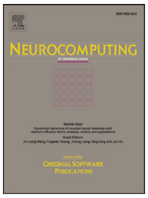
Neurocomputing 324 (2019) 31–42

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



# Wave2Vec: Deep representation learning for clinical temporal data

Ye Yuan[a,b,c,*], Guangxu Xun[d], Qiuling Suo[d], Kebin Jia[a,b,c], Aidong Zhang[d]

[a] *College of Information and Communication Engineering, Beijing University of Technology, Beijing 100124, China*
[b] *Beijing Laboratory of Advanced Information Networks, Beijing University of Technology, Beijing 100124, China*
[c] *Advanced Innovation Center for Future Internet Technology, Beijing University of Technology, Beijing 100124, China*
[d] *Department of Computer Science and Engineering, State University of New York at Buffalo, NY 100124, USA*

ARTICLE INFO

*Article history:*
Received 13 December 2017
Revised 22 February 2018
Accepted 12 March 2018
Available online 22 May 2018

*Keywords:*
Deep learning
Semantic learning
Representation learning
Biosignals

## ABSTRACT

Representation learning for time series has gained increasing attention in healthcare domain. The recent advancement in semantic learning allows researcher to learn meaningful deep representations of clinical medical concepts from Electronic Health Records (EHRs). However, existing models cannot deal with continuous physiological records, which are often included in EHRs. The major challenges for this task are to model non-obvious representations from observed high-resolution biosignals, and to interpret the learned features. To address these issues, we propose Wave2Vec, an end-to-end deep representation learning model, to bridge the gap between biosignal processing and semantic learning. Wave2Vec not only jointly learns both inherent and temporal representations of biosignals, but also allows us to interpret the learned representations reasonably over time. We propose two embedding mechanisms to capture the temporal knowledge within signals, and discover latent knowledge from signals in time-frequency domain, namely component-based motifs. To validate the effectiveness of our model in clinical task, we carry out experiments on two real-world benchmark biosignal datasets. Experimental results demonstrate that the proposed Wave2Vec model outperforms six feature learning baselines in biosignal processing. Analytical results show that the proposed model can incorporate both motif co-occurrence information and time series information of biosignals, and hence provides clinically meaningful interpretation.

## 1. Introduction

Recently, representation learning for time series has gained great attention in many scientific disciplines, including human activity recognition [1], natural language processing (NLP) [2], and protein localization [3]. Among them, in the healthcare domain, learning meaningful representations for complex clinical time series, such as Electronic Health Records (EHRs), has become a key challenge for a variety of applications [4,5]. The recent advances in deep representation learning enable researchers to capture the semantics of health data. For instance, Med2Vec [6] is able to learn interpretable representations for medical codes from EHRs using a neural-network-based architecture. Med2Vec is inspired by the idea of word embeddings [7], since the medical codes (e.g., medication, diagnosis, and procedure codes) are discrete and can be directly treated as words. However, it is not applicable to continuous physiological records. With the development of pervasive sensing technologies, continuous time series data captured from sensors are often included in EHRs. The advanced sensors, including implantable, wearable, and ambient sensors [8], allow for continuous monitoring to prevent serious outcomes caused by several medical diseases. Waveform biosignals, such as electroencephalogram (EEG), electrocardiogram (ECG), Electromyography (EMG), and Electrooculography (EOG), contain hidden information about physiological phenomena and thus reflect human health and wellbeing [9]. This necessitates the development of semantic analysis tools for continuous biosignal data to discover latent patterns and their relationships.

Word embeddings represent each word as a fixed-length vector, and words with similar semantics would be mapped to close positions in the vector space by learning the context information. However, unlike in text data (or medical codes), in continuous time series data, the notion of a word is non-obvious. Thus, the major challenge is to find a way to represent words from continuous signals. One simple way to tackle this is that a word could

[*] Corresponding author at: College of Information and Communication Engineering, Beijing University of Technology, Beijing 100124, China.

*E-mail addresses:* yuanye91@emails.bjut.edu.cn (Y. Yuan), kebinj@bjut.edu.cn (K. Jia).

https://doi.org/10.1016/j.neucom.2018.03.074
0925-2312/© 2018 Elsevier B.V. All rights reserved.

be directly specified as a segmented window of data itself. But the fact is that most continuous valued time series segments do not repeat exactly [10]. This renders discovering similar patterns from continuous time series data a popular research topic. In recent studies, on the one hand, some researchers defined these patterns as time series motifs [11,12]. They applied machine learning methods to discover motifs which are similar subsequences or frequently occurring patterns. On the other hand, some researchers [13–15], proposed various deep learning methods to learn a dictionary of biosignals. In particular, they assumed that each signal fragment consists of several base signal patterns (motifs) drawn from multinomial distributions. The signal dictionary is a set of all high-order signal words (base signal patterns), and each fragment can be viewed as a combination of signal words in the dictionary. In this way, a signal fragment can be sampled as a single word according to the normalized feature weights. In this paper, we follow the second view since biosignal is often formed by several template waveforms, and hence such component-based motifs are more reasonable than the subsequence-based motifs in healthcare domain.

Despite the promising results reported by these feature learning approaches, there are still several challenges to be addressed. First, since the existing deep learning-based dictionary methods learn representations separately, they cannot guarantee consistent good performance with such multi-stage training procedure to make the components work together. Second, it is infeasible to interpret the representations of biosignals learned by standard deep learning models, while the interpretability of the representation in healthcare applications is essential. Finally, the time-domain patterns in biosignals vary significantly across patients over time, rendering it a challenging task to develop a cross-patient feature extractor.

To tackle the aforementioned challenges, in our previous work [16], we proposed a preliminary model, namely Wave2Vec, to jointly learn inherent and temporal features of biosignals. Specifically, we first utilize a fixed-length sliding window to segment the entire biosignals into fragments, and adopt wavelet transform as preprocessing to express time-frequency information. We then train a unified model, which consists of an inherent layer and an embedding layer, to maintain both latent and temporal characteristics of the biosignals. Moreover, both deep features and signal handcrafted features are taken into consideration by the proposed model.

In this paper, we further extend our previous work [16] to learn low dimensional representations of biosignal patterns in an unsupervised fashion. We experiment with two types of semantic learning mechanisms: (i) score-based, and (ii) softmax-based, to calculate the temporal relationship between signal fragments. We then adopt an interpretation module to uncover latent patterns of biosignals in time-frequency domain, named component-based motifs, and their semantic meaning in real-world clinical tasks. We demonstrate that the proposed Wave2Vec model achieves better performance compared to the state-of-the-art feature learning methods on two real-world biosignal datasets. Finally, we conduct qualitative analysis to evaluate the interpretability of the learned representations. In summary, compared with our preliminary model of Wave2Vec [16], the main contributions of this paper are as follows:

- We propose Wave2Vec, an end-to-end interpretable deep learning model, to jointly learn inherent and temporal features of biosignals. We employ two embedding mechanisms to calculate the semantic relationship between signal fragments. This method bridges the gap between signal (biosignal) processing and semantic learning.
- Wave2Vec models biosignals in time-frequency domain, and employs a two-level interpretation module to discover component-based motifs and analyze which ones are crucial to clinical tasks.
- We empirically demonstrate that the proposed Wave2Vec outperforms existing feature learning methods on two real-world biosignal datasets. The analytical results show that the learned interpretable representations can identify the influential clinical concepts of biosignals from both motif and embedding levels.

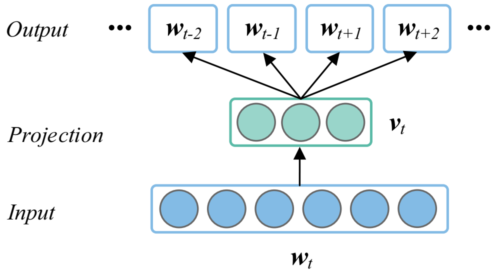


**Fig. 1.** Illustration of the Skip-gram model. $v_t$ is a vector representation of word $w_t$. The basic idea of the Skip-gram model is to learn word representations that are useful for predicting the surrounding words inside a context window of a predefined size.

The rest of the paper is organized as follows: In the next section, we introduce the preliminary ideas. Our proposed methodology is then described in Section 3. Section 4 presents and discusses the experimental results for our method followed by the related works in Section 5. Then we conclude this paper in Section 6.

## 2. Background

Before diving into the details of Wave2Vec, we first introduce two preliminary ideas used in our biosignals representation learning method.

### 2.1. Skip-gram for representation learning

Learning word representations using neural networks has garnered great attention recently since the learned vectors are able to capture linguistic regularities and patterns in language. Among these works, Mikolov et al. [7] proposed an efficient method for learning high-quality vector representations of words from text data, namely Skip-gram.

Fig. 1 illustrates the basic structure of the Skip-gram model. Formally, given a sequence of training words $w_1, w_2, \ldots, w_T$, the objective function is to maximize the following average log probability:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\ne 0}\log p(w_{t+j}|w_t), \tag{1}$$

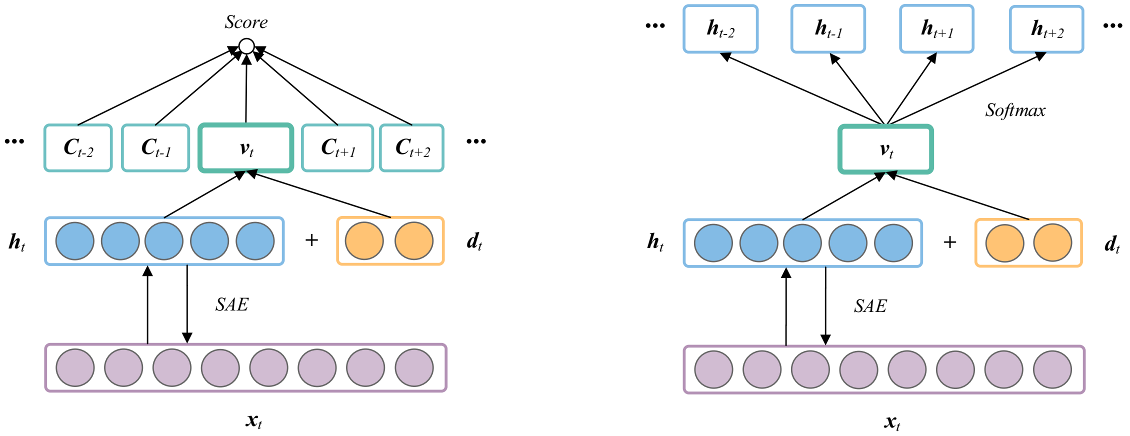
where $c$ denotes the context window size, and the conditional probability $\log p(w_{t+j}|w_t)$ is defined by the softmax function:

$$p(w_O|w_I) = \frac{\exp v_{w_O}^{\prime T} v_{w_I}}{\sum_{w=1}^{W}\exp v_{w_O}^{\prime T} v_{w_I}}, \tag{2}$$

where $v_w$ and $v_w^{'}$ are the input and output vector representations of word $w$, respectively. Here $W$ denotes the number of words in the vocabulary.

### 2.2. Deep learning for clinical tasks

In clinical settings, disease diagnosis through visual inspection demands highly professional knowledge. In addition, long-term



(a) The Wave2Vec$_{sc}$ model (b) The Wave2Vec$_{so}$ model

**Fig. 2.** Schematic illustrations of the proposed Wave2Vec model. (a) represents the structure of the Wave2Vec$_{sc}$ model and (b) represents the structure of the Wave2Vec$_{so}$ model.

biosignal visual inspection is extremely time-consuming and laborious for physicians [17]. This has motivated researchers to design automated machine learning approaches for several clinical tasks [18,19].

Deep learning is a machine learning technique founded on learning data representations. The typical architecture of deep learning is a multi-layer neural network and the extracted features become progressively abstract as the layer goes higher. The training strategy of deep learning consists of pre-training and fine-tuning [20]. A greedy layer-wise training method is adopted for pre-training each layer individually to initialize the weight matrix. Fine-tuning is then adopted to tune the weights of all layers simultaneously. The superiority of deep learning architectures over traditional hand-engineered approaches in terms of feature extraction has been validated by previous studies for a wide range of applications [21]. Therefore, numerous approaches have been proposed applying different deep learning models to various types of biosignals [22–24].

## 3. Methodology

In this section, we discuss our proposed Wave2Vec methodology. Wave2Vec contains two models, namely Wave2Vec$_{sc}$ and Wave2Vec$_{so}$, respectively. We first give an overview of the proposed models, then describe the details of the main components.

### 3.1. Framework

The motivation of our proposed algorithms arises from the inability of a single feature to reach the robust and accurate results. Fig. 2 describes the high-level overview of the two proposed Wave2Vec models. Specifically, given a signal segment $x_t$, this continuous data can be encoded into a low dimensional inherent representation $h_t$ using deep learning. In our model, we adopt a sparse autoencoder (SAE) layer to extract the latent characteristics of biosignals. The inherent representation $h_t$ is then concatenated with a vector of demographic information $d_t$, and converted to the final embedding representation $v_t$. For the Wave2Vec$_{sc}$ model, $v_t$ is trained to detect the relationship by scoring with its neighboring context representations $\{\ldots, c_{t-2}, c_{t-1}, c_{t+1}, c_{t+2}, \ldots\}$. For the Wave2Vec$_{so}$ model, $v_t$ is trained to predict its neighboring inherent representations $\{\ldots, h_{t-2}, h_{t-1}, h_{t+1}, h_{t+2}, \ldots\}$ through a softmax layer. The proposed two neural networks can be both trained end-to-end.

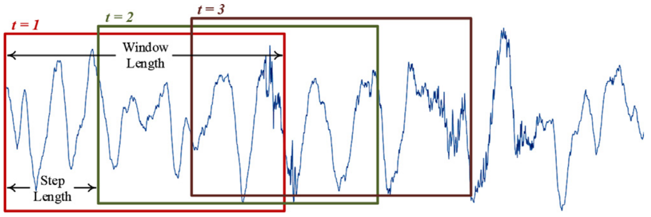


**Fig. 3.** Example of EEG segmentation, in which the length of the sliding window is fixed to $l = 3$ s and the step length $s = 1$ s.

### 3.2. Signal preprocessing

#### 3.2.1. Segmentation

The proposed Wave2Vec approach is designed for biosignal processing. Since the waveform cannot be explicitly segmented into sub-fragments associated with physiological meanings, in our model, we segment EEG signals into several fixed-length slots by sliding a $l$-length window with $s$-length step. Fig. 3 illustrates an example of EEG segmentation, in which the length of the sliding window is fixed to $l = 3$ s and the step length $s = 1$ s. It is worth noting that increasing window length $l$ can enhance feature representation while may cause delay in real-time applications.

#### 3.2.2. Scalogram

For biosignal processing, the sampled waveform data expressed in time-frequency domain are more meaningful than time domain. According to the paper [25], time-frequency waveform provides a sense of the occurrence of each frequency at each signal fragment, namely bags of frequencies, which is suitable for the subsequent deep representation learning model.

In our model, we use a similar preprocessing strategy, referred to scalogram, to express interpretable information of signals in time-frequency domain. In addition, scalogram also provides noise cancellation of signal over time and hence enables a more robust classifier against data shifting. Formally, given a signal fragment $x(t)$, the scalogram can be calculated by squaring the magnitude of complex-valued time-frequency transform (take wavelet [26] as example), as follows:

$$scalogram_x(a,\tau) = |CWT_x(a,\tau)|^2 = \left| \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\Psi^*\left(\frac{t-\tau}{a}\right)dt \right|^2, \quad (3)$$

where the asterisk is the function of complex conjugate, and $\Psi$ denotes the mother wavelet. The translation parameter $\tau$ determines its shifting position, and the dilation parameter $a$ determines the oscillatory frequency and the length of the wavelet. In this way, we can obtain the time-varied frequency content of biosignals, and further learn deep representations using our proposed model. In our model, Morlet is adopted as the mother wavelet to generate scalograms in the following sections.

### 3.3. Signal inherent representations

In the task of deep learning, learning on low dimensional latent characteristics can often avoid overfitting problem. In our model, we propose to unsupervisedly extract inherent features through a SAE layer to further express signals in a high-level space. In general, SAE is an unsupervised neural network that attempts to reconstruct the input by learning an encoder and a decoder with sparse constraint, respectively [20]. Formally, given a signal scalogram $x_t \in \mathbb{R}^n$, the inherent representation $h_t \in \mathbb{R}^s$ can be obtained based on forward-propagation:

$$h_t = f(W_h x_t + b_h), \quad (4)$$

where $W_h \in \mathbb{R}^{s\times n}$ and $b_h \in \mathbb{R}^s$ denote the weight matrix and bias vector, and $f(\cdot)$ is the activation function. The inherent representation $h_t$ is then mapped back into a reconstruction $\hat{x}_t$ of the same shape as $x_t$ through a similar transformation:

$$\hat{x}_t = f(W_h^{'} h_t + b_h^{'}), \quad (5)$$

where $b_h^{'} \in \mathbb{R}^n$ is the bias vector, and $W_h^{'} = W_h^{\top}$ is referred to as tied weights. The training objective of SAE is to minimize the reconstruction error of the input data. Given an unlabeled signal scalogram sequence $\{x_t, t = 1, 2, \ldots, T\}$, the cost function with respect to parameters $(W_h, b_h, b_h^{'})$ is defined as:

$$J_{SAE}(W_h, b_h, b_h^{'}) = \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_{\mathbb{H}}(x_t, \hat{x}_t) + \alpha \sum_{j=1}^{s} KL(\rho \parallel \hat{\rho}_j), \quad (6)$$

where $\mathcal{L}_{\mathbb{H}}$ is measured by the cross-entropy loss, as follows:

$$\mathcal{L}_{\mathbb{H}}(x_t, \hat{x}_t) = -\sum_{j=1}^{n} \left[x_{tj} \log \hat{x}_{tj} + (1 - x_{tj}) \log(1 - \hat{x}_{tj})\right]. \quad (7)$$

The second term of Eq. (6) is the sparsity penalty term in order to avoid overfitting. Here $s$ is the number of hidden units, and $\alpha$ is the hyper-parameter that controls the weight of the sparsity constraint. $KL(\rho \parallel \hat{\rho}_j)$ is the Kullback–Leibler (KL) divergence to measure the difference between a Bernoulli random variable with mean $\hat{\rho}_j$ and a Bernoulli random variable with mean $\rho$, defined as:

$$KL(\rho \parallel \hat{\rho}_j) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j}, \quad (8)$$

where $\rho$ is the sparsity parameter (a small value close to zero), and $\hat{\rho}$ denotes the average activation of hidden unit $j$ (averaged over the training set). In this way, high dimensional raw inputs are integrated into low dimensional inherent characteristics for semantic learning.

### 3.4. Signal embedding representations

To extract semantic features, inspired by Skip-gram [7], we adopt a similar way to capture semantics from scalogram sequence in the inherent feature space. The main idea is to infer the current inherent representation $h_t$ based on its context expressed as fixed-length embeddings. In this way, signal inherent features with similar semantics would be mapped to close positions in a vector space by learning the context information. Specifically, we concatenate the demographic information $d_t \in \mathbb{R}^d$ to $h_t$ and create the final embedding representation $v_t \in \mathbb{R}^p$, as follows:

$$v_t = f(W_v[h_t, d_t] + b_v), \quad (9)$$

where $W_v \in \mathbb{R}^{p\times(s+d)}$ denotes the weight matrix, and $b_v \in \mathbb{R}^p$ is the bias vector. We employ two mechanisms to learn signal embeddings.

#### 3.4.1. Score-based embeddings

One way to capture the temporal information is to measure the relationship between embedding representation $v_t$ and its neighbor context representations, as follows:

$$\mathcal{L}_{Emb} = -\sum_{\substack{-w\le i\le w,\\ i\ne 0}} \left[\log f(v_t^T c_i) - \frac{1}{k}\sum_{j=1}^{k} \log f(v_t^T c_j)\right], \quad (10)$$

where $c_t$ is the context representation, defined as:

$$c_t = f(W_c[h_t, d_t] + b_c), \quad (11)$$

where $W_c \in \mathbb{R}^{p\times(s+d)}$ and $b_c \in \mathbb{R}^p$ are the weight matrix and bias vector for the context representation. Since the first term of Eq. (10) is only concerned with learning high-quality vector representations, we conduct negative sampling to retain the quality. Here $k$ is the number of random negative samples.

#### 3.4.2. Softmax-based embeddings

Another way to calculate the dynamic information is to train a softmax classifier that predicts the inherent features of signals within a context window. Formally, given a embedding representation $v_t$, we minimize the cross entropy error, as follows:

$$\mathcal{L}_{Emb} = -\sum_{\substack{-w\le i\le w,\\ i\ne 0}} \left[h_{t+i}^T \log \hat{y}_t + (1 - h_{t+i})^T \log(1 - \hat{y}_t)\right], \quad (12)$$

where

$$\hat{y}_t = \frac{\exp(W_s v_t + b_s)}{\sum_{j=1}^{n} \exp(W_s[j,:]v_t + b_s[j])}.$$

Here $W_s \in \mathbb{R}^{n\times p}$ and $b_s \in \mathbb{R}^n$ are the parameters to be learned, $\exp(\cdot)$ is the element-wise exponential function, 1 denotes an all-ones vector, and $w$ is the predefined context window size.

### 3.5. Unified training procedure

The unified model can be obtained by adding the two cost functions $J_{SAE}$ and $J_{Emd}$. In particular, the final cost function of the Wave2Vec$_{sc}$ model in terms of parameters $W_{h,\,v,\,c}$, $b_{h,\,v,\,c}$ is defined as:

$$
\begin{aligned}
&J_{w2vsc}(W_{h,v,c}, b_{h,v,c}) \\
&= \frac{1}{T}\sum_{t=1}^{T}\left\{-\beta_1\sum_{j=1}^{n}\left[x_{tj}\log\hat{x}_{tj} + (1-x_{tj})\log(1-\hat{x}_{tj})\right]\right. \\
&\quad \left. +\beta_2\sum_{\substack{-w\le i\le w,\\ i\ne 0}}\left[-\log f(v_t^T c_i) + \frac{1}{k}\sum_{j=1}^{k}\log f(v_t^T c_j)\right]\right\} \qquad (13) \\
&\quad +\alpha\sum_{j=1}^{s}KL(\rho \parallel \hat{\rho}_j),
\end{aligned}
$$

and the final cost function of the Wave2Vec$_{so}$ model in terms of parameters $W_{h,\ v,\ s}$, $b_{h,\ v,\ s}$ is defined as:

$$
\begin{aligned}
&J_{w2vso}(W_{h,v,s}, b_{h,v,s}) \\
&= \frac{1}{T}\sum_{t=1}^{T}\left\{-\beta_1\sum_{j=1}^{n}\left[x_{tj}\log\hat{x}_{tj} + (1-x_{tj})\log(1-\hat{x}_{tj})\right]\right. \\
&\quad \left. +\beta_2\sum_{\substack{-w\le i\le w,\\ i\ne 0}}\left[-h_{t+i}^T\log\hat{y}_t - (1-h_{t+i})^T\log(1-\hat{y}_t)\right]\right\} \qquad (14) \\
&\quad +\alpha\sum_{j=1}^{s}KL(\rho \parallel \hat{\rho}_j),
\end{aligned}
$$

where $\beta_1$ and $\beta_2$ are the hyper-parameters that control the contribution of each cost function. For our model, the activation function is denoted as the sigmoid logistic function $f(z) = 1/(1+\exp(-z))$. By combining the two cost functions, we learn joint features from both inherent representations and embedding representations of biosignals at the same time. According to the training strategy of deep learning, during the pre-training step, we train the SAE layer individually to obtain a good weight initialization. After that, fine-tuning step is adopted to train all the layers as a unified model.

### 3.6. Interpretation

In healthcare domain, interpreting the learned representations is important to understand the clinical meaning. Since the input of our model is waveform data, we focus on discovering motifs and analyzing which ones are crucial to clinical tasks.

#### 3.6.1. Interpreting motifs from inherent representations

Since the proposed model is based on neural networks, it is easy to find component-based motifs by extending the dictionary learning method proposed in [14]. We use the learned $f(W_h^T) \in \mathbb{R}^{n\times s}$, a non-negative matrix, to represent a motif dictionary. Then, we define each dimension of the $n$-dimensional inherent feature space as a motif, as follows:

$$motif_i = f(W_h^T)[:, i] \in \mathbb{R}^n, \qquad (15)$$

where $W_h^T[:, i]$ represents the $i$th column or dimension of $W_h^T$. By analyzing the selected motifs, we can obtain the clinical interpretation in biosignals.

#### 3.6.2. Interpreting embedding representations

To interpret learned embedding vectors, we can use the same idea in [6]. For the $i$th dimension of the $p$-dimension embedding feature space, we rank top-$k$ dimensions from learned motif dictionary that have the strongest values, as follows:

$$\mathrm{argsort}(f(W_v^T)[:, i])[1:k], \qquad (16)$$

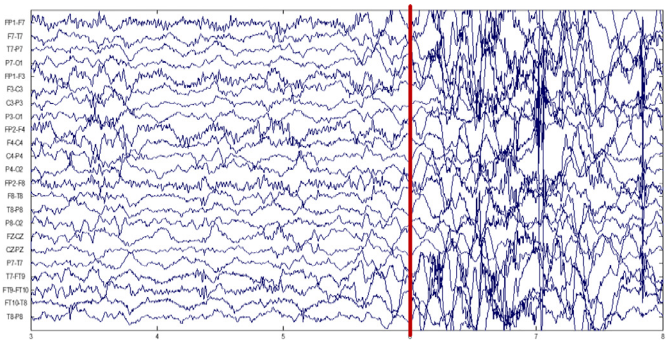


**Fig. 4.** Sample of multi-channel scalp EEG signals on the CHB-MIT dataset. The red bar marks the onset of a seizure. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where argsort$(\cdot)$ returns a vector that index its values in a descending order. Thus, we can combine the knowledge learned from inherent representations to understand how signal fragments are associated with a group of motifs over time.

## 4. Experiments

In this section, we experimentally evaluate the performance of our proposed Wave2Vec model on two real-world biosignal datasets, compare its performance with other state-of-the-art feature learning methods, and show that it achieves significantly better results on different evaluation criteria.

### 4.1. Dataset description

*EEG dataset.* The first dataset we use is the CHB-MIT dataset collected from the Childrens Hospital Boston [27]. This dataset is open access available and can be downloaded at the PhysioNet [28]. In the CHB-MIT dataset, the EEG signals are recorded from 23 patients with intractable seizures. The EEG signals of each patient contains 23 channels, and the data of each channel is recorded at 256 Hz with 16-bit resolution. Moreover, the beginning and end of each seizure are both annotated in the ground truth. Fig. 4 illustrates an example of EEG seizure onset within a patient on the CHB-MIT dataset. Note that the seizure onset reflects distinctive rhythmic patterns in different channels, and this variability makes the seizure detection problem even more difficult.

*ECG dataset.* The second dataset is the MIT-BIH arrhythmia database [29], which is used for evaluation of proposed model on the task of ECG arrhythmia detection. This dataset contains ECG recordings of 48 individuals, each containing a half-hour excerpt in two channels. Specifically, the recordings are digitized at 360 Hz with 11-bit resolution.

### 4.2. Baselines

Since the goal of our model is to learn features in an unsupervised fashion, we employ several widely used feature learning algorithms as baselines. For the sake of fairness, we feed the learned features of each models into a linear support vector machine (SVM) classifier [30] for evaluation.

*Principal Component Analysis (PCA).* Following the handcrafted feature engineering, we use PCA [31] as baseline. We select top-$p$ components as features from signals, in order to reduce the number of dimensions. This procedure significantly enhances the performance by excluding irrelevant and redundant information from waveform data.

*Stacked Sparse Autoencoders (SSAEs).* SSAEs is one of the most popular unsupervised representation learning algorithms [32], and

**Table 1**
Confusion matrix definition.

| | | Actual class | |
|---|---|---|---|
| | | Positive | Negative |
| Predicted class | Positive | TP | FP |
| | Negative | FN | TN |

is a multi-layer neural network consisting of several basic SAE layers. We concatenate the signal fragment $x_t$ and its demographic information $d_t$ as the input, and train a 2-layer SSAEs to minimize the reconstruction error.

*Skip-gram* [7]*.* Since the Skip-gram model is a word embedding method, it cannot be applied directly to waveform data. Thus, we first employ a word representation method for each signal fragment. The signal fragment $x_t$ is sampled as a single word representation according to a pooling strategy. This approach was proven very effective for EEG seizure detection in [14]. For the sake of fairness, we extend the Skip-gram model by employing both our embedding strategies, namely Skip-gram$_{so}$ and Skip-gram$_{sc}$, respectively.

*Med2Vec* [6]*.* Med2Vec is a simple and robust algorithm to efficiently learn features, which also follows the idea of Skip-gram. Comparing to Skip-gram, the difference of Med2Vec is that the input of Med2Vec is vector space instead of word space, which means that we do not need to use word representation method. We perform the same embedding process as Skip-gram, namely Med2Vec$_{so}$ and Med2Vec$_{sc}$, respectively.

### 4.3. Evaluation criteria

Since the evaluation tasks belong to classification problem, the evaluation criteria are calculated based on the confusion matrix of classification results. We present the definition of confusion matrix in Table 1. To validate our proposed method, we use *F*1*-score* and *Accuracy* for evaluation. *F*1*-score* is used to evaluate the results by considering both the precision and recall, defined as:

$$F1\text{-}score = \frac{2 \times Prec \times Rec}{Prec + Rec}, \tag{17}$$

where $Prec = TP/(TP + FP)$ is denoted as the precision, and $Rec = TP/(TP + FN)$ is denoted as the recall. Based on the confusion matrix, *Accuracy* is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{18}$$

Moreover, receiver operator characteristic (ROC) curves are plotted to generally illustrate the diagnostic ability of seizure detector. Since in most medical problems we usually care about the fraction of examples classified as abnormal cases that are truly abnormal, precision-recall (PR) curves are also plotted. In the experiments, we use area-under-the-curve (AUC) to numerically evaluate the quality of each method. Note that AUC scores are in the range of [0,1]. The higher AUC score, the better the performance.

### 4.4. Experiment setup

We implement all the approaches with Theano [33]. During the whole training step, we use Adadelta [34] with mini-batch to minimize the cost function. We perform 100 iterations and report the best performance for each method. Some training strategies including normalization, dropout (the dropout rate is 0.5), and regularization ($L_2$ penalty with the coefficient 0.001) are also adopted for all the approaches. To fit the input of the clinical tasks, in the experiments, we set the same $h = 80$ and $p = 60$ for baselines and our models.

For EEG seizure detection task, we denote the class label as interictal and ictal states for each EEG fragment according to the ground truth. We use gender and age as the demographic information in the input data. For ECG arrhythmia detection task, we use QRS detection, a common method in ECG preprocessing [35,36], to segment ECG signals into beats. Each segment is labeled as normal or arrhythmia according to the dataset annotations. To evaluate our method as a general algorithm, we randomly hold-off 20% data to train the SVM classifier after the whole-data representation learning. Specifically, we combine 4,302 23-channel EEG fragments from nine different patients and 21,681 2-channel ECG fragments from ten different patient, receptively. We divide the held-off data to training and testing folds with ration 4:1. Considering the scarcity of abnormal events, we trim our test data to balance the number of normal and abnormal fragments.

### 4.5. Performance on clinical tasks

We compare the clinical performance of our proposed models Wave2Vec$_{sc}$ and Wave2Vec$_{so}$ with the aforementioned baseline methods. The experimental results on the CHB-MIT and the MIT-BIH datasets are listed in Table 2. We can easily observe that both our Wave2Vec$_{sc}$ and Wave2Vec$_{so}$ models outperform the baselines on all four different evaluation measurements. Among them, our proposed Wave2Vec$_{so}$ model not only achieves better F1-score and Accuracy, but also obtains higher AUC-PR and AUC-ROC than baselines on both two datasets.

Feature representation is crucial for both EEG seizure detection and ECG arrhythmia detection. From Table 2, all the methods except Med2Vec-based models on the MIT-BIH dataset perform better results than those on the CHB-MIT dataset. The reason is that the rhythmic patterns in 2-channel ECG signals are more observable than those in 23-channel EEG signals. There are significant differences in amplitude and frequency of features in ECG that can be easily learned. In this situation, both the Accuracy and F1-score of our proposed Wave2Vec method are achieved the best of 99%.

Given the results of baselines, the Med2Vec-based methods perform worse than the others. This is because the waveform data is unsuitable for embedding directly, and hence the Med2Vec-based methods are failed to learn proper features from continuous physiological records. The results of the Skip-gram-based methods demonstrate that the models benefit from the signal word representations. The reason is that by incorporating the pooling strategies, noise interference is reduced in effectively, and hence Skip-gram can capture more powerful information from handcrafted features. We can also observe that SSAEs generates better result than the other embedding-based baselines. This results from the inherent features extracted by deep learning model. Furthermore, given the results of Wave2Vec-based methods which consider both inherent and temporal features, we can conclude that our models are able to obtain meaningful representations from biosignals.

From the results, the comparison between the proposed Wave2Vec$_{so}$ and Wave2Vec$_{sc}$ demonstrates that the softmax-based embedding mechanism is more suitable for the two detection tasks. We arrive at a conclusion that both inherent and temporal features are playing an important role to identify critical patterns, and the combination feature learning provides complementary information towards each other, which is crucial for clinical performance.

Fig. 5 illustrates the PR and ROC curves on the CHB-MIT dataset, respectively. We can see that Wave2Vec$_{so}$ yields the best AUC-PR and AUC-ROC in terms of the ability of cross-patient seizure detection. Given the Fig. 5a, our Wave2Vec$_{so}$ model achieves the best AUC of 0.9509 compared with the models such as SSAEs, Skip-gram$_{so}$, Med2Vec$_{so}$ and PCA with AUC of 0.8721, 0.8162, 0.4900,

**Table 2**
Performance comparisons on two clinical biosignal datasets.

| Method | EEG seizure detection | | | | ECG arrhythmia detection | | | |
|---|---|---|---|---|---|---|---|---|
| | AUC-ROC | AUC-PR | F1-score | Accuracy | AUC-ROC | AUC-PR | F1-score | Accuracy |
| PCA | 0.8552 | 0.7411 | 0.8638 | 0.7976 | 0.9860 | 0.9864 | 0.9602 | 0.9622 |
| SSAEs | 0.9171 | 0.8721 | 0.9144 | 0.8704 | 0.9912 | 0.9865 | 0.9765 | 0.9779 |
| Med2Vec$_{sc}$ | 0.7629 | 0.4872 | 0.7690 | 0.6879 | 0.6127 | 0.7005 | 0.5191 | 0.6599 |
| Med2Vec$_{so}$ | 0.7643 | 0.4900 | 0.7712 | 0.6899 | 0.5564 | 0.5533 | 0.4652 | 0.6278 |
| Skip-gram$_{sc}$ | 0.8546 | 0.7480 | 0.9057 | 0.8524 | 0.8569 | 0.6794 | 0.7643 | 0.8174 |
| Skip-gram$_{so}$ | 0.9039 | 0.8162 | 0.9145 | 0.8684 | 0.7810 | 0.8617 | 0.7108 | 0.8059 |
| Wave2Vec$_{sc}$ | 0.9677 | 0.9170 | 0.9506 | 0.9242 | 0.9965 | 0.9950 | 0.9926 | 0.9939 |
| Wave2Vec$_{so}$ | **0.9833** | **0.9509** | **0.9605** | **0.9392** | **0.9968** | **0.9955** | **0.9931** | **0.9943** |

and 0.7411, respectively. Moreover, from the Fig. 5b, we can see that the true positive rate of the Wave2Vec$_{so}$ model increases fast from the start. This means that Wave2Vec$_{so}$ can obtain critical information to separate data effectively and hence results in the best of F1-score and Accuracy.

## 4.6. Feature representation analysis

To analyze the feature representation of our proposed Wave2Vec model, we employ t-Distributed Stochastic Neighbor Embedding (t-SNE) [37] to visualize the learned features in a two-dimensional space. Figs. 6 and 7 show the subpopulation distribution of the CHB-MIT dataset and the MIT-BIH dataset using the features learned by the SSAEs baseline method and our proposed two Wave2Vec models, respectively. Data points are randomly subsampled for better viewing.

From the given results on the EEG dataset, Fig. 6a shows that the positive (ictal) and negative (non-ictal) samples extracted by the SSAEs baseline model are mixed up together. It fails to suggest a good separation between ictal and non-ictal states. In contrast, according to Fig. 6b and Fig. 6c, it is obvious that both our proposed Wave2Vec$_{sc}$ and Wave2Vec$_{so}$ models succeed at separating the non-ictal and ictal states for the task of seizure detection. These results demonstrate that our models are able to learn representative features of the non-ictal and ictal states. In addition, we can also see that our models can uncover underlying patterns in the non-ictal labels. We assume these patterns may reveal the transition periods between ictal and non-ictal states.

Given the given results of the ECG dataset, Fig. 7a shows the visualization of the features learned by the SSAEs baseline model. Such many scattered small clusters suggest that the features are not expressive enough to construct a large representative cluster in the space. In contrast, as can be observed in Fig. 7b and7c, our proposed models are able to learn more separable features. We can further analyze each cluster and discover underlying differences for these groups to explain the arrhythmia patterns.

## 4.7. Sensitivity analysis

To further evaluate the performance of our proposed Wave2Vec model, we conduct sensitivity analysis to study the impact of parameter configuration. Specifically, we study three main aspects: the number of inherent units $s$, the dimensionality of embeddings $p$, and the size of context window $w$. We plot the Accuracy and AUC-ROC results using different settings of parameters. Note that the basic configuration for Wave2Vec is $s = 80, p = 60, w = 1$. In each step, we vary one hyper-parameter while keeping others fixed to the basic configuration.

*Inherent unit number s.* Fig. 8 shows the change of Accuracy and AUC-ROC for different numbers of inherent representations $s$ on the CHB-MIT and MIT-BIH datasets, respectively. From the figure, we can see that when the number of hidden units is small, both of our models lack the capability of capturing representative features, resulting in limited performance on both Accuracy and AUC-ROC. As we increase the number of hidden units, our models show an increasing modeling power. However, when the number is too large, we have insufficient samples to train the network, which results in a worse performance and stability. In our experiment, we choose 80 as the inherent unit number.

*Embedding dimensionality p.* We report the experimental results of $p$ in Fig. 9. The proposed model gets the best performance when $p$ is 60. From the figure, we can see that the dimension is reduced effectively and 60 hidden units are enough to capture the dynamic information of the inherent features. While too few hidden units might result in the proposed models being unable to extract enough features, too many hidden units might also put the proposed model at the risk of the curse of dimensionality.

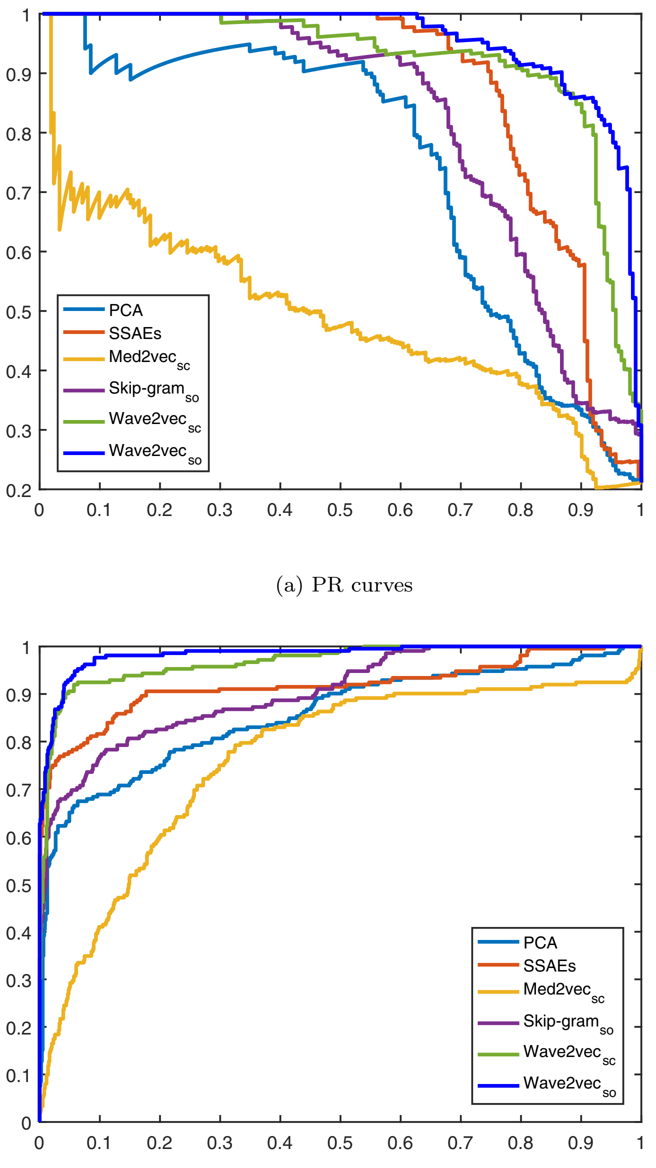
*Context window size w.* Fig. 10 illustrates the influence of $w$. From the figure, we can see that the performance decreases when setting $w$ to a too large value. Increasing the context window size degrades the performance of both of the models, as it leads to overfitting. For the Wave2Vec$_{sc}$ model, increasing the context window size $p$ seems to have a stronger influence on its performance than the Wave2Vec$_{so}$ model. This is due to the different embedding learning scenarios.

In summary, the performance of Wave2Vec$_{so}$ is less sensitive to the different choices of parameters than Wave2Vec$_{sc}$ on the EEG seizure detection, while Wave2Vec$_{so}$ is more sensitive to parameters on the ECG Arrhythmia detection. This results from the different nature of two clinical tasks. These results also indicate that with the help of negative sampling, the Wave2Vec$_{sc}$ model may perform more robustly than the Wave2Vec$_{so}$ model when the training samples are relatively large. Despite of the influence, it is obvious that our proposed Wave2Vec models consistently beat the baselines with different parameter settings.

## 4.8. Interpretation analysis

### 4.8.1. Motifs discovery

The interpretability of biosignal motifs is important in healthcare. In order to analyze the motifs discovered by our proposed models, we visualize some EEG motifs selected from the learned inherent representation matrix $W_h^T$ of our Wave2Vec$_{so}$ model. Since there are 23 EEG channels on the CHB-MIT dataset, we learn 1840 EEG motifs in time-frequency domain. From Fig. 11, we can clearly observe that the shape and range of discovered EEG motifs from all twelve dimensions are different, which represent different frequency energy distributions. From the perspective of medical concern, these discovered motifs are the common patterns that reflect the brain abnormality in EEG [38]. In this way, we can demonstrate the characteristic of each column and map each dimension from the motif-level space to the bags of frequencies. It is worth noting that according to the results of Section 4.6, the motifs learned from both of the Wave2Vec$_{so}$ and Wave2Vec$_{sc}$ models



(a) PR curves

(b) ROC curves

**Fig. 5.** PR and ROC curves of the proposed method and the baselines on the CHB-MIT dataset.

are similar, since the feature distributions and learning scenarios are similar.

### 4.8.2. Embedding representation

To demonstrate the benefit of applying embedding mechanisms in biosignal processing, we analyze the embedding weights $W_v$ learned from one of the proposed approach, Wave2Vec$_{so}$, which uses the softmax-based embedding mechanism. Table 3 shows a case study for EEG seizure detection of a patient on the CHB-MIT dataset. Since the discovered EEG motifs are the components of raw waveform data, we select the top-1 dimension of embedding representation $W_v$ by ranking the activation values of Wave2Vec's vector $v_t$. In this way, we can then interpret motif groups by analyzing the meaning of the selected embedding dimension. For instance, for embedding dimension 467, the top-3 strongest motifs connected to it are $motif_{638}$, $motif_{618}$, and $motif_{562}$, which are strongly activated in channel-7.

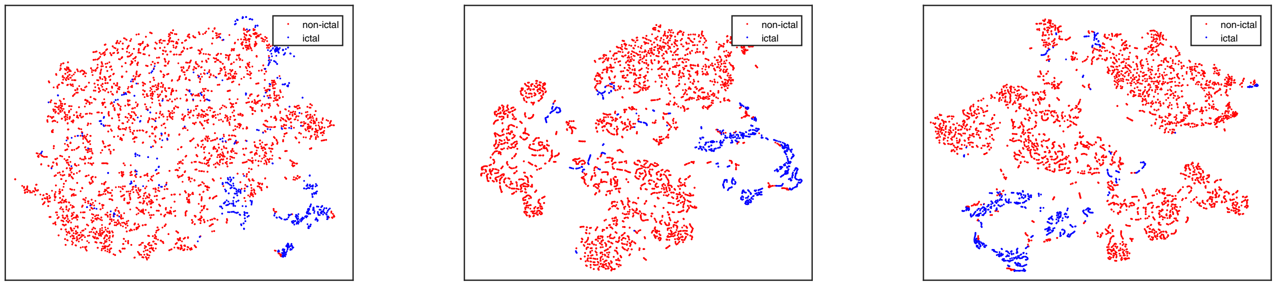
From Table 3, we can observe that the patient suffered from general seizures from segment *No*.9 to segment *No*.53. When the patient was in non-ictal state, the embedding dimension 1288 connected with $motif_{1743}$ was most activated. In Fig. 11k, we can observe that the frequency distribution of $motif_{1743}$ is spread evenly with relatively low energy. During the seizures, we can see that the seizure was firstly started from the embedding dimension 1197, which are strongly connected to $motif_{1579}$ in channel-19. Then the embedding dimension 1197 and 467 appeared in turns, which are strongly connected to $motif_{1579}$ in channel-19 and $motif_{638}$ in channel-7, respectively. Finally, the embedding dimension 665 with $motif_{956}$ in channel-11 appeared at last. Some of these mentioned motifs are shown in Fig. 11. We can observe that these motifs represent different frequency energy distributions which demonstrate different seizure patterns. In summary, the analysis results indicate that the interpretable representations learned by Wave2Vec not only improve the detection performance, but also identify the influential clinical concepts of biosignals both in motif-level and embedding-level.

## 5. Related work

Research on representation learning for clinical temporal data is literally the most vital part of science for humans, as it refers to any biosignals that can be continually measured and monitored from living organisms. Since our work is associated with time-frequency analysis, deep feature learning, and embeddings, we summarize the literatures into the following three categories.

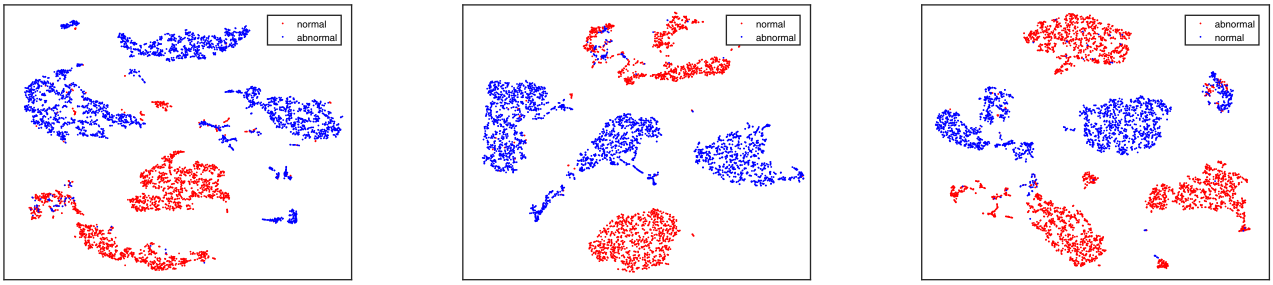
### 5.1. Time-frequency analysis of biosignals

From a variety of existing literatures, time-frequency representations are more informative than raw input, and are powerful to



(a) The SSAEs model (b) The Wave2Vec$_{sc}$ model (c) The Wave2Vec$_{so}$ model

**Fig. 6.** Feature visualization of learned representations using different models on the CHB-MIT dataset. Colors show different clusters according to the ground truth (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

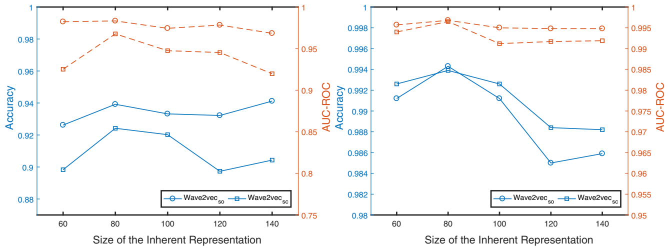


(a) The SSAEs model

(b) The Wave2Vec$_{sc}$ model

(c) The Wave2Vec$_{so}$ model

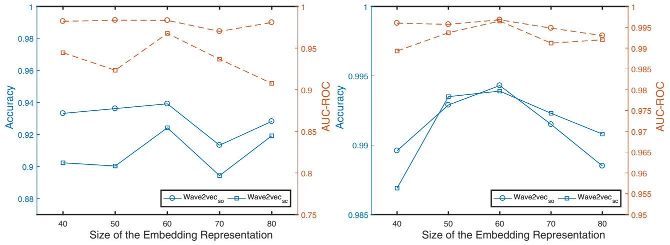
**Fig. 7.** Feature visualization of learned representations using different models on the MIT-BIH dataset. Colors show different clusters according to the ground truth (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).



(a) EEG

(b) ECG

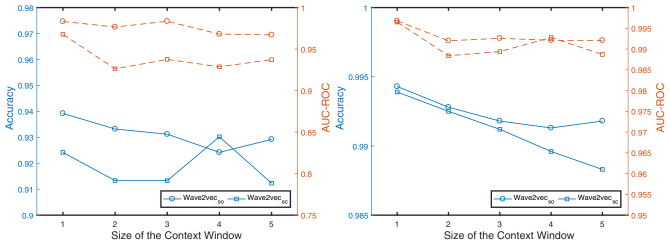
**Fig. 8.** Performance variations with different numbers of inherent units on the CHB-MIT and the MIT-BIH datasets.



(a) EEG

(b) ECG

**Fig. 9.** Performance variations with different dimensions of embeddings on the CHB-MIT and the MIT-BIH datasets.



(a) EEG

(b) ECG

**Fig. 10.** Performance variations with different sizes of context window on the CHB-MIT and the MIT-BIH datasets.

characterize non-stationary biomedical signals. There has been a growing interest in adopting time-frequency descriptors for biosignal processing. To determine an optimum classification scheme, Übeyli [39] presented a ECG beats classification approach using discrete wavelet transform and wavelet coefficients. Bachler et al. [40] and Zhao and Zhang [41] extracted a set of wavelet coefficients from the ECG signals using wavelet transform. Chen et al. [42] studied the most difficult epileptic EEG classification task using different decomposition level of discrete wavelet transform. The authors concluded that decomposition level effects the localization accuracy more significantly than mother wavelets. Şengür et al. [43] presented a texture descriptor for EEG time-frequency images to detect epileptic seizure patterns. Samiee et al. [44] extracted localized image features relying on rational functions in time-frequency domain. Instead of adopting handcrafted engineering for feature extraction, in our approach, we combine time-frequency transform with deep learning to enhance feature representation, and hence make improvement of performance for different clinical tasks.

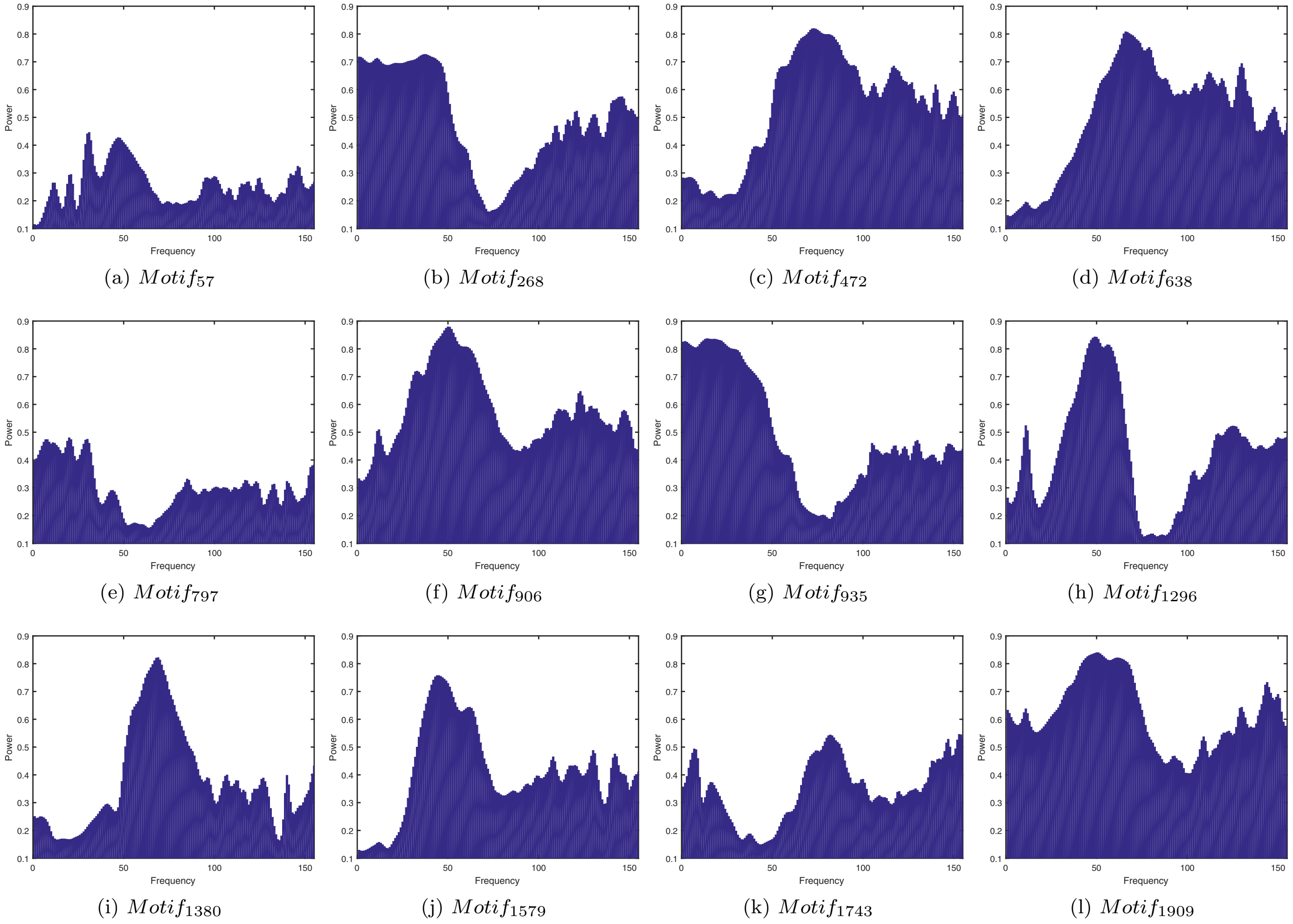
### 5.2. Deep learning for biosignals

In recent years, the existing literatures of biosignal processing through deep learning methods are diverse and roughly follow three lines of research. First, some researchers employed deep belief networks (DBN) to automated learn and identify features from biosignals. Yan et al. [45] used DBN to learn features from ECG signals for two-lead heart beat classification problem. Jia et al. [46] and Turner et al. [47] applied DBN to model EEG signals for affective state recognition and seizure detection, respectively. Second, there are some methods related to stacked autoencoders. Yang et al. [48] adopted sparse autoencoder with softmax regression classifier to differentiate premature ventricular contraction (PVC) beats and non-PVC beats using ECG signals. Lin et al. [49] adopted stacked autoencoders to detect seizures utilizing unsupervised EEG features. Finally, convolution neural networks (CNN) were also adopted for biosignal analysis. Antoniades et al. [50] adopted a CNN model to generate features from intra-cranial EEG signals in time domain. Kiranyaz et al. [51] proposed a patient-specific ECG classification using adaptive one-dimensional CNN. To sum up, all the aforementioned deep learning strategies take raw data as input and ignore the temporal relationship between each fragment.

Moreover, Yuan et al. [52] proposed a multi-view deep learning model to capture brain abnormality from EEG signals. The authors generated time-frequency EEG fragments as preprocessing. This method combines handcrafted engineering and deep learning, but ignored the temporal features and was not an end-to-end model. With the help of time-frequency information, we proposed an end-to-end model to jointly learn inherent and temporal features of biosignals.

### 5.3. Semantic learning for biosignals

Recently, discovering efficient representations of complex clinical continuous time series data has become an important indicator in healthcare data mining. There has been a growing interest in introducing semantic learning models to biosignal processing. First, word embedding is an efficient approach to



(a) $Motif_{57}$ (b) $Motif_{268}$ (c) $Motif_{472}$ (d) $Motif_{638}$

(e) $Motif_{797}$ (f) $Motif_{906}$ (g) $Motif_{935}$ (h) $Motif_{1296}$

(i) $Motif_{1380}$ (j) $Motif_{1579}$ (k) $Motif_{1743}$ (l) $Motif_{1909}$

**Fig. 11.** Twelve samples of discovered EEG seizure motifs, where frequency decreases across the *x*-axis and energy decreases down the *y*-axis. The discovered EEG motifs from all twelve dimensions represent different frequency energy distributions, which are the common medical patterns that the brain abnormality in EEG is reflected by frequency changes and increased amplitudes.

**Table 3**
EEG seizure detection in each fragment for a patient in the case study.

| Segment *no.* | State | Embedding dimension (top-1) | Embedding motifs (top-3) | Channel-graphic (top-1) |
|---|---|---|---|---|
| 1–8 | non-ictal | 1288 | *motif*$_{1743}$, *motif*$_{1702}$, *motif*$_{1686}$ | 21 |
| 9–14 | ictal | 1197 | *motif*$_{1579}$, *motif*$_{1564}$, *motif*$_{1540}$ | 19 |
| 15–16 | ictal | 1335 | *motif*$_{1825}$, *motif*$_{1836}$, *motif*$_{1817}$ | 22 |
| 17–22 | ictal | 467 | *motif*$_{638}$, *motif*$_{618}$, *motif*$_{562}$ | 7 |
| 23 | ictal | 1252 | *motif*$_{1672}$, *motif*$_{1627}$, *motif*$_{1622}$ | 20 |
| 24 | ictal | 467 | *motif*$_{638}$, *motif*$_{618}$, *motif*$_{562}$ | 7 |
| 25 | ictal | 1197 | *motif*$_{1579}$, *motif*$_{1564}$, *motif*$_{1540}$ | 19 |
| 26 | ictal | 467 | *motif*$_{638}$, *motif*$_{618}$, *motif*$_{562}$ | 7 |
| 27 | ictal | 1197 | *motif*$_{1579}$, *motif*$_{1564}$, *motif*$_{1540}$ | 19 |
| 28–31 | ictal | 467 | *motif*$_{638}$, *motif*$_{618}$, *motif*$_{562}$ | 7 |
| 32–33 | ictal | 1055 | *motif*$_{1380}$, *motif*$_{1383}$, *motif*$_{1438}$ | 17 |
| 34–35 | ictal | 1335 | *motif*$_{1385}$, *motif*$_{1396}$, *motif*$_{1377}$ | 22 |
| 36 | ictal | 467 | *motif*$_{638}$, *motif*$_{618}$, *motif*$_{562}$ | 7 |
| 37 | ictal | 1335 | *motif*$_{1825}$, *motif*$_{1836}$, *motif*$_{1817}$ | 22 |
| 38 | ictal | 1197 | *motif*$_{1579}$, *motif*$_{1564}$, *motif*$_{1540}$ | 19 |
| 33–42 | ictal | 233 | *motif*$_{267}$, *motif*$_{281}$, *motif*$_{313}$ | 3 |
| 43–44 | ictal | 665 | *motif*$_{956}$, *motif*$_{960}$, *motif*$_{933}$ | 11 |
| 45–47 | ictal | 1197 | *motif*$_{1579}$, *motif*$_{1564}$, *motif*$_{1540}$ | 19 |
| 48 | ictal | 1288 | *motif*$_{1743}$, *motif*$_{1702}$, *motif*$_{1686}$ | 21 |
| 49–50 | ictal | 665 | *motif*$_{956}$, *motif*$_{960}$, *motif*$_{933}$ | 11 |
| 51 | ictal | 1197 | *motif*$_{1579}$, *motif*$_{1564}$, *motif*$_{1540}$ | 19 |
| 52 | ictal | 1288 | *motif*$_{1743}$, *motif*$_{1702}$, *motif*$_{1686}$ | 21 |
| 53 | ictal | 233 | *motif*$_{267}$, *motif*$_{281}$, *motif*$_{313}$ | 3 |
| 54–60 | non-ictal | 1288 | *motif*$_{1743}$, *motif*$_{1702}$, *motif*$_{1686}$ | 21 |

derive temporal representative features from biosignals. Xun et al. [14] and Yuan et al. [13] proposed deep learning-based methods to learn a dictionary of EEG signals, and applied the CBOW model to extract the temporal features. Li et al. [15] proposed an orthogonal sparse autoencoder model to extract EEG word representations with the same strategy, and applied Skip-gram to learn feature vector for each composite sentence. With any methods adopted, all the above discussed models learned word and embedding representations separately, and hence are not end-to-end models. Second, topic modeling is another approach in NLP that can be used in biosignal data mining. Esbroeck et al. [53] explored an application of topic models for heart rate time series to identify functional sets of heart rate sequences. The authors applied QRS detection and symbolic aggregate approximation (SAX) to convert ECG signals into symbolic sequences. Although this method makes sense in ECG word representation, it is not a general way for biosignal representation learning. Saria et al. [10] proposed a time series topic model. The authors used a graph model (LDA-based) to discover latent topics (disease), words (motifs), and their distribution at the same time. With any strategy adopted, our proposed Wave2Vec model distinguishes itself from all the above discussed methods in both the method and the aimed task.

## 6. Conclusion

In this paper, we propose a novel model, namely Wave2Vec, to address the challenges of modeling biosignals using semantic learning techniques and interpreting the learned representations. Wave2Vec is a unified model that jointly learns both inherent and embedding representations of biosignals at the same time. Two embedding mechanisms are developed to interpret the learned temporal features with discovered latent component-based motifs reasonably over time. Experimental results on two real-world biosignal datasets justify the effectiveness of our proposed Wave2Vec model in clinical tasks. Analytical results show that the the proposed Wave2Vec model can incorporate both inherent motif co-occurrence information and embedding information of biosignals, which improves the clinical interpretability. Throughout a case study, we demonstrate that the learned component-based motifs and embedding representations are meaningful.

## Acknowledgments

This paper is supported by the National Natural Science Foundation of China (61672064, 81370038), the Beijing Natural Science Foundation (4172001), the China Postdoctoral Science Foundation (2016T90022, 2015T80030, 2015M580029), the Science and Technology Project of Beijing Municipal Education Commission (KZ201610005007), Beijing Laboratory of Advanced Information Networks (040000546617002), and the China Scholarship Council Fund (201606540008).

## References

[1] I. Sutskever, G. Hinton, Learning multilevel distributed representations for high-dimensionalsequences, Artif. Intell. Stat. (2007) 548–555.
[2] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, A neural probabilistic language model, J. Mach. Learn. Res. 3 (2003) 1137–1155.
[3] L. Wei, Y. Ding, R. Su, J. Tang, Q. Zou, Prediction of human protein subcellular localization using deep learning, J. Parallel Distrib. Comput. (2017).
[4] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, J. Gao, Dipole: diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks, in: Proceedings of the Twenty-Third ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2017, pp. 1903–1911.
[5] Q. Suo, F. Ma, G. Canino, J. Gao, A. Zhang, P. Veltri, A. Gnasso, A multi-task framework for monitoring health conditions via attention-based recurrent neural networks, in: Proceedings of the AMIA 2017 Annual Symposium (AMIA'17), 2017.
[6] E. Choi, M.T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, J. Sun, Multi-layer representation learning for medical concepts, in: Proceedings of the Twenty-Second ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 1495–1504.
[7] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of the Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.
[8] A.E.W. Johnson, M.M. Ghassemi, S. Nemati, K.E. Niehaus, D.A. Clifton, G.D. Clifford, Machine learning and decision support in critical care, Proc. IEEE 104 (2) (2016) 444–466.
[9] G.-Z. Yang, M. Yacoub, Body Sensor Networks, volume 1, Springer, 2006.
[10] S. Saria, D. Koller, A. Penn, Learning individual and population level traits from clinical temporal data, in: Proceedings of the Neural Information Processing Systems (NIPS), Predictive Models in Personalized Medicine workshop, Citeseer, 2010.
[11] S. Torkamani, V. Lohweg, Survey on time series motif discovery, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 7 (2) (2017) 1–8.
[12] H.A. Dau, E. Keogh, Matrix profile v: A generic technique to incorporate domain knowledge into motif discovery, in: Proceedings of the Twenty-Third ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2017, pp. 125–134.
[13] Y. Yuan, G. Xun, K. Jia, A. Zhang, A novel wavelet-based model for EEG epileptic seizure detection using multi-context learning, in: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2017, pp. 694–699.
[14] G. Xun, X. Jia, A. Zhang, Context-learning based electroencephalogram analysis for epileptic seizure detection, in: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2015, pp. 325–330.
[15] X. Li, X. Jia, G. Xun, A. Zhang, Improving EEG feature learning via synchronized facial video, in: Proceedings of the IEEE International Conference on Big Data (Big Data)„ IEEE, 2015, pp. 843–848.
[16] Y. Yuan, G. Xun, Q. Suo, K. Jia, A. Zhang, Wave2vec: learning deep representations for biosignals, in: Proceedings of the IEEE International Conference on-Data Mining (ICDM), IEEE, 2017, pp. 1159–1164.
[17] F. Mormann, R.G. Andrzejak, C.E. Elger, K. Lehnertz, Seizure prediction: the long and winding road, Brain 130 (2) (2007) 314–333.
[18] Q. Zou, J. Zeng, L. Cao, R. Ji, A novel features ranking metric with application to scalable visual and bioinformatics data classification, Neurocomputing 173 (2016) 346–354.
[19] C. Lin, W. Chen, C. Qiu, Y. Wu, S. Krishnan, Q. Zou, Libd3c: ensemble classifiers with a clustering and dynamic selection strategy, Neurocomputing 123 (2014) 424–435.
[20] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.
[21] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.
[22] P. Mirowski, D. Madhavan, Y. LeCun, R. Kuzniecky, Classification of patterns of EEG synchronization for seizure prediction, Clin. Neurophys. 120 (11) (2009) 1927–1940.
[23] M. Längkvist, L. Karlsson, A. Loutfi, A review of unsupervised feature learning and deep learning for time-series modeling, Pattern Recognit. Lett. 42 (2014) 11–24.
[24] A. Supratak, C. Wu, H. Dong, K. Sun, Y. Guo, Survey on feature extraction and applications of biosignals, in: Machine Learning for Health Informatics, Springer, 2016, pp. 161–182.
[25] P. Smaragdis, M. Shashanka, B. Raj, Topic models for audio mixture analysis, in: Proceedings of the NIPS Workshop on Applications for Topic Models: Text and Beyond, 2009.
[26] S. Mallat, A Wavelet Tour of Signal Processing, Academic press, 1999.
[27] A.H. Shoeb, Application of machine learning to epileptic seizure onset detection and treatment, Massachusetts Institute of Technology, 2009 Ph.D. thesis.
[28] A.L. Goldberger, L.A.N. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, H.E. Stanley, Physiobank, physiotoolkit, and physionet, Circulation 101 (23) (2000) e215–e220.
[29] G.B. Moody, R.G. Mark, The impact of the MIT-BIH arrhythmia database, IEEE Eng. Med. Biol. Mag. (2001).
[30] C. Cortes, V. Vapnik, Support-vector networks, Machine Learn. 20 (3) (1995) 273–297.
[31] I. Jolliffe, Principal Component Analysis, Wiley Online Library, 2002.
[32] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion, J. Mach. Learn. Res. 11 (Dec) (2010) 3371–3408.
[33] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, Y. Bengio, Theano: a CPU and GPU math compiler in python, in: Proceedings of the Ninth Python in Science Conference, 2010, pp. 1–7.
[34] M.D. Zeiler, Adadelta:an adaptive learning rate method, arXiv:1212.5701(2012).
[35] P.S. Hamilton, W.J. Tompkins, Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database, IEEE Trans. Biomed. Eng. (1986).
[36] S.H. Jambukia, V.K. Dabhi, H.B. Prajapati, Classification of ECG signals using machine learning techniques: A survey, in: Proceedings of the International Conference on Advances in Computer Engineering and Applications (ICACEA), IEEE, 2015, pp. 714–721.
[37] M.L. van der, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (Nov) (2008) 2579–2605.
[38] J. Gotman, D. Flanagan, J. Zhang, B. Rosenblatt, Automatic seizure detection

in the newborn: methods and initial evaluation, Electroencephalogr. and Clin. Neurophys. 103 (3) (1997) 356–362.

[39] E.D. Übeyli, Ecg beats classification using multiclass support vector machines with error correcting output codes, Digital Signal Process. 17 (3) (2007) 675–684.

[40] M. Bachler, C. Mayer, B. Hametner, S. Wassertheurer, A. Holzinger, Online and offline determination of QT and PR interval and QRS duration in electrocardiography, in: Proceedings of the Joint International Conference on Pervasive Computing and the Networked World, Springer, 2012, pp. 1–15.

[41] Q. Zhao, L. Zhang, Ecg feature extraction and classification using wavelet transform and support vector machines, in: International Conference on Neural Networks and Brain, 2005. ICNN&B'05., volume 2, IEEE, 2005, pp. 1089–1092.

[42] D. Chen, S. Wan, F.S. Bao, Epileptic focus localization using EEG based on discrete wavelet transform through full-level decomposition, in: Proceedings of the IEEE Twenty-Fifth International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, 2015, pp. 1–6.

[43] A. Şengür, Y. Guo, Y. Akbulut, Time–frequency texture descriptors of EEG signals for efficient detection of epileptic seizure, Brain Inf. 3 (2) (2016) 101–108.

[44] K. Samiee, P. Kovacs, M. Gabbouj, Epileptic seizure classification of EEG time-series using rational discrete short-time Fourier transform, IEEE Trans. Biomed. Eng. 62 (2) (2015) 541–552.

[45] Y. Yan, X. Qin, Y. Wu, N. Zhang, J. Fan, L. Wang, A restricted Boltzmann machine based two-lead electrocardiography classification, in: Proceedings of the IEEE Twelfth International Conference on Wearable and Implantable Body Sensor Networks (BSN), IEEE, 2015, pp. 1–9.

[46] X. Jia, K. Li, X. Li, A. Zhang, A novel semi-supervised deep learning framework for affective state recognition on eeg signals, in: Proceedings of the IEEE International Conference on Bioinformatics and Bioengineering (BIBE), IEEE, 2014, pp. 30–37.

[47] J.T. Turner, A. Page, T. Mohsenin, T. Oates, Deep belief networks used on high resolution multichannel electroencephalography data for seizure detection, in: Proceedings of the AAAI Spring Symposium Series, 2014.

[48] J. Yang, Y. Bai, G. Li, M. Liu, X. Liu, A novel method of diagnosing premature ventricular contraction based on sparse auto-encoder and softmax regression, Bio-medical materials and engineering 26 (s1) (2015) S1549–S1558.

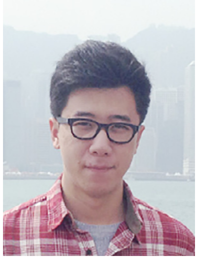
[49] Q. Lin, S.-q. Ye, X.-m. Huang, S.-y. Li, M.-z. Zhang, Y. Xue, W.-S. Chen, Classification of epileptic EEG signals with stacked sparse autoencoder based on deep learning, in: Proceedings of the International Conference on Intelligent Computing, Springer, 2016, pp. 802–810.

[50] A. Antoniades, L. Spyrou, C.C. Took, S. Sanei, Deep learning for epileptic intracranial EEG data, in: Proceedings of the IEEE Twenty-Sixth International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, 2016, pp. 1–6.

[51] S. Kiranyaz, T. Ince, M. Gabbouj, Real-time patient-specific ECG classification by 1-d convolutional neural networks, IEEE Trans. Biomed. Eng. 63 (3) (2016) 664–675.

[52] Y. Yuan, G. Xun, K. Jia, A. Zhang, A multi-view deep learning method for epileptic seizure detection using short-time Fourier transform, in: Proceedings of the Eight ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, ACM, 2017, pp. 213–222.

[53] A. Van Esbroeck, C.-C. Chia, Z. Syed, Heart rate topic models, in: The Twenty-Sixth AAAI Conference on Artificial Intelligence, volume 1001, 2012, p. 48109.



**Ye Yuan** is currently working toward the Ph.D. degree in the College of Information and Communication Engineering, Beijing University of Technology. His main research interests include data mining, machine learning, and bioinformatics.

**Guangxu Xun** is working toward the Ph.D. degree in computer science at the State University of New York at Buffalo. His main research interests include data mining, machine learning, language modeling, deep learning and bioinformatics.

**Qiuling Suo** is working toward the Ph.D. degree in computer science at the State University of New York at Buffalo. Her main research interests include data mining and health informatics.

**Kebin Jia** is Professor and Director of First-Class Disciplines Construction Office of Beijing University of Technology. He is a full Professor in the College of Information and Communication Engineering and the Director of the Digital Multimedia Information Processing Laboratory (MIPL). He received his M.S. degree and Ph.D. degree in Information and Communication Engineering from University of Science and Technology of China in 1990 and 1998, respectively. His research interests include multimedia and database systems, content-based image/video retrieval, image/video coding and processing, data mining, and pattern recognition. He has published over 200 research publications and authored 2 books in these areas. He has served as PI for more than 15 research projects from The National Natural Science Foundation of China (NSFC), 973 National Basic Research Program, and 863 Program. Dr. Jia is a senior member of the Chinese Institute of Electronics. His group has received the award of Outstanding and Innovator Group Award of Committee of Education of Beijing.

**Aidong Zhang** is a SUNY Distinguished Professor of Computer Science and Engineering at the State University of New York (SUNY) at Buffalo where she served as Department Chair from 2009 to 2015. She is currently on leave and serving as Program Director in the Information & Intelligent Systems Division of the Directorate for Computer & Information Science & Engineering, National Science Foundation. Her research interests include data analytics/data science, bioinformatics, and health informatics, and she has authored over 300 research publications in these areas. Dr. Zhang currently serves as the Editor-in-Chief of the IEEE Transactions on Computational Biology and Bioinformatics (TCBB). She served as the founding Chair of ACM Special Interest Group on Bioinformatics, Computational Biology and Biomedical Informatics during 2011–2015 and is currently Chair of its advisory board. She is also the founding and steering chair of ACM international conference on Bioinformatics, Computational Biology and Health Informatics. She has served as editor for several other journal editorial boards, and has also chaired or served on numerous program committees of international conferences and workshops. Dr. Zhang is an ACM Fellow and an IEEE Fellow.