4 Segmentation

The non-stationary property of PSG signals is not compatible with lots of signal processing algorithms, which suppose the processing signals are stationary. This problem is met in plenty of applications of signal processing and a widely used solution is to segment the signals into small epochs in the time domain such that they can be considered as being nearly stationary (a so-called assumption of quasi-stationarity) within each signal (Motamedi-Fakhr et al., 2014). In this project, the dataset also labelled the data based on 30-second epochs. Hence, rather more practically the study of PSG signals needs to identify the approximate temporal location or time range of events, as a result, naturally any processing scheme uses data corresponding to a certain finite duration window.

3.2 Feature Extraction

Features are parameters that provide information about the underlying structure of the signal of interest. Feature extraction is an important step in the whole procedure, which can significantly affect the final classification outcome. There are different methods to extract features from ECG and EEG signals respectively. As for the ECG signal, autoregressive (AR) coefficients, Shannon entropy, wavelet leader estimates are applied. Regarding the EEG data, both temporal features and spectral features are used. Specifically, frequency domain features are the powers in the frequency band of delta wave, theta wave, alpha wave, sigma wave, and beta wave. On the top of that, Hjorth parameters are estimated in the time domain, including activity, mobility, and complexity.

3.2.1 Autoregressive Coefficient

Autoregressive coefficients, also known as reflection coefficients, are obtained from the autoregressive (AR) model. In an AR model of order p, the output of current is a linear combination of the outputs of past p stages plus a white noise input. The weights on the p past outputs minimize the mean squared prediction error of the autoregression. Its mathematical formula is:

$$y(n) + \sum_{k=1}^{p} a(k)y(n-k) = x(n) \tag{1}$$

where $y(n)$ is the current value of the output; $a(k)$ is autocorrelation coefficient and $x(n)$ is a zero-mean white noise input. The reflection coefficients are the partial autocorrelation coefficients scaled by -1. They indicate the time dependence between $y(n)$ and $y(n-k)$ after subtracting the prediction based on the intervening $k-1$ time steps. In this project, the order of the AR model is 4, and Burg's method is used to calculate the coefficients. It estimates the reflection coefficients and uses the reflection coefficients to estimate the AR parameters recursively (Liang et al., 2012).

3.2.2 Wavelet Leader

Discrete Wavelet Transform (DWT) is a linear operator that decomposes the original signal into two components: approximation coefficients (ACs, which are the low frequency, high scale information of the initial signal) and detail coefficients (DCs, which capture the high frequency, low scale information in the original signal). Then, the DCs remain while the ACs are

recursively decomposed into new DCs and ACs. Owing to its great time and frequency localization ability, DWT can reveal the local characteristics of the input ECG signal. In addition, the multi-level decomposition of an ECG signal into different scales by DWT generates multi-scale features, each of which represents particular characteristics of the signal (Li and Zhou, 2016).

Wavelet leaders are derived from the critically sampled discrete wavelet transform (DWT) coefficients. Wavelet leaders offer significant theoretical advantages over wavelet coefficients in the multifractal formalism. Wavelet leaders are time- or space-localized suprema of the absolute value of the discrete wavelet coefficients. The time localization of the suprema requires that the wavelet coefficients are obtained using a compactly supported wavelet. The Holder exponents, which quantify the local regularity, are determined from these suprema. The singularity spectrum indicates the size of the set of Holder exponents in the data (Serrano and Figliola, 2009; Wendt and Abry, 2007).

3.2.3 Shannon Entropy

In this section, Shannon entropies of the wavelet packet coefficients of discrete wavelet packet transform (WPT) of the ECG signal are estimated. Instead of DWT, which is used in the last section, WPT, an extension of DWT is applied. Since DWT decomposes ACs only at each level, it is hard to extract distinctive information from DCs. On the other hand, in the WPT, the filtering operations are applied to not only ACs but also DCs. Therefore, the WPT has the same frequency bandwidths in each resolution while DWT does not. This property makes WPT not increase and lose information compared to original signals, resulting in the features from WPT having more discrimination power than those from DWT. The mathematical representation of WPT is:

$$\begin{cases} d_{0,0}(t) = y(t), \\ d_{i,2j-1}(t) = \sqrt{2}\sum_{k} h(k) d_{i-1,j}(2t-k), \\ d_{i,2j}(t) = \sqrt{2}\sum_{k} g(k) d_{i-1,j}(2t-k), \end{cases} \tag{2}$$

where $h(k)$ and $g(k)$ are high-pass and low-pass filter respectively, and $d_{i,j}$ is the reconstruction coefficients of WPT at the ith level for the jth node.

Although the coefficients by DWT or WPT can reveal the local characteristics of an ECG signal, the number of such coefficients is usually so huge that it is hard to use them as features for

classification directly. Therefore, Shannon entropy, which measures the uncertainty of the information contained in given systems, may derive from these coefficients for better classification. Mathematically, the Shannon entropy *E* is represented as:

$$E = -\sum_i {d_i}^2 \log\left({d_i}^2\right) \tag{3}$$

where d_i is the wavelet packet coefficients, with the convention 0log(0) = 0 (Li and Zhou, 2016).

3.2.4 Power in Frequency Band

Powers in the frequency band of various waves in EEG signal are estimated, including delta wave (0-4 Hz), theta wave (5-7 Hz), alpha wave (8-12 Hz), sigma wave (13-15 Hz) and beta wave (16-30 Hz). As 2.1.1 illustrated, different waves may have different amplitude, during different sleep stages. Thus, their power in the frequency band can reveal the current sleep stage. To calculate the power of different waves, Fast Fourier transform (FFT) is first conducted, to convert the signal from the time domain to the frequency domain. FFT is a rapid discrete Fourier transform algorithm factorizing the transform matrix into a product of sparse factors. Hence, it reduced the computational complexity from $\mathrm{O}(N^2)$ to $\mathrm{O}(N \log N)$, where *N* is the size of the data. The mathematical formula of discrete Fourier transform is:

$$Y_k = \sum_{n=0}^{N-1} y_n e^{-i2\pi kn/N} \tag{4}$$

where *k* equals to 0 to *N*-1; $e^{i2\pi/N}$ is a primitive *N*th root of 1. After the Fourier transform, the power of different waves (different frequency range) can be obtained by:

$$P = \frac{1}{T}\int_b^a |Y(f)|^2 \mathrm{d}f \tag{5}$$

where *T* is the time range (30 seconds); *a* and *b* are upper bound and lower bound of the wave frequency range, respectively.

3.2.5 Hjorth Parameters

The interpretation of EEG data, consisting of a sequence of observed electrical potentials, is complicated by the lack of a sufficient model to explain how states of the central nervous system are reflected in the measured signals. To resolve this problem and effectively represent the EEG

data, Hjorth parameters, including activity, mobility, and complexity, are proposed (Hjorth, 1973).

Activity parameter is the variance of the signal in the time domain, indicating the surface of the power spectrum in the frequency domain, which means the value of activity parameter is proportional to the number of high-frequency components. Its mathematical definition is:

$$activity = var(y(t)) \tag{6}$$

Mobility parameter is described as the square root of the ratio of the variance of the derivative of the signal and that of the signal. It positively proportions to standard deviation of the power spectrum. Its formula is:

$$mobility = \sqrt{\frac{var(y'(t))}{var(y(t))}} \tag{7}$$

Complexity parameter is the shape similarity between a signal and a pure sine wave. Its value converges to 1 as the shape of the signal becomes more similar to a sine wave.

$$complexity = \frac{mobility(y'(t))}{mobility(y(t))} \tag{8}$$

While these three parameters contain information about the frequency spectrum of a signal, they also help analyze signals in the time domain. In addition, they have lower computational complexity.

3.3 Dimension Reduction

Dimension reduction in this project is implemented with principal component analysis (PCA) and weighted PCA.

3.3.1 Principal Component Analysis

Principal component analysis (PCA) is a common linear dimension reduction algorithm. The idea of the PCA is to find the most important components in the data through linear transformation and to sort the obtained principal component vectors according to their explained variances. Based on the required number of components, the first serval components are remained, while other components are discarded. The original high dimensional data is finally transformed into a linearly independent representation of each dimension of the group, revealing the intrinsic structure behind the high dimensional data. The dimensionality reduction

of the data through PCA can achieve the effect of removing noise and redundancy and simplifying the model while maintaining the information of the original data to the greatest extent. Its optimization goal is maximizing the explained variance of the data in low dimensional space. It can be represented in the formula:

$$w = argmax \frac{w^T X^T X w}{w^T w} \tag{9}$$

where the w is the parameter, and the X is the data (Alickovic and Subasi, 2018).

3.3.2 Weighted Principal Component Analysis

There are briefly two types of weighted PCA (WPCA), sample-wise weighted PCA and variable-wise weighted PCA. In order to adapt to the trend and continuously learn the new information, sample-wise weighted PCA adopt moving window, recursing or EWMA (Exponentially Weighted Moving Average) filter and so on (Fan et al., 2011). As for variable-wise weighted PCA, the weightings are determined by a customized formula, with consideration given not to over-weight variables. In this project, the variable-wise weighted PCA is adopted, and mutual information is used as weights while performing the principal components analysis. The weights are defined as:

$$weight_i = I(X_i, Y) = \sum_{j,k} P(X_i = x_j, Y = y_k) \log \frac{P(X_i = x_j, Y = y_k)}{P(X_i = x_j) P(Y = y_k)} \tag{10}$$

where I is the mutual information of feature variable X_i and label Y. As the features are continuous variables. They are discretized into 256 bins or the number of unique values in the variable if it is less than 256. The function finds optimal bivariate bins for each pair of variables using the adaptive algorithm (Darbellay and Vajda, 1999).

3.4 Classification

In the classification stage, the leave-one-out classification based on the 30s epochs is used. First, the data splitting is based on leave-one-out, which means in each experiment one subject is used as the testing set, while others are used as the training set. For instance, in a dataset of 10 subjects, the training set is subject 2 to subject 10, while subject 1 is the testing set; in the next experiment, subject 2 is specified as the testing set, and others (subject 1 and subject 3 to 10) are defined as the training set. Then, according to different classification type, the label of the data has to be modified. In this project, three types of classification are investigated, including

three classification (Wake, REM, and NREM), four classification (Wake, REM, LS, SWS), and full six classification (Wake, REM, S1, S2, S3, and S4). After the specification of the classifier, the relabelled training data can be imported into the classifier and start training. Depending on the classifier, the computing period may vary. Next, the testing dataset is predicted by the trained model. By comparison of predicted values and actual values, evaluation metrics (confusion matrix and average accuracy, in this project) can be obtained. A visualisation of hypnogram can be obtained by plotting the classification result in temporal sequence (sample result can be found in Figure 10 in Chapter 4).

Despite it is critical to select a suitable classifier to achieve better performance, the emphasis of this project is on dimension reduction instead of classification. Therefore, only several classical classifiers are used, and delicate deep learning classifiers are not included. Classification only serves as a downstream task, which is used to evaluate the result of the previous procedure.

3.4.1 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is an algorithm that finds a linear combination of features that characterizes or separates two or more classes of objects. It also frequently used on dimension reduction. It is suitable when the data of different classes are Gaussian distributed and have similar variances. The model predict classifies so as to minimize the expected classification cost:

$$\hat{y} = \underset{y=1\ldots K}{\operatorname{argmin}} \sum_{k=1}^{K} \widehat{P}(k|x) C(y|k), \tag{11}$$

where $\hat{y}$ is the predicted classification; K is the number of classes; $\hat{P}(k|x)$ is the posterior probability of class k for observation x; and $C(y|k)$ is the cost of classifying an observation as y when its true class is k (Singh et al., 2016).

3.4.2 Quadratic Discriminant Analysis

Quadratic Discriminant Analysis (QDA) is similar to LDA. It also assumes the data of different classes are normally distributed, but QDA can tackle various means and covariances, which is more difficult to handle. Unlike LDA, which separates the classes using a linear surface, QDA separates the classes using a quadratic surface (i.e., a conic section) (Singh et al., 2016).

3.4.3 *k*-Nearest Neighbour

k-Nearest Neighbour (*k*NN) is a non-parametric classification method. The predicted class of specific unlabelled sample point is depending on the most common class of its k nearest neighbours (k is a positive integer). If $k = 1$, the sample is simply assigned to the class of the nearest neighbour. The nearest neighbour can be measured in multiple distances, but Euclidean distance is usually used. It is suitable for classification with multiple labels as well as multi-modal classification. However, it is simple and low efficient. The selection of optimal hyperparameter k is also a problem. The algorithm is sensitive to noise and unstable on performance (Singh et al., 2016).

3.4.4 Naïve Bayes Classifier

Naïve Bayes (NB) is a classifier based on Bayes' theorem, assuming strong independence between the features. Its mathematical formula is:

$$\hat{y} = \underset{k=1\ldots K}{\operatorname{argmax}} P(C_k) \prod_{i=1}^{n} P(x_i|C_k) \tag{12}$$

where $\hat{y}$ is the predicted classification; K is the number of classes; C_k is the kth class; and x_i is features of the ith sample. Naïve Bayes classifier is highly scalable. It requires a number of parameters linear in the number of features in a learning problem, taking less computational time (Singh et al., 2016).

3.4.5 Support Vector Machine

Support Vector Machine is a sophisticated algorithm, providing high accuracy with an appropriate kernel. It maps training samples to points in high dimensional space to maximise the gap width between classes and separates different classes with a hyperplane. In the testing stage, new samples are mapped into the same space and categorised based on the region they located. Unlike some algorithms (e.g., *k*-NN), the performance of SVM is independent of the size of the data as well as feature dimension, but the number of training cycles. It provides a high generalisation ability, preventing overfitting problem theoretically. However, it requires relatively longer computational time, and its performance also depends on the parameters (Singh et al., 2016).

Chapter 4: Results and Discussion

To validate the design and implementation described in the last chapter, several comparison experiments are conducted.

4.1 Classifier

According to the last chapter, Linear Discriminant Analysis (LDA), Quadratic Discriminant

Table 2 Comparison result of different classifiers

LDA	Wake	REM	S1	S2	S3	S4
Wake	1107	134	524	78	1	1
REM	216	1322	616	761	11	7
S1	311	246	475	286	1	1
S2	106	210	372	2980	158	125
S3	54	67	91	851	224	434
S4	64	31	62	246	141	950

(a)

QDA	Wake	REM	S1	S2	S3	S4
Wake	533	73	221	99	9	15
REM	503	1434	969	967	13	13
S1	668	193	457	176	1	5
S2	84	214	358	2727	153	139
S3	38	51	66	806	195	310
S4	32	45	69	427	165	1036

(b)

*k*NN	Wake	REM	S1	S2	S3	S4
Wake	826	218	468	191	6	10
REM	228	825	410	455	13	20
S1	500	436	566	612	29	52
S2	267	484	618	3277	274	417
S3	15	15	37	266	59	166
S4	22	32	41	401	155	853

(c)

NB	Wake	REM	S1	S2	S3	S4
Wake	552	57	191	72	6	4
REM	421	1367	802	822	2	1
S1	664	184	511	151	0	0
S2	137	331	534	3584	278	254
S3	35	17	21	129	41	72
S4	49	54	81	444	209	1187

(d)

SVM	Wake	REM	S1	S2	S3	S4
Wake	1080	106	474	95	3	5
REM	198	1069	427	299	0	3
S1	338	273	462	171	2	7
S2	223	539	739	4374	364	492
S3	0	0	0	0	0	0
S4	19	23	38	263	167	1011

(e)

ACC	WRN	WRLS	ALL
LDA	74.1	62.5	53.2
QDA	68.3	55.7	48.1
***k*NN**	71.4	60.0	48.3
NB	71.0	61.9	54.6
SVM	**77.2**	**68.4**	**60.3**

(f)

Analysis (QDA), *k*-Nearest Neighbour (*k*NN), Support Vector Machine (SVM), and Naïve Bayes Classifier (NB) can be applied in downstream classification. However, different classifiers can have diverse performance under various scenarios. The comparison results of different classifiers in the same settings (only EEG data, without dimension reduction) is shown in Table 2. Sub-table (a) to (e) show the confusion matrixes (summation of all subjects) of different classifiers, and sub-table (f) shows the average accuracies (the number of correct epochs divided by the number of all epochs) of classifiers in WRN (Wake, REM, and NREM), WRLS (Wake, REM, LS, SWS), and ALL (Wake, REM, S1, S2, S3, and S4), respectively. Despite SVM cannot distinguish S3, SVM significantly outperforms other classifiers in accuracy in all types of classification. Therefore, SVM is selected for follow-up experiments.

4.2 Pre-processing

In order to verify the effectiveness of the pre-processing, three variants of the procedure are implemented. The first one is the raw data without any pre-processing (RAW). The second one is the data with only filtering (FLT). The third one is the data with only removing bad segments (RMD). Same settings (SVM classifier, only EEG data, without dimension reduction) are used

Table 3 Comparison result of different pre-processing

RAW	Wake	REM	S1	S2	S3	S4
Wake	3097	339	840	323	31	69
REM	415	1172	538	429	6	3
S1	269	294	284	169	0	0
S2	796	1116	1654	5826	543	1360
S3	0	0	0	0	0	0
S4	145	95	87	238	93	558

(a)

FLT	Wake	REM	S1	S2	S3	S4
Wake	3032	354	829	303	23	76
REM	446	1247	547	383	0	0
S1	288	348	278	184	0	0
S2	798	1000	1663	5734	491	867
S3	0	0	0	0	0	0
S4	158	67	86	381	159	1047

(b)

RMD	Wake	REM	S1	S2	S3	S4
Wake	1164	131	535	121	6	6
REM	208	1347	519	379	8	5
S1	370	246	430	179	1	1
S2	257	577	762	4258	374	670
S3	0	0	0	0	0	0
S4	27	35	51	300	147	836

(c)

ACC	WRN	WRLS	ALL
RAW	72.5	61.8	52.6
FLT	73.0	64.2	54.5
RMD	75.4	65.4	57.6
FULL	**77.2**	**68.4**	**60.3**

(d)

in three variants except for the difference in pre-processing. Table 3 shows the final classification results of these variants and fully pre-processed data (FULL). Sub-table (a) to (c) show the confusion matrixes of different pre-processing, and sub-table (d) shows a summary of the average accuracy of each pre-processing. As the confusion matrix of FULL is the same as Table 1 (e), it is not shown in Table 3. According to Table 3, it is obvious that FULL performs better than others in all three classification types, indicating that both pre-processing, filtering and removing bad segments, are necessary.

4.3 Dimension Reduction

In this section, the influence of dimension reduction of different numbers of components is investigated. In order to facilitate the experiment, a pre-processed sub-dataset of 10 subjects is used in the experiments described in this section. Figure 7 shows how explained variance decrease with the reduction of the number of components in three types of data, EEG, ECG and COM (linear combination of EEG and ECG). The blue curve shows the change with conventional PCA, while the orange curve shows the weighted. Although the explained variance monotonically decreases, it changes slightly at the beginning and remains a high value after significantly reducing the number of components, which demonstrated the necessity of dimension reduction. Moreover, the orange curve is always above the blue curve, proving that weighted PCA helps information compact in first serval components better than the conventional one.

Figure 8 describes the relationship between the classification performance and the number of components with different data and different classification types. Sub-figure (a) shows the results of ALL (six classes full classification), sub-figure (b) shows WRLS (Wake, REM, LS, and SWS), and (c) shows WRN (Wake, REM, and NREM). In each sub-figure, the result of EEG, ECG, and COM (linear combination of EEG and ECG) data are shown in the Left, Middle, and Right, respectively. In the figure, blue curves represent the results of PCA, while orange

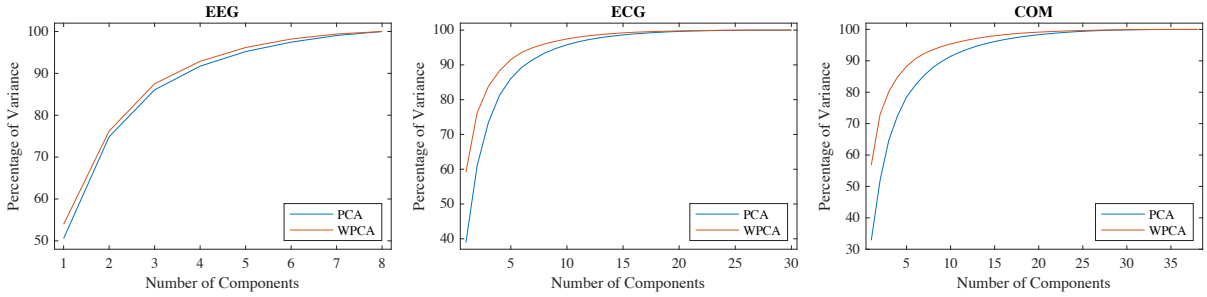


Figure 7 Percentage of variance verse number of components

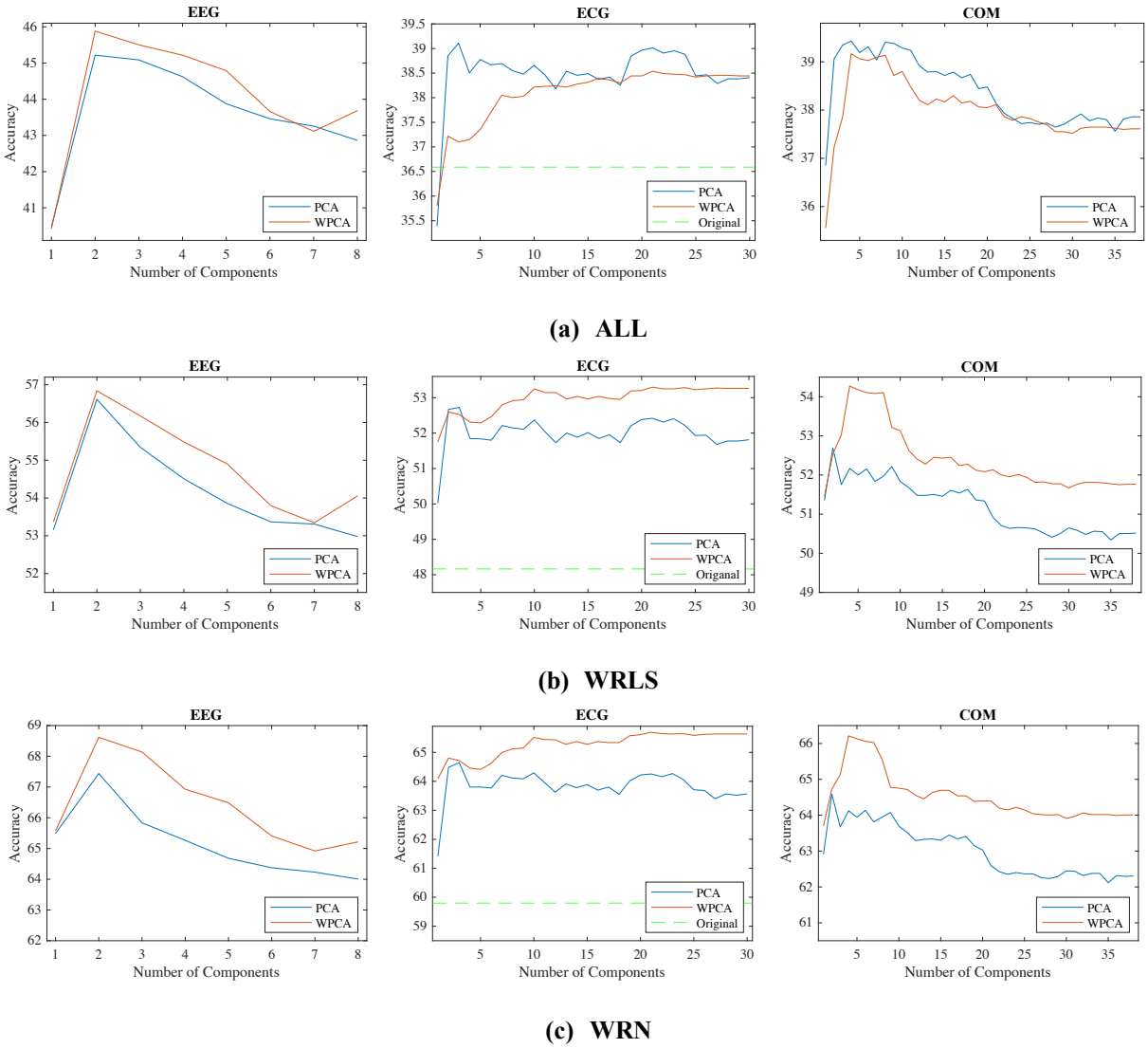


Figure 8 Classification accuracy verse number of components

curves represent WPCA's, and green horizontal lines show the results before dimension reduction. Table 4 shows the comparison optimal results of different data in different classification types.

From Figure 8 and Table 4, several conclusions can be drawn. First, in most curves, the optimal result is achieved with a relatively small number of components instead of maximum. The decrease in the performance with the increasing number of components could result from the noise contained in the latter components. It demonstrates the necessity and superiority of dimension reduction in general circumstances. Second, in the cases of relatively poor performance (e.g., ECG data shown in Figure 8 Middles and bold accuracy in Table 4), dimension reduction helps improve the performance of downstream classification. However, in the circumstances that original data has already achieved a high performance, the accuracy can

Table 4 Comparison of optimal accuracy of different dimension reduction methods

	WRN			WRLS			ALL		
	EEG	**ECG**	**COM**	**EEG**	**ECG**	**COM**	**EEG**	**ECG**	**COM**
Original	75.8	59.8	74.1	69.1	48.2	66.0	60.3	36.6	55.2
PCA	67.4	64.6	64.6	56.6	52.7	52.7	45.2	**39.1**	39.4
WPCA	68.6	**65.7**	66.2	56.8	**53.3**	54.3	45.9	38.5	39.2

drop after dimension reduction. The conclusion further proves the necessity and superiority of dimension reduction in specific scenarios. Third, the weighted PCA (WPCA) designed in this project has improved the classification accuracy in most cases. However, the performance can still decline in the case of both high dimension with many classes (e.g., Middle and Right in sub-figure (a)). In general, the effectiveness of weighted PCA has been verified.

4.4 Hypnogram

Hypnogram is a graph that plot the sleep stage results in a temporal consequence, which allows the different stages of sleep to be identified during the sleep cycle as was explained in Chapter 2. Figure 2 provides a sample of conventional hypnogram, an ideal model, which has perfect cycles of sleep. However, in reality, the sleep cycle of patients may not that predictable. Figure 9 shows the hypnogram of a sample in the dataset. The sleep could frequently switch between two or more stages within a short period of time, and the sleep cycle may not follow the constant sequence. Hence, the hypnogram can be very messy, and hard to extract useful information, especially in those periods of irregular stage switching. Fortunately, it is easy to obtain by analysis that the vertical lines in the hypnogram are the main reason for the confusion, but they seldom contain information related to sleep stages. Therefore, in this project, points, instead of

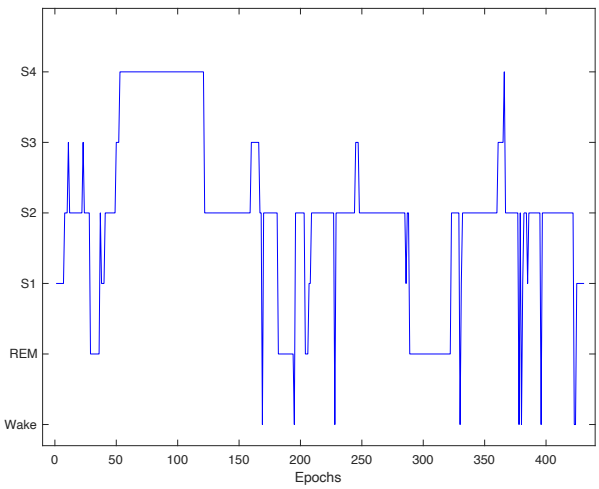


Figure 9 Conventional Hypnogram

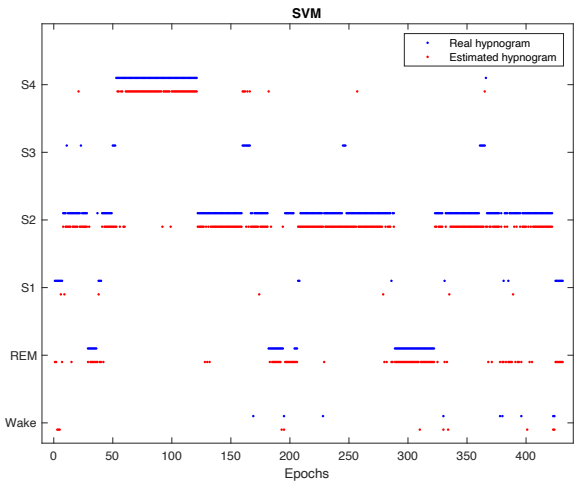


Figure 10 Comparison Hypnogram

lines, are used to represent the sleep stage in a specific time range, and the sample result is shown in Figure 10. Sleep stages are more clearly shown in the figure.

Figure 10 shows the hypnogram of overnight sleep of subject 5. the red points represent the classification result (Estimated hypnogram), while the blue points are ground truth results labelled by an expert (Real hypnogram). The accuracy reaches 77% in this subject. According to the figure, the result is relatively favourable, which is consistent with its accuracy. The estimated hypnogram has basically restored the real hypnogram, but approximate stages which are too hard to distinguish (e.g., S3 and S4) are sometimes confused. The effectiveness of the procedure has been proved.

Chapter 5: Conclusion and Further Work

5.1 Conclusion

In this project, the complete procedure of sleep stage analysis based on PSG data is investigated, including data pre-processing, feature extraction, dimension reduction, and stage classification, emphasizing dimension reduction. A series of experiments demonstrate the achievements of the project. The pre-processing of the dataset is effective, significantly improving the quality of the data and downstream classification performance. The features extracted from the data successfully represented the necessary information in both EEG and ECG signals for staging. Dimension reduction using PCA and designed weighted PCA has conserved the computational resource and facilitate the analysis process, even improved analysis performance in some cases. Classification of different models are compared, and the optimal one reaches satisfactory matrices. The result of such sleep stage analysis procedure can be used in further assistance for clinic diagnosis of sleep disorders (e.g., sleep apnea), reducing inter- and intra-observer variability and decreasing the need for interpreting multiple signals. In addition to the sleep stage analysis procedure, the pre-processed dataset and extracted feature data of St. Vincent's University Hospital / University College Dublin Sleep Apnea Database have established and can be used in further study as well. Furthermore, a novel form of hypnogram is proposed. It is more straightforward compared to a traditional hypnogram.

5.2 Further Work

Although the project is completed to the largest extent since the limitation of project duration and other unavoidable factors, further work of the project can be conducted from several aspects.

1. In this project, only conventional statistic machine learning classifiers are used. However, in recent years, deep learning has been applied in plenty of scenarios and achieved surprising performance, such as computer vision and natural language processing and so on. Introducing deep learning may significantly improve the result of the staging.
2. This project has not considered the temporal relations between the epochs in the signal. However, PSG data has a strong temporal dependence. To take information in previous epochs into account, Hidden Markov Model (HMM) can be used. Combining with deep learning, the neural network model which considered temporal relation (e.g., recurrent

neural network, LSTM, etc.) can be attempted as well.

3. In order to simplify the problem, only EEG and ECG signal in PSG is used in this project. However, other signals in PSG, such as EOG, can provide additional information related to sleep. The synthesis of these data may also enhance the outcome of the procedure.
4. Dataset adopted in this project is collected with suspicious patients. However, a normal sleeper can have a more typical sleep pattern. Adding normal sleepers' data can help model better learn the physiological signal pattern of each sleep stage.

References

Agarwal R and Gotman J (2001) Computer-assisted sleep staging. *IEEE Transactions on Biomedical Engineering* 48(12): 1412–1423. DOI: 10.1109/10.966600.

Alickovic E and Subasi A (2018) Ensemble SVM Method for Automatic Sleep Stage Classification. *IEEE Transactions on Instrumentation and Measurement* 67(6): 1258–1265. DOI: 10.1109/tim.2018.2799059.

Bajaj V and Pachori RB (2013) Automatic classification of sleep stages based on the time-frequency image of EEG signals. *Computer Methods and Programs in Biomedicine* 112(3): 320–328. DOI: 10.1016/j.cmpb.2013.07.006.

Baldwin CM, Griffith KA, Nieto FJ, et al. (2001) The Association of Sleep-Disordered Breathing and Sleep Symptoms with Quality of Life in the Sleep Heart Health Study. *Sleep* 24(1): 96–105. DOI: 10.1093/sleep/24.1.96.

Berry RB, R. B, E. GC, et al. (2015) The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications. American Academy of Sleep Medicine. Available at: http://www.aasmnet.org/.

Boostani R, Karimzadeh F and Torabi-Nami M (2016) A Comparative Review on Sleep Stage Classification Methods.pdf. Available at: https://hal.archives-ouvertes.fr/hal-01390384.

Darbellay GA and Vajda I (1999) Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory* 45(4): 1315–1321. DOI: 10.1109/18.761290.

Delorme A and Makeig S (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*.

Fan Z, Liu E and Xu B (2011) Weighted Principal Component Analysis. In: *Artificial Intelligence and Computational Intelligence*, 1 September 2011.

Faust O, Razaghi H, Barika R, et al. (2019) A review of automated sleep stage scoring based on physiological signals for the new millennia. *Computer Methods and Programs in Biomedicine* 176: 81–91. DOI: 10.1016/j.cmpb.2019.04.032.

Goldberger AL, Amaral LAN, Glass L, et al. (2000) PhysioBank, PhysioToolkit, and PhysioNet. *Circulation* 101(23): E215-20. DOI: 10.1161/01.cir.101.23.e215.

Hassan AR and Bhuiyan MIH (2017) Automated identification of sleep states from EEG signals by means of ensemble empirical mode decomposition and random under sampling boosting. *Computer Methods and Programs in Biomedicine* 140: 201–210. DOI: 10.1016/j.cmpb.2016.12.015.

Hillman DR and Lack LC (2013) Public health implications of sleep loss: the community burden. *Medical Journal of Australia* 199(S8): S7–S10. DOI: 10.5694/mja13.10620.

Hjorth B (1973) The physical significance of time domain descriptors in EEG analysis. *Electroencephalography and Clinical Neurophysiology* 34(3): 321–325. DOI: 10.1016/0013-4694(73)90260-5.

Kumar JS and Bhuvaneswari P (2012) Analysis of Electroencephalography (EEG) Signals and Its Categorization–A Study. *Procedia Engineering* 38: 2525–2536. DOI: 10.1016/j.proeng.2012.06.298.

Li T and Zhou M (2016) ECG Classification Using Wavelet Packet Entropy and Random Forests. *Entropy* 18(8): 285. DOI: 10.3390/e18080285.

Liang S-F, Kuo C-E, Hu Y-H, et al. (2012) Automatic Stage Scoring of Single-Channel Sleep EEG by Using Multiscale Entropy and Autoregressive Models. *IEEE Transactions on Instrumentation and Measurement* 61(6): 1649–1657. DOI: 10.1109/tim.2012.2187242.

Long X, Foussier J, Fonseca P, et al. (2014) Analyzing respiratory effort amplitude for automated sleep stage classification. *Biomedical Signal Processing and Control* 14: 197–205. DOI: 10.1016/j.bspc.2014.08.001.

Mendez MO, Matteucci M, Castronovo V, et al. (2010) Sleep staging from Heart Rate Variability: time-varying spectral features and Hidden Markov Models. *International Journal of Biomedical Engineering and Technology* 3(3/4): 246. DOI: 10.1504/ijbet.2010.032695.

Michielli N, Acharya UR and Molinari F (2019) Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals. *Computers in Biology and Medicine* 106: 71–81. DOI: 10.1016/j.compbiomed.2019.01.013.

Motamedi-Fakhr S, Moshrefi-Torbati M, Hill M, et al. (2014) Signal processing techniques applied to human sleep EEG signals—A review. *Biomedical Signal Processing and Control* 10: 21–33. DOI: 10.1016/j.bspc.2013.12.003.

Noviyanto A, Isa SM, Wasito I, et al. (2011) Selecting features of single lead ECG signal for automatic sleep stages classification using correlation-based feature subset selection. *International Journal of Computer Science Issues* 8(5): 139.

Phan H, Chén OY, Koch P, et al. (2019) Towards More Accurate Automatic Sleep Staging via Deep Transfer Learning. *arXiv*. DOI: 10.1109/tbme.2020.3020381.

Rahman MM, Bhuiyan MIH and Hassan AR (2018) Sleep stage classification using single-channel EOG. *Computers in Biology and Medicine* 102: 211–220. DOI: 10.1016/j.compbiomed.2018.08.022.

Safont G, Salazar A, Vergara L, et al. (2014) Mixtures of Independent Component Analyzers for Microarousal Detection. *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*: 752–755. DOI: 10.1109/bhi.2014.6864473.

Serrano E and Figliola A (2009) Wavelet Leaders: A new method to estimate the multifractal singularity spectra. *Physica A: Statistical Mechanics and its Applications* 388(14): 2793–2805. DOI: 10.1016/j.physa.2009.03.043.

Singh A, Thakur N and Sharma A (2016) A Review of Supervised Machine Learning Algorithms. In: *Proceedings of International Conference on Computing for Sustainable Global Development*, New Delhi, India, 18 March 2016. IEEE.

Stutts JC, Wilkins JW, Osberg JS, et al. (2003) Driver risk factors for sleep-related crashes. *Accident Analysis & Prevention* 35(3): 321–331. DOI: 10.1016/s0001-4575(02)00007-6.

Sulistyo B, Surantha N and Isa SM (2017) Sleep Apnea Identification using HRV Features of ECG Signals. *International Journal of Electrical and Computer Engineering (IJECE)* 8(5): 3940–3948. DOI: 10.11591/ijece.v8i5.pp3940-3948.

Wendt H and Abry P (2007) Multifractality Tests Using Bootstrapped Wavelet Leaders. *IEEE Transactions on Signal Processing* 55(10): 4811–4820. DOI: 10.1109/tsp.2007.896269.

Willemen T, Deun DV, Verhaert V, et al. (2014) An Evaluation of Cardiorespiratory and Movement Features With Respect to Sleep-Stage Classification. *IEEE Journal of Biomedical and Health Informatics* 18(2): 661–669. DOI: 10.1109/jbhi.2013.2276083.

Yücelbaş Ş, Yücelbaş C, Tezel G, et al. (2018) Automatic sleep staging based on SVD, VMD, HHT and morphological features of single-lead ECG signal. *Expert Systems with Applications* 102: 193–206. DOI: 10.1016/j.eswa.2018.02.034.

Acknowledgement

Throughout my undergraduate study, I have received a great amount of encouragement and assistance.

I would first like to thank my three supervisors in this project whose knowledge was invaluable in formulating the research methodology and questions. Your insightful comment pushed me to perfect my thought and took my work to a higher level, despite in a remote form due to the COVID-19 pandemic. Hope we can meet physically one day in Valencia.

In addition, I would like to acknowledge Prof. Philip S. Yu at the University of Illinois at Chicago and Prof. Chuan Shi at Beijing University of Posts and Telecommunications (none of my supervisors of this project) for their leading to the next episode of my study.

Moreover, I would like to offer my special thanks to my parents for their visionary counsel and continuous financial assistance. I could not have such an achievement without your support. Finally, I would like to extend my sincere thanks to my classmate Jingjing Wei who provided inspiring discussions during the project, as well as my friend Xinran Zhuang for her assistance in the early stage of the project.

Appendix

Appendix A: Project Specification Form (Supervisor)

北京邮电大学 本科毕业设计（论文）任务书

Project Specification Form

Part 1 – Supervisor

论文题目 Project Title	Dimension Reduction Methods Applied to Sleep Stage Analysis		
题目分类 Scope	Data Science and Artificial Intelligence	Research	Software
主要内容 Project description	This project will implement several techniques to reduce the dimensionality of physiological data used to classify stages of sleep in subjects with sleep disorders. The data will consist of features extracted from electroencephalographic (EEG) and electrocardiographic (ECG) signals measured from subjects while sleeping. The sleep pass through a cycle composed by several stages: wake, stage 1, stage 2, stage 3, and rapid eye movement (REM) sleep. Those stages perform different behavior that is captured in physiological signals. Sleep cycle is repeated several times at night and the proportion between stage duration as well as the content of physiological signals is used to determine sleep disorders. The number of features that could be extracted from the EEG and ECG signals can be very high compared to the number of data records. In these conditions, results of sleep stage classification could produce an overfitting of the classification model to the data and be invalid. To deal with this problem, dimension reduction methods should be applied to obtain data for classification with a smaller number of features than the original one. This project will implement several classification cases using different number of features and the classification accuracy index will be used to evaluate the quality of the results, in combination with clinical annotation information.		
关键词 Keywords	Signal processing, Dimension reduction, EEG, ECG, Classification, Sleep analysis		
主要任务 Main tasks	**1** Study of methods for sleep stage classification: preprocessing, feature extraction, dimension reduction, and classifiers.		
	2 Design and implementation of the procedures of preprocessing, dimension reduction, and classification.		
	3 Experimentation: definition of the database; tuning and debugging of the methods; implementation of figures of merit.		
	4 Evaluation and reporting of the results.		
主要成果 Measurable outcomes	**1** Software of the implementation of the dimension reduction processing step of EEG and ECG data (and report on the implemented methods).		
	2 Software of the implementation of the classification processing step (and report on the implemented methods).		
	3 Software for obtaining results: classification accuracy, result comparison, confusion matrices (and reports on the results).		

Appendix B: Project Specification Form (Student)

北京邮电大学 本科毕业设计（论文）任务书

Project Specification Form

Part 2 - Student

学院 School	International School	**专业 Programme**	**Internet of Things Engineering**		
姓 Family name	Liu	**名 First Name**	Zekuan		
BUPT 学号 BUPT number	2017213176	**QM 学号 QM number**	171043570	**班级 Class**	2017215120
论文题目 Project Title	Dimension Reduction Methods Applied to Sleep Stage Analysis				
论文概述 Project outline **Write about 500-800 words** **Please refer to Project Student Handbook section 3.2**	Sleep plays a critical role in our daily life. Sleep disorders will not only harm personal health but also reduce the productivity among society. Nowadays, overnight Polysomnography (PSG) is widely used in sleep stage analysis and further address sleep disorders. However, multiply measurements in PSG add data redundancy, and features extracted from collected data can have a very high dimension. To address this problem, we are going to reduce the dimension of collected data, classifying different sleep stages with fewer features. For sleep staging, we will implement various classifiers, and finally find an optimal solution. There are mainly four tasks in this project: **1. Study of methods for sleep stage classification: preprocessing, feature extraction, dimension reduction, and classifiers.** To begin with, preprocessing and feature extraction approaches specified in EEG and ECG data will be investigated. Multiple dimension reduction methods, which are the key of this project, including but not limited to principal component analysis, embedding, and autoencoder, will be compared. In addition, both conventional statistic classifiers and deep learning classifiers will be studied. **2. Design and implementation of the procedures of preprocessing, dimension reduction, and classification.** All the procedures mentioned in the first task will be designed and implemented within this task. In this project, MATLAB is briefly used, adding with necessary toolkits (e.g., Signal Processing Toolkit, etc.). If time is sufficient, we may able to implement other state-of-the-art models in order to conduct a baseline experiment with the procedure developed by us. The outcome of this task will be a software of the implementation of the dimension reduction processing step and the classification processing step of EEG and ECG data (and report on the implemented methods). **3. Experimentation: definition of the database; tuning and debugging of the methods; implementation of figures of merit.** The dataset used in this project is St. Vincent's University Hospital/University College Dublin Sleep Apnea Database, which contains 25 full overnight polysomnograms with simultaneous three-channel Holter ECG, from adult subjects with suspected sleep-disordered breathing. Since the existing dataset is used in this project, hardware is not required. Preprocessing and feature extraction will be conducted on this dataset. Then, the extracted data will be stored for further dimension reduction and classification. After the				

1

	implementation of dimension reduction methods and classifiers, tuning of these models is required to get better results. Moreover, metrics of obtained results will be calculated to reveal the performance. Besides, if plausible, we can also conduct a comparison experiment between classification with dimension reduction and classification without dimension reduction, and baseline experiments with other state-of-the-art sleep staging models. The outcome of this task is software for obtaining results: classification accuracy, result comparison, confusion matrices (and reports on the results). **4. Evaluation and reporting of the results.** Evaluation of different dimension reduction methods and diverse classifiers will be conducted. Among all the methods, the optimal solution will be selected, and the reason for the obtained results will be analyzed. Furthermore, we will conclude the performance of whole procedures and compare it to state-of-the-art models, and the synthesis and reporting of the results is also indispensable. **Bibliography** [1] Faust, O., Razaghi, H., Barika, R., Ciaccio, E. J., & Acharya, U. R. (2019). A review of automated sleep stage scoring based on physiological signals for the new millennia. *Computer Methods and Programs in Biomedicine*, *176*, 81–91. https://doi.org/10.1016/j.cmpb.2019.04.032 [2] Safont, G., Salazar, A., Vergara, L., Gomez, E., & Villanueva, V. (2014). Mixtures of Independent Component Analyzers for Microarousal Detection. *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 752–755. https://doi.org/10.1109/bhi.2014.6864473 [3] Agarwal, R., & Gotman, J. (2001). Computer-assisted sleep staging. *IEEE Transactions on Biomedical Engineering*, *48*(12), 1412–1423. https://doi.org/10.1109/10.966600 [4] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P.C., Mark, R., Mietus, J.E., Moody, G.B., Peng, C.K. and Stanley, H.E., 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.
道德规范 **Ethics**	Please confirm that you have discussed ethical issues with your Supervisor using the ethics checklist (Project Handbook Appendix 2). [YES]
	Summary of ethical issues: N/A
中期目标 **Mid-term target.** **It must be tangible outcomes, E.g. software, hardware or simulation.** **It will be assessed at the mid-term oral.**	Before the mid-term oral, most of the project coding should be done. Specifically, a complete procedure prototype of the project without fine tuning should be established. The procedure mainly includes preprocessing, feature extraction, dimension reduction, classification, etc. In addition, a well-defined database can be applied to the procedure.

2

Work Plan (Gantt Chart)

Fill in the sub-tasks and insert a letter X in the cells to show the extent of each task

	Nov 1-15	Nov 16-30	Dec 1-15	Dec 16-31	Jan 1-15	Jan 16-31	Feb 1-15	Feb 16-29	Mar 1-15	Mar 16-31	Apr 1-15	Apr 16-30
Task 1 Study of methods for sleep stage classification												
Study of preprocessing	X	X										
Study of feature extraction		X	X									
Study of dimension reduction			X	X								
Study of classifiers				X	X							
Task 2 Design and implementation of the procedures												
Design and implementation of preprocessing			X	X								
Design and implementation of feature extraction				X	X							
Design and implementation of dimension reduction					X	X						
Design and implementation of classifiers						X	X					
Task 3 Experimentation												
Definition of the database	X	X										
Tuning and debugging of the methods								X	X			
Implementation of figures of merit									X	X		
Comparison experiment								X	X	X	X	
Task 4 Evaluation and reporting of the results.												
Result evaluation								X	X	X	X	
Result reporting								X	X	X	X	
Viva preparing										X	X	X
Essay writing								X	X	X	X	X

3

Appendix C: Project Early-term Progress Report

北京邮电大学 本科毕业设计（论文）初期进度报告

Project Early-term Progress Report

学院 School	International School	**专业 Programme**	**Internet of Things Engineering**		
姓 Family name	Liu	**名 First Name**	Zekuan		
BUPT 学号 BUPT number	2017213176	**QM 学号 QM number**	171043570	**班级 Class**	2017215120
论文题目 Project Title	Dimension Reduction Methods Applied to Sleep Stage Analysis				

已完成工作 Finished work:

1. Study of methods for sleep stage classification

1.1. Study of preprocessing

Electroencephalogram (EEG) and electrocardiogram (ECG) signals are recorded by sensitive electrodes. Due to the complication of the sensing environment, artifacts are inevitable. Basically, there are two types of artifacts, biological artifacts and environmental artifacts. As the EEG electrodes are placed on the sculp, it will not only record brain neuron electrical activities, but also those from eyes (ocular artifacts), muscles (muscle artifacts), even heart (ECG artifacts). These are called biological artifacts. On the other hand, environmental artifacts refer to the artifacts originated outside of body, e.g., body movement, 50 or 60 Hz artifacts from power, etc. Artifacts may lead to significant variation to the signal. To reduce the influence of artifacts, three methods can be used in preprocessing of the raw data, rejecting bad data, filtering and independent component analysis (ICA).

1.2. Study of feature extraction

Feature extraction is an important step in the whole procedure, which can significantly affect the final classification outcome. There are different methods to extract features from ECG and EEG signals respectively. As for the ECG signal, from simple heart rate (HR), heart rate variability (HRV), to relatively complex singular value decomposition (SVD), variational mode decomposition (VMD), Hilbert Huang transform (HHT) and morphological features, plenty of methods can be used to extract features from ECG data. Regarding the EEG data, time-domain features, spectral features, time-frequency features and non-linear features can be used.

1.3. Study of dimension reduction

As per the requirement of supervisor, only principal component analysis (PCA) and weighted PCA will be used.

PCA: Principal component analysis is a common linear dimension reduction algorithm. Its optimization goal is maximizing the variance of the data in low dimensional space,

$$w = \text{argmax}\ \frac{w^T X^T X w}{w^T w}$$

where the w is the parameter, and the X is the data.

Weighted PCA: There are briefly two types of weighted PCA, sample-wise weighted PCA and variable-wise weighted PCA. In order to adapt to the trend and continuously learn the new information, sample-wise weighted PCA adopt moving window, recursing or EWMA (Exponentially Weighted Moving Average) filter and so on. As for variable-wise weighted PCA, the weightings are determined by a customized formula, with consideration given not to over-weight variables.

1.4. Study of classifiers

1

According to the requirement of supervisor, only classifiers, such as linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), logistic regression, and random forest, will be used in the project. Linear discriminant analysis is an algorithm which finds a linear combination of features that characterizes or separates two or more classes of objects. It also frequently used on dimension reduction. It is suitable when the data of different classes are Gaussian distributed and have similar variances, while quadratic discriminant analysis can tackle various variances, which is more difficult to handle. Their corresponding principals are,

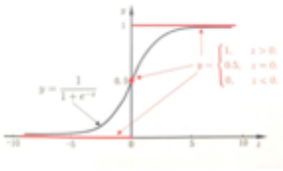
LDA

$$\log\frac{P(C_1|x)}{P(C_2|x)} = \log\frac{f_1(x)}{f_2(x)} + \log\frac{\pi_1}{\pi_2}$$
$$= x^T\Sigma^{-1}(\mu_1-\mu_2) - \frac{1}{2}(\mu_1+\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2) + \log\frac{\pi_1}{\pi_2}$$

QDA

$$\log\frac{P(C_1|x)}{P(C_2|x)} = \log\frac{f_1(x)}{f_2(x)} + \log\frac{\pi_1}{\pi_2}$$
$$= x^T(\Sigma_1^{-1}-\Sigma_2^{-1})x + x^T\Sigma^{-1}(\mu_1-\mu_2) - \frac{1}{2}(\mu_1+\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2) + \log\frac{\pi_1}{\pi_2} - \frac{1}{2}\log\frac{|\Sigma_1|}{|\Sigma_2|}$$

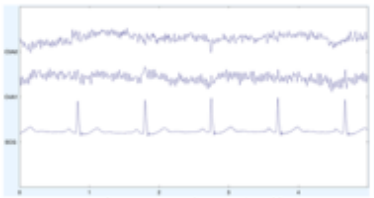
where π_k refers to the proportion of sample number of class C_k, μ_k refers to the sample mean of class C_k, and $\Sigma = \frac{1}{N-K}\Sigma_{k=1}^{K}\Sigma_{x\in C_k}(x-\mu_k)(x-\mu_k)^T$. Logistic regression is an algorithm which uses a logistic function to model a binary dependent variable. The figure of the logistic function is shown on the right. Random forest is an ensemble learning algorithm, which apply a set of decision trees during training and outputting the mode of the classes or average prediction of the individual trees. It applies bootstrap algorithm.



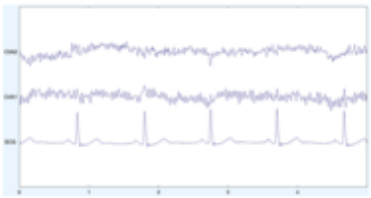
2. Design and implementation of the procedures

2.1. Design and implementation of preprocessing

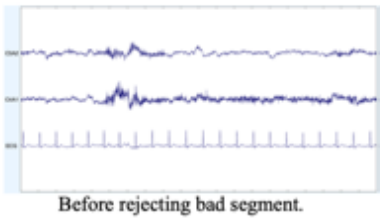
By far, I have applied EEGLAB (a toolbox developed by UCSD) to implemented preprocessing, including selecting channels, removing DC offset, filtering, rejecting bad segment, etc. As the selected comparisons show below, the baseline component of the signal is removed. Then, those segments with significant artifacts are rejected (red lines in the figure represent the original position of bad segments). Finally, independent component analysis is conducted. ECG artifacts are largely reduced in EEG signal.



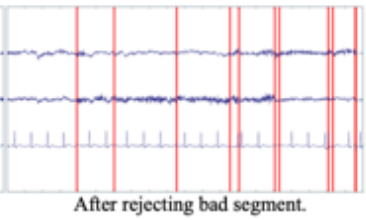
Before removing DC offset.



After removing DC offset.

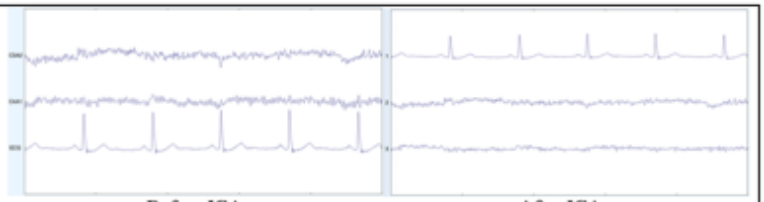


Before rejecting bad segment.



After rejecting bad segment.

2



Before ICA. After ICA.

In addition, in order to fulfill the requirement of sleep staging, the signals are segmented into 30s segments.

2.2. Design and implementation of feature extraction

By using EEG Feature Extraction Toolbox (https://github.com/JingweiToo/EEG-Feature-Extraction-Toolbox), 30 types of EEG features can be extracted easily, including Hjorth parameters, band power, entropy, etc. Similarly, ECG feature extraction is done by the Open-Source Electrophysiological Toolbox.

Bibliography

[1] En.wikipedia.org. 2021. *Electroencephalography*. [online] Available at: <https://en.wikipedia.org/wiki/Electroencephalography> [Accessed 11 January 2021].

[2] Es.mathworks.com. 2021. *Principal Component Analysis Of Raw Data - MATLAB Pca-Mathworks*. [online] Available at: <https://es.mathworks.com/help/stats/pca.html?lang=en> [Accessed 11 January 2021].

[3] Delorme, A. and Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of neuroscience methods*, *134*(1), pp.9-21.

[4] Faust, O., Razaghi, H., Barika, R., Ciaccio, E.J. and Acharya, U.R., 2019. A review of automated sleep stage scoring based on physiological signals for the new millennia. *Computer methods and programs in biomedicine*, *176*, pp.81-91.

[5] Yücelbaş, Ş., Yücelbaş, C., Tezel, G., Özşen, S. and Yosunkaya, Ş., 2018. Automatic sleep staging based on SVD, VMD, HHT and morphological features of single-lead ECG signal. *Expert Systems with Applications*, *102*, pp.193-206.

[6] Wold, S., Esbensen, K. and Geladi, P., 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems*, *2*(1-3), pp.37-52.

[7] Yue, H.H. and Tomoyasu, M., 2004, December. Weighted principal component analysis and its applications to improve FDC performance. In *2004 43rd IEEE Conference on Decision and Control (CDC)(IEEE Cat. No. 04CH37601)* (Vol. 4, pp. 4262-4267). IEEE.

[8] Oh, S.H., Lee, Y.R. and Kim, H.N., 2014. A novel EEG feature extraction method using Hjorth parameter. *International Journal of Electronics and Electrical Engineering*, *2*(2), pp.106-110.

[9] R. Sameni, OSET: The open-source electrophysiological toolbox. Version 3.14, URL: http://www.oset.ir

是否符合进度？On schedule as per GANTT chart?

YES

下一步 Next steps:

Before the next mid-term progress report, I will preliminarily implement the whole procedure of the project, including the further dimension reduction and classification, without hyperparameter turning and finely optimization.

Appendix D: Project Mid-term Progress Report

北京邮电大学 本科毕业设计（论文）中期进度报告

Project Mid-term Progress Report

学院 School	International School	**专业 Programme**	**Internet of Things Engineering**		
姓 Family name	Liu	**名 First Name**	Zekuan		
BUPT 学号 BUPT number	2017213176	**QM 学号 QM number**	171043570	**班级 Class**	2017215120
论文题目 Project Title	Dimension Reduction Methods Applied to Sleep Stage Analysis				

是否完成任务书中所定的中期目标？ Targets met (as set in the Specification)?
YES

已完成工作 Finished work:

1. Study of methods for sleep stage classification

1.1. Study of preprocessing

Electroencephalogram (EEG) and electrocardiogram (ECG) signals are recorded by sensitive electrodes. Due to the complication of the sensing environment, artifacts are inevitable. Basically, there are two types of artifacts, biological artifacts and environmental artifacts. As the EEG electrodes are placed on the sculp, it will not only record brain neuron electrical activities, but also those from eyes (ocular artifacts), muscles (muscle artifacts), even heart (ECG artifacts). These are called biological artifacts. On the other hand, environmental artifacts refer to the artifacts originated outside of body, e.g., body movement, 50 or 60 Hz artifacts from power, etc. Artifacts may lead to significant variation to the signal. To reduce the influence of artifacts, three methods can be used in preprocessing of the raw data, rejecting bad data, filtering and independent component analysis (ICA). In addition to the preprocessing stated above, other common preprocessings are performed, including normalisation, calibration, detrending and equalisation. Besides, the signals are required to be divided into uniform 30 seconds epochs for later classification.

1.2. Study of feature extraction

Features are parameters which provide information about the underlying structure of the signal of interest. Feature extraction is an important step in the whole procedure, which can significantly affect the final classification outcome. There are different methods to extract features from ECG and EEG signals respectively. As for the ECG signal, autoregressive (AR) coefficients, Shannon entropy, wavelet variance estimates are applied. Regarding the EEG data, temporal features and spectral features are used. Specifically, time domain features are the powers in the frequency band of delta wave (0-4 Hz), theta wave (5-7 Hz), alpha wave (8-12 Hz), sigma wave (13-15 Hz) and beta wave (16-30 Hz). On the top of that, Hjorth parameters are estimated in the frequency domain, including activity, mobility and complexity.

1.3. Study of dimension reduction

As per the requirement of supervisor, only principal component analysis (PCA) and weighted PCA will be used.

PCA: Principal component analysis is a common linear dimension reduction algorithm. Its optimization goal is maximizing the variance of the data in low dimensional space. It can be represented in following formula.

$$w = argmax \frac{w^T X^T X w}{w^T w}$$

where the w is the parameter, and the X is the data.

Weighted PCA: There are briefly two types of weighted PCA, sample-wise weighted PCA and variable-wise weighted PCA. In order to adapt to the trend and continuously learn the new

1

information, sample-wise weighted PCA adopt moving window, recursing or EWMA (Exponentially Weighted Moving Average) filter and so on. As for variable-wise weighted PCA, the weightings are determined by a customized formula, with consideration given not to over-weight variables.

1.4. Study of classifiers

According to the requirement of supervisor, following classifiers are researched.

Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA): Linear discriminant analysis is an algorithm which finds a linear combination of features that characterizes or separates two or more classes of objects. It also frequently used on dimension reduction. It is suitable when the data of different classes are Gaussian distributed and have similar variances, while quadratic discriminant analysis can tackle various variances, which is more difficult to handle. Their corresponding principals are,

LDA

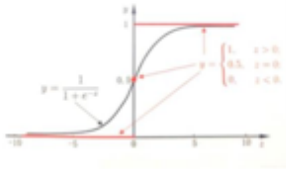
$$\log\frac{P(C_1|x)}{P(C_2|x)} = \log\frac{f_1(x)}{f_2(x)} + \log\frac{\pi_1}{\pi_2}$$
$$= x^T\Sigma^{-1}(\mu_1-\mu_2) - \frac{1}{2}(\mu_1+\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2) + \log\frac{\pi_1}{\pi_2}$$

QDA

$$\log\frac{P(C_1|x)}{P(C_2|x)} = \log\frac{f_1(x)}{f_2(x)} + \log\frac{\pi_1}{\pi_2}$$
$$= x^T(\Sigma_1^{-1}-\Sigma_2^{-1})x + x^T\Sigma^{-1}(\mu_1-\mu_2)$$
$$-\frac{1}{2}(\mu_1+\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2)$$
$$+\log\frac{\pi_1}{\pi_2} - \frac{1}{2}\log\frac{|\Sigma_1|}{|\Sigma_2|}$$

where π_k refers to the proportion of sample number of class C_k, μ_k refers to the sample mean of class C_k, and $\Sigma = \frac{1}{N-K}\Sigma_{k=1}^{K}\Sigma_{x\in C_k}(x-\mu_k)(x-\mu_k)^T$.

Logistic Regression: Logistic regression is an algorithm which uses a logistic function to model a binary dependent variable. The figure of the logistic function is shown on the right.

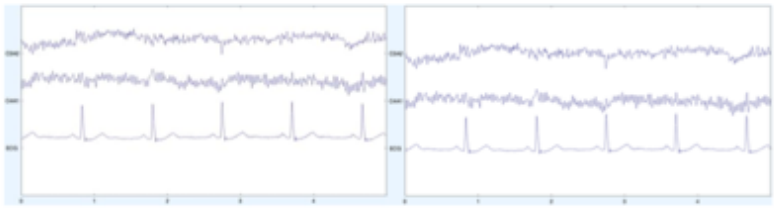


Random Forest: Random forest is an ensemble learning algorithm, which apply a set of decision trees during training and outputting the mode of the classes or average prediction of the individual trees. It applies bootstrap algorithm.

2. Design and implementation of the procedures

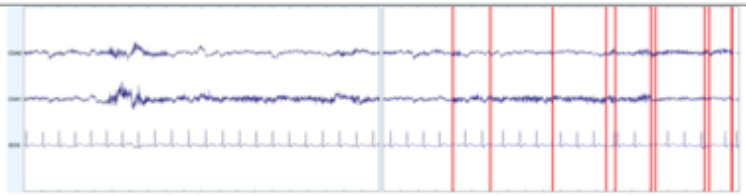
2.1. Design and implementation of preprocessing

By far, I have applied EEGLAB (a toolbox developed by UCSD) to implemented preprocessing, including selecting channels, removing DC offset, filtering, rejecting bad segment, etc. As the selected comparisons show below, the baseline component of the signal is removed. Then, those segments with significant artifacts are rejected (red lines in the figure represent the original position of bad segments). Finally, independent component analysis is conducted. ECG artifacts are largely reduced in EEG signal.

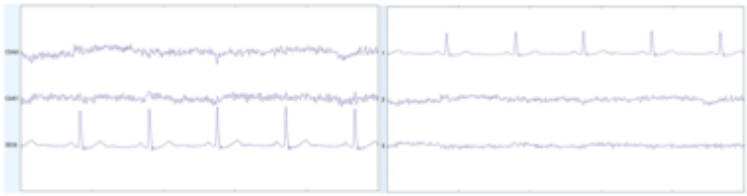


Before removing DC offset. After removing DC offset.

2



Before rejecting bad segment. After rejecting bad segment.

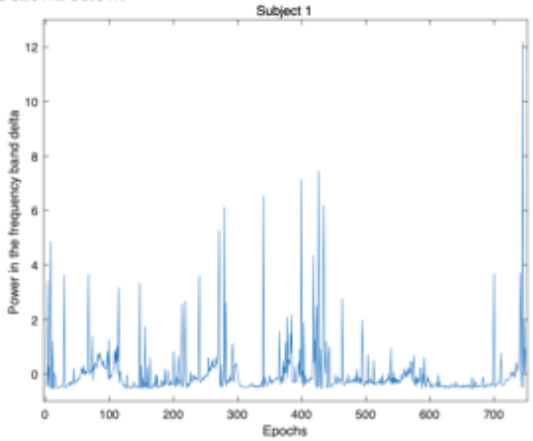


Before ICA. After ICA.

After the steps stated above, the data is retrieved from EEGLAB and segmented into uniform 30s epochs in order to fulfill the requirement of sleep staging.

2.2. Design and implementation of feature extraction

By using EEG Feature Extraction Toolbox (https://github.com/JingweiToo/EEG-Feature-Extraction-Toolbox), 30 types of EEG features can be extracted easily, including Hjorth parameters, band power, entropy, etc. Similarly, ECG feature extraction is done by the Open-Source Electrophysiological Toolbox. A sample of an extracted feature (power of delta wave of subject 1) is shown below.

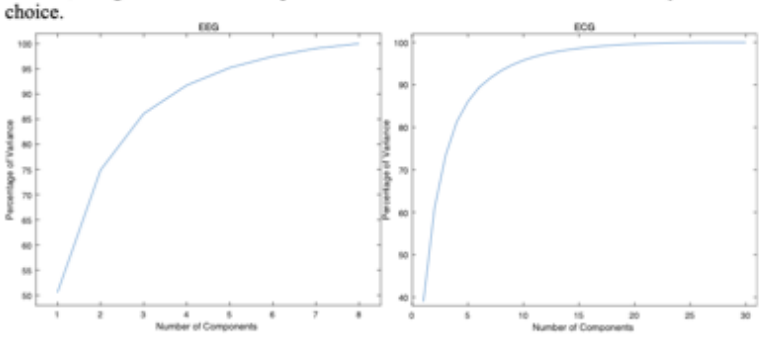


2.3. Design and implementation of dimension reduction

By far, the dimension reduction is implemented with principal component analysis function in MATLAB Statistics and Machine Learning Toolbox. The input of the function is the extracted feature data, while the outputs are principal component coefficients, principal component scores, principal component variances, Hotelling's T-squared statistic, percentage of total variance explained and estimated means. As the figure shows, with the decrease of the number

3

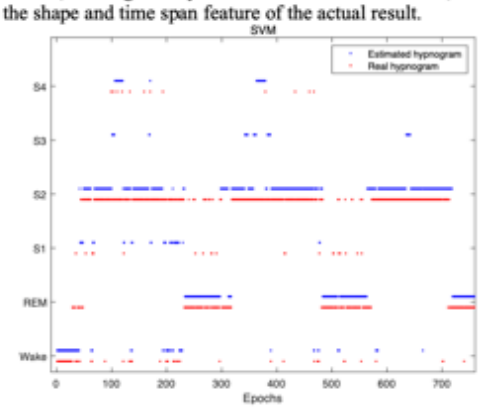
of components the percentage of the total variance explained by each principal component decline as well. To reduce the dimension without losing too much information contained in the data, using the number of components 12 for ECG and 6 for EEG is a relatively reasonable choice.



2.4. Design and implementation of classifiers

The classification is implemented by MATLAB Statistic and Machine Learning Toolbox. The leave-one-out classification based on the 30s epochs is used. The procedure is shown in the following flow diagram.

First, the data splitting is based on leave-one-out, which means in each experiment one subject is used as testing set, while other is used as training set. For instance, the training set is subject 1 to subject 9, while subject 10 is the testing set. Then, according to different classification type (e.g., awake-asleep binary classification, full six classification, etc.), the label of the data has to be modified. After specification of the classifier, the relabelled training data can be imported into the classifier and start training. Depending on the classifier, the computing period may vary. Next, the testing dataset are predicted by trained model. By comparison of predicted values and actual values, evaluation metrics can be obtained. The following figure shows the visualization comparison between SVM predicted value and actual value for a subject. As we can see, although many exact values are inconsistent, the prediction result basically restore the shape and time span feature of the actual result.



4

3. Experimentation

3.1. Definition of the database

The raw database of this project is St. Vincent's University Hospital / University College Dublin Sleep Apnea Database. But the latter part of this project is based on provided extracted feature data of this database, including 8 EEG features and 30 ECG features of 10 subject. Specifically, EEG power in the frequency band of delta, theta, alpha, sigma and beta, EEG Hjorth parameters (activity, mobility and complexity), ECG Autoregressive (AR) coefficients (four values), ECG Shannon entropy (16 values) and ECG Wavelet variance estimates (10 values).

3.2. Tuning and debugging of the methods

The process and methods are implemented without significant problem. The basic process of the project runs normally. However, without fine tuning and careful code review, potential bugs and problems may exist, and the result of experiments may be affected. As the experiment continue, tuning and debugging will follow.

3.3. Implementation of figures of merit

Following metrics are implemented.

Confusion Matrix: As the following table shows, the row of the matrix represents the predicted class, while the column represents the actual class. In the confusion matrix, *tp* and *tn* denote the number of positive and negative instances that are correctly predicted. Meanwhile, *fp* and *fn* denote the number of misclassified negative and positive instances, respectively. In multiclass problems, the result is only identified as true when actual class and predicted class are the same. Otherwise, the result is false. In multiclass problem, the rows and columns of the matrix can also be classes of the data.

	Actual Positive	**Actual Negative**
Predicted Positive	True Positive (*tp*)	False Negative (*fn*)
Predicted Negative	False Positive (*fp*)	True Negative *(tn*)

Accuracy: Accuracy measures the ratio of instances that are correctly classified over total number of instances.

$$acc = \frac{tp + tn}{tp + fp + tn + fn}$$

Error Rate: On the contrary of accuracy, error rate measures the ratio of instances that are wrongly classified over total number of instances.

$$err = \frac{fp + fn}{tp + fp + tn + fn} = 1 - acc$$

Precision: Precision measures the positive patterns that are correctly predicted from the total predicted patterns in a positive class.

$$p = \frac{tp}{tp + fp}$$

Recall rate: Recall measures the fraction of positive patterns that are correctly classified.

$$r = \frac{tp}{tp + tn}$$

F1: F1 is the harmonic mean between recall and precision values. In multiclass problems, two kind of F1 can be estimated, micro-F1 and macro-F1. Macro-F1 computes the F1 independently for each class and then take the average (hence treating all classes equally), whereas a micro-F1 will aggregate the contributions of all classes to compute the average metric.

$$F1 = \frac{2 * p * r}{p + r}$$

Bibliography

[1] En.wikipedia.org. 2021. *Electroencephalography*. [online] Available at: <https://en.wikipedia.org/wiki/Electroencephalography> [Accessed 11 January 2021].
[2] Es.mathworks.com. 2021.*Principal Component Analysis Of Raw Data - MATLAB Pca-Mathworks*. [online] Available at: <https://es.mathworks.com/help/stats/pca.html?lang=en> [Accessed 11 January 2021].
[3] Delorme, A. and Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis.*Journal of neuroscience methods*, *134*(1), pp.9-21.
[4] Faust, O., Razaghi, H., Barika, R., Ciaccio, E.J. and Acharya, U.R., 2019. A review of automated sleep stage scoring based on physiological signals for the new millennia. *Computer methods and programs in biomedicine*, *176*, pp.81-91.
[5] Yücelbaş, Ş., Yücelbaş, C., Tezel, G., Özşen, S. and Yosunkaya, Ş., 2018. Automatic sleep staging based on SVD, VMD, HHT and morphological features of single-lead ECG signal. *Expert Systems with Applications*, *102*, pp.193-206.
[6] Wold, S., Esbensen, K. and Geladi, P., 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems*, *2*(1-3), pp.37-52.
[7] Yue, H.H. and Tomoyasu, M., 2004, December. Weighted principal component analysis and its applications to improve FDC performance. In *2004 43rd IEEE Conference on Decision and Control (CDC)(IEEE Cat. No. 04CH37601)* (Vol. 4, pp. 4262-4267). IEEE.
[8] Oh, S.H., Lee, Y.R. and Kim, H.N., 2014. A novel EEG feature extraction method using Hjorth parameter. *International Journal of Electronics and Electrical Engineering*, *2*(2), pp.106-110.
[9] R. Sameni, OSET: The open-source electrophysiological toolbox. V ersion 3.14, URL: http://www.oset.ir
[10] Hossin, M. and Sulaiman, M.N., 2015. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), p.1.
[11] Motamedi-Fakhr, S., Moshrefi-Torbati, M., Hill, M., Hill, C.M. and White, P.R., 2014. Signal processing techniques applied to human sleep EEG signals—A review. *Biomedical Signal Processing and Control*, *10*, pp.21-33.
[12] Li, T. and Zhou, M., 2016. ECG classification using wavelet packet entropy and random forests. *Entropy*, *18*(8), p.285.
[13] Hjorth, B., 1973. The physical significance of time domain descriptors in EEG analysis. Electroen Clin Neuro 34, 321–325. https://doi.org/10.1016/0013-4694(73)90260-5

尚需完成的任务 Work to do:

3.4. Comparison experiment

On-going comparison experiments for different classifiers, number of data components, etc.

4. Evaluation and reporting of the results.

4.1. Result evaluation

The figures of merit are implemented. Pending for the results of experiments.

4.2. Result reporting

The figures of merit are implemented. Pending for the results of experiments.

4.3. Viva preparing

The mid-term viva is in preparation.

4.4. Essay writing

The essay writing is at the beginning stage.

存在问题 Problems:

Problem 1: The experimental results between subjects in data vary a lot.

Problem 2: Until now, the classifiers only involve conventional machine learning classifiers. However, novel deep learning may produce a considerable improvement.

Problem 3: Theory part of the project is not complete yet.
拟采取的办法 Solutions: **Solution 1:** Reproduce the experimental result and seek the potential bugs. **Solution 2:** If time is sufficient, deep learning classifiers will be attempted. For instance, Recurrent Neural Networks (RNN) take time sequence into account. **Solution 3:** Enhance literature review and theory summary.
论文结构 Structure of the final report: Abstract Chapter 1: Introduction Chapter 2: Background Chapter 3: Design and Implementation 3.1 Pre-processing 3.2 Feature Extraction 3.3 Dimension Reduction 3.4 Classification Chapter 4: Results and Discussion Chapter 5: Conclusion and Further Work References Acknowledgement Appendix Risk and Environmental Impact Assessment

7

Appendix E: Project Supervision Log

北京邮电大学 本科毕业设计（论文）教师指导记录表

Project Supervision Log

学院 School	International School	**专业 Programme**	**Internet of Things Engineering**		
姓 Family name	Liu	**名 First Name**	Zekuan		
BUPT 学号 BUPT number	2017213176	**QM 学号 QM number**	171043570	**班级 Class**	2017215120
论文题目 Project Title	Dimension Reduction Methods Applied to Sleep Stage Analysis				

Please record supervision log using the format below:

Date: dd-mm-yyyy
Supervision type: face-to-face meeting/online meeting/email/other (please specify)
Summary:

Date: 13-08-2020
Supervision type: email
Summary: defined programming language and introduced the project briefly

Date: 21-09-2020
Supervision type: email
Summary: defined dataset and provided bibliography

Date: 06-10-2020
Supervision type: group meeting
Summary: external project kick-off meeting

Date: 13-10-2020
Supervision type: email
Summary: confirmed project timeline

Date: 11-11-2020
Supervision type: group meeting
Summary: discussed the project Gannt chart and latest progress

Date: 16-11-2020
Supervision type: email
Summary: discussed the draft specification and ethical issue

Date: 19-11-2020
Supervision type: email
Summary: Q&A

Date: 03-12-2020
Supervision type: email
Summary: discussed the Chinese translation of the project title

Date: 06-01-2020
Supervision type: group meeting

1

Summary: discussed latest progress and project plan for next step

Date: 18-01-2021
Supervision type: email
Summary: discussed the draft early-term progress report

Date: 02-02-2021
Supervision type: email
Summary: provided extracted feature data and the description

Date: 04-02-2021
Supervision type: email
Summary: Q&A

Date: 05-02-2021
Supervision type: email
Summary: provided additional bibliography

Date: 15-02-2021
Supervision type: email
Summary: discussed the classification process

Date: 23-02-2021
Supervision type: email
Summary: discussed the mid-term progress report and program implementation

Date: 26-02-2021
Supervision type: group meeting
Summary: discussed the mid-term progress and mid-term viva

Date: 09-03-2021
Supervision type: email
Summary: discussed the experiment result

Date: 09-04-2021
Supervision type: group meeting
Summary: discussed late-term progress, draft report and mock viva

Date: 21-04-2021
Supervision type: email
Summary: discussed final report

2

Risk and environmental impact assessment

This project applied the collected open-source PSG dataset in the training period, without interaction of human patients. But in the application period, data may have to be collected from human patients and assist doctors to diagnose sleep disorders. Therefore, several risks are assessed:

Description of Risk	Description of Impact	Likelihood rating	Impact rating	Preventative actions
Unreliable dataset	The performance of the classifier could significantly decrease and provides false information to doctors.	2 (Unlikely)	3 (Very serious)	Apply only trusted dataset.
Inexperienced doctor	The automatic sleep stage analysis result is for reference only. Absolutely obedience to the result could result in a false diagnosis.	2 (Unlikely)	3 (Very serious)	Acknowledge users of potential mistakes.
Inconsistent software environment	The program may not be able to work under different software environment (e.g., operating system, library version).	3 (Moderate)	2 (Serious)	Configure an identical environment before utilizing the program.
Unfamiliar sleep environment	Patients who are sensitive to in-lab sleep environment and the PSG equipment may be diagnosed with more severe sleep apnea.	4 (Likely)	2 (Serious)	Improve the sleep environment and assist with drugs.