

Research Article

Robust Deep Network with Maximum Correntropy Criterion for Seizure Detection

Yu Qi,^{1,2} Yueming Wang,^{1,2} Jianmin Zhang,³ Junming Zhu,³ and Xiaoxiang Zheng^{1,4}

¹ Qiushi Academy for Advanced Studies, Zhejiang University, Hangzhou 310027, China

² Department of Computer Science, Zhejiang University, Hangzhou 310027, China

³ Second Affiliated Hospital of Zhejiang University, College of Medicine, Hangzhou 310000, China

⁴ Department of Biomedical Engineering, Zhejiang University, Hangzhou 310027, China

Correspondence should be addressed to Yueming Wang; yumingwang@zju.edu.cn

Received 27 March 2014; Accepted 4 June 2014; Published 6 July 2014

Academic Editor: Ting Zhao

Copyright © 2014 Yu Qi 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.

Effective seizure detection from long-term EEG is highly important for seizure diagnosis. Existing methods usually design the feature and classifier individually, while little work has been done for the simultaneous optimization of the two parts. This work proposes a deep network to jointly learn a feature and a classifier so that they could help each other to make the whole system optimal. To deal with the challenge of the impulsive noises and outliers caused by EMG artifacts in EEG signals, we formulate a robust stacked autoencoder (R-SAE) as a part of the network to learn an effective feature. In R-SAE, the maximum correntropy criterion (MCC) is proposed to reduce the effect of noise/outliers. Unlike the mean square error (MSE), the output of the new kernel MCC increases more slowly than that of MSE when the input goes away from the center. Thus, the effect of those noises/outliers positioned far away from the center can be suppressed. The proposed method is evaluated on six patients of 33.6 hours of scalp EEG data. Our method achieves a sensitivity of 100% and a specificity of 99%, which is promising for clinical applications.

1. Introduction

Epilepsy is a common and serious brain disorder, which affects about 50 million people worldwide [1]. Epileptic seizures are characterized by convulsions, loss of consciousness, and muscle spasms resulting from excessive synchronization of neuronal activities in the brain [2]. The abnormal neuronal discharges lead to epileptic patterns such as closely spaced spikes and slow waves in electroencephalogram (EEG). In seizure diagnosis and evaluation, visual inspection of these epileptic patterns from long-term EEG is a routine job for the doctors, which could be highly tedious and time-consuming [3]. Therefore, reliable seizure detection system that identifies seizure events automatically would facilitate seizure diagnosis and has great potential in clinical applications.

There are two key points in automatic seizure detection. One is how to capture the diverse patterns of seizure EEG. For different individuals, the morphologies of seizure patterns could vary considerably. Therefore, effective feature extraction plays a key role in seizure detection and lots of efforts

have been made. In order to characterize the changes in amplitude and energy in epileptic EEG, Saab and Gotman [4] proposed to use three measures, relative average amplitude, relative scale energy, and coefficient of variation of amplitude. Similarly, Majumdar and Vardhan [5] utilized the variance of differentiation of time window to detect significant changes in EEG signals. To identify the sharp waves which typically appear in seizure signals, Yadav et al. [6] introduced a morphology-based detector based on the slopes of the half-waves of signals. To characterize the intrinsic time-frequency components of seizure patterns, Ghosh-Dastidar et al. [7] used principal component analysis and Zandi et al. [8] applied wavelet transform to decompose the EEG signal for feature enhancement. To encode the changes in dynamics of epileptic signal, Jouny and Bergey [9] utilized nonlinear measures of sample entropy and Lempel-Ziv complexity. To describe the topology state of epilepsy, Santaniello et al. [10] transformed the multichannel EEG data into a cross-power matrix, and eigenvalues of the matrix are used for seizure detection. The other key point is how to reduce the effect

of noise. The noises caused by electromyography (EMG) or electrode movements commonly appear in EEG signal and are prone to trigger false alarms. These artifacts could bring impulsive changes with large amplitudes in EEG signal and lead to outlying values in the feature space. Some existing methods simply assumed these noises to be Gaussian [11, 12] and thus would be fragile given large amounts of outliers. Other approaches applied specific false alarm avoidance methods against these noises [4–6].

Although existing methods have shown some strengths in specific EEG datasets, the following problems have not yet been well explored. First, most existing features are designed according to the observation of a few seizure patterns, which seems too empirical to cover a wide range of seizure patterns; thus the features are usually suboptimal. Second, existing methods could be sensitive to the noises in EEG signals. Artifacts caused by EMG or electrode movements probably lead to a EEG signal shape similar to that of seizure states. A simple Gaussian assumption for the noises can be incorrect and the approaches designed based on this can cause high false alarms [11, 12]. Finally, most methods design the feature and classifier individually. Few efforts have been made to study the relationship between them or simultaneously optimize both of the two parts to maximize the abilities of them.

Inspired by the great success of deep network in image retrieval, speech recognition, and computer vision [13–21], this paper proposes a deep model framework to deal with the above issues. The main contributions of our work can be summarized as follows.

- (i) Instead of manually designing a feature, we propose a network called robust stacked autoencoder (R-SAE) to automatically learn a feature to represent seizure patterns. The reconstruction error is first used to learn an initial feature.
- (ii) To reduce the effect of noises on EEG signals, we formulate a maximum correntropy criterion (MCC) to the R-SAE network. Unlike the traditional autoencoder model which uses the mean square error (MSE) as the reconstruction cost, the output of the new kernel MCC increases more slowly than that of MSE when the input goes away from the center. Thus, the effect of those noises/outliers positioned far away from the center can be suppressed.
- (iii) The R-SAE part and classification part are integrated to a new deep network. The objective of the network is the best seizure classification accuracy. Thus, both the initial feature and the classifier could be optimized according to the detection objective so that the whole detection system could be as optimal as possible. Besides, the optimal feature is completely data-driven. Given enough training data, the optimal feature learned by our method is able to represent various seizure patterns.

Our method is evaluated on 33.6 hours of EEG signals from six patients. With the MCC-based R-SAE model, robust features are extracted from noisy EEG signal that the sensitivity and specificity increase by 14% and 1% compared with

TABLE 1: Patient information and selected frequency bands.

Patient	Sex	Chan. #	Sei. #	Hours	Freq. band
Pt01	Female	28	2	5.6	14–30 Hz
Pt02	Female	28	2	5.6	8–13 Hz
Pt03	Female	28	3	5.6	4–7 Hz
Pt04	Male	28	3	5.6	14–30 Hz
Pt05	Male	28	3	5.6	8–13 Hz
Pt06	Male	28	3	5.6	8–13 Hz

the traditional stacked autoencoder (S-SAE). By supervised joint optimization of our deep model, the features are further optimized with better separability in the feature space and the sensitivity and specificity increase by 8% and 15%, respectively. In comparison with other methods, the proposed R-SAE model outperforms the competitors and achieves a high sensitivity of 100% and a specificity of 99%.

The rest of this paper is organized as follows. Section 2 presents the detail of the R-SAE deep model. The experimental results and discussions are shown in Section 3. Finally, we draw the conclusions in Section 4.

2. Materials and Methods

The framework of our method is shown in Figure 1. The multichannel EEG signals are firstly divided into short-time segments, and we calculate the cross-power matrix for each segment to reveal the spatial patterns of the brain. Then, compact features are extracted from the cross-power matrix by a deep network cascaded to a softmax classifier. In our method, the deep network is first pretrained with the R-SAE model to extract useful features, and then the features are further optimized jointly with the classifier to obtain optimal seizure detection system.

2.1. EEG Data. Scalp EEG data of six patients are used in this study. The EEG data were recorded during long-term presurgical epilepsy monitoring using NicoletOne amplifier at Second Affiliated Hospital of Zhejiang University, College of Medicine. A total of 28 channels were acquired at the sample rate of 256 Hz according to 10–20 electrode placement systems. The detail of the EEG data is given in Table 1. For each patient, all the available seizure EEG signals are used, and we randomly choose two 2.8-hour-long EEG segments as the nonseizure data segmentation and data preparation.

2.2. Segmentation and Data Preparation. In the preprocessing stage, the multichannel EEG data are divided into 5-second-long segments with a sliding window. For each patient, a total of 4000 segments of nonseizure data and 1000 segments of seizure data are divided from the EEG signals. There is no overlap between nonseizure segments, while, for seizure segments, the proportion of overlap is configured considering the total length of the seizure signal and number of segments required.

After segmentation, all the segments are disordered and we randomly pick 750 seizure segments and 750 nonseizure

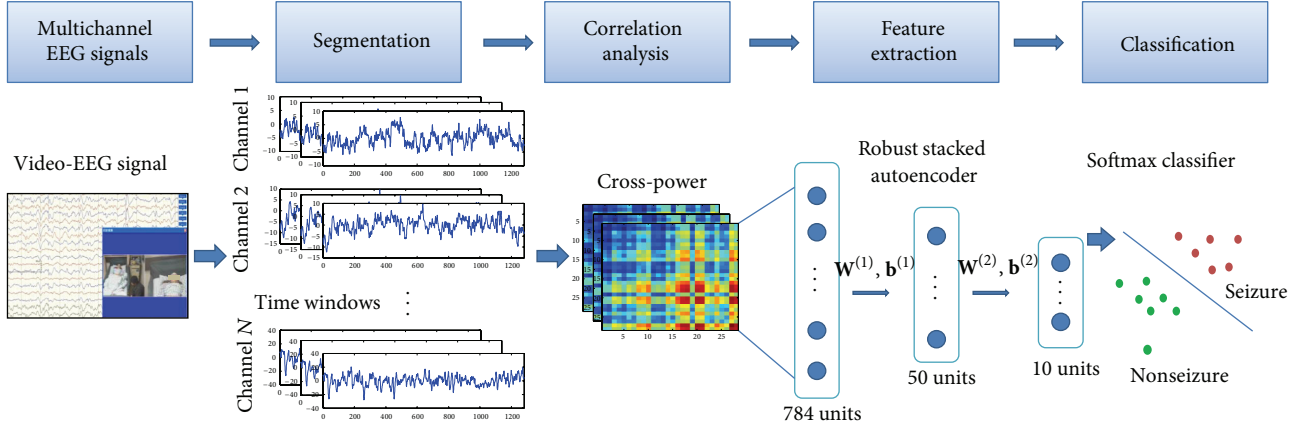


FIGURE 1: Framework of our method.

segments as the training set and the rest 3500 segments are used as the testing set. All the experiments are carried out on the same training and testing set.

2.3. Multichannel Analysis. Studies have shown that the correlation structure of all pairs of EEG channels could reflect the spatiotemporal evolution of electrical ictal activities [22–24]. By characterizing the spatiotemporal patterns, it is possible to identify seizures and analyze seizure dynamics.

In this study, we adopt cross-power matrix [10] to reflect the spatial patterns of the brain. For each time window with N channels, the cross-power matrix \mathbf{A} is $N \times N$. Each element a_{ij} in \mathbf{A} is defined by the cross-power [10] between the two EEG channels i and j in a given frequency band of $[lb, ub]$ as follows:

$$a_{ij} = \int_{lb}^{ub} P_{ij}(\omega) d\omega, \quad (1)$$

where $P_{ij}(\omega)$ is the cross-power spectral density of channels i and j at frequency ω .

2.4. Frequency Band Selection. Considering the diversity of epileptic patterns among patients, we choose the frequency band patient specifically from theta (4–7 Hz), alpha (8–13 Hz), and beta (14–30 Hz) bands. In order to select the frequency band that best reflects the difference between seizure and nonseizure states, we adopt Fisher's discriminant ratio (FDR) [25] as the criterion as follows:

$$C = \frac{(\mu_s - \mu_n)^2}{\sigma_s^2 + \sigma_n^2}, \quad (2)$$

where μ_s and σ_s^2 are means and covariance, respectively, of cross-power matrix of seizure segments and μ_n and σ_n^2 are those of nonseizure segments. For each patient, only the training segments are utilized for frequency band selection, and the frequency band with the highest FDR is used for seizure detection. The frequency band selected for each patient is shown in Table 1.

2.5. Robust Stacked Autoencoder. After multichannel analysis, each time window is represented by a cross-power matrix of $N \times N$, where N denotes the number of EEG channels. We propose to employ robust stacked autoencoders to extract reliable and compact features from the cross-power matrix.

In this section, first, we briefly introduce the basic autoencoder. Then, the robust autoencoder with MCC is presented to improve the feature learning ability under noises. Finally, we stack the robust autoencoders into a deep model for compact feature extraction.

2.5.1. Basic Autoencoder. Here, we begin with the traditional standard stacked autoencoder model (S-SAE). An autoencoder is a three-layer artificial network including an encoder and a decoder. The encoder takes an input vector \mathbf{x} and maps it to a hidden representation \mathbf{x}' through a nonlinear function as follows:

$$\mathbf{x}' = s(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}), \quad (3)$$

where $s(\cdot)$ is the sigmoid function. Suppose \mathbf{x} and \mathbf{x}' are d -dimensional and d' -dimensional vectors, respectively; then $\mathbf{W}^{(1)}$ is a $d' \times d$ weight matrix and $\mathbf{b}^{(1)}$ is a d' -dimensional bias vector.

Then, the vector \mathbf{x}' is mapped back to a reconstruction vector \mathbf{y} by the decoder as follows:

$$\mathbf{y} = s(\mathbf{W}^{(2)}\mathbf{x}' + \mathbf{b}^{(2)}), \quad (4)$$

where the output vector is d -dimensional, $\mathbf{W}^{(2)}$ is $d \times d'$, and $\mathbf{b}^{(2)}$ is a d -dimensional bias vector.

The parameter set $\theta = \{\mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \mathbf{W}^{(2)}, \mathbf{b}^{(2)}\}$ is optimized by minimizing the average reconstruction error as follows:

$$\theta = \operatorname{argmin}_{\theta} \frac{1}{n} \sum_{i=1}^n L(\mathbf{x}_i, \mathbf{y}_i), \quad (5)$$

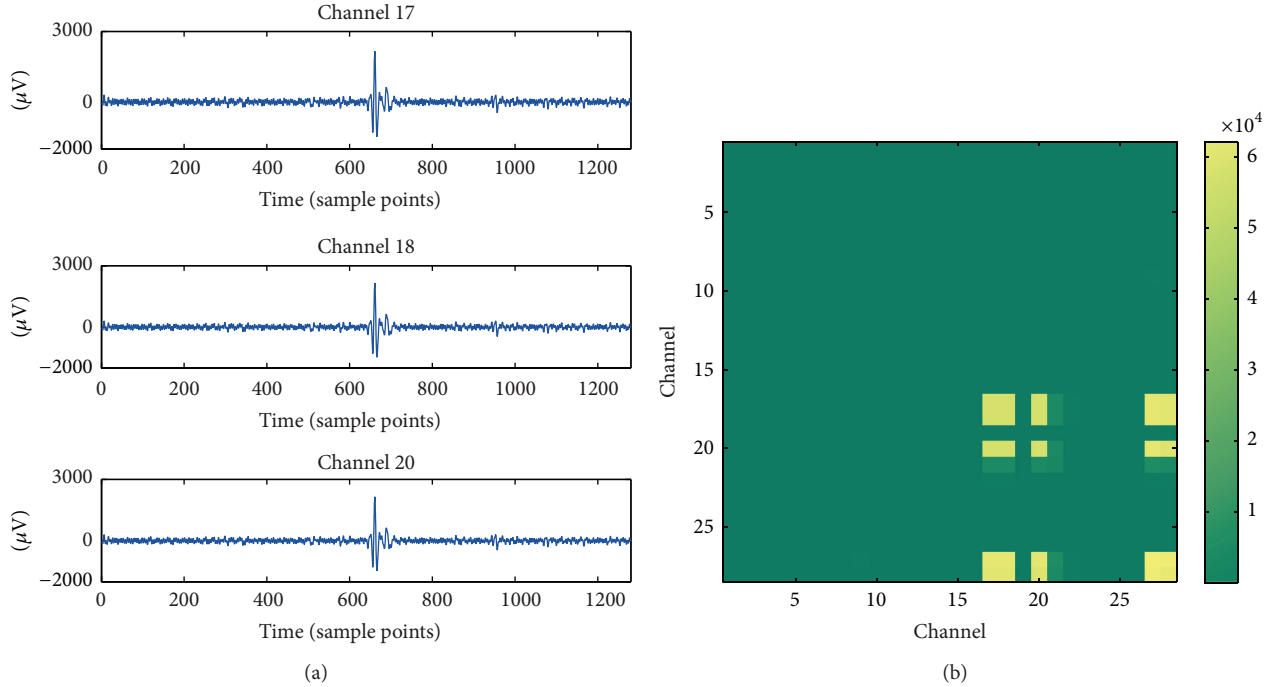


FIGURE 2: An EEG segment with impulsive noises. (a) EMG artifacts cause short-term burst noises in some channels of EEG signal; (b) visualization of the cross-power matrix of the segment with noises. The vertical and horizontal axes denote the channels and each *point* (i, j) in this figure is the cross-power value of channel i and channel j . The cross-power matrix contains outliers with large values. Because of the noise, the cross-power between channel 17 and channel 18 is far away from the interquartile range value (5.41×10^4 versus 395.3).

where L is the loss function. Mostly, the mean square error (MSE) is used as

$$\begin{aligned}
 J_{\text{MSE}}(\theta) &= \frac{1}{n} \sum_{i=1}^n L_{\text{MSE}}(\mathbf{x}_i, \mathbf{y}_i) \\
 &= \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{2} \|\mathbf{y}_i - \mathbf{x}_i\|^2 \right).
 \end{aligned} \tag{6}$$

2.5.2. Robust Autoencoder. The traditional autoencoder model based on MSE loss is not suitable for stable feature learning in EEG signals. In EEG, especially in scalp EEG signals, the large amount of noises caused by EMG artifacts or electrode movements could bring abrupt changes in EEG signal and lead to outliers in both time and frequency domain. A typical example is shown in Figure 2. In this time window, the EEG signals are noised by short-term EMG artifacts which lead to abrupt large-amplitude vibrations in some of the channels as shown in Figure 2(a). In the cross-power domain, such artifacts lead to outlying large values as in the light blocks in Figure 2(b). In the example illustrated, the cross-power between channel 17 and channel 18 is 5.41×10^4 , which is far away from the interquartile range value of 395.3. In this situation, the MSE-based cost of the traditional autoencoder model could be dominated by these outliers so that the feature learning ability is weakened.

In order to learn robust features from EEG signals, we replace the loss function of the autoencoder model with correntropy-based criterion to build robust autoencoder.

Maximum Correntropy Criterion. Correntropy is defined as a localized similarity measure [26] and it has shown good outlier suppression ability in studies [27, 28]. For two random variables X and Y , the correntropy is defined as

$$V_{\sigma}(X, Y) = E[\kappa_{\sigma}(X - Y)], \tag{7}$$

where $E[\cdot]$ is the mathematical expectation and $\kappa_{\sigma}(\cdot)$ is the Gaussian kernel with kernel size of σ as follows:

$$\kappa_{\sigma}(\cdot) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\cdot)^2}{2\sigma^2}\right). \tag{8}$$

The correntropy induces a new metric that, as the distance between X and Y gets larger, the equivalent distance evolves from 2-norm to 1-norm and eventually to zero-norm when X and Y are far apart [29]. Compared with second-order statistics such as MSE, correntropy is less sensitive to outliers. Figure 3 compares the second-order cost and correntropy cost. As the input x goes further from the center, the second-order cost increases sharply, so that it is sensitive to outliers. By contrast, the correntropy is only sensitive in a local range and the increase of the cost is extremely slow when the input value goes out of the central area. Therefore, the correntropy measure is particularly effective in outlier suppression.

In practice, the joint probability density function is unknown and usually only a finite set of samples of

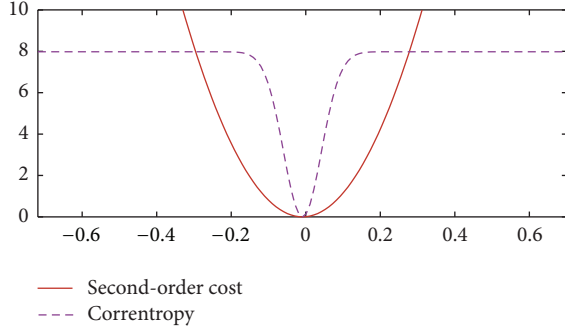


FIGURE 3: Illustration of second-order cost (red solid line) and correntropy cost (purple dashed line).

$\{(x_i, y_i)\}_{i=1}^N$ is available for both X and Y ; then the estimated correntropy can be calculated by

$$\tilde{V}_\sigma(X, Y) = \frac{1}{N} \sum_{i=1}^N \kappa_\sigma(x_i - y_i). \quad (9)$$

The maximum of correntropy error in (9) is called the maximum correntropy criterion (MCC) [29]. Due to the good outlier rejection property of correntropy, MCC is suitable for robust algorithm design.

Robust Autoencoder Based on MCC. In order to improve the antinoise ability of traditional autoencoders, we measure the reconstruction loss between the input vector \mathbf{x} and the output vector \mathbf{y} by MCC instead of MSE. In the MCC-based robust autoencoder, the cost function J is formulated as

$$\begin{aligned} J_{\text{MCC}}(\theta) &= \frac{1}{n} \sum_{i=1}^n L_{\text{MCC}}(\mathbf{x}_i, \mathbf{y}_i) \\ &= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m \kappa_\sigma(x_i^j - y_i^j), \end{aligned} \quad (10)$$

where n is the number of training samples and m is the length of each training sample. The optimal parameter θ is obtained when $J_{\text{MCC}}(\theta)$ is maximized.

In order to encourage the deep model to capture more implicit patterns, a sparsity-inducing term is adopted. Studies of sparse coding have shown that the sparseness seems to play a key role in learning useful features [30, 31]. Xie et al. [32] combined the virtues of sparse coding and deep networks into a sparse stacked denoising autoencoder to achieve better feature learning and denoising performance. In our model, we regularize the reconstruction loss by a sparsity-inducing term defined as in [32] as follows:

$$J_{\text{sparse}}(\theta) = \beta \sum_{i=1}^{s_2} \text{KL}(\rho \parallel \hat{\rho}_i), \quad (11)$$

where β is the weight adjustment parameter, s_2 is the number of units in the second layer, $\hat{\rho}_i$ is the activation value for the i th hidden layer unit, and ρ is a small number. The sparsity-inducing term constrains that the value of $\hat{\rho}_i$ should be near ρ under Kullback-Leibler divergence.

Also, a weight decay term $J_{\text{weight}}(\theta)$ is added to avoid overfitting. It is defined as follows:

$$J_{\text{weight}}(\theta) = \frac{\lambda}{2} \sum_{l=1}^2 \sum_{i=1}^{s_l} \sum_{j=1}^{s_{l+1}} (w_{ji}^{(l)})^2, \quad (12)$$

where $w_{ji}^{(l)}$ represents an element in $W^{(l)}$, λ is the parameter to adjust the weight of $J_{\text{weight}}(\theta)$, and s_l denotes number of units in layer l . Therefore, the cost function of the proposed robust autoencoder is defined as

$$J_{\text{R-SAE}}(\theta) = -J_{\text{MCC}}(\theta) + J_{\text{weight}}(\theta) + J_{\text{sparse}}(\theta). \quad (13)$$

By minimizing the cost of $J_{\text{R-SAE}}(\theta)$, the parameter set θ could be optimized.

2.5.3. Stacking Robust Autoencoders into Deep Network. In order to learn more effective features for seizure classification, we stack the robust autoencoders into a deep model. Stacking the robust autoencoders works in the same way as stacking the ordinary autoencoders [17] and the output from the highest layer is cascaded to a softmax classifier for seizure detection. Such a model aims at the best seizure classification accuracy, and it is able to simultaneously optimize the feature and classifier.

The training process of the deep network includes two stages: unsupervised pretraining and supervised fine-tuning. In the pretraining stage, the network is trained layer-wisely by the proposed robust autoencoder model to learn useful filters for feature extraction. A well pretrained network yields a good starting point for fine-tuning [33]. In the fine-tuning stage, a softmax classifier is added to the output of the stack, and the parameters of the whole system are tuned to minimize the classification error in a supervised manner. The network is globally tuned through back-propagation and all the parameters of both feature extraction and classification are jointly optimized. After fine-tuning, the deep network is well configured to obtain optimal overall classification performance.

3. Results and Discussion

In this section, experiments are carried out to evaluate the seizure detection performance of our model. The experiments include four parts: (1) we compare the unsupervised feature learning performance of the modified R-SAE model and the standard stacked autoencoder (S-SAE); (2) we compare the features before and after supervised fine-tuning to demonstrate the strength of joint optimization; (3) we compare the seizure detection performance of R-SAE model with other methods; (4) we evaluate the influence of parameters in the R-SAE model on the seizure detection performance.

In our experiments, the seizure detection performance is evaluated with the two commonly used criteria, sensitivity and specificity. Sensitivity is defined as the percentage of true seizure segments detected and specificity is the proportion of nonseizure segments correctly classified.

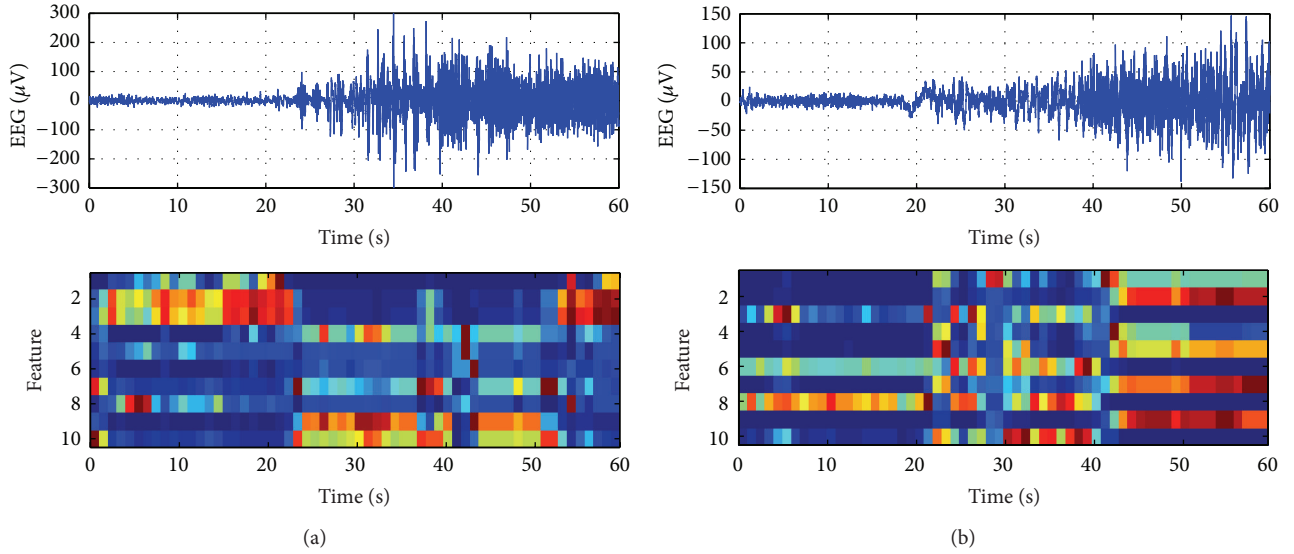


FIGURE 4: Unsupervised feature learning results by R-SAE model for patient pt03 (a) and pt04 (b). For each subfigure, the top is the original EEG signal from one channel and the bottom is the features extracted by the R-SAE model.

3.1. Performance of Feature Learning. In this experiment, we evaluate the unsupervised feature learning ability of the R-SAE model with EEG signals. In our method, we train the R-SAE model to learn compact features from the cross-power matrix. After the layer-wised self-taught training, the deep network is well configured to learn useful features. The feature extraction results of the proposed R-SAE model are illustrated in Figure 4. For both illustrations, the seizure begins at about the 20th second. After seizure onset, the patterns of features extracted by R-SAE model show clear differences from nonseizure ones.

The feature learning performance of R-SAE and S-SAE is compared using EEG signal. In order to evaluate the ability of the features quantitatively, we utilize the classification performance as the criterion. In this experiment, the cost function of the S-SAE model is as follows:

$$J_{\text{S-SAE}}(\theta) = J_{\text{MSE}}(\theta) + J_{\text{weight}}(\theta) + J_{\text{sparse}}(\theta), \quad (14)$$

where the loss function $J_{\text{MSE}}(\theta)$ is formulated with MSE-based loss function as in (6) and $J_{\text{weight}}(\theta)$ and $J_{\text{sparse}}(\theta)$ are formulated the same as R-SAE.

We stack two autoencoders to constitute a three-layer network with 784 input units, 50 hidden units, and 10 output units. The same stacked architectures are applied for both R-SAE and S-SAE. The networks are initialized randomly and trained layer-wisely using back-propagation to minimize the cost functions. The parameters are set as $\lambda = 0.003$, $\beta = 3$, and $\rho = 0.1$ for both methods and $\sigma = 0.05$ for R-SAE.

The seizure detection results of both R-SAE model and S-SAE model are shown in Table 2. In order to eliminate the effects of randomness in network initialization, we present all the results averaged over 10 trials. Results show that the average sensitivity of R-SAE is 97%, which demonstrates 14% improvement compared with S-SAE. With specificity, the average result is 92% for R-SAE which is also higher than that

of S-SAE. Thus, R-SAE outperforms S-SAE in both sensitivity and specificity.

In the analysis of the detection results, we find that S-SAE fails mostly on EEG segments with impulsive noises such as the segment illustrated in Figure 2. Since such abrupt artifacts could appear frequently in EEG signals, the S-SAE model could not be well trained because the MSE-based cost could be dominated by the large outliers. Thus, these EEG segments could not be well represented by the S-SAE model. By contrast, the MCC in the R-SAE model is more robust to large outliers. Therefore, the proposed R-SAE method could handle noises in EEG signal well, and it provides more robust feature extraction performance than S-SAE.

3.2. Performance of Joint Feature Optimization. In this experiment, we test the effects of joint feature optimization. After the MCC-based unsupervised learning, the deep network is well configured to extract useful features from EEG signals. On this basis, the deep model is fine-tuned through back-propagation to jointly optimize both feature and classifier, so that the optimal overall classification performance could be achieved. In this experiment, the parameters of R-SAE are set the same as in Section 3.1 that only the unit number of the output layer is set to 3 for visualization convenience.

The visual comparison of features before and after fine-tuning is illustrated in Figure 5. In Figures 5(a) and 5(b), the red circles denote features of seizure segments while the blue stars are nonseizure ones. It can be seen that, after fine-tuning, the seizure and nonseizure segments are more separable in the feature space. We quantitatively analyze the separability of the features before and after fine-tuning with the FDR criterion as in (2) using the first four patients. As illustrated in Figure 5(c), the fine-tuned features achieve about ten times higher FDR than do the original ones, which strongly indicates that the joint optimization could help to

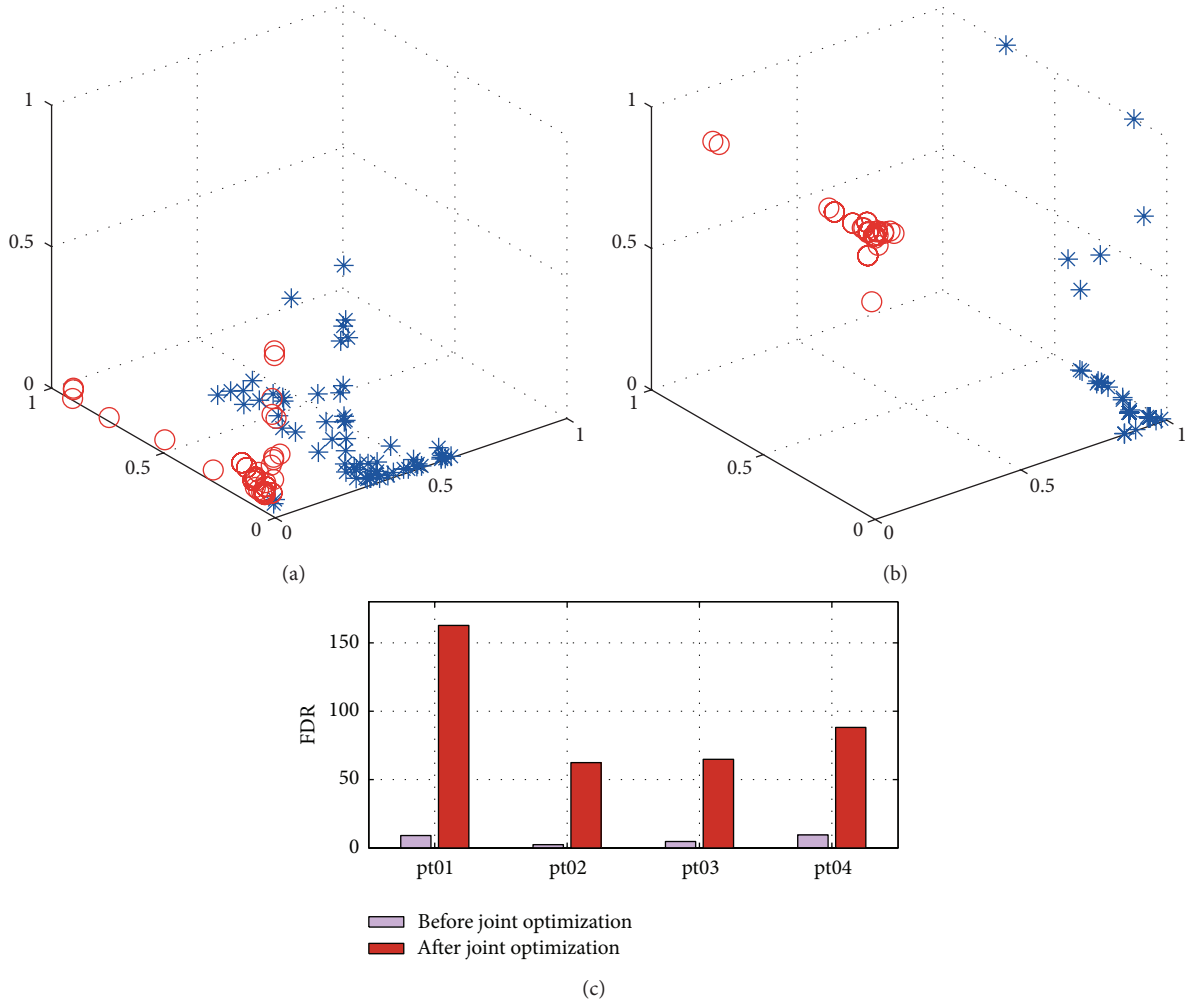


FIGURE 5: Comparison between features before and after joint optimization. (a-b) Visualization of features for seizure and nonseizure segments. The red circles denote features of seizure segments while the blue stars are nonseizure ones. (c) The FDR value of features before and after joint optimization.

TABLE 2: Comparison between R-SAE and S-SAE (before fine-tuning).

Patient	R-SAE		S-SAE	
	Sensitivity	Specificity	Sensitivity	Specificity
Pt01	$0.99 \pm 8.1 \times 10^{-3}$	$0.96 \pm 2.0 \times 10^{-2}$	$0.83 \pm 2.0 \times 10^{-1}$	$0.92 \pm 3.2 \times 10^{-2}$
Pt02	$0.96 \pm 2.0 \times 10^{-2}$	$0.91 \pm 1.7 \times 10^{-2}$	$0.90 \pm 1.0 \times 10^{-1}$	$0.83 \pm 7.6 \times 10^{-2}$
Pt03	$0.96 \pm 1.5 \times 10^{-2}$	$0.93 \pm 2.1 \times 10^{-2}$	$0.82 \pm 2.8 \times 10^{-1}$	$0.93 \pm 3.2 \times 10^{-2}$
Pt04	$0.95 \pm 3.7 \times 10^{-2}$	$0.91 \pm 1.3 \times 10^{-2}$	$0.91 \pm 4.1 \times 10^{-2}$	$0.92 \pm 2.6 \times 10^{-2}$
Pt05	$0.98 \pm 1.7 \times 10^{-2}$	$0.94 \pm 1.5 \times 10^{-2}$	$0.70 \pm 1.3 \times 10^{-1}$	$0.96 \pm 2.4 \times 10^{-2}$
Pt06	$0.97 \pm 2.9 \times 10^{-2}$	$0.84 \pm 5.5 \times 10^{-2}$	$0.82 \pm 1.7 \times 10^{-1}$	$0.90 \pm 1.5 \times 10^{-2}$
Avg.	$0.97 \pm 1.3 \times 10^{-2}$	$0.92 \pm 3.8 \times 10^{-2}$	$0.83 \pm 6.9 \times 10^{-2}$	$0.91 \pm 4.0 \times 10^{-2}$

learn superior features with high separability, so that the seizure detection performance could be improved.

The seizure detection performance of features before and after fine-tuning is presented in Table 3. After joint feature learning, the average sensitivity of six patients increases by 8% and the specificity increases by 15%. Therefore, the joint learning process enhances the separability of features

between the two classes and greatly facilitates seizure detection performance.

3.3. Performance of Seizure Detection. In this experiment, seizure detection performance of the proposed R-SAE model is evaluated and compared with singular value decomposition-(SVD-) based method. The SVD method is the most popular

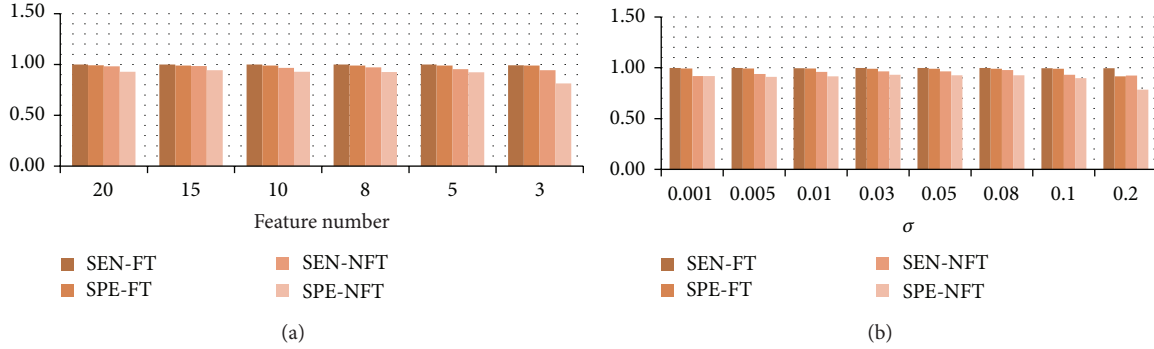


FIGURE 6: Model analysis of two important parameters of R-SAE. (a) Seizure detection performance under different feature numbers; (b) seizure detection performance with different selections of σ . In this figure, SEN-FT and SPE-FT are sensitivity and specificity after fine-tuning and SEN-NFT and SPE-NFT are those before fine-tuning.

TABLE 3: Comparison of seizure detection performance before and after fine-tuning (FT).

Feature	Sensitivity	Specificity
Before FT	$0.90 \pm 1.0 \times 10^{-1}$	$0.84 \pm 1.4 \times 10^{-1}$
After FT	$0.98 \pm 3.1 \times 10^{-2}$	$0.99 \pm 4.7 \times 10^{-3}$

tool for correlation matrix analysis. Studies have shown that the seizure EEG signals commonly lead to a lower-complexity state which could be well reflected by the eigenvalues from SVD of the correlation matrix [10, 22].

To provide a benchmark for the comparison, we also test the seizure detection performance with the original cross-power matrix without further feature extraction. The methods included in the comparison are configured as follows.

- (i) *SVM*: in SVM, the cross-power matrices of time windows are reshaped to vectors and fed into an SVM classifier with RBF kernel. The parameters of the SVM model are selected using 3-fold cross-validation.
- (ii) *SVD(p) + SVM*: for each time window, the cross-power matrix is decomposed by SVD, and the first p eigenvalues are adopted as the features. The feature vectors are then classified by an SVM classifier with RBF kernel. The parameters of the SVM model are selected using 3-fold cross-validation.
- (iii) *R-SAE(q)*: the R-SAE model is configured with 784 input units, 50 hidden units, and q output units. The parameters are set as $\lambda = 0.003$, $\beta = 3$, $\rho = 0.1$, and $\sigma = 0.05$. For this method, all results are averaged over 10 trials.

The seizure detection results of the three methods are given in Table 4. For both SVD + SVM and R-SAE, we test the seizure detection performance under two different choices of parameters of p and q , respectively. Results show that, with the original cross-power matrix classified by SVM, high sensitivities of above 0.99 are achieved for all six patients and the average specificity is 0.91. By the SVD + SVM method with $p = 3$, uneven performance is shown in different patients. For pt03, high sensitivity of 0.96 is reached with

0.99 of specificity. However, low sensitivities are obtained for pt01, pt05, and pt06. For SVD + SVM method with $p = 10$ where more features are preserved, better sensitivities and specificities are achieved. However, the uneven performance over patients still exists, and the average sensitivity is only 0.83. Since the feature extraction process of the SVD-based method loses much useful information, lower performance is obtained compared with SVM benchmark. Besides, the seizure detection performance sees a decrease when fewer eigenvalues are used. By contrast, the proposed R-SAE method achieves better performance than the benchmark SVM method. In R-SAE with $q = 10$, high sensitivities of 1.00 and specificities of 0.99 are achieved for all patients. Equally high performance is obtained with $p = 3$. The R-SAE model keeps robust seizure detection ability even with such small dimension of features.

3.4. Model Analysis. In this experiment, we test the influence of the two important parameters on the seizure detection performance. The first parameter is the output feature number, that is, the number of units of the output layer of the R-SAE model, and the second parameter is the kernel size σ in MCC. The experiment is carried out using the first four patients.

3.4.1. Analysis of Feature Number. The feature number is tuned by the parameter q in Section 3.3. In order to test the influence of q on seizure detection, all the other parameters are fixed as in Section 3.3 and we gradually tune q from 20 to 3. Figure 6(a) illustrates the seizure detection results averaged over four patients under different choices of q . The result shows that the seizure detection performance of R-SAE before fine-tuning sees a slight decrease with the decrease of feature number. However, after the fine-tuning, the seizure detection performance is greatly enhanced that high sensitivities and specificities up to 99% are achieved even with small feature numbers.

3.4.2. Analysis of σ . In the MCC, the kernel size σ serves as an important parameter that an appropriate choice of σ can effectively suppress the outliers and noises. The kernel size or bandwidth is a free parameter that its selection is still an

TABLE 4: Comparison with other methods.

Method	Pt01		Pt02		Pt03		Pt04		Pt05		Pt06		Avg	
	SEN*	SPE*	SEN	SPE	SEN	SPE	SEN	SPE	SEN	SPE	SEN	SPE	SEN	SPE
SVM	1.00	0.96	1.00	0.89	1.00	0.93	0.99	0.95	1.00	0.96	1.00	0.78	1.00	0.91
SVD(3) + SVM [10]	0.45	1.00	0.72	0.99	0.96	0.99	0.84	0.95	0.46	0.98	0.64	0.97	0.68	0.98
SVD(10) + SVM [10]	0.61	1.00	0.76	0.99	0.99	1.00	0.80	0.95	0.84	0.93	0.95	0.96	0.83	0.97
R-SAE(3)	1.00	0.99	0.99	0.99	0.98	0.99	1.00	0.99	1.00	0.99	0.92	0.98	0.98	0.99
R-SAE(10) (ours)	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.99	0.99	0.97	1.00	0.99

*SEN indicates sensitivity and SPE is specificity.

open issue in ITL [26, 29, 34]. In practice, the parameter σ can be selected with Silverman's rule [35]. In the experiments of Sections 3.1–3.3, we simply set $\sigma = 0.05$.

Here, we test the influence of parameter σ on overall seizure detection performance. Also, all the other parameters are fixed as in Section 3.3. Figure 6(b) illustrates the seizure detection results under different selections of σ averaged over four patients. Results show that high seizure detection performance could be achieved under a wide choice of σ . Better results are obtained with small σ , and when σ increases from 0.1 to 0.2, the seizure detection performance becomes worse. In practice, the choice of σ should be small to keep good local property of the MCC.

4. Conclusions

In this paper, we have presented a novel deep model which is capable of extracting robust features under large amounts of outliers. Experimental results show that the proposed R-SAE model could learn effective features in EEG signals for high performance seizure detection, and it is promising for clinical applications.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research was supported by Grants from the National Natural Science Foundation of China (no. 61031002), National 973 Program (no. 2013CB329500), National High Technology Research and Development Program of China (no. 2012AA020408), National Natural Science Foundation of China (no. 61103107), and Zhejiang Provincial Science and Technology Project (no. 2013C03045-3).

References

- [1] *Epilepsy*, Factsheet no. 999, World Health Organization, Geneva, Switzerland, 2012.
- [2] R. S. Fisher, W. Van Emde Boas, W. Blume et al., "Epileptic seizures and epilepsy: definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE)," *Epilepsia*, vol. 46, no. 4, pp. 470–472, 2005.
- [3] F. Mormann, R. G. Andrzejak, C. E. Elger, and K. Lehnertz, "Seizure prediction: the long and winding road," *Brain*, vol. 130, no. 2, pp. 314–333, 2007.
- [4] M. E. Saab and J. Gotman, "A system to detect the onset of epileptic seizures in scalp EEG," *Clinical Neurophysiology*, vol. 116, no. 2, pp. 427–442, 2005.
- [5] K. Majumdar and P. Vardhan, "Automatic seizure detection in ECoG by differential operator and windowed variance," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 19, no. 4, pp. 356–365, 2011.
- [6] R. Yadav, A. K. Shah, J. A. Loeb, M. N. S. Swamy, and R. Agarwal, "Morphology-based automatic seizure detector for intracerebral EEG recordings," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 7, pp. 1871–1881, 2012.
- [7] S. Ghosh-Dastidar, H. Adeli, and N. Dadmehr, "Principal component analysis-enhanced cosine radial basis function neural network for robust epilepsy and seizure detection," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 2, pp. 512–518, 2008.
- [8] A. S. Zandi, M. Javidan, G. A. Dumont, and R. Tafreshi, "Automated real-time epileptic seizure detection in scalp EEG recordings using an algorithm based on wavelet packet transform," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1639–1651, 2010.
- [9] C. C. Jouny and G. K. Bergey, "Characterization of early partial seizure onset: Frequency, complexity and entropy," *Clinical Neurophysiology*, vol. 123, no. 4, pp. 658–669, 2012.
- [10] S. Santaniello, S. P. Burns, A. J. Golby, J. M. Singer, W. S. Anderson, and S. V. Sarma, "Quickest detection of drug-resistant seizures: an optimal control approach," *Epilepsy and Behavior*, vol. 22, supplement 1, pp. S49–S60, 2011.
- [11] D. Liu and Z. Pang, "Epileptic seizures predicted by modified particle filters," in *Proceedings of the IEEE International Conference on Networking, Sensing and Control (ICNSC08)*, pp. 351–356, IEEE, Sanya, China, April 2008.
- [12] D. Liu, Z. Pang, and Z. Wang, "Epileptic seizure prediction by a system of particle filter associated with a neural network," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, Article ID 638534, 2009.
- [13] P. Vincent, H. Larochelle, Y. Bengio, and P. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th International Conference on Machine Learning*, pp. 1096–1103, ACM, July 2008.
- [14] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

- [15] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *The American Association for the Advancement of Science. Science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [16] Y. Boureau and Y. Cun, "Sparse feature learning for deep belief networks," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 1185–1192, 2007.
- [17] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–27, 2009.
- [18] R. Salakhutdinov, J. B. Tenenbaum, and A. Torralba, "Learning with hierarchical-deep models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1958–1971, 2013.
- [19] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th International Conference on Machine Learning (ICML '09)*, pp. 609–616, Montreal, Canada, June 2009.
- [20] G. Hinton, L. Deng, D. Yu et al., "Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.
- [21] H. Lee, L. Yan, P. Pham, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS '09)*, vol. 9, pp. 1096–1104, December 2009.
- [22] K. Schindler, H. Leung, C. E. Elger, and K. Lehnertz, "Assessing seizure dynamics by analysing the correlation structure of multichannel intracranial EEG," *Brain*, vol. 130, no. 1, pp. 65–77, 2007.
- [23] K. A. Schindler, S. Bialonski, M. Horstmann, C. E. Elger, and K. Lehnertz, "Evolving functional network properties and synchronizability during human epileptic seizures," *Chaos*, vol. 18, no. 3, Article ID 033119, 2008.
- [24] C. Rummel, M. Müller, G. Baier, F. Amor, and K. Schindler, "Analyzing spatio-temporal patterns of genuine cross-correlations," *Journal of Neuroscience Methods*, vol. 191, no. 1, pp. 94–100, 2010.
- [25] B. Scholkopf and K. Mullert, "Fisher discriminant analysis with kernels," 1999.
- [26] L. Weifeng, P. P. Pokharel, and J. C. Principe, "Correntropy: a localized similarity measure," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '06)*, pp. 4919–4924, July 2006.
- [27] K. Jeong, W. Liu, S. Han, E. Hasanbelliu, and J. C. Principe, "The correntropy MACE filter," *Pattern Recognition*, vol. 42, no. 5, pp. 871–885, 2009.
- [28] R. He, B. Hu, W. Zheng, and X. Kong, "Robust principal component analysis based on maximum correntropy criterion," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1485–1494, 2011.
- [29] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: properties and applications in non-Gaussian signal processing," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5286–5298, 2007.
- [30] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.
- [31] H. Lee, A. Battle, R. Raina, and A. Ng, "Efficient sparse coding algorithms," *Advances in Neural Information Processing Systems*, vol. 19, pp. 801–808, 2007.
- [32] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS '12)*, vol. 25, pp. 350–358, December 2012.
- [33] P. Vincent, H. Larochelle, I. Lajoie, and P. Manzagol, "Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.
- [34] R. He, W. Zheng, B. Hu, and X. Kong, "A regularized correntropy framework for robust pattern recognition," *Neural Computation*, vol. 23, no. 8, pp. 2074–2100, 2011.
- [35] B. Silverman, *Density Estimation for Statistics and Data analysis*, vol. 26, CRC Press, 1986.