UNIVERSITAT POLITÈCNICA DE VALÈNCIA

_ TELECOM ESCUELA TÉCNICA VLC SUPERIOR DE INGENIERÍA DE TELECOMUNICACIÓN

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

School of Telecommunications Engineering

A Study of Late Fusion Methods for Multimodal Classification

End of Degree Project

Bachelor's Degree in Telecommunication Technologies and Services Engineering

AUTHOR: Zou, Dejian

Tutor: Vergara Domínguez, Luís

External cotutor: SALAZAR AFANADOR, ADDISSON

ACADEMIC YEAR: 2022/2023

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

_ TELECOM ESCUELA TÉCNICA VLC SUPERIOR DE INGENIERÍA DE TELECOMUNICACIÓN

A STUDY OF LATE FUSION METHODS FOR MULTIMODAL CLASSIFICATION

Resumen

La fusión tardía es un tipo de técnica popular para mejorar la fiabilidad del sistema de reconocimiento y el análisis de datos multimodal puede reducir la incertidumbre de la información y mejorar el rendimiento de los modelos. En este proyecto, se utiliza un conjunto de datos de ECG y EEG para diferenciar entre dos etapas del sueño, vigilia y sueño. Estos datos en bruto son pre-procesados y segmentcorrectamente luego se puede construir el conjunto de entrenamiento y el conjunto de prueba. Probamos dos esquemas diferentes para dividir todo el conjunto de datos en conjunto de entrenamiento y conjunto de prueba. Posteriormente, implementamos varios métodos de extracción de características, clasificadores y métodos de fusión tardía. Los métodos de extracción de características incluyen coeficientes de regresión automática, entropía de Shannon, energía de banda y parámetros de Hjorth. Los clasificadores contienen KNN, LDA, QDA, MLP, DT y SVM. Los métodos de fusión tardía incluyen voto mayorit, métodos baye, teoría de Dempster-Shafer e Integral difusa. Además, se diseñdiferentes experimentos para que se puedan obtener comparaciones entre clasificadores, entre métodos de fusión tardía y entre clasificadores individuales y métodos de fusión tardía. Además, se lleva a cabo un estudio de ablación para analizar los efectos de los componentes principales de forma que el modelo pueda optimizarse gradualmente mediante la eliminación de los componentes inútiles.

Mostramos que los métodos de fusión tardía no siempre son mejores que los clasificadores individuales como ANN. Cuando se construye un conjunto de entrenamiento y un conjunto de pruebas, se supone que el cambio de covarianza se nota y disminuye. Por estudio de ablación, encontramos que todo el sistema es aparentemente afectado por clasificadores individuales y la eliminación de clasiinferiores puede ser un método válido para mejorar el rendimiento. Dado que el mismo clasificador puede tener un rendimiento diferente en diferentes modalidades, los datos multimodno son absolutamente adecuados para la clasificación utilizando métodos de fusión tardía.

Resum

La fusió tardana és una mena de tècnica popular per millorar la fiabilitat del sistema de reconeixement i l'anàlisi multimodal de dades pot reduir la incertesa de la informació i

millorar el rendiment dels models. En aquest projecte s'utilitza un conjunt de dades de l'ECG-EEG per diferenciar entre dues etapes de son WAKE i SLEEP. Aquestes dades en brut es preprocessen i segmenten correctament i, a continuació, es pot construir un conjunt d'entrenament i un conjunt de proves. Intentem dos esquemes diferents per dividir tot el conjunt de dades en conjunt d'entrenament i conjunt de proves. Després, implementem diversos mètodes d'extracció de característiques, classificadors i mètodes de fusió tardana. Els mètodes d'extracció de característiques inclouen coeficients de regressió automàtica, entropia de Shannon, energia de banda i paràmetres de Hjorth. Els classificadors contenen KNN, LDA, QDA, MLP, DT i SVM. Els mètodes de fusió tardana inclouen el vot majoritari, els mètodes bayesians, la teoria de Dempster-Shafer i la integral difusa. A més, es dissenyen diferents experiments de manera que es puguin obtenir comparacions entre classificadors, entre mètodes de fusió tardana i entre classificadors individuals i mètodes de fusió tardana. A més, es tracta d'un estudi d'ablació per analitzar els efectes dels components principals de manera que el model es pugui optimitzar gradualment eliminant components inútils.

Demostrem que els mètodes de fusió tardana no sempre són millors que els classificadors individuals com ANN. Quan es construeix el conjunt d'entrenament i el conjunt de proves, se suposa que el canvi de covariància s'ha de notar i disminuir. Mitjançant l'estudi d'ablació, trobem que tot el sistema està aparentment afectat per classificadors individuals i l'eliminació de classificadors inferiors pot ser un mètode vàlid per millorar el rendiment. Atès que un mateix classificador pot tenir un rendiment diferent en diferents modalitats, les dades multimodals no són absolutament adequades per a la classificació mitjançant mètodes de fusió tardana.

Abstract

Late fusion is a kind of popular technique to improve the reliability of recognition system and multimodal data analysis can reduce information uncertainty and improves models' performance. In this project, an ECG-EEG dataset is used to differentiate between two sleeping stage WAKE and SLEEP. These raw data are properly pre-processed and segmented then training set and test set can be built. We try two different schemes to split the whole dataset into training set and test set. Afterwards, we implement several feature extraction methods, classifiers and late fusion methods. Feature extraction methods

include Auto-regression coefficients, Shannon entropy, Band energy and Hjorth parameters. Classifiers contain KNN, LDA, QDA, MLP, DT and SVM. Late fusion methods include Majority Voting, Bayesian methods, Dempster-Shafer Theory and Fuzzy Integral. Moreover, different experiments are designed so that comparisons among classifiers, among late fusion methods and between individual classifiers and late fusion methods can be gained. Additionally, an Ablation Study is involved to analyze the effects of main components so that the model can be gradually optimized by removing useless components.

We show that late fusion methods are not always better than individual classifiers such as ANN. When constructing training set and test set, covariance shift is supposed to be noticed and diminished. By Ablation Study, we find that the whole system is apparently affected by individual classifiers and removing inferior classifiers can be a valid method to improve performance. Since the same classifier can have different performance on different modality, multimodal data are not absolutely suitable for classification using late fusion methods.

Table of contents

Chapter 1. Introduction

Late fusion, which refers to combination of expert opinions from different classifiers before taking final decision, has been used to increase the reliability of recognition systems [1]. In medical area, a more comprehensive view can be provided for doctors of specific diseases when combining different modalities. Accordingly, multimodal medical data analysis has potential to reduce information uncertainty and upgrade models' performance [2]. To study the performance of late fusion methods on multimodality, much work has been done in this report.

In this report, a medical dataset called University College Dublin Sleep Apnea Database (UCDSAD) was finally used to distinguish between two different sleeping stage WAKE and SLEEP. It contains electroencephalogram (EEG) and electrocardiogram (ECG) signals collected from 25 subjects. Both ECG and EEG experienced pre-processing including data cleaning and segmentation. Meanwhile, different schemes of partitioning dataset into training set and test set were used. Moreover, since features can reflect properties of signals and benefit classification eventually, several feature extraction methods were implemented for ECG and EEG signal: Auto-regression coefficients and Shannon entropy for ECG; Band energy and Hjorth parameters for EEG.

In this report, many classifiers were tried and implemented. The classifier group included K-Nearest Neighbours Algorithm (KNN), decision tree (DT), linear discriminant analysis (LDA), support vector machine (SVM), quadratic discriminant analysis (QDA) and multilayer perceptron (MLP). In the meantime, several late fusion methods were also implemented. These methods consisted of Majority Voting (Normal and Weighted), Bayesian Methods (Max, Mean and Median), Dempster-Shafer Theory and Fuzzy Integral (Choquet integral and Sugeno integral).

Additionally, several experiments were designed and carried out so as to compare results among different classifiers and late fusion methods. To evaluate the performance of classifiers and late fusion methods, some metrics including accuracy, standard deviation, Kappa Index, AUROC, AUPR, ROC curve, PR curve and time were used. Ablation study was also finished to analyse the effects of main components on different late fusion methods and the model can be optimised during this procedure.

The expected output of the project can be summarized as follows: (1) A MATLAB codebase was constructed to implement data pre-processing, feature extraction, classification and late fusion. (2) In this codebase, several experiments were carried out to evaluate late fusion methods and results were recorded automatically.

The results illustrate that late fusion methods are not always better than individual classifiers and removing inferior classifiers can be a valid method to improve performance of a late fusion system. The rest of this report is organized as follows: Chapter 2 introduces background of modality and late fusion methods in this project. Chapter 3 introduces dataset, pre-processing methods, feature extraction methods, classifiers, late fusion methods and design of experiments. Chapter 4 demonstrates results of experiments with tables and figures and results are also discussed. Eventually, conclusion and possible outlook of research are listed in Chapter 5.

Chapter 2. Background

Many researchers have been attracted by medical data analysis with the help of computers. Common medical data used for clinical diagnosis includes magnetic resonance imaging (MRI), computerized tomography (CT), electroencephalogram (EEG), electrocardiogram (ECG) and so on. Since various important information can be extracted from every single modality, the combination of different modalities can provide a more comprehensive view of patients and their diseases. When we utilize different modalities, methods of integrating information derived from them are supposed to be decided [2]. Late fusion or decision-level fusion is a popular technique. When using late fusion, we train a single classifier by using each modality as a single input. Then the outputs we preliminarily gain of individual classifiers are going to be combined so as to gain the final result.

In this section, modalities including EEG and ECG and several late fusion methods used in this project and their related work will be introduced and discussed. The details are as follows: EEG and ECG are popular medical data used for diagnosis. As is explained in Wikipedia [3][5], ECG is an electrogram of the heart which is a graph reflecting voltage versus time of the electrical activity of the heart. Electrodes placed on the skin are common tools used to detect electrical activity of heart. Small electrical changes existing in heart are a consequence of cardiac muscle depolarization followed by repolarization during each cardiac cycle (heartbeat). These electrodes detect these subtle electrical changes. ECG can be used to measure some physiological data such as the rate and rhythm of heartbeats, the size and position of the heart chambers, the presence of any damage to the heart's muscle cells or conduction system, the effects of heart drugs, and the function of implanted pacemakers [4]. EEG is electrogram which illustrates spontaneous electrical activity of the brain. The bio-signals detected by EEG have been shown to represent the postsynaptic potentials of pyramidal neurons in the neocortex and allocortex. A lot of abnormal electrical discharges exist in brain such as spikes, sharp waves or spike-and-wave complexes that are seen in people with epilepsy and they can be detected by EEG. Therefore, it is often used to supply information to the medical diagnosis. EEG can detect the onset and spatial-temporal (location and time) evolution of seizures and the presence of status epilepticus. Moreover, diagnosis of brain diseases such as depth of anesthesia, sleep disorders, encephalopathies, cerebral hypoxia after cardiac arrest and coma can be assisted by EEG. Many researchers have used ECG or EEG data

for classification. K.Padmavathia et al. [6] has used ECG to detect Atrial fibrillation, which is a type of arrhythmia that causes death in the adults. Additionally, E. Parvinnia et al. [7] use EEG to judge whether patients have Schizophrenia, which is a severe and persistent psychiatric disorder. Examples of EEG and ECG are shown in Figure 1.

Various late fusion techniques have been proposed for combining multiple classifiers by researchers and they were used and tested in this project. Majority Voting is one of the most used methods to combine classifiers. Majority Voting methods utilize counting or weighted counting to map a sample to the correct class. An example of model using Majority Voting to differentiate handwritten characters was introduced in [8]. In this paper, proposed model outperformed other classifier ensembles such as boosting and bagging. Bayesian fusion method uses posteriori probabilities generated by the individual classifier. Max rule, Mean rule or Median rule can be chosen for Bayesian fusion method. In the paper [9], a face recognition system was built based on it and Bayesian fusion method has a better performance than individual PCA based Distance Measure Classifier. Dempster-Shafer Theory [10] is a generalized framework for reasoning with uncertainty. It is considered to have connections to other frameworks such as possibility, probability, and imprecise probability theories. In [11], researchers used this method on iris data and the final showed that the performance of combined classifier was better than single classifier and the improvement is especially apparent when the features for different member classifiers are heterogeneous. The fuzzy integral is a family of nonlinear functional which is defined with respect to a fuzzy measure. Fuzzy measure is a special concept and it generalizes probability measure. Objective evidence supplied by the classifiers is combined during the classifier fusion process. Fuzzy integral used for decision making on handwritten numeral recognition was discussed in [12] and the performance of this multi-classifier fusion method outperformed that of other conventional fusion techniques.

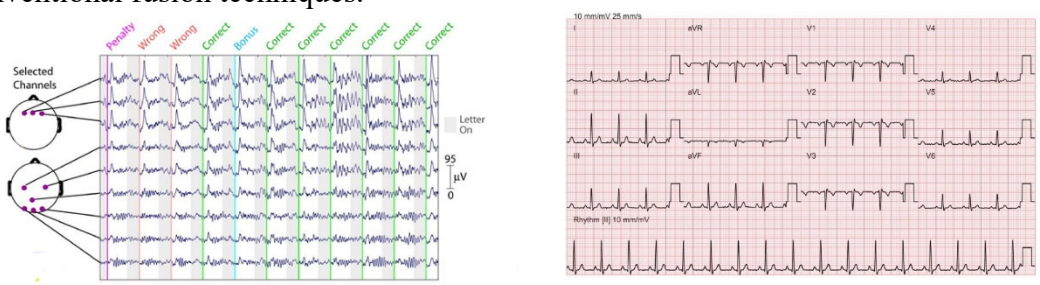


Figure 1. Examples of EEG and ECG. The left picture shows an EEG graph and the right picture shows an ECG graph.

Chapter 3. Design and Implementation

3.1 Dataset

In order to assess late fusion methods implemented, University College Dublin Sleep Apnea Database (UCD database) [13] was used, which contains 25 full overnight polysomnograms with simultaneous three-channel Holter ECG, from adult subjects (21 males and 4 females) with suspected sleep-disordered breathing. The overview of general properties of subjects are shown in Table 1. Expert reference annotations of sleep stages for every 30-second epoch are based on Rechtschaffen and Kales (R&K) rules: WAKE, REM, NREM (Stage 1, Stage 2, Stage 3, Stage 4).

In this dataset, Polysomnograms, which is used to diagnose sleep disorders, were obtained using the Jaeger-Toennies system. Plenty of signals were recorded and they were listed as follows: EEG (C3-A2), EEG (C4-A1), left EOG, right EOG, submental EMG, ECG (modified lead V2), oro-nasal airflow (thermistor), ribcage movements, abdomen movements (uncalibrated strain gauges), oxygen saturation (finger pulse oximeter), snoring (tracheal microphone) and body position. Meanwhile, a Reynolds Lifecard CF system was used by researchers to record three-channel Holter ECGs (V5, CC5, V5R). For this project, ECG (V2) and EEG (C3-A2) sampled by 128Hz were used.

Table 1: General properties of subjects within UCD database

	Mean ± Std	Range
Age	49.96 ± 9.55 years	28.0 - 68.0 years
BMI	31.60 ± 4.03 kg/m^2	25.1 - 42.5 kg/m^2
AHI	24.24 ± 20.29	2.0 - 91.0

3.2 Data Pre-processing

The resulting EEG and EEG of each subject are segmented into epochs of 60 seconds with 30- second overlap. In order to decide the annotation of the 60-second epoch, we combine both 30-second epoch reference annotations, prioritizing in the case of a tie WAKE over SLEEP and Artifact or Indeterminate over others. 0 was used to annotate WAKE and 1 was used to annotate SLEEP (REM and NREM). Moreover, in total of 19 samples were removed since original class is Artifact or Indeterminate. Therefore, in total of 20945 samples were constructed from original dataset. Then, two schemes were used to divide the whole dataset into two parts: a. 50% subjects were used as training set and 50% subjects were used as test set. b. 50% segments of each subject were used as training

set and 50% segments of each subject were used as test set. For scheme a, since these subjects are selected from patients independently and randomly, the first 13 subjects were used as training set and the rest subjects were used as test set. For scheme b, segments of each subject were randomly split into two sections and then two sections were put into training set or test set.

3.3 Feature extraction

Features can reflect properties of ECG and EEG signals. Therefore, they are suitable for classification task. In this project, two features were firstly chosen for ECG and EEG signal respectively: Auto-regression coefficients and Shannon entropy for ECG, and Band energy and Hjorth parameters for EEG.

3.3.1 Auto-regression coefficients

Autoregressive model utilizes the property of linear prediction [14]. For time series F(n), The p-th order autoregressive time series of it is shown below:

$$F(n) = \sum_{i=1}^{p} \alpha_i F(n-i) + \varepsilon(n) \tag{1}$$

Where p is the model order, $\varepsilon(x)$ is error and it is assumed to be white Gaussian noise with zero mean. The AR model parameters α_i are calculated so that the MSE shown in equation (2) is minimized when $\frac{\partial E}{\partial \alpha_i} = 0$. In this project, p was equal to 10.

$$E = \sum_{n=1}^{N} (F(n) - \sum_{i=1}^{p} \alpha_i F(n-i)) \tag{2}$$

3.3.2 Shannon entropy

Shannon entropy is used to measure uncertainty related to random variables in information theory. Based on the probability, it can be calculated as Equation (3):

$$SE = -\sum_{k=1}^{N} p_k \log(p_k) \tag{3}$$

To attain Shannon entropy, Wavelet Packet Decomposition (WPD) is used [15]. The WPD originates from DWT (Discrete Wavelet Transform) which is a renowned technique for signal processing but it extends DWT. Compared WPD with DWT, the main difference is that the

former decomposes not only the detailed coefficients but also the approximation coefficients simultaneously. A two-level WPD tree for ECG signal is shown in Figure 2. In this project, a five-level WPD binary tree with 32 leaves was established.

We use wavelet energy to measure the information of the k-th coefficient of the j-th node at i-th level and it is defined as follows:

$$E_{i,j,k} = \| d_{i,j,k} \|^2 \tag{4}$$

Then, we can calculate sum of energy for the j-th node at i-th level according to Equation (5):

$$E_{i,j} = \sum_{k=1}^{N} E_{i,j,k} \tag{5}$$

Where N represents the quantity of the corresponding coefficients in the node. To calculate the probability of the k-th coefficient at its corresponding node. The rate of wavelet energy is used as is shown in Equation (6):

$$p_{i,j,k} = \frac{E_{i,j,k}}{E_{i,j}} \tag{6}$$

Therefore, Shannon entropy can be computed by Equation (7):

$$SE_{i,j} = -\sum_{k=1}^{N} p_{i,j,k} \log(p_{i,j,k}) \tag{7}$$

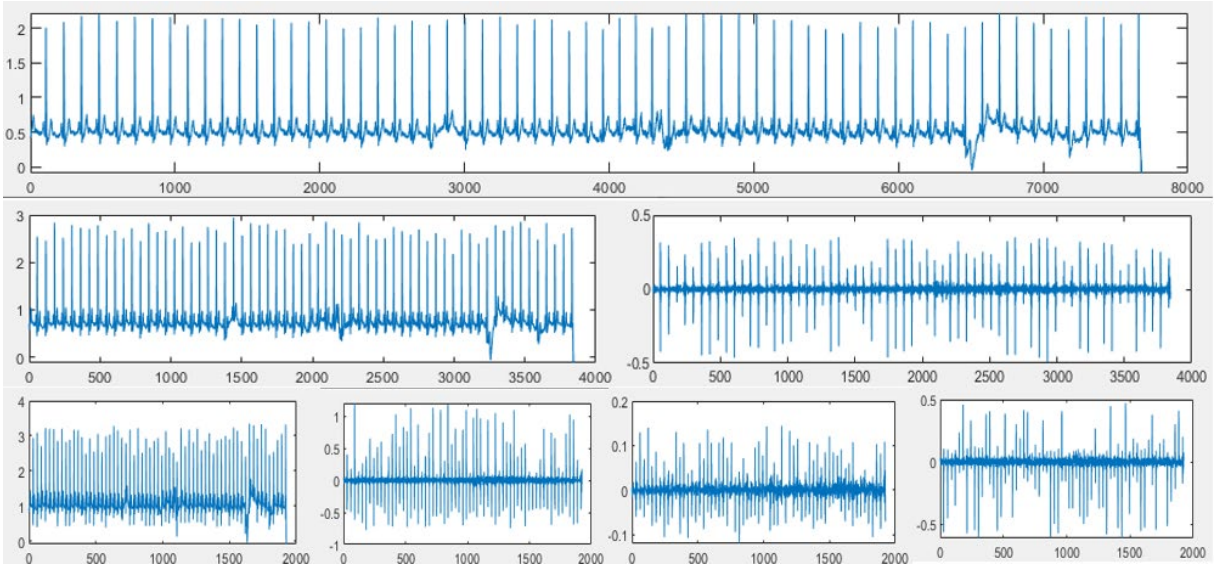


Figure 2. Two-level WPD tree for ECG signal

3.3.3 Band energy

The EEG signal contains several different frequency components, EEG is normally classified as delta = [less than 4 Hz], theta = [4–8 Hz], alpha = [8–13 Hz], beta = [13–30 Hz], and gamma = [more than 30 Hz] [7]. Moreover, the frequency ranges of subjective and electrical artifacts scarcely exceed 50 Hz, and motion or other electrical activity are implied by signals below 0.5 Hz [16]. Therefore, EEG signals were firstly filtered by band-pass FIR filter (Butterworth filter of order five) where the lower cut-off frequency was 0.5 Hz and higher cut-off frequency was 50 Hz. Then by using a band-pass filter, signals were filtered in specific frequency ranges.

3.3.4 Hjorth parameters

Hjorth parameters are famous indicators of statistical properties of signals. It was proposed by Hjorth in 1970 and used for analysing electroencephalogram signals [17]. Hjorth parameters consist three parts: Hjorth Activity, Hjorth Mobility and Hjorth Complexity. And these are calculated as follows:

$$Activity = \mathrm{var}(y(t)) \tag{8}$$

$$Mobility = \sqrt{\frac{\mathrm{var}(\frac{dy(t)}{dt})}{\mathrm{var}(y(t))}} \tag{9}$$

$$complexity = \frac{Mobility(\frac{dy(t)}{dt})}{Mobility(y(t))} \tag{10}$$

3.4 Classifiers

In total of six classifiers were preliminarily implemented in this project. According to the performance, some of them were elected for further experiments. Classifiers involved are as follows: k-nearest Neighbours algorithm (KNN), Linear Discriminant Analysis (LDA), Quadratic discriminant analysis (QDA), Multi-layer Perceptron (MLP), Decision Tree (DT) and Support Vector Machine (SVM). *k-nearest Neighbours algorithm*

3.4.1 k-nearest Neighbours algorithm

KNN algorithm was first developed by Evelyn Fix and Joseph Hodges in 1951 [18] and it is a typical non-parametric supervised learning method. For each sample in the test set, distance between it and each sample in the training set are calculated and distances are sorted. Then k nearest samples in the training set are selected. In this project, k is equal

to 15. For classification, a majority voting scheme meaning that this sample is assigned to the class which outnumber other classes in k samples is used. The KNN algorithm can be demonstrated in the Figure 3.

Normally, Euclidean distance is used for distance metrics:

$$D = \| \vec{x}_1 - \vec{x}_2 \|_2 \tag{11}$$

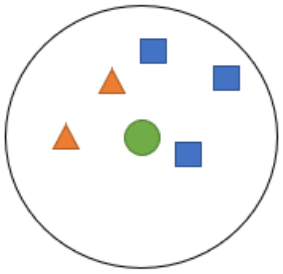


Figure 3. Example of k-NN classification. The green dot represents a test sample and it should be classified either to blue squares or to orange triangles. When k = 5, it is assigned to blue squares.

3.4.2 Linear Discriminant Analysis & Quadratic discriminant analysis

LDA is a common method for dimensionality reduction. All original samples are transformed into a subspace by linear projection so that we can distinguish among samples of different classes [19]. For original dataset $D = \{(\vec{x}_1, y_1), ..., (\vec{x}_N, y_N)\}$, where $\vec{x}$ is a n-dimension vector, we use a projection matrix W to transform $\vec{x}$ into d dimension:

$$\vec{z} = W^T \vec{x} \tag{12}$$

We compute projection matrix based on criteria: Maximize the central distance among classes and minimize in-class variance. To gain projection matrix, we firstly compute two matrixes S_ω, S_b, let D_k represent samples belonging to kth class, then:

$$S_\omega = \sum_{k=1}^{K} \left(\frac{\sum_{\vec{x} \in D_k} \vec{x}\vec{x}^T}{N_k} - \vec{m}_k \vec{m}_k{}^T \right), \vec{m}_k = \frac{\sum_{\vec{x} \in D_k} \vec{x}}{N_k} \tag{13}$$

$$S_b = \sum_{i \neq j} (\vec{m}_i - \vec{m}_j)(\vec{m}_i - \vec{m}_j)^T \tag{14}$$

Afterwards, we compute matrix $S_\omega^{-1} S_b$ and its eigenvalues and eigenvectors. After sorting all eigenvalues, we choose top d largest eigenvalues and corresponding eigenvectors. Finally, $W = (\vec{w}_1, ..., \vec{w}_d)$. In this project, final dimension size is equal to 50% of original

dimension size.

The training process is computing projection matrix with training set. For classifying samples in testing set, we use Maximum likelihood classification (ML) method [20], which means that the data is assumed to follow a distribution according to a previously defined probability model. We assume samples undergoing dimensionality reduction in each class obey gaussian distribution:

$$f_k(\vec{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\sum_k|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\vec{x}-\vec{\mu}_k)^T \sum^{-1}{}_k (\vec{x}-\vec{\mu}_k)) \quad (15)$$

For LDA method, we assume covariance matrixes of each class are equal, so we assign label k to a sample and it is described in equation 10, where π_k represents prior probability.

$$k = \arg\max_k(-\frac{1}{2}(\vec{x}-\vec{\mu}_k)^T \sum^{-1}(\vec{x}-\vec{\mu}_k) + \log(\pi_k)), \pi_k = \frac{N_k}{N} \quad (16)$$

For QDA method, it is closely related to linear discriminant analysis. However, in QDA it is not assume that all classes have identical covariance matrix. As a result, we assign label k to a sample as follows:

$$k = \arg\max_k(-\frac{1}{2}\log|\sum_k| - \frac{1}{2}(\vec{x}-\vec{\mu}_k)^T \sum_k{}^{-1}(\vec{x}-\vec{\mu}_k) + \log(\pi_k)) \quad (17)$$

3.4.3 Multi-layer Perceptron

A multilayer perceptron (MLP) is a kind of fully-connected artificial neural network which mimics the information processing and knowledge learning ability of human. At least three layers containing neuron nodes compose a common MLP: an input layer, a hidden layer and an output layer [21]. A typical structure of MLP is shown in Figure 4.

Except for the input nodes, each node is a neuron that consists of weight, offset and activation function:

$$o = f(\vec{w}^T \vec{x} + b) \quad (18)$$

For hidden layer, sigmoid function is a common activation function which is shown as follows:

$$f(x) = \frac{1}{1+e^{-x}} \quad (19)$$

For output layer, softmax function is a common function. It converts final score vector

into a probability distribution:

$$f_i(\vec{s}) = \frac{e^{s_i}}{\sum_{k=1}^{N} e^{s_k}} \tag{20}$$

The learning process of MLP is based on gradient descent and back propaganda. For each iteration, the weight is updated:

$$W_{t+1} = W_t - \alpha \frac{\partial L}{\partial W_t} \tag{21}$$

Where α represents the learning rate which is selected carefully so as to ensure that the weights quickly converge to a response without oscillations. Backpropagation is an algorithm based on chain rule and is used to compute the gradient of the loss function with respect to the weights of the network. In this project, MLP was trained by SGD with one 10-neutron hidden layer and loss function is cross entropy loss.

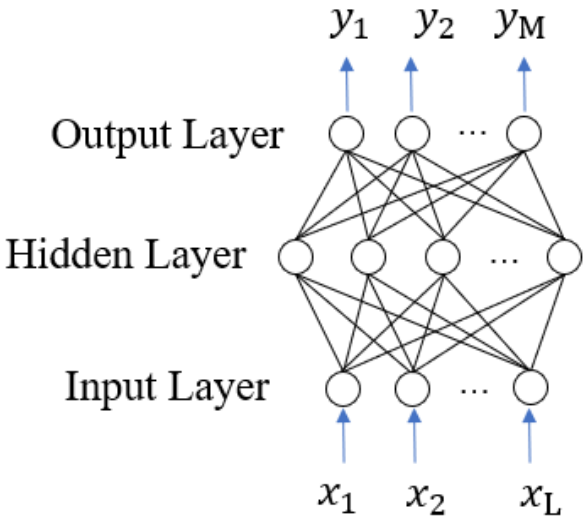


Figure 4. Model of multilayer perceptron.

3.4.4 Decision Tree

A decision tree is a classifier that uses a model resembling a tree which comprises decisions and their possible consequences [22]. One example of decision tree is demonstrated in Figure 5.

ID3 (Iterative Dichotomiser 3) is a typical algorithm for decision tree learning. The ID3 algorithm starts with the original set S. On each iteration of the algorithm, it iterates through every unused attribute of the set S. The attribute which owns largest information gain is selected for decision making. Afterwards the set S is split by the selected attribute

and subsets of the data is constructed. The information gain is defined as equation 22. In this project, parameters of decision tree were automatically optimized

$$I(X,Y) = H(Y) - H(Y \mid X)$$
$$I(X,Y) = -\sum_{y} p(y)\log p(y) - \sum_{x} p(x)H(Y \mid X = x) \quad (22)$$

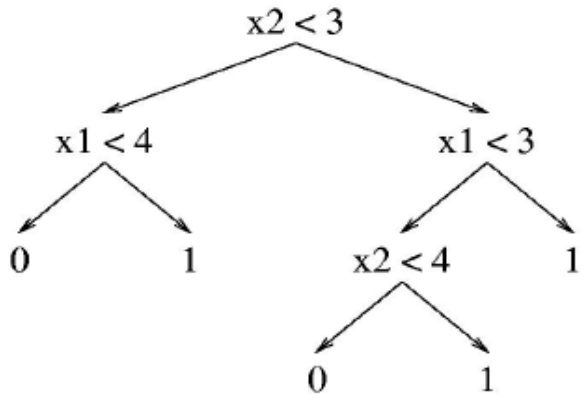


Figure 5. A decision tree for binary classification.

3.4.5 Support Vector Machine

The SVM is also a non-parametric classifier and it was firstly proposed by Vapnik and Chervonenkis in 1971 [23]. SVM finds a hyperplane and allots training examples to points in space in order to maximise the width of the gap between edges of two categories. Figure 6 is an example of SVM model.

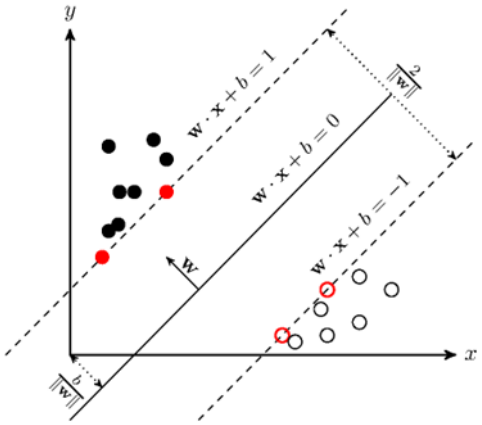


Figure 6. A SVM for binary classification. Red points refer to support vectors.

For linear SVM, the decision function can be defined as follows:

$$f(\vec{x}) = \vec{w}^T\vec{x} + b \quad (23)$$

It is obvious that we should find appropriate $\vec{w}$ and b when training SVM so that when label is 1, output is larger than 1 and when label is -1, output is smaller than -1. There

are two cases called hard-margin and soft-margin. Hard-margin means training data are linearly separable, then the task can be described below:

$$\min_{w,b} \| \vec{w} \|^2, \quad s.t. y_i(\vec{w}^T \vec{x}_i + b) \geq 1, 1 \leq i \leq N \tag{24}$$

As for soft-margin case, the data are not linearly separable. The hinge loss constrains SVM.

3.5 Late Fusion Methods

Combining classifiers to gain a system with better performance is a popular idea. A final decision is derived from outputs of multiple classifiers. The process of late fusion can be summarized as Figure 7. In this project, four late fusion methods are used for classifier combination: Majority voting (Normal and Weighted), Bayesian methods (Max, Mean and Median rule), Dempster-Shafer Theory and Fuzzy Integral (Choquet and Sugeno Integral).

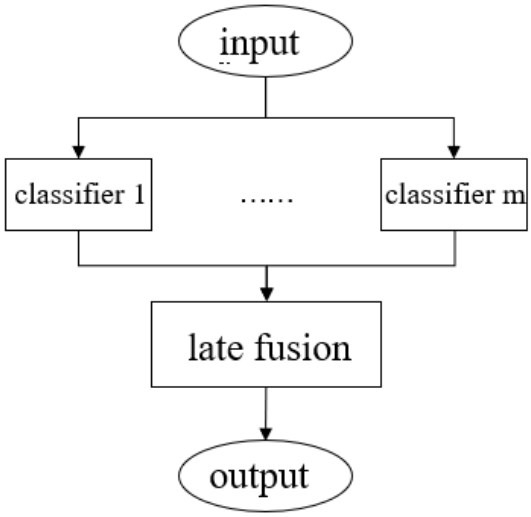


Figure 7. Process of late fusion.

3.5.1 Majority Voting

Majority voting is a common method for classifier combination [8]. We assume there are K classifiers and samples have M classes. The decision of kth classifier:

$$\vec{d}_k = [d_{k1}, d_{k2}, ..., d_{kM}], d_{km} \in \{0,1\}, 1 \leq k \leq K, 1 \leq m \leq M \tag{26}$$

For normal majority voting, the final decision which is the label assigned to the input sample is gained:

$$y_p = \arg\max_j \sum_{k=1}^{K} d_{kj} \tag{27}$$

Sometimes we use weighted majority voting since classifiers have different accuracy:

$$y_p = \arg\max_j \sum_{k=1}^{K} \omega_k d_{kj} \tag{28}$$

Where ω_k represents weighted of classifier k. In this project, weight is defined as follows:

$$\omega = \log(\frac{p}{1-p}) \tag{29}$$

Where p represents accuracy of a classifier.

3.5.2 Bayesian Methods

Bayesian methods are based on the fact that many classifiers can generate posterior probability of each class based on likelihood when one sample is input [9]. Then we compute final probability of each class by mean, median or max value from all classifiers. Finally, we assign the label with largest probability:

$$y_p = \arg\max_j f(\vec{P}_j) \tag{30}$$

3.5.3 Dempster-Shafer Theory

Dempster-Shafer Theory is a framework for dealing with uncertainty based on evidence [24]. We suppose elements in frame Θ which is a finite set of classes are mutually exclusive. Then the basic probability assignment on power set of Θ $m: 2^{\Theta} \to [0,1]$ has some properties:

$$\sum \{m(A) \mid A \subseteq 2^{\Theta}\} = 1, m(\varphi) = 0 \tag{31}$$

Then two functions called belief function and plausibility function are defined:

$$Bel(A) = \sum_{B \subseteq A} m(B) \tag{32}$$

$$pl(A) = \sum_{A \cap B \neq \varphi} m(B) \tag{33}$$

It is considered that the probability of A P(A) is between Bel(A) and pl(A). When we combine mass functions of all classifiers, the rule is as follows:

$$m(A) = m_1 \oplus m_2 \oplus ... \oplus m_n = \frac{\sum_{\cap A_j = A} \prod_{1 \le i \le n} m_i(A_j)}{\sum_{\cap A_j \neq \varphi} \prod_{1 \le i \le n} m_i(A_j)} \quad (34)$$

There are many methods for final decision. For example, we can use belief function [11]:

$$j = \arg\max_j (Bel(A_j)) \quad (35)$$

3.5.4 Fuzzy Integral

Fuzzy Integral is related to the concept fuzzy measures which can be understood as grade or degree of importance [12]. The classifier having greater the fuzzy density is considered to be more important. A fuzzy integral can be considered as the maximal grade of agreement between the objective evidence and the expectation in a classifier system.

Suppose that there are M classifiers and N classes. The fuzzy density of classifier e^k about class i is denoted as $g^{i/k}$. The confusion matrix of a single classifier can be denoted as $C_k = [c^k{}_{ij}]$, where i refers to actual class and j refers to predicted class. Then fuzzy density is computed as follows:

$$g^{i/k} = \frac{c^k{}_{ii}}{\sum_{n=1}^{N} c^k{}_{in}} \quad (36)$$

Afterwards, the λ-fuzzy measures $g_{\lambda_i}(A_k)$, where $A_k = \{e^1, e^2, ..., e^k\}$, can be constructed as follows:

$$\begin{aligned} &g_{\lambda_i}(A_1) = g^{i/1}, i = 1, 2, ..., N \\ &g_{\lambda_i}(A_k) = g^{i/k} + g_{\lambda_i}(A_{k-1}) + \lambda_i g^{i/k} g_{\lambda_i}(A_{k-1}), i = 1, 2, ..., N, k = 1, 2, ..., M \end{aligned} \quad (37)$$

λ_i was obtained using following formula:

$$\lambda_i = \prod_{k=1}^{M} (1 + \lambda_i g^{i/k}), \lambda_i > -1 \wedge \lambda_i \neq 0 \quad (38)$$

The fuzzy integral value is set as overall confidence for each class and it is calculated using the Sugeno fuzzy integral or Choquet fuzzy integral. These two methods are described by formula 39 and 40 respectively:

$$S^{i}_{g} = \max_{k}[\min(f_i(e^k), g_{\lambda_i}(A_k))] \tag{39}$$

$$C^{i}_{g} = \sum_{k=1}^{M} f_i(e^k)[g_{\lambda_i}(A_k)) - g_{\lambda_i}(A_{k-1})] \tag{40}$$

Where $f_i(e^k)$ represents confidence of class i given by classifier e^k. Finally, the sample is assigned to class having largest fuzzy integral.

3.6 Experiments

3.6.1 Environments

Several experiments have been designed in this project and all of them were carried out on MATLAB platform. Additionally, Symbolic Math Toolbox, Wavelet Toolbox, Statistics and Machine Learning Toolbox, Deep Learning Toolbox and Signal Processing Toolbox were involved. The hardware configuration are as follows: CPU is Intel(R) Core (TM) i9-9900K CPU@ 3.60GHz. RAM is 16.0 GB. GPU is NVDIA GeForce RTX 2080.

3.6.2 Pipelines

What is shown in Figure 8 is the overview of late fusion process for multimodality ECG and EEG classification. ECG and EEG firstly experienced feature extraction. Then, data were split into training set and test set. Training set was used for training classifiers and test set was used to assess performance of late fusion methods.

However, to compare performance of individual classifiers and late fusion methods, two experiments were designed and carried out. In experiment 1, performance of individual classifiers when using different features with respect to ECG or EEG was assessed. The pipeline of experiment 1 is described in Figure 9. In experiment 2, classifiers with relatively higher accuracy were elected from experiment 1. Moreover, only one better feature was used for ECG or EEG. The pipeline of experiment 2 is described in Figure 10. Meanwhile, two schemes of dataset splitting mentioned in section 3.2 were tried in experiment 1 and 2. By two schemes, we can observe variance of classifiers and late fusion methods.

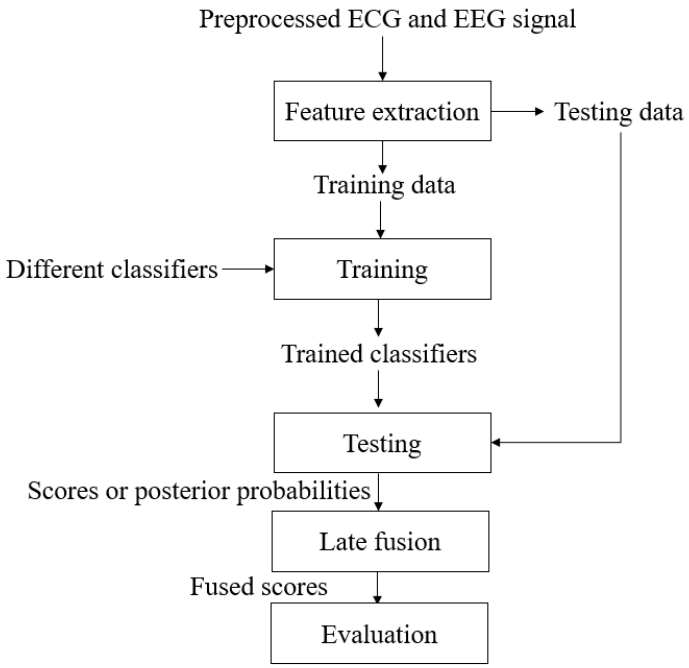


Figure 8. Framework of model.

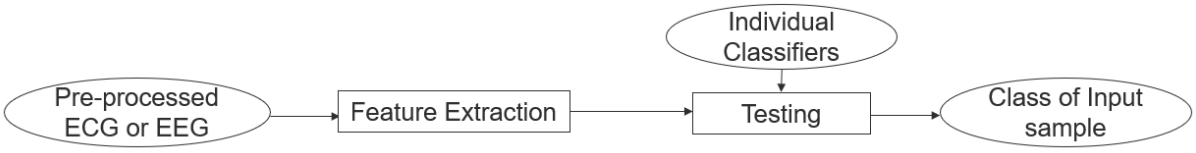


Figure 9. The pipeline of experiment 1.

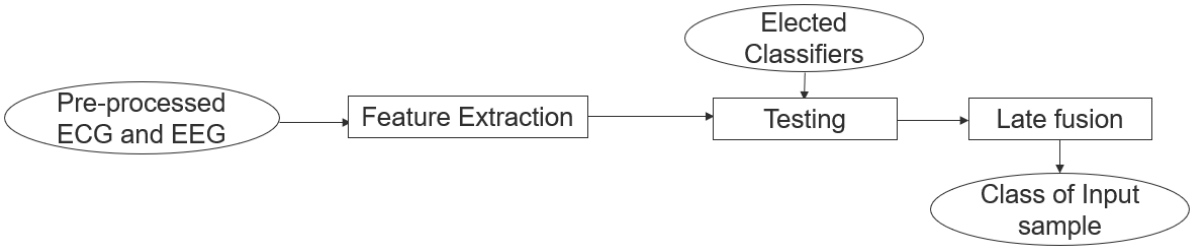


Figure 10. The pipeline of experiment 2.

3.6.3 Evaluation Metrics

Evaluation Metrics were used to assess performance, which are detailed as follows: (1) **Accuracy** is used to assess the general classification ability and derived from the confusion matrix. Accuracy = (TP+TN)/(TP+TN+FP+FN). (2) **Time** is used to reflect the temporal efficiency of a model. (3) **ROC curve** reflects the relationship between FPR and TPR when changing threshold. TPR = TP/(TP+FN), FPR = FP/(FP+TN). (4) **PR curve** reflects the relationship between Recall (TPR) and Precision when changing

threshold. Precision = TP/(TP+FP). (5) **AUROC** is the Area under the ROC curve. It is a threshold independent metric since it is derived when changing threshold of classification. (6) **AUPR** is the Area under the PR curve. It is another threshold independent metric. (7) **Kappa Index is** an index used for consistency test and can also be used to measure the performance of classification. When the number of samples in different classes is unbalanced, the accuracy may be not very suitable for evaluating a model. Let C represents confusion matrix, kappa = (po – pe)/(1 - pe), where po refers to accuracy and pe = $(\sum_{i=1}^{n}(\sum_{j=1}^{n} Cij) \times (\sum_{j=1}^{n} Cji))/(\sum_{i=1}^{n}\sum_{j=1}^{n} Cij)$. Normally kappa is between -1 and 1, the larger kappa is, the better performance the model has. (8) **Standard deviation of accuracy** can measure the stability and variability when experiments are repeated several times.

3.6.4 Ablation Study and Model Optimisation

After carrying out experiment 1 and 2, we surprisingly found that most late fusion methods did not have a better performance than some classifiers. These results will be detailed in Chapter 4. Therefore, Ablation Study was done to analyse the effects of main components so that the model can be gradually optimised by removing useless components. Tactics of Ablation Study can originate from two potential aspects: a. We only use one kind of signal (ECG or EEG). b. We analyse effect of each classifier by removing classifiers consecutively.

Chapter 4. Results and Discussion

4.1 Results of Experiment 1

To test and evaluate the performance of all classifiers, the dataset introduced in Section 3.1 was used to distinguish between WAKE and SLEEP stage. Pre-processed ECG and EEG segments were utilized to extract features by in total of four methods and then individual classifiers were fed by them. Eventually, metrics (Accuracy, AUROC, AUPR) were gained and their average value with respect to classifiers and features were also calculated so that these classifiers and features can be easily analysed. Moreover, two schemes of dataset splitting were tried to analyse the effect and results are displayed in Table 2 and Table 3 respectively. Through the process of the experiment, the fluctuation of results was found. To avoid accidental factors, the experiment 1 was repeated for 5 times and mean values was recorded in final tables.

In Table 2, for classifiers, it can be concluded that SVM has the highest average accuracy and MLP has the largest average AUROC and AUPR. The results also show that the performance of KNN and LDA are apparently inferior compared with other classifiers. As for features, Band Energy of EEG has the best performance since all metrics of it outnumbers those of other features. Shannon Entropy of ECG also has a good performance compared with another feature AR Parameters of ECG.

Table 2: Performance (ACC, AUROC, AUPR: %) of individual classifiers using scheme a.

	Features				
Classifiers	AR Parameters	Shannon Entropy	Band Energy	Hjorth Parameters	Average
KNN	26.76/48.98/72.06	58.54/50.50/73.87	32.84/68.78/83.96	39.05/46.46/71.79	39.30/53.68/75.42
LDA	36.46/45.23/70.35	32.50/53.89/74.37	35.88/76.73/85.53	30.85/48.98/72.06	33.92/56.21/75.58
QDA	73.24/49.56/73.72	73.24/43.30/67.98	73.24/49.44/67.52	26.76/48.98/72.06	61.62/47.82/70.32
MLP	56.48/53.62/76.97	71.95/60.31/78.34	80.77/83.79/91.28	80.42/84.05/91.68	72.40/**70.44**/**84.57**
DT	54.65/54.17/75.68	66.46/58.32/77.32	76.01/84.08/93.08	79.02/81.26/90.95	69.03/69.46/84.26
SVM	73.62/50.23/73.53	73.45/53.29/73.99	80.13/83.47/91.18	77.34/83.24/90.82	**76.13**/67.56/80.38
Average	53.54/50.30/73.72	62.69/53.27/74.31	**63.15**/**74.38**/**85.43**	55.57/65.49/81.56	

Table 3: Performance (ACC, AUROC, AUPR: %) of individual classifiers using scheme b

	Features				
Classifiers	AR Parameters	Shannon Entropy	Band Energy	Hjorth Parameters	Average
KNN	21.18/48.51/77.47	70.45/52.42/79.32	22.49/62.62/84.29	38.49/32.67/51.36	38.15/49.05/73.11
LDA	24.96/49.47/77.49	23.78/49.41/76.07	78.82/75.29/89.98	21.18/48.51/77.47	37.19/55.67/80.25
QDA	78.82/52.45/80.12	78.82/42.48/73.37	78.82/62.90/80.17	21.18/48.51/77.47	64.41/51.59/77.78
MLP	70.83/63.61/85.53	75.07/58.59/82.16	82.71/79.90/90.80	83.53/80.35/90.64	78.14/**70.62**/**87.28**
DT	60.44/59.61/83.08	73.13/54.96/79.19	81.73/83.39/93.55	81.83/78.49/90.55	74.23/69.12/86.59
SVM	78.82/57.95/82.34	79.51/58.99/83.02	82.49/76.88/87.93	82.01/79.91/90.14	**80.59**/68.43/85.86
Average	55.84/55.27/81.00	66.79/52.81/78.86	**71.10**/**73.50**/**87.79**	52.52/61.41/79.61	

Table 3 shows similar results to Table 2. However, the results indicate that performance of most classifiers and features generally ascends when using scheme b except KNN. The reason is that the covariate shift, which refers to the change of distribution of data in training set and test set, is smaller, since the subject group of training set and test set are identical.

4.2 Results of Experiment 2

As is mentioned in Section 3.6.2. Shannon Entropy of ECG and Band Energy of EEG were selected from four types of features. In the meantime, KNN and LDA were removed from individual classifier group due to their bad performance. As a results, in total of eight trained classifiers were integrated into a late fusion system. For each sample, synchronous ECG and EEG segment were used to extract features. Afterwards, two features were input to corresponding classifiers. Finally, late fusion methods were used so that the final decision was acquired and we can calculate metrics (Accuracy, AUROC, AUPR, Time) to assess performance of these late fusion methods. Two schemes of dataset splitting were also tried and it was the same with experiment 1. Table 4 and Table 5 shows results of experiment 2. To avoid accidental factors, the experiment 2 was also repeated for 5 times and mean values was recorded in final tables. In addition, ROC curves and PR curves of experiment 2 were sketched and they are illustrated in Figure 10, 11, 12, 13. To compare late fusion methods with individual classifiers, we add curves of MLP trained on Band Energy of EEG in each graph. For further comparison between individual classifiers and late fusion methods, we also calculate Kappa index and standard deviation of accuracy of 8 classifiers and late fusion methods. This time we only use scheme b to avoid time cost.

Table 4: Performance (ACC, AUROC, AUPR: %, TIME: s) of late fusion methods using scheme a.

		ACC	AUROC	AUPR	TIME
Majority Voting	Normal	76.61	78.52	87.95	1395.49
	Weighted	76.39	79.65	88.38	1425.88
Bayesian Method	Max	**78.24**	57.68	72.17	1380.01
	Median	75.07	80.20	91.59	1384.27
	Mean	77.60	82.34	88.03	1402.14
Dempster-Shafer Theory		74.34	**84.66**	**92.41**	**1360.01**
Fuzzy Integral	Choquet	35.59	63.56	79.54	2928.82
	Sugeon	55.37	39.30	66.36	2929.09

Table 5: Performance (ACC, AUROC, AUPR: %, TIME: s) of late fusion methods using scheme b.

Late Fusion Methods		ACC	AUROC	AUPR	TIME
Majority Voting	Normal	81.20	75.81	89.50	1445.99
	Weighted	80.70	76.91	89.78	**1441.60**
Bayesian Method	Max	**82.09**	55.75	76.08	1448.10
	Median	79.95	78.37	91.90	1497.37
	Mean	82.03	75.97	87.35	1477.29
Dempster-Shafer Theory		79.03	**82.07**	**92.99**	1462.02
Fuzzy Integral	Choquet	35.58	56.79	79.99	3174.60
	Sugeon	79.87	78.11	90.67	3171.49

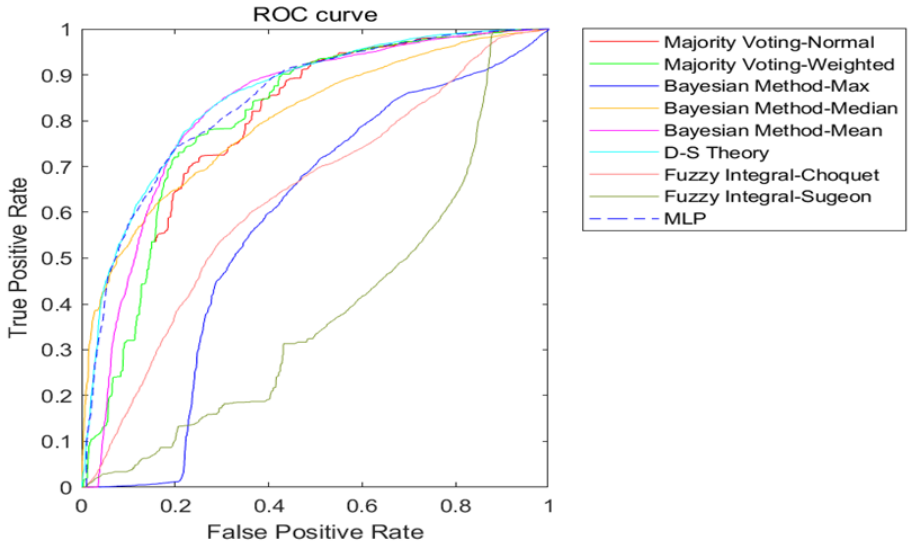


Figure 11. ROC curves of late fusion methods using scheme a.

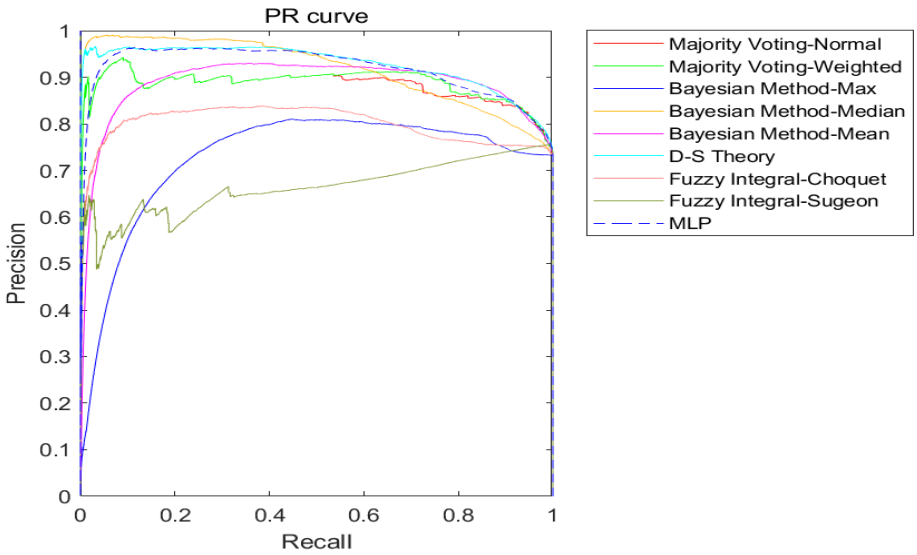
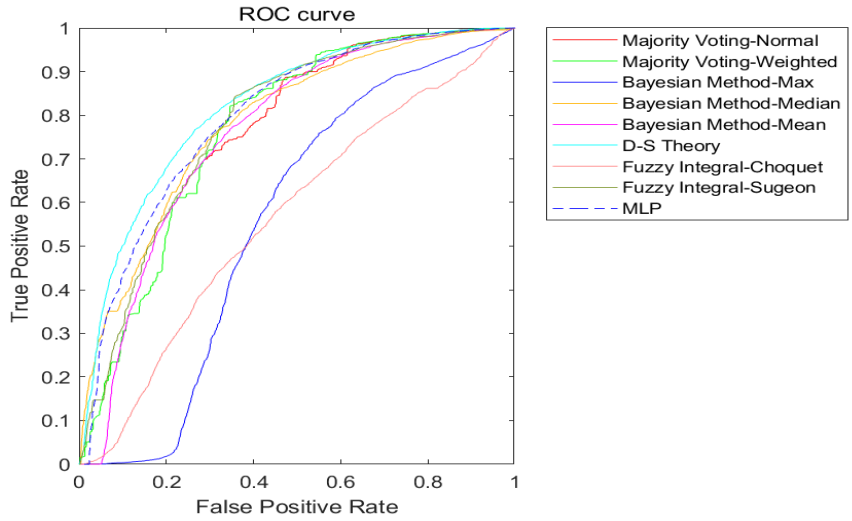


Figure 12. PR curves of late fusion methods using scheme a.



. **Figure 13. ROC curves of late fusion methods using scheme b.**

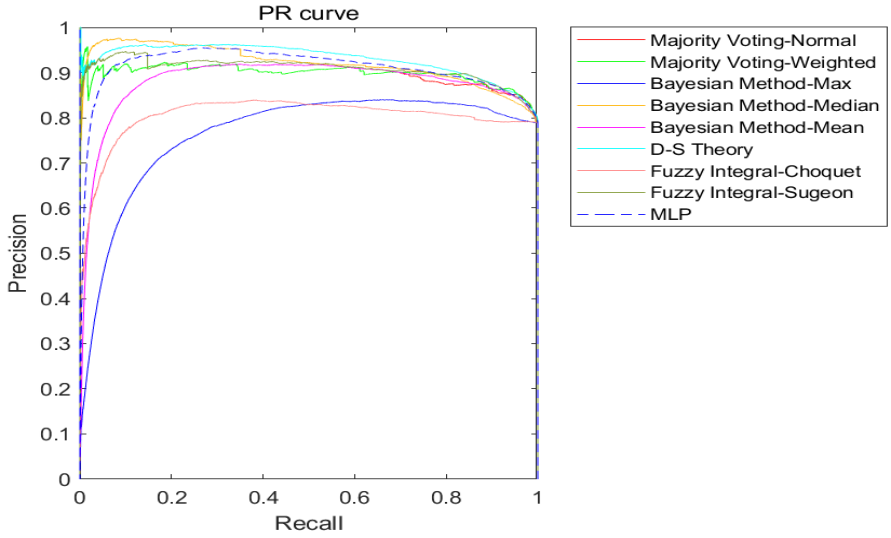


Figure 14. PR curves of late fusion methods using scheme b.

In Table 4, the results show that Bayesian Method with Max rule has the highest accuracy. Then Dempster-Shafer Theory has the largest AUROC and AUPR, and it costs the least time. Moreover, it is easily to find that performance two Fuzzy Integral methods are bad, since the accuracies are pretty low and much more were spent. Table 5 demonstrates similar results to Table 4. Some differences are that Weighted Majority Voting costs the least time and accuracy of Sugeon Fuzzy Integral is acceptable. The results also indicate that performance of most late fusion methods generally ascends when using scheme b except Fuzzy Integral - Sugeon. The accuracy of each late fusion method was gained when we use 0.5 as threshold. With the assistance of ROC and PR curve in Figure 11, 12, 13, 14, we can observe relationship between FPR and TPR or between Recall and Precision when threshold varies from 0 to 1. These ROC and PR curves are consistent

with corresponding AUROC and AUPR values, since AUROC and AUPR are areas under ROC curve and PR curve respectively.

In Table 6, we find that DT trained on Band Energy of EEG has the maximal Kappa Index 61.98. And MLP and SVM trained on Band Energy of EEG also have good Kappa Index. Moreover, standard deviation of accuracies of QDA and SVM are close to 0, which indicates that these two classifiers are pretty stable. In Table 7, it is apparent that Bayesian Method-Max has the largest Kappa Index. Then Bayesian Method-Max and Bayesian Method-Mean are most stable late fusion methods.

However, both Table 4 and 5 imply that all late fusion methods do not outperform individual classifiers within the current model. For example, in Table 3, MLP trained on Band Energy of EEG has the highest accuracy 82.71% among eight used classifiers while accuracies of all late fusion methods cannot transcend it. Compare Table 7 with Table 6, we can also notice that none of late fusion methods have a better Kappa Index.

Table 6: Kappa Index (%) and standard deviation of accuracy of individual classifiers.

	Features	
Classifiers	Shannon Entropy	Band Energy
QDA	44.08/≈ **0**	44.08/≈ **0**
MLP	47.34/1.51×10^{-2}	59.79/8.03×10^{-4}
DT	44.61/7.58×10^{-3}	**61.98**/3.46×10^{-3}
SVM	46.37/≈ **0**	57.33/≈ **0**

Table 7: Kappa Index (%) and standard deviation of accuracy of late fusion methods.

Late Fusion Methods		KAPPA	STDA
Majority Voting	Normal	51.46	1.44×10^{-4}
	Weighted	50.10	9.65×10^{-5}
Bayesian Method	Max	**54.33**	≈ **0**
	Median	47.61	1.44×10^{-4}
	Mean	53.61	≈ **0**
Dempster-Shafer Theory		45.81	5.78×10^{-4}
Fuzzy Integral	Choquet	19.82	2.40×10^{-4}
	Sugeon	47.33	1.00×10^{-2}

4.3 Results of Ablation Study

In this section, effects of main components are validated so that potential methods of optimising the current model can be tried. Only scheme b is used in this section because scheme of dataset partition is not relevant to Ablation Study. Impacts of individual ECG and EEG signal is discussed in Section 4.3.1 and how individual classifiers contribute to this system is discussed in Section 4.3.2.

4.3.1 Effects of ECG and EEG

Classifiers and Features used in this section were the same as those in Section 4.2. The only difference was that the late fusion model comprised four classifiers when using ECG or EEG. Then performance of all late fusion methods was assessed and results were recorded in Table 8.

According to Table 8, it is easily to find that all late fusion methods using EEG globally have better accuracies than those using ECG. Moreover, compared with Table 3, Table shows that late fusion methods surprisingly begin to outperform individual classifiers. According to Table 3, the highest accuracy amid four used classifiers trained on EEG is 82.71% of MLP. Both Majority Voting-Normal and Majority Voting-Weighted have a better accuracy 83.01%, and Bayesian Method has a better accuracy 82.85%. Additionally, Sugeon Fuzzy Integral trained on EEG have better AUROC and AUPR than those in Table 3 (83.39% and 93.55% of DT). It is also apparent that the gap between accuracies of Choquet Fuzzy Integral method when using different signals is huge.

Table 8: Performance (ACC, AUROC, AUPR: %) of late fusion methods using ECG or EEG.

Late Fusion Methods		ACC	AUROC	AUPR
			ECG/EEG	
Majority Voting	Normal	77.69/**83.01**	56.89/75.13	80.69/88.57
	Weighted	**79.49**/**83.01**	57.07/75.12	80.72/88.57
Bayesian Method	Max	79.40/81.65	53.51/72.57	79.82/86.97
	Median	79.01/82.85	**60.12**/83.04	**83.22**/94.16
	Mean	79.25/82.52	58.53/82.87	81.25/93.31
Dempster-Shafer Theory		79.25/79.99	43.18/21.18	74.07/63.93
Fuzzy Integral	Choquet	64.24/82.47	53.30/66.48	80.05/84.34
	Sugeon	79.35/81.65	55.90/**83.59**	80.21/**94.54**

4.3.2 Effects of Classifiers

Classifiers and Features used in this section were also the same as those in Section 4.2. To analyse effects of individual classifiers, four classifiers were removed from original eight classifiers one by one. It is assumed that the classifier with relatively lower accuracy may have negative effects on our model. Therefore, the classifier with the lowest accuracy was removed from the rest of classifiers each time. These experiments are discussed in Table 9.

Table 9: Ablation Study on individual classifiers. From EXP #1 to EXP #4, DT, MLP, QDA trained on ECG and QDA trained on EEG are consecutively removed.

#	ECG				EEG				Late Fusion Methods		ACC	AUROC	AUPR
	QDA	MLP	DT	SVM	QDA	MLP	DT	SVM					
1	√	√		√	√	√	√	√	Majority Voting	Normal	80.27	76.23	89.51
										Weighted	80.24	76.97	89.62
									Bayesian Method	Max	82.09	55.75	76.08
										Median	79.77	71.62	86.05
										Mean	**82.10**	76.28	87.27
									D-S Theory		79.27	**83.64**	**93.53**
									Fuzzy Integral	Choquet	74.61	48.41	73.89
										Sugeon	78.87	81.45	92.60
2	√			√	√	√	√	√	Majority Voting	Normal	**82.63**	75.57	88.73
										Weighted	**82.63**	75.58	88.73
									Bayesian Method	Max	82.09	55.75	76.08
										Median	79.34	70.51	84.80
										Mean	82.11	76.17	86.85
									D-S Theory		79.04	**84.75**	**94.81**
									Fuzzy Integral	Choquet	72.33	43.78	71.36
										Sugeon	78.87	83.43	93.87
3				√	√	√	√	√	Majority Voting	Normal	**82.58**	75.33	88.61
										Weighted	**82.58**	75.32	88.61
									Bayesian Method	Max	82.10	70.51	85.48
										Median	82.33	80.11	92.36
										Mean	82.41	82.50	93.15
									D-S Theory		80.02	**84.83**	**94.93**
									Fuzzy Integral	Choquet	70.35	34.84	69.31
										Sugeon	78.82	83.58	94.55
4				√		√	√	√	Majority Voting	Normal	**83.51**	76.02	88.99
										Weighted	82.98	76.00	88.98
									Bayesian Method	Max	83.03	71.44	85.66
										Median	83.13	82.46	93.10
										Mean	82.80	82.68	93.27
									D-S Theory		83.45	**84.91**	**94.86**
									Fuzzy Integral	Choquet	72.33	76.97	88.34
										Sugeon	81.44	84.30	94.68

As is shown in Table 9, we can easily find that accuracies of most late fusion methods increase except Choquet Fuzzy Integral when inferior classifiers are gradually removed. As for AUROC and AUPR, the situation is complicated. AUROC and AUPR of many methods (three Bayesian Methods, two Fuzzy Integral methods and Dempster-Shafer Theory) have an upward trend while AUROC and AUPR of two Majority Voting methods fluctuate from EXP #1 to EXP #4. Moreover, EXP #4 shows that lots of late fusion methods with two Fuzzy Integral methods as exception are more accurate than the best individual classifier MLP trained on EEG. And better AUROC and AUPR exist in these late fusion methods from EXP #2 to EXP #4.

4.3.3 Discussion

Two experiments of Ablation Study simultaneously indicate that integrating more classifiers into a late fusion model do not always build a model with better performance, since inferior classifiers in the classifier group can mislead the final decision and have a negative influence on the whole system. There is a problem-solving principle called Occam's razor, which is renowned for “Entities must not be multiplied beyond necessity” [25]. It advocates that if competing hypotheses about the same prediction exist simultaneously, we should prefer the one that requires fewest assumptions. In our model, individual classifiers can be sources of assumptions and too much classifiers can be useless. In addition, Section 4.3.1 implies that multimodality data are not always suitable for classification. The main reason is that even the same type of classifier can have obviously different performance on multimodality data. And choices of feature extraction methods can also affect performance of classifiers.

Chapter 5. Conclusion and Further Work

5.1 Conclusion

Existing literature has shown late fusion methods can help to construct a more reliable pattern recognition system and multimodal data can provide extra information for decision making. In this project, a raw ECG and EEG dataset was pre-processed so that it became suitable for classification of sleeping stage. Moreover, various feature extraction methods, classifiers and late fusion methods were implemented on MATLAB after looking through plenty of papers. Several experiments were also carried out in this project. Experiment 1 was used to assess performance of individual classifiers trained on different features and Experiment 2 was used to assess and compare performance of different late fusion methods with selected classifiers and features. Both experiments can reflect whether scheme of dataset partition can affect results. Ablation Study was used to analyse effects of components of our late fusion model so that we could find possible methods to improve our model.

For Experiment 1, we find that identical classifiers trained on different modality can have different performance and feature extraction methods can be one factor affecting performance. For Experiment 2, we notice that all late fusion methods do not outperform the single classifier MLP trained on Band Energy of EEG according to accuracy. And it indicates that problems exist in the preliminary system. Both Experiment 1 and Experiment 2 reflect that scheme b mentioned in Section 3.2 can benefit experiment results. For Ablation Study, we find that when inferior classifiers were removed, performance of most late fusion methods are improved and late fusion methods begin to outperform the individual classifiers used for late fusion. And classifiers trained on Band Energy of EEG are generally better than those trained on Shannon Entropy of ECG so that EEG should be chosen if only one type of signal is allowed to use for late fusion.

In conclusion, late fusion methods are not always better than individual classifiers since current classifiers, such as artificial neural network, can have prominent performance. When selecting classifiers for late fusion, we should care about their performance because the whole system is affected by individual classifiers. Moreover, some factors such as type of modality and feature extraction methods can also affect performance of classifiers and eventually affect the system. Therefore, multimodal data are not absolutely suitable for classification using late fusion methods. When constructing training set and test set,

covariance shift is supposed to be noticed and diminished.

5.2 Further Work

In this project, some practical problems can be barriers to me and they really waste my time. For example, I used to utilize function in MATLAB to construct components in the whole system. However, I gradually found that the code structure was not reasonable since it was not efficient to carry out experiments. After browsing other programmers' codebase in GitHub, I found that object-oriented programming is available in MATLAB. As a result, my code became efficient and user-friendly. To be honest, I would prefer to use python to finish this project if I were given a chance to do it again because python is more suitable for object-oriented programming and many libraries for data processing and machine learning can be available.

Due to limitation of time and sources, only one dataset was chosen to used. In the future, other datasets can be used in our system so that we can test the robustness of these classifiers and late fusion methods on different classification tasks. Extra work for more dataset is that new data pre-processing methods and feature extraction methods should be implemented. In the end, I have to admit that the current late fusion system is less competitive compared with increasingly outstanding deep learning techniques.

References

[1] Mohamed Mohandes, Mohamed Deriche, & Salihu O. Aliyu (2018). Classifiers Combination Techniques: A Comprehensive Review. *IEEE Access, 10.1109/ACCESS.2018.2813079.*

[2] Fatemeh Behrad, & Mohammad Saniee Abadeh (2022). An overview of deep learning methods for multimodal medical data mining. *Expert Systems with Applications,* 200 (2022) 117006.

[3] *Electrocardiography,* Retrieved March 11, 2023, from Wikipedia: https://en.wikipedia.org/wiki/Electrocardiography

[4] Braunwald, Eugene, ed. (1997). *Heart Disease: A Textbook of Cardiovascular Medicine (5th ed.)*. Philadelphia: Saunders. p. 118.

[5] *Electroencephalography,* Retrieved March 11, 2023, from Wikipedia: https://en.wikipedia.org/wiki/Electroencephalography

[6] K.Padmavathia, & K.Sri, Ramakrishnab (2015). Classification of ECG signal during Atrial Fibrillation using Autoregressive modelling. *Procedia Computer Science*, 46 (2015) 53 – 59.

[7] E. Parvinnia, M. Sabeti a M. Zolghadri Jahromi, & R. Boostani. (2013). Classification of EEG Signals using adaptive weighted distance nearest neighbor algorithm. *Journal of King Saud University – Computer and Information Sciences (2014)*, 26, 1–6.

[8] C. De Stefano, F. Fontanella, & A. Scotto di Freca (2012). A Novel Naive Bayes Voting Strategy for Combining Classifiers. In *2012 International Conference on Frontiers in Handwriting Recognition.*

[9] Rizoan Toufiq, & Md. Rabiul Isalm (2016). Face Recognition System Using Soft-output Classifier Fusion Method. In *2nd International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE)*.

[10] G. Shafer (1976). *A Mathematical Theory of Evidence*, Princeton: Princeton University Press.

[11] Deqiang Han, Chongzhao Han, & Yi Yang (2007). Combination of Heterogeneous Multiple Classifiers based on Evidence Theory. In *2007 International Conference on Wavelet Analysis and Pattern Recognition, Beijing, China, 2-4 Nov. 2007*.

[12] Tuan D. Pham (2002). Combination of Multiple Classifiers using Adaptive Fuzzy Integral. In *2002 IEEE International Conference on Artificial Intelligence Systems (ICAIS'02)*.

[13] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P.C., Mark, R., Mietus, J.E., Moody, G.B., Peng, C.K. and Stanley, H.E. (2000). Retrieved March 11, 2023, from St. Vincent's University Hospital / University College Dublin Sleep Apnea Database: https://physionet.org/content/ucddb/1.0.0/

[14] J. G. Proakis (2001). *Digital signal processing principles algorithms and application* Pearson Education India.

[15] Taiyong Li, & Min Zhou (2016). ECG Classification Using Wavelet Packet Entropy and

Random Forests. *Entropy 2016, 18, 285*.

[16] M.H. Libenson (2009). *Practical approach to electroencephalography*, First ed., Saunders,

United States.

[17] Hjorth, Bo, & Elema-Schönander. AB. (1970). EEG analysis based on time domain properties. *Electroencephalography and Clinical Neurophysiology*. *29*, 306–310.

[18] Fix, Evelyn, & Hodges, Joseph L. (1951). *Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties*. USAF School of Aviation Medicine, Randolph Field, Texas.

[19] Suman Kumar Bhattacharyya, & Kumar Rahul. (2014) Face Recognition by Linear Discriminant Analysis, *International Journal of Communication Network Security*, ISSN: 2231 – 1882, Volume-2, Issue-3.

[20] Jens Keuchela et al. (2003). Automatic land cover analysis for Tenerife by supervised classification using remotely sensed data, *Remote Sensing of Environment, 86,* 530–541

[21] Hastie, Trevor et al. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

[22] Otukei, J.R., & Blaschke, T. (2010). Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms.

International Journal of Applied Earth Observation and Geoinformation, Vol.S12, pp S27–S31.

[23] Vapnik, W.N., & Chervonenkis, A.Y. (1971). On the uniform convergence of the relative frequencies of events to their probabilities. T*heory of Probability and its Applications*, Vol.17, pp. 264–280.

[24] Rajib Ghosh, Pradeep Kumar, & Partha Pratim Roy (2019). A Dempster–Shafer theory-based classifier combination for online Signature recognition and verification systems, *International Journal of Machine Learning and Cybernetics,10*, 2467–2482

[25] Ariew, Roger (1976). *Ockham's Razor: A Historical and Philosophical Analysis of Ockham's Principle of Parsimony*. Champaign-Urbana, University of Illinois.