Research Article

Collaborative Sleep Electroencephalogram Data Analysis Based on Improved Empirical Mode Decomposition and Clustering Algorithm

Xiangwei Zheng ¹, Xiaochun Yin,² Xuexiao Shao ³, Yalin Li,⁴ and Xiaomei Yu ¹

¹School of Information Science and Engineering, Shandong Normal University, Ji'nan, China

²Facility Horticulture Laboratory of Universities in Shandong, WeiFang University of Science & Technology, ShouGuang, China

³School of Information Science and Engineering, Lanzhou University, Lanzhou, China

⁴School of Information Engineering, Shandong Management University,
Key Laboratory of TCM Data Cloud Service in Universities of Shandong, Jinan 250357, China

Correspondence should be addressed to Xuexiao Shao; 1377699626@qq.com and Xiaomei Yu; yxm0708@126.com

Received 23 January 2020; Accepted 28 March 2020; Published 13 June 2020

Guest Editor: Xuyun Zhang

Copyright © 2020 Xiangwei Zheng et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Sleep-related diseases seriously affect the life quality of patients. Sleep stage classification (or sleep staging), which studies the human sleep process and classifies the sleep stages, is an important reference to the diagnosis and study of sleep disorders. Many scholars have conducted a series of sleep staging studies, but the correlation between different sleep stages and the accuracy of classification still needs to be improved. Therefore, this paper proposes an automatic sleep stage classification based on EEG. By constructing an improved empirical mode decomposition and K-means experimental model, the concept of “frequency-domain correlation coefficient” is defined. In the process of feature extraction, the feature vector with the best correlation in the time-frequency domain is selected. Extraction and classification of EEG features are realized based on the K-means clustering algorithm. Experimental results demonstrate that the classification accuracy is significantly improved, and our proposed algorithm has a positive impact on sleep staging compared with other algorithms.

1. Introduction

Sleep is of extraordinary significance to human beings and is closely related to people's life. It plays an important role in the maintenance of human body functions because it can enhance assimilation and reduce the level of alienation [1, 2]. Sleep is a state of rapid reversibility characterized by loss of consciousness and diminished response to external stimuli [3–5]. For humans, one of the principal causes of medical problems is sleep-related diseases, which seriously affect the life quality of patients. The purpose of sleep staging is to classify sleep stages, which is essential for sleep studies and the diagnosis of sleep disorders. Traditionally, according to the Rechtschaffen and Kales recommendations or the new guidelines developed by the American Academy of Sleep

Medicine (AASM), experts manually analyze night polysomnography (PSG) records to perform visual scoring. Later, based on the improvement of the AASM rules, the S3 and S4 phases were merged into slow-wave sleep (SS), and the sleep was divided into five stages: the W, S1, S2, SS, and REM periods [6, 7].

An electroencephalograph (EEG) is a record that reflects the regular electric action of brain cell groups and it contains a large amount of physiological and pathological information. It is helpful for clinicians to improve the reliability and accuracy of diagnosis and detection of neurological injury in the brain [8–10]. At the same time, it provides an effective method for the diagnosis of brain diseases. An EEG signal is a waveform that contains a variety of frequency components, and it is usually divided into β waves (13–40 Hz), α waves

(8–13 Hz), θ waves (4–7 Hz), and δ waves (0–4 Hz). Therefore, sleep stages can be classified based on different brainwave frequencies and data characteristics in the EEG signal. The manual classification of 8-hour PSG recordings (whole records) takes approximately 2 to 4 hours. Moreover, manual marking results have a strong subjective consciousness, which easily affects the classification accuracy. Therefore, the study of automatic sleep stage classification is imperative [11]. Through the analysis on PSG records, automatic sleep stage classification (ASSC) can be achieved with computers, thus solving the problem of time-consuming and laborious manual marking [12].

Considering the nonlinear and unsteady timing complexity of EEG data signals, we propose an automatic sleep stage classification method based on EEG. The improved complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) algorithm is used to extract the features of EEG data and calculate the IMF components. By calculating the frequency-domain correlation coefficients of the IMF components, an appropriate number of IMF components are selected, and a new feature vector is formed as the input for the next stage classifier. The clustering of extracted and selected features is performed by an improved K-means algorithm. ASSC is finally achieved based on the selection of relevant distances and cluster centers.

The main contributions of this study are as follows:

- (1) An improved CEEMDAN incorporating the frequency-domain correlation coefficient is proposed to extract EEG feature and calculate the IMF components. By calculating the frequency-domain correlation coefficients of the IMF components, an appropriate number of IMF components are selected and the new feature vector is formed.
- (2) An improved K-means clustering algorithm with density based on correlation coefficient is proposed. The correlation coefficient is first defined as the distance metric based on the temporal and spatial correlation of specific time series data. Then, the density is used to select the clustering center of K-means and clustering centers are iteratively updated by calculating the average of all the points.
- (3) For the nonlinear and unsteady timing complexity of EEG data signals, we propose an automatic sleep stage classification method based on the improved empirical mode decomposition and K-means clustering algorithm. Through the innovative improvement of the feature extraction method and the classifier algorithm, the classification accuracy of ASSC is obviously improved, and better experimental results are obtained.

The remainder is organized as follows. Section 2 reviews on the state-of-the-art automatic sleep stage classification and Section 3 introduces the EMD and its variants briefly. Section 4 presents the proposed ASSC based on improved CEEMDAN and K-means in detail including overall framework, improved CEEMDAN incorporating the frequency-domain correlation coefficient, and improved

K-means clustering algorithm with density based on correlation coefficient. Section 5 analyzes the experimental setting and results on sleep staging accuracy and clustering effectiveness and also discusses the finding in this study. At last, Section 6 concludes the paper and discusses future opportunity of the research.

2. Related Work

To further improve the efficiency and accuracy of ASSC, researchers have conducted a large number of experimental studies to achieve better ASSC results by improving the relevant algorithms [13–15]. Among them, feature selection is applied to enhance the ability of classifying the training data and can improve the efficiency and accuracy of data classification [16–18]. Recently, feature analysis and extraction methods have been increasingly studied, and classical or modern signal processing methods have been adopted to analyze EEG data. For example, Anderson et al. [19] used an autoregressive (AR) model to extract features of EEG signals and used two- and three-layer forward neural network to perform 10-fold cross-validation on 4 subjects with 5 cognition items. To achieve better results, Yang et al. [20] proposed an EEG signal feature extraction method based on wavelet packet decomposition to classify two different thinking activities. Fell et al. [21] conducted a comparative study of frequency domain and nonlinear methods and divided the sleep process into four stages: S1, S2, SS, and REM. Extracted frequency-domain features included power, spectral edge, and D2. However, these studies ignore the changes in the timing characteristics of EEG and local feature signal along with self-adaptation. Fortunately, the empirical mode decomposition (EMD) algorithm solves the problem of nonstationary signals in EEG data [22], and data feature extraction can be solved adaptively by decomposing the data into an intrinsic mode function (IMF). Therefore, scholars have conducted extensive studies on applying EMD and EMD algorithms into signal filtering and detection, fault analysis, and physiological signal processing. These results are difficult to achieve by traditional methods such as Fourier transform and wavelet transform [23].

Due to the good self-adaptability of the EMD algorithm, researchers have conducted extensive studies on EMD. MuFeng and Yuyu [24] proposed an improved EMD filtering algorithm and used fast Fourier transformation (FFT) to perform simple spectrum analysis on the signal. If there are high-frequency noises in the signal, then the first-order IMF component decomposed by EMD is processed and can achieve a better filtering effect. Zhang et al. [25] proposed a fast wavelet transform (FWT), which can achieve high computational speed and improve computational accuracy at the same time. Wu and Huang [26] proposed a new ensemble empirical mode decomposition (EEMD) algorithm in 2005. It is an improved EMD algorithm that effectively solves the EMD mixing phenomenon. Later, CEEMDAN algorithm was proposed with added adaptive noise so that the modal effect is further reduced. Compared with the former algorithms, it has better convergence and

Hassan and Bhuiyan [27] applied it to the analysis of EEG data to achieve ASSC.

At present, there are many classification methods applied to EEG signals, including clustering, SVM, neural network, and decision tree [28, 29]. The K-means algorithm was proposed by Macqueen in 1967 [30]. It is a numerical clustering algorithm and requires the simultaneous extraction of N features. The original K-means is a distance-based iterative clustering algorithm. The advantages are that it is fast and simple and has high efficiency. This paper focuses on the research of clustering algorithms and hopes to apply it to the research of sleep staging through the improved K-means algorithm. For example, Günes et al. [31] proposed the combined structure of feature weighting and a C4.5 decision tree based on K-means clustering in sleep stage classification. The clustering algorithm on the ASSC not only solves the disadvantages of time-consuming and laborious manual marks but can also effectively improve the efficiency of staging operations and the accuracy of ASSC [32].

3. EMD and Its Variants

EMD is a novel and adaptive signal time-frequency processing method proposed by Huang et al. [33] in 1998. It is especially suitable for the analysis of nonlinear nonstationary signals. In 1999, Huang et al. [34] improved EMD and introduced Hilbert spectral analysis to improve the data processing capability. It was considered to be a breakthrough in linear and steady-state spectral analysis based on the Fourier transform in 2000. EMD aims to generate a highly localized time-frequency estimation of a signal in a data-driven fashion by decomposing it into a finite sum of IMF or modes. Each mode must satisfy two conditions:

- (1) The number of extrema and the number of zero crossings must be the same or different by at most one
- (2) At any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero

For input signal, EMD iteratively decomposes an N -point EEG epoch into amplitude and frequency-modulated IMFs according to the following steps:

Step 1: initializing $k = 1$ and $l = 1$.

Step 2: identifying the local maxima and minima of input EEG data x .

Step 3: obtaining the envelope of local maxima $v_{\max}$ and local minima $v_{\min}$ using cubic spline interpolation.

Step 4: generating the local mean curve m with the upper and lower envelopes:

$$m_k = \frac{v_{\max} + v_{\min}}{2}. \quad (1)$$

Step 5: computing h_k by subtracting the local mean curve from x :

$$h_k = x - m_k. \quad (2)$$

Step 6: if h_k satisfies the two conditions of IMF, then IMF _{l} is obtained; otherwise, set $x = h_k$, $k = k + 1$, and go to step 2, repeating steps 2–5 until h_k satisfies the two conditions of IMF, and finally, IMF _{l} is obtained as follows:

$$\text{IMF}_l = h_k - m_k. \quad (3)$$

Step 7: setting $c_l = \text{IMF}_l$ as the current mode.

Step 8: finding the residue, $r_l = x - c_l$, $x = r_l$, and $l = l + 1$. Steps 2–8 are known as sifting.

Step 9: repeating steps 2–8 to find the rest of the IMFs $c_2, c_3, \dots, c_L$.

Thus, the input signal can be decomposed into L IMFs until the residue becomes a monotonic function such that further extraction of an IMF is not possible. The input x can be reconstructed from all the IMFs as follows:

$$x = \sum_{j=1}^L c_j + r_L, \quad (4)$$

where r_L is the residue of the L -th iteration. EMD and its variants such as bivariate EMD and multivariate EMD are widely used for EEG and other physiological signal analysis [35]. However, EMD and its extensions suffer from a mode mixing problem. Later, to eliminate the problem of mode mixing, a noise-aided adaptive data analysis and extension method based on EMD was proposed and named EEMD [36]. In EEMD, low-level random noise is added as input to the EMD decomposition process. The process of implementing EMD on the modified signal is called a test. The test was repeated many times to obtain the final pattern. The biggest improvement of EEMD is that it incorporates Gaussian white noise on the basis of EMD as follows:

$$x^i(n) = x(n) + w^i(n), \quad (5)$$

where $w^i(n)$ with $i = 1, 2, \dots, I$ represents different realizations of white Gaussian noise.

Although EMD is data-driven, it is affected by the problem of mode mixing, which leads to different oscillations in the same mode, or similar oscillations in different modes. The structure of this algorithm is too simple to solve the problem. Although EEMD effectively solves the problem of pattern blending, EEMD decomposition produces residual noise and is also computationally expensive.

4. ASSC Based on Improved CEEMDAN and K-Means

4.1. Overall Framework. We design an improved CEEMDAN and K-means combination algorithms to implement ASSC as shown in Figure 1. On the basis of time-domain feature extraction, we further add frequency-domain feature calculations and select the most effective IMF components.

FIGURE 1: Automatic sleep stage classification based on EEG.

This approach eventually achieves the ASSC through K-means clustering.

The process of the proposed approach is as follows:

- (1) Input: data preprocessing removes artifacts and interference noise by the wavelet denoising, and EEG data are obtained as experimental sample data.
- (2) Feature extraction: we use improved CEEMDAN algorithm to extract features from the sample data. Through the calculation of the time-domain correlation of the EEG signal, the sample data are subjected to EMD decomposition to obtain the IMF component. Then, we apply an iterative calculation to extract features from the EEG signals.
- (3) Feature selection: on the basis of feature extraction and IMF component acquisition, we transform each IMF component into its corresponding frequency-domain feature through FFT, and its “frequency-domain correlation coefficient” is obtained through empirical selection and effective data analysis methods; then, IMF components with high-frequency-domain correlation are selected, and all IMF components are reconstructed.
- (4) Classifier: the reconstructed EEG is selected as input to achieve sleep staging based on the improved K-means clustering algorithm.
- (5) Output: the clustering result is directly output, and then the result is compared with manual markers.

4.2. Improved CEEMDAN Algorithm. CEEMDAN algorithm is improved from the original EMD algorithm. The added Gaussian white noise is subjected to EMD decomposition to obtain the IMF component based on the

bootstrap aggregation method, and then, iterative calculations are performed to realize the decomposition of the dataset. It is worth mentioning that the performance of the traditional time-frequency transform method based on wavelet transformation is influenced by choice of the best basis function. CEEMDAN is data-driven and does not require prebase functionality, which makes it an attractive option for dealing with highly nonlinear and nonstationary signals such as sleep EEG signals [37–39].

In this study, an improved CEEMDAN incorporating the frequency-domain correlation coefficient is proposed in which the IMF’s frequency-domain correlation coefficient is redefined and the correlation of each component of the IMF with the original signal in the frequency domain is embodied. Then, an appropriate number of IMF components are selected and the new feature vector is formed by calculating the frequency-domain correlation coefficients of the IMF components. Improved CEEMDAN algorithm can be described as follows:

An operator $E_j(\cdot)$ is defined that produces the j -th mode of EMD.

Step 1: calculate $X_i(n)$:

$$X_i(n) = x(n) + \varepsilon_0 w^i(n), \quad (6)$$

where $X_i(n)$ is obtained by adaptively adding white noise sequence to $x(n)$, $w^i(n)$ with $i = 1, 2, \dots, I$ represents different realizations of white Gaussian noise, and ε_0 is the standard deviation of white Gaussian noise.

Step 2: decompose the above signals by EMD to obtain their first modes.

Step 3: compute the first mode $\widehat{\text{IMF}}_1$ of CEEMDAN with

$$\widetilde{\text{IMF}}_1 = \frac{1}{I} \sum_{i=1}^I \text{IMF}_1^i(n) = \overline{\text{IMF}_k(n)}. \quad (7)$$

Step 4: obtain the first residue:

$$r_1(n) = x(n) - \widetilde{\text{IMF}}_1(n). \quad (8)$$

Step 5: decompose realizations $r_1(n) + \varepsilon_1 E_1(w^i(n))$ with $i = 1, 2, \dots, I$ up to their first EMD mode. ε_k ($k = 1$ for this stage) is the standard deviation of white Gaussian noise at k -th stage. $\widetilde{\text{IMF}}_2(n)$ can be calculated as follows:

$$\widetilde{\text{IMF}}_2(n) = \frac{1}{I} \sum_{i=1}^I E_1(r_1(n)) + \varepsilon_1 E_1(w^i(n)). \quad (9)$$

Step 6: compute the k -th residue for $k = 2, 3, \dots, K$:

$$r_k(n) = r_{k-1}(n) - \widetilde{\text{IMF}}_k(n). \quad (10)$$

Step 7: decompose realizations $r_1(n) + \varepsilon_1 E_1(w^i(n))$ with $i = 1, 2, \dots, I$ up to their first EMD mode and define the $(k+1)$ th mode as follows:

$$\widetilde{\text{IMF}}_{k+1}(n) = \frac{1}{I} \sum_{i=1}^I E_1(r_k(n)) + \varepsilon_k E_k(w^i(n)). \quad (11)$$

Step 8: go to step 6 for the next k .

Steps 6 to 8 are repeated until the residue becomes a monotonic function such that further extraction of an IMF is not possible. K is the total number of modes and $r_k(n)$ is the final residue. Then, $k+1$ EMD modes are obtained through iterative calculation as shown in formula (11). At this point, we have multiple intrinsic modal functions (IMFs), that is, $\text{IMF}_1, \text{IMF}_2, \dots, \text{IMF}_K$.

Step 9: for each IMF component, fast Fourier transformation (FFT) is performed to transform the EEG signal from the time domain to the frequency domain. Its definition is as follows:

$$\begin{aligned} x(k) &= \sum_{n=0}^{N-1} x(n) W_N^{nk}, \quad k = 0, \dots, N-1, \\ x(n) &= \frac{1}{N} \sum_{k=0}^{N-1} x(k) W_N^{-nk}, \quad n = 0, \dots, N-1, \end{aligned} \quad (12)$$

where $W_N = e^{-j(2\pi/N)}$.

The frequency spectrum analysis of the IMF component is performed by using FFT, and the corresponding frequency-domain form in the time-domain feature is calculated to obtain the frequency and amplitude.

In the experiment, we directly call the FFT function in MATLAB. Among them, we set the sampling frequency

$f_s = 100$ Hz and sample time $t = N/f_s$; N is the data length. The amplitude and frequency are shown in Figure 2.

Step 10: to obtain effective feature components after CEEMDAN decomposition, we redefine the IMF's frequency-domain correlation coefficient index by the improvement of the concept of "frequency-domain correlation coefficient" proposed in [40]:

$$\rho_{x, \text{IMF}_i} = \left| \frac{E[(\text{IMF}_i(f) - \mu_{\text{IMF}_i})(x(f) - \mu_x)]}{\sigma_{\text{IMF}_i} \sigma_x} \right|. \quad (13)$$

In formula (13), $\text{IMF}_i(f)$ and $x(f)$ represent the frequency-domain form of $\widetilde{\text{IMF}}_k(n)$ and $x(n)$ in formula (7) respectively; μ_{IMF_i} and μ_x are the frequency-domain mean of $\widetilde{\text{IMF}}_k(n)$ and $x(n)$; σ_{IMF_i} and σ_x represent the standard deviation of the frequency domain.

The frequency-domain correlation coefficient of the IMF reflects the correlation of each component of the IMF with the original signal in the frequency domain. Namely, $\rho_{x, \text{IMF}_i} \in [0, 1]$ indicates the correlation between the IMF components and the original signal in the frequency domain. Therefore, according to the value of ρ_{x, IMF_i} , the most effective IMF component can be selected from the multiple IMF components obtained by the decomposition, and the IMF can be recorded according to the frequency band from high to low: $\text{IMF} = \{\text{IMF}_1, \text{IMF}_2, \dots, \text{IMF}_K\}$. The specific details of IMF component screening are described in the experimental section and in Figure 3.

Step 11: the input $x(n)$ can be reconstructed from all the IMFs as follows:

$$x(n) = \sum_{k=1}^K \widetilde{\text{IMF}}_k(n) + r_k(n). \quad (14)$$

We calculate the mean (μ), variance (σ^2), skewness (γ), and kurtosis (κ) for the reconstructed $x(n)$, as defined mathematically in Table 1.

IMF time-frequency-domain feature vector is calculated for N -band EEG data, and the feature vector is composed as follows:

$$S = \{S_{x(1)}, S_{x(2)}, \dots, S_{x(n)}\}. \quad (15)$$

The input $x(n)$ can be reconstructed from an N -dimensional vector feature set into a new $x(n)$ set as input to the next stage classifier.

4.3. Improved K-Means Algorithm. K-means is a basic clustering algorithm; however, it is also limited by some practical applications and its own mechanisms. First, K must be given in advance and the choice of K value is very difficult to estimate. It is necessary to determine an initial partition based on the initial clustering center. It can be seen from K-means algorithm framework that the time complexity of the algorithm is large, the application of the time series data

FIGURE 2: The amplitude and corresponding frequency domain.

FIGURE 3: Frequency-domain correlation coefficient of IMF frequency.

TABLE 1: The feature vector used in this work.

Name	Mathematical formulation
Mean (μ)	$\mu = \frac{1}{N} \sum_{i=1}^N x_i$
Variance (σ^2)	$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$
Skewness (γ)	$\gamma = \frac{1}{N} \sum_{i=1}^N \left(\frac{x_i - \mu}{\sigma} \right)^3$
Kurtosis (κ)	$\kappa = \frac{1}{N} \sum_{i=1}^N \left(\frac{x_i - \mu}{\sigma} \right)^4$

is vulnerable to outliers, and the uncertainty of K value also leads to a decline in the quality of the cluster [41]. To overcome the disadvantage of K-means converging to the local problem and improve the influence of K value on the clustering quality due to inaccuracy, we improve the K-means algorithm as follows.

In this paper, the initial center of clustering is selected by a density concept based on the correlation coefficient and correlation distance. For EEG sample data, there is a certain

correlation between the time series of EEG data. The correlation coefficients are defined as follows:

$$P_{x_i, x_j} = \frac{\text{cov}(x_i, x_j)}{\sqrt{D(x_i)}\sqrt{D(x_j)}} = \frac{E((x_i - E_{x_i})(x_j - E_{x_j}))}{\sqrt{D(x_i)}\sqrt{D(x_j)}}. \quad (16)$$

In formula (16), $\text{cov}(x_i, x_j)$ is the covariance of x_i, x_j . $D(x_i)$ and $D(x_j)$ are the variance of x_i and x_j , respectively. P_{x_i, x_j} is called the correlation coefficient, which is used to measure the degree of correlation between random variables. $P_{x_i, x_j} \in [-1, 1]$ indicates that greater correlation coefficients are associated with greater correlation between the variables x_i and x_j . When P_{x_i, x_j} is 1 or -1, there is a definite linear correlation between x_i and x_j .

The relevant distance is calculated as follows:

$$d_{x_i, x_j} = 1 - P_{x_i, x_j}. \quad (17)$$

For these data relationships, the density is defined as a number of data points randomly distributed within a certain range. Now, setting $D = \{x_1, x_2, \dots, x_n\}$, the density of x_i is defined as follows:

$$\rho_i = \max\{d[x_i, x_j]\}. \quad (18)$$

Among them, x_j belongs to the set of points closest to x_i . In the clustering process, ρ minimum point is the first clustering center. When the next cluster center is determined, the set of clusters formed by the first ρ minimum is removed from the dataset D . In the remaining sets, the smallest point is chosen to form a new clustering center until K clustering centers are selected. Most related algorithms choose Euclidean distance to calculate the cluster center. However, Euclidean distance neglects the correlation between the time series data, and thus, it is not suitable to EEG signal analysis. Therefore, improved K-means uses the correlation coefficient

as the distance metric based on the temporal and spatial correlation of specific time series data and takes full account of the correlations in the time series data.

Improved K-means clustering will iterate to update the prototype after selecting the initial center and calculate the average of all the points in the class as the new clustering center. The mean vector of the new clustering center is defined as follows:

$$u_i = \frac{1}{|D_i|} \sum_{x \in D_i} x. \quad (19)$$

For a certain set of data, the data distribution law conforms to a certain normal distribution law. Thus, the normally distributed sample data are discussed in detail to be accurate to the threshold of each segment. In addition, it is used in conjunction with the idea of a piecewise function. Now, the normal distribution probability and data correlation coefficient are equivalent:

$$P_{x_i, x_j} = \phi\left(\frac{x - u}{\sigma}\right), \quad (20)$$

where u is the mean of the distance from the interior point to the center of the cluster and σ is the distance standard deviation from the interior point to the center of the cluster.

4.4. Algorithmic Flowchart and Description. The flowchart of the proposed ASSC is depicted in Figure 4.

The algorithmic description of the proposed ASSC is summarized in Algorithm 1.

5. Experiments and Discussion

5.1. Performance Metrics. We use the accuracy rate (ACC) as an objective evaluation metric, which means that the number of correct samples is divided by the number of all samples. Its definition is as follows:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\%, \quad (21)$$

where TP and TN represent, respectively, the positive samples and negative samples that are classified into the right types. ACC, as one of the evaluation indexes of the classification accuracy, can effectively measure the classification accuracy of the experimental results. Therefore, we calculated the ACC to better evaluate the experimental results.

To evaluate K-means clustering algorithm, we used SSE as its performance metric to evaluate the clustering quality. SSE expresses the sum of squared errors of the fitting data and the corresponding points of the original data. The smaller the SSE is, the smaller the error of the sample and the center is. Its definition is as follows:

$$\text{SSE} = J(k, u) = \sum_{i=1}^N \|x^{(i)} - u_k(i)\|^2, \quad (22)$$

where N indicates the number of data sources, K represents the number of cluster centers, and u_k indicates the cluster

center. The purpose is to make the data in each class different from each cluster.

5.2. Datasets and Preprocessing. Experimental data are obtained from the PhysioNet Data Bank's Sleep-EDF database. The experimental samples were from Caucasian men and women (21–35 years) who did not take any medications. The first four records (marked as sc*) were obtained from volunteers with healthy respiration within 24 hours of normal daily life in 1989. The data records in the database (marked as st*) were obtained in 1994 from subjects who were light sleepers but relatively healthy. One example is shown in Figure 5.

There were only 8 sample datasets in the original Sleep-EDF database, including 4 from healthy volunteers. To achieve the ASSC, our experimental samples selected sleep data from three healthy volunteers, sc1, sc2, and sc3 (corresponding to the database data sc4002e0/sc4012e0/sc4112e0); each sample contained the level of EOG, Fpz-Cz, and Pz-Oz EEG data, and each sampled at 100 Hz. EEG signals from the Pz-Oz channel produced better classification performance than the Fpz-Cz channel [42, 43]. The Pz-Oz channel was chosen for our study by comparing with the literature [44].

Experts scored the EEG data and generated the PSG based on R&K recommendations as shown in Figure 6. The interval for each period in this study was defined as 30 s, or $30 \times 100 = 3000$ data points. In addition, we calculated the effective sleep time for each sample, marked the 24-hour EEG signal, and selected 9 hours of sleep validity as the final test sample through the manual marker results (.hyp) in the Polyman statistics.

For comparison with the manual marking results, we designed two classification criteria. One was defined according to the R&K sleep stages including AWA, S1, S2, S3, S4, and REM, and the other was to define sleep stages as five categories (S3 and S4 were combined into SS periods).

Data preprocessing was performed to reduce noise interference. In this experiment, the function *wthcoef* was used to perform the threshold processing on wavelet decomposition coefficients, and then, the threshold-processed wavelet coefficients were used for reconstruction to denoise. First, wavelet function “db5” was used to perform 3-layer decomposition of the signal. Second, the scale vector n was set to 1, 2, and 3, and the threshold vector p was set to 100, 90, and 80. Next, the modified wavelet decomposition structure was reconstructed. Finally, we obtained the processed data as shown in Figure 7.

5.3. Results. FFT was calculated based on the frequency-domain correlation, according to the principle that the larger the ρ value, the greater the correlation between the IMF components and the original signal in the frequency domain. We found that, after sorting, the ρ values of IMF₁ to IMF₁₂ gradually decreased as shown in Figure 8.

We set the threshold method (when $\rho > 0.3$, the frequency-domain correlation was more obvious; otherwise, if $\rho \leq 0.3$, we ignored it) and selected 7 IMF components,

FIGURE 4: Flowchart based on improved CEEMDAN and K-means.

Require:

The original EEG signal is processed with the wavelet denoising algorithm.

Ensure:

The clustering results indicate that EEG signal is divided into different sleep stages.

- (1) Define $x(n)$ an N -point EEG epoch X .
- (2) Variable $Nstd$ is the noise standard deviation; NR is the number of realizations; $MaxIter$ is the maximum number of sifting iterations allowed.
- (3) By improved CEEMDAN decomposition, the first mode IMF_1 and the first residual component $r_1(n)$ are obtained, as in formulas (6)–(9).
- (4) **for** $k = 2, \dots, K$ **do**
- (5) Calculate the k -th IMF component and residual component $r_k(n) = r_{k-1}(n) - \widetilde{IMF}_k(n)$.
- (6) Decomposing to achieve the new mode as in formula (11).
- (7) **end for**
- (8) The initial cluster center is divided.
- (9) According to formulas (16) and (17), the correlation distance d_{x_i, x_j} between the data points and the density ρ of each point are calculated, and the smallest ρ is used as the first cluster center to obtain set D .
- (10) The data remaining in set D are allocated to the nearest class according to the distance from the nearest cluster center.
- (11) According to formula (19), the distance is calculated from each point to the center in each class; u and σ are calculated according to different σ segment calculations. The smallest ω is obtained as the new clustering center.
- (12) Recalculate and assign individual sample objects until the cluster center no longer changes.

ALGORITHM 1: ASSC based on improved CEEMDAN and K-means.

which were reconstructed by improved CEEMDAN and used as classifier input as shown in Figure 9.

According to the clinical requirements of sleep staging, the sleep process was divided into 30 s as a stage. If two phases of data appeared in the same data area, more than half of the sleep phase time represented the sleep phase as this segment of data. We observed the changes of classification accuracy of sleep grading by adjusting the number of cluster centers ($K = 5$ or $K = 6$) as shown in Figures 2, 9, and 10.

The objective evaluation metric of ACC is introduced to compare and analyze the results of manual markers. The proportion of accurately classified samples to the total samples was calculated, and the classification accuracy of the experiment was obtained.

The experimental results show that different K values lead to different results in Tables 2 and 3. When $K = 6$, the R&K sleep staging criteria were used to define six sleep stages. At this time, S3 and S4 were counted independently, and the classification accuracy reached 81%. When $K = 5$,

FIGURE 5: The features extracted such as horizontal EOG, Fpz-Cz, and Pz-Oz.

FIGURE 6: Electroencephalogram and manual marking results of Pz-Oz channels in the Polyman statistics.

stages S3 and S4 were considered as a single stage. The number of cluster centers was reduced, and the experimental complexity was reduced. With this approach, the accuracy rate was 83%.

Table 4 shows the SSE values of the two algorithms after the first iteration. The SSE value of the improved algorithm was less than that of the original clustering algorithm, and the smaller the SSE, the better the clustering results.

Due to the uncertainty of different channels of EEG data, various methods of ASSC make it difficult to establish a unified standard for comparison. To make the results more meaningful, the results obtained from several different methods on the same dataset were used for comparison. The accuracy values in Table 5 are the best for a given method. There are no relevant accuracy values in some literature studies, so the missing cases are represented by “—” in

FIGURE 7: Before and after denoising on data to reduce noise interference.

FIGURE 8: Box-plots of sample data processed by improved CEEMDAN.

FIGURE 9: The IMF component in the time domain.

FIGURE 10: Different K values for clustering experiments.

TABLE 2: Five-category staging results.

Subjects	Classification accuracy					Total (%)
	W (%)	S1 (%)	S2 (%)	SS (%)	REM (%)	
SC1	78	65	76	96	93	83
SC2	75	35	76	76	90	79
SC3	82	83	87	94	95	89

TABLE 3: Six-category staging results.

Subjects	Classification accuracy					Total (%)
	W (%)	S1 (%)	S2 (%)	S3 (%)	S4 (%)	
SC1	75	71	78	93	64	80
SC2	87	15	71	82	43	78
SC3	80	88	82	95	76	84

TABLE 4: Comparison of SSE values of the two algorithms after the first iteration.

Method	SSE		
	SC1	SC2	SC3
Original K-means	1103.77	1221.05	1097.54
Improved K-means	1100.03	1197.96	1015.73

Table 5. Therefore, the proposed method was compared with the original K-means algorithm and other algorithms. The experimental results show that obvious increase was achieved in metrics of accuracy and efficiency.

5.4. Discussion. It is not difficult to find that the average accuracy greatly improved using the proposed method. We also found that there is a certain correlation between different stages such as W and S1 stages. In the blinking state, the alpha wave is weak, and its characteristics are similar to those of S1. This is indistinguishable from the

TABLE 5: Comparison of the results of our proposed approach with other relevant literature studies.

Method	6-class (%)	5-class (%)
Original K-means	67	71
Improved K-means [45]	—	76
Wavelet transform + SVM [46]	—	76.5
PSD + ANN [47]	76.7	—
Proposed method	81	83

morphological diversity of EEG waveforms, but it does not mean that all phase values will be reduced. Especially, when using the six-phase classification, the S3 and S4 phases are significantly improved in comparison with the results in the literature [44]. This improvement may be related to our choice of features and improvements in the correlation coefficient of the clustering algorithm. Moreover, previous studies may have overlooked the correlations between different stages and these issues will be studied in the future.

6. Conclusion

This paper proposes an ASSC method based on improved CEEMDAN and K-means. First, improved CEEMDAN algorithm is applied to time series data processing. The appropriate time-frequency is selected based on frequency-domain correlation analysis and calculation. Domain features are used as feature vectors, which are reconstructed by EMD methods to reduce the data dimensionality of the original EEG signals and to improve the computational efficiency. Second, we have improved the classification accuracy of the clustering algorithm based on the innovation of density definition. Finally, we find that the correlation between different sleep stages is significantly improved based on the improved clustering algorithm.

Although the proposed method can improve the classification accuracy of sleep staging, there are still some limitations that should be taken into consideration [48–50]. One of the disadvantages is the relatively low classification accuracy. We will employ deep learning to solve this problem in the future study [51, 52]. On the other hand, we will further explore the correlation between different sleep stages and differentiate them to better improve the classification accuracy [53, 54].

Data Availability

Data used in preparation of this article were obtained from the PhysioNet Data Bank's Sleep-EDF database (<https://www.physionet.org/content/sleep-edfx/1.0.0/>). The investigators within the Sleep-EDF database contributed to the design and implementation of the ASSC method and/or provided data but did not participate in analysis or writing this report. The dataset is described in Kemp et al. [55]. It can be downloaded from <https://ieeexplore.ieee.org/document/867928>. This dataset has been supported by Goldberger et al. [56].

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors are grateful for the support of the National Natural Science Foundation of China (61373149 and 61672329). The authors would also like to acknowledge sleep experts for their invaluable support in data acquisition and visually scoring the EEG recordings of the Sleep-EDF dataset.

References

- [1] Z. Cui, X. Zheng, X. Shao, and L. Cui, "Automatic sleep stage classification based on convolutional neural network and fine-grained segments," *Complexity*, vol. 2018, Article ID 9248410, 13 pages, 2018.
- [2] X. W. Zheng, B. Hu, D. J. Lu, Z. H. Chen, and H. Liu, "Energy-efficient virtual network embedding in networks for cloud computing," *International Journal of Web and Grid Services*, vol. 13, no. 1, pp. 75–93, 2017.
- [3] Y. Liu, S. Wang, M. S. Khan, and J. He, "A novel deep hybrid recommender system based on auto-encoder with neural collaborative filtering," *Big Data Mining and Analytics*, vol. 1, no. 3, pp. 211–221, 2018.
- [4] L. Qi, Y. Chen, Y. Yuan, S. Fu, X. Zhang, and X. Xu, "A QoS-aware virtual machine scheduling method for energy conservation in cloud-based cyber-physical systems," *World Wide Web*, vol. 23, no. 2, pp. 1275–1297, 2020.
- [5] W. Gong, L. Qi, and Y. Xu, "Privacy-aware multidimensional mobile service quality prediction and recommendation in distributed fog environment," *Wireless Communications and Mobile Computing*, vol. 2018, Article ID 3075849, 8 pages, 2018.
- [6] X. Zheng, X. Xiao, and Y. Zhang, "A multidomain survivable virtual network mapping algorithm," *Security and Communication Networks*, vol. 2017, Article ID 5258010, 12 pages, 2017.
- [7] X. Zheng, X. Yu, Y. Li, and H. Liu, "An enhanced multi-objective group search optimizer based on multi-producer and crossover operator," *Journal of Information Science and Engineering*, vol. 33, no. 1, pp. 37–50, 2017.
- [8] X. Yu, H. Wang, X. Zheng, and Y. Wang, "Effective algorithms for vertical mining probabilistic frequent patterns in uncertain mobile environments," *International Journal of Ad Hoc and Ubiquitous Computing*, vol. 23, no. 3/4, p. 137, 2016.
- [9] X. Zheng, J. Tian, X. Xiao, X. Cui, and X. Yu, "A heuristic survivable virtual network mapping algorithm," *Soft Computing*, vol. 23, no. 5, pp. 1453–1463, 2019.
- [10] X. Zheng, Y. Li, H. Liu, and H. Duan, "A study on a cooperative character modeling based on an improved NSGA II," *Multimedia Tools and Applications*, vol. 75, no. 8, pp. 4305–4320, 2016.
- [11] Y.-L. Hsu, Y.-T. Yang, J.-S. Wang, and C.-Y. Hsu, "Automatic sleep stage recurrent neural classifier using energy features of eeg signals," *Neurocomputing*, vol. 104, pp. 105–114, 2013.
- [12] K. A. Aboalayon, H. T. Ocbagabir, and M. Faezipour, "Efficient sleep stage classification based on EEG signals," in *Proceedings of the IEEE Long Island Systems, Applications and Technology (LISAT) Conference 2014*, IEEE, Farmingdale, NY, USA, pp. 1–6, 2014.
- [13] J. Zhou, J. Sun, P. Cong et al., "Security-critical energy-aware task scheduling for heterogeneous real-time MPSoCs in IoT," *IEEE Transactions on Services Computing (TSC)*, In press, 2020.
- [14] L. Qi, X. Zhang, S. Li, S. Wan, Y. Wen, and W. Gong, "Spatial-temporal data-driven service recommendation with privacy-preservation," *Information Sciences*, vol. 515, pp. 91–102, 2020.
- [15] X. Yu, W. Feng, H. Wang et al., "An attention mechanism and multi-granularity-based Bi-LSTM model for Chinese Q&A system," *Soft Computing*, vol. 24, no. 8, pp. 5831–5845, 2020.
- [16] H. Liu, H. Kou, C. Yan, and L. Qi, "Link prediction in paper citation network to construct paper correlated graph," *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, no. 1, 2019.
- [17] L. Qi, Q. He, F. Chen et al., "Finding all you need: web APIs recommendation in web of things through keywords search," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 5, pp. 1063–1072, 2019.
- [18] A. Ramlatchan, M. Yang, Q. Liu, M. Li, J. Wang, and Y. Li, "A survey of matrix completion methods for recommendation systems," *Big Data Mining and Analytics*, vol. 1, no. 4, pp. 308–323, 2018.

- [19] C. W. Anderson, E. A. Stolz, and S. Shamsunder, "Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks," *IEEE Transactions on Biomedical Engineering*, vol. 45, no. 3, pp. 277–286, 1998.
- [20] Y. G. Yang Banghua and Y. Rongguo, "Feature extraction based on wavelet packet optimal basis in brain-computer interface," *Journal of Shanghai Jiaotong University*, vol. 39, no. 11, pp. 1879–1882, 2015.
- [21] J. Fell, J. Rösche, K. Mann, and C. Schäffner, "Discrimination of sleep stages: a comparison between spectral and nonlinear EEG measures," *Electroencephalography and Clinical Neurophysiology*, vol. 98, no. 5, pp. 401–410, 1996.
- [22] A. R. Hassan, S. K. Bashar, and M. I. H. Bhuiyan, "Automatic classification of sleep stages from single-channel electroencephalogram," in *Proceedings of the 2015 Annual IEEE India Conference (INDICON)*, IEEE, Jamia Millia Islamia, India, pp. 1–6, December 2015.
- [23] H. Zhang and Q. Cai, "Time series similar pattern matching based on wavelet transform," *Chines Journal of Computer*, vol. 26, no. 3, pp. 373–377, 2003.
- [24] C. F. Mu Feng and J. Yuyu, "Signal filtering based on improved EMD algorithm," *Journal of Shandong University (Engineering Science)*, vol. 45, no. 3, pp. 35–42, 2015.
- [25] Z. Zhang, H. Kawabata, and Z.-Q. Liu, "Electroencephalogram analysis using fast wavelet transform," *Computers in Biology and Medicine*, vol. 31, no. 6, pp. 429–440, 2001.
- [26] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Advances in Adaptive Data Analysis*, vol. 01, no. 01, pp. 1–41, 2009.
- [27] A. R. Hassan and M. I. H. Bhuiyan, "Computer-aided sleep staging using complete ensemble empirical mode decomposition with adaptive noise and bootstrap aggregating," *Biomedical Signal Processing and Control*, vol. 24, pp. 1–10, 2016.
- [28] H. Cecotti and A. Graser, "Convolutional neural networks for p300 detection with application to brain-computer interfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 433–445, 2010.
- [29] B. Şen, M. Peker, A. Çavuşoğlu, and F. V. Çelebi, "A comparative study on classification of sleep stage based on eeg signals using feature selection and classification algorithms," *Journal of Medical Systems*, vol. 38, no. 3, p. 18, 2014.
- [30] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297, Oakland, CA, USA, 1967.
- [31] S. Güneş, K. Polat, S. Yosunkaya, and M. Dursun, "A novel data pre-processing method on automatic determining of sleep stages: K-means clustering based feature weighting," in *Proceedings of the Computational Science and Its Applications-ICCSA*, pp. 112–117, Normandy, France, 2009.
- [32] T. Lajnef, S. Chaibi, P. Ruby et al., "Learning machines and sleeping brains: automatic sleep stage classification using decision-tree multi-class support vector machines," *Journal of Neuroscience Methods*, vol. 250, pp. 94–105, 2015.
- [33] N. E. Huang, Z. Shen, S. R. Long et al., "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998.
- [34] N. E. Huang, Z. Shen, and S. R. Long, "A new view of nonlinear water waves: the hilbert spectrum," *Annual Review of Fluid Mechanics*, vol. 31, no. 1, pp. 417–457, 1999.
- [35] F. Riaz, A. Hassan, S. Rehman, I. K. Niazi, and K. Dremstrup, "Emd-based temporal and spectral features for the classification of eeg signals using supervised learning," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 1, pp. 28–35, 2015.
- [36] M. Långkvist, L. Karlsson, and A. Loutfi, "A review of unsupervised feature learning and deep learning for time-series modeling," *Pattern Recognition Letters*, vol. 42, pp. 11–24, 2014.
- [37] X. Xu, S. Fu, L. Qi et al., "An iot-oriented data placement method with privacy preservation in cloud environment," *Journal of Network and Computer Applications*, vol. 124, pp. 148–157, 2018.
- [38] X. Xu, Q. Liu, Y. Luo et al., "A computation offloading method over big data for iot-enabled cloud-edge computing," *Future Generation Computer Systems*, vol. 95, pp. 522–533, 2019.
- [39] L. Qi, X. Zhang, W. Dou, C. Hu, C. Yang, and J. Chen, "A two-stage locality-sensitive hashing based approach for privacy-preserving mobile service recommendation in cross-platform edge environment," *Future Generation Computer Systems*, vol. 88, pp. 636–643, 2018.
- [40] Q. Tang, *Study on automatic sleep staging based on EEG*, Ph.D. thesis, Guangdong University of Technology, Guangzhou, China, 2016.
- [41] X. Zheng, B. Hu, and D. Lu, "A multi-objective virtual network embedding algorithm in cloud computing," *Journal of Internet Technology*, vol. 17, no. 4, pp. 633–642, 2016.
- [42] G. Zhu, Y. Li, and P. Wen, "Analysis and classification of sleep stages based on difference visibility graphs from a single-channel EEG signal," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 6, pp. 1813–1821, 2014.
- [43] C. Berthomier, X. Drouot, M. Herman-Stoica et al., "Automatic analysis of single-channel sleep EEG: validation in healthy individuals," *Sleep*, vol. 30, no. 11, pp. 1587–1595, 2007.
- [44] A. R. Hassan and M. I. Hassan Bhuiyan, "Automatic sleep scoring using statistical features in the EMD domain and ensemble methods," *Biocybernetics and Biomedical Engineering*, vol. 36, no. 1, pp. 248–255, 2016.
- [45] S. Xiao, B. Wang, J. Zhang, Q. Zhang, and J. Zou, "Automatic sleep stage classification based on an improved K-means clustering algorithm," *Journal of Biomedical Engineering*, vol. 33, no. 5, pp. 847–854, 2016.
- [46] J. M. S. Pascualvaca, C. Fernandes, A. Guillén et al., "Sleep stage classification using advanced intelligent methods," in *International Work-Conference on Artificial Neural Networks*, pp. 604–612, Springer, Berlin, Germany, 2013.
- [47] M. Ronzhina, O. Janoušek, J. Kolářová, M. Nováková, P. Honzík, and I. Provazník, "Sleep scoring using artificial neural networks," *Sleep Medicine Reviews*, vol. 16, no. 3, pp. 251–263, 2012.
- [48] Y. Wang, C. Zhao, q. Xu, Z. Zheng, Z. Chen, and Z. Liu, "Fair secure computation with reputation assumptions in the mobile social networks," *Mobile Information Systems*, vol. 2015, Article ID 637458, 8 pages, 2015.
- [49] L. Zhang, Y. Wang, F. Li, Y. Hu, and M. H. Au, "A game-theoretic method based on q-learning to invalidate criminal smart contracts," *Information Sciences*, vol. 498, pp. 144–153, 2019.
- [50] Y. Wang, G. Yang, T. Li, F. Li, and X. Yu, "Belief and fairness: a secure two-party protocol toward the view of entropy for iot devices," *Journal of Network and Computer Applications*, vol. 161, Article ID 102641, 2020.

- [51] J. Zhou, X. S. Hu, Y. Ma, J. Sun, T. Wei, and S. Hu, "Improving availability of multicore real-time systems suffering both permanent and transient faults," *IEEE Transactions on Computers*, vol. 68, no. 12, pp. 1785–1801, 2019.
- [52] J. Zhou, J. Sun, X. Zhou et al., "Resource management for improving soft-error and lifetime reliability of real-time mpsoes," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 38, no. 12, pp. 2215–2228, 2019.
- [53] R. Liu, H. Wang, and X. Yu, "Shared-nearest-neighbor-based clustering by fast search and find of density peaks," *Information Sciences*, vol. 450, pp. 200–226, 2018.
- [54] B. Hu, H. Wang, X. Yu, W. Yuan, and T. He, "Sparse network embedding for community detection and sign prediction in signed social networks," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 1, pp. 175–186, 2019.
- [55] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Obery, "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 9, pp. 1185–1194, 2000.
- [56] A. L. Goldberger, L. A. N. Amaral, and L. Glass, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2003.