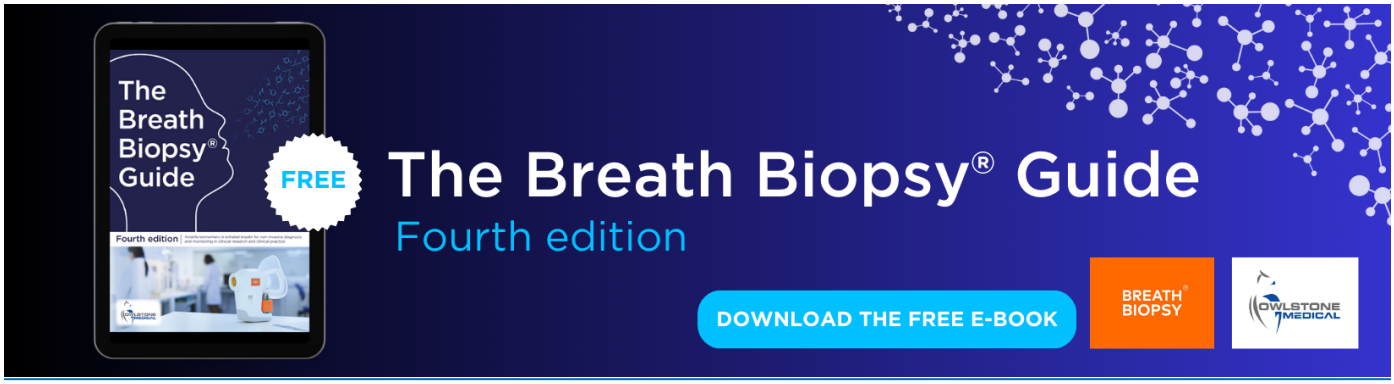
Physiological Measurement

PAPER

Feature extraction of EEG signals based on functional data analysis and its application to recognition of driver fatigue state



IOP Publishing Physiol. Meas. 41 (2020) 125004 https://doi.org/10.1088/1361-6579/abc66e

Physiological Measurement

IPEM Institute of Physics and Engineering in Medicine

PAPER

RECEIVED
21 July 2020

REVISED
26 October 2020

ACCEPTED FOR PUBLICATION
30 October 2020

PUBLISHED
28 December 2020

Feature extraction of EEG signals based on functional data analysis and its application to recognition of driver fatigue state

Pengpeng Shangguan, Taorong Qiu, Tao Liu, Shuli Zou, Zhuo Liu and Siwei Zhang

Department of Computer, Nanchang University, Nanchang Jiangxi, 330029, People's Republic of China

E-mail: tliu@ncu.edu.cn

Keywords: electroencephalography, feature extraction, kernel principal component analysis, functional data analysis

Abstract

Objective: Our objective is to study how to obtain features which can reflect the continuity and internal dynamic changes of electroencephalography (EEG) signals and study an effective method for fatigued driving state recognition based on the obtained features. *Approach*: A method of EEG signalfeature extraction based on functional data analysis is proposed. Combined with kernel principal component analysis method, the obtained features are applied to the recognition of driver fatigue state, and a corresponding recognition model of fatigued driving state is constructed. *Main results*: The recognition model is tested on the real collected driver fatigue EEG signals by selecting a suitable classifier. The test results show that the proposed driver fatigue state recognition method has good recognition effect, especially on the classifier based on decision tree, with an average accuracy of 99.50%. *Significance*: The extracted features well reflect the continuityand internal dynamic changes of the EEG signals, and it is of great significance and application value to study an effective method of fatigued driver state recognition based on the features.

1. Introduction

Electroencephalography (EEG) signals are spontaneous electrical activities of brain cell groups recorded by electrodes from the scalp. They are a kind of physiological signal with strong randomness. The acquisition of EEG signals is of great significance to the detection of brain function and the diagnosis of brain diseases. At present, the main methods of EEG signal analysis are linear and non-linear analysis. The linear analysis method is mainly the time-frequency analysis method (Polat *et al* 2019, Ramos-Aguilar *et al* 2020). Because the EEG system is a non-linear system, some non-linear EEG analysis methods such as artificial neural network (Sharif *et al* 2018, George *et al* 2020) and entropies (Acharya *et al* 2019) such as sample entropy (Song and Zhang 2016), fuzzy entropy (Cao and Lin 2017), approximate entropy (Li *et al* 2016), and wavelet entropy (Mooij *et al* 2016) have been applied to the analysis of EEG signals.

Driver fatigue is a major cause of traffic accidents, and has a significant impact on road safety. Data shows that 35%–45% of traffic accidents are caused by driver fatigue (Li *et al* 2017). Therefore, it is of realistic and far-reaching significance to accurately detect whether a driver is in a fatigued state in time.

At present, a series of studies has been conducted on the recognition of driver fatigue status at home and abroad. Most of them detect driver fatigue through the following two methods. Subjective detection methods such as a questionnaire, driver self-recording table, the Stanford Sleepiness Scale (Hoddes *et al* 1973), and Epworth Sleepiness Scale (Janssen *et al* 2017) have been used to determine a driver's fatigue state. Objective detection methods judge the fatigue state of drivers by measuring physiological signals like EEG (Correa *et al* 2014), electrocardiograph (Wang *et al* 2016), and electrooculogram (EOG) (Ma *et al* 2014), as well as eyelid blinking (Acioglu and Erçelebi 2016), head displacement, driving manipulation behavior, and vehicle trajectory (Shi and Yang 2017). As a kind of direct indicator of the brain state, EEG is considered as the 'gold' standard to identify driver fatigue.

Hu (2017) extracted the features of single-channel EEG signals by four kinds of entropy methods, sample, fuzzy, approximate, and spectral entropy, and then different classifiers were used to classify and recognize the fatigue state. Chai *et al* 2016 proposed a classification method based on two-stage EEG for

driver fatigue detection. An independent component of entropy rate bound minimization analysis is used in this method for the source separation, autoregressive modeling for feature extraction, and Bayesian neural network for the classification algorithm. Wang *et al* (2018) proposed a new real-time driver fatigue detection method based on EEG signals. This method can judge a driver's fatigue state by observing a prediction curve of power spectral density and sample entropy of EEG signals. Zhao *et al* 2016 used graph theory to compare the changes of brain functional networks under normal and fatigued states for fatigue detection. Hajinoroozi *et al* 2016 proposed a novel channelwise convolutional neural network (CCNN) for prediction of driver fatigue states from EEG signals, and compared CCNN and CCNN-R, a CCNN variation that uses a Restricted Boltzmann Machine to replace the convolutional filter with conventional convolutional neural networks (CNNs) and deep neural networks (DNNs). The results showed that CCNN and CCNN-R had better robustness and performance than conventional DNN and CNN. Ye *et al* (2018) proposed a driver fatigue recognition method based on sample entropy and kernel principal component analysis (KPCA), and compared it with those based on three kinds of entropies including sample, fuzzy, and combination entropies that merged with KPCA. The results showed that the method is effective. Luo *et al* 2019 proposed an adaptive multi-scale entropy method based on K-means for driving fatigue detection; the results showed the method to be superior to single-scale entropy and have a good effect on forehead EEG. Dimitrakopoulos *et al* (2018) employed graph theoretical analysis to investigate fatigue-related reorganization of functional brain networks using two different fatigue-inducing paradigms: simulated driving and process verification test (PVT) tasks. Their results provide some of the first quantitative evidence to show the complex nature of fatigue-related neural mechanisms and present the feasibility of whole-brain network analysis in assessing the fatigue processes and the effectiveness of using connectivity features for monitoring of mental fatigue. Han *et al* (2019) introduced complex network theory to study the evolution of brain dynamics under different rhythms of EEG signals during several periods of simulated driving. Their result shows that driver fatigue can cause brain complex network characteristics to change significantly for certain brain regions and certain rhythms. Zou *et al* (2020) proposed a method based on the combination of shortest path tree for constructing a functional brain network (denoted as CSP-FBN), which is applied to fatigued driving state recognition and neural mechanism analysis of fatigued driving. The results showed that the functional brain network constructed by the combined shortest path tree in fatigue state recognition is better than the functional brain network constructed by other methods, with the accuracy of 10-fold cross validation reaching 99.17%. Zou *et al* (2020) proposed a method based on the empirical mode decomposition (EMD) of multi-scale entropy on the recorded forehead EEG signals to recognize fatigued driving. Their results indicated that the classification recognition rate of EMD multi-scale fuzzy entropy features is up to 88.74%, which is 23.88% higher than single-scale fuzzy entropy and 5.56% higher than multi-scale fuzzy entropy.

The feature extraction methods of EEG proposed by these scholars do not consider processing EEG as a continuous and non-periodic continuous data, and fail to mine or describe the intrinsic change information and its relationship to EEG signals effectively. The concept of functional data was first proposed by the Canadian scholar Ramsay (1982). Subsequently, Ramsay and Dalzell (1991) formally put forward the concept of functional data analysis (FDA). More abundant information can be mined by the method of FDA through the analysis of derivative or differential curves, such as the analysis of first or higher derivative curves to explore the difference among curves and the dynamic change pattern within curves, etc. FDA is the branch of statistics which focuses on data that can be seen as the observed value of a functional random variable (Ferraty and Vieu 2006). However, from a practical point of view, most of the data we measure are discrete data, and there are measurement errors. The measured data may be affected by fright, cough, environmental changes, etc. Although the collected data is preprocessed, there will still be measurement errors for discrete data. Therefore, a key step in FDA is to estimate continuous function data from discrete observations.

Sangalli *et al* (2009) proposed a smoothing technique, based on free-knot splines, that was shown to provide very accurate estimates of multidimensional curves and their derivatives, even when the curves are characterized by spatial inhomogeneities. Pigoli and Sangalli (2012) proposed a wavelet-based method to obtain accurate estimates of curves in more than one dimension and of their derivatives, and applied it to multi-lead electrocardiogram records. Their results showed the method to be particularly attractive when the curves to be estimated present strongly localized features. After smoothing the discrete data function, the analysis of the previous discrete data is transformed into analysis of the function itself. The EEG data created in event-related potential (ERP) experiments have a complex high-dimensional structure. Each stimulus presentation, or trial, generates an ERP waveform which is an instance of functional data. For traditional EEG analysis, this structure needs to be simplified many times to improve the signal-to-noise ratio. By identifying the key features of ERPs and averaging them across trials, the functional and longitudinal components can be effectively disintegrated. Hasenstab *et al* 2017 proposed a multi-dimensional functional principal component analysis technologywhich does not collapse any dimension of the ERP data. The results

show that the proposed method is helpful for modeling the longitudinal trends of the ERP function, leading to novel insights into the learning patterns of children with Autism Spectrum Disorder.

Therefore, in this paper, the FDA method is applied to the feature extraction of EEG signals, which treats EEG signals as continuous data to mine the intrinsic feature information through functional differences between EEG signals in normal and fatigued states. It well reflects the continuity and internal dynamic changes of EEG signals, and has good effect on driver fatigue recognition.

2. Relevant theories and methods

2.1. Functional data analysis

FDA is a development and extension of traditional statistical analysis methods. The general process of FDA is as follows:

Given a sequence of observed time series data $y = (y_1, y_2, \ldots, y_n)$,

$$y_i = x(t_i) + \epsilon_i, \quad i = 1, 2, \ldots, n \tag{1}$$

where $x(t_i)$ denotes the value of the function $x(t)$ at t_i of the observation data sequence, while the upward ε_i denotes noise, representing disturbance factors, errors, or other exogenous factors in the observation data.

In order to estimate the value of $x(t_i)$ in formula (1), a set of basis functions $\Phi(t) = (\phi_1(t), \phi_2(t), \ldots, \phi_K(t))$ are chosen to converts discrete data into a linear combination of basis functions, i.e.,

$$x(t_i) = \sum_{k=1}^{K} c_k \phi_k(t_i), \quad i = 1, 2, \ldots, n. \tag{2}$$

After choosing the basis function, a function is determined uniquely by a set of values of the coefficient vector $C = (c_1, c_2, \ldots, c_K)^T$. The most direct way to solve the coefficient vector is the least square method, that is, to minimize the sum of squares of errors or residuals

$$SMSSE(y \mid c) = \sum_{i=1}^{n} [y_i - \sum_{k=1}^{K} c_k \phi_k(t_i)]^2. \tag{3}$$

For periodic observation data, Fourier basis is often used as the basis function, and for non-periodic observation data, B-spline basis is often used as the basis function. In this paper, B-spline basis is chosen as the basis function because EEG data is non-periodic.

2.2. B-Splines basis function

Boor *et al* 1978 gives a complete introduction to B-spline:

Given a sequence of incremental nodes $\{t_1, \ldots, t_i, \ldots, t_{i+k}, \ldots, t_n\}$, requires $t_i < t_{i+k}$, the first order spline is defined as:

$$\begin{cases} B_i^1(x) = 1, \ t_i \le x < t_{i+1}, \\ B_i^1(x) = 0, \ \text{otherwise} \end{cases}. \tag{4}$$

K-order splines are given by recursive definitions:

$$B_i^k(x) = \frac{x - t_i}{t_{i+k-1} - t_i} B_i^{k-1}(x) + \frac{t_{i+k} - x}{t_{i+k} - t_{i+1}} B_{i+1}^{k-1}(x). \tag{5}$$

So $B_i^k(x)$ is a piecewise polynomial function defined on the interval $[t_i, t_{i+k}]$, and $B_i^k(x) = 0$ outside the domain. The B-spline basis can show the local details of the function well, and is compactly supported.

3. Algorithm description and model construction

3.1. Algorithm description

Different fitting functions (the fitting function and its parameters are determined in the next chapter) can be obtained according to the EEG data of different states using FDA. There are differences in the curve shape characteristics of different fitting functions, and these differences near the extremum interval are the basis on which samples can be classified. Therefore, the EEG data in different states can be distinguished only by extracting the data with great differences in function near the extremum interval. The feature extraction algorithm of EEG data based on FDA is as follows:

Table 1. Algorithm description.

1:Initialize: $p_1 = [t_1, t_2, \ldots, t_i, \ldots, t_m] \leftarrow f_1'(t) = 0$
2: $p_2 = [t_1^*, t_2^*, \ldots, t_j^*, \ldots, t_n^*] \leftarrow f_2'(t) = 0$
/* p_1, p_2 store extreme points in normal and fatigue states, respectively, m and n are the total number of extreme points in normal and fatigue states */
3:Initialize: V, $V' = [\,]$, $d = 15$, p_3, $p_4 = [\,]$
/* d is threshold, V and V' represent the eigenvector in normal and fatigue states, p_3 and p_4 store required extreme points in normal and fatigue states respectively */
/* T is the sampling period */
4:for $i = 1$:m:
5: for $j = 1$:n:
6: if $|t_i - t_j^*| \leq 3$ T *and* $|f_1(t_i) - f_2(t_j^*)| \geq d$:
7: p_3.append(t_i)
8: p_4.append(t_j^*)
9: L = length(p_3)
10: for $k = 1$:L:
11: $\alpha_k = p_3[k]$
12: $\beta_k = p_4[k]$
13: $V_k = (f_1(\alpha_k - 3\,\mathrm{T}), f_1(\alpha_k - 2\,\mathrm{T}), f_1(\alpha_k - T), f_1(\alpha_k), f_1(\alpha_k + T), f_1(\alpha_k + 2\,\mathrm{T}), f_1(\alpha_k + 3\,\mathrm{T}))$
14: $V_k' = (f_2(\beta_k - 3\,\mathrm{T}), f_2(\beta_k - 2\,\mathrm{T}), f_2(\beta_k - T), f_2(\beta_k), f_2(\beta_k + T), f_2(\beta_k + 2\,\mathrm{T}), f_2(\beta_k + 3\,\mathrm{T}))$
15: V.extend(V_k)
16: V'.extend(V_k')
return V, V'

Input of the algorithm: EEG data fitting function $f_1(t)$ in normal state and EEG data fitting function $f_2(t)$ in fatigue state.

Output of the algorithm: the eigenvector V in normal state and the eigenvector V' in fatigue state.

Step 1: calculate the derivatives of $f_1(t), f_2(t)$ and record them as $f_1'(t), f_2'(t)$.

Step 2: let the equation $f_1'(t), f_2'(t) = 0$, and get all extreme points t_i, t_j^* $(i = 1, 2, \ldots, m; j = 1, 2, \ldots, n)$ of the curve $f_1(t), f_2(t)$.

Step 3: select the adjacent t_i, t_j^*; the proximity principle is based on the time interval between t_i and t_j^* being no more than three sampling periods, that is, $|t_i - t_j^*| \leq 3T$ (T is the sampling period) until all t_i or t_j^* are completely selected.

Step 4: for each pair of adjacent extremum points t_i, t_j^*, we can obtain a set of extremum coordinates $(t_i, f_1(t_i))$, $(t_j^*, f_2(t_j^*))$. The vertical distance of each set of extremum is calculated, the extremum coordinates $(t_i, f_1(t_i))$, $(t_j^*, f_2(t_j^*))$ which are larger than the threshold d are screened out, and the number of the groups is counted as L, so that the required extremum coordinate groups can be replaced by $(\alpha_k, f_1(\alpha_k))$, $(\beta_k, f_2(\beta_k))$, where $k = 1, 2, \ldots, L$.

Step 5: for each pair of adjacent α_k, β_k, the neighborhood $U(\alpha_k, 3T)$, $U(\beta_k, 3T)$ is the extremum interval in normal and fatigue states, respectively. Within this interval, the eigenvectors under normal and fatigue states are obtained by taking the sampling period T as the time interval. Eigenvectors in normal state $V_k = (f_1(\alpha_k - 3T), f_1(\alpha_k - 2T), f_1(\alpha_k - T), f_1(\alpha_k), f_1(\alpha_k + T), f_1(\alpha_k + 2T), f_1(\alpha_k + 3T))$ and eigenvectors in fatigue state $V_k' = (f_2(\beta_k - 3T), f_2(\beta_k - 2T), f_2(\beta_k - T), f_2(\beta_k), f_2(\beta_k + T), f_2(\beta_k + 2T), f_2(\beta_k + 3T))$, where $k = 1, 2, \ldots, L$. Therefore, the eigenvectors at normal state on a certain electrode are $V = (V_1, V_2, \ldots, V_k, \ldots, V_L)$, the eigenvectors at fatigue state on the same certain electrode are $V' = (V_1', V_2', \ldots, V_k', \ldots, V_L')$; the algorithmic description is shown in table 1.

3.2. Model construction

The model framework of the proposed method is shown in figure 1. It mainly includes the following components.

3.2.1. Feature extraction of EEG signal based on FDA

- EEG signal fitting: the EEG signals of normal and fatigue states on each electrode were fitted with cubic B-spline basis, which were recorded as $f_1^i(t), f_2^i(t)$. $f_1^i(t)$ and $f_2^i(t)$ are fitting functions at normal and fatigue states on the ith electrode.
- Feature extraction based on extrenum interval: after the fitting function is obtained, according to the algorithm description, the feature vectors V_1^i and $V_1^{i'}$ can be obtained under normal and fatigue states respectively, in which $V_1^i = (v_{11}^i, v_{12}^i, \ldots, v_{1C}^i)$ is the eigenvector of the normal state on the ith electrode

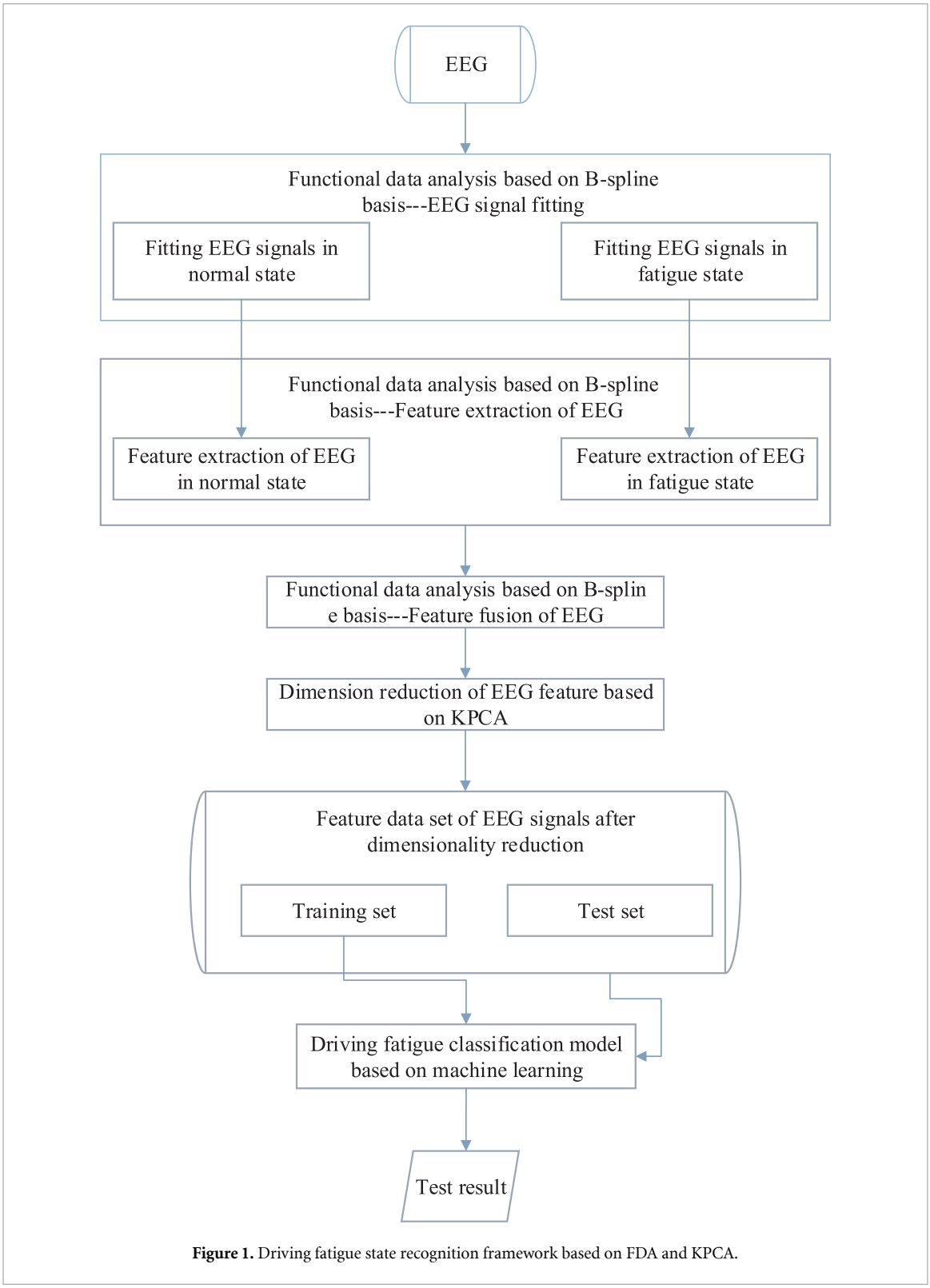


Figure 1. Driving fatigue state recognition framework based on FDA and KPCA.

and $V_1^{i'} = (v_{11}^{i'}, v_{12}^{i'}, \ldots, v_{1C}^{i'})$ is the eigenvector of the fatigue state on the ith electrode. C is the average value of the minimum feature points of the sample on the selected electrode and rounds down. This is mainly because the extreme points on each electrode are different, so the feature points of the eigenvectors on each electrode are not necessarily the same. In order to ensure the consistency of the feature points on each eigenvector, C is taken as the average of the smallest feature points on the electrode as the final feature points, that is, $C = \frac{a_1+a_2+\cdots+a_m}{m}$ (where a_i is the smallest feature point of the ith individual, $i = 1, 2, \ldots, m$; and m is the total number of individuals). If the minimum feature point on an electrode is less than C, the eigenvector of the electrode is discarded. Therefore, the eigenmatrix $M(P^*C)$ and $N(P^*C)$ under normal

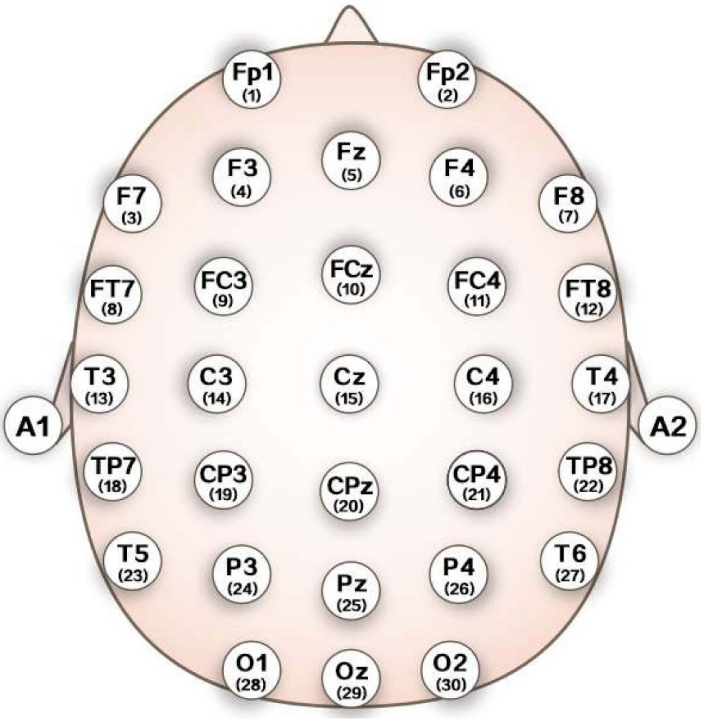


Figure 2. Position of electrodes according to International 10–20 System standards.

and fatigue states can be obtained, where P represents the final number of electrodes and C represents the feature points on each electrode.

- Feature fusion: the eigenmatrix M and N constitute the total eigenmatrix K ($2P^*C$).

3.2.2. Dimension reduction of feature based on KPCA

Considering that the eigenmatrix K is still large according to the non-linearity of EEG data, in this paper, we reduce the dimension of the eigenmatrix K by KPCA (Ye *et al* 2018). The kernel function is Gauss Radial Basis Kernel Function, and the cumulative contribution rate is set as 0.9; then we get the final eigenmatrix K'.

3.2.3. Construction of driver fatigue state recognition model based on machine learning

There are many kinds of classification models. In the following test section, two classifiers are selected to classify and identify the test data and analyze the results.

4. Model testing and result analysis

4.1. Test environment and data introduction

Test environment: the main equipment of this driving simulation experiment includes a vehicle driving simulator, a 32-electrode EEG collecting cap (the sampling frequency is 1000 Hz), and a preprocessing software (Neuroscan 3.2). The position of the electrodes is shown in figure 2. The EEG data analysis software-MATLAB R2016b (Windows 10, 64 bit) was used in this experiment.

Data collection: the experimental data of EEG signals in this article comes from the voluntary participation of 25 subjects (11 males, 14 females) who are between 18 and 30 years old, have no history of brain disease, no bad habits, and are healthy. Before the start of the experiment, after ensuring that the subjects did not drink alcohol or tea, and did not use a series of hair cosmetics, such as hair wax, that interfere with the electrode caps, the experiment administrator began to explain the experiment process and precautions, and ensured the subjects were familiar with the experimental environment. After the subjects indicated they were familiar with all the experimental procedures, operations, and rules, the preparation work was over. Next, the administrator let the subjects calm down and enter a normal state. The experiment instructor used the software to record the EEG signals of the subjects in the resting state, and the recording time was 5 min. The data obtained in this state was used as the EEG data in the resting state. The subjects then entered the simulated driving state and were required to simulate driving for at least 1 h. After subjectively experiencing symptoms such as weakness in limbs, inability to concentrate, and heavy eyes, they

Table 2. GCV under different k and λ.

k	$\lambda = 1e^{-12}$	$\lambda = 1e^{-11}$	$\lambda = 1e^{-10}$	$\lambda = 1e^{-9}$	$\lambda = 1e^{-8}$
851	1.1895	1.1883	1.1800	1.2834	3.3230
861	1.2355	1.2341	1.2247	1.3217	3.3376
871	1.1159	1.1146	1.1062	1.2226	3.3019
881	0.9792	0.9780	0.9710	1.1153	3.2722
891	1.1128	1.1114	1.1017	1.2189	3.2992

filled out a questionnaire about fatigue. After investigating the questionnaire, it could be judged whether the subject had entered a state of fatigue (Shergis *et al* 2016). After confirming that the subject had entered the fatigue state, the experimental administrator used the software to record the EEG data in the current state. The recording time wass 5 min to obtain the EEG data of the subject in the fatigue state.

Test data set: the experimental data selected in this paper are exactly the same as those of Ye *et al* (2018). The EEG data are 32 electrode, 600 s time series at a sampling rate of 1000 Hz, consisting of 300 s of normal state and 300 s of fatigue state. This paper conducted two sets of experiments. The first set took data from 25 subjects at 60 s for each person (the first 30 s in normal state and the other 30 s in fatigue state). The other set took data from 16 subjects (8 males and 8 females) at 600 s for each person (the first 300 s in normal state and the other 300 s in fatigue state) of the prefrontal electrode (FP1, FP2).

Data preprocessing: we used Neuroscan 4.5 to preprocess the collected data. The EEG signal sampling frequency was 1000 Hz, and the frequency range was 0.15–45 Hz. The main steps of data preprocessing include: drift removal, EOG removal, artifact removal, baseline correction, and filtering (Mu *et al* 2017). In view of the abnormal conditions that may appear in the experimental process, such as sneezing, coughing, being suddenly frightened, and so on, the EEG drift was removed by an artificial method. Obvious EOG, mainly vertical EOG, was deleted. We used transform-artifact rejection to remove artifacts in EEG signals. We chose the time domain (time) according to our experience, which is in the range of $\pm$50–$\pm$100 ms. For the data that does not appear in the baseline after processing, one linear correction or two baseline corrections are usually needed. The main purpose of digital filtering is to get the EEG data of the main frequency band. In this paper, 1.5–70 Hz bandpass filter was used.

4.2. Setting of the number of basis functions and smoothing factor

Considering the computing time requirement of the function smoothness, Cubic B-spline Basis is chosen in this paper when extracting EEG data based on FDA. However, the number of basis functions k and smoothing factor λ (the smoothing factor determines the level of smoothing and the speed of response to the difference between the predicted value and the actual result) need to be determined according to generalized cross validation (GCV). GCV was proposed by Golub *et al* (1972) in view of the fact that traditional cross-validation is very sensitive to calculations; especially when there are too many sample points, the method is not stable enough and prone to overfitting when minimizing cross-validation values. GCV can avoid the shortage whereby cross validation needs to be smoothed n times, and it also can avoid overfitting effectively. GCV (Golub *et al* 1972) is usually expressed as

$$GCV(\lambda) = \frac{n^{-1}||I - A(\lambda)y||^2}{\left(n^{-1}\mathrm{trace}(I - A(\lambda))\right)^2} \tag{6}$$

where I is the $n \times n$ identity, $A(\lambda)$ is a $n \times n$ influence matrix, and y is a column n-vector. In this paper, by changing the number of basis functions k and smoothing factor λ, we choose k and λ, which minimize the value of the GCV function. The test results are shown in table 2.

It can be seen from table 2 that when 881 basis functions are selected, the GCV is smallest when the smoothing factor λ is $1e^{-10}$, as is shown in figure 3.

Therefore, 881 cubic B-spline bases are used to fit the data, and the smoothing factor λ is $1e^{-10}$.

4.3. Constructing the EEG data eigenvector

We compute the eigenvectors of EEG Data based on FDA and FDA + KPCA, respectively. After determining the number of basis functions k and the smoothing factor λ, EEG data in different states can generate different fitting functions, and there are differences among different function curves. In view of these differences, we can extract the eigenvector of EEG data in different states of each electrode by algorithm description and, according to the constructed model, the eigenmatrix K' can be obtained.

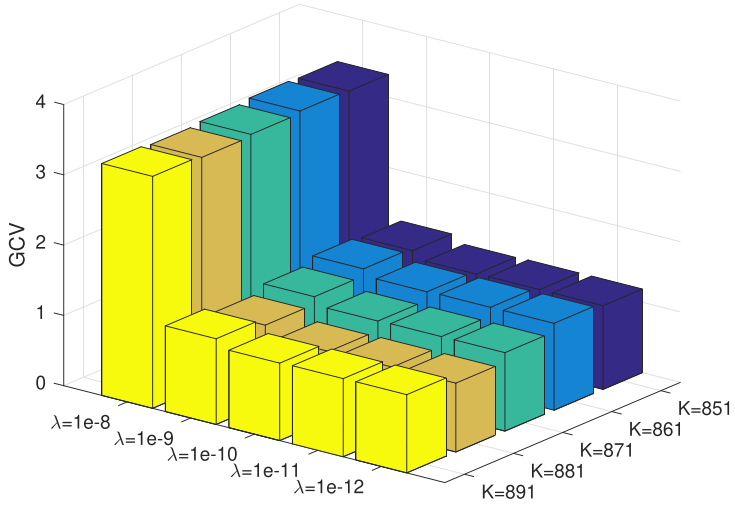


Figure 3. GCV under Different K and λ.

Table 3. Classification accuracy of two classifiers under different thresholds without KPCA.

Classifiers	$d=0$	$d=5$	$d=10$	$d=15$	$d=20$	Avg+Var
DT	93.98%	94.38%	94.31%	94.03%	94.04%	94.14% + 0.0335
RF	94.50%	94.40%	94.73%	94.58%	94.82%	94.60% + 0.0289
KNN	86.01%	88.28%	87.27%	83.55%	85.75%	86.17% + 3.1821
SVM	95.15%	96.09%	94.44%	93.94%	93.26%	94.58% + 1.1941

Table 4. Classification accuracy of two classifiers under different thresholds with KPCA.

Classifiers	$d=0$	$d=5$	$d=10$	$d=15$	$d=20$	Avg+Var
DT	99.45%	99.45%	99.54%	99.69%	99.39%	99.50% + 0.0137
RF	98.77%	99.18%	99.10%	99.30%	98.70%	99.01% + 0.0687
KNN	77.83%	83.85%	78.62%	79.15%	82.46%	80.38% + 6.8701
SVM	89.30%	91.86%	90.45%	89.43%	83.66%	88.94% + 9.7637

4.4. Model testing and result analysis

In this experiment, firstly, the FDA method is used to extract EEG data features. Then KPCA is used to reduce the dimension of features. Finally, classification and recognition are performed under four classifiers: decision tree (DT), random forest (RF), K-nearest neighbor (KNN), and support vector machines (SVM). The K of KNN is 5. The number of trees for RF is 10. Radial basis function (RBF) is selected for SVM, the penalty factor is 3, and the size of nuclear is −7. We test the model with a 10-fold cross-validation and calculate the average (Avg) accuracy rate. The accuracy results under four classifiers are shown in tables 3 and 4.

The experimental results show that the classification accuracy variances (Var) of KNN and SVM are higher than those of DT and RF, and the average accuracy rates (Avg) of KNN and SVM are lower than or close to those of DT and RF, both with KPCA and without KPCA. The stability of KNN and SVM is poor in this experiment. In the case of KPCA, the classification accuracy rates of DT and RF are significantly higher than that without KPCA. The Avg on DT can reach 99.50%, and the Var is 0.0137 in the case of KPCA. When the threshold is 15, the accuracy both on DT and RF are the highest with KPCA. Therefore, the DT and RF are chosen as the classifier and the optimal threshold is 15 in this paper, as shown in figures 4–6.

In order to confirm the stability of this method and consider the physiological differences between males and females, which lead to changes in EEG signals, we selected the EEG data of 16 subjects (8 males and 8 females) in normal and fatigue states of the prefrontal electrode (FP1, FP2). The reason why we chose the prefrontal electrode is mainly based on the following two points: first, the EEG signal in this area is less affected by hair and other factors, and the collected data is less interfered with by noise; second, it is convenient to wear the EEG acquisition device in this area. We choose the classifier DT with the highest average classification accuracy for this experiment. The experimental results are shown in table 5.

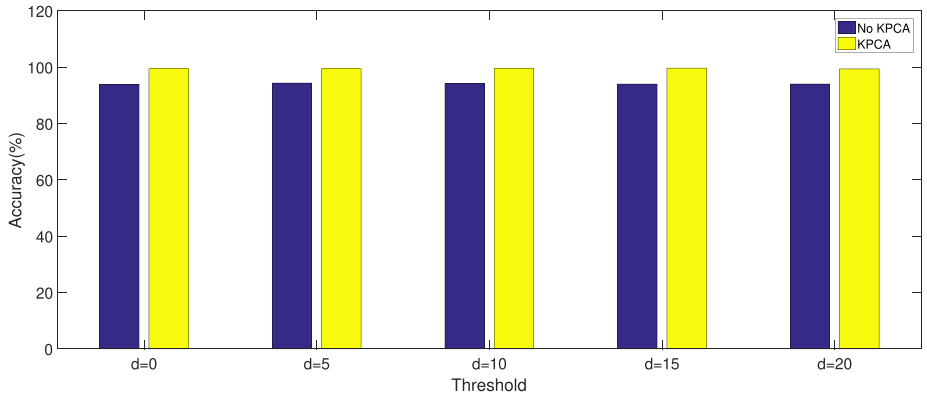


Figure 4. Classification effect of different thresholds on DT.

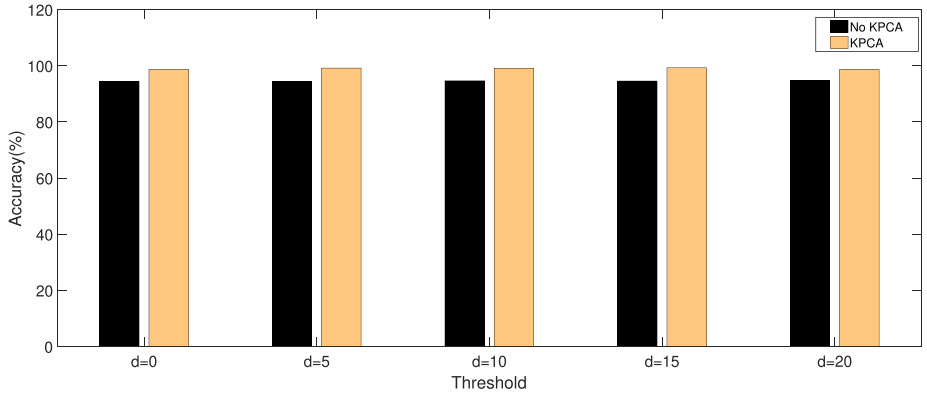


Figure 5. Classification effect of different thresholds on RF.

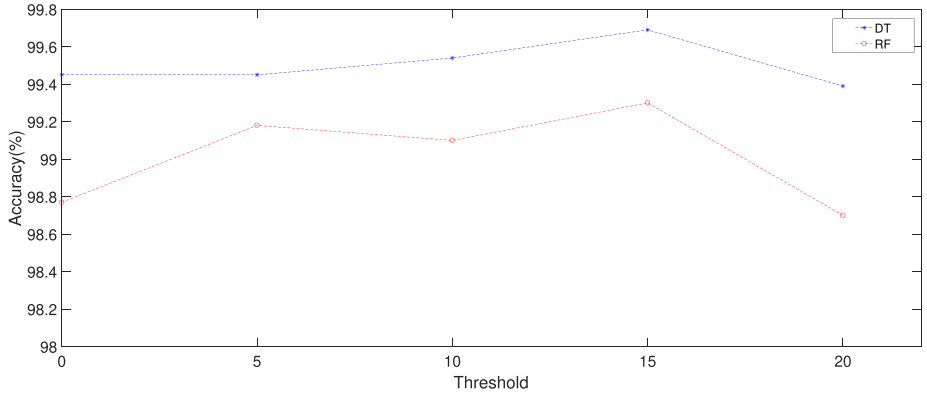


Figure 6. Comparison of accuracy between DT and RF.

Table 5. Accuracy of driver fatigue detection in 8 males and 8 females using forehead electrode based on FDA and KPCA under DT.

Gender	$d=0$	$d=5$	$d=10$	$d=15$	$d=20$	Avg+Var
Male	98.71%	99.49%	99.20%	99.02%	98.51%	98.99% + 0.1509
Female	99.45%	99.50%	98.89%	99.06%	99.22%	99.22% + 0.0664

The experimental results show that although the physiological differences between males and females have an impact on the driver fatigue detection rate, the impact is small. At the same time, the results show that the method proposed in this paper is relatively stable, and the driver fatigue detection rate is high in both male and female individuals, with less fluctuation in accuracy.

Table 6. Algorithm comparison.

Publications	Feature extraction method	EEG electrode	Accuracy
Hu (2017)	Fuzzy entropy	CP4	96.60%
Ye *et al* (2018)	Sample entropy	30 electrodes	98.33%
Luo *et al* (2019)	Adaptive multi-scale entropy	Fp1, Fp2	95.37%
This paper	Functional data analysis	30 electrodes	99.50%

Hu 2017 used the method of fuzzy entropy to extract EEG features and, only using the CP4 electrode, their accuracy was 96.60%. Ye *et al* (2018) used the method of sample entropy to extract EEG features, and then used KPCA to reduce dimension; their accuracy was 98.33%. Luo *et al* 2019 used the method of adaptive multi-scale entropy to extract EEG features; while they only used the FP1 and FP2 electrodes, their accuracy was 95.37%, as shown in table 6. However, the feature-extraction method of EEG proposed by these scholars basically does not consider processing EEG as a continuous and non-periodic continuous data, and fails to mine or describe the intrinsic change information and its relationship to EEG signals effectively. In this paper, in which the method of FDA is applied to extract EEG features and KPCA is used to reduce dimension, the accuracy can reach 99.50%. This method well reflects the continuity and internal dynamic changes of EEG signals, has good recognition efficacy in driver fatigue recognition, and is relatively stable.

5. Discussion and conclusions

In this paper, we propose a novel method combining FDA and KPCA for the detection of driver fatigue. The stability and application convenience of the proposed method are discussed. Because of the non-linear characteristics of EEG signals, we propose a feature-extraction method based on FDA, in which the discrete data are functionalized to express the continuity of human EEG data. Then, according to the difference of functional curves near the extremum, the feature intervals and eigenvectors are determined. The results show that this method can extract EEG features well. Because of the high-dimensional feature of functional EEG, KPCA is used to reduce the dimension of the eigenmatrix extracted from FDA. Our results show that this method achieves good classification results on RF and DT classifiers. Finally, in order to verify the stability and application convenience of the proposed method, we applied it to the data collected from forehead-mounted electrodes (FP1, FP2) on men and women. The results show that the proposed method has good stability. To sum up, the method proposed in this paper reflects the continuity and internal dynamic changes of EEG signals, and has good efficacy in driver fatigue recognition. At the same time, compared with other methods, this method also gives more stable classification results. However, this study still has the following limitations: (1) we have not found a specific method to determine the threshold, and choosing different thresholds to obtain the optimal threshold has some randomness. In the follow-up work, we will do more in-depth research on the method of obtaining the threshold and propose an adaptive threshold determination method. (2) When choosing the extremum interval and calculating the eigenvectors, we neglect the influence of different time intervals on the experiment. We plan to determine the extreme value interval and calculate feature vectors by increasing different time intervals in subsequent experiments, and conduct in-depth comparison tests and result analysis. (3) The number of experimental samples is not enough, and the generalization of the method needs to be improved. The experimental results may have certain limitations because of the insufficient datasets. We plan to collect more experimental samples in the follow-up work, and broaden the age range of experimental samples to verify and improve the stability and application convenience of the proposed method.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Nos. 61841201, 81460769, 61662045 and 61762045). Thanks to Jianfeng Hu's team for providing EEG experiment data.

Ethical statement

This study meets the ethical principles of the Declaration of Helsinki and the current legal requirements. Ethical approval for this work was obtained from the Academic Ethics Committee of the Jiangxi University of Technology. All subjects signed an informed consent.

References

Acioglu A and Erçelebi E 2016 Real time eye detection algorithm for PERCLOS calculation *24th IEEE Signal Processing and Communication Conf. (SIU)* pp 1641–4

Beige Y, Qiu T, Bai X and Liu P 2018 Research on recognition method of driving fatigue state based on sample entropy and kernel principal component analysis *Entropy* **20** 701

Cao Z and Lin C-T 2017 Inherent fuzzy entropy for the improvement of EEG complexity evaluation *IEEE Trans. Fuzzy Syst.* **26** 1032–5

Chai R, Naik G R, Nguyen T N, Ling S H, Tran Y, Craig A and Nguyen H T 2016 Driver fatigue classification with independent component by entropy rate bound minimization analysis in an EEG-based system *IEEE J. Biomedical Health Inform.* **21** 715–24

Correa A G, Orosco L and Laciar E 2014 Automatic detection of drowsiness in EEG records based on multimodal analysis *Med. Eng. Phys.* **36** 244–9

De Boor C 1978 *A Practical Guide to Splines* (New York: Springer) vol 27

Dimitrakopoulos G N, Kakkos I, Dai Z, Wang H, Sgarbas K, Thakor N, Bezerianos A and Sun Y 2018 Functional connectivity analysis of mental fatigue reveals different network topological alterations between driving and vigilance tasks *IEEE Trans. Neural Systems Rehabilitation Eng.* **26** 740–9

Ferraty F and Vieu P 2006 *Nonparametric Functional Data Analysis: Theory and Practice* (Berlin: Springer) (https://doi.org/10.1007/0-387-36620-2)

George S T, Subathra M S P, Sairamya N J, Susmitha L and Premkumar M J 2020 Classification of epileptic EEG signals using PSO based artificial neural network and tunable-q wavelet transform *Biocybern. Biomed. Eng.* **40** 709–28

Golub G H, Heath M and Wahba G 1979 Generalized cross-validation as a method for choosing a good ridge parameter *Technometrics* **21** 215–23

Hajinoroozi M, Mao Z, Jung T-P, Lin C-T and Huang Y 2016 EEG-based prediction of driver's cognitive performance by deep convolutional neural network *Signal Process. Image Commun.* **47** 549–55

Han C, Sun X, Yang Y, Che Y and Qin Y 2019 Brain complex network characteristic analysis of fatigue during simulated driving based on electroencephalogram signals *Entropy* **21** 353

Hasenstab K, Scheffler A, Telesca D, Sugar C A, Jeste S, DiStefano C and Şentürk D 2017 A multi-dimensional functional principal components analysis of EEG data *Biometrics* **73** 999–1009

Hoddes E, Zarcone V, Smythe H, Phillips R and Dement W C 1973 Quantification of sleepiness: a new approach *Psychophysiology* **10** 431–6

Hu J 2017 Comparison of different features and classifiers for driver fatigue detection based on a single EEG channel *Computat. Math. Methods Med.* **2017** 5109530

Janssen K C, Phillipson S, O'Connor J and Johns M W 2017 Validation of the Epworth sleepiness scale for children and adolescents using Rasch analysis *Sleep Med.* **33** 30–5

Jia-Xin M, Shi Li-C and Bao-Liang L 2014 An EOG-based vigilance estimation method applied for driver fatigue detection *Neurosci. Biomed. Eng.* **2** 41–51

Luo H, Qiu T, Liu C and Huang P 2019 Research on fatigue driving detection using forehead EEG based on adaptive multi-scale entropy *Biomed. Signal Process. Control* **51** 50–8

Mooij A H, Frauscher B, Amiri M, Otte W M and Gotman J 2016 Differentiating epileptic from non-epileptic high frequency intracerebral EEG signals with measures of wavelet entropy *Clin. Neurophysiol.* **127** 3529–36

Pigoli D and Sangalli L M 2012 Wavelets in functional data analysis: estimation of multidimensional curves and their derivatives *Computat. Stat. Data Anal.* **56** 1482–98

Polat H, Alulu M U and Zerdem M S 2020 Evaluation of potential auras in generalized epilepsy from EEG signals using deep convolutional neural networks and time-frequency representation *Biomed. Tech.* **65** 379–91

Rajendra Acharya U, Hagiwara Y, Deshpande S N, Suren S, Wei Koh J E, Lih Oh S, Arunkumar N, Ciaccio E J and Lim C M 2019 Characterization of focal EEG signals: a review *Future Gener. Comput. Syst.* **91** 290–9

Ramos-Aguilar R, Olvera-López J A, Olmos-Pineda I and Snchez-Urrieta S 2020 Feature extraction from EEG spectrograms for epileptic seizure detection *Pattern Recognit. Lett.* **13** 202–9

Ramsay J O 1982 When the data are functions *Psychometrika* **47** 379–96

Ramsay J O and Dalzell C J 1991 Some tools for functional data analysis *J. R. Stat. Soc.* B **53** 539–61

Sangalli L M, Secchi P, Vantini S and Veneziani A 2009 Efficient estimation of three-dimensional curves and their derivatives by free-knot regression splines, applied to the analysis of inner carotid artery centrelines *J. R. Stat. Soc.* C **58** 285–306

Sharif M S, Naeem U, Islam S and Karami A 2018 Functional connectivity evaluation for infant EEG signals based on artificial neural network *Proc. of SAI Intelligent Conf.* (Berlin: Springer) pp 426–38

Shergis J L, Xiaojia N, Jackson M L, Zhang A L, Guo X, Yan Li, Lu C and Xue C C 2016 A systematic review of acupuncture for sleep quality in people with insomnia *Complement. Ther. Med.* **26** 11–20

Shi X and Yang C 2017 Research on driver fatigue state recognition technology based on vehicle trajectory characteristics *China New Commun.* **19** 158–60

Song Y and Zhang J 2016 Discriminating preictal and interictal brain states in intracranial EEG by sample entropy and extreme learning machine *J. Neurosci. Methods* **257** 45–54

Wang C-Y, Yu S-C, Lin Y-C and Lin Y-H 2016 Fatigue detection system based on indirect-contact ECG measurement *2016 Int. Conf. on Advanced Robotics and Intelligent Systems (ARIS)* (Piscataway, NJ: IEEE) pp 1–1

Wang H, Dragomir A, Abbasi N I, Junhua Li, Thakor N V and Bezerianos A 2018 A novel real-time driving fatigue detection system based on wireless dry EEG *Cogn. Neurodyn.* **12** 365–76

Xiaoling Li, Jiang Y, Hong J, Dong Y and Yao L 2016 Estimation of cognitive workload by approximate entropy of EEG *J. Mech. Med. Biol.* **16** 1650077

Zhao C, Zhao M, Yang Y, Gao J, Rao N and Lin P 2016 The reorganization of human brain networks modulated by driving mental fatigue *IEEE J. Biomed. Health Inf.* **21** 743–55

Zhendong M, Hu J and Min J 2017 Driver fatigue detection system using electroencephalography signals based on combined entropy features *App. Sci.* **7** 150

Zou S, Qiu T, Huang P, Bai X and Liu C 2020 Constructing multi-scale entropy based on the empirical mode decomposition (EMD) and its application in recognizing driving fatigue *J. Neurosci. Methods* **341** 108691

Zou S, Qiu T, Huang P, Luo H and Bai X 2020 The functional brain network based on the combination of shortest path tree and its application in fatigue driving state recognition and analysis of the neural mechanism of fatigue driving *Biomed. Signal Process. Control* **62** 102129

Zuojin Li, Li S E, Renjie Li, Cheng B and Shi J 2017 Online detection of driver fatigue using steering wheel angles for real driving conditions *Sensors* **17** 495