



北京邮电大学
Beijing University of Posts and Telecommunications



Queen Mary
University of London



UNIVERSIDAD
POLITÉCNICA
DE VALENCIA

Undergraduate Project Report 2020/21

Synthetic data methods applied to sleep stage analysis

Name: Jingjing Wei
School: International School
Class: 2017215120
QM Student No.: 171044142
BUPT Student No.: 2017213152
Programme: Internet of Things
Engineering

Date: 3-5-2021

Table of Contents

Abstract.....	2
Chapter 1: Introduction	4
1.1 Motivation	4
1.2 Workflow	4
1.3 Contribution	5
Chapter 2: Background.....	7
2.1 EEG and ECG singals	7
2.2 Pre-processing of EEG and ECG singals	7
2.2.1 EEG Artefact Removal	8
2.2.2 EEG and ECG signal segmentation	9
2.3 Feature Extraction.....	10
2.3.1 EEG Signal Feature Extraction.....	10
2.3.2 ECG Signal Feature Extraction.....	11
2.4 Sleep Stage Classification	15
2.4.1 Sleep Stage Classes	15
2.4.2 Classification Methods	16
2.5 Generate synthetic data	16
Chapter 3: Design and Implementation	18
3.1 Dataset.....	18
3.1.1 Dataset Description	18
3.1.2 Training and Test Dataset.....	20
3.2 Generate synthetic data	21
3.2.1 Adding Gaussian noise.....	21
3.2.2 Using Surrogates.....	22
3.3 Sleep Stage Classification	25
3.2.1 Classification Method	25
3.2.2 Evaluation	26
3.3.3 Experiment Design.....	28
Chapter 4: Results and Discussion.....	29
4.1 Synthetic data.....	29
4.2 Sleep Stage Classification	31
Chapter 5: Conclusion and Further Work.....	45
References	47
Acknowledgement	49
Appendix	50
Risk and environmental impact assessment.....	67

Abstract

This paper applied synthetic data in the sleep stage classification problem. This paper used two methods to generate synthetic data, which are adding Gaussian noise and using surrogates. Pre-processing and feature extraction were done on an open-source EEG and ECG signals dataset published by University College Dublin. [16] The well-defined dataset contains EEG and ECG features of 10 subjects. The six classes of sleep stage labelled with numbers from 0 to 5. Training set and test set division following the "leave one out" procedure. Four classification methods were used, which are linear discriminant analysis, quadratic discriminant analysis, linear support vector machine, and quadratic support vector machine. The results showed that Linear SVM performs better. Different amounts of synthetic data generated by the two methods were combined with the original data to form new training sets. Applied linear SVM to these new training sets. 6 classes, 2 classes, 3 classes, and 4 classes classification problems were studied. The experiment results demonstrated that when the categories of classification problems are not too many, for instance, 2 classes and 3 classes, the accuracy of the sleep stage classification model can be improved by adding Gaussian replicates. If the categories of classification problems are too many, the accuracy of the sleep stage classification model decreased by adding Gaussian replicates. Surrogate samples of EEG-ECG features can be combined with real data of EEG-ECG features without a great decrease of classification performance of sleep staging when the added synthetic data accounts for no more than 50%. Moreover, adding a few surrogate samples, usually 5% of the original data can even improve the accuracy of the sleep stage classification model. Therefore, synthetic data can be used in the sleep stage classification problem when the sleep EEG and ECG data size is small.

摘要

本文在睡眠阶段分类问题中应用了合成数据。本文采用两种方法来生成合成数据，即加入高斯噪声和使用代理方法。在一个由都柏林大学出版的开源的数据集上进行了 EEG 和 ECG 数据的预处理和特征提取。[16] 定义好的数据集包含 10 名受试者的 EEG 和 ECG 信号特征。睡眠阶段的六个类别用 0 到 5 的数字标注。按照“留下一个”的原则划分训练集和测试集。在睡眠阶段分类问题中采用了四种分类方法，这四种方法分别为：线性判别分析、二次判别分析、线性支持向量机和二次支持向量机。结果表明，线性支持向量机具有更好的性能。将两种方法产生的不同数量的合成数据与原始数据相结合，形成新的训练集。将线性支持向量机应用于这些新的训练集。本文研究了 6 类、2 类、3 类和 4 类分类问题。实验结果表明，当分类问题的类别不太多，例如 2 类和 3 类时，加入高斯复制的数据可以提高睡眠阶段分类模型的精度。如果分类问题的类别太多，增加高斯复制的数据会降低睡眠阶段分类模型的精度。当添加的合成数据不超过 50% 时，EEG 和 ECG 特征的替代样本可以与 EEG 和 ECG 特征的真实数据相结合，而不会大大降低睡眠分期的分类性能。另外，在原始数据中加入少量的替代样本，通常为原始数据的 5%，甚至可以提高睡眠阶段分类模型的精度。因此，合成数据可以在 EEG 和 ECG 数据较少的情况下用于睡眠阶段的分类问题。

Chapter 1: Introduction

1.1 Motivation

Sleep apnea is a disease marked by abnormal breathing during sleep. The subject can have several micro-awakenings, also called microarousals, during sleep. Subject with sleep apnea has pauses in breath many times during their sleep period. These suddenly breathing pauses reduce the supply of oxygen, which lower the quality of sleep, leading to serious health consequences. We can detect microarousal through sleep stage classification, so as to find sleep apnea and evaluate the sleep quality of patients, providing support for the following treatment.

Electroencephalogram (EEG) and Electrocardiograph (ECG) signals are often used for sleep stages classification problem. [9] However, the acquisition of EEG and ECG signals is not easy, and the labelling of the sleep stage also takes a lot of time and manpower. In the process of EEG signals acquisition, it is often interfered by artefact from many sources. [1] In order to solve the scarce of EEG and ECG signals that can be used in sleep classification, this paper will use synthetic data. By observing the accuracy of the sleep stage classification model trained with the synthetic data, we can study whether the synthetic data can be applied to the sleep stage classification problem. The classification results may have little difference or might be improved.

1.2 Workflow

Figure 1 Workflow of this project shows the workflow of this project.

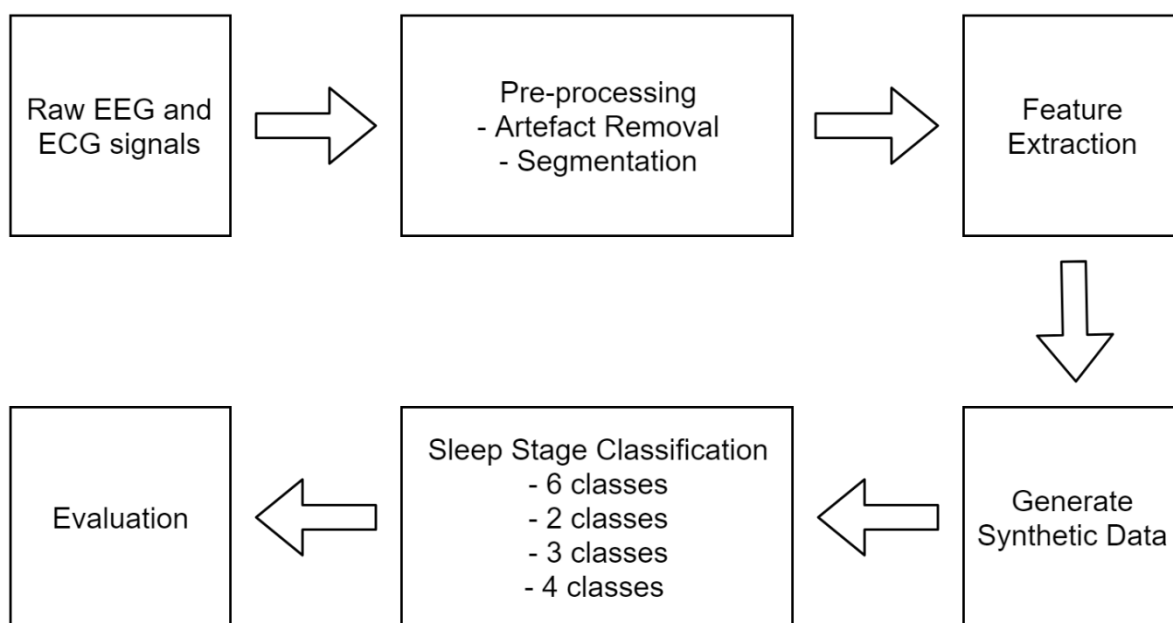


Figure 1 Workflow of this project

Synthetic data methods applied to sleep stage analysis

First, we get raw EEG and ECG signals from the open-source EEG and ECG signals dataset published by University College Dublin. [16] Then, do the pre-processing on raw EEG and ECG signals. Here, we mainly carry out the artefact removal of EEG signals and segmentation of both EEG and ECG signals. For the segmentation, each EEG and ECG epoch are 30 second long. Then we extract features from the EEG and ECG signals. Different methods are used to extract different features of EEG and ECG signals. After that, this paper used two data synthesis methods, which are adding Gaussian noise, and using surrogates to generate synthetic data on the defined dataset.

According to the American Academy of Sleep Medicine (AASM), sleep stages consist of 6 classes: wake (W), Stage I (N1), Stage II (N2), Stage III (N3), Stage IV (N4), and REM. [11] Different classification methods were used to classify the sleep stages into 6 or 2 or 3 or 4 classes. For 6 classes classification, class 0 is wake (W), class 1 is REM, class 2 is Stage I (N1), class 3 is Stage II (N2), class 4 is Stage III (N3), and class 5 is Stage IV (N4). For 2 classes classification, class 0 is wake (W), and class 1 is Stage I (N1), Stage II (N2), Stage III (N3), Stage IV (N4), and REM. For 3 classes classification, class 0 is wake (W), class 1 is REM, and class 2 is Stage I (N1), Stage II (N2), Stage III (N3), and Stage IV (N4). For 4 classes classification, class 0 is wake (W), class 1 is REM, class 2 is Stage I (N1) and Stage II (N2), and class 3 is Stage III (N3) and Stage IV (N4). The training data uses original data and the original data combined with different numbers of synthetic data.

Then, we evaluate the classification models. The accuracy and confusion matrix are used as the evaluation basis. The performance of classification models trained with original data and different numbers of synthetic data are compared so that the performance of using synthetic data in sleep stage classification problems can be studied.

1.3 Contribution

This paper applied synthetic data methods to generate synthetic sleep EEG and ECG data to solve the lack of the sleep EEG and ECG data and balance the stage imbalance in EEG and ECG signal. Four classification methods are used in the sleep stage classification, which are linear discriminant analysis, quadratic discriminant analysis, linear support vector machine, and quadratic support vector machine. The performances of these four classification methods in sleep stage classification are compared. This paper compares the performance of using the original dataset and the combined dataset which contains different numbers of synthetic data in the 6 classes, 4 classes, 3 classes, and 2 classes sleep stage classification problems.

Synthetic data methods applied to sleep stage analysis

Through the experimental results, we can conclude that linear SVM is the best of the four classification methods. The experiment results demonstrated that when the categories of classification problems are not too many, for instance, 2 classes and 3 classes, the accuracy of the classification model can be improved by adding Gaussian replicates. If the categories of classification problems are too many, the accuracy of the classification model decreased by adding Gaussian replicates. Surrogate samples of EEG and ECG features can be combined with real data of EEG and ECG features without a great decrease of classification performance of sleep staging when the added synthetic data accounts for no more than 50%. Moreover, adding a few surrogate samples, usually 5% of the original data can even improve the accuracy of the classification model. Therefore, synthetic data can be used in the sleep stage classification problem.

Chapter 2: Background

2.1 EEG and ECG signals

Electroencephalogram (EEG) reflects the electrophysiological activity of nerve cells in the cerebral cortex or scalp surface, which are located on the brain. EEG signal contains a lot of physiological and disease information. In clinical, EEG signals can not only provide diagnostic information for some diseases including sleep diseases but also provide effective treatment. **Electrocardiograph (ECG)** is a technique that uses electrocardiograph to record the electrical activity changes of the heart in each cardiac cycle. This paper uses EEG and ECG signals of subjects to classify sleep stage.

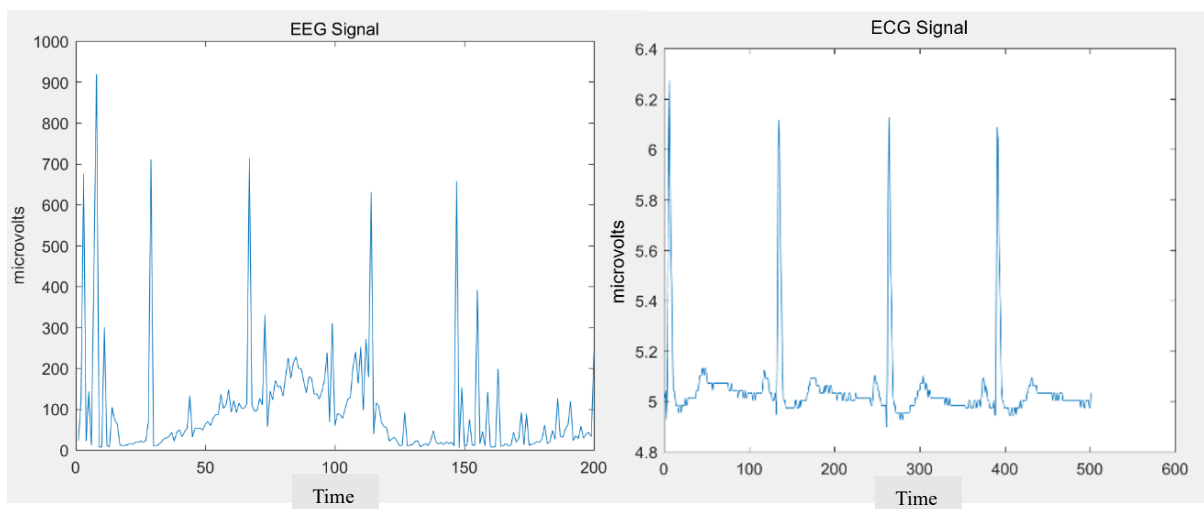


Figure 2 Example of EEG and ECG signal shows the example of EEG and ECG signals.

Figure 2 Example of EEG and ECG signal

2.2 Pre-processing of EEG and ECG signals

Before using EEG and ECG signals as train or test datasets for classification, it is necessary to pre-processing EEG and ECG signals. These pre-processing operations include normalization, calibration, detrending and equalization, etc. There are two important operations of pre-processing. The first is artefact removal. EEG signals are hard to interpret because of the artefacts. Thus, remove artefacts from EEG signals is essential for the following classification tasks. In signal processing, it is usually necessary to divide the non-stationary continuous signal into segmentations which are approximately stationary. After that, following operations such as feature extraction can be done on each segmentation. The following content introduces the

artefact removal of EEG signal and the segmentation of EEG and ECG signal.

2.2.1 EEG Artefact Removal

Sleep EEG recording is often subject to many different types of artefacts. The reasons for the occurrence of artefacts usually due to: (1) Eye movements which impact the electrical field. (2) Muscle movement. (3) Cardiac muscle depolarization causes electrical field changes. (ECG interference) (4) Hardware, such as power line interference. (5) Body part movement, such as head and chest movement. (6) Physiological reaction interference, such as sweat. These artefacts are unwanted, because they will have a negative impact on the classification problem. Thus, we need to detect and remove artefacts. Figure 3 Common EEG artefacts. shows the common artefact of EEG signals. [1]

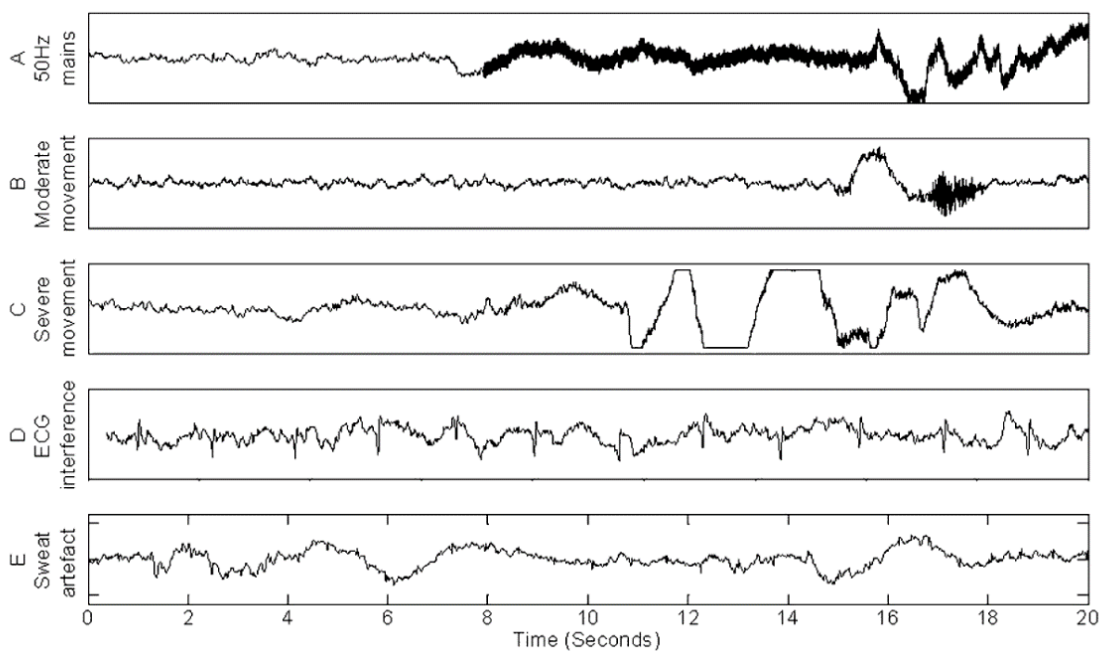


Figure 3 Common EEG artefacts.

The easiest way to remove artefacts is to detect the period of artefacts and then delete the artefact data during this time period. A way to distinguish artefact from normal EEG data is that if the amplitude of sleep EEG signal changes suddenly and lasts for more than 2 seconds, artefact appears. Therefore, artefacts can be found by finding the EEG signal with large amplitude variation. However, the EEG signal itself is non-stationary, which means norm EEG data also has amplitude change. Therefore, the energy operator can be used as an indicator of amplitude sudden change (such as peak), because energy is sensitive to instantaneous fluctuation but insensitive to subtle changes in the signal spectrum.

Another way to remove artefacts is to use filters of different frequency bands, such as low pass, high pass, bandpass and bandstop filters. However, the premise of this method is that EEG

Synthetic data methods applied to sleep stage analysis

signal and artefact are obviously distinct in the frequency domain. However, take EMC interference as an example, because EMG signals have a wide spectrum, it is difficult to separate them from EEG signals.

Another way to reduce the artefact is to measure the contaminating artefact signal first, and then remove the artefact by processing EEG signals and the reference signal. This method can be done through plenty ways including using time domain region, frequency domain region or adaptive filter. [1] The results of this method are very related to the quality of the reference signal, and the performance of this method in practical application is limited.

Another method is based on ICA (independent component analysis), which means decomposing EEG signals into random variables that are independently. EEG signals and artefacts are generated from different sources, thus they are independent signals. By comparing the components of these signals, artefact and EEG signal can be distinguished. Only the component signals related to EEG signals were needed when reconstruct EEG signals without artefacts. Many previous works use ICA in artefact reduction of EEG signals. [1]

2.2.2 EEG and ECG signal segmentation

The precondition of many signal processing methods is that the signal is stationary. But EEG and ECG signals are non-stationary. In order to meet the precondition of signal processing, one solution is to carry out signal segmentation in the time domain, each sub-section is approximately stationary. Signal segmentation can be uniform or non-uniform. For EEG and ECG signals, they usually divided into fixed duration segments, where duration is usually 30s for EEG signal and 30s and 5 minutes for ECG signal. [1, 4]

Non-uniform segmentation divide signal to suitable signal length. When the signal is stationary, segmentation will be longer, while the signal changes rapidly, segmentation will be shorter, so that each segment is approximately stationary, without sharp changes. Although non-uniform segmentation is a more reasonable method to divide signals, it has two disadvantages. The first is that this method increases the computational burden, and needs more computing resources. The other problem is this method requires signal analysis first, which leads to the fact that signal segmentation cannot be a separate pre-processing process, and the signal needs to be analysis thoroughly. The EEG and ECG datasets used in this paper use uniform segmentation and each segment is 30 seconds.

2.3 Feature Extraction

2.3.1 EEG Signal Feature Extraction

When we do the sleep stage analysis, we need to extract features from EEG signals. Features are the parameters that contain some information of EEG signals and can represent the structure of EEG signals. EEG signals have different kinds of features. The following section will introduce different feature extraction techniques according to different features.

(1) Temporal features

Temporal features are the parameters gained from the signal in the time domain. Instantaneous statistics are the simplest features, which can be obtained from a time series. These features include the measures obtained from the signal wave form, including mean absolute amplitude, variance, skewness and kurtosis, as well as measures related to the probability density function of the waveform, such as pattern, median or entropy.

Also, we can use Hjorth parameters to represent the EEG signals. There are three Hjorth parameters, defining activity, mobility and complexity of EEG signals respectively. [5] This paper used 3 Hjorth parameters: activity, mobility and complexity as 3 features.

Detrended fluctuation analysis (DFA) can be used to measure the long range temporal correlation in a time series. It is suitable for the analysis of non-stationary signal, for instance EEG signals. [6]

(2) Spectral features

These features are the most common extracted features of EEG signals, which are obtained from the frequency domain. EEG signals have 5 frequency bands, which are delta (0-4Hz), theta (5-7Hz), alpha (8-12Hz), sigma (13-15Hz), and beta (16-30Hz). The wake stage has high frequency band (>12 Hz) and low amplitude of energy. Non-REM sleep stage is characterized by high amplitude of energy (>100 μV) and low frequency band (<5 Hz). REM sleep stage is characterized by abundant 5-7 Hz frequency (theta) in EEG activity, and its amplitude is less than 100 μV . Therefore, it is very effective to analyze the sleep stage by analyzing the frequency of EEG signals. This paper used power in the frequency bands delta, theta, alpha, sigma and beta as 5 features separately.

There are numerous methods that can be used for extracted spectral features. Non-parametric spectral estimation methods are a form of spectral analysis. They are implemented through the Fast Fourier transform (FFT) algorithm. The normalised form of the cross-spectrum is

Synthetic data methods applied to sleep stage analysis

coherence, which can be applied on multiple signals. From this method, we can gain the level of synchronization between the frequency components of two EEG signals and the brain functional connectivity. [7]

(3) Time-frequency features

EEG signals can be decomposed into both time and frequency through time-frequency analysis. These features represent the frequency change according to time. Many sleep events, such as sleep arousals, sleep spindles, can be reflected from the time-frequency features. We can use short time Fourier transform (STFT), which is the commonest method of time-frequency analysis. First, EEG signals are segmented into uniform and short time period. Then, do the Fourier transform on each segment. From the result, we can know the spectral change from one time period to another.

(4) Nonlinear features

From traditional view, EEG signals are generated from stochastic processes, thus statistical features can represent those signals. However, a newly point of view suggests that EEG signals may be generated from a deterministic nonlinear process. Since EEG signals are usually non-stationary, analyze the nonlinear features can reveal features that can well-represent EEG signals. The fractal dimension (FD) measures the complexity of the EEG signals. Simple time series have lower fractal dimension, which means those time series have higher complexity. Another method of nonlinear features measurement is entropy measures. Entropy measure complexity statistically. Compute the entropy from the multivariable time series need to calculate the joint probability density function. However, those calculation need heavy computation resource, we can use approximate entropy (ApEn). If the signals have low complexity, they will have low ApEn value. [8]

2.3.2 ECG Signal Feature Extraction

As we described in 2.2.1, we need to do the feature extraction before signal analysis. Thus, we also need to extract features from ECG signals. Some feature extraction methods can be used both in EEG and ECG signals analysis. The following section will introduce some feature extraction methods of ECG signals.

(1) Singular value decomposition (SVD)

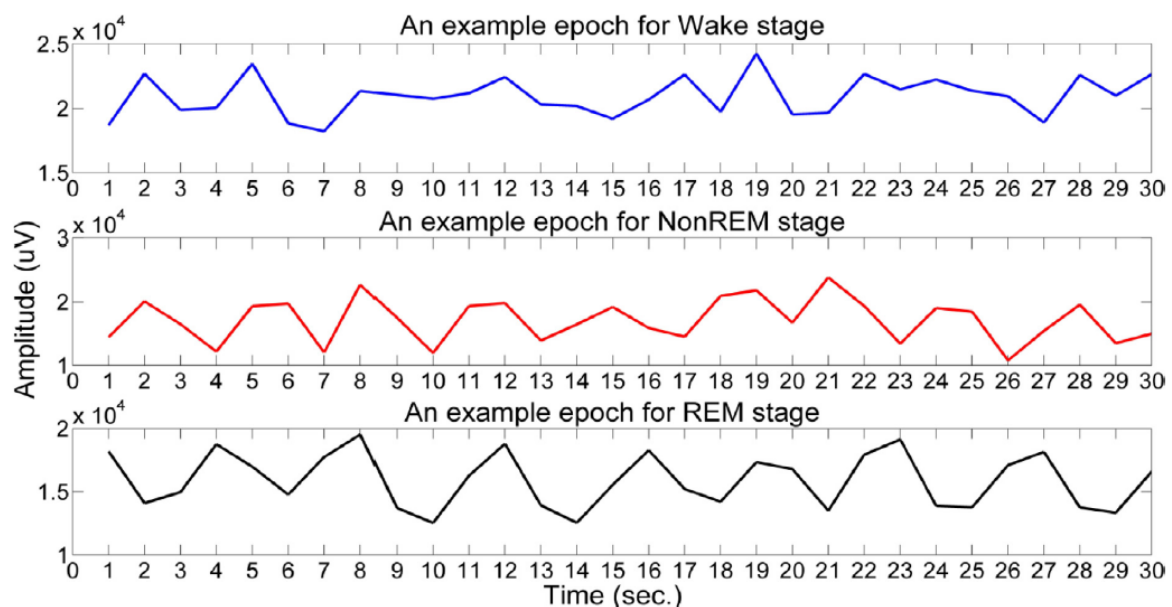
Matrix can be decomposed into three sub-matrix through this method. Then, matrix can be recomposed based on the three parts. Suppose A is the original matrix, and U , Λ and V is

Synthetic data methods applied to sleep stage analysis

the decomposed matrix. The following equation shows the relationship between the three matrices and the original matrix.

$$A = U\Lambda V^T \quad (1)$$

More specifically, Λ is the $m \times n$ diagonal matrix, where the diagonal components are the singular values of A . U is the $m \times m$ orthogonal matrix, where the columns of U are the left singular vectors. V is the $n \times n$ orthogonal matrix, where the columns of V are the right singular vectors. [9] This method transforms the input matrix into a more discernible matrix. When this method is applied to ECG signal feature extraction, SVD is carried out for each 1 second time period, aiming obtain more easily identified eigenvalues. Figure 4 An example SVD application for Wake, NonREM and REM stage. shows the



example of using SVD to extract features from Wake, NonREM and REM stage. [9]

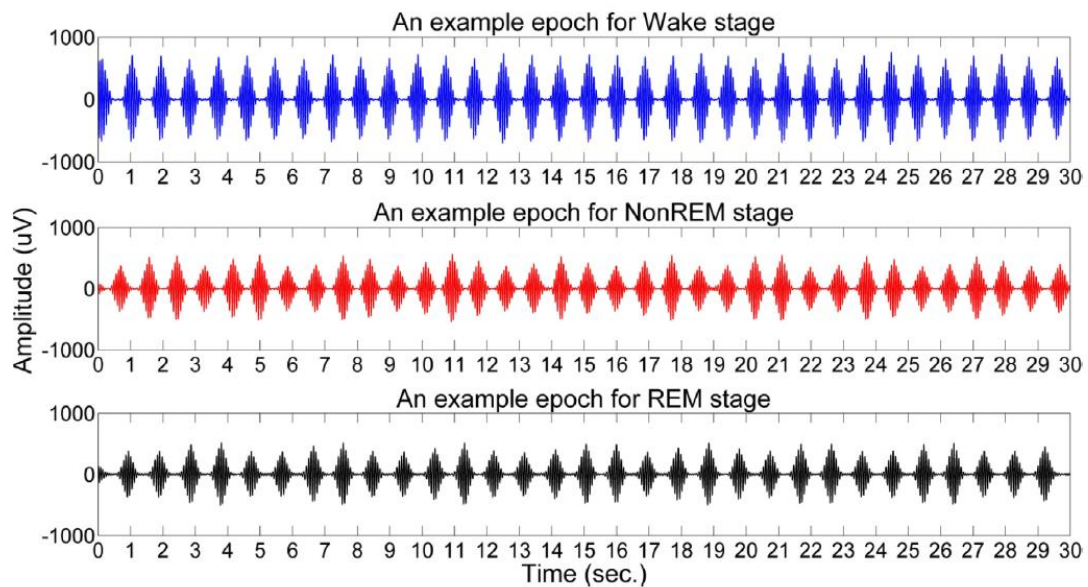
Figure 4 An example SVD application for Wake, NonREM and REM stage.

(2) Variational mode decomposition (VDM)

VMD decomposes the nonstationary signals in the time domain into subcomponents called Intrinsic Mode Function (IMF) iteratively. Let u_k be the initial mode and it contains low frequency components. The following modes contain higher frequency components. The purpose of variational mode decomposition is to decompose the signal into sub signals with different frequencies. Figure 5 An example VDM application for Wake, NonREM and REM

Synthetic data methods applied to sleep stage analysis

stage. shows the example of using VDM to extract features from Wake, NonREM and REM



stage. [9]

Figure 5 An example VDM application for Wake, NonREM and REM stage.

(3) Hilbert Huang transform (HHT)

HHT is an efficient and effective time-frequency analysis method. [9] Before apply HHT to non-stationary signals, we need to apply Empirical Mode Decomposition (EMD) to the signals. Through EMD, the original signals are decomposed into IMF components. Then, perform HHT on the IMF components. Therefore, HHT analysis the components of each time period of non-stationary signals.

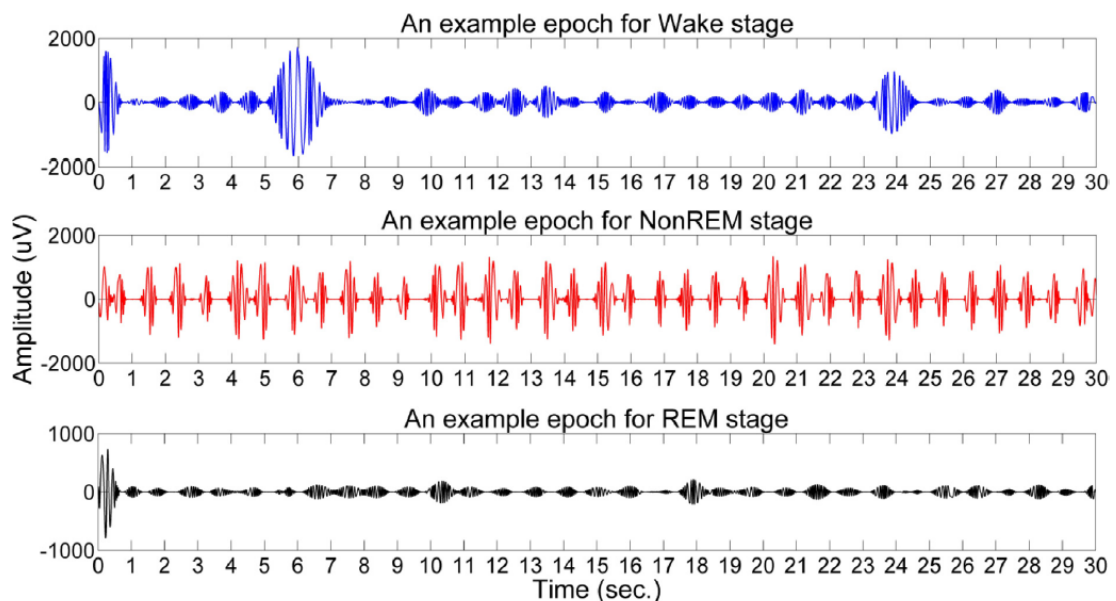


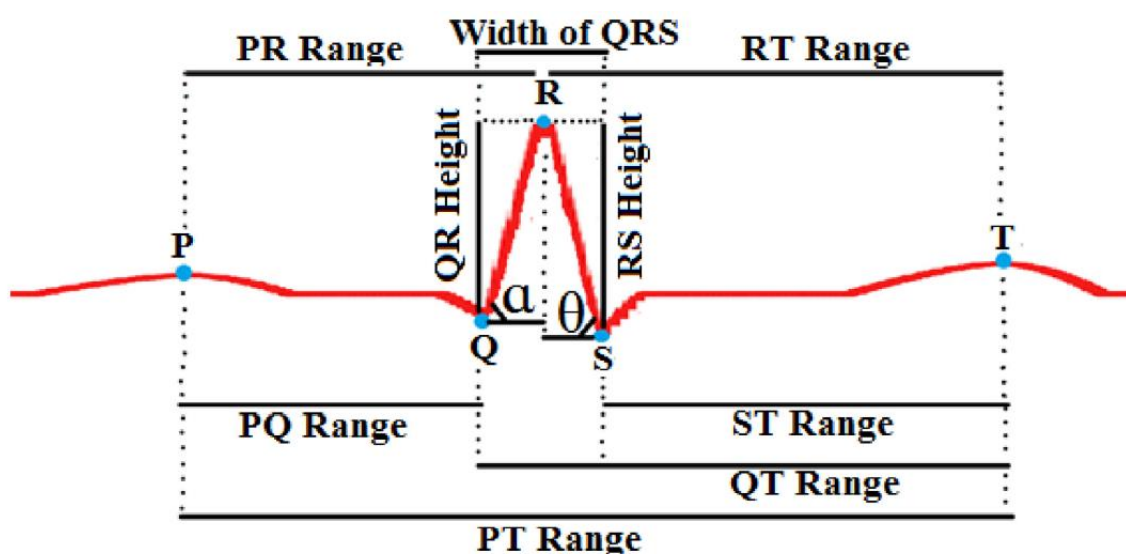
Figure 6 An example HHT application for Wake, NonREM and REM stage.

Figure 6 An example HHT application for Wake, NonREM and REM stage.错误!未找到

引用源。 shows the example of using HHT to extract features from Wake, NonREM and REM stage. [9]

(4) Morphological Method

According to the shape of ECG signals, an ECG signal has five points P, Q, R, S, and T. R is the peak of the ECG signal, which can be detected through Pan-Tompkins algorithm. [10] Then, P, Q, S and T can be identified. Using the x and y coordinates of each point, we can extract features from each segment, for instance, range between two points, height between two points, and the width between two points. As shown in Figure 7 Morphological components of an example ECG signal, 5 points divide the ECG signal and form different



signal segments. [9]

Figure 7 Morphological components of an example ECG signal

(5) Discrete Wavelet Transformation (DWT)

This method is a powerful time-frequency transformation method. It decomposes the original signal into two components, detail coefficients and approximation coefficients. Keep the detail coefficients unchanged, and decomposed the approximation coefficients into detail coefficients and approximation coefficients. Repeat this decompose process according to the need. This method reveals the local features of each time period. Wavelet packet decomposition (WPD) is based on DWT. WPD decomposes detail coefficients and approximation coefficients simultaneously. Therefore, WPD will not lose or change the information of the original signal.[15] Figure 8 Wavelet packet decomposition (WPD) for an ECG signal with three-level wavelet packet decomposition shows the example of using

three level WPD decomposition to extract features.

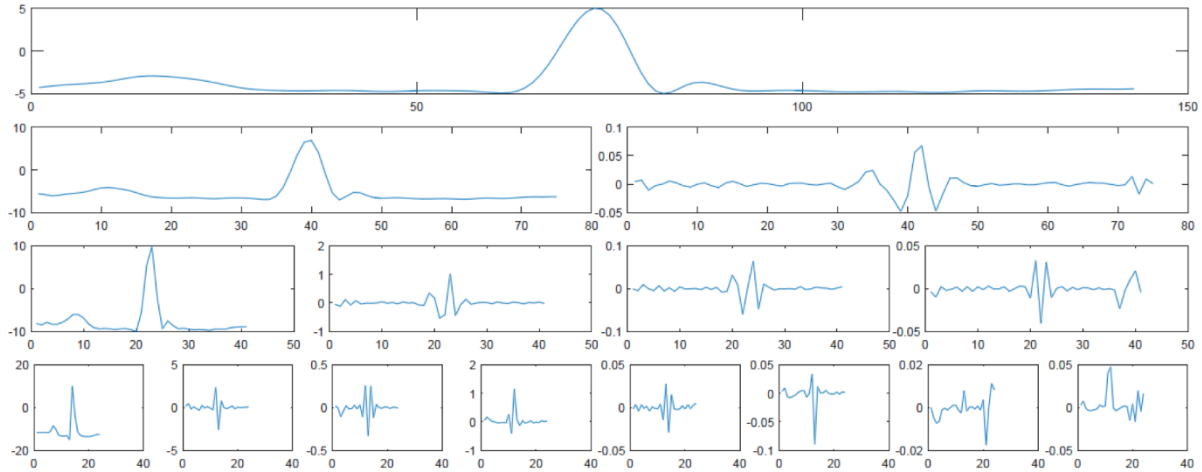


Figure 8 Wavelet packet decomposition (WPD) for an ECG signal with three-level wavelet packet decomposition

(6) Wavelet Packet Entropy (WPE)

Although the DWT and the WPD extract the local features of ECG signals, those features are large amount, so it is difficult to use them into classification. Thus, we need a high-level feature to describe ECG signals. Entropy is a suitable parameter to measure the uncertainty of the information content and it is a high-level feature. Typical type of entropy includes Shannon entropy (SE), and etc. WPE is a type of entropy used to extract features from WPT, which is computed based on energy.

2.4 Sleep Stage Classification

2.4.1 Sleep Stage Classes

Sleep contains two main stages, which are rapid eye movement (REM) and non-rapid eye movement (NREM). According to the American Academy of Sleep Medicine (AASM), non-rapid eye movement can be divided into 4 stages, which are Stage I (N1), Stage II (N2), Stage III (N3), and Stage IV (N4). [11] The 6 classes defined by AASM are wake (W), Stage I (N1), Stage II (N2), Stage III (N3), Stage IV (N4), and REM.

For 6 classes classification, class 0 is wake (W), class 1 is REM, class 2 is Stage I (N1), class 3 is Stage II (N2), class 4 is Stage III (N3), and class 5 is Stage IV (N4). For 2 classes classification, class 0 is wake (W), and class 1 is Stage I (N1), Stage II (N2), Stage III (N3),

Synthetic data methods applied to sleep stage analysis

Stage IV (N4), and REM. For 3 classes classification, class 0 is wake (W), class 1 is REM, and class 2 is Stage I (N1), Stage II (N2), Stage III (N3), and Stage IV (N4). For 4 classes classification, class 0 is wake (W), class 1 is REM, class 2 is Stage I (N1) and Stage II (N2), and class 3 is Stage III (N3) and Stage IV (N4).

2.4.2 Classification Method

The purpose of a classification problem is to find a function to determine the category of input data, which can be a binary classification problem or a multiple classes classification problem. classification problem predicts which class is the input value would belong to. Compared with the regression problem, the outputs of the classification problem are not continuous values, but discrete values. In this paper, we mainly study the classification of sleep stage, including six classes classification problem, 4 classes classification problems, 3 classes classification problem, and binary classification problems.

In the past, many papers mainly used supervised learning classification methods, including decision tree, Bayesian, support vector machine, and random forest.[9] The experiments of this paper use the following four classification methods, (i) linear discriminant analysis, (ii) quadratic discriminant analysis, (iii) linear support vector machine, and (iv) quadratic support vector machine.

LDA often produces robust classification results, while it is easy to implement due to its decision boundary is linear. However, linear decision boundaries may not suitable for too many classes. When the dimension is high, LDA needs too many parameters. QDA is an extension algorithm of LDA. It assumes that each class has its own covariance matrix. When the number of parameters is large, we need to use more parameters in QDA because for each class we need to calculate the covariance matrix. Support vector machine is another simple algorithm but can produce high accuracy with less computation power. Quadratic support vector machine is an extension algorithm of SVM.

2.5 Generate synthetic data

Synthetic data developed rapidly in the field of data science. As the word "synthesis" implies, synthetic data are generated through programs, not by documents of real events. The main problem in data science is data collection and processing. Usually, obtaining a large amount of data that can train an accurate model is not easy. Manually labelling data is a costly and slow way to obtain data. Especially for sleep EEG and ECG signals, the cost of collecting, pre-processing, and labelling the sleep stage is very high. However, these problems can be solved

Synthetic data methods applied to sleep stage analysis

by using synthetic data. By comparing the results of sleep stage classification using the original data and the combined data contains original data and synthetic data, we can verify that the synthetic data can be used in EEG and ECG signal analysis, including but not limited to sleep stage classification.

This paper mainly uses two methods to generate synthetic data. One is adding Gaussian noise to real data to generate replicates samples of the original data. The probability density function of Gaussian noise following Gaussian distribution. Using Gaussian replicates can improve the performance of the estimator used for classification problems. [13] The other is using surrogates to generate synthetic data samples, which is a way of synthesizing multivariable time series with prescribed covariance function and marginal distributions, obtained from empirical results. The synthetic data has a similar covariance function and marginal distributions with the input multivariable time series.

Chapter 3: Design and Implementation

3.1 Dataset

3.1.1 Dataset Description

In this paper, I used an open-source EEG and ECG signals dataset published by University College Dublin. [16] This database contains 25 full overnight EEG signals and three-channel ECG signals of 28 subjects. Signals recorded in this dataset that will be used are 8 channel EEG signals and three-channel Holter ECG. The dataset stores EEG signals and ECG signals in EDF format. The class labels of this dataset are 0 for Wake, 1 for REM, 2 for Stage I, 3 for Stage II, 4 for Stage III, 5 for Stage IV, 6 for Artifact, and 7 for Indeterminate. We were only interested in class 0 to class 5. We use the EEG and the ECG signals in this dataset, applying the pre-processing and the feature extraction processes.

The well-defined dataset contains the extracted features of EEG and ECG signals from the dataset described above. There are 10 subjects. The six classes labelled with numbers were shown in Table 1: Labels of 6 sleep stage classes. 0 for wake, 1 for REM, 2 for stage I, 3 for stage II, 4 for stage III, and 5 for stage IV.

Table 1: Labels of 6 sleep stage classes

Sleep Stage	Label
Wake (W)	0
REM	1
Stage I (N1)	2
Stage II (N2)	3
Stage III (N3)	4
Stage IV (N4)	5

Synthetic data methods applied to sleep stage analysis

The following variables were extracted from the dataset and each sleep epoch (one epoch 30s) is classified. The first variable is “classes_sub”, which contains class information of each epoch of each subject. “features_ecg” includes all the ECG features of 10 subjects, and each epoch has 30 features. “features_eeg” includes all the EEG features of 10 subjects, and each epoch has 8 features. “features_sub” contains 2 columns. The first column contains 8 EEG features and the second column contains 30 ECG features. “subs” represent 10 subjects. Figure 9 Dataset Structure shows the dataset structure.

名称 ^	值
classes_sub	10x1 cell
features_ecg	8415x30 double
features_eeg	8415x8 double
features_sub	10x2 cell
subs	[1,2,3,4,5,6,7,8,9,10]

Figure 9 Dataset Structure

The 8 features of EEG signals were shown in Table 2: 8 EEG features. The 30 features of ECG signals were shown in Table 3: 30 EEG features.

Table 2: 8 EEG features

Index 1	Power in the frequency band delta (0-4 Hz)
Index 2	Power in the frequency band theta (5-7 Hz)
Index 3	Power in the frequency band alpha (8-12 Hz)
Index 4	Power in the frequency band sigma (13-15 Hz)
Index 5	Power in the frequency band beta (16-30 Hz)
Index 6	Hjorth parameter: activity
Index 7	Hjorth parameter: mobility
Index 8	Hjorth parameter: complexity

Table 3: 30 EEG features

Index 1-4	Autoregressive (AR) coefficients
Index 5-20	Shannon entropy
Index 21-30	Wavelet variance estimates

Synthetic data methods applied to sleep stage analysis

As described in 2.3.1, EEG features used in this paper with index from 1 to 5 are the spectral features extracted from the frequency domain. This paper used power in the frequency band delta, theta, alpha, sigma, and beta as five EEG features separately. Moreover, this paper used 3 Hjorth parameters: activity, mobility, and complexity as three EEG features indexed from 6 to 8. These features are temporal features extracted from the time domain.

EEG features indexed from 1 to 4 are the autoregressive coefficients. Multivariable time series can be estimated by a linear weighted sum of previous terms in the series. The weights are the autoregression coefficients. AR coefficients can be computed through the least-squares method. EEG features indexed from 5 to 20 are the Shannon entropy. The entropy measures the uncertainty of the information content and it is a high-level feature. Shannon entropy (SE) is a typical type of entropy. EEG features indexed from 21 to 30 are obtained from wavelet variance analysis. The wavelet variance decomposes the variance of a time series.

3.1.2 Training and Test Dataset

Before using EEG and ECG signal features to classify the sleep stage, it is necessary to divide the dataset into the training set and the test set. The training set is used to train the model, and the parameters of the model are fitted by the data of the training set. The test set is used to test the accuracy of the trained model. Reasonable division of training set and test set can prevent overfitting of model and improve the accuracy of prediction results.

The training set and test set division in this paper follows one principle called the “leave one out” procedure. This procedure means EEG and ECG features of each subject turn out to be a test set, and the rest of the features of the rest subjects are used as a training set. Thus, we obtain ten different data set splitting results. We can obtain the results of classification for each subject one by one. Table 4: Training sets and test sets shows the ten different training sets and test sets. The number of the test subject defines those ten different training sets and test set combinations. We will use the test subject number to indicate the training set and test set shown in Table 4: Training sets and test sets in the following part of the paper.

Table 4: Training sets and test sets

	Training set (Subjects number)	Test set (Subjects number)
1	2, 3, 4, 5, 6, 7, 8, 9, 10	1
2	1, 3, 4, 5, 6, 7, 8, 9, 10	2

3	1, 2, 4, 5, 6, 7, 8, 9, 10	3
4	1, 2, 3, 5, 6, 7, 8, 9, 10	4
5	1, 2, 3, 4, 6, 7, 8, 9, 10	5
6	1, 2, 3, 4, 5, 7, 8, 9, 10	6
7	1, 2, 3, 4, 5, 6, 8, 9, 10	7
8	1, 2, 3, 4, 5, 6, 7, 9, 10	8
9	1, 2, 3, 4, 5, 6, 7, 8, 10	9
10	1, 2, 3, 4, 5, 6, 7, 8, 9	10

3.2 Generate synthetic data

This paper uses two methods to generate synthetic data, which are (i) adding Gaussian noise to real data to produce replicates of them, and (ii) to generate synthetic data samples using surrogates, which is a method of synthesizing multivariable time series with prescribed properties that obtained from input multivariable time series. Specifically, the empirical properties are covariance function and marginals. The principle and specific implementation of the two methods is as follows.

3.2.1 Adding Gaussian noise

This method adds Gaussian noise to real EEG and ECG data to produce replicates of them. The probability density function of Gaussian noise following Gaussian distribution. The Gaussian distribution function is shown in equation (2), where μ is mean and σ^2 is variance.

$$P(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}} \quad (2)$$

Before adding Gaussian noise to the original data, we need to determine the mean and variance, which determines the different decisions of Gaussian noise (different signal to noise ratios). The bandwidth of Gaussian noise is infinite, and the noise is evenly distributed in the whole frequency band, so from the spectrum point of view, it just raises the spectrum of the original signal a little.

Use “`imnoise(I, 'gaussian', M, V)`” command in MatLab can add Gaussian noise to the dataset. This command means adding Gaussian white noise with constant mean M and variance V. For example, we take the EEG and ECG features of subject 3 as the test set, and the EEG and ECG

Synthetic data methods applied to sleep stage analysis

features of the remaining 9 subjects as the training set. We add Gaussian noise with constant mean $M = 0.1$ and variance $V = 0.002$ to the training set.

Before we add gaussian noise data, the data of the dataset have to be normalized using the command “zscore”.

3.2.2 Using Surrogates Method

This method used in this paper generates stationary multivariate time series with prescribed covariance function and marginal distributions.[14] Covariance function and marginal distributions can be prescribed either through a mathematic model or by empirical properties that obtained from measured multivariable time series. This paper uses the original EEG and ECG data as input to obtain the empirical covariance function and marginal distributions used as the prescribed properties. Suppose $X(n)$ is a multivariate time series that has M component. $x_j(n)$ ($n=0, \dots, N-1; j=1, \dots, M$) is the j^{th} component. The $X(n)$ is shown as followed.

$$X(n) = \begin{pmatrix} v_{0,0} & \cdots & v_{0,N-1} \\ \vdots & \ddots & \vdots \\ v_{M-1,0} & \cdots & v_{M-1,N-1} \end{pmatrix} \quad (3)$$

The multivariate surrogate Algorithm 0 is described below.[14]

Input: $A_{x_j}(f)$: Amplitude. $\Psi_{x_j}(f)$: Phase ($j=1, \dots, M$) (From empirical properties of measured series)

1. Generate a random phase $\Phi(f)$, which follows independent and identically distributed of uniform distribution in $[0, 2\pi]$.
2. For each component j , obtain $S_j(n)$ through the Fourier Transform and add the $\Phi(f)$.
3. Generate the multivariate time series surrogate $S(n) = [S_1(n), \dots, S_M(n)]^t$

Output: S

The pre-algorithm of multivariate surrogate method with prescribed covariance function and marginals distribution is Algorithm IAAFT as shown in below. [14]

Input: $A_{x_j}(f)$: Amplitude. $\Psi_{x_j}(f)$: Phase ($j=1, \dots, M$). $x_j(n)$ ($n = 0, \dots, N - 1, j = 0, \dots, M$) (From empirical properties of measured series)

1. Generate surrogate $r_j(n)$ from Algorithm 0 described above. The prescribed values $x_j(n)$

Synthetic data methods applied to sleep stage analysis

are the rank-ordered values of x , which means ($x = \text{sort}(x_j(n))$). Apply two steps iteratively. For iteration l the two step is shows below.

(1): Project on the prescribed covariance. Replace the amplitudes by the $A_{x_j}(f)$, while keep the phase $\Psi_{x_j}(f)$ of this iteration. Obtain s_j .

(2): Project on the prescribed marginal distributions. For each component, apply the rank ordering mapping with the prescribed values $x_j(n)$ to get r_j . Rank ordering mapping means that for S_j , $\text{rank}(S_j(n) = k)$, if $S_j(n)$ is the k^{th} small value.

2. Stop iterations when convergence, which means $R \simeq S$, or R or S does not change. Define the surrogates $R(n) = [r_1, \dots, r_M]^t$ and $S(n) = [s_1, \dots, s_M]^t$.

Output: S and/or R .

The multivariate surrogate Algorithm with prescribed covariance function and marginals is described in the following Algorithm 1. [14]

Input: $C_{j_k}(n)$: covariance. $p_j(v)$: marginal distributions. (From empirical properties of measured series)

1. Compute $A_{x_j}(f)$ and $\Psi_{x_j}(f)$ through Fourier transform of each component j ($j=1, \dots, M$).
2. Draw $x_j(n)$ ($n = 0, \dots, N - 1, j = 0, \dots, M$) from $p_j(v)$.

Sort values: $x_j = \text{sort}(x_j)$

3. Get surrogate R from Algorithm 0.
4. Strat Algorithm IAAFT with R .
5. Stop iterations when convergence.

Output S and/or R

错误!未找到引用源。 shows an example of using Algorithm 1 to generate synthetic data. [14]

(a) shows a time series ($N=500$). Red lines in (b) are the prescribed marginal distribution. Black lines in (c) and (d) are surrogates and red lines are prescribed covariances. Specifically, (c) shows auto-covariances and (d) shows cross-covariance. [14]

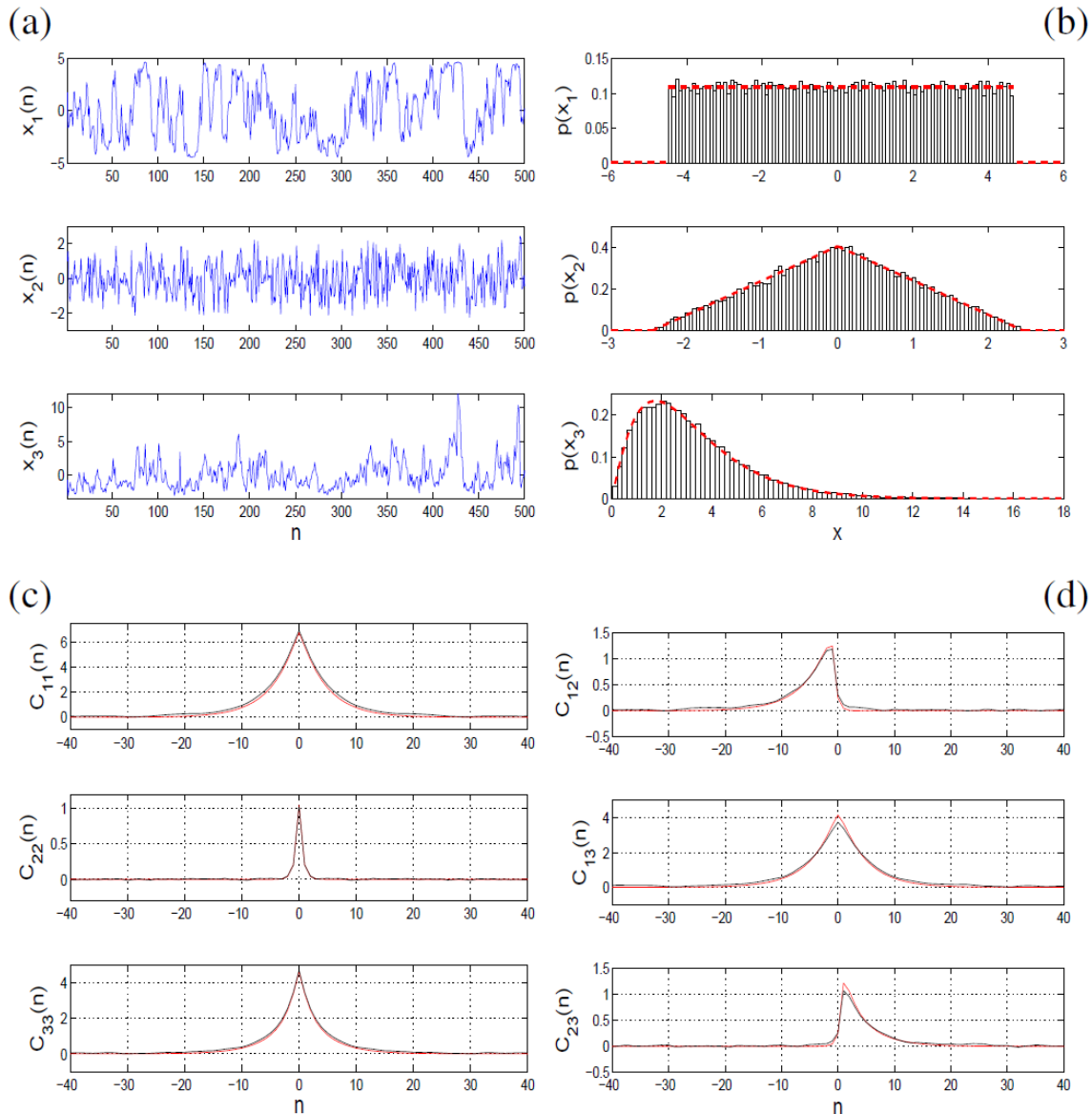


Figure 10 Example results of using Algorithm 1

Use the open-source MatLab code of the Algorithm 1 mention above to generate synthetic data. The open-source code can be found in [17]. For example, we take the EEG and ECG features of subject 3 as the test set, and the EEG and ECG features of the remaining 9 subjects as the training set. Similarly, apply the command “zscore” on the dataset to normalize the data. Take the test set as the input multivariable time series to obtain the empirical properties. Then, synthetic data can be generated with prescribed covariance function and marginal distributions obtained from the last step.

3.3 Sleep Stage Classification

3.3.1 Classification Method

Four classification methods are used for sleep stage classification in this paper, which are linear discriminant analysis, quadratic discriminant analysis, linear support vector machine and quadratic support vector machine. Detailed of these algorithms are shown in below.

(1) Linear Discriminant Analysis (LDA)

This is a linear classifier that divides the features into different classes by hyper-planes. The basic idea of LDA is to project high-dimensional pattern samples into the best discriminant vector space to extract classification information and compress the dimension of feature space. After projection, the maximum inter class distance and minimum intra class distance of pattern samples in the new subspace are guaranteed, that is, the pattern has the best separability in the space. LDA is easy to implement and performs well in classifying linear data. LDA and the classification method based on LDA are widely used in EEG and ECG signals analysis.

(2) Quadratic Discriminant Analysis

Quadratic discriminant analysis is an extension algorithm of LAD, which is suitable for the classification problem with quadratic data. Similar to linear discriminant analysis, quadratic discriminant analysis is another linear discriminant analysis algorithm. They have similar algorithm characteristics. The only difference is that linear discriminant analysis is used when the covariance matrix of different classification samples is the same; quadratic discriminant analysis is used when the covariance matrix of different classification samples is different. That means each class has its own covariance matrix.

(3) Linear Support Vector Machine (SVM)

Support vector machine (SVM) is a kind of generalized linear classifier that classifies data in the way of supervised learning. The objective of the support vector machine algorithm is to find a hyperplane in N-dimensional space (N is the number of features) that classifies the data samples. Its decision boundary is the maximum margin hyperplane of learning data samples. While SVM is a linear classifier, it is associated with the nonlinear function of the features and in the original feature space, the decision boundaries are nonlinear. This method is flexible and robust. Many automatic sleep staging problems using SVM in EEG and ECG signals analysis. [12]

(4) Quadratic Support Vector Machine

Quadratic support vector machine is an extension algorithm of SVM, which is suitable for the classification problem with quadratic data. It uses quadratic decision function which can separate data nonlinearly. In order to separate data nonlinearly, dual optimization form and kernel technique must be used.

In this paper, the four classification methods are implemented through MatLab. Figure 11 Part of the code of Linear SVM shows part of the MatLab code that implement linear SVM.

```

1  function [trainedClassifier, validationAccuracy] = trainClassifier(trainingData)
2
3  inputTable = array2table(trainingData, 'VariableNames', {'column_1', 'column_2', 'column_3', 'colu
4
5  predictorNames = {'column_1', 'column_2', 'column_3', 'column_4', 'column_5', 'column_6', 'column
6  predictors = inputTable(:, predictorNames);
7  response = inputTable.column_39;
8  isCategoricalPredictor = [false, false, false, false, false, false, false, false, false, false, f
9
10 % Train a classifier
11 % This code specifies all the classifier options and trains the classifier.
12 template = templateSVM(...
13     'KernelFunction', 'linear', ...
14     'PolynomialOrder', [], ...
15     'KernelScale', 'auto', ...
16     'BoxConstraint', 1, ...
17     'Standardize', true);
18 classificationSVM = fitcecoc(...
19     predictors, ...
20     response, ...
21     'Learners', template, ...
22     'Coding', 'onevsone', ...
23     'ClassNames', [0; 1; 2; 3; 4; 5]);
24
25 % Create the result struct with predict function
26 predictorExtractionFcn = @(x) array2table(x, 'VariableNames', predictorNames);
27 svmPredictFcn = @(x) predict(classificationSVM, x);
28 trainedClassifier.predictFcn = @(x) svmPredictFcn(predictorExtractionFcn(x));
    
```

Figure 11 Part of the code of Linear SVM

3.3.2 Evaluation

Accuracy can be used to indicate the performance of the classification model. The accuracy is calculated by comparing the predict classes with the correct classes. The following equation shows how to compute accuracy.

$$accuracy = \frac{correct}{correct+false} \quad (4)$$

“correct” means the number of predict classes that match the correct classes, while “false” means the number of predict classes that do not match the correct classes. The higher the

Synthetic data methods applied to sleep stage analysis

accuracy, the more accurate the prediction of sleep stage can be obtained and the model has better performance. The prediction results and the correct results can be compared through MatLab code, and the accuracy of each model can be calculated. Figure 12 shows part of code that calculate accuracy.

```

%%% subject 3
correct=0;
false=0;
for i=1:813 %sub1-749 sub2-882 sub3-813
    if result(i)==testclass(i)%比较预测结果和正确结果
        correct=correct+1;
    else
        false=false+1;
    end
end
acc1_3=correct/813;%sub1-749 sub2-882 sub3-813

```

Figure 12 part of code that calculate accuracy

Another indicator used to evaluate the model is the confusion matrix. Confusion matrix is a matrix that used for accuracy evaluation, which has n rows and n columns. Each column of the confusion matrix represents each class separately, and the total number of each column represents the number of data predicted to be this class. Each row represents the correct class of input data, and the total number of each row represents the number of data belongs to this class. The values in each column and each row represent the number of data that belongs to the class of this row predicted to be the class of this column. For the most common binary classification, its confusion matrix is 2 rows and 2 columns. [错误!未找到引用源。](#) shows

Confusion matrix		Predicted Class	
		Predicted Value : Positive (+)	Predicted Value Negative (-)
Actual Class	Actual Value : Positive (+)	TP True Positive	FN False Negative
	Actual Value : Negative (-)	FP False Positive	TN True Negative

the confusion matrix of binary classification.

Figure 13 confusion matrix of binary classification

The confusion matrix can be generated using “confusionmat(predict_result, correct_result)”

Synthetic data methods applied to sleep stage analysis

command in MatLab. “predict_result” represent the predict classes generated through classification model and test set. “correct_result” represent the correct classes of each epoch in the test set.

3.3.3 Experiment Design

Four classification methods are applied to 10 training sets described in 3.1.2. By comparing the accuracy of sleep class prediction results of the models trained on the same training set, the performance of the four classification methods in sleep stage 6 classes classification problem using EEG and ECG features are analyzed.

After that, the classification method with better performance and a training set and test set with higher performance are selected for the later classification problem. Applied two methods of generating synthetic data to the original EEG-ECG features. Different numbers of the synthetic EEG-ECG features are combined with the original EEG-ECG features, forming new training sets. Apply the classification method with higher accuracy on 6 classes, 2 classes, 3 classes, and 4 classes sleep stage classification problems. After that, sleep stage classes were predicted using the test set. Compare the results of using original data as the training set, using original data and synthetic data generated by adding Gaussian noise, and using original data and synthetic data generated by using the surrogate method as the training set. Then, analyze the performance of adding different numbers of synthetic EEG-ECG features in different sleep stage classification problems. So that, we can find out whether the synthetic data can be used in the sleep stage classification problem.

Chapter 4: Results and Discussion

4.1 Synthetic Data

错误!未找到引用源。 shows part of the original data, while 错误!未找到引用源。 shows part of synthetic data by adding Gaussian noise with constant mean $M = 0.1$ and variance $V = 0.002$.

	1	2	3	4	5	6	7	8	9	10
1	-0.4955	-0.5108	-0.5756	-0.4422	0.4849	-0.3992	1.6370	0.7483	-0.2815	0.2559
2	-0.0175	-0.3252	-0.3322	0.1407	1.8855	0.0392	1.5266	1.9491	-0.2380	0.1428
3	2.7267	1.4967	3.4898	4.0287	4.3072	2.7363	0.1348	1.9328	0.5363	-1.2734
4	-0.4997	-0.4637	-0.4748	-0.3843	-0.1732	-0.3884	0.0504	1.1015	-0.3771	0.3680
5	0.0994	-0.3478	-0.4408	-0.3946	-0.1319	2.2584	-1.2523	1.9500	0.0849	-0.2978
6	-0.5497	-0.5203	-0.5303	-0.4884	-0.2433	-0.3562	-0.2697	1.3398	-0.2558	0.3006
7	1.6130	1.9175	0.2801	0.7282	0.9484	4.3031	-0.7845	2.2602	0.5347	-1.3691
8	3.9278	0.6366	0.5106	0.7625	2.3928	2.3537	-0.5322	1.6433	0.1735	-0.6187
9	-0.5702	-0.4445	-0.1457	-0.4829	-0.2069	-0.4420	1.1432	1.1832	0.5370	-0.8745

Figure 14 Part of the original data

	1	2	3	4	5	6	7	8	9	10
1	0.1240	0.0800	0.0743	0.1621	0.5368	0.0536	1	0.8106	0.1067	0.3465
2	0.1820	0.1010	0.1076	0.2415	1	0.1878	1	1	0.0723	0.2227
3	0.9990	1	1	1	1	1	0.2804	1	0.6781	0.1028
4	0.1386	0.1037	0.1122	0.1128	0.0881	0.1816	0.1248	1	0.0679	0.5384
5	0.2137	0.1010	0.0618	0.1207	0.1732	1	0.0798	1	0.2415	9.0704e-04
6	0.0415	0.1160	0.1264	0.0814	0.1187	0.1315	0.1264	1	0.0750	0.3796
7	1	1	0.3727	0.8699	1	1	0.0773	1	0.6219	0.0836
8	1	0.7108	0.5185	0.8654	1	1	0.1528	1	0.2353	0.0674
9	0.2600	0.1157	0.1086	0.1483	0.1193	0.1550	1	1	0.6044	0.0994

Figure 15 Part of the synthetic data generated by adding Gaussian noise

错误!未找到引用源。 shows part of the synthetic data generated by using surrogates. The

	1	2	3	4	5	6	7	8	9	10
1	-0.4138	-0.0930	-0.4947	0.0981	0.0105	-0.3949	1.0166	0.3077	-0.2989	0.5416
2	-0.5472	-0.5323	-0.2484	-0.3006	-0.2500	-0.4120	1.7835	1.5389	-0.0920	0.2562
3	-0.4635	-0.2901	-0.7872	-0.0185	0.0261	-0.4182	0.7015	1.6253	0.0998	0.3359
4	-0.5269	-0.0249	1.5568	-0.0340	0.1549	-0.3622	0.5506	1.5561	-0.0944	0.2040
5	-0.5935	0.1183	2.7260	0.9271	-0.2271	-0.5177	0.1843	0.4015	-0.2351	0.2374
6	-0.2074	-0.3607	-0.4698	-0.0708	-0.2829	-0.2233	1.7202	1.7046	-0.1958	0.1994
7	-0.1571	0.2173	-0.0837	-0.5373	-0.1502	-0.2030	-0.1314	0.0954	-0.2354	0.4211
8	-0.2093	0.1134	-0.3600	-0.0349	-0.7566	-0.4227	1.1848	-0.2882	-0.1385	0.3575
9	-0.3803	0.0139	-0.5726	-0.3414	-0.1275	-0.3090	0.4395	-0.1819	-0.1169	0.2435
10	-0.1961	-0.3748	0.1204	-0.6149	-0.2658	-0.0627	0.4318	0.8706	8.1921e-04	0.1464
11	-0.0889	-0.4910	0.0798	-0.4068	-0.2059	-0.1558	0.1431	1.0611	0.1555	0.0895
12	0.5012	0.0990	0.5126	-0.1485	-0.1002	-0.4005	0.0945	-0.4503	-0.0346	0.1917
13	-0.4499	-0.1308	-0.2979	-0.0699	-0.3663	3.1637	0.0583	-0.6010	-0.1630	0.0522
14	-0.0016	-0.4067	0.0802	-0.3497	-0.6789	1.0089	0.9477	-0.3525	-0.1953	0.1877

Synthetic data methods applied to sleep stage analysis

original data is shown in 错误!未找到引用源。 .

Figure 16 Part of the synthetic data generated by using surrogates

Synthetic data methods applied to sleep stage analysis

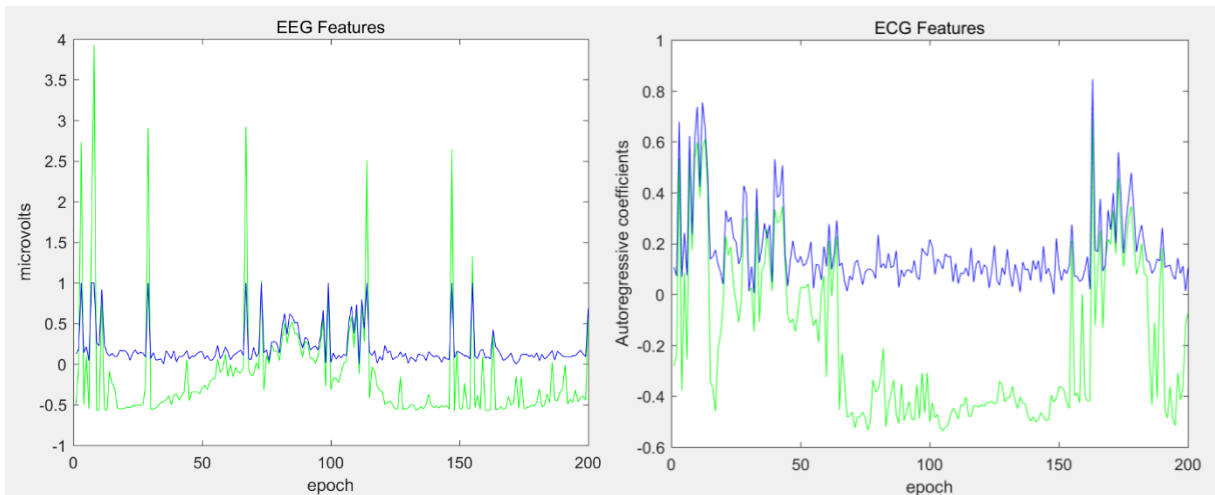


Figure 17 Part of original EEG and ECG features and synthetic EEG and ECG features generated by adding Gaussian noise

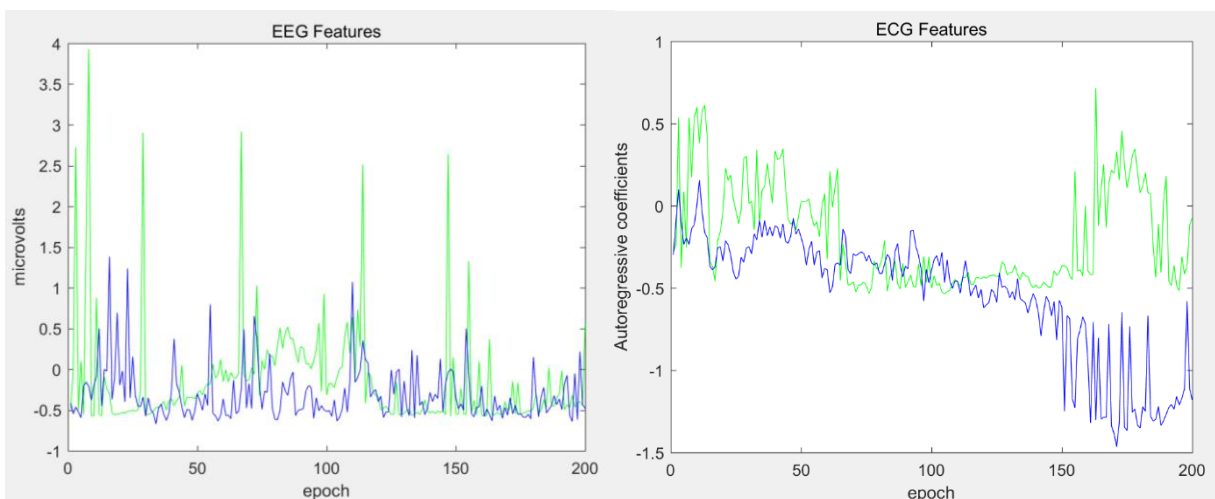


Figure 18 Part of original EEG and ECG features and synthetic EEG and ECG features generated by surrogate

Figure 19 shows part of original EEG and ECG features and synthetic EEG and ECG features generated by adding Gaussian noise. The figure on the left shows 200 EEG features with index 1. The green line is the original features, and the blue line is the synthetic features generated by adding Gaussian noise. The figure on the right shows 200 ECG features with index 1. The green line is the original features, and the blue line is the synthetic features generated by adding Gaussian noise. Figure 19 shows part of original EEG and ECG features and synthetic EEG and ECG features by surrogate. The figure on the left shows 200 EEG features with index 1. The green line is the original features, and the blue line is the synthetic features generated by surrogate. The figure on the right shows 200 ECG features with index 1. The green line is the original features, and the blue line is the synthetic features generated by surrogate.

4.2 Sleep Stage Classification

The accuracy of each classification method applied to 10 training sets described in 3.1.2 was shown in Table 5: . The results came from EEG-ECG features.

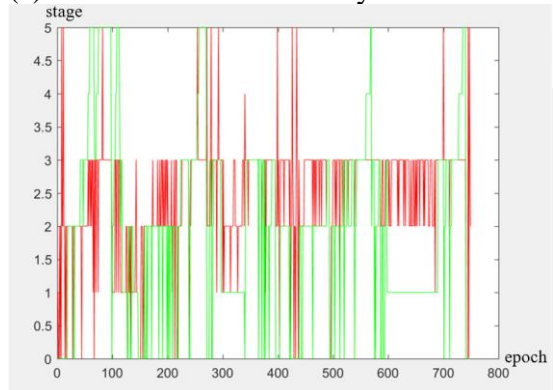
Table 5: Accuracy of 4 classification method using 10 training set (6 classes classification)

Test subject	Linear Discriminant Analysis	Quadratic Discriminant Analysis	Linear Support Vector Machine	Quadratic Support Vector Machine
1	0.374	0.280	0.455	0.369
2	0.564	0.460	0.620	0.610
3	0.642	0.221	0.705	0.604
4	0.202	0.156	0.213	0.208
5	0.405	0.277	0.352	0.404
6	0.140	0.396	0.383	0.334
7	0.529	0.155	0.589	0.603
8	0.476	0.486	0.445	0.493
9	0.463	0.252	0.571	0.324
10	0.512	0.131	0.574	0.319
Average Accuracy	0.431	0.281	0.491	0.427

From Table 5: , we know that Linear SVM performs better in those four classification methods. In the experimental accuracy results, the average accuracy of linear SVM is 0.491, which is the highest. At the same time, linear SVM has the highest accuracy in 6 prediction results. Using different training set and test set partition methods, the classification results have obvious differences. When using the Linear Discriminant Analysis method, the highest accuracy is obtained through test subject 3. When using the quadric discriminant analysis as the classifier, the highest accuracy is gained from test subject 8. When using the linear SVM for classification, the highest accuracy achieved through test subject 3. Test subject 2 achieves the highest accuracy when using the quadric SVM.

Synthetic data methods applied to sleep stage analysis

(a) Linear discriminant analysis

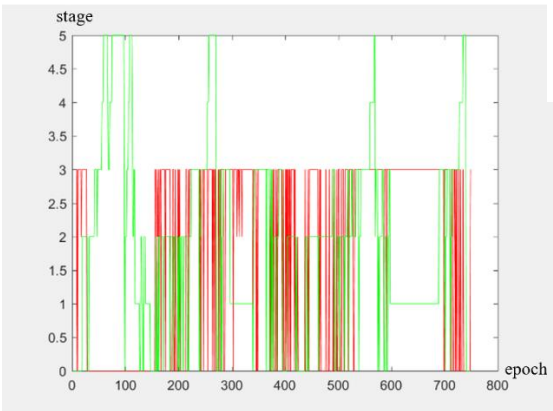


```
>> confusionmat(result, testclass)
```

```
ans =
```

34	0	2	1	1	0
2	20	10	6	2	2
62	61	78	15	1	2
15	74	123	147	24	53
1	0	0	0	0	0
8	0	0	3	1	1

(b) Quadratic discriminant analysis

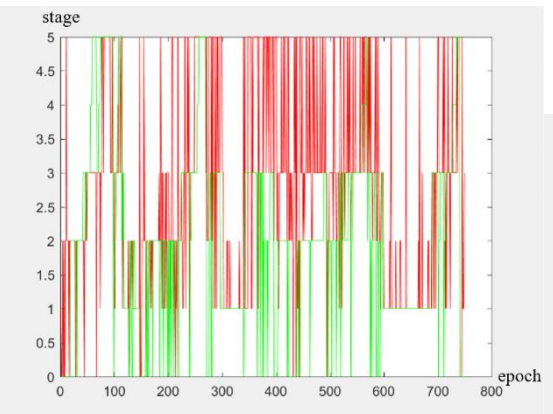


```
>> confusionmat(result2, testclass)
```

```
ans =
```

82	21	89	47	18	44
0	0	0	0	0	0
3	3	3	0	0	0
37	131	121	125	11	14
0	0	0	0	0	0
0	0	0	0	0	0

(c) Linear support vector machine

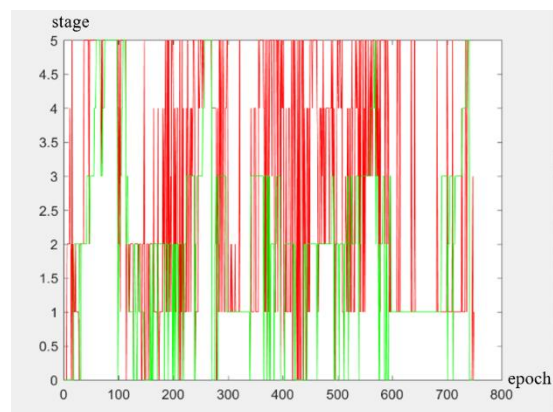


```
>> confusionmat(result3, testclass)
```

```
ans =
```

24	0	2	1	1	0
5	128	25	17	1	0
51	14	61	7	0	0
16	7	87	94	17	19
0	0	0	0	0	0
26	6	38	53	10	39

(d) Quadratic support vector machine



```
>> confusionmat(result4, testclass)
```

```
ans =
```

43	2	6	1	0	0
5	123	75	59	6	1
29	19	33	2	0	0
2	0	7	19	2	1
23	1	43	27	3	2
20	10	49	64	18	54

Figure 19 Sleep stage prediction results (Test subject: 1)

Synthetic data methods applied to sleep stage analysis

Figure 19 shows the sleep stage prediction results of four classification methods applied on the test subject 1. The red lines are predicted classes, while the green lines are correct classes. The figures on the right side are the confusion matrix.

Select training set and test set number 3 for the following experiments. Generate synthetic data through two methods described in 3.2 based on the EEG-ECG features in training set number 3. Then, combine different numbers of synthetic data generated by the two methods with the original data separately, forming new training sets. Table 6: Accuracy of three training set using Linear SVM (6 classes classification) shows the accuracy of predicted results of 6 classes using different training sets, forming with different numbers of synthetic data and different methods to generate synthetic data. The classifier is Linear SVM and the results came from EEG-ECG features.

Table 6: Accuracy of three training set using Linear SVM (6 classes classification)

Number of synthetic data	Original Data	Original data and synthetic data generated by adding Gaussian noise	Original data and synthetic data generated by using surrogate
76 (1%)	0.705	0.700	0.706
380 (5%)	0.705	0.689	0.711
760 (10%)	0.705	0.691	0.689
1140 (15%)	0.705	0.688	0.677
1900 (25%)	0.705	0.686	0.679
3801 (50%)	0.705	0.689	0.681
5701 (75%)	0.705	0.690	0.634
7602 (100%)	0.705	0.695	0.567

From Table 6: Accuracy of three training set using Linear SVM (6 classes classification), we can see that when the added synthetic data accounts for 10% or more of the original data, the accuracies of models trained with the synthetic data generated by adding Gaussian noise are higher than the accuracies of classification models that trained with the synthetic data generated by using surrogate. The accuracies of models trained with the synthetic data generated by adding Gaussian noise are lower than the accuracy of the model that trained with the original data. The accuracies of classification models that trained with the synthetic data generated by

Synthetic data methods applied to sleep stage analysis

using surrogate are higher than the accuracy of the model trained with the original data when the added synthetic data accounts for 1% and 5% of the original data.

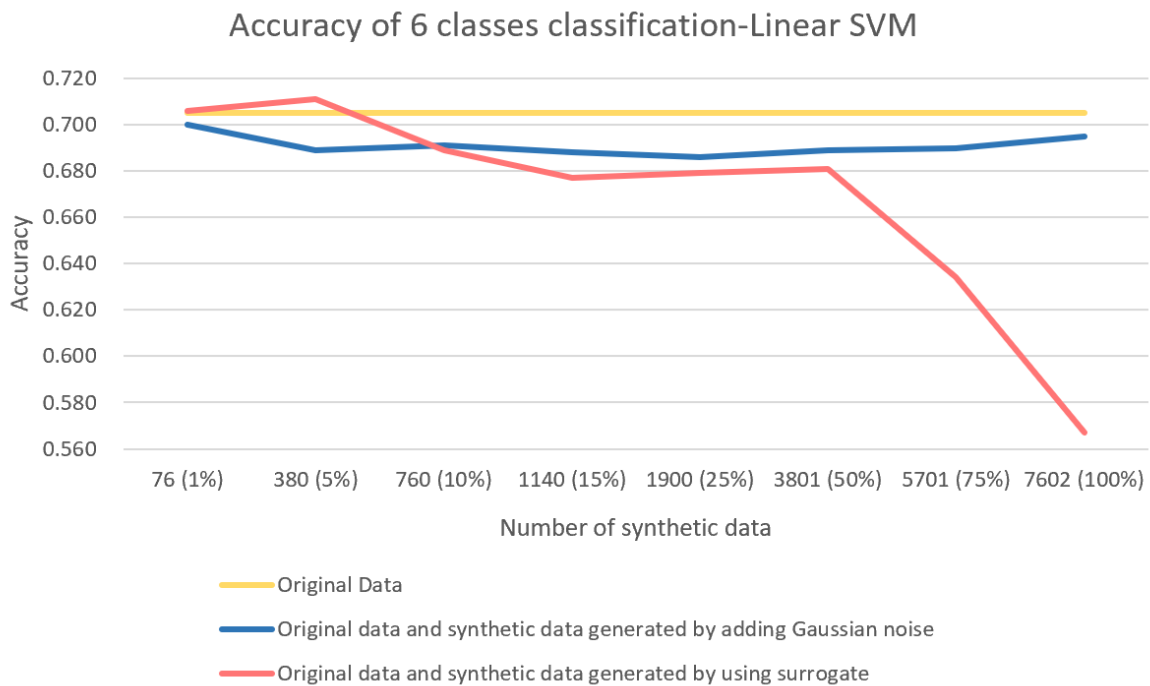
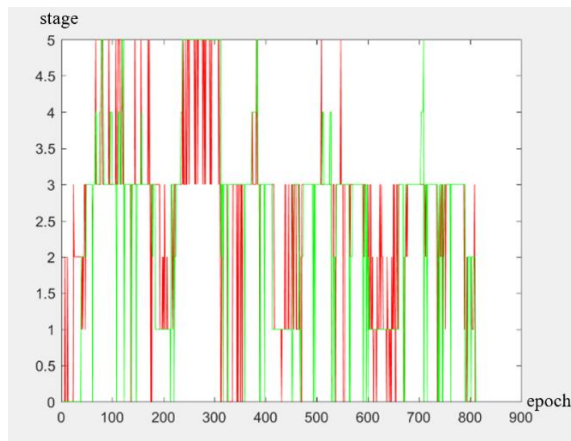


Figure 20 Accuracy of 6 classes classification. Classifier: Linear SVM

Figure 20 Accuracy of 6 classes classification. Classifier: Linear SVM shows the accuracy of the 6 classes classification. We can see the variation of accuracy depending on the number of synthetic data that added to the original data. From the figure, we can see that when the added synthetic data accounts for 1% and 5% of the original data, the accuracies of classification models that trained with the synthetic data generated by using surrogate are higher than the accuracy of the model trained with the original data. When the added surrogate samples account for more than 50% of the original data, the accuracy of adding more synthetic data is lower. Figure 21 Sleep stage prediction results using Linear SVM (Test subject: 3) shows the sleep stage prediction results using Linear SVM, where subject 3 is test set. The added synthetic data number is 7602, which is 100% of original data. The red lines are predicted classes, while the green lines are correct classes. The figures on the right side are the confusion matrix.

Synthetic data methods applied to sleep stage analysis

(a) Original Data

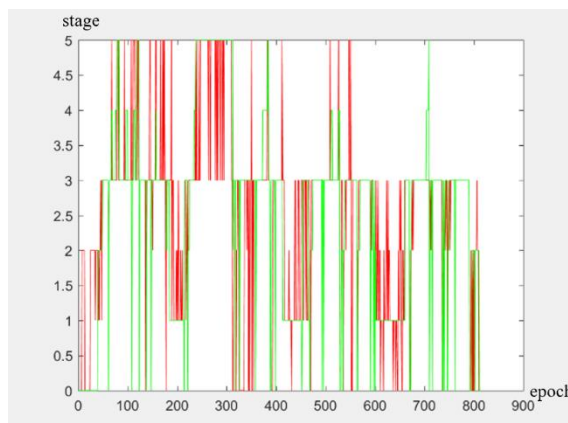


```
>> confusionmat(result3_3, testclass)
```

```
ans =
```

51	9	12	21	0	1
0	80	7	9	0	0
22	15	14	7	0	0
11	28	23	368	32	26
0	0	0	0	1	1
0	0	0	9	6	60

(b) Original data and synthetic data generated by adding Gaussian noise

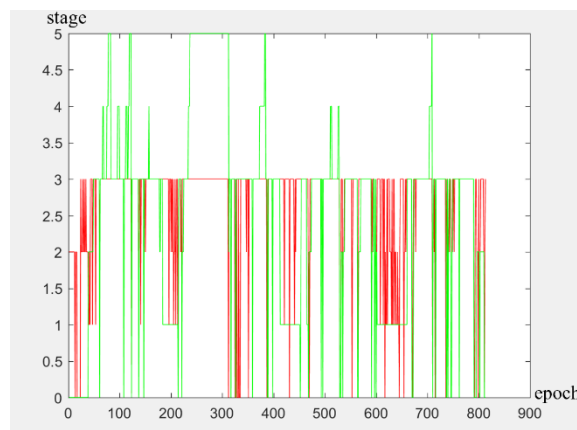


```
>> confusionmat(result3_3_noise, testclass)
```

```
ans =
```

49	9	8	25	0	1
1	76	4	7	0	0
27	17	23	14	0	0
7	29	21	352	31	22
0	0	0	0	0	0
0	1	0	16	8	65

(c) Original data and synthetic data generated by using surrogate



```
ans =
```

34	4	2	13	0	1
0	20	4	2	0	0
35	19	16	8	0	0
15	89	34	391	39	87
0	0	0	0	0	0
0	0	0	0	0	0

Figure 21 Sleep stage prediction results using Linear SVM (Test subject: 3)

Then, we do the binary classification to observe the predicted results of using synthetic data. Binary classification problem classified wake (W) and Sleep (S). The new classes are class 0: label 0; class 1: label 1, 2, 3, 4, 5. Table 6: Accuracy of three training set using Linear SVM (6 classes classification) shows the accuracy of predicted results of 2 classes using different

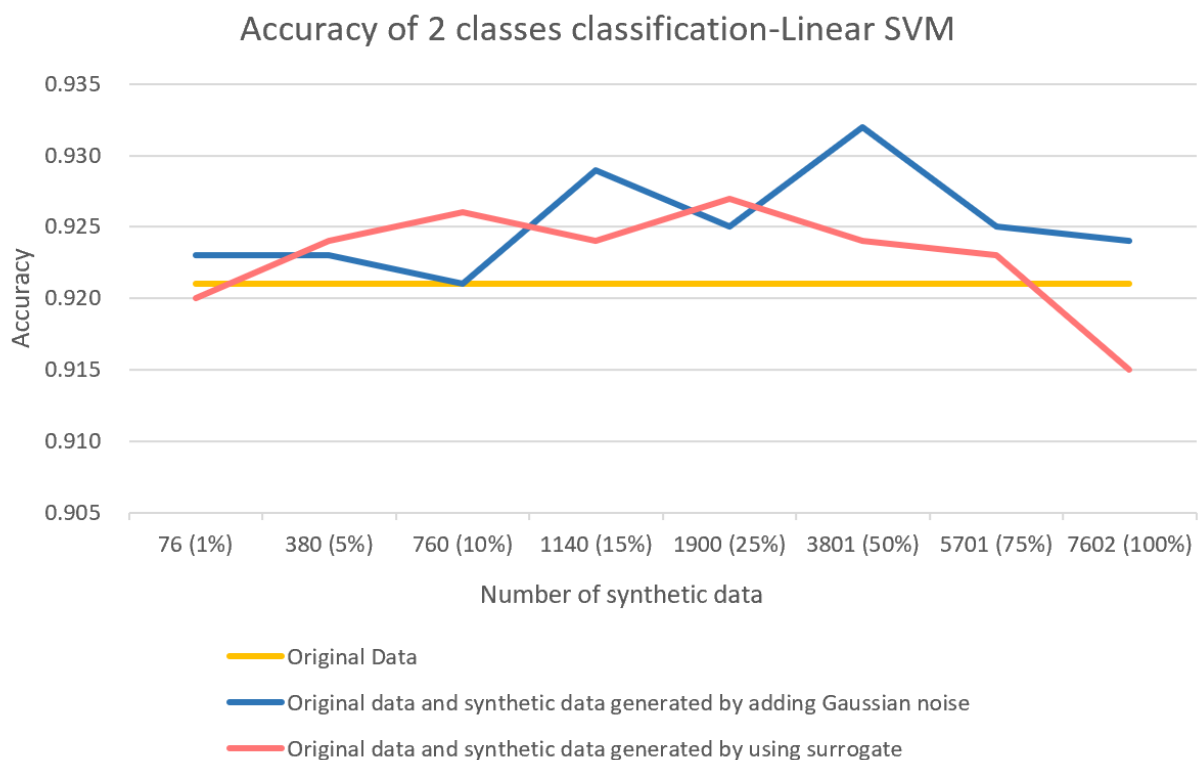
Synthetic data methods applied to sleep stage analysis

training sets, forming with different numbers of synthetic data and different methods to generate synthetic data. The classifier is Linear SVM and the results came from EEG-ECG features. The test subject is 3.

Table 7: Accuracy of three training set using Linear SVM (2 classes classification)

Number of synthetic data	Original Data	Original data and synthetic data generated by adding Gaussian noise	Original data and synthetic data generated by using surrogate
76 (1%)	0.921	0.923	0.920
380 (5%)	0.921	0.923	0.924
760 (10%)	0.921	0.921	0.926
1140 (15%)	0.921	0.929	0.924
1900 (25%)	0.921	0.925	0.927
3801 (50%)	0.921	0.932	0.924
5701 (75%)	0.921	0.925	0.923
7602 (100%)	0.921	0.924	0.915

From Table 6: Accuracy of three training set using Linear SVM (6 classes classification), we



Synthetic data methods applied to sleep stage analysis

can see that the accuracies of models trained with the synthetic data generated by adding Gaussian noise are higher than or at least the same as the classification model that trained with the original data. The accuracies of models trained with the synthetic data generated by adding Gaussian noise are higher than the accuracy of the model that trained with the original data when the added synthetic data accounts for 5%, 10%, 15%, 25%, 50%, and 75% of the original data. The accuracies of classification models that trained with the synthetic data generated by using surrogate are lower than the accuracy of the model trained with the original data when the added synthetic data accounts for 1%, 15%, 50%, 75%, and 100% of the original data. Otherwise, the accuracies of models that trained with the surrogate samples are higher.

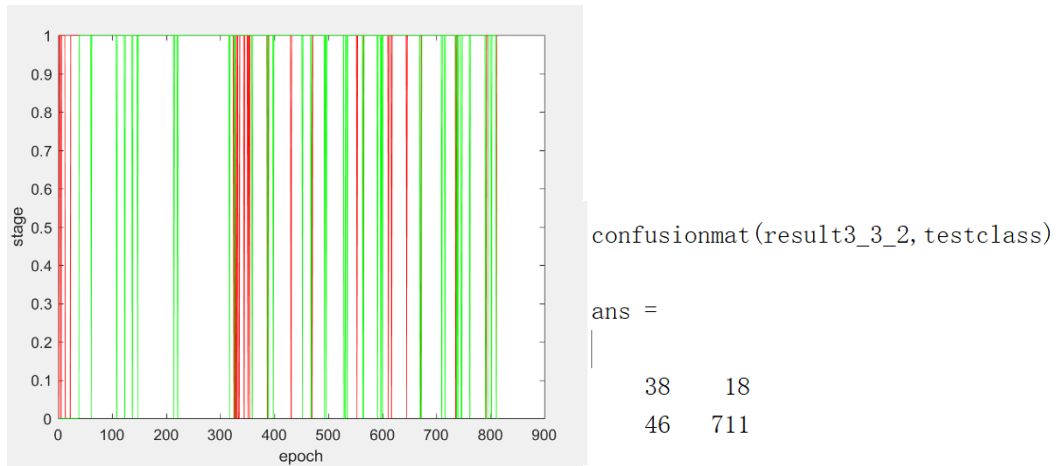
Figure 22 Accuracy of 2 classes classification. Classifier: Linear SVM

Figure 20 Accuracy of 6 classes classification. Classifier: Linear SVM shows the accuracy of 2 classes classification. We can see the variation of accuracy depending on the number of synthetic data that added to the original data. The accuracies of models trained with the synthetic data generated by adding Gaussian noise are higher than or at least the same as the classification model that trained with the original data. When the added synthetic data accounts for 5%, 10%, 15%, 25%, 50%, and 75% of the original data, the accuracies of classification models that trained with the synthetic data generated by using surrogate are higher than the accuracy of the model trained with the original data. When the added surrogate samples account for more than 25% of the original data, the accuracy of adding more synthetic data is lower.

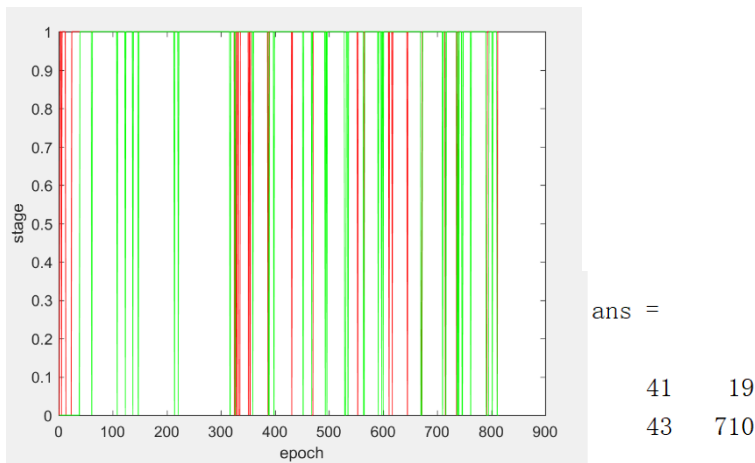
Figure 21 Sleep stage prediction results using Linear SVM (Test subject: 3) shows the sleep stage prediction results using Linear SVM, where subject 3 is test set. The added synthetic data number is 7602, which is 100% of original data. The red lines are predicted classes, while the green lines are correct classes. The figures on the right side are the confusion matrix.

Synthetic data methods applied to sleep stage analysis

(a) Original Data



(b) Original data and synthetic data generated by adding Gaussian noise



(c) Original data and synthetic data generated by using surrogate

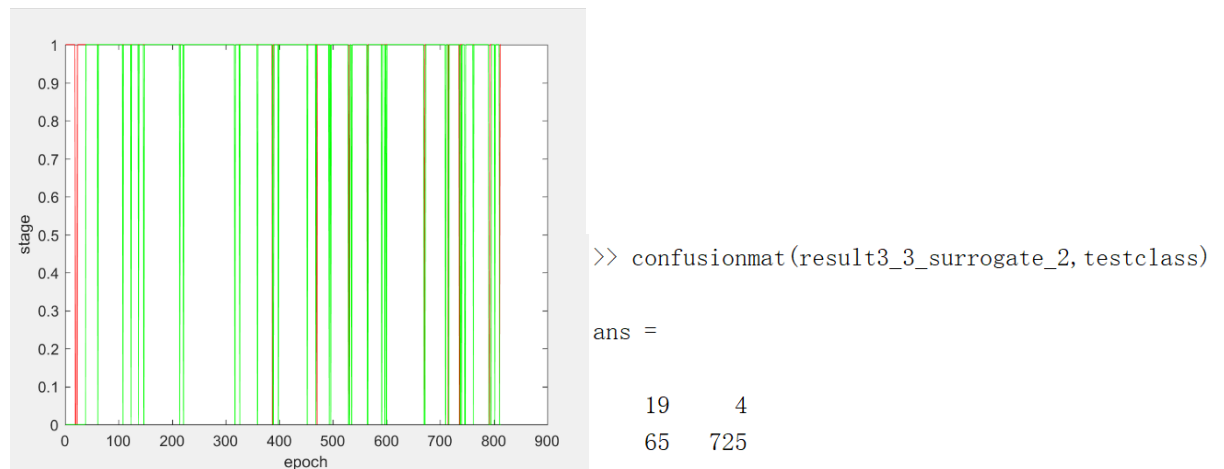


Figure 23 Sleep stage prediction results using Linear SVM (Test Set: Number 3)

Then, we do the 3 classes classification to observe the predicted results of using synthetic data. Three classes classification problem classified wake (W), REM and NonREM. The new classes are class 0: label 0; class 1: label 1; class 2: label 2, 3, 4, 5. Table 6: Accuracy of three training

Synthetic data methods applied to sleep stage analysis

set using Linear SVM (6 classes classification) shows the accuracy of predicted results of 3 classes using different training sets, forming with different numbers of synthetic data and different methods to generate synthetic data. The classifier is Linear SVM and the results came from EEG-ECG features. The test subject is 3.

Table 8: Accuracy of three training set using Linear SVM (Binary class classification)

Number of synthetic data	Original Data	Original data and synthetic data generated by adding Gaussian noise	Original data and synthetic data generated by using surrogate
76 (1%)	0.841	0.846	0.854
380 (5%)	0.841	0.850	0.846
760 (10%)	0.841	0.851	0.846
1140 (15%)	0.841	0.852	0.833
1900 (25%)	0.841	0.846	0.793
3801 (50%)	0.841	0.850	0.768
5701 (75%)	0.841	0.851	0.757
7602 (100%)	0.841	0.860	0.741

From Table 6: Accuracy of three training set using Linear SVM (6 classes classification), we can see that the accuracies of models trained with the synthetic data generated by adding Gaussian noise are higher than the classification model that trained with the original data. The accuracies of models trained with the synthetic data generated by adding Gaussian noise are higher than the accuracy of the model that trained with the original data when the added synthetic data accounts for 1%, 5%, and 10% of the original data. The accuracies of classification models that trained with the synthetic data generated by using surrogate are lower than the accuracy of the model trained with the original data when the added synthetic data accounts for 5%, 10%, 15%, 25%, 50%, 75%, and 100% of the original data. Otherwise, the accuracies of models that trained with the surrogate samples are higher.

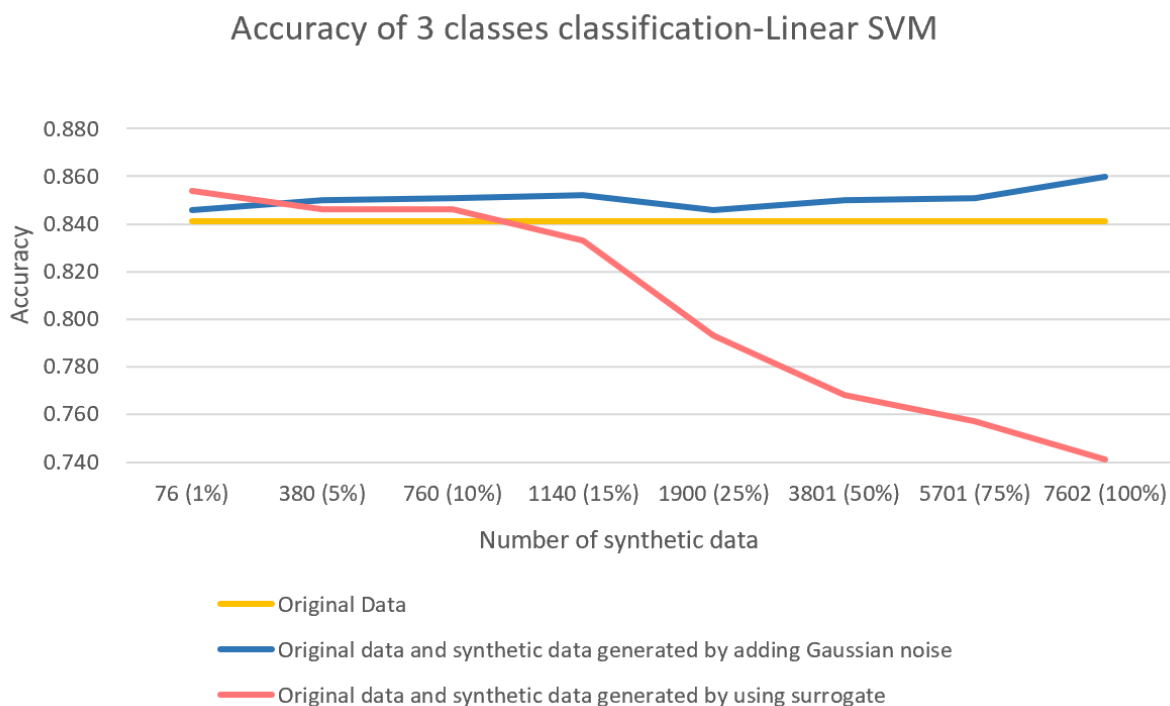


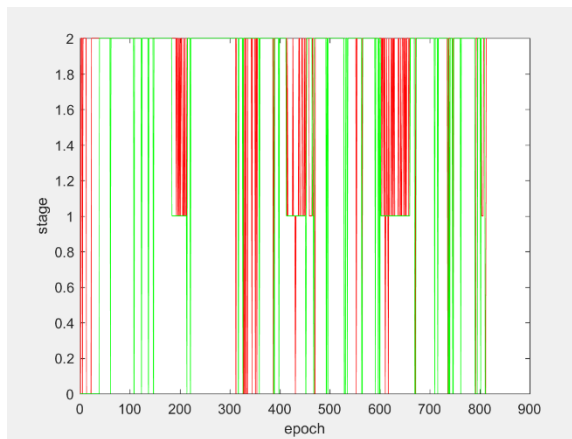
Figure 24 Accuracy of 3 classes classification. Classifier: Linear SVM

Figure 20 Accuracy of 6 classes classification. Classifier: Linear SVM shows the accuracy of the 3 classes classification. We can see the variation of accuracy depending on the number of synthetic data that added to the original data. The accuracies of models trained with the synthetic data generated by adding Gaussian noise are higher than the classification model that trained with the original data. When the added synthetic data accounts for 1%, 5% and, 10% of the original data, the accuracies of classification models that trained with the synthetic data generated by using surrogate are higher than the accuracy of the model trained with the original data. When the added surrogate samples account for more than 10% of the original data, the accuracy of adding more synthetic data is lower.

Figure 21 Sleep stage prediction results using Linear SVM (Test subject: 3) shows the sleep stage prediction results using Linear SVM, where subject 3 is test set. The added synthetic data number is 1140, which is 15% of original data. The red lines are predicted classes, while the green lines are correct classes. The figures on the right side are the confusion matrix.

Synthetic data methods applied to sleep stage analysis

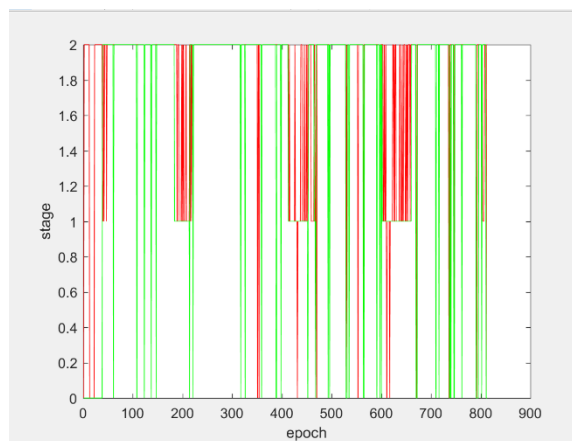
(a) Original data



ans =

39	3	15
0	73	10
45	56	572

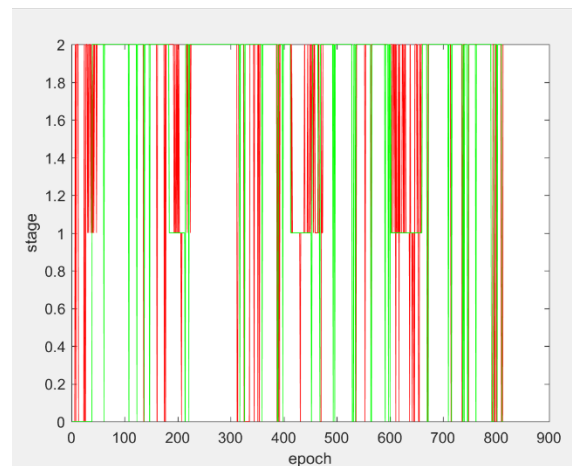
(b) Original data and synthetic data generated by adding Gaussian noise



ans =

33	3	5
0	84	16
51	45	576

(c) Original data and synthetic data generated by using surrogate



ans =

51	10	33
4	76	14
29	46	550

Figure 25 Sleep stage prediction results using Linear SVM (Test Subject: 3)

Then, we do the 4 classes classification to observe the predicted results of using synthetic data. Four classes classification problem classified wake (W), REM, Stage I and II, and Stage III and IV. The new classes are class 0: label 0; class 1: label 1; class 2: label 2, 3; class 3: label 4, 5. Table 6: Accuracy of three training set using Linear SVM (6 classes classification) shows the

Synthetic data methods applied to sleep stage analysis

accuracy of predicted results of 4 classes using different training sets, forming with different numbers of synthetic data and different methods to generate synthetic data. The classifier is Linear SVM and the results came from EEG-ECG features. The test subject is 3.

Table 9: Accuracy of three training set using Linear SVM (Binary class classification)

Number of synthetic data	Original Data	Original data and synthetic data generated by adding Gaussian noise	Original data and synthetic data generated by using surrogate
76 (1%)	0.774	0.770	0.781
380 (5%)	0.774	0.775	0.774
760 (10%)	0.774	0.768	0.763
1140 (15%)	0.774	0.771	0.749
1900 (25%)	0.774	0.763	0.725
3801 (50%)	0.774	0.779	0.605
5701 (75%)	0.774	0.774	0.609
7602 (100%)	0.774	0.780	0.605

From Table 6: Accuracy of three training set using Linear SVM (6 classes classification), we can see that the accuracies of models trained with the synthetic data generated by adding Gaussian noise are higher than the classification model that trained with the original data when the added synthetic data accounts for 5%, 50%, and 100% of the original data. The accuracies of models trained with the synthetic data generated by adding Gaussian noise are higher than or at least the same as the accuracy of the model that trained with the original data when the added synthetic data accounts for 1% and 5% of the original data. The accuracies of classification models that trained with the synthetic data generated by using surrogate are lower than the accuracy of the model trained with the original data when the added synthetic data accounts for 5%, 10%, 15%, 25%, 50%, 75%, and 100% of the original data. Otherwise, the accuracies of models that trained with the surrogate samples are higher. Figure 20 Accuracy of 6 classes classification. Classifier: Linear SVM shows the accuracy of the 4 classes classification. We can see the variation of accuracy depending on the number of synthetic data that added to the original data. The accuracies of models trained with the synthetic data generated by adding Gaussian noise are higher than classification models that trained with the original data when the added synthetic data accounts for 5%, 50% and 100% of the original

Synthetic data methods applied to sleep stage analysis

data. When the added synthetic data accounts for 1% of the original data, the accuracies of classification models that trained with the synthetic data generated by using surrogate are higher than the accuracy of the model trained with the original data. When the added surrogate samples account for more than 5% and less than 50% of the original data, the accuracy of adding more synthetic data is lower.

Figure 26 Accuracy of 4 classes classification. Classifier: Linear SVM

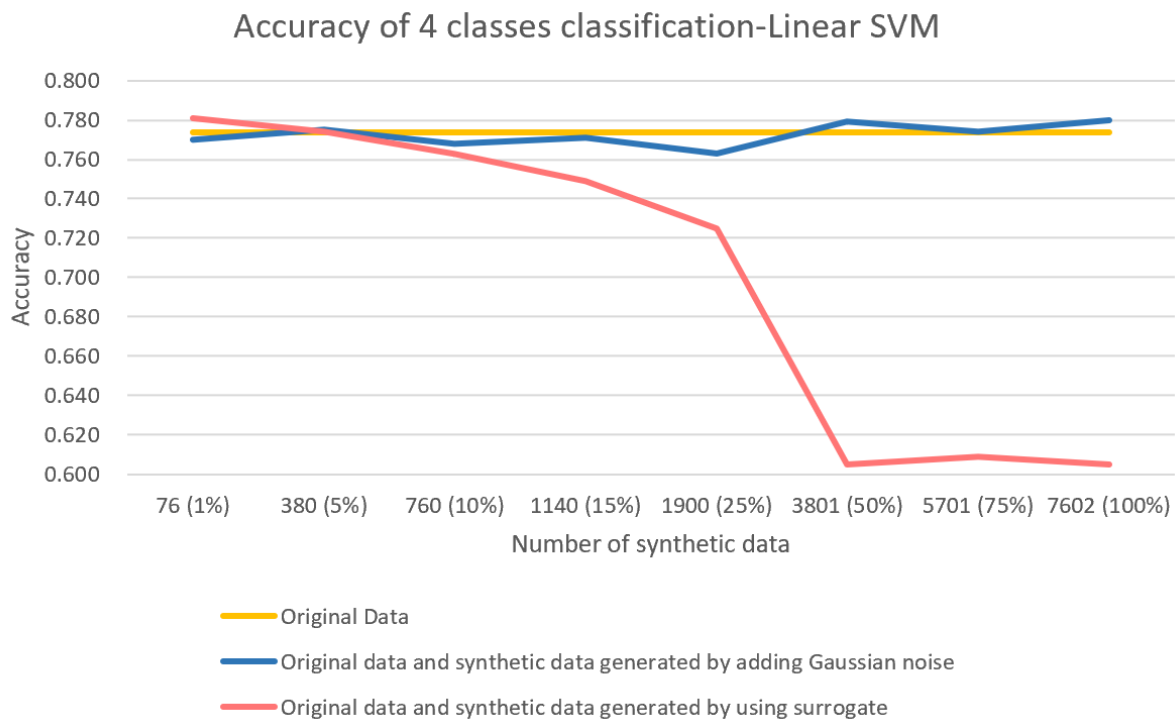
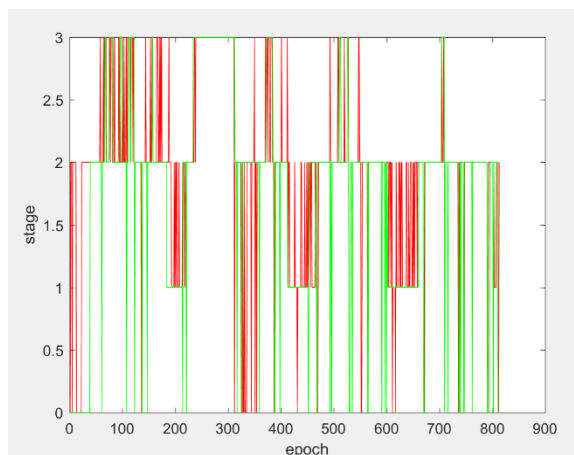


Figure 21 Sleep stage prediction results using Linear SVM (Test subject: 3) shows the sleep stage prediction results using Linear SVM, where subject 3 is the test set. The added synthetic data number is 1140, which is 15% of the original data. The red lines are predicted classes, while the green lines are correct classes. The figures on the right side are the confusion matrix.

(a) Original data

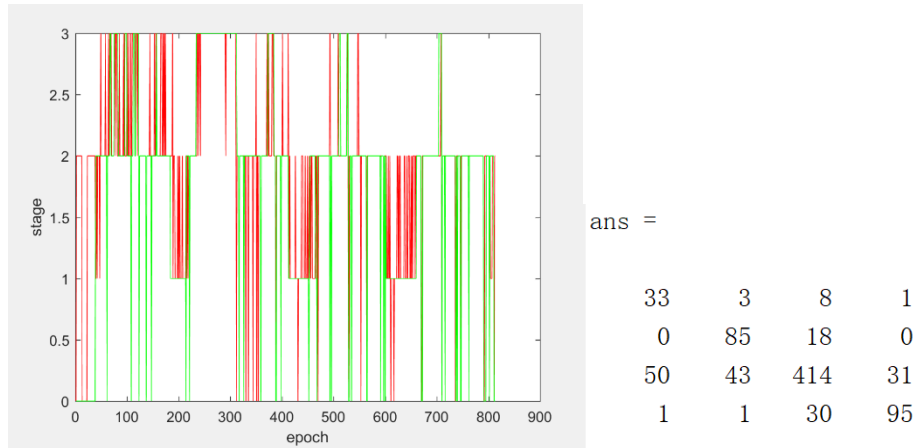


ans =

39	3	18	1
0	78	12	0
44	50	409	23
1	1	31	103

Synthetic data methods applied to sleep stage analysis

(b) Original data and synthetic data generated by adding Gaussian noise



(c) Original data and synthetic data generated by using surrogate

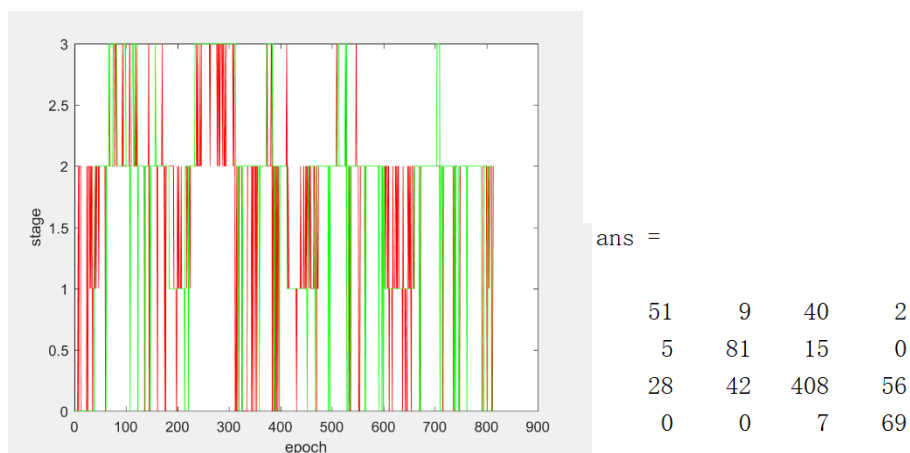


Figure 27 Sleep stage prediction results using Linear SVM (Test Subject: 3)

Through the experiment results shown in Table 6: Accuracy of three training set using Linear SVM (6 classes classification), Table 7: , Table 7: and Table 7: , we know that the performances of applying synthetic data generated by adding Gaussian noise to the 2 classes and 3 classes classification problem are better than using the original data. When the added synthetic data accounts for 5%, 50%, 75% and 100% of the original data, the performances of applying synthetic data generated by adding Gaussian noise to the 4 classes classification problem are better than or at least the same as using the original data. However, when the classification problem comes to 6 classes classification, the accuracies of the model trained with the synthetic data generated by adding Gaussian noise are lower than the accuracy of the model that trained with the original data. The fewer categories of classification problems, the more accuracy improvement can be obtained by adding Gaussian replicates.

The performances of applying synthetic data generated by surrogate to the 2 classes classification problem are better than using the original data when the added synthetic data accounts for 5%, 10%, 15%, 25%, 50%, and 75% of the original data. The performances of

Synthetic data methods applied to sleep stage analysis

applying synthetic data generated by surrogate to the 3 classes classification problem are better than using the original data when the added synthetic data accounts for 1%, 5%, and 10% of the original data. When the added synthetic data accounts for 1% and 5% of the original data, the accuracies of the model trained with the synthetic data generated by surrogate are higher than or at least the same as the accuracy of the model that trained with the original data. The accuracies of the model trained with the synthetic data generated by surrogate are higher than the accuracy of the model that trained with the original data when the added synthetic data accounts for 1% and 5% of the original data. Moreover, when the added surrogate samples account for more than 50% of the original data, the accuracy of adding more synthetic data is lower. Therefore, it is better not to add too many surrogate samples in the classification problem. When the added synthetic data accounts for 5% of the original data, the accuracy of the classification model usually improved.

Chapter 5: Conclusion and Further Work

This paper applied synthetic data in the sleep stage classification problem. This paper uses two methods to generate synthetic data, which are (i) adding Gaussian noise to real data to generate replicates of the original data, and (ii) generating synthetic data samples using surrogates, which is a way of synthesizing multivariable time series with prescribed covariance function and marginal distributions, obtained from empirical results.

In this paper, an open-source EEG and ECG signals dataset published by University College Dublin is used. [16] We do the pre-processing and feature extraction on this dataset. The well-defined dataset contains EEG and ECG features of 10 subjects and the sleep stage of each epoch of each subject. The six classes of sleep stage labelled with numbers from 0 to 5. After that, divide the dataset into the training set and the test set following the "leave one out" procedure.

Four classification methods are used in the 6 classes of sleep stage classification, which are linear discriminant analysis, quadratic discriminant analysis, linear support vector machine, and quadratic support vector machine. These methods are applied to 10 different training sets. By comparing the accuracy of 40 models, we can conclude that the performance of linear SVM in sleep classification is the best of the four classification methods. In addition, the results of using different training sets are quite different. We need a more reasonable dataset division method rather than the "leave one out" procedure.

After that, we select subject 3 as the test set and the other subjects as the training set for the following experiments. Two data synthesis methods are applied to EEG-ECG features in the training set separately. Combine different numbers of synthetic data with the original data to form new training sets. Then, apply linear SVM on the original training set and the new training sets with different numbers of synthetic data. In order to better observe the impact of synthetic data on sleep stage classification problems. We transform the 6 classes classification problem into the binary classification problem, 3 classes classification problem, and 4 classes classification problem. Apply linear SVM on the original training set and the new training sets with different numbers of synthetic data. Through the experiment results, we know that the performances of applying synthetic data generated by adding Gaussian noise to the 2 classes and 3 classes classification problem are better than using the original data. When the added synthetic data accounts for 5%, 50%, 75% and 100% of the original data, the performances of applying synthetic data generated by adding Gaussian noise to the 4 classes classification problem are better than or at least the same as using the original data. However, when the

Synthetic data methods applied to sleep stage analysis

classification problem comes to 6 classes classification, the accuracy of the model trained with the synthetic data generated by adding Gaussian noise is lower than the accuracy of the model that trained with the original data. Thus, we can conclude that the fewer categories of classification problems, the more accuracy improvement can be obtained by adding Gaussian replicates.

The performances of applying synthetic data generated by surrogate to the 2 classes classification problem are better than using the original data when the added synthetic data accounts for 5%, 10%, 15%, 25%, 50%, and 75% of the original data. The performances of applying synthetic data generated by surrogate to the 3 classes classification problem are better than using the original data when the added synthetic data accounts for 1%, 5%, and 10% of the original data. When the added synthetic data accounts for 1% and 5% of the original data, the accuracies of the model trained with the synthetic data generated by surrogate are higher than or at least the same as the accuracy of the model that trained with the original data. The accuracies of the model trained with the synthetic data generated by surrogate are higher than the accuracy of the model that trained with the original data when the added synthetic data accounts for 1% and 5% of the original data. Moreover, when the added surrogate samples account for more than 50% of the original data, the accuracy of adding more synthetic data is lower. Therefore, it is better not to add too many surrogate samples in the classification problems. When the added synthetic data accounts for 5% of the original data, the accuracy of the classification model usually improved.

In conclusion, when the categories of classification problems are not too many, for instance, 2 classes and 3 classes, the accuracy of the classification model can be improved by adding Gaussian replicates. If the categories of classification problems are too many, the accuracy of the classification model decreased by adding Gaussian replicates.

Surrogate samples of EEG and ECG features can be combined with real data of EEG and ECG features without a great decrease of classification performance of sleep staging from sleep disorder patients when the added synthetic data accounts for no more than 50%. Moreover, adding a few surrogate samples, usually 5% of the original data can even improve the accuracy of the classification model. The above might be interesting for practical studies when the sample size is small.

For future research, we can change the training set and test set dividing method to improve the classification results.

References

- [1] Motamedi-Fakhr, S., Moshrefi-Torbati, M., Hill, M., Hill, C. and White, P., 2014. Signal processing techniques applied to human sleep EEG signals—A review. *Biomedical Signal Processing and Control*, 10, pp.21-33.
- [2] Hae-Jeong Park, Do-Un Jeong and Kwang-Suk Park, 2002. Automated detection and elimination of periodic ECG artifacts in EEG using the energy interval histogram method. *IEEE Transactions on Biomedical Engineering*, 49(12), pp.1526-1533.
- [3] Widrow, B., Glover, J., McCool, J., Kaunitz, J., Williams, C., Hearn, R., Zeidler, J., Eugene Dong, J. and Goodlin, R., 1975. Adaptive noise cancelling: Principles and applications. *Proceedings of the IEEE*, 63(12), pp.1692-1716.
- [4] 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.
- [5] Hjorth, B., 1970. EEG analysis based on time domain properties. *Electroencephalography and Clinical Neurophysiology*, 29(3), pp.306-310.
- [6] Lee, J., Kim, D., Kim, I., Park, K. and Kim, S., 2002. Detrended fluctuation analysis of EEG in sleep apnea using MIT/BIH polysomnography data. *Computers in Biology and Medicine*, 32(1), pp.37-47.
- [7] Kamiński, M., Blinowska, K. and Szelenberger, W., 1997. Topographic analysis of coherence and propagation of EEG activity during sleep and wakefulness. *Electroencephalography and Clinical Neurophysiology*, 102(3), pp.216-227.
- [8] Pincus, S., 1991. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, 88(6), pp.2297-2301.
- [9] 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.
- [10] Pan, J. and Tompkins, W., 1985. A Real-Time QRS Detection Algorithm. *IEEE Transactions on Biomedical Engineering*, BME-32(3), pp.230-236.
- [11] Wolpert, E., 1969. A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects. *Archives of General Psychiatry*, 20(2), p.246.
- [12] Gudmundsson, S., Runarsson, T. and Sigurdsson, S., n.d. Automatic Sleep Staging using

Synthetic data methods applied to sleep stage analysis

Support Vector Machines with Posterior Probability Estimates. International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06).

[13] Dl.acm.org. 2021. ICA using spacings estimates of entropy | The Journal of Machine Learning Research. [online] Available at: <<https://dl.acm.org/doi/10.5555/945365.964306>> [Accessed 21 April 2021].

[14] Borgnat, P., Abry, P. and Flandrin, P., 2012. Using surrogates and optimal transport for synthesis of stationary multivariate series with prescribed covariance function and non-gaussian joint-distribution. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),.

[15] Li, T. and Zhou, M., 2016. ECG Classification Using Wavelet Packet Entropy and Random Forests. Entropy, 18(8), p.285.

[16] Physionet.org. 2021. St. Vincent's University Hospital / University College Dublin Sleep Apnea Database v1.0.0. [online] Available at: <<https://physionet.org/content/ucddb/1.0.0/>> [Accessed 15 April 2021].

[17] Perso.ens-lyon.fr. 2021. Pierre Borgnat - Matlab Codes. [online] Available at: <<http://perso.ens-lyon.fr/pierre.borgnat/codes.html>> [Accessed 12 April 2021].

Acknowledgement

Thanks to my supervisor for guiding my project. When I encounter problems, he will patiently answer my doubts and correct my mistakes. Moreover, I would like to thanks for my classmate and friend Zekuan Liu. He was supervised by the same supervisor as mine. During the project period, we discussed a lot about EEG and ECG signal pre-processing and feature extraction, and explained many papers to each other.

Appendix

- Specification, part 1 and part 2

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	Synthetic data methods applied to sleep stage analysis		
题目分类 Scope	Data Science and Artificial Intelligence	Research	Software
主要内容 Project description	<p>This project will implement several techniques to augment the sample size of categories of data in classification problems using synthetic data. These techniques seek to improve the imbalance that can be found in data populations, when one or some of the data categories are scarce in number of samples. Thus, the a priori probability of the categories of data could be approximated to be equally-probable. We will assume that synthetic data enables a smoothed version of the estimators for classification to be obtained, and thus, classification performance might be improved. The application considered in this project is sleep apnea which is a disease that behaves that a subject can have several micro-awakenings, also called microarousals, during sleep. Indeed, there is a great imbalance between the time that the subject remains asleep and awake. The data will consist of features extracted from electroencephalographic (EEG) and electrocardiographic (ECG) signals measured from subjects while sleeping. The synthetic data can be obtained, for instance, as distorted replicas of the original data or surrogate data that follows a similar distribution of the original ones. This project will implement several classification cases considering to increase the sample size with different amounts of synthetic data. The classification accuracy index will be used to evaluate the quality and comparison of the results, in combination with clinical annotation information.</p>		
关键词 Keywords	Signal processing, Synthetic data, EEG, ECG, Classification, Sleep analysis		
主要任务 Main tasks	1 Study of methods for microarousal detection: preprocessing, feature extraction, synthetic data, and classifiers.		
	2 Design and implementation of the procedures of preprocessing, synthetic data, 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 synthetic data 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).		

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	Wei	名 First Name	Jingjing		
BUPT 学号 BUPT number	2017213152	QM 学号 QM number	171044142	班级 Class	2017215120
论文题目 Project Title	Synthetic data 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	<p>1. An initial analysis of user requirements, and how data will be collected</p> <p>As there is a great imbalance between the time that the subject remains asleep and awake, using synthetic data to augment the sample size of categories of data in classification problems of sleep apnea is an effective way to increase the accuracy of the classification result of sleep apnea.</p> <p>The electroencephalographic (EEG) and electrocardiographic (ECG) data will used in this project is an open-source EEG and ECG signals dataset published by University College Dublin [1]. This database contains 25 full overnight polysomnograms with simultaneous three-channel Holter ECG, from adult subjects with suspected sleep-disordered breathing. Three-channel Holter ECGs (V5, CC5, V5R) were recorded using a Reynolds Lifecard CF system.</p> <p>2. The algorithms, methodologies and other techniques to be employed</p> <p>2.1 Sleep apnea, sleep staging classification</p> <p>The first step of this project is to study the sleep apnea and sleep staging classification. There are many existing methods to do this, for instance, sleep staging based on Singular Value Decomposition (SVD), Variational Mode Decomposition (VMD), Hilbert Huang Transform (HHT), sleep apnea identification using HRV features of ECG signals, and etc. The detailed classification methods will be investigated and studied. Meanwhile, the EEG and ECG data and signal processing techniques, including preprocessing, feature extraction, synthetic and classifiers using Matlab will be studied.</p> <p>2.2 Generate synthetic samples</p> <p>The most essential task of this project is to generate synthetic EEG and ECG samples to compensate on the imbalance of the neuroscience data. This project will primarily use two techniques to generate synthetic data samples: (i) adding Gaussian noise to real data to produce replicates of them; and (ii) to generate synthetic data samples using surrogates, which is a way of synthesizing series matching empirical properties of some observed data.</p> <p>Specifically, I may use surrogates for synthesis of stationary multivariate</p>				

<p>series. This method was proposed by Pierre Borgnat, Patrice Abry, and Patrick Flandrin. [2] This method synthesizes many different multivariate time-series, where the covariance function and marginal or joint-distributions are prescribed either through a model or by empirical properties of measured series. To obtain more realizations of the same series, we can iterate the proposed algorithms. The open-source Matlab code of this synthesis method will be used.</p> <p>3. Experiments that should be done to prove the project hypotheses</p> <p>3.1 Pre-processing of the EEG and ECG data</p> <p>The first step of experiments is to pre-processing the data. Feature extraction will be done through this process. Then we will use the processed data for further synthesis and sleep apnea classifications.</p> <p>3.2 Generate synthetic samples</p> <p>Different synthetic algorithm will be used in this part of experiments. I will implement or using open-source code of different methods to generate synthetic data. Moreover, different synthetic result will be compared.</p> <p>3.3 Sleep apnea classification</p> <p>Sleep apnea classification will be done by using different datasets with different amounts of synthetic data. The accuracy results of the classification will be calculated and compared between each other. Moreover, the classification accuracy result using synthetic data will also be compared with the classification accuracy result that doesn't use synthetic data.</p> <p>3.4 Tuning and debugging and evaluation of the methods</p> <p>Based on the 3 processing steps mentioned above, I will debug the methods and find synthetic method that achieved best result.</p> <p>4. Programming language and software package to be used</p> <p>Matlab. The third-party libraries, such as Signal Processing Toolkit used for signal processing.</p> <p>5. A list of background material consulted including World Wide Web pages</p> <p>[1] St. Vincent's University Hospital / University College Dublin Sleep Apnea Database v1.0.0 (2020). Available at: https://physionet.org/content/ucddb/1.0.0/ (Accessed: 6 November 2020).</p> <p>[2] Borgnat, P., Abry, P. and Flandrin, P., 2012. Using surrogates and optimal transport for synthesis of stationary multivariate series with prescribed covariance function and non-gaussian joint-distribution. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),.</p> <p>[3] Sulistyono, B., Surantha, N. and Isa, S., 2018. Sleep Apnea Identification using HRV Features of ECG Signals. International Journal of Electrical and Computer Engineering (IJECE), 8(5), p.3940.</p>

Synthetic data methods applied to sleep stage analysis

	<p>[4] Index of /papers/volume4/learned-miller03a (2020). Available at: https://www.jmlr.org/papers/volume4/learned-miller03a/ (Accessed: 6 November 2020).</p> <p>[5] Jayaraj, R., 2017. A Review on Detection and Treatment Methods of Sleep Apnea. JOURNAL OF CLINICAL AND DIAGNOSTIC RESEARCH.</p> <p>[6] 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.</p>
<p>道德规范 Ethics</p>	<p>Please confirm that you have discussed ethical issues with your Supervisor using the ethics checklist (Project Handbook Appendix 2). [YES]</p>
	<p>Summary of ethical issues: (put N/A if not applicable)</p> <p>[N/A]</p>
<p>中期目标 Mid-term target.</p> <p>It must be tangible outcomes, E.g. software, hardware or simulation.</p> <p>It will be assessed at the mid-term oral.</p>	<ol style="list-style-type: none"> 1. A well-defined dataset should be created and write the code to visualize the dataset 2. Realize both data synthetic methods: (i) adding Gaussian noise, and (ii) using surrogates to generate synthetic samples. 3. Realize the sleep apnea classification algorithm and evaluate the accuracy of the classification results. 4. Compared the classification results using different synthetic data. 5. Based on the above achievement more advanced data synthetic methods can be used and optimized to achieve higher accuracy or computational efficiency of sleep apnea classification.

Work Plan (Gantt Chart)

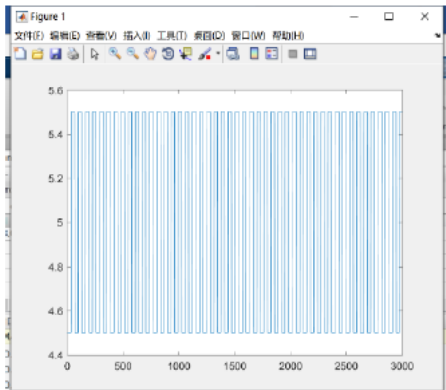
Fill in the sub-tasks and insert a letter X in the cells to show the extent of each task

	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 apnea classification												
Learn the preprocessing of EEG and ECG signal	X	X										
Learn the feature extraction of EEG and ECG signal			X									
Learn the synthetic methods of EEG and ECG signal				X	X	X	X					
Learn the classification methods of sleep apnea								X	X			
Task 2 Define the dataset												
Visualize the data set	X											
Do the feature extraction of the dataset		X										
Deal with data format		X										
Well define the dataset that will be used for further experiment		X										
Task 3 Study and implement two synthetic data algorithm												
Adding Gaussian noise to the dataset to generate synthetic data				X	X							
Evaluate the synthetic data					X							
Using Surrogates to generate the synthetic data						X	X					
Evaluate the synthetic data							X					
Task 4 Realize the sleep apnea classification algorithm												
Realize the sleep apnea classification algorithm								X	X			
Evaluate the accuracy of the classification results.									X			
Compare the classification result									X			
Optimize the algorithm										X	X	

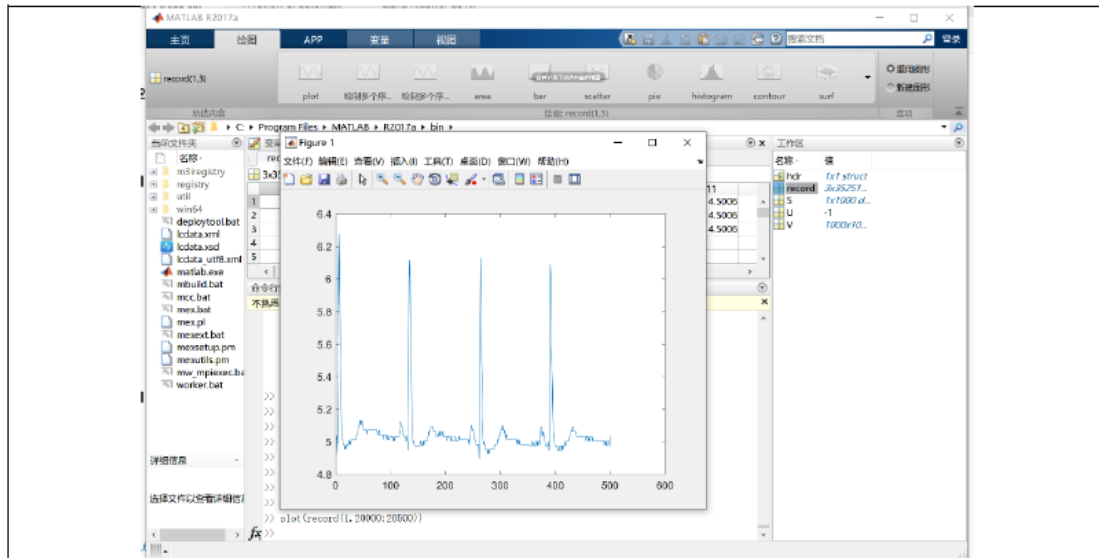
- Early-term Progress Report

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

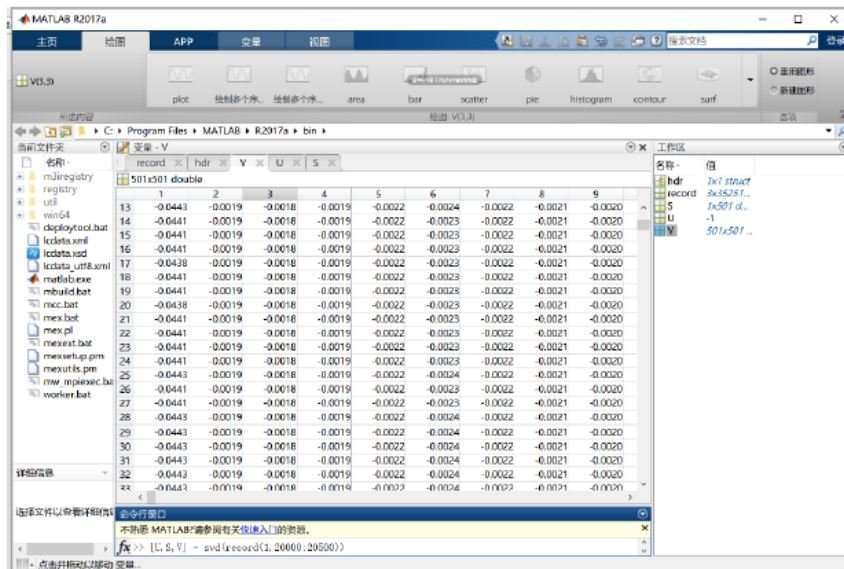
学院 School	International School	专业 Programme	Internet of Things Engineering		
姓 Family name	Wei	名 First Name	Jingjing		
BUPT 学号 BUPT number	2017213152	QM 学号 QM number	171044142	班级 Class	2017215120
论文题目 Project Title	Synthetic data methods applied to sleep stage analysis				
<p>已完成工作 Finished work:</p> <p>Task 1:</p> <p>1.1 Learn the pre-processing ECG signal</p> <p>The following materials have helped me to have a successful understanding of pre-processing of ECG signal. Paper: Automatic sleep staging based on SVD, VMD, HHT and morphological features of single-lead ECG signal. These ECG data were assessed by taking epochs that included 30 s of data for each subject. The epochs that were detected to include errors (noise, electrode disconnection, etc.) were deleted from the data record. University College Dublin Sleep Apnea Database contain the ECG signal in EDF format named with _lifecard.edf for 28 subjects. In record uccdb002, only two distinct ECG signals were recorded; the second ECG signal was also used as the third signal. The first part of these signals is square wave. I deleted these meaningless square wave signals.</p> <div style="text-align: center;">  </div> <p>1.2 Learn the feature extraction of ECG and EEGsignal</p> <p>I learned four methods of feature extraction, which are singular value decomposition (SVD), variational mode decomposition (VMD), the Hilbert- Huang transform (HHT), and morphological feature extraction.</p> <p>SVD is a factorization of a matrix. Any matrix $A = [a_1, a_2, a_3, \dots, a_n]$ with dimensions $m \times n$ is transformed into matrices U, S and V through SVD. $A = USV^T$. By using the SVD command of the MATLAB, the maximum of the singular values obtained from each window. VMD is an iterative technique that can be applied to nonlinear and nonstationary signals like EEG and EOG. Through VMD, the signal is decomposed into subcomponents called IMF. HHT is an efficient time-frequency analysis method which is effective for the analysis of nonlinear and nonstationary signals. We can apply Empirical Mode Decomposition (EMD) on nonlinear and nonstationary data, such as EEG signals, then the signal is decomposed into components known as IMF. For morphological feature extraction, an ECG signal is composed of 5 major waves, as P, Q, R, S, and</p>					

Synthetic data methods applied to sleep stage analysis



2.2 Do the feature extraction of the dataset.

Using SVD command to obtain the maximum of the singular values. Then save the matrix using SaveEDF command.



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

YES

下一步 Next steps:

Determine the final use of which way to extract features of the data and methods. Learning two methods of data synthesis, adding Gaussian noise and using surrogates, then use data synthesis methods on the determined data set to obtain the synthetic data.

- 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	Wei	名 First Name	Jingjing		
BUPT 学号 BUPT number	2017213152	QM 学号 QM number	171044142	班级 Class	2017215120
论文题目 Project Title	Synthetic data 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					
<p>已完成工作 Finished work:</p> <p>Task 1:</p> <p>1.1 Learn the pre-processing EEG and ECG signals Similar to Early term. The following materials have helped me to have a successful understanding of pre-processing of ECG signal. Paper: Automatic sleep staging based on SVD, VMD, HHT and morphological features of single-lead ECG signal. [1]These ECG data were assessed by taking epochs that included 30 s of data for each subject. The epochs that were detected to include errors (noise, electrode disconnection, etc.) were deleted from the data record. University College Dublin Sleep Apnea Database contain the ECG signal in EDF format named with _lifecard.edf for 28 subjects. In record ucddb002, only two distinct ECG signals were recorded; the second ECG signal was also used as the third signal. The first part of these signals is square wave. I deleted these meaningless square wave signals.</p> <p>1.2 Learn the feature extraction of ECG and EEG signal I learned four methods of feature extraction, which are singular value decomposition (SVD), variational mode decomposition (VMD), the Hilbert- Huang transform (HHT), and morphological feature extraction. [1] SVD is a factorization of a matrix. Any matrix $A = [a_1, a_2, a_3, \dots, a_n]$ with dimensions $m \times n$ is transformed into matrices U, S and V through SVD. $A = USV^T$. By using the SVD command of the MATLAB, the maximum of the singular values obtained from each window. VMD is an iterative technique that can be applied to nonlinear and nonstationary signals like EEG and EOG. Through VMD, the signal is decomposed into subcomponents called IMF. HHT is an efficient time-frequency analysis method which is effective for the analysis of nonlinear and nonstationary signals. We can apply Empirical Mode Decomposition (EMD) on nonlinear and nonstationary data, such as EEG signals, then the signal is decomposed into components known as IMF. For morphological feature extraction, an ECG signal is composed of 5 major waves, as P, Q, R, S, and T. We can detect R-peaks of the ECG signals using the Pan-Tompkins algorithm, and then P- Q-S-T waves were identified based on the R-peaks.</p> <p>1.3 Learn the synthetic methods of EEG and ECG signal I learned surrogates and optimal transport for synthesis of stationary multivariate series, proposed by Pierre Borgnat, Patrice Abry, and Patrick Flandrin.[2] This method synthesis many different multivariate time-series, where the covariance function and</p>					

marginal or joint-distributions are prescribed either through a model or by empirical properties of measured series. To obtain more realizations of the same series, we can iterate the proposed algorithms. The open-source Matlab code of this synthesis method will be used.

I learned using of Gaussian replicates to improve the performance of the estimator used for classification in ICA Using Spacings Estimates of Entropy [3]. The implementation of replicates using Gaussian noise can be implemented in Matlab using `innoise` command.

1.4 Learn the classification methods of sleep apnea

I have read the following references. Computer-assisted sleep staging [4]; Mixtures of Independent Component Analyzers for Microarousal detection [5]; Automatic sleep staging based on SVD, VMD, HHT and morphological features of single-lead ECG signal [1]; Sleep Apnea Identification using HRV Features of ECG Signals [6]; A review of automated sleep stage scoring based on physiological signals for the new millennia [7].

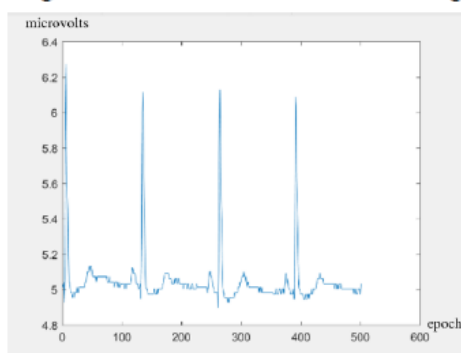
I have used four classification methods, which are linear discriminant analysis, quadratic discriminant analysis, linear SVM and quadratic SVM. The basic idea of LDA is to project high-dimensional pattern samples into the best discriminant vector space to extract classification information and compress the dimension of feature space. After projection, the maximum inter class distance and minimum intra class distance of pattern samples in the new subspace are guaranteed, that is, the pattern has the best separability in the space. Similar to linear discriminant analysis, quadratic discriminant analysis is another linear discriminant analysis algorithm. They have similar algorithm characteristics. The only difference is that linear discriminant analysis is used when the covariance matrix of different classification samples is the same; quadratic discriminant analysis is used when the covariance matrix of different classification samples is different. Support vector machine (SVM) is a kind of generalized linear classifier which classifies data in the way of supervised learning. Its decision boundary is the maximum margin hyperplane of learning samples.

Task 2 Define the dataset

2.1 Visualize the dataset

Similar to Early term

Read the `.mat` file using `load` command. Plot the data using `plot` command.



2.2 Do the feature extraction of the dataset.

Similar to Early term

Use the extracted EEG and ECG signal data set, its name is `data_features.mat`. There are 10 subjects. The data set contains five variables. The first variable is `classes_`, and

each sleep epoch (one epoch 30s) is classified. There are six categories: 0 for wake, 1 for REM, 2 for stage 1, 3 for stage 2, 4 for stage 3, and 5 for stage 4. features_ECG includes all the ECG features of 10 subjects, and each epoch extracts 30 features. features_EEG includes all the EEG features of 10 subjects, and each epoch extracts 8 features. features_Sub includes EEG and ECG signals of 10 subjects. subs represent 10 subjects.

2.3 dataset pre-processing

The EEG and ECG signal features of 10 subjects were divided into training set and test set. Here, follow a "leave - one out" procedure. This means that each subject turns out to be a test set, and the rest of the subjects' data is used as a training set.

Task 3 Study and implement two synthetic data algorithm

3.1 Adding Gaussian noise to the dataset to generate synthetic data

Use innoise command in MatLab can add Gaussian noise to the dataset. For example, we take the EEG and ECG features of subject 3 as the test set, and the EEG and ECG features of the remaining 9 subjects as the training set. We add Gaussian noise to the training set. Before we add gaussian noise data, the variables of the dataset have to be normalized using the command "zscore". The following figures show the original data and the synthetic data by adding Gaussian noise.

	1	2	3	4	5	6	7	8	9	10
1	-0.4955	-0.5108	-0.5756	-0.4422	0.4849	-0.3992	1.6370	0.7483	-0.2815	0.2559
2	-0.0175	-0.3252	-0.3322	0.1407	1.8855	0.0392	1.5266	1.9491	-0.2380	0.1428
3	2.7267	1.4967	3.4898	4.0287	4.3072	2.7363	0.1348	1.9328	0.5363	-1.2734
4	-0.4997	-0.4637	-0.4748	-0.3843	-0.1732	-0.3884	0.0504	1.1015	-0.3771	0.3680
5	0.0994	-0.3478	-0.4408	-0.3946	-0.1319	2.2584	-1.2523	1.9500	0.0849	-0.2978
6	-0.5497	-0.5203	-0.5303	-0.4884	-0.2433	-0.3562	-0.2697	1.3398	-0.2558	0.3006
7	1.6130	1.9175	0.2801	0.7282	0.9484	4.3031	-0.7845	2.2602	0.5347	-1.3691
8	3.9278	0.6366	0.5106	0.7625	2.3928	2.3537	-0.5322	1.6433	0.1735	-0.6187
9	-0.5702	-0.4445	-0.1457	-0.4829	-0.2069	-0.4420	1.1432	1.1832	0.5370	-0.8745
10	-0.5705	-0.3793	0.1352	-0.4395	-0.1757	-0.4499	1.5338	0.6678	0.5984	-0.9553

Part of the original data

	1	2	3	4	5	6	7	8	9	10
1	0.1240	0.0800	0.0743	0.1621	0.5368	0.0536	1	0.8106	0.1067	0.3465
2	0.1820	0.1010	0.1076	0.2415	1	0.1878	1	1	0.0723	0.2227
3	0.9990	1	1	1	1	1	0.2804	1	0.6781	0.1028
4	0.1386	0.1037	0.1122	0.1128	0.0881	0.1816	0.1248	1	0.0679	0.5384
5	0.2137	0.1010	0.0618	0.1207	0.1732	1	0.0798	1	0.2415	9.0704e-04
6	0.0415	0.1160	0.1264	0.0814	0.1187	0.1315	0.1264	1	0.0750	0.3796
7	1	1	0.3727	0.8699	1	1	0.0773	1	0.6219	0.0836
8	1	0.7108	0.5185	0.8654	1	1	0.1528	1	0.2353	0.0674
9	0.2600	0.1157	0.1086	0.1483	0.1193	0.1550	1	1	0.6044	0.0994

Part of the synthetic data by adding Gaussian noise

3.2 Using Surrogates to generate the synthetic data

I download the implementation of the method proposed by Pierre Borgnat, Patrice Abry, and Patrick Flandrin. Read the code.

Task 4 Implement the sleep apnea classification algorithm

4.1 Implement the sleep apnea classification algorithm

Four classification methods are used, which are linear discriminant analysis, quadratic discriminant analysis, linear SVM and quadratic SVM. These methods are applied to 10 training sets in task 2. For example, take subject 1 as the test set and the other 9 subjects as the training set. The trained model is used in the test set to do the classification of the category. The classification results of each method are shown below. The red line is classified classes, while the green line is correct classes. The confusion matrix is generated using confusionmat command.

- Linear discriminant analysis

Synthetic data methods applied to sleep stage analysis

```
>> confusionmat(result, testclass)
```

ans =

34	0	2	1	1	0
2	20	10	6	2	2
62	61	78	15	1	2
15	74	123	147	24	53
1	0	0	0	0	0
8	0	0	3	1	1

● Quadratic discriminant analysis

```
>> confusionmat(result2, testclass)
```

ans =

82	21	89	47	18	44
0	0	0	0	0	0
3	3	3	0	0	0
37	131	121	125	11	14
0	0	0	0	0	0
0	0	0	0	0	0

● Linear SVM

```
>> confusionmat(result3, testclass)
```

ans =

24	0	2	1	1	0
5	128	25	17	1	0
51	14	61	7	0	0
16	7	87	94	17	19
0	0	0	0	0	0
26	6	38	53	10	39

● Quadratic SVM

```
>> confusionmat(result4, testclass)
```

ans =

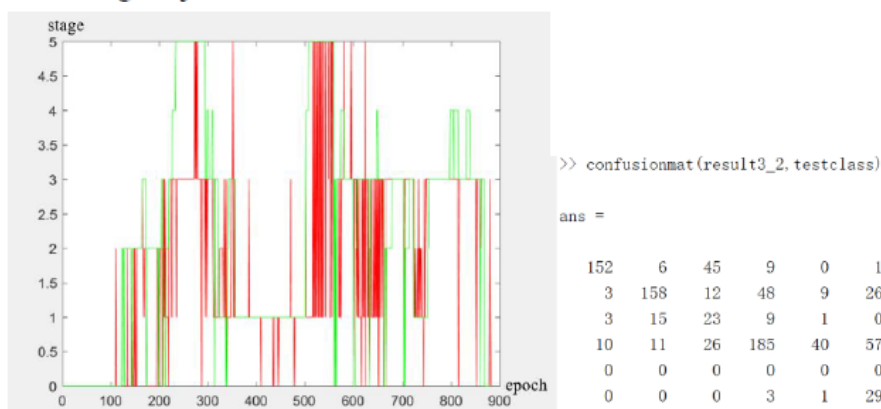
43	2	6	1	0	0
5	123	75	59	6	1
29	19	33	2	0	0
2	0	7	19	2	1
23	1	43	27	3	2
20	10	49	64	18	54

4.2 Evaluate the accuracy of the classification results.

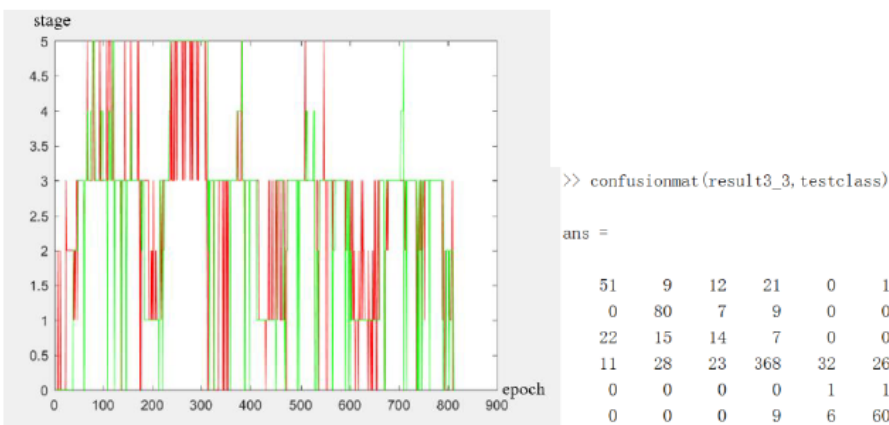
The accuracy is calculated by comparing the classified results with the correct results. $accuracy = \frac{correct}{correct + false}$. For the example of using the data of subject 1 as the test set and other data as the training set, the classification accuracies of the four methods are 0.37, 0.28, 0.46, 0.36. After comparing the classification results of different training sets and different classification methods, we can know that the classification result of linear SVM is the best in those four classification methods.

4.3 Compare the classification result

After that, we compare the classification results of different training sets trained by linear SVM on different test sets. The following figure shows the classification results using subject 2 as the test set and the other subjects as the training set. The accuracy is 0.62, which is better than using subject 1 as the test set.

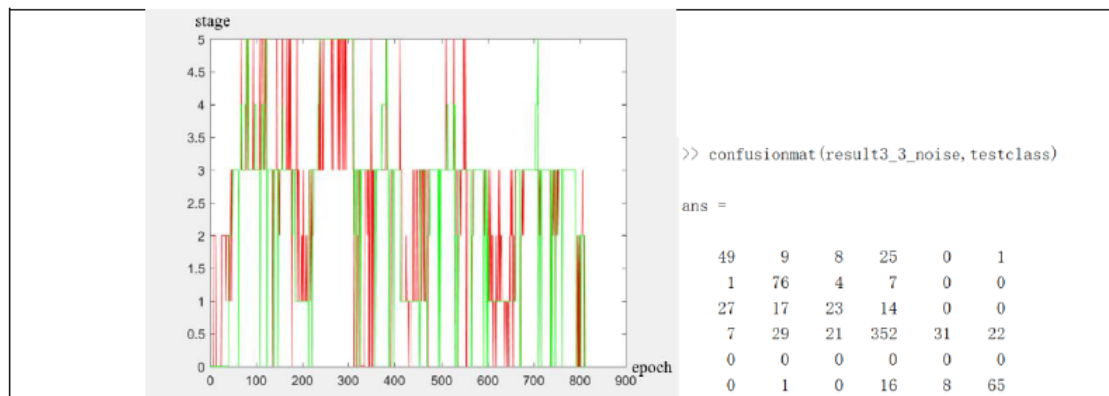


The following figure shows the classification results using subject 3 as the test set and the other subjects as the training set. The accuracy is 0.7.



Using different training set and test set partition methods, the classification results have obvious differences.

Then the synthesized data (Generated in task 3, at present, only use synthetic data with Gaussian noise) and the original data are combined together as a new training set, and the training set is trained by linear SVM. For example, we use data from subjects other than subject 3 as the training set. Then, subject 3 is used as the test set to classify the sleep categories. The classification result is shown below. The accuracy is 0.695.



尚需完成的任务 Work to do:

Task 3 [Study and implement two synthetic data algorithm]

3.3 Using Surrogates to generate the synthetic data

Read the sample code, complete the code to generate synthetic data

Task 4 [Evaluate the performance of the method and compare it with alternative approaches]

4.4. Optimize the algorithm

Improve classification algorithm. I can combine the classification algorithm and improve the accuracy of classification

Essay Writing.

Reference:

- [1] Ş. Yücelbaş, C. Yücelbaş, G. Tezel, S. Özşen and Ş. Yosunkaya, "Automatic sleep staging based on SVD, VMD, HHT and morphological features of single-lead ECG signal", *Expert Systems with Applications*, vol. 102, pp. 193-206, 2018. Available: 10.1016/j.eswa.2018.02.034.
- [2] P. Borgnat, P. Abry and P. Flandrin, "Using surrogates and optimal transport for synthesis of stationary multivariate series with prescribed covariance function and non-gaussian joint-distribution", *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012. Available: 10.1109/icassp.2012.6288727 [Accessed 11 March 2021].
- [3] E. G. Learned-Miller and J. W. Fisher III, "ICA using spacings estimates of entropy", *The Journal of Machine Learning Research*, vol. 4, 2003. Available: 10.5555/945365.964306 [Accessed 11 March 2021].
- [4] R. Agarwal and J. Gotman, "Computer-assisted sleep staging", *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 12, pp. 1412-1423, 2001. Available: 10.1109/10.966600.
- [5] G. Safont, A. Salazar, L. Vergara, E. Gomez and V. Villanueva, "Mixtures of independent component analyzers for microarousal detection", *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2014. Available: 10.1109/bhi.2014.6864473 [Accessed 11 March 2021].
- [6] B. Sulistyono, N. Surantha and S. Isa, "Sleep Apnea Identification using HRV Features of ECG Signals", *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 5, p. 3940, 2018. Available: 10.11591/ijece.v8i5.pp3940-3948.
- [7] O. Faust, H. Razaghi, R. Barika, E. Ciaccio and U. Acharya, "A review of automated sleep stage scoring based on physiological signals for the new millennia", *Computer Methods and Programs in Biomedicine*, vol. 176, pp. 81-91, 2019. Available: 10.1016/j.cmpb.2019.04.032.

Synthetic data methods applied to sleep stage analysis

存在问题 Problems:

1. Using synthetic data, the accuracy of classification is not greatly improved
2. Using different partition methods of training set and test set, the classification results of trained models are very different. A better way to divide data sets may be needed

拟采取的办法 Solutions:

1. Try to add different decibels of Gaussian noise (different signal to noise ratios) and analyze the experimental results. Use other methods to generate synthetic data.
2. Using more reasonable data set partition method, changing the proportion of synthetic data that are combined with original data in training stage.
3. Try to implement different classification problems other different classification problems in order to analyze classification results, by combining sleep stage classes.

论文结构 Structure of the final report:

- A. Introduction
- B. Dataset
 - a) Dataset Description
 - b) Dataset visualization
- C. Generate synthetic data
 - a) Adding Gaussian noise
 - b) Using Surrogates
- D. Classification
 - a) Linear discriminant analysis
 - b) Quadratic discriminant analysis
 - a) Linear SVM
 - b) Quadratic SVM
- D. Result
 - a) Evaluation
 - b) Analyses
- E. Conclusion
- F. Reference

Synthetic data methods applied to sleep stage analysis

- 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	Wei	名 First Name	Jingjing		
BUPT 学号 BUPT number	2017213152	QM 学号 QM number	171044142	班级 Class	2017215120
论文题目 Project Title	Synthetic data 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: 20-10-2020 Supervision type: email Summary: provide references and open-source data sets. Describes what the entire project needs to do					
Date: 16-11-2020 Supervision type: email Summary: received written feedback on the draft specification					
Date: 12-1-2020 Supervision type: email Summary: discussed the early term progress and the draft version of the early term report					
Date: 2-2-2020 Supervision type: email Summary: received the database containing the extracted features from EEG and ECG signals and some tasks that need to do.					
Date: 2-26-2020 Supervision type: email Summary: discussed the mid-term progress and the draft version of the mid-term report					

Risk and environmental impact assessment

Risk and environmental impact assessment are shown in below.

(1) Prevents the successful completion of the project

Event	Scores for level of likelihood L	Scores for level of consequence C	Result R	Rating of risk	Action
There is a problem with the equipment that collects EEG and ECG signals, which leads to collecting wrong data in the open-source dataset	1	2	2	Low Risk	Use other open-source EEG and ECG dataset
The computer crashed while training the model	2	1	2	Low Risk	Replace the computer and retrain the model
In the process of collecting EEG and ECG signals, the staff improper operate equipment, leading to collecting wrong data in the open-source dataset	1	2	2	Low Risk	Use other open-source EEG and ECG dataset

Synthetic data methods applied to sleep stage analysis

(2) Causes potential harm to people and /or animals

Event	Scores for level of likelihood L	Scores for level of consequence C	Result R	Rating of risk	Action
Impact the health of subject due to the improper operation during EEG and ECG signal collecting process	0	2	0	No Risk	No action

(3) Causes potential harm to the environment

Event	Scores for level of likelihood	Scores for level of consequence	Result R	Rating of risk	Action
The equipment for collecting EEG and ECG signals used in open-source dataset is damaged and needs to be discarded.	2	1	2	Low Risk	Suggested the University that publishes this open-source dataset address the damaged equipment in an environment friendly manner to reduce the impact on the environment

Synthetic data methods applied to sleep stage analysis

(4) Causes potential financial loss to the project or to other individuals or organizations.

Event	Scores for level of likelihood	Scores for level of consequence	Result R	Rating of risk	Action
The equipment for collecting EEG and ECG signals used in open-source dataset is damaged and needs to be discarded.	2	1	0	Low Risk	Suggested the University that publishes this open-source dataset buy the new equipment. Use other open-source EEG and ECG dataset