University of Twente

Master Thesis - Confidential Draft

Investigation on factors limiting the performance of deep sleep classification

Author:
Yuan Lu

Supervisor:
Mannes Poel
Pedro Miguel Fonseca
Gary Garcia Molina
Femke Nijboer

Human-Media Interaction Program
Faculty of Electrical Engineering, Mathematics and Computer Science

The content of this document is intended for reviewing by the supervisors stated above.

April 2014

Contents

Chapter 1

Introduction

Sleep is important for people. People spend nearly half of their lifetime in sleep. Sleep problems are harmful for health. Lots of people suffer from sleep problems, such as sleep apnea, insomnia and sleep deprivation etc. These sleep problems interfere with normal physical and mental functioning of human body. About 17-24% of North American adults are affected by obstructive sleep apnea, which increases the risk of sudden death [1]. A lot of people, who do not have sleep disorders, still struggle with tiredness and lack of motivation resulted from sleep. Therefore, sleep studies emphasizes on identifying elements interfering with sleep in order to design interventions aiming at relieving symptoms of sleep disorders and improving sleep quality.

Conventional sleep monitoring procedure and interpretation of sleep recordings are cumbersome tasks. Automatic sleep staging system is therefore developed to remove human factors in the interpretation of sleep recordings. Other modalities for sleep monitoring are investigated as well, such as cardiorespiratory signals which exhibit sufficient ability in sleep staging. The new monitoring modality shows the possibility for the development of unobtrusive sleep monitoring.

Among different sleep stages, deep sleep is an important sleep stage. The importance of deep sleep in learning have been proved [2][3]. The features differentiating sleep stages and various classifiers are largely investigated for the sleep classification. Yet studies on the elements limiting the performance of sleep classification has received little attention so far.

1.1 Objectives

An early investigation of the deep sleep classification was conducted prior to this research. The early investigation aimed at utilizing features, which were derived from cardiorespiratory signals, on the deep sleep classification. The result of the early investigation indicates some aspects which may limit the performance of deep sleep classification.

Two aspects are explicitly stated that influence the classification performance. Although a series of features, which are dedicated to the discrimination between deep sleep and non-deep sleep, are extracted from the cardiorespiratory signals, there are still characteristics of deep sleep that are not captured by these features. For example, the discontinuity of deep sleep, which illustrates a considerable correlation with the classification performance (Spearman $\rho = -0.5$), is not reflected in the extracted features. Therefore, we suspect the classification performance depends on not only the quality of the extracted features, but also some characteristics of deep sleep external to the extracted features. Another aspect limiting the classification performance is the subject-independent property of the constructed classification model. This property ensures that the model performs good in general, but is not able to consider the characteristics in the features for specific subjects. Thus, we suspect the classification performance can be improved if a personalized classifier is constructed.

Based on the two aspects introduced above, two hypotheses are formulated accordingly. The first hypothesis deals with the relationships between classification performance and deep sleep characteristics. The second hypothesis concerns the improvement brought by personalized classifier. The scope of this master project is focused on the validation of the two hypotheses. The validation of the two hypotheses can provide guidance for the future work of deep sleep classification.

1.2 Methods

The validation of the two hypotheses uses a dataset of 45 subjects. For the validation of the first hypothesis, the first-night data of the 45 subjects are segmented into sleep cycles [4] to ensure the data sufficiency for the analysis. For each cycle, several variables are defined according to sleep characteristics and observations from the result of the early investigation. Multiple linear regression analysis [5] is employed to quantify the relationship between the defined variables and the classification performance. The validity of the regression model is assessed by residual analysis [6]. The significance of the variables and the model is justified by student's t-test [7] and F-test [8] respectively.

The adequacy of the model is evaluated by R-squared [5]. Two-night recordings of the 45 subjects and segmented two-night recordings of 13 subjects, who are selected among the 45 subjects, are utilized for the validation of the second hypothesis. Training and evaluation of the trained classifier is performed using Leave-One-Out Cross-Validation (LOOCV) [9]. The classification performance is evaluated using the area under the *Precision-Recall* curve, which is a metric more suitable than the area under the ROC curve in the case where class-imbalance problem exists. Partial data and ground-truths of the test subject/cycle are used for the construction of personalized classifier. The performance of the personalized classifier is compared to that of a subject-independent classifier.

1.3 Organization

Chapter 2 gives an overview of the sleep physiology. It also summarizes the past work on unobtrusive sleep staging. The classification framework is introduced in detail as well. Furthermore, the result of an early investigation on deep sleep classification is explained in this chapter. In the end of Chapter 2, the two hypotheses that need to be validated are presented. Chapter 3 describes the methods and procedures of the validation of the two hypotheses. The description of the experiment setup and dataset employed in this research are also introduced. The result of validations and detailed discussion are presented in Chapter 4. Chapter 5 and Chapter 6 describe the conclusions and future work respectively.

Chapter 2

Background

2.1 Sleep physiology

Sleep is a period during which the brain is disconnected from external environment. When in sleep, people remain in quiescence and the vigilance are reduced. Although the precise function of sleep is not fully understood, it has been proved that the brain benefits from sleep [10].

Sleep is not a homogeneous state, but a period during which both the brain and the body enter different states. There are two broad types of sleep: rapid eye movement (REM) sleep and non-rapid eye movement (NREM or non-REM) sleep. NREM sleep can be further divided into four stages according to the standardized criteria by Allan Rechtschaffen and Anthony Kales in the R&K sleep scoring manual [11]: S1, S2, S3 and S4. In 2007, the R&K scoring system was reviewed by the American Academy of Sleep Medicine (AASM) [12], which resulted in several changes of which the most significant being the combination of S3 and S4 into NREM stage 3 (N3). In the R&K criteria, S3 and S4 denote *deep sleep* stage while N3 denotes *deep sleep* stage in the AASM criteria.

The dominant brainwaves featuring different stages are presented in electrocephalography (EEG). EEG records the electrical activity within the brain [13]. Sleep studies employ EEG, along with other methodologies, such as accelerometer and electromyography (EMG), to identify or rule out sleep disorders. With the help of electrooculography (EOG) and submental EMG, EEG can also provide ground-truth for automatic sleep stage classification. Characteristics of different sleep stages are summarized below (AASM guideline for sleep staging are adopted in the following explanation).

- **N1**.

NREM stage 1 (N1) is a stage between sleep and wakefulness. The dominant brainwaves in N1 stage is theta waves. Theta waves oscillates in the 4-8 Hz frequency bands and its amplitudes are approximately 10 microvolts, which differentiate them from the dominant brainwaves when awake.

- **N2**.

  NREM stage 2 (N2) follows N1 where theta wave activity continues. Besides that, two other abrupt activities, sleep spindles and K-complexes, start to happen. A K-complex is an EEG waveform and it is the "largest event in healthy human EEG" [14]. K-complexes help suppressing cortical arousal in response to external stimuli and aiding sleep-based memory consolidation [14]. A sleep spindle is a brain activity occurring during N2. It consists of 12-14Hz waves that occur for at least 0.5s [15]. Sleep spindles function to help the brain inhibiting processing, ensuring the sleeper stay in calm.

  N1 and N2 are considered to be *light sleep*. People may not admit to be asleep if they are awoken during *light sleep*.

- **N3**.

  NREM stage 3 (N3), also called slow-wave sleep (SWS), is the deep sleep stage. The dominant brainwaves for N3 are delta waves. A delta wave is a high amplitude brainwave with a period approximately equal to 1 second. N3 is the most difficult sleep stage to wake sleepers when the brain shows the least reaction to environmental stimuli. As sleepers move into N3, their breaths become smooth and steady and their heart rates become more regular. If somesone is awoken during deep sleep stage, he or she will feel sleepy or disoriented.

  N3 is thought to be an important part of sleep. It is closely related to learning and memory. The brainwaves, which dominate slow-wave sleep, are essential for memory consolidation. It is proved that stimulations, which are given during slow-wave sleep, enhance the retention and consolidation of declarative memory, while stimulations out of slow-wave sleep leave it unchanged [16][17][18].

- **REM**.

  REM is a sleep stage accompanied by rapid eye movements. Most dreams occur in this sleep stage. In REM, the exhibiting brainwaves are similar to those in the wake state. REM differentiates from the wake state in EOG and submental EMG. There are eye blinks and movements in EOG when awake, while REM exhibits episodic REMs. In the submental EMG, the muscle activities are relatively reduced when in REM stage while the muscle activities are high in the wake state [19].

Stage	EEG Frequency	EEG Amplitude
Awake	8-13 Hz	Low
N1	4-8 Hz	Low
N2	4-7 Hz Sleep spindles and K complexes	Medium
N3	1-3 Hz	High
REM	More than 10 Hz	Low

TABLE 2.1: Brainwaves featuring different sleep stages.

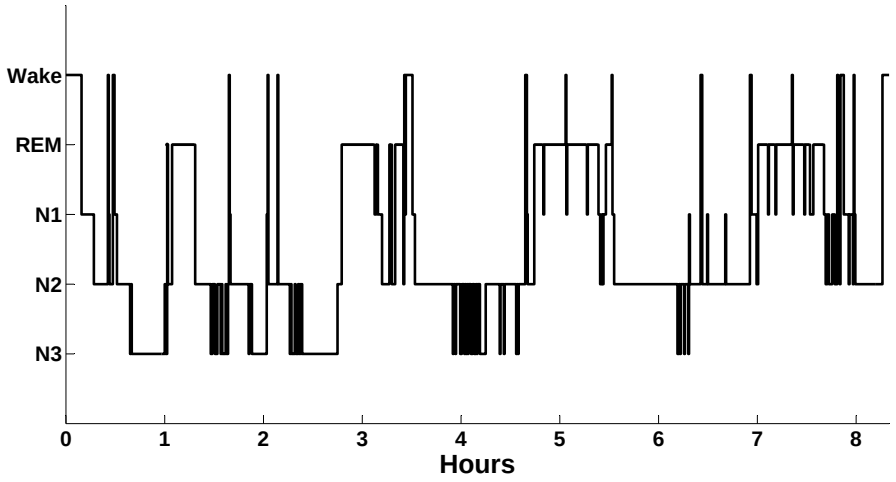


FIGURE 2.1: Hypnogram of a healthy subject.

Both the intensity and duration of deep sleep (N3) are greater in the first half of the night, while the proportion of REM sleep increases as sleep progresses. Considering the total sleep time, a great amount of time is spent in N2, approximate 45-55% of total sleep time. Only 5-15% of total sleep time is consisted of N3. N1 and REM represents 2-5% and 20-25% of total sleep time respectively.

In sleep study, the golden standard for sleep quality assessment and sleep disorder identification is overnight polysomnography (PSG). Such a clinical procedure is carried out in a specialized hospital-based laboratory. Well-trained sleep technicians manually annotate sleep stages by visually examining electroencephalography (EEG, brainwaves), eletrooculography (EOG, eye movements) and electromyography (EMG, muscle activity). Annotations are represented by a graph called hypnogram. Hypnogram represents the stages of sleep as a function of time. In both R&K and AASM standards, annotation is performed on non-overlapping 30-second epochs. Thus, a healthy subject (a healthy subject in the context of this research is a subject who does not have sleep disorder), whose sleep lasts for 6-9 hours, will have a hypnogram consisting of 720-1080 epochs. Fig.2.1 shows a typical hypnogram of a healthy subject in one night scored by PSG. It

could be seen from Fig.2.1 that the subject cycles between NREM and REM sleep. A full-night sleep is comprised of several sleep cycles [4]. The notion of a sleep cycle is fuzzy though. Roughly speaking, it is limited in time by a sequence NREM-REM. A sleep cycle begins with a period of NREM sleep followed by a period of REM sleep. One sleep cycle takes approximate 90 to 100 minutes, thus there are four to five complete sleep cycles for an average sleep time of 8 hours.

2.2 Unobtrusive sleep staging

As mentioned earlier, traditional sleep study requires overnight PSG and manually scoring sleep stages for the recordings. Both procedures are cumbersome tasks. The PSG-based methods for acquiring sleep data may be uncomfortable for subjects and may interfere with their normal sleep. Also, the obtrusive measuring make it difficult for long-term monitoring of several months. Due to these reasons, unobtrusive sleep monitoring methods have received significant attentions from the research community [20][21][22].

In addition to brainwaves, there are other physiological characteristics, which can be captured unobtrusively, that are capable in sleep staging, such as cardiac activity, respiratory activity and body movements. For example, the average duration of the inter-beat intervals is longer in deep sleep than light sleep and the cardiac signals are more regular in deep sleep than in other stages [23]. The monitoring of cardiorespiratory signals, compared to the monitoring of the EEG signals, is easier to implement. For example, in order to monitor cardiac activity, a Lead II configuration can be used to measure the voltage between the right arm and the left leg. Although monitoring cardiac activity is also obtrusive, compared to EEG, it causes less discomfort in subjects. With advanced technology, cardiac activity can be unobtrusively monitored by a ballistocardiography-based (BCG-based) system. BCG is non-invasive technique that can be measured with sensors installed in the bed to monitor the vibrations of human body which are caused by cardiac and respiratory activities [24][25][26]. Therefore, it can be expected that, by extracting elaborate features, it is possible to distinguish sleep stages from ECG signals and respiratory effort. In the research of sleep staging, it has been shown that cardiorespiratory signals are able to differentiate deep sleep stage from other stages [27][28][29]. Willemen et al. [27] reports a non subject-specific deep sleep versus light sleep classification result of 0.55 in Cohen's κ [30]. The easier signal acquiring procedure and the good performance in sleep staging make the cardiorespiratory signal a competitive candidate for unobtrusive sleep monitoring and staging.

Although this research is still based on obtrusively acquired cardiorespiratory signals, the results can be translated to the fully unobtrusive case as long as the sensors used can give the same information regarding cardiac and respiratory activities in order to extract essentially equivalent features.

Most studies, which mention deep sleep classification, focus on multiple-sleep staging using ECG and/or respiratory signals [27][28][29]. Few published studies exclusively focus on deep sleep classification using cardiorespiratory signals. Only one research, done by Shinar et al. [31], used heart rate variability information on deep sleep classification, achieving an accuracy of 80%.

2.3 Framework for deep sleep classification

This section summarizes the steps within the classification process. The described process follows a common routine of machine learning.

Sleep stages are defined at a 30-second resolution according to AASM guideline. In other words, a sleep stage is assigned to each 30-second long segment (i.e. one epoch) of sleep for each subject. In this framework, each epoch is assigned to either 'deep sleep' or 'non-deep sleep'. The goal of the $classifier$ is to find a mapping between the sleep recordings and sleep stages. The classification result of a classifier is compared to *annotations* given by sleep experts. The result of the classifier is expected to match as closely as possible that of a manual scoring.

There are four main steps within the classification process, which are shown in Fig.2.2.

- Feature extraction. The goal of feature extraction is to extract relevant information from the recordings. The distinction between deep sleep and non-deep sleep can hardly be seen from the raw sensor recordings. To better describes the difference between the two classes, a series of features are designed to extract the characteristics from the recordings.
- Feature normalization. Features extracted from the raw recordings are further processed to adapt to the requirements of the classifier, referred to as feature normalization. Feature normalization helps to reduce the between-subject variability in features. Thus, the performance of a generalized classifier is improved. The classifier maps a set of normalized features to a class.
- Classifier training. During training, a classifier is tuned to a selected feature subset, so that the generalization ability of the classifier is maximized. Feature selection is also performed, prior to the classifier training, in order to lower the complexity.

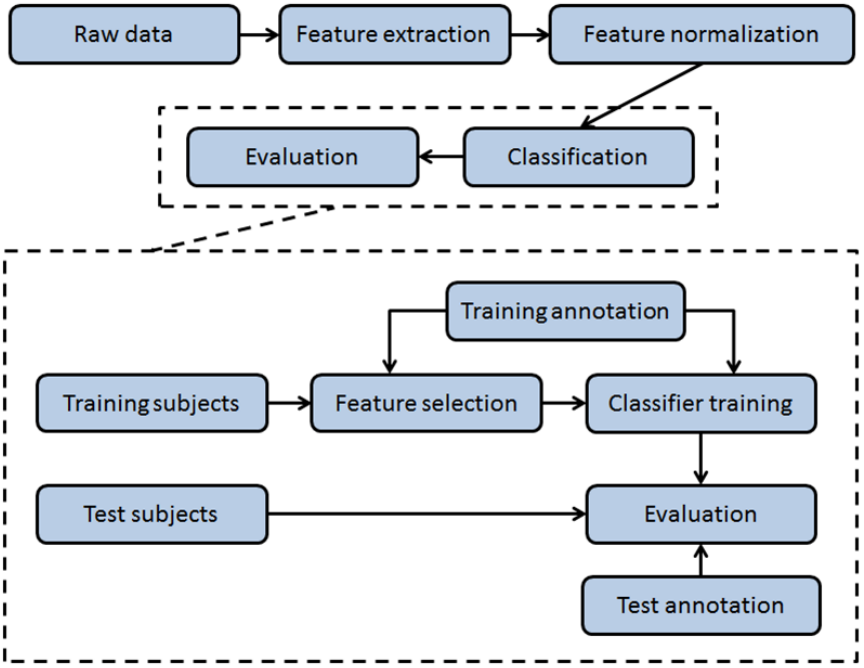


FIGURE 2.2: Flow-process diagram of the deep sleep classification framework.

- Performance evaluation. The training and evaluation of the classifier are performed utilizing separate sets of subjects. Evaluation is usually done by splitting the dataset in two, one is for classifier training and the other is for evaluation. In practice, this method may suffer from lack of data. Therefore, we employ the Leave-One-Subject-Out Cross-Validation (LOSOCV) paradigm for training and evaluation, which will be explained in Chapter 3.

2.3.1 Feature overview

Features are computed from the cardiorespiratory signals. Each feature describes a characteristic of the cardiac and respiratory system. For example, the shapes of the respiratory signals during deep sleep are more similar than those during the non-deep sleep. Thus, a features expressing this characteristic would be a good candidate feature. Though the sleep recordings are recorded at different sampling rates, features are required to describe each epoch with a single value, either an integer or a real number. The computation of the features is usually restricted to a single epoch. For features which describe the long-term change in cardiac and respiratory system, the computation will not only consider the current epoch but also take the adjacent epochs into account. Fig.2.3 gives an example of extracted features, which depicts the change of mean heart rate over the entire night.

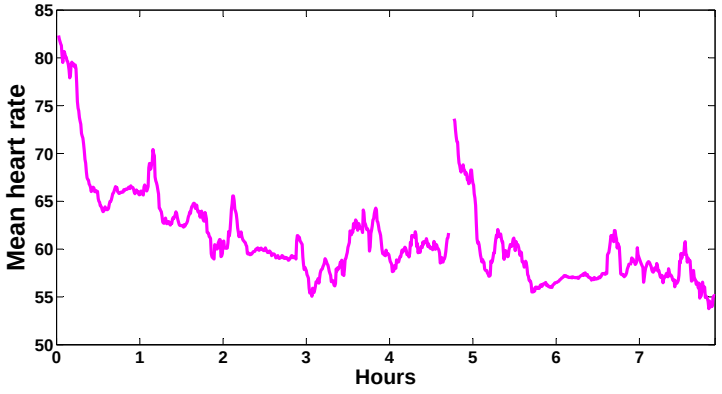


FIGURE 2.3: Mean heart rate over the entire night for one subject.

Two problems, which occurs frequently in features, are demonstrated in Fig.2.3: missing values and trends in features. The missing values in features will cause problem for those classifiers whose classification score is based on a linear combination of feature values. Removing the epochs with missing values is not allowed since this will disrupt the time series, and data coverage is decreased as well. To deal with the missing value problem, cubic interpolation is employed. The values in epochs with invalid data are interpolated according to the values of adjacent epochs. Since cubic interpolation aims at interpolating using a smooth curve, it sometimes assigns extreme values for the missing epochs (extreme values means that the assigned values far exceed the range of the before-interpolation features). To avoid outliers in the after-interpolation features, only values within the 2th percentile and 98th percentile of the total feature values are considered as valid values. For those epochs whose values are outside the range, their values will be set to the 2th percentile or the 98th percentile of the total feature values accordingly. Trends in features is another problem since it is difficult for many classifiers to adapt to the changes over time within the classifier itself. Thus, detrending is performed for some features, most of them are features describing heart rate.

2.3.2 Cardiorespiratory features

This research utilizes many features which are claimed to be good at discriminating deep sleep and non-deep sleep in literature. Thus, for each subject, 125 features are computed. There are correlated features among the 125 features. Though the features contain mutual information, they are still useful since they may carry unique information that is good for classification but hard to isolated. Most features are ECG features, which occupies 84 of 125.

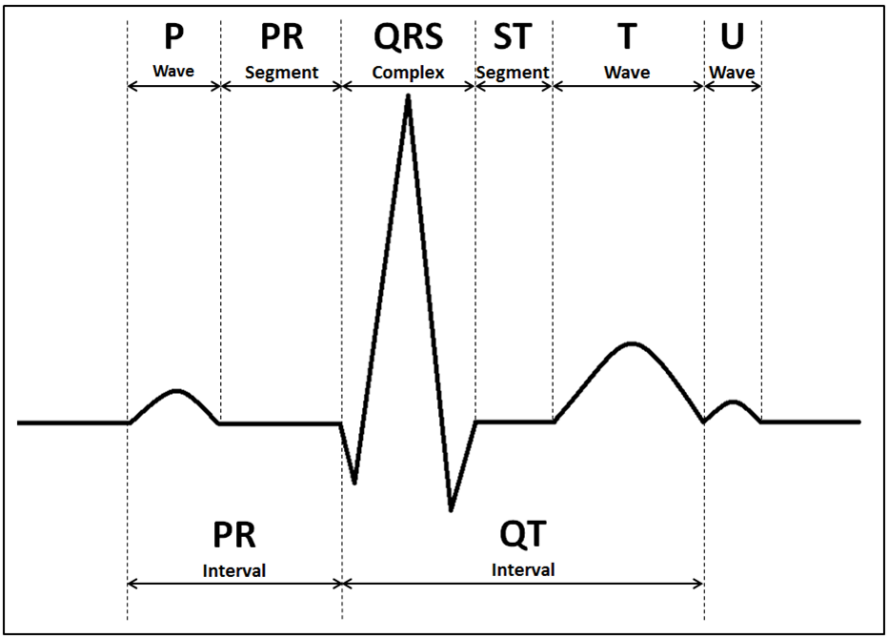


FIGURE 2.4: Schematic diagram of ECG waveform and attributes.

Since features are not the main focus of this research, we only provide a short description of the most important features.

ECG features

To introduce ECG features, we first introduce ECG (Electrocardiogram) signal. ECG interprets the electrical activity of the heart, which is generated by the polarization and depolarization of cardiac tissue [32]. A single heartbeat ECG waveform and its attributes are given in Fig.2.4. Features can be derived from RR interval, ECG-derived respiration (EDR) information and raw ECG signal.

RR interval represents the time between two successive R waves. It captures the instantaneous heart rate. Features derived from RR interval describe the heart rate variability. Some example features, expressing the heart rate variability in the time domain, include [33]:

- Mean of the normalized RR interval duration.
- Standard deviation of the normalized RR intervals.
- Range of RR intervals.

Features derived from RR interval in the frequency domain using power spectral density (PSD) analysis. For cardiac analysis, three frequency bands are employed to the RR interval time series, which are Very Low Frequency (VLF, 0.005-0.04 Hz), Low Frequency

(LF, 0.04-0.15 Hz) and High Frequency (HF, 0.15-0.45 Hz) [34]. The power in each of these frequency bands are shown to be good for discriminating deep sleep [35][31].

Another two types of ECG features that are proved to be useful for deep sleep classification employ detrended fluctuation analysis (DFA) and multiscale sample entropy analysis respectively [36][37][38][39]. DFA eliminates trends in time series in order to study long-range correlations in the data. Multiscale sample entropy analysis conducts a multi-scale (temporal and spatial scales) study to discover the interactions in the physiologic systems.

Respiratory Features

Respiratory features measure the time domain as well as the frequency domain information in the respiratory system. Similar to ECG signal, the extracted respiratory features describe the respiration variability and contents in the same frequency bands (VLF, LF and HF) as the cardiac features. In addition, some features related to respiration amplitudes were extracted. Some important respiratory features are [40]:

- Standardized mean and median value of peaks and troughs of the respiration amplitudes.
- Median breath volume over time.
- Approximate entropy for peaks and troughs of the respiration amplitudes.

2.3.3 Feature normalization

Extracted features exhibit significant between-subject variability. For example, Fig.2.5 represents a box plot illustrating the distribution of a feature describing heart rate variability. Since this framework generates a generalized classifier, the generalized classifier is obtained given various subjects. Such big between-subject variability makes the generalized classifiers difficult to achieve the optimal performance across the 45 subjects. Its classification performance is limited by the variation between subjects. With the feature in Fig.2.5, the generalized classifier is unable to find a constant threshold that could achieve a good separation between two classes for all the subjects. Therefore, features are normalized to eliminate the between-subject variability.

It is displayed in Fig.2.6 the normalized heart rate variability feature. The normalized feature of each subject is comparable. For each feature, one additional feature is generated after feature normalization. Thus, each subject has 250 features in total.

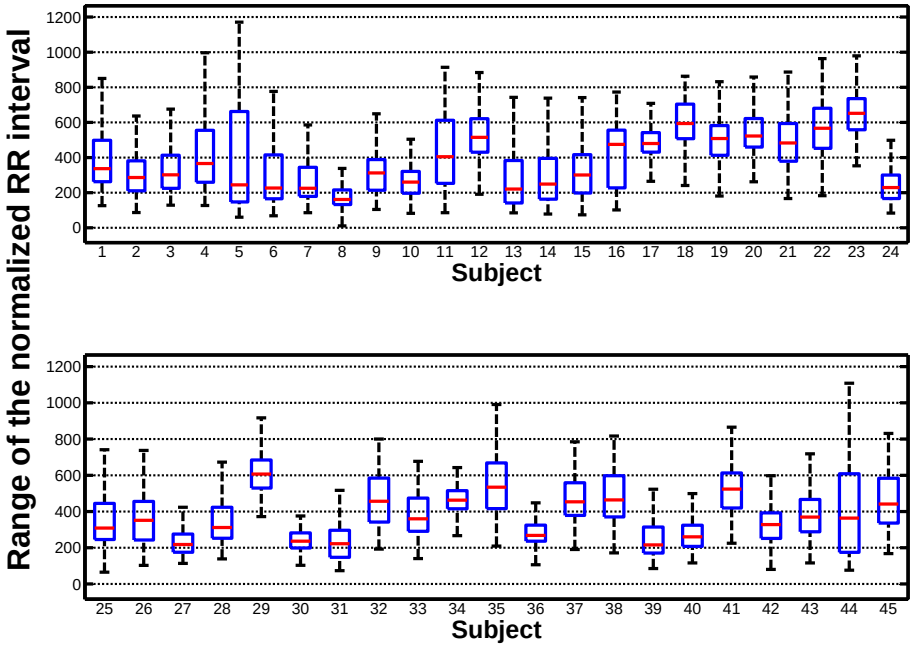


FIGURE 2.5: Distribution of the heart rate variability feature of 45 subjects.

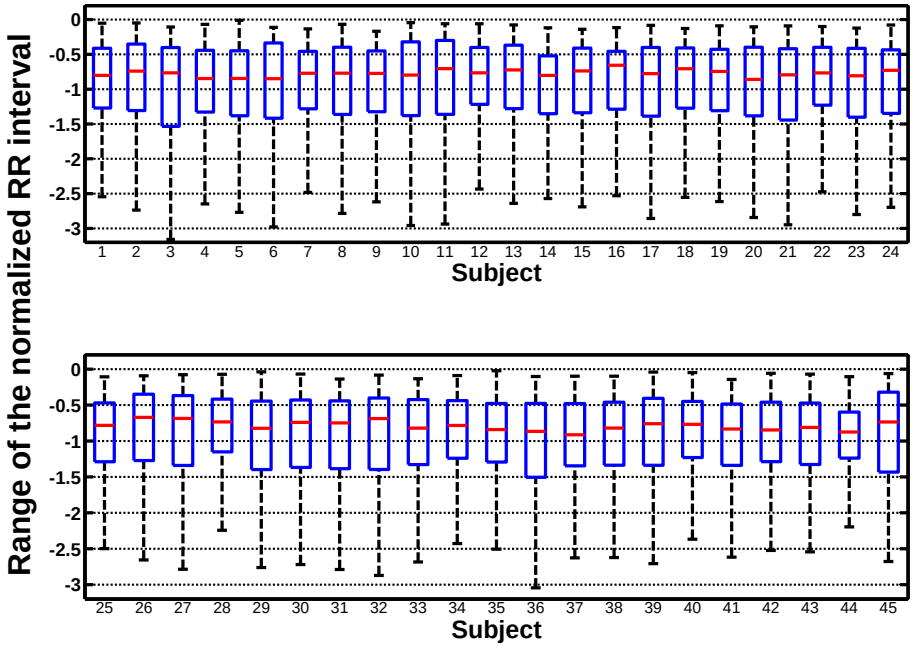


FIGURE 2.6: Distribution of the heart rate variability feature of 45 subjects after feature normalization.

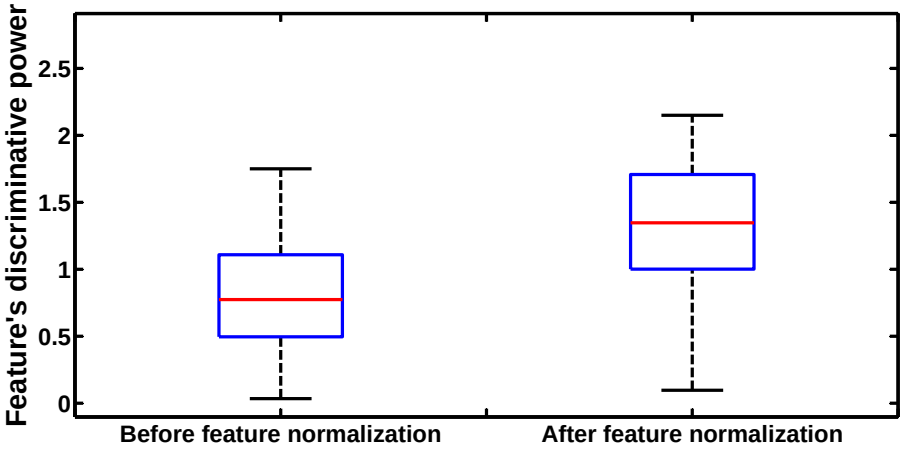


FIGURE 2.7: Distribution of the discriminative power of the heart rate variability feature among 45 subjects before and after feature normalization. Feature normalization improves the discriminative power of the original feature.

2.3.4 Classification

Classification includes feature selection and classifier training. Feature selection helps to remove redundant information in features and improve the efficiency and efficacy for the classifier. The classifier realizes the mapping between a set of features and classes. Detail explanations of the classification process are in Chapter 3.

2.4 Initial results

This part summarizes the work of an early research investigation conducted prior to this master project. The results of the early investigation are the initial results and the conclusions drawn from the results motivate the work of this master project. The work of the early investigation mainly targets on two tasks: one is to resolve the class-imbalance problem in the dataset and the other is to test the performance of different classifiers on deep sleep classification.

When introducing deep sleep, we have mentioned deep sleep only occupies 5 - 15% of total sleep time. Therefore, in deep sleep classification, the positive class is the minority class. In the dataset, on average, the ratio of the number of positive instances to the number of negative instances is 1/7. Such a skewed data distribution makes the classification problem even harder. The majority of the majority class are instances from N2 stage, which is a stage that is the most easily been confused with deep sleep stage. The severe class-imbalance problem in the dataset can result in lots of errors in the classification

results. Although there are more false negatives than false positives in the errors, the harm of the false positives is bigger than the false negatives. In the introduction of deep sleep, we have mentioned the stimulation should be given in accordance with the progress of deep sleep. Otherwise, the stimulation will only interrupt the sleep process. These false positives make the timing of stimulation inaccurate. Therefore, solving the class-imbalance problem in the dataset is one major task of the early investigation.

We approached the imbalance problem from the data level. One can also think of solving the imbalance problem from the algorithm level. On the algorithm level, the technique is called cost-sensitive learning, which is to introduce a cost matrix to describe the cost of misclassification in order to compensate and/or penalize the misclassification of minority class instances and majority class instances. In our case, the objective of solving the class-imbalance problem is to achieve a balanced data distribution. The reason we approach the class-imbalance problem from the data level is that the classifier used for deep sleep classification in this framework can be considered as solving the class-imbalance problem from the algorithm level. The trained classifier models each class separately, implicitly achieving the goal to ignore the percentage of data for each class in the dataset. We therefore adopted methods from the data level as another way to solve the problem.

On the data level, the objective is to achieve a balanced data distribution in the training set. Thus, the trained classifier will not be biased in favor of the majority class. The re-sampling technique we adopted is down-sampling the instances from the majority class. Up-sampling the instances from the minority class is another way to achieve a balanced data distribution. However, up-sampling results in more instances for training, therefore overload the training system with more calculations. Two down-sampling techniques have been employed, one is random down-sampling the instances from majority class, another is informative down-sampling the instances from majority class. Random down-sampling method is simple. It randomly selects a part of instances from the majority class, to make sure in the training data, the number of instances from the majority class is the same as the number of instances from the minority class. Informative down-sampling method is more complex. The instances from the majority class sampled by it are those instances which are called the representative instances. The representative instances are the majority instances that are difficult to differentiate from the minority instances. The training set, which is comprised of the minority instances and the representative majority instances, makes the trained classifier learn the difference between these hard-to-classify instances.

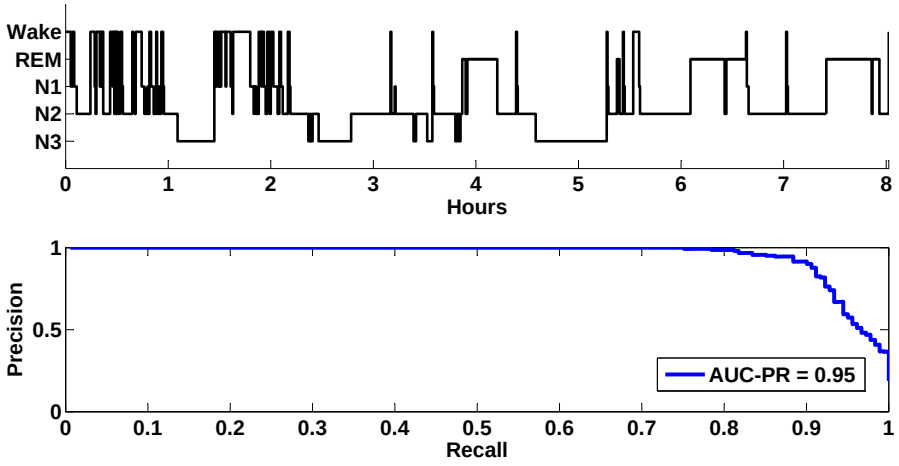
The quality of the re-sampling techniques is evaluated by the performance of the classifier

trained on the obtained training set. The obtained training sets are trained with a k-Nearest Neighbor (kNN) classifier, and then tested on the test set. The choice for the kNN classifier is motivated by its simplicity. kNN classifier suffers greatly if the training data has severe class-imbalance issue. Therefore, it is appropriate to evaluate the quality of the re-sampling techniques. The performance of the kNN classifier is compared with the performance of a linear discriminant (LD) classifier. The LD classifier is trained on the original training set (which is not re-sampled). The impact of skewed data distribution on a LD classifier is small. The goodness of a LD classifier depends on the extent to which the data meets the assumptions of linear discriminant. Although it holds the assumption of shared covariance matrix, in practice, it can achieve good performance as long as both classes approximate Gaussian distribution even for skewed data distribution. Although we proposed a way to deal with the class imbalance problem in the dataset, the simple LD classifier still performed the best on the classification task.

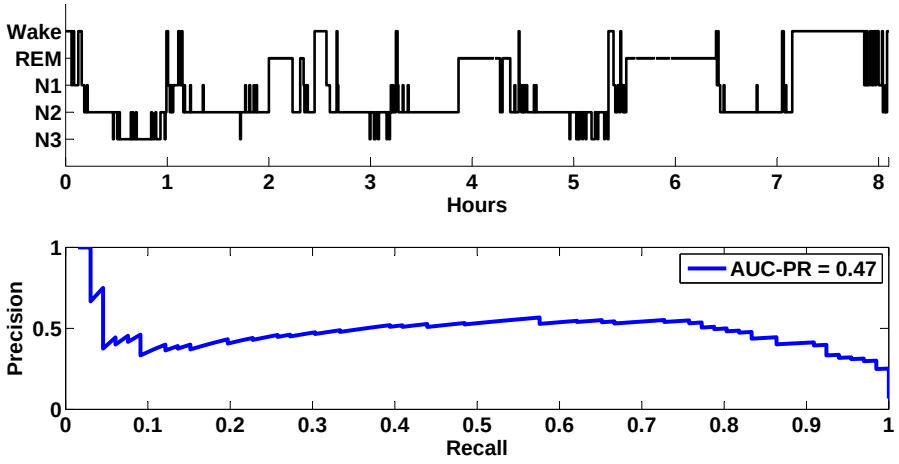
Although the informative re-sampling technique helps to preserve the representative majority instances and make the amount of both classes comparable, the re-sampled data still exhibits highly degree of mixture between classes, which makes the classification problem still difficult. Thus, the conclusions we drawn are it is easy to achieve a balanced data distribution with re-sampling techniques but it is extremely difficult to get a classifier with the re-sampled data who can perform better than a classifier immune to skewed data distribution.

Since resolving the class-imbalance problem is unable to give a better classifier, we switched the work to test the performance of different classifiers on deep sleep classification. We have experimented with linear classifier and non-linear classifier. After comparing their performance, linear discriminant (LD) turned out to outperform other classifier on deep sleep classification. Classifiers which takes context into account might perform better in deep sleep classification since sleep exhibits dependence between stages. For example, some use Hidden Markov Model in sleep staging [41][42][43]. Nevertheless, we restricted our test to classifiers who do not try to model the relationship between instances. Under such circumstances that the best classifier for deep sleep classification is LD.

After examining the classification results, we have noticed some issues. When we put the classification results and the hypnogram of the test subject together, we found there might be a connection between the classification result and characteristics of the deep sleep. Fig.2.8 illustrates hypnograms of two subjects and corresponding classification results. By looking at Fig.2.8(a) and Fig.2.8(b), we have a rough impression that the classification performance might be related to the continuity of deep sleep. The performance of subject *a* is better and his/her deep sleep has seldom been interrupted by other



(a) Hypnogram and classification result of subject *a*.



(b) Hypnogram and classification result of subject *b*.

FIGURE 2.8: Examples of two hypnograms and corresponding classification results. N3 is the deep sleep stage. Classification result is given by the *Precision-Recall* curve and area under the *Precision-Recall* curve (AUC-PR).

stages. The performance of subject *b* is worse and his/her sleep has been interrupted by other stages frequently. Fig.2.9 shows a hypnogram and four features of a subject. Both of these features are with strong discriminative power for deep sleep classification for the subject. The deep sleep in the first half of the night is more continuous than those in the second half of the night. And in the feature plots, the difference between positive epochs and negative epochs in the first half of the night is obvious while it is difficult to differentiate between positive epochs and negative epochs in the second half of the night. It seems the feature expression is affected by the continuity of the

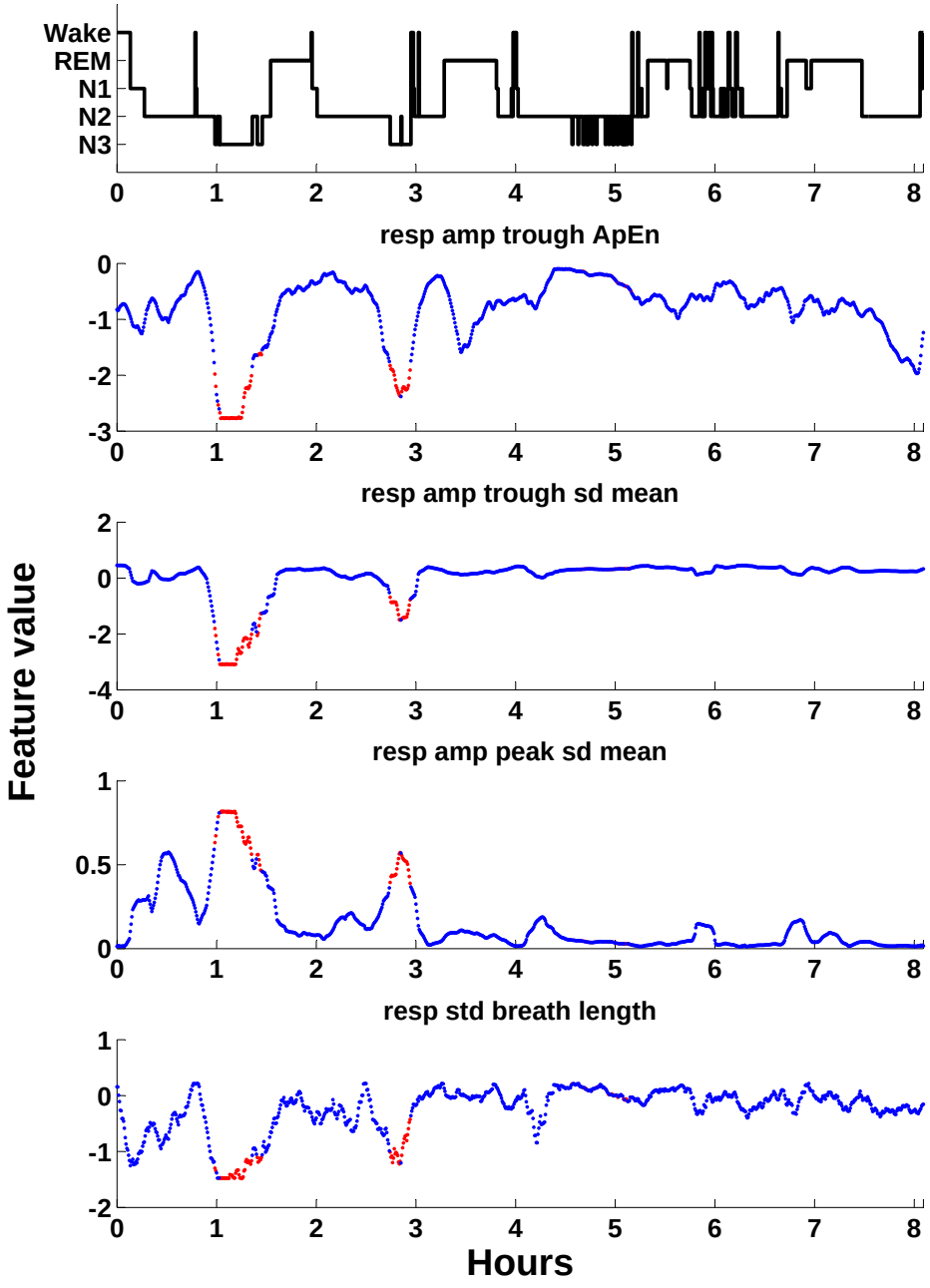


FIGURE 2.9: An example of hypnogram and features. Feature names are indicated in the plot. resp_amp_trough_ApEn computes the approximate entropy for troughs in the amplitudes of the respiratory signals. resp_amp_trough/peak_sd_mean computes the standardized mean trough/peak value in the amplitudes of the respiratory signals. resp_std_breath_length computes the standardized breath lengths in the respiratory signals. Red points represent positive epochs while blue points represent negative epochs.

deep sleep interval. This property that may have impact on the performance of deep sleep classification is not encoded in the features but is the characteristic of the deep sleep interval itself. There might be other factors which may also affect the deep sleep classification, such as the age of the subject and the duration of deep sleep intervals. Therefore, we suspect that the performance of deep sleep classification is affected by the characteristics of deep sleep and the subjects, which is the first hypothesis that needs to be validated.

Another issue is the limitation of the subject-independent model. Though the performance of the subject-independent model is good in general, its performance on specific subjects can be low. Instead of blaming the classifier, we turned the attention to the features. It is observed that the selected features, which have strong discriminative power over all, behave badly for specific subjects. Fig.2.10 illustrates the distribution of the discriminative power (*ASMD* scores) of the selected features for each subject. It is clear that the range of scores varies significantly between subjects. The subject-independent model, which is not tailored to specific subjects, is unable to discover what is the real good features for them (By real good features, we mean features with strong discriminative power.). In other words, the subject-independent property of the constructed model limits its ability on accounting for the personal characteristics in features. Therefore, the second hypothesis is formed, which is for a specific subject, a personalized model is able to provide better performance on deep sleep classification than a subject-independent model.

So far, the two hypotheses, which are motivated by the work done before the master project, are stated and will be validated in the following chapters. The validation of the two hypotheses will provide new directions for the research work on improving the performance of deep sleep classification using features derived from cardiorespiratory signals.

Hypothesis I:

The performance of deep sleep classification is affected by sleep characteristics which are not captured in features.

Hypothesis II:

Personalized classifier can help improve the performance of deep sleep classification.

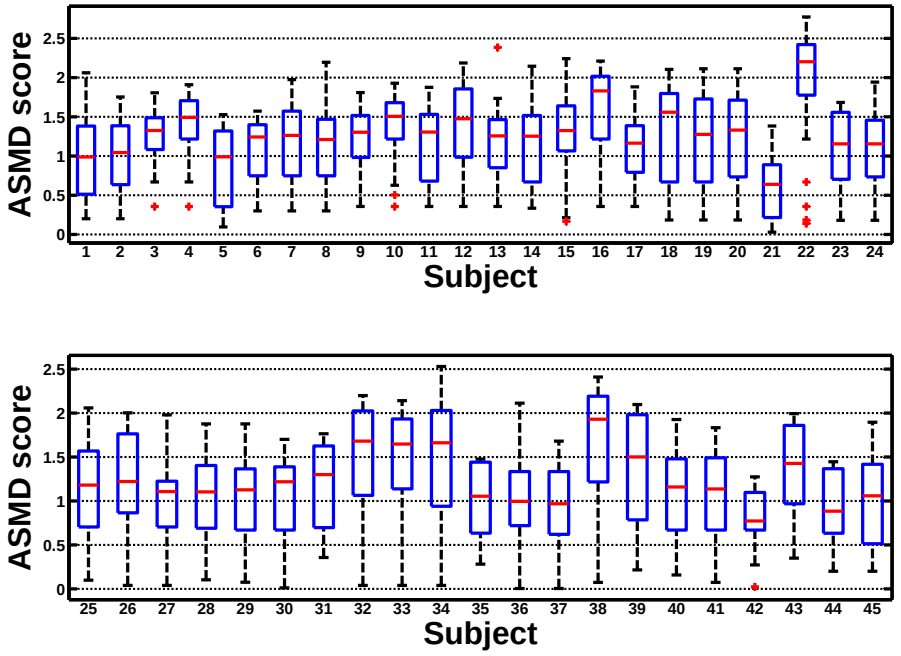


FIGURE 2.10: Distribution of the discriminative power (*ASMD* scores) of the selected features for each subject.

Chapter 3

Methods

In Chapter 2, two hypotheses have been presented, which will be validated in this research. These two hypotheses mention two aspects that may influence the performance of deep sleep classification. The first hypothesis states the relationships between classification performance and one sleep characteristic or a series of sleep characteristics which are not encoded in features. The second hypothesis focuses on the importance of personalization of deep sleep classifier. In this chapter, methods and steps for hypotheses validation are explained. The dataset used in the research is also introduced in this chapter.

3.1 Experimental setup and Dataset

Experimental Setup

The dataset used in this research is from the sleep database created during the EU Siesta project (1997-2000) [44]. This project was organized as a multi-center study, which comprised 8 clinical partners and 8 engineering groups located in Europe. In current study, the dataset is comprised of recordings from 7 separate labs. Each subject spent two consecutive nights in the same sleep lab using the same sensors. Different sensors were used in different labs, therefore variation in the recordings between subjects caused by varying sensors readings are to be expected.

For each subject, lead II ECG configuration was used on the chest to acquire ECG signals. The respiratory effort signal is obtained by measuring the chest circumference using respiratory inductance plethysmography.

Dataset

	$Mean \pm std$	Range
Age	43.6 ± 17.1	20 - 86
Time in bed (hours)	7.9 ± 0.4	6.7 - 9.3
Sleep efficiency (%)	84.7 ± 6.5	73.9 - 97.5
Deep sleep (%)	14.2 ± 4.4	6.7 - 24.8
Wake (%)	15.4 ± 6.5	2.5 - 26.1
REM (%)	16.8 ± 3.5	10.8 - 26.3
30 female subjects and 15 male subjects.		

TABLE 3.1: Population statistics among the 45 subjects

The dataset includes recordings of 45 subjects two-night sleep. Subjects are selected for their healthy sleep patterns and high data coverage. Subjects those do not exhibit a healthy sleep pattern are excluded from the dataset. Healthy subjects whose recordings have missing values of more than 50 continuous epochs are excluded. None of the 45 subjects have known sleep disorders. For the purpose of hypothesis validation, original recordings are segmented into sleep cycles. The first-night recordings after cycle segmentation of the 45 subjects are used in the validation of the first hypothesis. The validation of the second hypothesis utilizes both nights recordings of the 45 subjects. 13 subjects are selected from the 45 subjects, whose both nights recordings are segmented into cycles for the validation of the second hypothesis as well. The selection criteria for the 13 subjects are explained later in the chapter. The validation of the two hypotheses will be the guidance for the future work on improving the performance of deep sleep classification using features derived from cardiorespiratory signals. An overview of the demographics and sleep statistics of the first-night recordings of the 45 subjects is given in Table.3.1.

3.2 Hypothesis 1 validation: Regression analysis

The first hypothesis that needs to be validated is: performance of deep sleep classification is affected by sleep characteristics which are not captured in features. These sleep characteristics are called $influencial\ factors$. Regression analysis is chosen to validate the first hypothesis since it is a statistical approach in order to investigate the relationship among a set of variables. It is used to model and analyze the relationship between response variable (Y, also called dependent variable) and explanatory variables (X_is, also called independent variables), which helps to understand how the response variable changes when any of the explanatory variables varies. In our case, the response variable is the classification performance and the explanatory variables are $influencial\ factors$. A more detail description of $influencial\ factors$ is given in the next chapter.

Linear regression is the easiest category of regression analysis [5]. This is because the assumed linearity among variables makes the model easy to fit and the statistical properties of the estimated parameters are easy to interpret. The constructed model is a linear combination of explanatory variables in order to predict the outcome of response variable. Linear regression could also be used to quantify the strength of the relationship between Y and X_is. It reveals which X_i has a stronger impact on Y and which X_i has little impact on Y. The model could have one explanatory variable, the case which is called simple linear regression, or multiple explanatory variables, the case which is called multiple linear regression. In later discussion, we are focusing on multiple linear regression.

To determine whether it is appropriate to use linear regression modeling the data, it is essential to plot the data first. Scatter plots are recommended to determine the relationship among variables. If the variables appear to be linearly related, a linear regression model can be used then. If the variables are not linearly related, one can tackle the problem by transforming the data or adopting a non-linear model to establish the relationship among variables.

3.2.1 Equation and parameter estimation

Linear regression assumes the relationship between Y and X_is is linear. Given a data set $\{Y, x_1, x_2,...,x_{n-1}, x_n\}$, Y is defined as response variable or dependent variable while x_is are defined as explanatory variables or independent variables. In linear regression, the simplest relationship between Y and X_i can be modelled as:

$$Y = \sum_{1 \leq i \leq n} \beta_i x_i + b + \varepsilon,$$

where β_i is the linear coefficient corresponding to x_i, and ε is the error term, which denotes for the unobserved random variable which adds noise to the model. Such model poses a big advantage in terms of interpretability, however can be over simplistic in some cases. In this thesis, a more flexible linear regression model that allows between-variable interaction is adopted:

$$Y = \sum_{1 \leq i \leq n} \beta_i x_i + \sum_{1 \leq i,j \leq n} \gamma_{i,j} x_i x_j + \varepsilon,$$

where $\gamma_{i,j}$ models the influence of the *interaction term* $x_i x_j$ on response variable. The interaction term $x_i x_j$ assumes that the independent variable $x_i(x_j)$ has a effect on Y depending on the values of the independent variable $x_j(x_i)$. This model, called the two-way interaction linear regression, provides a more flexible way to model sophisticated

scenarios. The two-way interaction model can be further extended to a more generic and flexible form, allowing more interactions among variables, e.g., three-way linear regression model. As interpretability comes as an expense of flexibility, we restrict the discussion within the scope of two-way linear regression.

In order to estimate coefficients of independent variables, ordinary least squares (OLS) is employed. This method minimizes the sum of squared vertical distances between the response variables in the dataset and the response variables predicted by the model. Assume the response variable in the dataset is Y and the response variable predicted by the model is Y_{fit}, OLS attempts to estimate unknown parameter by minimizing $\Sigma(Y - Y_{fit})^2$. $(Y - Y_{fit})$ is also called residual, thus OLS attempts to estimate unknown parameters in the linear regression model by minimizing the sum of squared residuals.

Assumptions

In order to validate the use of linear regression model, a series of assumptions should be met by the variables. If the assumptions are violated, then the results are not trustworthy. According to [45], the assumptions are:

- The measurement of dependent variable and independent variables is error free.
- Dependent variable and independent variables are linearly related.
- The independent variables should be independent of each other.
- Dependent variable should be normally distributed given each value of the independent variables.
- The variance of the residuals is constant for all values of the independent variables, dependent variables and fitted values of the dependent variables (homoscedasticity).
- The residuals are independent of each other.
- The residuals should be normally distributed.

Among these assumptions, the first four assumptions should be met by the data and the last three assumptions should be met by the residuals. The assumptions can be summarized as: independence, linearity, normality and homoscedasticity. To validate a linear regression model, the residuals of the model should be normally and randomly distributed.

3.2.2 Evaluation of a linear regression model

Evaluation of a linear regression model consists of evaluating whether the assumptions of linear regression are violated and the adequacy of the model.

Residual analysis is used to check for the assumptions. To check for normality, a quantile-quantile plot (Q-Q plot) can be used. Q-Q plot is used to compare two probability distributions [46]. If the two distributions being compared are similar, the points in the Q-Q plot should approximately lie on the line $y = x$. In order to check whether the residuals are normally distributed, the y-axis has been fixed with quantiles from a normal distribution with mean 0 and standard deviation 1. The modified Q-Q plot is called normal probability plot. By examining the deviation from the line $y = x$ in the normal probability plot, one can check for the normality of residuals from a linear regression model.

To check for linearity and homoscedasticity, one could produce a scatter plot, plotting residuals against predicted response variable. Ideally, residuals are randomly scattered around 0 and have a relatively even distribution, exhibiting a shape resembling centrally dense cloud. To quantitatively test for the homoscedasticity, one can use a Breusch-Pagan test [47]. It is a test for heteroscedasticity in a linear regression model. It can detect whether the heteroscedasticity in a linear regression model comes from the linear dependence between the residuals and the independent variables in the model. Breusch-Pagan test employs OLS to construct a regression model for the squared residuals with a linear combinations of the independent variables. The hypothesis of homoscedasticity can be rejected when an F-test suggests that the regression model is statistically significant.

Some statistical properties are derived to assess the goodness-of-fit of a linear regression model [5]. Typically, people use R-square and/or adjusted R-square, root mean squared error (RMSE) and marginal sums of squares (Type III SS) to test whether the fitted model is good. R-square measures how many variances in the dependent variable has been explained by the fitted model. Values of R-square are in the range 0 to 1. R-square = 1 indicates that the fitted model accounts for all variances in the response variable. Thus, R-square is desired to be as high as possible. However, a high R-square does not necessarily indicate that the model has a good fit. R-square alone does not tell the entire story. In order to have a comprehensive view of the constructed model, R-square values as well as residual plots and other model statistics should be evaluated. When the model has many explanatory variables, one may use adjusted R-square instead. Because R-square will always increase with the increase of the number of explanatory

variables. Adjusted R-square takes into account the number of explanatory variables when calculating. Thus adjusted R-square is always lower than R-square.

Root mean squared error (RMSE) measures the difference between estimated values of response variables and true values of response variables. RMSE is desired to be small, meaning the predicted values of response variable are close to the true value of response variable.

Type III SS measures the sum of squares in the error which can be obtained for each explanatory variable if it was the last variable entering the model. The effect of each variable is evaluated after all other variables have been considered. The goal of this metric is to test the significance of each independent variable in the fitted model. A variable is considered statistically significant when it has a Type III SS p-value of 0.05 or less.

3.3 Hypothesis 2 validation

The second hypothesis which will be validated in the research is: personalized classifier can improve the performance of deep sleep classification. The experiment procedures for validating the second hypothesis are explained in this part. A thorough description of the methods for feature selection and classifier is also given in this part.

3.3.1 Classification

As mentioned in Chapter 2, classification includes feature selection and classifier training. The feature selection method employed in the framework is explained as well as the classifier trained for deep sleep classification.

Feature selection

The goal of feature selection is to select a feature subset from original feature space in order to build a classification model afterwards. The objective of feature selection is to reduce the amount of redundant features and irrelevant features in the feature space. Having many redundant features will cause overfitting problem for the constructed model. The constructed model will have difficulty in generalizing to new, unseen data if it has been trained with too many equal examples of the same thing, while irrelevant features will not contribute to the constructed model. Therefore, feature selection is expected to tackle these two problems. Moreover, feature selection helps to identify key features for solving a problem, making the constructed model more easily interpretable.

Since feature selection reduce the amount of features used in model construction, both the training time and test time will be reduced.

Feature selection method could be sorted into two broad types: filter methods and wrapper methods. Both methods attempt to find one feature subset which contributes most according to some evaluation metrics. Filter methods measure the "usefulness" of features, such as the mutual information between features, the correlation between features etc. Therefore, feature subset selected by filter methods is not tuned to a specific classification model. Wrapper methods use a classification model to score feature subsets based on their performance on a given test set. Since wrapper methods consider the interaction between the classification model and the training set, it is thought to achieve the best possible performance with a particular learning algorithm on a particular training set. However, wrapper methods could not guarantee the selected subset is the best in terms of generalization since overfitting problem might occur in some cases [48][49].

Since for each candidate feature subset, wrapper methods need a model to be trained, it is a computational intensive task. Moreover, we are interested in not only the performance of selected feature subset on the given dataset with a particular classification model but also the quality of features. Hence we adopt filter methods for their simpler computational methods and better generalization ability [50].

Correlation Feature Selection

Correlation Feature Selection (CFS) [51] is a supervised feature selection method which makes use of the ground-truths of training examples in the selection process. The hypothesis adopted by CFS is that good feature subsets should contain features that are highly correlated with ground-truths while the correlations between features are weak. The merit of each feature subset is calculated as:

$$Merit_S = \frac{k \cdot corr(c, S)}{\sqrt{k + k(k-1)corr(S, S)}},$$

where c is the ground-truths of the features in set S and k is the number of features in S. $corr(c, S)$ is the correlation between the features in set S and c. $corr(S, S)$ is the average pairwise correlation between features in S. The correlation is defined by the Pearson's linear correlation coefficient. Different search algorithms could be used, such as forward search or backward search, to find the feature subset with the highest merit.

$$CFS = \arg\max_S Merit_S.$$

Classifier

Classifier assigns an instance to a category (class) according to its characteristics. Characteristics will be represented by a series of *features*. The objective of classifier is to find a mapping between *features* and class labels. A classifier can handle binary class classification problem as well as multi-class classification problem. The classification problem in this research is a binary class problem which decides whether an instance belongs to deep sleep or non-deep sleep. Many classifiers have been investigated and achieve varying degrees of success in sleep stage analysis [52][53]. Some used simple classifiers, such as linear discriminant classifier, while some used more complex classifiers, such as neural network. Simple classifiers are easier to interpret and implement but they might not be able to learn all the useful information from the data. Complex classifier may well model the data but when the data is limited it will probably cause overfitting problem [54][55]. In this research, linear discriminant classifier is utilized.

Linear Discriminant

Linear Discriminant classifier (LD) is a simple and robust classifier. It was first introduced by Fisher [56]. LD classifier attempts to construct a function which is a weighted combination of features. Hence the decision boundary drawn by LD is a hyperplane in a high-dimensional feature space. LD assumes that each instance is generated from a multi-variate Gaussian distribution and the covariance matrix of each class is the same. Though in many real-life cases these two assumptions are violated, it also performs well in those cases.

The aim of LD classifier is to find a hyperplane that maximally separates two classes. The output of LD classifier is a numerical value which could be interpreted as the difference in standardized distances between the two classes. An instance will be assigned to the class that is closest to it. LD classifier uses probability to measure the distance between an instance and classes.

Given two classes p and n, indicating *positive* and *negative* class respectively, and an instance x with unknown label, the task is to find the most probable class that this instance belongs to. This can be achieved by comparing the posterior probability of $P(p|x)$ and $P(n|x)$:

$$y = P(p|x) - P(n|x).$$

If y is greater than T, then the instance x will be assigned to *positive* class, otherwise it will be assigned to *negative* class. T then is a threshold for the scores given by LD classifier that can be manually set, which makes the classifier more flexible in order to achieve a better performance towards the priority class (in our case, deep sleep is the priority class).

Directly computing y is difficult since it needs to model posterior probability. But with Bayes theorem, posterior probability can be expressed by prior probability and likelihood:

$$P(p|x) = \frac{P(x|p)P(p)}{P(x)}, P(n|x) = \frac{P(x|n)P(n)}{P(x)}.$$

y is therefore expressed as:

$$y = \frac{P(x|p)P(p) - P(x|n)P(n)}{P(x)}.$$

Prior probability can be easily computed as the number of times instances belonging to one class against the number of all instances. With LD classifier's two assumptions, likelihood can also be computed. We denote the Gaussian distribution of the positive class be $f_p \sim N(\mu_p, \Sigma)$, whose mean is μ_p and covariance matrix being shared covariance matrix Σ. And the Gaussian distribution of the negative class be $f_n \sim N(\mu_n, \Sigma)$, whose mean is μ_n and covariance matrix being shared covariance matrix Σ. Thus, likelihood of $P(x|p)$ is:

$$P(x|p) = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} exp(-\frac{1}{2}(x-\mu_p)^T\Sigma^{-1}(x-\mu_p)),$$

where k denotes x has k features. $P(x|n)$ can be expressed likewise.

When computing y, both $P(x)$ and the part outside of the exponential of likelihood can be left out because they will not contribute to the final score since they are constant terms over two classes. To simplify computing, we take the natural logarithm of y, then we have:

$$y = logP(x|p) - logP(x|n) + log\frac{P(p)}{P(n)},$$

$$y = -\frac{1}{2}(x-\mu_p)^T\Sigma^{-1}(x-\mu_p) + \frac{1}{2}(x-\mu_n)^T\Sigma^{-1}(x-\mu_n) + log\frac{P(p)}{P(n)}.$$

y is the score function of LD classifier. If we set the threshold to T, then LD classifier draws a decision boundary through the feature space along $y = T$. Note the prior probability could be manually set when the distribution of two classes are imbalanced in order to give more weights to minority class, which achieves the same effect as changing the value of T.

3.3.2 Experiment procedure

To validate the second hypothesis, we have designed several experiments to compare the performance of the personalized classifier and the subject-independent classifier. Because we have recordings of two-night sleep for 45 subject, the first experiment is to use each subject's one-night data for classifier training and the other night's data for

	$Mean \pm std$	Range
Age	39.8 ± 14.8	20 - 71
Time in bed (epochs)	941.6 ± 43.3	799 - 989
Sleep efficiency (%)	89.7 ± 5.2	81.6 - 97.5
Deep sleep (%)	16.5 ± 4.6	7.1 - 28.0
Wake (%)	10.3 ± 5.2	2.5 - 18.4
REM (%)	20.0 ± 4.3	12.2 - 29.2
10 female subjects and 3 male subjects.		

TABLE 3.2: Population statistics among the 13 subjects

classifier evaluation. The second experiment is to compare the performance of classifiers trained with segmented data. In the second experiment, subjects' data are segmented into sleep cycles. The purpose of cycle segmentation is to overcome the lack of training data problem. To ensure the validity of the experiments using cycles, the first step is to make a selection of subjects. This is step is to ensure the sleep characteristics variability between nights is not high within one subject. Sleep characteristics is defined as the percentage of different sleep stages, number of cycles, and duration of sleep. Each percentage of the characteristics of the selected subject between the two nights does not differ by more than 20%. Based on the criteria, only 13 subjects are selected from the 45 subjects. The validation of the hypothesis will prove the feasibility of constructing a personalized classifier which can therefore improve the performance of deep sleep classification. An overview of the demographics and sleep statistics of the 13 subjects is given in Table.3.2. Due to the limited number of selected subjects, the experiment is only a pilot study of personalized classifier on deep sleep classification.

Cycle segmentation

After subjects selection, we can proceed to cycle segmentation. The first cycle is between the sleep onset and the end of the first REM period. The subsequent cycles are defined between the end of REM sleep of the previous cycle and the next end of REM sleep. A REM period briefly interrupted by micro-arousals can be tolerated. If there is an awakening, which lasts for longer than 5 minutes, in the middle of a cycle, then the cycle is interrupted by this event. Each of the 13 selected subjects has $6 - 12$ sleep cycles. Sleep cycles without deep sleep stage are excluded. Thus each subject has $3 - 6$ valid sleep cycles.

3.3.3 Evaluation of hypothesis 2

In this part, the metric to evaluate features and classification performance are explained.

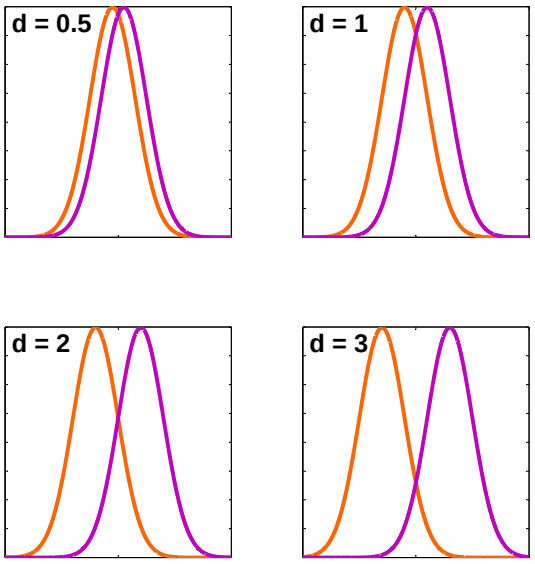


FIGURE 3.1: Gaussian densities illustrating various values of Cohen's d (ASMD)

3.3.3.1 Feature evaluation

Two metrics are adopted to evaluate the effectiveness of a feature: absolute standardized mean difference ($ASMD$) and area under the *Precision-Recall* curve ($AUCPR$).

Absolute Standardized Mean Difference

Absolute standardized mean difference ($ASMD$) is a measurement of the discriminative power of a feature between two classes. This metric is a variation of standardized mean difference, Cohen's d [57]. Cohen's d is a measurement for effect size based on distances between means of different groups. Since the magnitude of Cohen's d is of interest in this research, we use $ASMD$ instead.

For a feature f, its $ASMD$ is calculated as the absolute difference in the mean of two classes divided by the pooled standard deviation for both classes

$$ASMD(f) = \frac{|\mu_p - \mu_n|}{\sigma},$$

here μ_p and μ_n are the means of positive class and negative class in f, and σ is the standard deviation for f over two classes. Since we assume both classes are normally distributed, Figure 3.1 illustrates various values of Cohen's d ($ASMD$). It could be seen from the plot that a feature with a larger $ASMD$ value has a higher discriminative power in separating two classes. It is widely adopted Cohen's guideline to interpret d: small,

d=0.2; medium, d=0.5; and large, d=0.8. Therefore, we decide that if the *ASMD* value of a feature is larger than 1 then the discriminative power of the feature is strong, while the smaller than 0.5 *ASMD* value means a feature with weak discriminative power in separating two classes.

Area under the *Precision-Recall* curve

Precision-Recall curve is a visualization tool to examine classification performance, especially when there is class-imbalanced problem [58]. *Precision-Recall* curve and area under the *Precision-Recall* curve (*AUCPR*) are usually adopted to evaluate the performance of classifier. The construction of *Precision-Recall* curve requires classification to be done first. However, these metrics could also be used for evaluating features. As mentioned before, when a threshold T is set for the score function of a LD classifier, the binary classification result for the test set can be obtained. With varying T, a series of classification results can be obtained. Thus a series of *precision*s and *recall*s can be derived to make the *Precision-Recall* curve. If we want to make a *Precision-Recall* curve for a feature f, we only need to replace the classification score by the feature value of f since the feature value is numerical. Then we could have the same process of constructing *Precision-Recall* curve for a classifier to get the *Precision-Recall* curve for a feature. Details of *Precision-Recall* curve will be explained later in the chapter.

3.3.3.2 Classification evaluation

Classification performance could be used to evaluate a classifier and/or a selected feature subset. The dataset is split into training set which a classifier is built on and test set which is used for the performance evaluation for the trained classifier.

Cross-Validation

Usually we split the dataset into training set and test set for classifier modeling and evaluation. However, this approach is not appropriate in this research since we only have a dataset of 45 subjects. In order to test the classifier on as large as possible test set, while keeping the training set also large and independent of the test set, we consider cross-validation. Cross-validation does not need an independent test set. It partitions the current dataset, on the subject basis, into two disjoint subsets: one is used for training a classifier, whereas the other is used to evaluate the classifier. This procedure is repeated until all subjects in the initial set have been tested.

We applied the Leave-One-Subject-Out cross-validation (LOSOCV) scheme which is a special case of cross-validation where the test set consists of data from only one subject and data of the remaining subjects are used for classifier training. Thus we obtained 45

individually trained classifiers, each one followed by a test run. Results are reported by the averaged performance over the 45 classifiers.

Performance Measure

To evaluate the performance of a classifier and compare the performance between different classifiers, we need to make a choice for performance measurement. Accuracy or error rate is normally used in performance measurement. However, high accuracy does not necessarily imply better performance on target task, particularly there is a class-imbalanced problem. When the data distribution is imbalanced, accuracy will be biased in favour of the majority class. Since the classification problem in this research is a binary case, the explanations of the performance measurement will be restricted to the two-class case.

With a score given by the LD classifier and the threshold T, each test instance will receive a label of either Positive (P) indicating it is deep-sleep or Negative (N) indicating it is non-deep sleep. Given its original label and the classification result, there are four possible outcomes for each test instance:

- True Positive (TP): The original label of the instance is positive and the classifier correctly labels it positive.
- True Negative (TN): The original label of the instance is negative and the classifier correctly labels it negative.
- False Positive (FP): The original label of the instance is negative and the classifier erroneously labels it positive.
- False Negative (FN): The original label of the instance is positive and the classifier erroneously labels it negative.

These possible outcomes are usually shown in what is called a confusion matrix. The structure of a confusion matrix is shown in Table 3.3.

	Predicted P	Predicted N
Actual P	TP	FN
Actual N	FP	TN

TABLE 3.3: Example of confusion matrix

Several statistics can be derived from the confusion matrix:

- *Accuracy*: the percentage of predictions that are correct, $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$.

- *Precision*: also called *positive predictive value*, the number of correctly classified positive instances divided by the number of classified positive instances, $precision = \frac{TP}{TP+FP}$.
- *Recall*: also called *sensitivity* and *true positive rate (TPR)*, the number of correctly classified positive instances divided by the number of total positives, $recall = \frac{TP}{TP+FN}$.
- *False positive rate (FPR)*: the number of erroneously classified negatives divided by the number of total negatives, $FPR = \frac{FP}{TN+FP}$.

ROC Curve

In order to have a more comprehensive view for classification performance, one can use the receiver operating characteristic (ROC) curve. It is possible to compare the performance of different classifiers in ROC space as well. ROC curve is drawn as the TPR against the FPR of a classifier. It describes the trade-off between benefits and loss [59].

There are some important points in ROC space. The point (0,1) in top-left corner represents perfect classification because it achieves maximum benefit with no cost. The line connecting point (0,0) and point (1,1) represents the performance of a random classifier. Any point below the diagonal line represents the performance of a classifier is worse than a random guess. The closer the point to the top-left corner, the best the performance of classifier.

ROC curve could also be used for tuning the threshold. One could use the distance between the point on the curve and the point (0,1) or the area under the ROC curve (AUC-ROC) as the criteria for threshold selection. An example of ROC curve is shown in Figure 3.2.

Precision-Recall Curve

As mentioned earlier, interventions are only useful when subjects are in the deep sleep stage. If interventions are delivered outside of the deep sleep stage, they are no use for sleep [60][61][16]. Therefore, we want to reduce the amount of misclassified negative instances, i.e. false positives.

However, ROC curve is not a reliable tool for evaluation of deep sleep classification. ROC curve is only reliable when there is no class-imbalanced problem. It is unable to reliably capture the variations of false positives. It will overestimate the performance of classifier for tasks with a large skew in the class distribution which is exactly the case in deep sleep classification since there is a severe imbalance between positive class

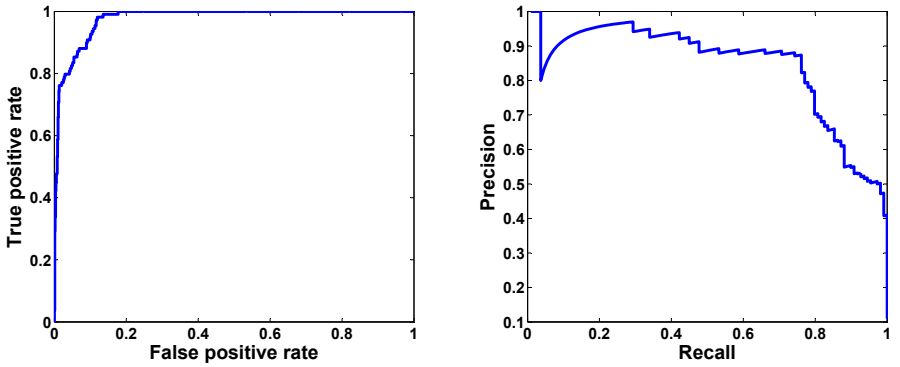


FIGURE 3.2: An example of ROC curve (left) and an example of *Precision-Recall* curve (right).

and negative class. Whereas *Precision-Recall* curve can well capture the change in false positives, which makes it a suitable tool for evaluation in deep sleep classification. ROC curve is *recall* (TPR) against FPR while *Precision-Recall* curve is *precision* against *recall*. When the number of negative instances greatly exceeds the number of positive instances, a large change in the number of false positives only lead to a small change in the FPR used in ROC curve. But *Precision-Recall* curve is able to capture the effect of the large number of negative examples on the performance of classifier.

In PR space, a point near upper-right-hand corner represents better performance. PR curve can also be used for tuning threshold. Two possible performance measure metrics are the distance between the point on the curve and the point (1,1) and the area under the *Precision-Recall* curve ($AUCPR$). An example of *Precision-Recall* curve is shown in Figure 3.2.

Chapter 4

Results and Discussion

4.1 Validating hypothesis 1

Although there are a series of elaborate extracted features for deep sleep classification, the classification performance is limited by other factors which are not captured by these features. In the initial results section, there are already some evidences illustrating the influence of deep sleep continuity on performance. In this section, a more thorough study is conducted to establish the quantitative relationship between classification performance and $influencial\ factors$. Multiple regression analysis is employed to model the relationship. The theory associated with multiple linear regression is well-understood so that the model constructed is easily-interpretable.

4.1.1 Variables

As mentioned in the previous chapter, sleep characteristics which are not encoded in features are defined as $influencial\ factors$. The $influencial\ factors$ in the regression analysis are: age, fragmentation index, lengths, number of epochs and cycle index.

In order to have sufficient data for the regression analysis, whole-night recordings of 45 subjects are segmented into sleep cycles. Instead of having 45 groups of variables, there are 135 groups of variables for the analysis after cycle segmentation. The cycle segmentation procedure is the same as those explained in Chapter 3.

- Age

  Age has an impact on sleep. The sleep process changes with age. As we age, the duration of sleep declines. And elderly people are more vulnerable to sleep

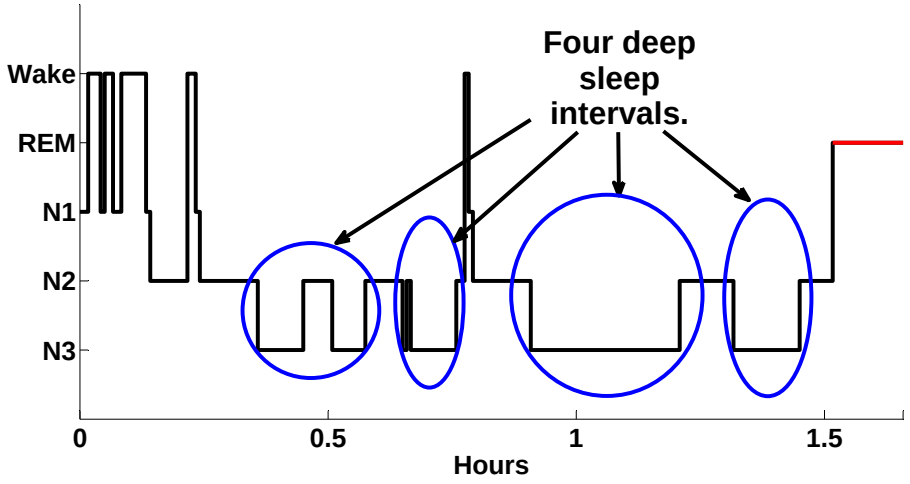


FIGURE 4.1: An example of one sleep cycle.

problems, such as insomnia and disrupted sleep. Even healthy elderly people who does not suffer from sleep problems, their deep sleep (slow-wave sleep) are reduced [17]. The cardiac and respiratory dynamics in healthy subjects across sleep stages are also affect by age [18]. Therefore, it can be expected that the classification performance for older subjects is worse than that of younger subjects.

Cycles within one subject receive the same value for the *age* variable. The values of variable *age* are normalized to 0-1 range. The normalized value of age_i is calculated as:

$$Normalized(age_i) = \frac{age_i - Age_{min}}{Age_{max} - Age_{min}}, \tag{4.1}$$

where Age_{min} is the youngest age among the 45 subjects and Age_{max} is the eldest age among the 45 subjects.

- Fragmentation index and Lengths

  The criteria evaluating the classification performance is the area under the precision-recall curve ($AUCPR$). Discontinuities in the deep sleep intervals can be characterized by a fragmentation index ($Findex$). Within a sleep cycle, a deep sleep interval is defined if the interval is not disrupted by non-deep sleep stages or the interval is disrupted but each disruption is no longer than 5 minutes (10 epochs). Therefore, there are four deep sleep intervals in Fig.4.1. Each deep sleep interval is assigned an f_value. f_value is called fragmentation value and is used to quantify the continuity of one deep sleep interval. The sleep cycle in Fig.4.1 therefore has

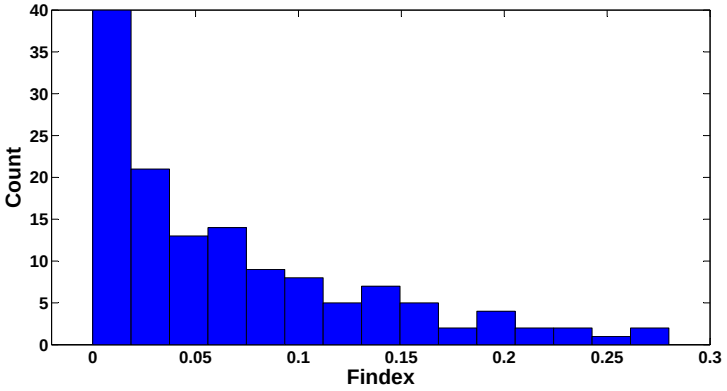


FIGURE 4.2: Distribution of $Findex$.

four f_values. f_value_i is calculated as:

$$f\_value_i = \frac{\#disruptions}{lengths\ of\ the\ i^{th}\ deep\ sleep\ interval\ in\ the\ sleep\ cycle\ (in\ epochs)}. \tag{4.2}$$

The fragmentation index $Findex$ is then the weighted average of the f_values. $Findex$ is computed as:

$$Findex = \sum_{i=1}^{n} w_i f\_value_i, \tag{4.3}$$

where n is the number of deep sleep intervals within one sleep cycle, w_i is the weight of the f_value_i and it is calculated as:

$$w_i = \frac{lengths\ of\ the\ i^{th}\ deep\ sleep\ interval}{total\ lengths\ of\ deep\ sleep\ intervals\ in\ current\ cycle}. \tag{4.4}$$

The calculation of $Findex$ captures the frequency of disruptions by non-deep sleep stages in the deep sleep interval. The value of $Findex$ can be interpreted as the probability of one deep-sleep epoch being followed by a non deep-sleep epoch. Fig.4.2 shows the distribution of $Findex$ in the 135 cycles.

Each cycle has variable $Lengths$. $Lengths$ variable is the weighted average lengths of deep sleep intervals within one sleep cycle. The calculation of $Lengths$ is :

$$Lengths = \sum_{i=1}^{n} w_i lengths_i, \tag{4.5}$$

where n is the number of deep sleep intervals within one sleep cycle, w_i is the weight of the $lengths_i$ and it is calculated using equation (4.4). Then the values of $Lengths$ are normalized to 0-1 range. The distribution of $Lengths$ in the 135 cycles is shown in Fig.4.3.

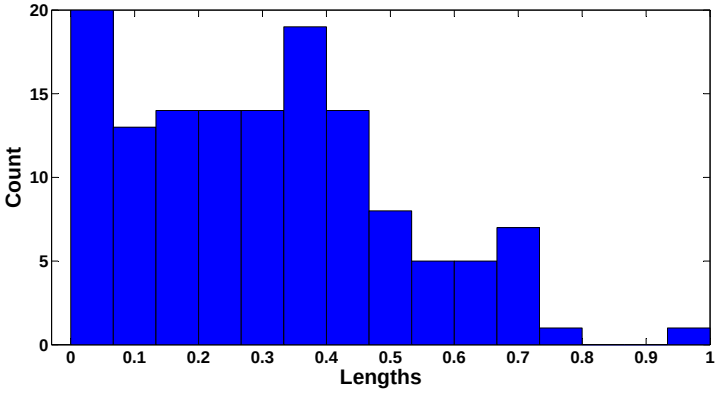


FIGURE 4.3: Distribution of *Lengths*.

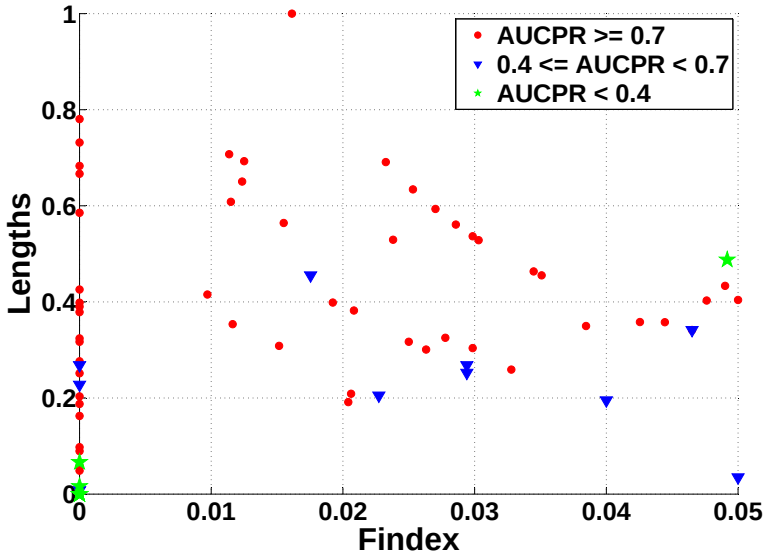
We have already had the impression that there is a link between classification performance, continuity and lengths of the deep sleep intervals from Fig.2.8. We can roughly draw the conclusion that the difference between deep sleep and non-deep sleep described in the features are larger when a deep sleep interval is continuous and longer. If a deep sleep interval is frequently interrupted by other stages, then we can hardly tell the difference between deep sleep and non-deep sleep from the feature expression. We attach another figure (Fig.4.4) here, further demonstrating the relationship between classification performance, lengths and continuity of deep sleep intervals.

Fig.4.4 clearly demonstrates the impact of $Findex$ and $Lengths$ on $AUCPR$. Majority of cycles which have worse classification performance are those cycles whose $Findex$ are below 0.1 and/or $Lengths$ are below 0.2. This observation is in accordance with the conclusion drawn from Fig.2.8. This is not unexpected. From the physiological point of view, it is too late to reflect the characteristics featuring deep sleep on the cardiorespiratory signals if the duration of deep sleep is too short or the transitions between deep sleep and non-deep sleep are constant.

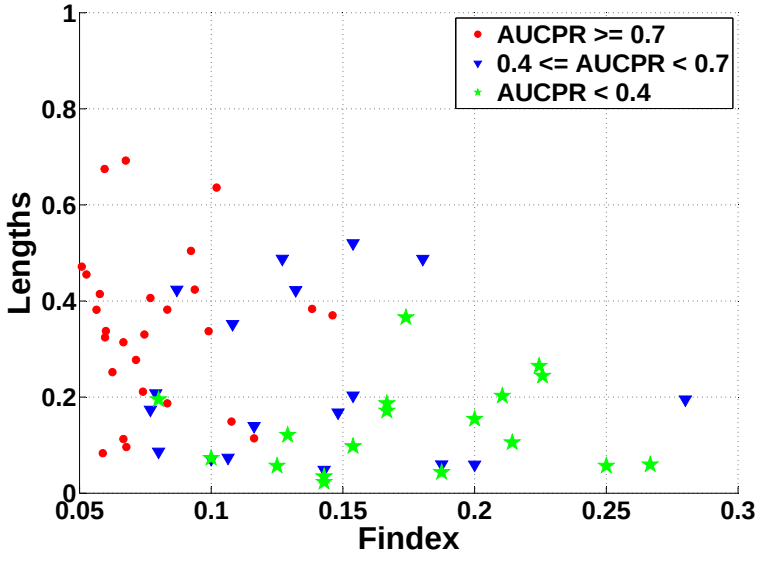
- Number of epochs

  Number of epochs (Nr_epochs) computes the total number of deep sleep epochs within one sleep cycle. Nr_epochs is normalized to 0-1 range. Though Nr_epochs seems a repetition to $Lengths$, Nr_epochs commits to describe the proportion of deep sleep within one sleep cycle. Two cycles can have the same value for $Findex$, but the one with bigger value of Nr_epochs may have better classification performance compared to the other with smaller Nr_epochs.

- Cycle index



(a)



(b)

FIGURE 4.4: Relationships between *Findex*, *Lengths* and *AUCPR*. (a) and (b) differentiate in the range of the *Findex*. *Findex* goes from 0 to 0.05 in (a) while it runs from 0.05 to 0.3 in (b). It can be seen from (a) that the longer and/or the more continuous the deep sleep interval, the better the classification performance. From (b), it can be seen that most misclassifications happen in the interrupted and/or short deep sleep intervals.

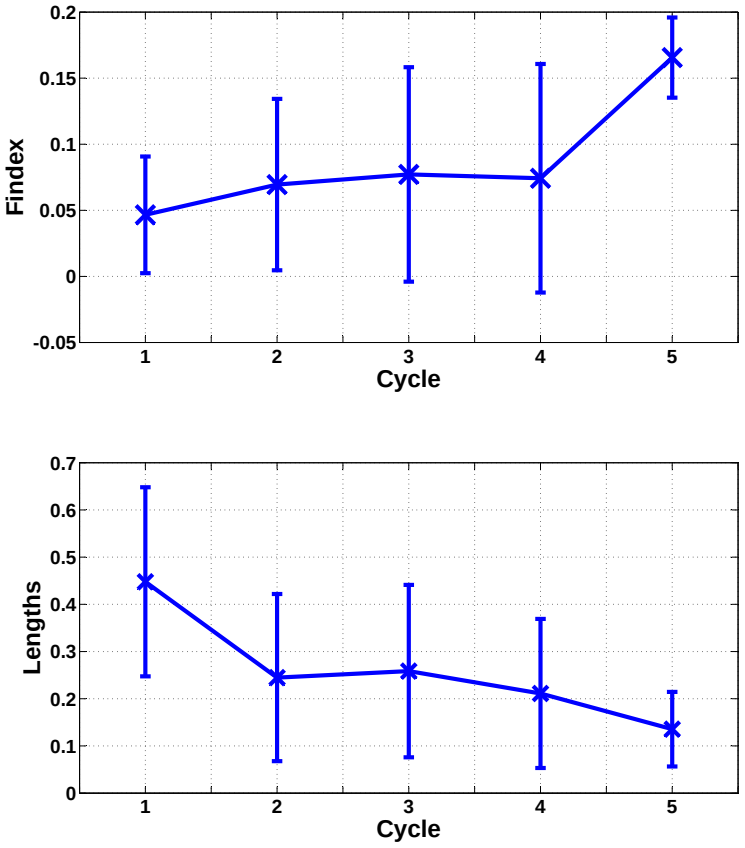


FIGURE 4.5: *Findex* and *Lengths* changing with time.

As mentioned in the sleep physiology, both the duration and intensity of deep sleep decrease as sleep progresses. Thus we can expect the first sleep cycle has a small *Findex* and big *Lengths* compared to the last sleep cycle within one night. Fig.4.5 depicts how *Findex* and *Lengths* evolve throughout the night. Therefore, we consider cycle index (*Cindex*) as one of the *influential factors* to the classification performance. *Cindex* is a categorical variable, with 1 indicating the first sleep cycle, 2 indicating the second sleep cycle and so on.

4.1.2 Regression analysis with one independent variable

Findex, *Lengths* and *Cindex* are employed respectively to predict the classification performance (*AUCPR*). A line is estimated to establish the relationship between *AUCPR* and these independent variables. The three independent variables are chosen because the spearman correlation coefficients for each of them and *AUCPR* are moderate ($|\rho| \geq 0.4$) with narrow 95% confidence intervals. The estimation function resembles the form: $y = ax + b$, a is the slope and b is the intercept. To validate the use of linear regression, residuals analysis is conducted after each fitting.

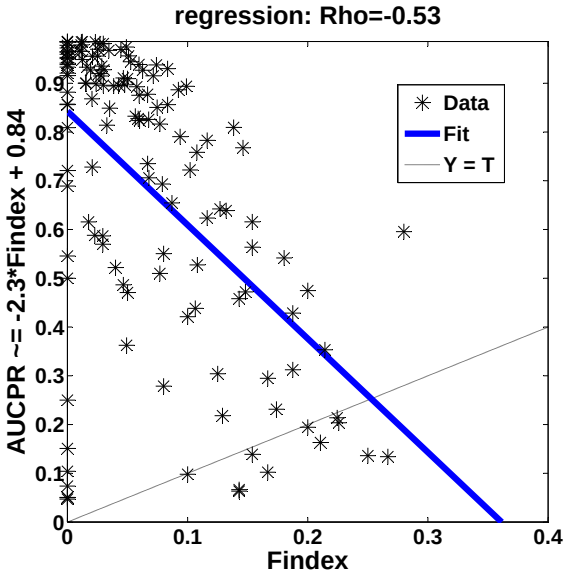


FIGURE 4.6: Model1, $AUCPR \sim Findex$.

	Estimate	tStat	pValue
Intercept	0.843	27.69	< 0.001
$Findex$	-2.332	-7.25	< 0.001
R-squared: 0.283			
F-statistic vs. Constant model: 52.6, p-value < 0.001			

TABLE 4.1: Statistics of the Model1

- Model1: $AUCPR \sim Findex$

  Fig.4.6 illustrates the estimated line of predicting $AUCPR$ by $Findex$. The black stars represent raw data and the blue line represents the estimation. The label of y-axis is the fitted function and the title of the plot is the spearman correlation coefficient of $AUCPR$ and $Findex$. Table 4.1 shows the anova statistics for Model1.

  The p-value of *F-statistic* indicates Model1 is significantly different from a constant model (Constant model uses a line with slope=0 to predict $AUCPR$.) $Findex$ is a significant variable when predicting classification performance (with p-value smaller than 0.001). Nearly 30% of the variance in the dependent variable has been explained by $Findex$. The result of linear regression coheres with the former observation. The classification performance of a cycle is negatively associated with its fragmentation index. The larger the $Findex$, the worse the classification performance, and vice versa.

  Fig. 4.7 shows the result of residual analysis of Model1. Linear regression assumes the residuals are normally distributed and the range of residuals is independent

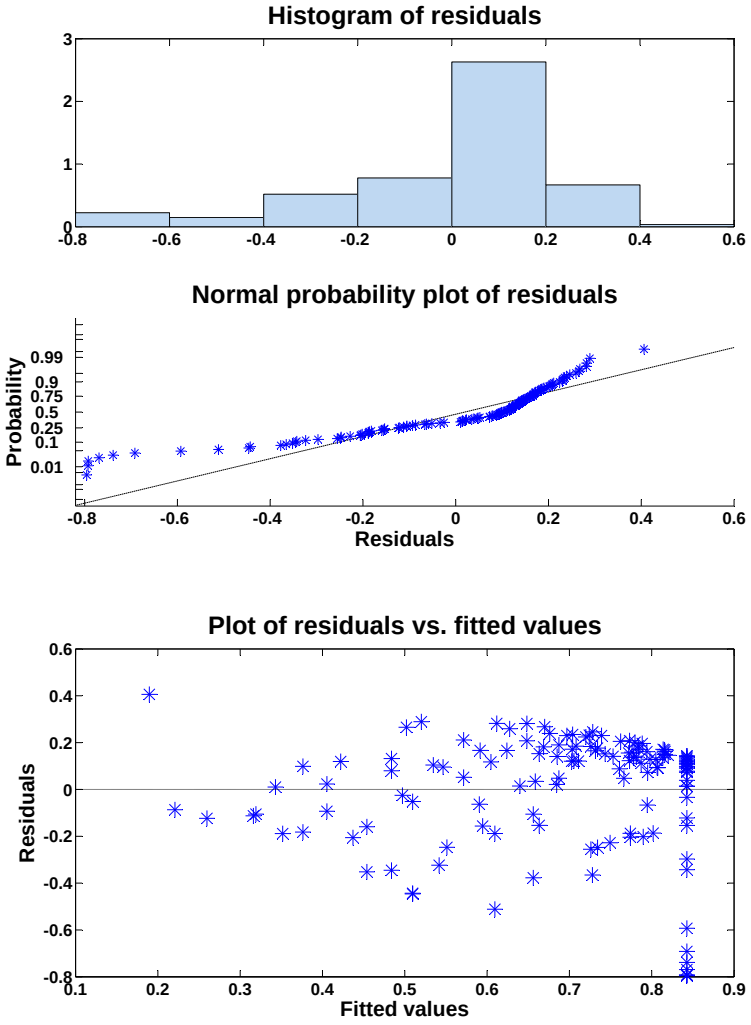


FIGURE 4.7: Results of residual analysis for Model1.

of independent variables or fitted variables (homogeneity of variance). It is plausible that the residuals are normally distributed judging from Fig.4.7. Both the histogram of residuals and the normal probability plot of residuals exhibit some degree of left skewing in the residuals. Judging from the plot of residuals against fitted values, it illustrates a nice 'cloud' shape except for data points on the right. The strange points comes from cycles whose $Findex$ are 0. $Findex = 0$ represents the deep sleep interval which is not interrupted by other stages. Though the general trend is that sleep cycles with smaller $Findex$ have better classification performance, using $Findex$ alone for prediction of classification performance is dangerous Consider the extreme case where there is only one epoch of deep sleep in a cycle. Though the $Findex$ for the cycle is 0, the classification performance is still worse since the number of positive class is too few.

- Model2: $AUCPR \sim Lengths$

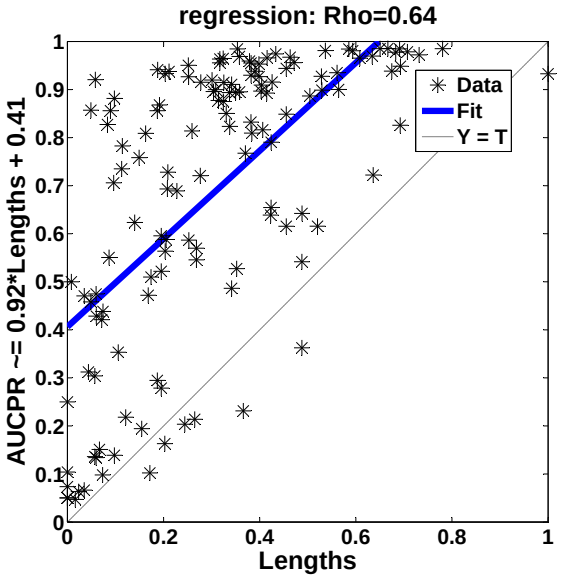


FIGURE 4.8: Model2, $AUCPR \sim Lengths$.

	Estimate	tStat	pValue
Intercept	0.406	11.45	< 0.001
Lengths	0.916	9.54	< 0.001
R-squared: 0.406			
F-statistic vs. Constant model: 91.0, p-value < 0.001			

TABLE 4.2: Statistics of the Model2

The estimated line of predicting $AUCPR$ by *Lengths* is shown in Fig.4.8. The black stars represent raw data and the blue line represents the estimation. The label of y-axis is the fitted function and the title of the plot is the spearman correlation coefficient of $AUCPR$ and $Findex$. Table.4.2 shows the anova statistics for Model2.

The < 0.001 p-values indicate both Model2 and variable *Lengths* are significant for performance prediction. *Lengths* explains more variance in the $AUCPR$ compared to $Findex$. Model2 confirms that the longer the deep sleep interval, the better the classification performance. This finding meets the expectation and makes sense from the physiological point of view. When deep sleep progresses, the characteristics of deep sleep are more remarkable in the signals to differentiate it from non-deep sleep stage, which makes the classification much easier for long deep sleep intervals.

Fig.4.9 shows the result of residual analysis for Model2. There still exists skewness in the residuals but the situation is better than Model1. However, the homoscedasticity is not met by Model2. The residual decreases as fitted value increases. Though the homoscedasticity is violated, unless the heteroscedasticity

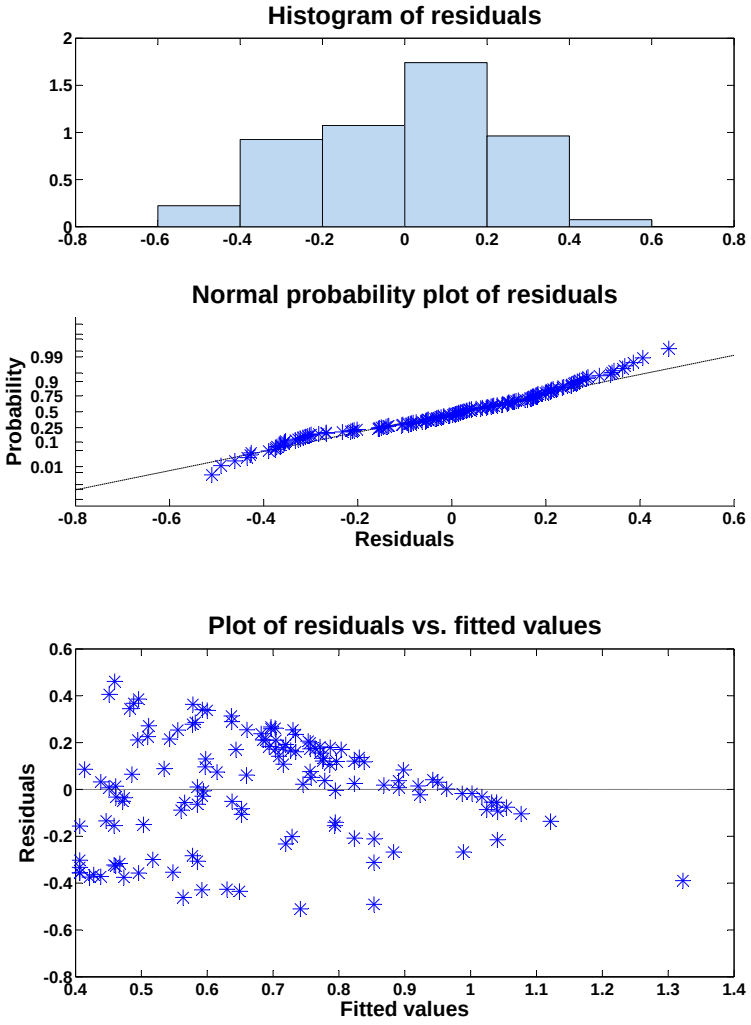


FIGURE 4.9: Results of residual analysis for Model2.

is pronounced, its effect will not be severe, which means the estimation is still unbiased. We therefore still consider Model2 validated.

- Model3: $AUCPR \sim Cindex$

  Fig.4.10 illustrates the estimated line of predicting $AUCPR$ by $Cindex$. The black stars represent raw data and the blue line represents the estimation. The label of y-axis is the fitted function and the title of the plot is the spearman correlation coefficient of $AUCPR$ and $Cindex$. Table.4.3 shows the anova statistics for Model3.

  From the statistics of Model3 we can see cycle index ($Cindex$) is a significant variable for predicting $AUCPR$. However, $Cindex$ is less important than $Lengths$ and $Findex$ since it only explains 15% of the variance in the dependent variable. This can be expected to since $Cindex$ is a categorical variable and itself is insufficient to model classification performance. $Cindex$ can only give a general impression

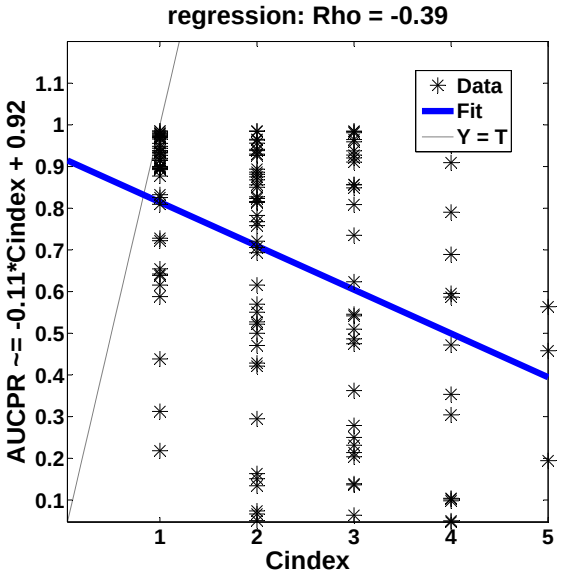


FIGURE 4.10: Model3, $AUCPR \sim Cindex$.

	Estimate	tStat	pValue
Intercept	0.920	17.34	< 0.001
Cindex	-0.105	-4.88	< 0.001
R-squared: 0.152			
F-statistic vs. Constant model: 23.8, p-value < 0.001			

TABLE 4.3: Statistics of the Model3

that the classification performance for earlier cycles are better than later cycles. In spite of this, the importance of *Cindex* cannot be overlooked. Though the direct effect of cycle index on classification performance is small, its interaction with other independent variables, such as *Findex* and *Lengths*, may still contribute to the prediction for *AUCPR*. It is reasonable to consider the impact of interactions between some independent variables on the regression model. There is more deep sleep in the first half of the night. And the less deep sleep appearing in the second half of the night is more vulnerable to disruptions. The interactions between independent variables will be discussed in the following subsection.

The residual analysis of Model3 is shown in Fig.4.11. The residuals illustrate a little left skewing. It is difficult to tell whether homoscedasticity has been violated since the categorical property of *Cindex*. No matter the validation of Model3, it reinforces the notion that time information plays an important role on the classification performance.

Both the correlation coefficients and the regression analysis confirms the significance of *Findex*, *Lengths* and *Cindex*. These variables are then used in multiple linear regression

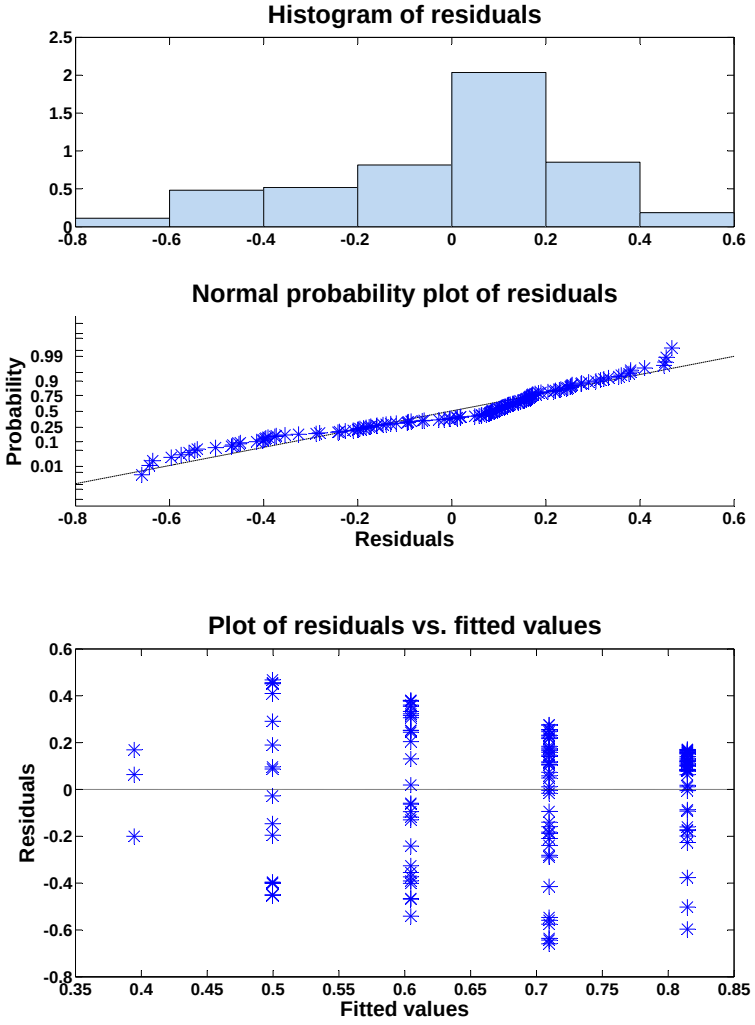


FIGURE 4.11: Results of residual analysis for Model3.

analysis to see their joint force when modeling $AUCPR$.

4.1.3 Regression analysis with more independent variables

In the last subsection, we have analyzed some significant $influencial\ factors$ for $AUCPR$ prediction. By conducting such analysis, we can determine the importance of each independent variable. Yet, none of the significant variables is sufficient to describe the dependent variable. Multiple linear regression allows to employ more than one independent variables to model the dependent variable. The final model is a linear combination of independent variables. The linear combination makes the model easy to interpret. Significant variables are obtained from the model, which can be compared to the former analysis. Moreover, the interactions between variables are allowed in the multiple linear regression. The constructed model is more comprehensive. Some variables, which may

	Estimate	tStat	pValue
Intercept	0.640	10.36	< 0.001
$Findex$	-1.614	-5.96	< 0.001
$Lengths$	0.705	7.46	< 0.001
$Cindex$	-0.028	-1.59	0.11
R-squared: 0.549			
F-statistic vs. Constant model: 53.1, p-value < 0.001			

TABLE 4.4: Statistics of the Model4, version1

	Estimate	tStat	pValue
Intercept	0.932	8.77	< 0.001
$Findex$	-2.658	-3.11	0.002
$Cindex$	-0.157	-3.99	< 0.001
$Lengths \times Cindex$	0.344	3.88	< 0.001
R-squared: 0.597			
F-statistic vs. Constant model: 31.7, p-value < 0.001			

TABLE 4.5: Statistics of the Model4, version2

not be important for the dependent variable by themselves, will have the ability to affect the dependent variable through other independent variables.

- Model4: $AUCPR \sim Findex + Lengths + Cindex$

  We start to make the multiple regression model with these three variables since they have been proved as significant for predicting $AUCPR$. There are two versions for Model4, one without interactions between independent variables in the model (version1), the other with interactions between independent variables in the model(version2). The regression model use multiplicative terms to represent variable interactions. The multiplicative terms are: $Findex \times Lengths$, $Findex \times Cindex$ and $Lengths \times Cindex$. The statistics of Model4, version1 are shown in Table.4.4. The statistics of Model4, version2 are shown in Table.4.5. Only terms which are significant in Model4, version2 are shown in the table.

  From Table.4.4, we find that under the condition that $Findex$ and $Lenghts$ are in the model, $Cindex$ is no longer a significant term. This does not exceed our expectation since numerical variables are more expressive than categorical variable when constructing a model. Nevertheless, the power of $Cindex$ is shown when interaction terms are considered. $Lengths$ itself is dumped by the model. Its impact on $AUCPR$ is replaced by the interaction between $Lengths$ and $Cindex$. According to previous analysis (Fig.4.5), the lengths of deep sleep interval decreases as time progresses, which validates this multiplicative terms in the linear regression model.

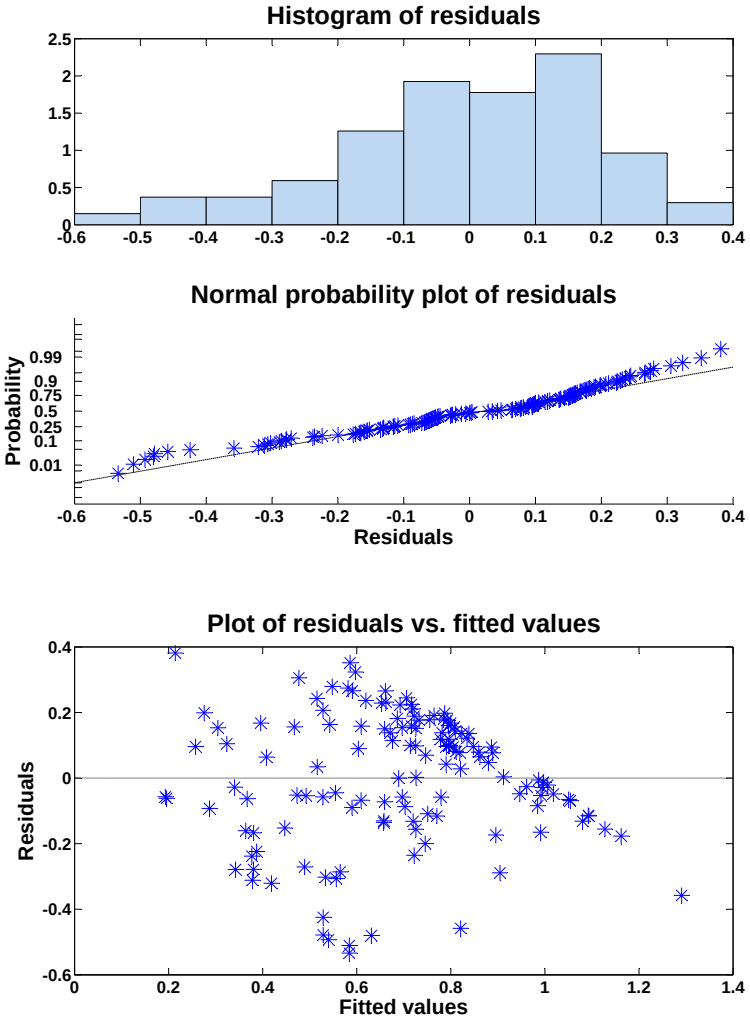


FIGURE 4.12: Results of residual analysis for Model4, version1.

Version2 is better than version1, which is reflected in two places. Version2 is able to explain more variances in $AUCPR$ than version1. And when we look at the residual analysis for both versions (Fig.4.12 and Fig.4.13), the residuals of version1 still exhibit some degree of left skewing while the residuals of version2 basically obeys normal distribution (except for some points on the left, which might be caused by outliers.). Instead of visually examining violation of homoscedasticity by both versions, we employ Breusch-Pagan test. It is a test for heteroscedasticity in a linear regression model. It tests whether the variance of the residuals are dependent on the values of the independent variables. For both versions, Breusch-Pagan test reject the hypothesis that the variance of the residuals are dependent on the values of the independent variables. Therefore, we are safe to say the homoscedasticity is not violated by both versions. And the results and analysis of both versions are validated.

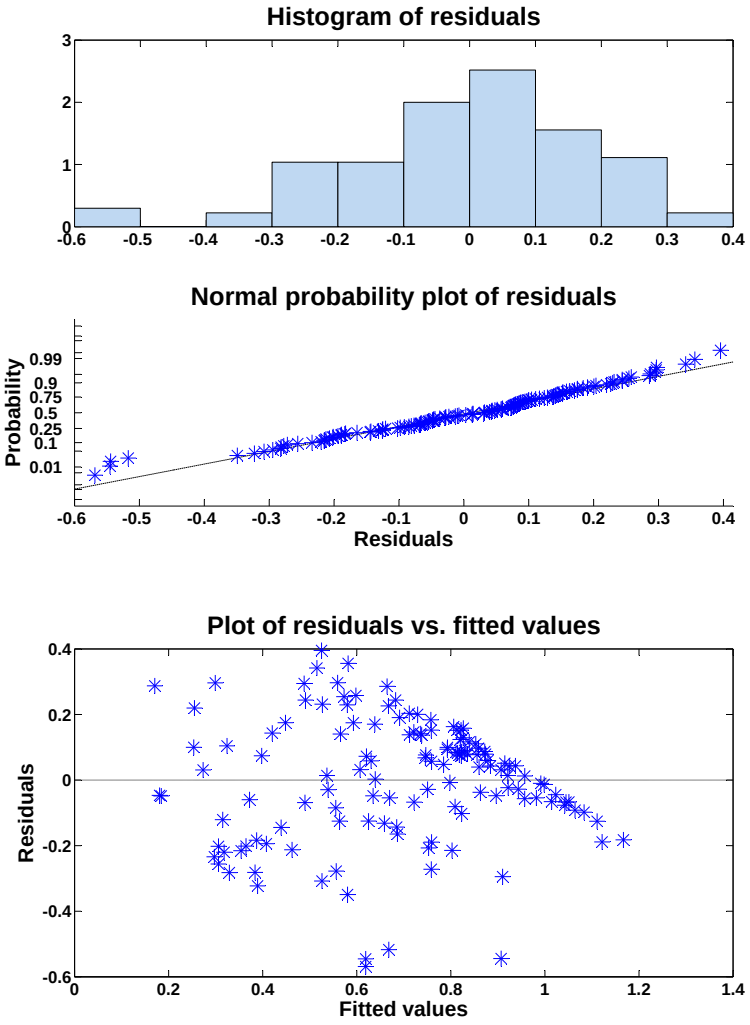


FIGURE 4.13: Results of residual analysis for Model4, version2.

- Model5: $AUCPR \sim Findex + Lengths + Cindex + Age + Nr\_epochs$

  The advantage brought by non-linear term to the regression model has been approved by Model4, version2. Hence, interactions between independent variables are under consideration in Model5. Model5 aims at including two more variables to the regression model: Age and Nr_epochs. Only by looking at the plot of $AUCPR$ against Age, we can hardly see its impact on classification performance. Yet, Age is valuable since it plays a role in the proportion of deep sleep. The proportion of deep sleep decreases as aging. Therefore, it is expected the impact of Age on $AUCPR$ can be reflected by other independent variables, might be $Lengths$ and $Findex$. To better describe the notion of the proportion of deep sleep, Nr_epochs is introduced. Thus, the interaction between Age and Nr_epochs is also of interest.

  Table.4.6 summarizes the statistics of Model5 (only significant terms are included.). Model5 is able to explain 73% variance in $AUCPR$, which means we can use Model5

	Estimate	tStat	pValue
Intercept	0.625	3.77	< 0.001
$Findex$	-2.359	-2.34	0.02
$Lengths$	1.383	3.23	0.002
$Cindex$	-0.124	-2.43	0.02
$Findex \times Cindex$	0.641	2.37	0.02
$Lengths \times Nr\_epochs$	-1.699	-3.83	< 0.001
$Cindex \times Nr\_epochs$	0.557	2.73	0.007
R-squared: 0.727			
F-statistic vs. Constant model: 21.1, p-value < 0.001			

TABLE 4.6: Statistics of the Model5

Age	Averaged $Findex$	Averaged $Lengths$
[20 30)	0.04 ± 0.05	0.35 ± 0.24
[30 50)	0.08 ± 0.07	0.32 ± 0.17
[50 80)	0.08 ± 0.07	0.24 ± 0.19

TABLE 4.7: $Findex$ and $Lengths$ change with Age

to predict $AUCPR$ and it would be reasonably accurate. However, Age does not appear in Model5, even through the interactions with other variables. Because of the sleep pattern criteria, which is used for selecting healthy subjects, the age effect on sleep is not significant for subjects. Table.4.7 shows how $Findex$ and $Lengths$ change with age increasing. Though we can still see the trend that sleep of young age is less likely to be interrupted and the duration of deep sleep is longer for young subjects, the difference between different ages is not significant. That explains why the impact of Age is not reflected in Model5. It is not surprising that $Lengths \times Nr\_epochs$ is a significant term since $Lengths$ can be thought of the weighted average of Nr_epochs, which explains the strong relationship between these two variables. Compared to Model4, version2, $Lengths \times Cindex$ has been replaced by $Nr\_epochs \times Cindex$, which also makes sense from the physiological point of view. Though Model5 contains two highly correlated variables, both variables ($Lengths$ and Nr_epochs) are kept in the model as they make the final model more comprehensive.

The result of residual analysis for Model5 (Fig.4.14) indicate the former analysis is validated. Breusch-Pagan test confirms the homoscedasticity of Model5.

To make the final regression model more easy to interpret, we therefore reduce the number of independent variables in the final model based on Model5. Age variable is ignored in the final model. Since $Lengths$ and Nr_epochs are highly correlated, only one of them will be employed in the final model. Model4, version2 is the one with $Lengths$

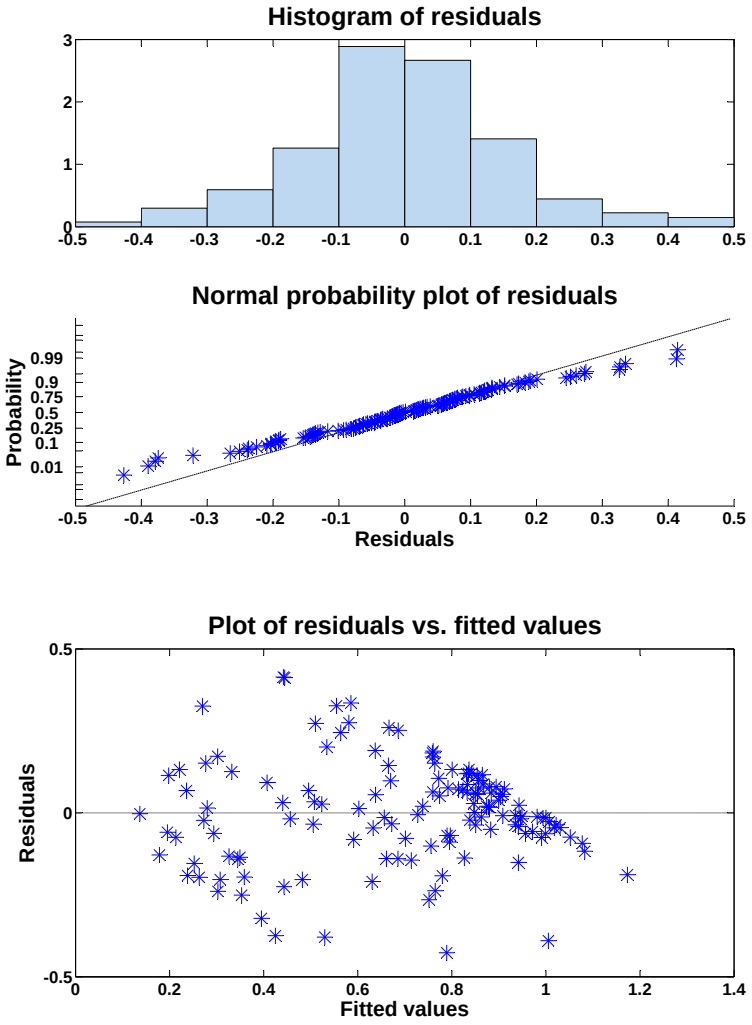


FIGURE 4.14: Results of residual analysis for Model5.

	Estimate	tStat	pValue
Intercept	1.002	9.99	< 0.001
$Findex$	-4.022	-4.64	< 0.001
$Cindex$	-0.197	-5.32	< 0.001
$Findex \times Cindex$	0.973	3.70	< 0.001
$Cindex \times Nr\_epochs$	0.534	6.01	< 0.001
R-squared: 0.653			
F-statistic vs. Constant model: 40.2, p-value < 0.001			

TABLE 4.8: Statistics of the Model6

as independent variable. Model6 is the one with Nr_epochs as independent variable (Table.4.8). Comparing the two models, Model6 is chosen as the final regression model.

With multiple linear regression, we are able to model the classification performance by a linear combination of $influencial\ factors$ and their interactions. These sleep characteristics are not reflected on features. The hypothesis, classification performance is

statistically significant related to sleep characteristics, has been proved. The significant *influential factors* are $Findex$, Nr_epochs and $Cindex$. The established relationship between the classification performance and the significant factors reflects the influence of the duration and continuity of deep sleep on the performance of deep sleep classification. This gives the hint for future works on deep sleep classification. Unless researchers can find a solution for this problem, the performance of deep sleep classification will be limited.

4.2 Validating hypothesis 2

Before moving on to discover new features and experiment with new classification algorithms to improve the performance of deep sleep classification, we would like to investigate how existing framework can be improved to have a better performance. The model constructed by current framework is a subject-independent model, which can guarantee its good performance on a large population. However, if we closely examine the built model, we find this model might not be applicable to specific subjects. The most obvious manifestation is in the selected features. In the framework, the leave-one-subject-out cross-validation (LOSOCV) paradigm is employed. However the selected features do not exhibit strong discriminative powers for some subjects. Therefore, we suspect lack of subject-dependent information in the current framework has prevented it to have a better performance. In this section, we will have a detail discussion.

4.2.1 Upper bound

In order to assess the best performance that can be obtained with the existing framework, which means we do not change the feature selection method and classification algorithm in current framework, the testing data is used for training. Although training a classifier with test data violates the rules of machine learning, this strategy can provide an estimate of the upper bound on the classification performance. Concretely, the testing data will be used in the feature selection stage. The purpose of the violation is to see how the classification performance can benefit from features that are optimal for specific subjects. To define what is a good feature, we use the evaluation metric Absolute Standardized Mean Difference ($ASMD$). It describes the ability of a feature in discriminating deep sleep and non-deep sleep.

The experiment is carried out as follows. In the process of classifier training, we employ LOSOCV paradigm. The data of the test subject is only used in the feature selection stage with correlation feature selection (CFS) method. Then the selected features are

	Initial Result	New Result
Averaged $AUCPR$	0.753	0.845
Standard Deviation	0.160	0.112

TABLE 4.9: Experiment Results

	Initial Result	Subject-dependent Result
Averaged $AUCPR$	0.753	0.634
Standard Deviation	0.160	0.213

TABLE 4.10: Comparison between subject-independent model and subject-dependent model

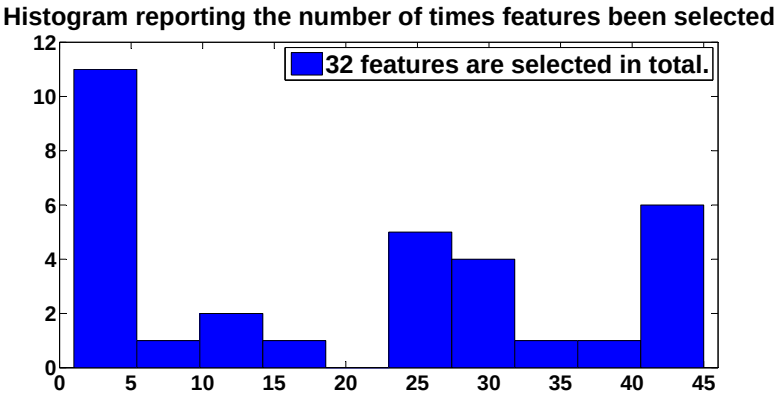
extracted from the training subjects to construct a linear discriminant (LD) classifier. The trained LD is tested on the test subject for evaluation. The evaluation metric for classifier is area under the precision-recall curve ($AUCPR$). Results of the experiment are compared to the initial results. Because the data from test subject is employed in the feature selection, we call it partial LOSOCV paradigm.

Table.4.9 summarizes the results of 45 subjects. With incorporating data of test subjects in the feature selection stage, the averaged performance is improved by nearly 10%. The one-sided Wilcoxon rank sum test indicates the improvement is statistically significant. We believe this is the upper bound of the performance with existing framework.

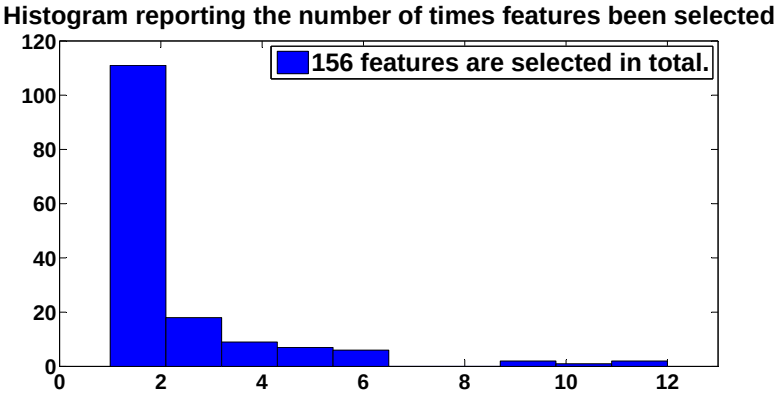
We also look at the divergence in the selected features between subjects. The lists of selected features are aggregated from 45 subjects. Fig.4.15 demonstrates the histograms of the selection frequency of the selected features. Much more features are selected in the partial LOSOCV paradigm while features selected in the LOSOCV paradigm exhibit low divergence. The LOSOCV paradigm is unable to capture some subject-dependent information which can help improve the classification performance for a specific subject. Fig.4.15 implies lack of subject-dependent information in current framework. In the following analysis, we are going to test whether subject-dependent information can help improve the classification performance in a valid machine learning procedure.

4.2.2 Subject-dependent model based on whole-night data

Since we have recordings of two-night sleep for 45 subjects, we are able to test the usefulness of subject-dependent information with these data. For each subject, one night data is used for feature selection and classifier construction. The obtained classifier is tested on the data of the other night. The results are compared to the initial results. The comparison is shown in Table.4.10. And the training performance and test performance of the subject-dependent models are shown in Fig.4.16.



(a) LOSOCV paradigm



(b) Partial LOSOCV paradigm

FIGURE 4.15: Histogram reporting the number of times a feature was selected.

Though the performance of the subject-dependent model is worse than the performance of the subject-independent model, we cannot make the conclusion that incorporating subject-dependent information in the framework does not help improving classification performance. In the subject-dependent scenario, there is only one-night data can be used in the training session. The performance of the subject-dependent model has suffered from lack of training data. With such little training data, the classifier is not able to learn sufficient knowledge to achieve the optimal performance. It is obvious in Fig.4.16 that the test performance and training performance have not converged in most cases. The learning process of the subject-dependent model is not completed. The model has not entered into the generalization stage so that its performance on unseen data is bad.

To continue the experiments on personalization, in the training session, one can still employ the LOSOCV scheme to avoid lack of training data problem, and integrating partial data from test subject into the training data. However, it can be expected the classification performance might not be significantly different from before. This is

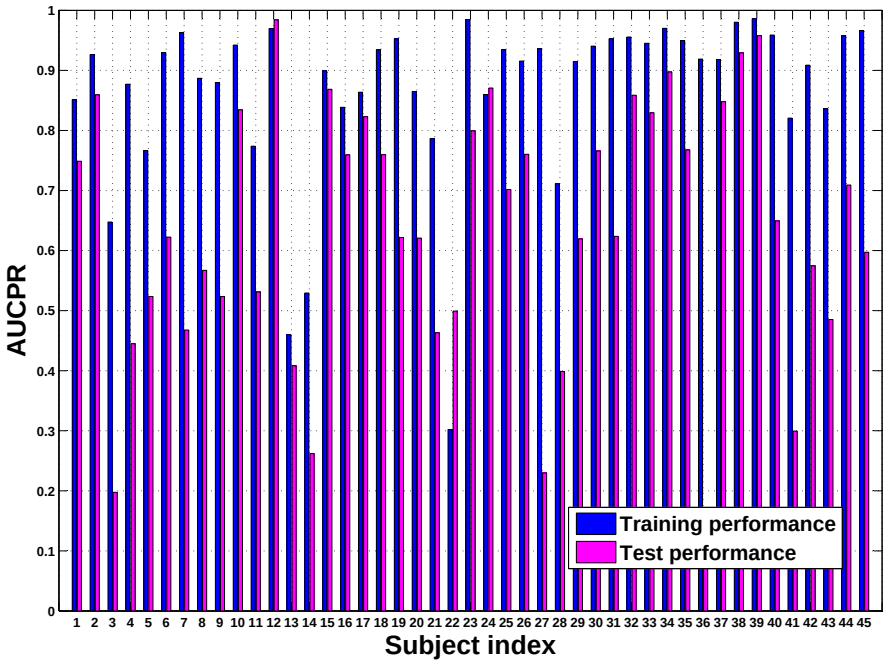


FIGURE 4.16: Training and test performance of subject-dependent model.

because the amount of data from test subject is not comparable to the amount of training data from other subjects. Thus the influence of the data from test subject might be tiny on the classification performance. Therefore, this is not an appropriate way to test the impact of subject-dependent information on the classification performance.

In order to solve the lack of training data problem meanwhile verifying the usefulness of personalized classifier, a feasible plan is to divide one-night recordings into sleep cycles for the experiments. Instead of training and test with whole-night data, the rest experiments are carried out with cycle data. A complete sleep is composed of several sleep cycles. It can be roughly assumed that sleep cycles are independent. Therefore, it is reasonable to divide whole-night data into sleep cycles and continue the experiments. After cycle segmentation, one sleep cycle is treated as one sample. Rather than having two samples, each subject will have more samples. Then we are able to re-implement the subject-dependent paradigm with cycles to see the impact of subject-dependent information on classification performance.

4.2.3 Subject-dependent model based on cycles

The test for subject-dependent model based on cycles is carried out under three conditions.

- ***Scenario a***

  Subject-independent paradigm. Each model is trained with data from all the subjects but the test subject and tested on the n cycles of the test subject.

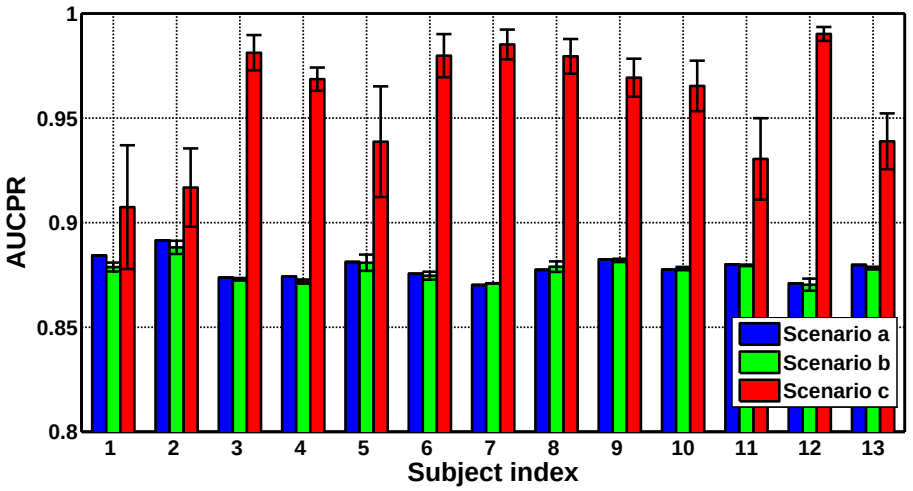
- ***Scenario b***

  Partial subject-dependent paradigm. Each model is trained with data from all the subjects but the test subject, and the data from $(n-1)$ cycles of the test subjects. The model is tested on the remaining cycle of the test subject. This process will repeat n times for each subject.

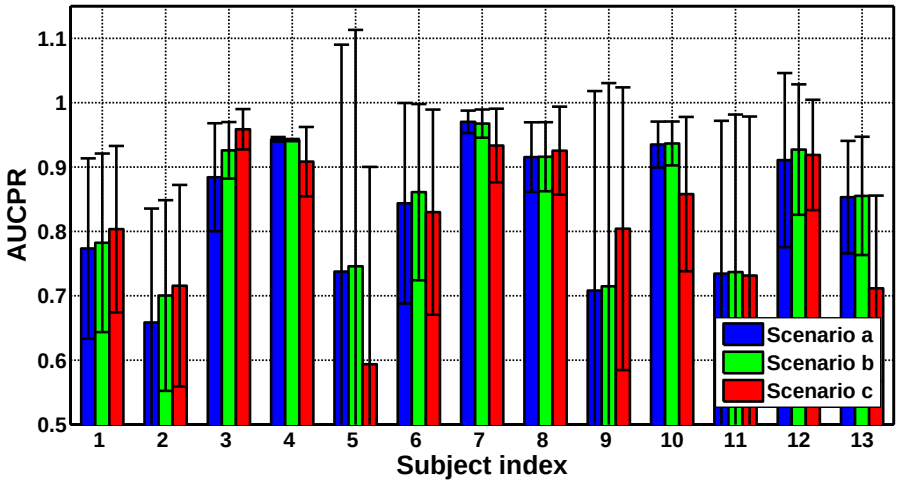
- ***Scenario c***

  Subject-dependent paradigm. Each model is trained with data from $(n-1)$ cycles of the test subjects and tested on the remaining cycle. This process will repeat n times for each subject.

Training and test performance under three scenarios are shown in Fig.4.17. Results of each scenario are reported by the averaged performance of n cycles with standard deviation. The training performance of ***Scenario c*** is significantly better than the other scenarios. However, the test performance of ***Scenario c*** in most cases is the worst in the three scenarios. This is inconsistent with our expectations. It is expected by training with data from test subjects themselves, the personalized model can outperform the model trained with data from other subjects because the model has learned unique characteristics featuring one's sleep. However, the big gap between training performance and test performance of ***Scenario c*** implies the reason is the lack of training data. Even the whole-night recordings have been segmented to ensure each subject has at least 3 samples, the number of samples is still limited for subjects. Fig.4.18 demonstrates the amount of data of each subject (in epochs). In average, each subject has 1000 epochs' data for training and test. With such limited data, the classifier can hardly generalize to have a reliable performance. We have checked that with how many sleep cycles in the training data a model is going to generalize to achieve a reliable performance. Fig.4.19 shows the learning curve of a model with y-axis represents the performance of the model and x-axis represent the number of cycles in the training data. It is observed that until there being at least 30 sleep cycles in the training data learning curves start to stabilize. This proves that the training data in ***Scenario c*** is insufficient to make a valid personalized model. For this reason, the comparison between ***Scenario c*** and ***Scenario a*** makes no sense in testing the usefulness of personalized classifier. The emphasis should be put on the comparison between ***Scenario a*** and ***Scenario b***.



(a) Training performance



(b) Test performance

FIGURE 4.17: Results under 3 scenarios.

	Averaged $AUCPR \pm$ std
Scenario a	0.836 ± 0.102
Scenario b	0.847 ± 0.098

TABLE 4.11: Averaged performance of ***Scenario a*** and ***Scenario b***

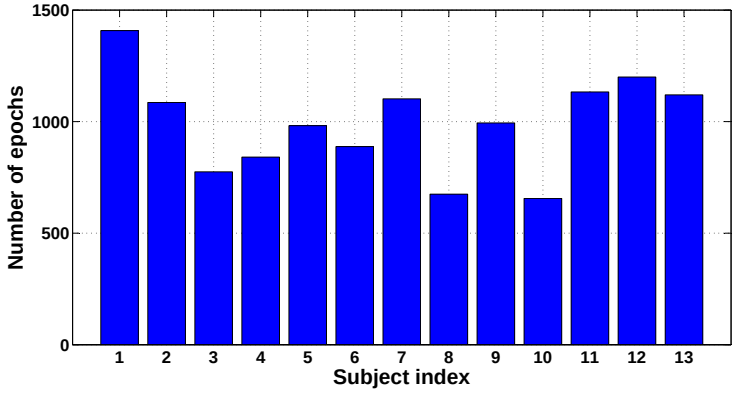


FIGURE 4.18: Amount of data for each subject.

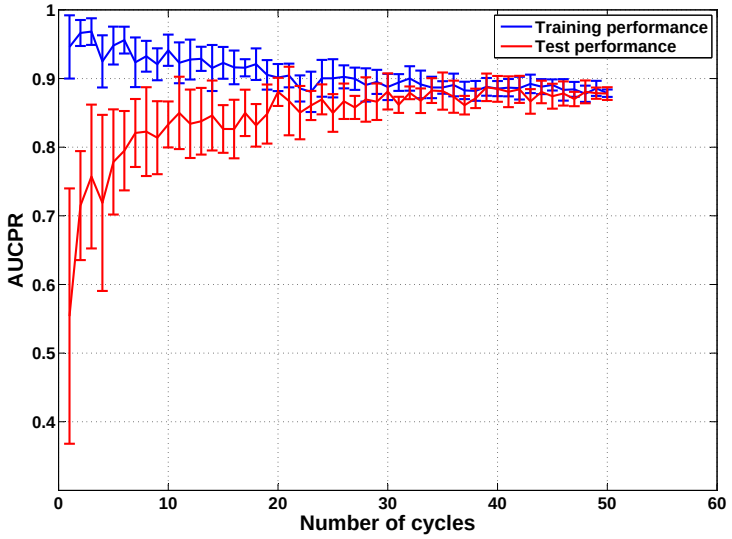


FIGURE 4.19: An example of learning curves in ***Scenario a***.

Comparing ***Scenario a*** and ***Scenario b***, the model integrating partial data from test subject is better than the subject-independent model (Table.4.11). The individual performance in ***Scenario b*** is at least no worse than the performance in ***Scenario a***. The one-sided Wilcoxon rank sum test confirms the improvement is statistically significant with $p - value = 0.003$.

Fig.4.20 is an example of the learning curves under ***Scenario a*** and ***Scenario b*** of one subject. It can be seen from Fig.4.20 that with personal data in the training session, the model starts to generalize faster than the subject-independent model. When looking at the performance when the training data contains no more than 20 sleep cycles, the partial subject-dependent model has been stabilized while the subject-independent model is still in the learning process. The performance of the personalized model is better than the

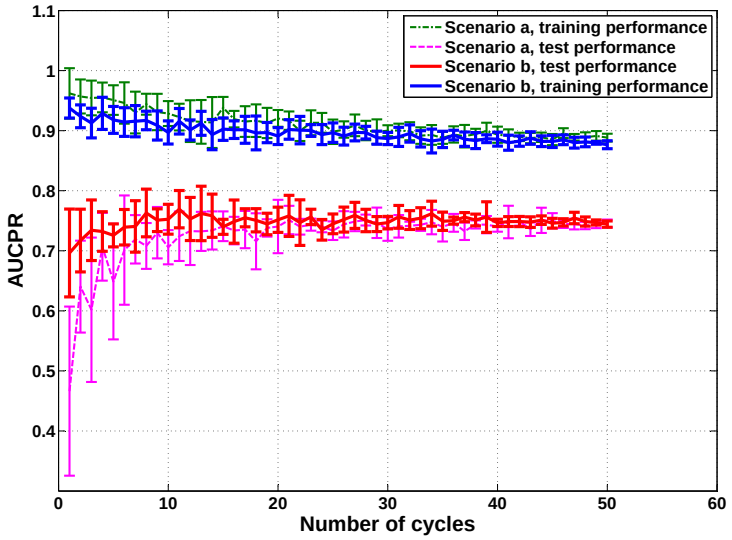


FIGURE 4.20: An example of learning curves in ***Scenario a*** and ***Scenario b***.

performance of subject-independent model overall. With more cycles in the training data, the performances of two models become almost the same. The reason for the same performance of two models is that the weights of subject-dependent data becomes small when adding data from other subjects. If continue adding new data from other subjects, the discrepancy in the amount of data from other subjects and data from test subject biases the model in favor of the subject-independent part. Nevertheless, the fast generalization and better performance of the personalized model suggests the importance of personalization and the second hypothesis has been validated. Only by incorporating less than 10% data from the test subject in the training session makes a better model in deep sleep classification. It can be expected if more data from the test subject can be obtained we are able to have a decent personalized classifier with existing framework. However, the personalized classifier needs data, especially ground-truths, from test subject. In practical, it is difficult to acquire labelled sleep data. Therefore, the results of our experiments only proves the feasibility of personalized model on deep sleep classification. Under the unsupervised circumstances, where data from the test subject are available but the ground-truths are unavailable, one can think of searching for subjects who are similar to test subject on the feature level. The data and ground-truths from the similar subjects can be used for personalization for the test subject. To define similarity on the feature level, one can think of comparing feature distributions.

Chapter 5

Conclusions

The hypotheses this report aims at validating are:

1) The performance of deep sleep classification is affected by $influencial\ factors$ (such as the continuity and duration of deep sleep intervals) that are not captured by extracted features.

2) The performance of deep sleep classification can be improved by personalized classifier.

Both of them are motivated by the result of our early investigation, in order to identify factors that limit the performance of deep sleep classification. In addition to features which are extracted from raw cardiorespiratory signals, $influencial\ factors$ also play an important role in deep sleep classification. The significant $influencial\ factors$ are the duration of deep sleep interval and the continuity of deep sleep interval. The relationship between $influencial\ factors$ and the classification performance are established by multiple linear regression. The classification performance can be predicted by the duration of deep sleep interval, the continuity of deep sleep interval and their interactions. Up to 65% of the variance in the classification performance can be explained by a linear combination of the continuity and the duration of deep sleep interval along with their interactions. If the duration of deep sleep is approximate 14 - 36 minutes with $Findex$ (a parameter to quantify deep sleep continuity) about 0 - 0.06, the expected classification performance justified by the area under the *Precision-Recall* curve is no worse than 0.6. The framework used in this research is able to achieve good performance ($AUCPR \geq 0.7$) of deep sleep classification for the subjects and/or cycles which have long and/or continuous deep sleep intervals. To further improve the performance of deep sleep classification, scientists need to target the subjects and/or sleep cycles whose deep sleep intervals are short and/or fragmented.

The performance of deep sleep classification can also be improved if subject-dependent information are considered by the constructed model. The procedure of incorporating subject-dependent information in the classifier is called personalization. The personalized classifier is able to capture personal sleep characteristics, therefore achieving a better classification performance. Within the selected subjects, the averaged classification performance justified by the area under the *Precision-Recall* curve is improved by 1.1% with a *p-value* of 0.003. Although we have proved that the personalized classifier can perform significantly better than the subject-independent model, in the validation process, the partial data and ground-truths of the test subject are utilized. Thus our validation is no more than a theoretical prove of the usefulness of personalized classifier. The validation of the second hypothesis suggests the personalization of the classification model is another possible direction for improving the performance of deep sleep classification.

Chapter 6

Future work

The validation of the two hypotheses gives hints for improving the performance of deep sleep classification. Having proved that factors, such as the continuity and duration of deep sleep interval, are significant variables for the classification performance, improving the classification performance can be exclusively training a model for a subject whose deep sleep is short and/or fragmented. For subjects and/or sleep cycles which have short and/or continuous deep sleep intervals, in order to improve their classification performance, the features can be specifically processed so that the differences between deep sleep and non-deep sleep in the feature expression are enlarged. Another possible solution for improving the classification performance for subjects and/or sleep cycles which contain the short and/or fragmented deep sleep intervals is to search for features that are robust to the duration and continuity of the deep sleep interval. The robust features are those features where the distinctions between deep sleep and non-deep sleep are large regardless of the duration and continuity of the deep sleep interval.

As mentioned in the conclusion, the partial data and ground-truths of the test subjects are used for constructing the personalized classifier. Yet, in the real situation where performing the deep sleep classification in realtime, though it is possible to acquire partial sleep recordings from the test subject, it is impossible to acquire the ground-truths of the test subject for personalization. The future work can target to incorporating subject-dependent information into the classification framework in an unsupervised way. A possible strategy could initiate from the feature selection process. An unsupervised personalized feature selection method, which compares the distributions of a feature in test subject and other subjects to determine the usefulness of the feature, might help improve the performance of deep sleep classification.

Bibliography

[1] Apoor S Gami, Daniel E Howard, Eric J Olson, and Virend K Somers. Day–night pattern of sudden death in obstructive sleep apnea. *New England Journal of Medicine*, 352(12):1206–1214, 2005.

[2] Matthew P Walker and Robert Stickgold. Sleep-dependent learning and memory consolidation. *Neuron*, 44(1):121–133, 2004.

[3] Robert Stickgold. Sleep-dependent memory consolidation. *Nature*, 437(7063):1272–1278, 2005.

[4] Michel Billiard and Jonathan Bruce Santo. ¡ i¿ sleep: Physiology, investigations and medicine¡/i¿. 2003.

[5] Raymond H Myers. *Classical and modern regression with applications*, volume 2. Duxbury Press Belmont, CA, 1990.

[6] Adrian Rodney Pagan and Anthony David Hall. Diagnostic tests as residual analysis. *Econometric Reviews*, 2(2):159–218, 1983.

[7] By Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908.

[8] George W Snedecor. Calculation and interpretation of analysis of variance and covariance. 1934.

[9] Frederick Mosteller and John W Tukey. *Data analysis, including statistics*. 1968.

[10] Giulio Tononi and Chiara Cirelli. Sleep function and synaptic homeostasis. *Sleep medicine reviews*, 10(1):49–62, 2006.

[11] Allan Rechtschaffen and Anthony Kales. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. 1968.

[12] Michael H Silber, Sonia Ancoli-Israel, Michael H Bonnet, Sudhansu Chokroverty, Madeleine M Grigg-Damberger, Max Hirshkowitz, Sheldon Kapen, Sharon A Keenan, Meir H Kryger, Thomas Penzel, et al. The visual scoring of sleep in adults. *J Clin Sleep Med*, 3(2):121–131, 2007.

[13] Ernst Niedermeyer and Fernando H Lopes da Silva. *Electroencephalography: basic principles, clinical applications, and related fields*. Wolters Kluwer Health, 2005.

[14] Sydney S Cash, Eric Halgren, Nima Dehghani, Andrea O Rossetti, Thomas Thesen, ChunMao Wang, Orrin Devinsky, Ruben Kuzniecky, Werner Doyle, Joseph R Madsen, et al. The human k-complex represents an isolated cortical down-state. *Science*, 324(5930):1084–1087, 2009.

[15] Luigi De Gennaro and Michele Ferrara. Sleep spindles: an overview. *Sleep medicine reviews*, 7(5):423–440, 2003.

[16] Björn Rasch, Christian Büchel, Steffen Gais, and Jan Born. Odor cues during slow-wave sleep prompt declarative memory consolidation. *Science*, 315(5817):1426–1429, 2007.

[17] Hans-Peter Landolt, Derk-Jan Dijk, Peter Achermann, and Alexander A Borbély. Effect of age on the sleep eeg: slow-wave activity and spindle frequency activity in young and middle-aged men. *Brain research*, 738(2):205–212, 1996.

[18] Aicko Y Schumann, Ronny P Bartsch, Thomas Penzel, Plamen Ch Ivanov, and Jan W Kantelhardt. Aging effects on cardiac and respiratory dynamics in healthy subjects across sleep stages. *Sleep*, 33(7):943, 2010.

[19] James D Geyer, Sachin Talathi, and Paul R Carney. Introduction to sleep and polysomnography.

[20] Reinder Haakma and Robbert Jan Beun. Unobtrusive sleep monitoring. *Measuring Behavior 2012*, page 122, 2012.

[21] S Ancoli-Israel, R Cole, C Alessi, M Chambers, W Moorcroft, and C Pollak. The role of actigraphy in the study of sleep and circadian rhythms. american academy of sleep medicine review paper. *Sleep*, 26(3):342–392, 2003.

[22] AM Adami, TL Hayes, and M Pavel. Unobtrusive monitoring of sleep patterns. In *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*, volume 2, pages 1360–1363. IEEE, 2003.

[23] JW Kantelhardt, S Havlin, and P Ch Ivanov. Modeling transient correlations in heartbeat dynamics during sleep. *EPL (Europhysics Letters)*, 62(2):147, 2003.

[24] Eduardo Pinheiro, Octavian Postolache, and Pedro Girão. Theory and developments in an unobtrusive cardiovascular system representation: ballistocardiography. *The open biomedical engineering journal*, 4:201, 2010.

[25] J Alihanka, K Vaahtoranta, and I Saarikivi. A new method for long-term monitoring of the ballistocardiogram, heart rate, and respiration. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 240(5):R384–R392, 1981.

[26] David C Mack, James T Patrie, Paul M Suratt, Robin A Felder, and Majd Alwan. Development and preliminary validation of heart rate and breathing rate detection using a passive, ballistocardiography-based sleep monitoring system. *Information Technology in Biomedicine, IEEE Transactions on*, 13(1):111–120, 2009.

[27] Tim Willemen, Dorien Van Deun, Vincent Verhaert, Marie Vandekerckhove, Vasileios Exadaktylos, Johan Verbraecken, S Huffel, Bart Haex, and Jos Vander Sloten. An evaluation of cardio-respiratory and movement features with respect to sleep stage classification. 2013.

[28] Yosuke Kurihara and Kajiro Watanabe. Sleep-stage decision algorithm by using heartbeat and body-movement signals. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 42(6):1450–1459, 2012.

[29] Sani M Isa, Ito Wasito, and Aniati Murni Arymurthy. Kernel dimensionality reduction on sleep stage classification using ecg signal. *International Journal of Computer Science Issues (IJCSI)*, 8(4), 2011.

[30] Jean Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254, 1996.

[31] Z Shinar, A Baharav, Y Dagan, and S Akselrod. Automatic detection of slow-wave-sleep using heart rate variability. In *Computers in Cardiology 2001*, pages 593–596. IEEE, 2001.

[32] Gari D Clifford, Francisco Azuaje, Patrick McSharry, et al. *Advanced methods and tools for ECG data analysis*. Artech house London, 2006.

[33] Joseph E Mietus. Time domain measures: from variance to pnnx. *Beth Israel Deaconess Medical Center, Harvard Medical School, Boston*, 2006.

[34] Marek Malik, J Thomas Bigger, A John Camm, Robert E Kleiger, Alberto Malliani, Arthur J Moss, and Peter J Schwartz. Heart rate variability standards of measurement, physiological interpretation, and clinical use. *European heart journal*, 17(3): 354–381, 1996.

[35] C-K Peng, Sergey V Buldyrev, Shlomo Havlin, M Simons, H Eugene Stanley, and Ary L Goldberger. Mosaic organization of dna nucleotides. *Physical Review E*, 49 (2):1685, 1994.

[36] Forrest Sheng Bao, Xin Liu, and Christina Zhang. Pyeeg: an open source python module for eeg/meg feature extraction. *Computational intelligence and neuroscience*, 2011, 2011.

[37] M Wiggins, A Saad, Brian Litt, and G Vachtsevanos. Evolving a bayesian classifier for ecg-based age classification in medical applications. *Applied soft computing*, 8 (1):599–608, 2008.

[38] Madalena Costa, Ary L Goldberger, and C-K Peng. Multiscale entropy analysis of complex physiologic time series. *Physical review letters*, 89(6):068102, 2002.

[39] Madalena Costa, Ary L Goldberger, and C-K Peng. Multiscale entropy analysis of biological signals. *Physical Review E*, 71(2):021906, 2005.

[40] Xi Long, Jérôme Foussier, Pedro Fonseca, Reinder Haakma, and Ronald M Aarts. Respiration amplitude analysis for rem and nrem sleep classification. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pages 5017–5020. IEEE, 2013.

[41] A Flexerand, Georg Dorffner, P Sykacekand, and Iead Rezek. An automatic, continuous and probabilistic sleep stager based on a hidden markov model. *Applied Artificial Intelligence*, 16(3):199–207, 2002.

[42] Martin Oswaldo Mendez, Matteo Matteucci, Vincenza Castronovo, Luigi Ferini-Strambi, Sergio Cerutti, and Anna Bianchi. Sleep staging from heart rate variability: time-varying spectral features and hidden markov models. *International Journal of Biomedical Engineering and Technology*, 3(3):246–263, 2010.

[43] LG Doroshenkov, VA Konyshev, and SV Selishchev. Classification of human sleep stages based on eeg processing using hidden markov models. *Biomedical Engineering*, 41(1):25–28, 2007.

[44] G Klosh, B Kemp, Th Penzel, A Schlogl, P Rappelsberger, E Trenker, G Gruber, J Zeithofer, B Saletu, WM Herrmann, et al. The siesta project polygraphic and clinical database. *Engineering in Medicine and Biology Magazine, IEEE*, 20(3): 51–57, 2001.

[45] Michael A Poole and Patrick N O'Farrell. The assumptions of the linear regression model. *Transactions of the Institute of British Geographers*, 52:145–158, 1971.

[46] Martin B Wilk and Ram Gnanadesikan. Probability plotting methods for the analysis for the analysis of data. *Biometrika*, 55(1):1–17, 1968.

[47] Trevor S Breusch and Adrian R Pagan. A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, pages 1287–1294, 1979.

[48] Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324, 1997.

[49] Ron Kohavi and Dan Sommerfield. Feature subset selection using the wrapper method: Overfitting and dynamic search space topology. In *KDD*, pages 192–197, 1995.

[50] Sanmay Das. Filters, wrappers and a boosting-based hybrid for feature selection. In *ICML*, volume 1, pages 74–81. Citeseer, 2001.

[51] Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[52] Claude Robert, Christian Guilpin, and Ayme Limoge. Review of neural network applications in sleep research. *Journal of Neuroscience methods*, 79(2):187–193, 1998.

[53] Guillaume Becq, Sylvie Charbonnier, Florian Chapotot, Alain Buguet, Lionel Bourdon, and Pierre Baconnier. Comparison between five classifiers for automatic scoring of human sleep recordings. In *Classification and Clustering for Knowledge Discovery*, pages 113–127. Springer, 2005.

[54] Tom Dietterich. Overfitting and undercomputing in machine learning. *ACM computing surveys (CSUR)*, 27(3):326–327, 1995.

[55] Igor V Tetko, David J Livingstone, and Alexander I Luik. Neural network studies. 1. comparison of overfitting and overtraining. *Journal of Chemical Information and Computer Sciences*, 35(5):826–833, 1995.

[56] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[57] Jacob Cohen. *Statistical power analysis for the behavioral sciences (rev.* Lawrence Erlbaum Associates, Inc, 1977.

[58] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM, 2006.

[59] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8): 861–874, 2006.

[60] Lisa Marshall, Halla Helgadóttir, Matthias Mölle, and Jan Born. Boosting slow oscillations during sleep potentiates memory. *Nature*, 444(7119):610–613, 2006.

[61] Hong-Viet V Ngo, Thomas Martinetz, Jan Born, and Matthias Mölle. Auditory closed-loop stimulation of the sleep slow oscillation enhances memory. *Neuron*, 2013.