EEG signal classification method based on improved empirical mode decomposition and SVM

Zihao Zhang1,*, Zhiyi Li1, Ting Ma1 and Jiayu Zhao1
1 Institute of Information and communication, North University of China, Taiyuan, Shanxi, 030000, China
*Corresponding author’s e-mail: 1805004416@st.nuc.edu.cn

Abstract. Epilepsy is a common phenomenon formed by abnormal discharges between brain neurons. The seizures of epilepsy are sudden and irregular. As a non-stationary signal, EEG signals can express its characteristics to a certain extent, and makes a significant difference in the monitoring and treatment of epilepsy diseases. This study employs empirical mode decomposition (EMD) to decompose the interictal and epileptic EEG signals into multiple eigenmode functions (IMF), and combines the correlation coefficient to screen the main IMF and extract its variance, fluctuation coefficient and Coefficient of variation and other features, combined with support vector machines for classification. Compared with the traditional empirical mode decomposition, this method has higher accuracy in the identification and classification of epileptic signals. The combination of this method not only provides a theoretical method for disease diagnosis and treatment, but also verifies the research and application value of EEG signals to a certain extent.

1. Introduction
Epilepsy is a common neurological disease. It is manifested as sudden abnormal discharge of brain neurons and transient brain dysfunction. Many people in the world are affected by this disease. According to statistics, about 0.8% to 1% of people in the world suffer from epilepsy [1].

In the last few years, with the continuous improvement of equipment accuracy and the continuous advance of medical standards, our research on EEG signals has become more and more extensive. The scalp Ephiacho-Graph (EEG) signal involves a large quantity of physiological and pathological information and has a high time resolution. During epileptic seizures, the human brain central nervous system will be abnormal, causing the synchronous neurons to suddenly discharge, thereby The corresponding EEG will show abnormal waves, so the study of epilepsy EEG has remarkable value and clinical meaning for the identification and detection of epilepsy diseases [2].

EEG signals have strong nonlinear characteristics. Research and detection of EEG signals have been extensively studied in recent years, and plenty of new detection methods have been put forward and put into practice [3]. Jones-Gotman M et al. proposed a method of decomposing the EEG signal into half-waves and after that extracting traits to process epileptic EEG signals [4]. Oweis et al. realized the classification of mode mixture by extracting the instantaneous information of each intrinsic modal function, and used it to distinguish the interictal and epileptic EEG [5]. Samiee et al. proposed a reasonable discrete short-time Fourier transform and an epilepsy classification feature extraction scheme based on statistical features [6]. Fu et al. realized Hilbert marginal spectrum analysis and SVM to detect epilepsy from EEG data [7].

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd
As a new type of adaptive signal time-frequency processing mean, Empirical Mode Decomposition (EMD) was first proposed by Huang in 1998. It is the core of the Hilbert-Huang Transformation Method (HHT) and is suitable for analyzing and processing nonlinear and non-linear signals. Stationary signal [8]. Empirical Mode Decomposition (EMD) can decompose epilepsy EEG into eigenmode functions (IMF) of different scales. The main IMF components obtained by the decomposition include characteristic trend information and signal fluctuation characteristics [9]. Based on the above time-frequency analysis and pattern recognition methods, this research proposes an EEG signal processing method that combines correlation coefficients to improve empirical mode decomposition and SVM to extract features of clinically collected epileptic seizure intermittent and epileptic EEG signals and classify them.

2. Experimental data selection
The experimental data of this study used the epilepsy patient data set publicly offered by the University of Bonn. The data set includes five subsets, namely A, B, C, D, and E. Each subset contains 100 single-channel EEG with a record length of 23.6 seconds and a sampling frequency of 173.61 Hz [10]. The scalp electrodes are distributed in the international 10-20 system. Subsets A and B were collected from a control group consisting of 5 healthy people. The fragment in A is the EEG when the subject is open, and the fragment in B is the EEG when the subject is closed. Subsets C, D, and E are intracranial EEG, collected from 5 patients who have been diagnosed before surgery. These patients have undergone partial hippocampal resection to make their epilepsy completely controllable. The resection area has been clinically verified as epileptic foci. Subset C contains EEG collected from the opposite side of the epileptic foci, and subset D contains EEG collected from the epileptic foci. Subsets C and D are collected during the epileptic seizure period. Subset E contains the seizure EEG collected by all intracranial electrodes. These fragments are manually cut from the long-range multi-channel EEG, and some possible interferences are removed at the same time, and they are widely used in public, and the experimental data has strong credibility.

3. Experimental method

3.1. Principle of EMD
Empirical Mode Decomposition (EMD) is an adaptive nonlinear signal handling method. The mean decomposes the signal according to the data’s time scale characteristics, and is completely driven by the data, we don't have to set the basis function in advance. The key to EMD is to resolve a complicated signal into a limited number of eigenmode functions (IMF), and every of the decomposed IMF components involves the local characteristic signals at different time scales of the original signal, which is conducive to highlighting the local traits of the EEG signal [11]. The conditions to be met by the decomposed IMF:

1. The number of extreme points and zero points is equal, or one difference at most.
2. The local mean of the upper envelope defined by the maximum and the lower envelope defined by the minimum is zero.

The specific steps of EMD are as follows, assuming that the original signal is \( x(t) \):

Step 1: Find all the maximum and minimum points of the original signal \( x(t) \).

Step 2: Fit all the maximum value points through the cubic spline function to fit the maximum value envelope \( m_1(t) \), fit all the minimum value points through the cubic spline function to fit the minimum value envelope \( m_2(t) \).

Step 3: Figure up the average value of the upper and lower envelopes \( a(t) \).

\[
a(t) = \frac{1}{2}[m_1(t) + m_2(t)]
\]  \hspace{1cm} (1)

Step 4: Subtract the envelope average value from the original signal to get a new signal sequence \( f(t) \).
\[ f(t) = x(t) - a(t) \]  

Step 5: If the new signal sequence satisfies the conditions of the IMF, it is the first IMF component, denoted as \( s_1(t) \). If it is not met, repeat steps 1, 2, and 3 until the two conditions of the IMF component are met. Since the average value of the upper and lower envelopes cannot be zero in actual situations, Huang believes that when the Cauchy screening stop criterion is met, the loop can be stopped.

\[ SD = \frac{\sum|l_u(t) - s(t)|}{\sum|l_u(t)|} \]  

Generally, when the value of SD is between 0.2 and 0.3, the screening can be stopped.

Step 6: Separate the IMF component \( s_1(t) \) from the original signal \( x(t) \) to obtain the residual signal \( r(t) \).

Step 7: Regard the residual signal \( r(t) \) as the new original signal, repeat the above steps \( n \) times until the residual function \( r_n(t) \) is a monotonic function. The EMD of the original signal are able to decomposed into:

\[ x(t) = \sum_{i=1}^{\infty} s_i(t) + r_i(t) \]  

EMD approximates an adaptive high-pass filter, and the IMF components of each order obtained by EMD decomposition are arranged in the order of frequency from high to low.

### 3.2 Correlation coefficient

Correlation coefficient is a statistical index which measures the linear relationship and the related degree between two variables\[12\]. In this paper, the main IMF components are extracted by combining the correlation coefficient calculation and the dependence relationship between the IMF components of each order and the original brain signal disintegrated by the EMD algorithm, and the relevant features are drawn and then screened.

The correlation coefficient is the earliest statistical indicator designed by statistician Karl Pearson. If there are random variables X, Y, they are defined as follows:

\[ r = \frac{Cov(X,Y)}{\sigma_x \sigma_y} \]  

\[ r = \frac{E[X-E(X)][Y-E(Y)]}{\sqrt{E[X-E(X)]^2 E[Y-E(Y)]^2}} \]  

Among them, Cov (X, Y) is the covariance function between two random variables, \( \sigma \) representing the standard deviation of the two random variables, and E represents the mean value of the function.

\( r \) represents the closeness between two variables. In this study, the correlation coefficient between the IMF components of each order and the original signal was calculated, and the calculated value was used as the screening basis, and the IMF components containing the main EEG signal characteristics were selected.

Step 1: The EMD decomposition of the original EEG signal is:

\[ x(t) = \sum_{i=1}^{\infty} s_i(t) + r_i(t) \]  

Step 2: By setting the threshold \( \varepsilon \) of \( r \), compare the calculated correlation coefficient \{\( C_i, i=1,2,3,4,5...n \}\} with its judgment, and calculate the \( s_i(t) \) whose correlation coefficient is greater than the threshold \( \varepsilon \) as the main IMF component.

### 3.3 Support Vector Machine (SVM)

SVM is an analysis data set method developed based on statistical learning theory to perform multi-class classification. The basic model is a linear classifier that defines the maximum interval in the feature space. An SVM model constructs a hyperplane to best divide the data classify into one of two
categories, and the distance from the hyperplane to the nearest data point is maximized on each side [13].

Assume that the test data set is linearly separable, the usual form of the hyperplane can be represented as:

$$\omega^T x + b = 0$$  \hspace{1cm} (8)

$\omega$ is the normal vector of the hyperplane and $b$ is the bias term. The follow equation must be met for each data point $X_i$:

$$d_i(\omega^T x_i + b) \geq 1$$  \hspace{1cm} 1 \leq i \leq n \hspace{1cm} (9)

The classification interval is equal to $2/\|\omega\|$, the interval is to be maximally equivalent to the minimum $\|\omega\|$. Then optimize $\omega$ and $b$ to set the best separation hyperplane, and maximize the margin between these two categories.

If the provided data set is not linearly separable, that is, when the two types of points cannot be completely separated by a hyperplane, slack variables $\xi_i (\xi_i \geq 0, i=1, 2, …, n)$ are introduced $\xi_i (\xi_i \geq 0, i=1, 2, …, n)$ to measure the degree of misclassification. When $\xi_i<1$, the sample point $X_i$ is correctly classified, and when $\xi_i \geq 1$, the sample point $X_i$ is misclassified. For this reason, the objective function is introduced:

$$\phi(\omega, \xi) = \frac{1}{2} \alpha^T \omega + C \sum_{i=1}^{n} \xi_i$$  \hspace{1cm} (10)

And modify the original question to:

$$\min \frac{1}{2} \alpha^T \alpha + C \sum_{i=1}^{n} \xi_i$$  \hspace{1cm} (11)

subject to  \hspace{1cm} $d_i(\omega^T \phi(x_i) + b) \geq 1 - \xi_i$  \hspace{1cm} $1 \leq i \leq n$  \hspace{1cm} (12)

Among them, C is a regularization parameter, which decides the penalty for data classification errors, and $\phi(x_i)$ maps $x_i$ into a higher-dimensional space to make it easier to separate such spaces.

The theorem shows that linear combinations with large dimensions $\omega$ can be written as training $\omega = \sum a_i \phi(x_i)$, Therefore, we can optimize $\alpha_i$ instead $\omega$, the decision function is changed into

$$f(x) = \sum a_i K(x_i, x) + b$$  \hspace{1cm} (13)

The kernel function is $K(x_i, x) = \phi(x_i)^T \phi(x)$. The new double question was revised to:

$$\min \sum \frac{1}{2} a_i a_j K(x_i, x_j)$$  \hspace{1cm} subject to  \hspace{1cm} $0 \leq a_i \leq C$, and $\sum a_i d_i = 0$ \hspace{1cm} (14)

Use Gaussian radial basis kernel function (RBF) $K(x_j, x_k)$. Finally, the classification discriminant function of nonlinear SVM can be expressed as:

$$f(x) = \text{sgn} \left( \sum a_i y_i K(x_i, x) + b \right)$$  \hspace{1cm} (15)

3.4 Feature selection

3.4.1 Coefficient of Variation. When analyzing the amplitude of EEG signals, the commonly used features include the average value, standard deviation, and coefficient of variation. The coefficient of variation can be used to measure the amplitude change of the epileptic EEG signal, and the coefficient of variation is defined as:

$$CV = \frac{\sigma^2}{\mu^2}$$  \hspace{1cm} (16)

among them:
\[\mu = \frac{1}{n} \sum_{i=1}^{n} |\text{IMF}_i| \] (17)
\[\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (|\text{IMF}_i - \mu|^2)} \] (18)

\(n\) is the length of the eigenmode function filtered by correlation coefficient after EMD decomposition.

3.4.2. **Relative energy.** Define the energy of the \(i\)-th IMF component as:
\[E_i = \frac{\Delta t}{n} \sum_{l} |c_l(i)| \] (19)
Where \(n\) is the length of the IMF component, \(\Delta t\) is the sampling time, and \(c_l(i)\) is the \(l\)th IMF component.
\[E_i = E_i / \sum_{j=1}^{m} E_j \] (20)
Where \(m\) is the number of IMF components.

3.4.3. **Volatility Index.** The volatility index can measure the intensity of changes in brain electrical signals. There will be different fluctuations during epileptic seizures and during the intermittent period. It is defined as:
\[F_1 = \frac{1}{n} \sum_{i=1}^{n} |\text{IMF}(i+1) - \text{IMF}(i)| \] (21)

\(\text{IMF}\) still represents the eigenmode function after screening, and \(n\) represents its length.

4. **Experiments and discussion of results**

4.1. **Epilepsy EEG signal**
In order to compare and verify, this study uses the D (seizure interval) data set and E data set (seizure period) two sets of data sets to conduct experiments. Here, this experiment selects one single-channel signal among the 100 single-channel signals of the epileptic interval and seizure period in the Bonn data set, each of which contains 4097 sample points.

![Figure 1 Interictal and ictal EEG signals](image)

From the figure, we not only have differences in amplitude between seizure periods and seizure periods, but there are also obvious differences in signal jitter and stable differences, which is convenient for us to perform EMD decomposition and feature extraction behind.
4.2. EMD decomposition and correlation coefficient calculation

After the two groups of 100 single channels are selected 20 sub-data sets corresponding to the average value, the EMD epilepsy signal is classically decomposed, and the correlation coefficient between the IMF obtained by each decomposition and the original signal is calculated with the help of Matlab software. The value is shown in the figure.

![Correlation coefficients of various levels of IMF and original EEG signals](image_url)

Figure 2: Correlation coefficients of various levels of IMF and original EEG signals

Through the calculation chart, this experiment selects IMF1~IMF5 after empirical mode decomposition to extract features and calculate.
Figure 3 selects one of the 100 single-channel data in the Bonn epilepsy data set D (interictal epilepsy) as a sample, and retains its IMF1-IMF5 eigenmode function after selection according to the correlation coefficient. Figure 2 selects one of the 100 single-channel data in the Bonn epilepsy data set E (seizure period) as a sample, and retains its IMF1-IMF5 eigenmode functions. From the decomposed IMF components, it can be seen that there are large differences in amplitude and frequency. Next, the retained IMF is classified and calculated according to the eigenvalues used.

4.3. Features calculation and classification

In this study, the difference is observed by calculating the coefficient of variation, relative energy, and fluctuation coefficient of the signal. The features obtained by decomposition are combined to form a feature vector, which is sent to a support vector machine (SVM) classifier for training and classification.

Select the Bonn epilepsy EEG signal data set and 1 single channel signal in 100 data sets of the epilepsy as an example, calculate the two sets of data and the characteristic values of the original EEG signal and list the following table.

Table 1. Calculation of the characteristics of epileptic signal seizure period and seizure period

| IMF component | average | variance | variable coefficient | relative energy | coefficient of fluctuation |
|---------------|---------|----------|----------------------|----------------|----------------------------|
| IMF1          | 0.0028  | 7.3802   | 0.00000107           | 0.012319186    | -0.000391447              |
| IMF2          | 0.0617  | 50.4001  | 0.0000754            | 0.084135733    | -0.000124146              |
| IMF3          | 0.0852  | 86.9201  | 0.000836             | 0.145101595    | 0.001166039               |
| IMF4          | 0.4546  | 226.3661 | 0.000913             | 0.378201584    | -0.010816136              |
| IMF5          | 0.1644  | 227.7681 | 0.0001186            | 0.380241903    | 0.005412189               |

| IMF component | average | variance | variable coefficient | relative energy | coefficient of fluctuation |
|---------------|---------|----------|----------------------|----------------|----------------------------|
| IMF1          | 5.34    | 115383.8912 | 0.000247194        | 0.338840356    | 0.006061233              |
It can be seen from the example data that the calculated values of the epilepsy EEG data concentrated in the interictal and during the seizures are quite different. We obtained the characteristics of the values in the 100 sets of interictal and the interictal data, and divided IMF1- The selected features of the IMF5 combination are added to the support vector machine for classification. The first 80 groups of data are used as training samples for training, and the remaining data is used to verify the data classification effect and observe the correct rate of the classifier.

The IMF components filtered by the correlation coefficient are used to extract relevant effective features to form a combined vector. After being sent to the SVM classifier for classification, the desired effect is obtained. The accuracy rate reaches 95.7%, which is higher than the traditional single method analysis. Combined with the method in this paper, the EEG signals in the inter-seizure period and the seizure period in the data set can be well classified.

5. Conclusion
There are abundant physiological and pathological signals in EEG signals. It is of great research value to use the analysis EEG signals to monitor diseases. Decomposing EEG signals into arrays of IMF by EMD algorithm is a method widely used in signal processing in recent years. This study uses a combination of correlation coefficients to screen out IMFs containing main relevant information, and then classifies EEG signals to reduce the interference of unnecessary signals and noises, the classification accuracy is improved, and it can provide certain auxiliary means for clinical monitoring and treatment, which is beneficial to the prevention and treatment of epilepsy diseases. It can also be extended to other biomedical signal processing fields for other physiological signals and EEG classification applications such as mood and depression. This research provides a certain reference method.

Acknowledgment
The support of the National Undergraduate Innovation and Entrepreneurship Training Program in 2020(No.202010110006).Thanks to University of Bonn, Germany, for the support of epilepsy EEG data!

References
[1] Adeli H, Zhou Z, Dadmehr N. Analysis of EEG records in an epileptic patient using wavelet transform [J]. J Neurosci Methods, 2003 Feb 15, 123( 1): 69 – 87.
[2] M.E. Saab, J. Gotman, A system to detect the onset of epileptic seizure in scalp EEG, Clin. Neurophysiol. 116 (2) (2005) 427–442.
[3] P. Kwan, M.J. Brodie, Combination therapy in epilepsy, Drugs 66 (14) (2006) 1817–1829.
[4] Jones-Gotman M, Bouwer MS, Gotman J. EEG slow waves and memory performance during the intracarotid amobarbital test [J]. Epilepsia, 1994, 35( 1): 61-69.
[5] Oweis R J, Abdulhay EW. Seizure classification in EEG signals utilizing Hilbert-Huang transform [J]. BioMedical Engineering OnLine, 2011, 10( 1): 1-15.
[6] K. Samiee, P. Kovacs, M. Gabbouj, Epileptic seizure classification of eegtime-series using rational discrete short-time fourier transform, IEEE Trans.Biomed. Eng. 62 (2) (2015) 541–552.
[7] K. Fu, J. Qu, Y. Chai, T. Zou, Hilbert marginal spectrum analysis forautomatic seizure detection in eeg signals, Biomed. Signal Process. Control 18 (2015) 179–185.
[8] N.E. Huang, Z. Shen, S.R. Long, M.L. Wu, H.H. Shih, Q. Zheng, N.C. Yen, C.C. Tung, H.H. Liu, The empirical mode decomposition and the Hilbert spectrum for nonlinear and nonstationary time series analysis, Proceedings of the Royal Society of London Series A 454(1998) 903–995.

[9] R.B. Pachori, V. Bajaj, Analysis of normal and epileptic seizure EEG signals using empirical mode decomposition, Comput. Methods Programs Biomed. 104 (3) (2011) 373–381.

[10] R. Andrzejak, G. Widman, K. Lehnertz, C. Rieke, P. David, C. Elger, The epileptic process as nonlinear deterministic dynamics in a stochastic environment: an evaluation on mesial temporal lobe epilepsy, Epilepsy Res. 44 (2) (2001) 129–140.

[11] HUANG N E, SHEN Z, LONG S R, et al. The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time Series Analysis [J]. Proceedings of the Royal Society of London, 1998, 454(1971): 903-995.

[12] WANG G, CHEN X Y, QIAO F L, et al. On intrinsic mode function [J]. Advances in Adaptive Data Analysis, 2010, 2(3): 277-293.

[13] M. Nandan, S.S. Talathi, S. Myers, et al., Support vector machines for seizure detection in an animal model of chronic epilepsy, J. Neural Eng. 7 (3) (2010) 036001.