# Unsupervised Feature Representation Based on Deep Boltzmann Machine for Seizure Detection

Tengzi Liu, Muhammad Zohaib Hassan Shah, Xucun Yan, and Dongping Yang

***Abstract*— The Electroencephalogram (EEG) pattern of seizure activities is highly individual-dependent and requires experienced specialists to annotate seizure events. It is clinically time-consuming and error-prone to identify seizure activities by visually scanning EEG signals. Since EEG data are heavily under-represented, supervised learning techniques are not always practical, particularly when the data is not sufficiently labelled. Visualization of EEG data in low-dimensional feature space can ease the annotation to support subsequent supervised learning for seizure detection. Here, we leverage the benefit of both the time-frequency domain features and the Deep Boltzmann Machine (DBM) based unsupervised learning techniques to represent EEG signals in a 2-dimensional (2D) feature space. A novel unsupervised learning approach based on DBM, namely DBM_transient, is proposed by training DBM to a transient state for representing EEG signals in a 2D feature space and clustering seizure and non-seizure events visually. The effectiveness of DBM_transient is demonstrated on a widely-used benchmark dataset from Bonn University (Bonn dataset) and a raw clinical dataset from Chinese 301 Hospital (C301 dataset), with a large fisher discriminant value, surpassing the abilities of other dimensionality reduction methods, including DBM converged to an equilibrium state, Kernel Principal Component Analysis, Isometric Feature Mapping, t-distributed Stochastic Neighbour Embedding, Uniform Manifold Approximation. Such feature representation and visualization can help physicians to understand better the normal versus epileptic brain activities of each patient and thus enhance their diagnosis and treatment abilities. The significance of our approach facilitates its future usage in clinical applications.**

***Index Terms*— Seizure detection, electroencephalogram (EEG), Deep Boltzmann Machine (DBM), discrete wavelet transform (DWT), Fisher's discriminant function.**

Manuscript received 2 July 2022; revised 23 December 2022 and 19 February 2023; accepted 5 March 2023. Date of publication 8 March 2023; date of current version 14 March 2023. The work of Dongping Yang was supported by the National Natural Science Foundation of China under Grant 12175242. *(Corresponding author: Dongping Yang.)*

Tengzi Liu is with the Research Center for Augmented Intelligence, Research Institute of Artificial Intelligence, Zhejiang Laboratory, Hangzhou 311100, China, also with the Quanzhou Institute of Equipment Manufacturing, Haixi Institute, Chinese Academy of Sciences, Quanzhou 362200, China, and also with the School of Electrical and Control Engineering, North University of China, Taiyuan 030000, China (e-mail: tengzi11134@163.com).

Muhammad Zohaib Hassan Shah is with the Research Center for Augmented Intelligence, Research Institute of Artificial Intelligence, Zhejiang Laboratory, Hangzhou 311100, China (e-mail: zohaib@zhejianglab.com).

Xucun Yan is with the School of Electrical and Information Engineering, The University of Sydney, Camperdown, NSW 2006, Australia (e-mail: xucun.yan@sydney.edu.au).

Dongping Yang is with the Research Center for Augmented Intelligence, Research Institute of Artificial Intelligence, Zhejiang Laboratory, Hangzhou 311100, China, also with the Quanzhou Institute of Equipment Manufacturing, Haixi Institute, Chinese Academy of Sciences, Quanzhou 362200, China, and also with the Fujian Institute of Research on the Structure of Matter, Chinese Academy of Sciences, Fuzhou 350002, China (e-mail: dpyang@zhejianglab.com).

Digital Object Identifier 10.1109/TNSRE.2023.3253821

## I. Introduction

AROUND 1% of the world population suffers from epilepsy, which brings huge trouble to patients, even life-threatening, while one-third of the patients are drug-resistant, requiring physical intervention [1], [2]. The Electroencephalogram (EEG) signals have been used as the ground truth for epileptic seizure detection for years to find the focal point of seizure and treat the injured brain tissues with medication and/or surgery. EEG signals carry informative features which can explain most of the normal and abnormal brain activities, especially epileptic seizures but require experienced specialists to annotate seizure events [3]. Clinically, visual scanning of EEG signals to identify seizure activities is time-consuming and error-prone, with low consistency among physicians [4].

In the literature, many supervised learning techniques have been developed as classifiers for seizure detection, achieving relatively higher accuracy, e.g., support vector machine (SVM), multi-layer perceptual (MLP) neural networks, deep neural networks (DNN), convolution neural networks (CNN), and so on [5], [6]. However, supervised seizure detection techniques require extensive monitoring and physicians' time-consuming evaluation to label EEG signals, which are commonly heavily under-represented and sometimes not sufficient enough to learn the seizure detection models with a large number of parameters [7]. Moreover, the disease of epilepsy can progress rapidly, and the corresponding EEG footprint patterns evolve, making the trained model impractical [8].

This issue can be naturally avoided by unsupervised learning, which explores the structure of complex data in extracting the essential latent properties and consequently reduce the dimension [7], [9]. It can lead to a better representation of complex data. Once a good representation is found, the inferred latent variables, rather than the data vector itself, can solve subsequent supervised learning tasks for seizure identification [9]. Furthermore, visualization of high-dimensional data in low-dimensional and visible feature space can help physicians to observe the disease progression and understand better the normal versus epileptic brain activities of each patient. Thus enhancing their diagnosis/treatment abilities [10].

To this end, we propose to leverage the benefits of both the spectral-temporal features and the Deep Boltzmann Machine (DBM) [11] based unsupervised data processing techniques in achieving a 2-dimensional (2D) representation and visualization of EEG signals for clustering seizure and non-seizure events visually.

Firstly, we extract informative features in the time-frequency domain, which carry sufficient and powerful information on seizure activities [12], [13], [14]. Continuous wavelet transforms (CWT) or discrete wavelet transforms (DWT) are generally used to extract spectral-temporal features of EEG signals as data preprocessing [15], [16], [17]. In our previous work [16], we have discovered significant low-dimensional spectral-temporal features of EEG signals and extracted their compact statistical information for seizure detection, achieving more reliable seizure detection performance in comparison to other features, e.g., frequency-domain features and intrinsic mode functions from empirical mode decomposition (EMD).

Secondly, a novel unsupervised learning approach based on DBM, trained to a transient state (DBM_transient) is employed to obtain the essential 2D latent features for data representation, visualization, as well as clustering seizure and non-seizure events. Unsupervised learning, such as Restricted Boltzmann Machine (RBM) or auto-encoder, using high-dimensional statistical space, acts as feature representation, from which the desired output can be robustly extracted by a linear readout and thus intuitively clustered [18]. Recently, Unsupervised learning techniques have also received much attention in signal analysis, e.g., Kernel Principal Component Analysis (KPCA) [19], Isometric Feature Mapping (Isomap) [20], t-distributed Stochastic Neighbour Embedding (t-SNE) [21], Uniform Manifold Approximation (UMAP) [22], Deep Boltzmann Machine (DBM) [11], Deep Belief Networks (DBN) [23], Kernel DBN [24], Generative adversarial networks (GAN) [25], and so on. Both DBM and DBN have shown a powerful ability for dimension reduction and served as highly effective methods to pre-train the deep network for supervised learning or reconstructing the data, e.g., the MNIST handwritten dataset and the Olivetti face dataset [18]. However, both of the methods require fine-tuning, which employs supervised back-propagation of error derivatives to achieve better representations [26]. Up till now, it remains elusive whether DBM without fine-tuning can show better representations of EEG signals and then facilitate clustering seizure and non-seizure events, particularly on the raw clinical EEG data [16].

Finally, the performance of our method is tested on both the widely used small-size benchmark dataset from Bonn University (Bonn dataset) [27] as well as a large-size raw clinical dataset from Chinese 301 hospital (C301 dataset) [16], and evaluated by Fisher's discriminant function [28] and classification accuracy of a subsequent linear SVM. Although the scope of this study is to achieve superior visualization in 2D feature space, our method is also comparable with the compared other methods in the classification performance with a subsequent linear SVM for both datasets.

The results demonstrate the effectiveness of our proposal in the 2D feature representation and clustering, in comparison to DBM trained to a converged state (DBM_converged), and other commonly used dimensionality reducing methods, e.g., KPCA [19], Isomap [20], t-SNE [21], and UMAP [22]. More significantly, the proposed DBM_transient can cluster seizure and non-seizure events clearly in a 2D feature space on both datasets, beyond the abilities of the compared methods. Therefore, the low-dimensional feature representation by the proposed DBM_transient is significant for the identification of seizure activities visually, indicating its future usage in clinical applications.

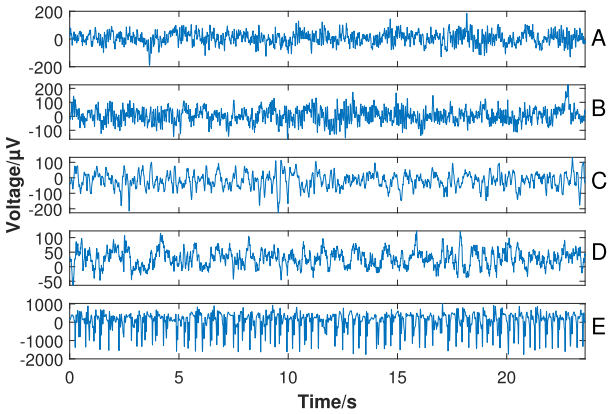


Fig. 1. EEG signals of 5 sets: A, B, C, D and E from Bonn dataset.

TABLE I
EXPERIMENT CASES ON BONN DATASET

| Classes | Case 1 | Case 2 | Case 3 | Case 4 | Case 5 | Case 6 | Case 7 |
|---|---|---|---|---|---|---|---|
| Seizure | E | E | E | E | E | E | E |
| Non-seizure | A | B | C | D | ACD | BCD | ABCD |

## II. DATASETS

In order to verify the general representation ability of DBM_transient, two datasets are used; one small-size credible and clean dataset–the widely-used benchmark Bonn dataset with only 500 samples, and the other large-size raw clinical dataset–C301 dataset of absence seizure with 10749 samples.

### A. Description of Bonn dataset

The dataset, available in [27], is recorded by attaching electrodes to the surface of awake subjects, using 10-20 international standard systems with a 128-channel amplifier [29]. It consists of seizure and non-seizure EEG signals, divided into five subsets (A - E) (see Fig. 1). Each subgroup is collected with a sampling frequency of 173.61 Hz [27], having 100 samples each with a duration of 23.6s (4096 data points) [30]. The data is denoised through a bandpass filter with a frequency band between 0.53 Hz and 40 Hz [27].

Subsets A and B were collected from the control group with eyes open and closed, respectively. Subsets C, D, and E were recorded from epileptic patients. The subsets C and D were collected in the epileptogenic interval, which is seizure-free. Data in subset E was acquired in lateral and nasal regions of the neocortex when a seizure occurs [27]. Seven common cases are explored in this work, with their details listed in Table I.

Subset C was collected by recording the signals in the hippocampal formation in the opposite hemispheres to the

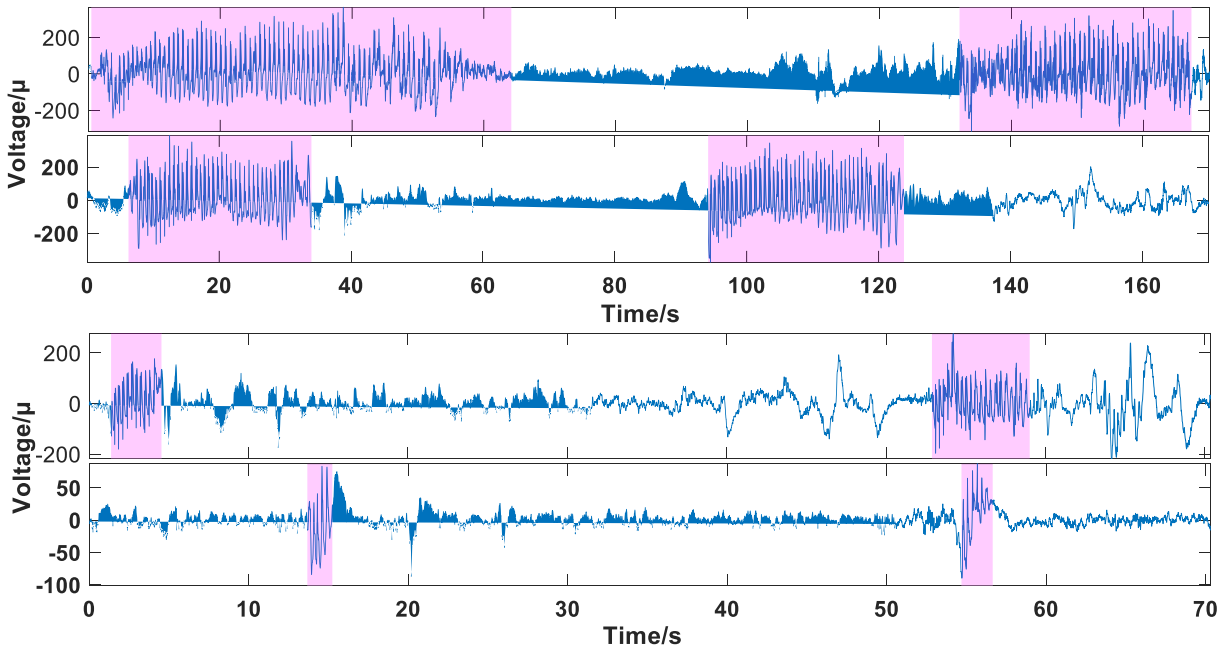


Fig. 2. Samples of EEG signals of 4 patients from C301 dataset. Pink shadows indicate seizure regions marked by physicians.

TABLE II
DATA INFORMATION OF C301 DATASET. $T_{MIN}$ AND $T_{MAX}$ DENOTE RESPECTIVELY THE MINIMUM AND MAXIMUM SEIZURE DURATIONS OF SEIZURE EVENTS OF EACH SUBJECT

| ID | Seizure events ($T_{min}$ - $T_{max}$) (s) | Total Seizure time (s) | Total Non-Seizure time (s) |
|---|---|---|---|
| 1 | 26 (7.4 - 84.4) | 980.0 | 2070 |
| 2 | 53 (2.0 - 21.8) | 300.8 | 2748 |
| 3 | 20 (12.1 - 45.3) | 499.8 | 2550 |
| 4 | 16 (1.6 - 45.3) | 89.4 | 2958 |
| 5 | 10 (1.6 - 8.3) | 45.3 | 3006 |
| 6 | 6 (5.1 - 14.1) | 55.5 | 2994 |
| 7 | 1 (7.8 - 7.8) | 7.8 | 3042 |
| 8 | 12 (3.5 - 8.6) | 78.3 | 2970 |
| 9 | 5 (3.1 - 18.6) | 51.4 | 3000 |
| 10 | 17 (1.2 - 12.5) | 129.2 | 2922 |
| 11 | 10 (2.3 - 11.3) | 80.8 | 2970 |
| 12 | 8 (2.0 - 7.2) | 44.3 | 3006 |
| 13 | 6 (2.7 - 9.4) | 34.4 | 3012 |
| 14 | 16 (2.3 - 42.1) | 179.8 | 2868 |
| 15 | 20 (2.3 - 13.7) | 133.5 | 2916 |
| 16 | 20 (2.3 - 11.7) | 133.5 | 2916 |
| 17 | 12 (2.0 - 15.6) | 95.4 | 2952 |
| 18 | 8 (1.6 - 7.4) | 41.1 | 3006 |
| 19 | 12 (8.6 - 22.5) | 206.4 | 2844 |
| Sum | 258 | 3186.6 | 54744 |

epileptogenic zones. While subset D was from the epileptogenic zones [29]. Their local hippocampal formation was resected to make their epilepsy completely controllable. The resection area was confirmed to be the Epileptogenic focus by clinical verification.

### B. Description of C301 dataset

C301 dataset was recorded on 19 absence seizure subjects, using 19 channel electrodes attached to the scalp with a sampling frequency of 256 Hz [16]. Only one of 19 channel electrodes is selected for processing, due to high spatial coherence of absence seizure EEG signals. The data information of each subject is summarized in Table II, where the seizure event numbers vary from 1 to 53 over different subjects, and the seizure duration is widely distributed. For each subject, the seizure-free time is generally dozens (around 17 times in average) of the seizure time, leading to a larger data-size unbalance for different classes than that of Bonn dataset. Classifying or clustering seizure and non-seizure data events in a highly unbalanced C301 dataset is extremely challenging for supervised or unsupervised training techniques, respectively.

Fig. 2 shows 4 samples of clinical EEG signals with seizure regions red-shaded by physicians. EEG signals of both seizure and non-seizure activities are diverse and complicated. Their durations and wave patterns of seizures vary from subject to subject, and non-seizure EEG patterns are diverse as well. It can be seen that the seizure activities in Fig. 2 show non-stereotyped wave patterns as well as stochastic and non-stationary characteristics. Furthermore, the clinical EEG signals often interfered with physiological artifacts (e.g., induced from ECG, pacemaker, eye movements, and sniffing) or non-physiological artifacts (e.g., from bed or chair movements and dropped electrodes) [16], [31], [32].

## III. METHODOLOGY

The above two EEG datasets are extracted, segmented, and labeled in Sec. III-A, and preprocessed by DWT to get meaningful statistics as input features in Sec. III-B. DBM is introduced in detail in Sec. III-C. Two methods evaluating clustering performance are introduced in Sec. III-E: Fisher's discriminant function and classification accuracy of a subsequent linear SVM.

### A. Data Preparation

For Bonn dataset, each sample with a duration of 23.6s is already prepared in the original data. In contrast, for C301 dataset, the continuous EEG recordings are segmented as usual

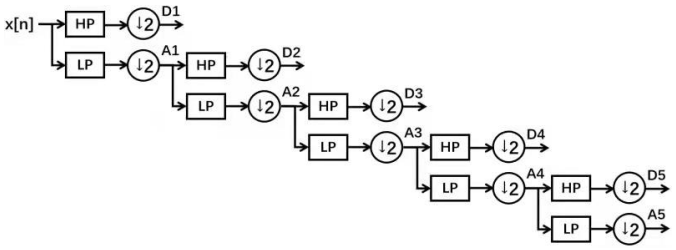


Fig. 3. Divisions of a signal $x[n]$ into different frequency bands by DWT, resulting in 5-level half down-sampled wavelet coefficient sets: A1, D1, A2, D2, A3, D3, A4, D4, A5 and D5. The signal $x[n]$ can be fully described by D1, D2, D3, D4, D5 and A5.

into half-overlapping 4s samples. As listed in Table II, there are many short sections of seizure activities with the duration less than 4s. To make a clean dataset without ambiguous segments with mix periods of seizure and non-seizure events, all samples are scanned visually. As a result, 10794 samples are picked out, where 9992 non-seizure samples are labeled as 0 and 757 seizure samples are labeled as 1.

### B. Data Preprocessing

Joint spectral-temporal features are believed to contain sufficient and powerful information about seizure activities [33]. The information in both time and frequency domains should be captured simultaneously using Wavelet transform techniques, which can effectively extract arbitrary signal bands by changing the size and position of the time window described as follows:

$$x_\omega(j,k) = 2^{-j/2} \int_{-\infty}^{+\infty} x(t)\psi\left(\frac{t-k2^j}{2^j}\right) dt, \qquad (1)$$

where $2^j$ and $k2^j$ represent the scale and time location or shifting parameters respectively, and $\psi(t)$ is the mother wavelet, set as Daubechies order-4 (db4) wavelet here [34].

CWT is employed to obtain the time-frequency images with high-dimension features, which are quite complicated and involves redundant information, leading to large training time and likely over-fitting. The essential low-dimensional information of these images has been discovered in our previous work [16]. One alternative is DWT, which uses low-pass and high-pass filters to select dynamic scales and positions without the redundant information obtained from CWT [34]. DWT divides a signal $x[n]$ with a sampling frequency $fs$ through the low pass (LP), and high pass (HP) filters into an approximate coefficient (A1) within a lower-frequency band $[0, fs/2]$ and a detail coefficient (D1) within a higher-frequency band $[fs/2, fs]$, respectively. A1 can be subdivided similarly into A2 and D2. In this way, as shown in Fig. 3, the signal $x[n]$ can be divided by DWT through 5 levels to obtain 5 wavelet coefficient sets: A1, D1, A2, D2, A3, D3, A4, D4, A5, and D5, having frequency bands $[0, fs/2]$, $[fs/2, fs]$, $[0, fs/4]$,$[fs/4, fs/2]$, $[0, fs/8]$, $[fs/8, fs/4]$, $[0, fs/16]$, $[fs/16, fs/8]$, $[0, fs/32]$ and $[fs/32, fs/16]$, respectively. After each division, the resulting coefficient sets are half-down sampled from the preceding coefficients, as indicated in circles in Fig. 3. Consequently, the signal $x[n]$ can be completely represented by the following wavelet coefficients: D1, D2, D3, D4, D5, and A5.

TABLE III
FREQUENCY RANGE OF THE DECOMPOSED SIGNAL FROM BOTH DATASETS

| Decomposed signal | Frequency range (Hz) | |
|---|---|---|
| | Bonn | C301 |
| D3 | 10.8 - 21.7 | 16 - 32 |
| D4 | 5.4 - 10.8 | 8 - 16 |
| D5 | 2.4 - 5.4 | 4 - 8 |
| A5 | 0 - 2.7 | 0 - 4 |

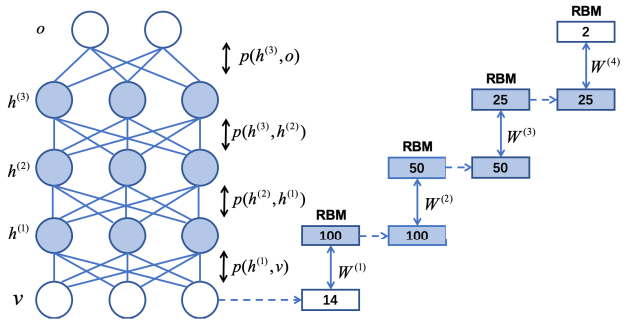


Fig. 4. DBM with five layers (left) as a stack of four RBMs (right). The vector $v$ has 14 input units; $h^{(i)}$ represents the unit vector of the $i$-th hidden layer, whose unit number are set to 100, 50 and 25 for 1st, 2nd and 3rd hidden layer, respectively; The output vector $o$ only has two units for feature representation; The matrix $W^{(i)}$ represents the connection weights of the $i^{th}$ RBM. Open and shadowed circles in DBM (or rectangles in RBM) represent linear-valued and binary units, respectively.

Only D3, D4, D5, and A5 are employed for further processing, as seizure activities are dominated by the low-frequency domains, whose frequency ranges for Bonn and C301 datasets are listed in Table III. Following the advise in [15] and [35], four statistics for the resulted wavelet coefficients are employed to represent the spectral-temporal features:

- Time-averaged absolute values for each coefficient: $\mathrm{E}[|D3(t)|]$, $\mathrm{E}[|D4(t)|]$, $\mathrm{E}[|D5(t)|]$, $\mathrm{E}[|A5(t)|]$;
- Time-averaged power for each coefficient: $\mathrm{E}[|D3(t)|^2]$, $\mathrm{E}[|D4(t)|^2]$, $\mathrm{E}[|D5(t)|^2]$, $\mathrm{E}[|A5(t)|^2]$;
- Standard deviation for each coefficient: $\mathrm{Std}[D3(t)]$, $\mathrm{Std}[D4(t)]$, $\mathrm{Std}[D5(t)]$, $\mathrm{Std}[A5(t)]$;
- Ratio of time-averaged absolute values of adjacent sub-band coefficients: $\frac{\mathrm{E}[|D4(t)|]}{\mathrm{E}[|D3(t)|]}$, $\frac{\mathrm{E}[|D5(t)|]}{\mathrm{E}[|D4(t)|]}$,

where $\mathrm{E}[S(t)]$ and $\mathrm{Std}[S(t)]$ represent time-average and corresponding standard deviation of time series $S(t)$, respectively. Here, each data sample is characterized by 14 feature variables, each of which is normalized into the range $[0, 1]$.

### C. Deep Boltzmann Machine for Feature Representation

As illustrated in Fig. 4, training DBM with multiple hidden layers $\left(h^{(1)}, h^{(2)}, h^{(3)}\right)$ can be regarded as learning a stack of RBMs by introducing a greedy, layer-by-layer unsupervised learning algorithm for each RBM [36], [37] and treating the hidden activities of one RBM as the data for training a higher-level RBM.

RBM is formed between two adjacent layers, with the lower as the visible layer and the higher as the hidden layer, as shown in Fig. 5. Taking the Bernoulli-Bernoulli RBM for example,

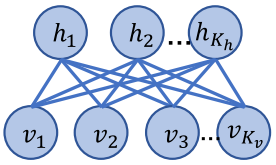


Fig. 5. RBM with $K_v$ binary variables in the visible layer and $K_h$ binary variables in the hidden layer.

it contains a set of visible units $\boldsymbol{v} = \{0, 1\}^{K_v}$ and a set of hidden units $\boldsymbol{h} = \{0, 1\}^{K_h}$. The energy of the state $\{\boldsymbol{v}, \boldsymbol{h}\}$ is defined as:

$$E(v, h) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_i \sum_j v_i w_{ij} h_j, \quad (2)$$

where $a_j$, $b_j$ and $w_{i,j}$ are parameters of RBM: $a_j$ is the bias of $j$-th visible unit $v_i$; $b_j$ is the bias of $j$-th hidden unit $h_j$; $w_{i,j}$ is the symmetrically coupling weight between $v_i$ and $h_j$. The joint probability of the state $\{\boldsymbol{v}, \boldsymbol{h}\}$ is defined as:

$$p(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{Z} \exp(-E(\boldsymbol{v}, \boldsymbol{h})), \quad (3)$$

with the partition function $Z = \sum_{\boldsymbol{v},\boldsymbol{h}} \exp(-E(\boldsymbol{v}, \boldsymbol{h}))$. Marginalizing out the hidden states leads to:

$$p(\boldsymbol{v}) = \frac{1}{Z} \sum_{\boldsymbol{h}} \exp(-E(\boldsymbol{v}, \boldsymbol{h})). \quad (4)$$

From (3) and (4), the conditional probability of $h_j = 1$ given the visible units can be obtained:

$$p\left(h_j = 1 \mid \boldsymbol{v}\right) = \sigma\left(b_j + \sum_i w_{ij} v_i\right), \quad (5)$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$ is a sigmoid function. Similarly, the conditional probability of $v_i = 1$ given hidden units is:

$$p\left(v_i = 1 \mid \boldsymbol{h}\right) = \sigma\left(a_i + \sum_j w_{ij} h_j\right). \quad (6)$$

The conditional probability distribution provides the basis for Gibbs sampling to update parameters in the following.

In RBM, given a set of training data $D$, the parameters $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{w}$ are learned to maximize the log-likelihood function:

$$L(D; \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{w}) = \frac{1}{N} \sum_{n=1}^{N} \log p\left(\boldsymbol{v}^{(n)}; \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{w}\right), \quad (7)$$

the gradient of which with respect to the model parameters leads to:

$$\frac{\partial L(D; \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{w})}{\partial a_i} = \langle v_i \rangle_{\text{data}} - \langle v_i \rangle_{\text{model}}, \quad (8)$$

$$\frac{\partial L(D; \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{w})}{\partial b_j} = \langle h_j \rangle_{\text{data}} - \langle h_j \rangle_{\text{model}}, \quad (9)$$

$$\frac{\partial L(D; \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{w})}{\partial w_{i,j}} = \langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{model}}, \quad (10)$$

where $\langle \bullet \rangle_{\text{data}}$ represents a data-dependent expectation with respect to the completed empirical data distribution, and $\langle \bullet \rangle_{\text{model}}$ denotes a data-independent expectation with respect to the distribution defined by the model. Then the parameters $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{w}$ are updated by the gradient ascending method:

$$a_i = a_i + \alpha \left(\langle v_i \rangle_{\text{data}} - \langle v_i \rangle_{\text{model}}\right), \quad (11)$$

$$b_j = b_j + \alpha \left(\langle h_j \rangle_{\text{data}} - \langle h_j \rangle_{\text{model}}\right), \quad (12)$$

$$w_{ij} = w_{ij} + \alpha \left(\langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{model}}\right), \quad (13)$$

with the learning rate $\alpha = 0.1$.

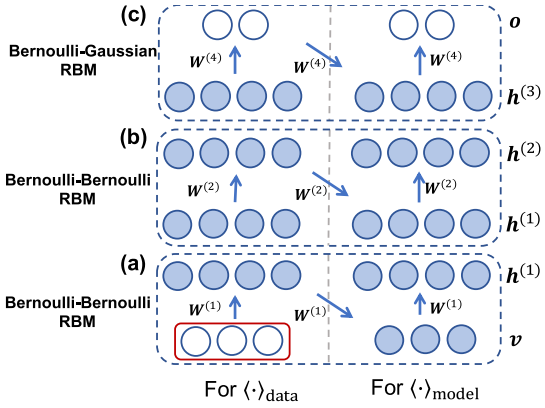


Fig. 6. Schematic diagram of CD-1 algorithm for training each RBM in the DBM. The empty and filled circles denote the real-valued and binary-valued units, respectively. Specifically, the red-framed input units are regarded as logistic units, for which only binary value 0 or 1 is sampled in MCMC.

The distribution from the completed model is intractable but can be approximated by Gibbs sampling in the way of Monte Carlo Markov Chains (MCMC). Contrastive divergence [38] (CD) has been developed to approximate the data-independent expectation under the RBM distribution, as described in Fig. 6. First, the dataset is partitioned into small batches, and the gradient ascent is performed sequentially over all these batches in random order. Starting from a data point, a random configuration of the hidden layer is sampled according to the conditional probability $p\left(h_j = 1 \mid v\right)$ for each $j$-th unit in Eq. 5, as depicted in Fig. 6. In turn, given this, a configuration of the visible layer is sampled according to Eq. (6) and so on. Thus, one can take advantage of the bipartite structure of RBM to draw a whole visible or hidden layer at once, thanks to the factorization of the conditional distribution in Eq. (6) or (5), respectively. The approximation using $n$ iterated sampling of both visible and hidden units is known as CD-$n$, and CD-1 as shown in Fig. 6 works surprisingly well in practice [18], [39]. Therefore, the sampled configurations can be used to estimate the data-independent expectation under the RBM model through CD-1 algorithm, as described in Fig. 6.

In Fig. 4, the input data is fed into the bottom layer, while the top layer outputs two non-linear features to represent the non-linear structure of the input data. Both are real-valued, while all hidden nodes are binary, only taking 0 or 1 to speed up the training [18], [40]. Thus, the above-introduced training method for the Bernoulli-Bernoulli RBM can be applied to the training of $\boldsymbol{h}^{(1)} \leftrightarrow \boldsymbol{h}^{(2)}$ and $\boldsymbol{h}^{(2)} \leftrightarrow \boldsymbol{h}^{(3)}$. $\boldsymbol{h}^{(1)} \leftrightarrow \boldsymbol{h}^{(2)}$ is shown in Fig. 6(b). The output of the top layer in Fig. 4 is set as stochastic real values drawn from a Gaussian to make good use of continuous variables, allowing data representation in a 2D feature space. Then $\boldsymbol{h}^{(3)} \leftrightarrow \boldsymbol{o}$ in Fig. 6(c) is regarded as

a Bernoulli-Gaussian RBM, whose energy function should be correspondingly modified as:

$$E(v,h) = -\sum_i a_i v_i - \sum_j \frac{(h_j - b_j)^2}{2\sigma_j^2} - \sum_i \sum_j v_i w_{ij} \frac{h_j}{\sigma_j}, \tag{14}$$

where $\sigma_j$ is the standard deviation of the Gaussian noise for $j$-th output unit. All $\sigma_j$ are set to 1 for our results in Sec. IV-B and IV-C. The stochastic update rule for the visible units remains the same except that each $h_j$ is divided by $\sigma_j$, and the update rule for hidden units $j$ is to sample from a Gaussian with mean $b_j + \sigma_j \sum_i v_i w_{ij}$ and variance $\sigma_j^2$.

As mentioned in Sec. III-B, the input values are here normalized into the range $[0, 1]$, and the visible units are regarded as logistic units in $\boldsymbol{v} \leftrightarrow \boldsymbol{h}^{(1)}$), as suggested in Hinton's seminal work [18], [40], [41]. If the visible units are set as linear units with Gaussian noise and $\boldsymbol{v} \leftrightarrow \boldsymbol{h}^{(1)}$) is regarded as a Gaussian-Bernoulli RBM, the training can not achieve a good 2D representation for clustering seizure and non-seizure events (data not shown here). Further explanation is presented in Sec. IV-A.

Therefore, our DBM model can be decomposed into one special Bernoulli-Bernoulli RBM for $\boldsymbol{v} \leftrightarrow \boldsymbol{h}^{(1)}$) (Fig. 6(a)), two normal Bernoulli-Bernoulli RBM for $\boldsymbol{h}^{(1)} \leftrightarrow \boldsymbol{h}^{(2)}$) and $\boldsymbol{h}^{(2)} \leftrightarrow \boldsymbol{h}^{(3)}$) (only $\boldsymbol{h}^{(1)} \leftrightarrow \boldsymbol{h}^{(2)}$) shown in Fig. 6(b)), and one Bernoulli-Gaussian RBM for $\boldsymbol{h}^{(3)} \leftrightarrow \boldsymbol{o}$ (Fig. 6(c)).

It is important to point out that DBM should be shortly trained into a transient state, not converging to an equilibrium state. DBM_transient are obtained by training each RBM of the DBM model only 10 rounds of MCMC Gibbs sampling, which is far away from reaching an equilibrium state ($\sim$ 200 rounds required for most cases). For DBM_converged, each RBM of the DBM model is trained to an equilibrium state. The similar results are achieved in many different trials with random initialization of the network parameters, implying that our results are statistically significant.

### D. The Compared Methods

Our proposed DBM_transient is compared to DBM_converged, KPCA, Isomap, t-SNE, and UMAP. Since the normalization is done linearly, Principal Component Analysis (PCA) is affected by non-Gaussian distributions of the features (see Fig. 7). Thus, PCA is not a good candidate for comparison and is replaced by KPCA, which employs a nonlinear transformation [28]. For KPCA, the RBF kernel with gamma equal to 0.5 is chosen. Isomap, t-SNE and UMAP can estimate the intrinsic geometry of the data manifold by exploring neighbors of each data point on the manifold, and embed high-dimensional data in a low-dimensional space for visualizaion. For Isomap and UMAP, the number of neighbors is set to 5 and 15, respectively. For t-SNE, the perplexity and number of iterations equal to 30 and 1000, respectively, producing the best performance for the datasets. Rest of the parameters are set to default or 'auto' as given in their corresponding python libraries. Following the Grid search cross-validation technique, the best parameters are selected for each method.

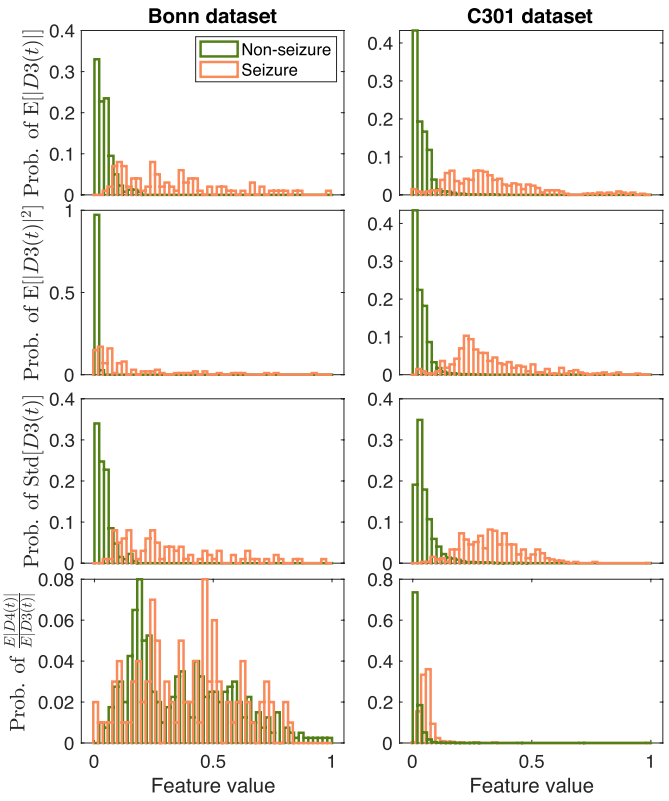


Fig. 7. Feature distributions of the input data for both Bonn dataset and C301 dataset. Here only $\mathrm{E}[|D3(t)|]$, $\mathrm{E}[|D3(t)|^2]$, $\mathrm{Std}[D3(t)]$ and $\frac{\mathrm{E}[|D4(t)|]}{\mathrm{E}[|D3(t)|]}$ are shown as examples to indicate the complicated distributions of the features.

DBM can adaptively learn complex nonlinear global patterns in the data (see Fig. 7) [36], which is beyond the ability of the compared nonlinear dimensionality reduction methods. KPCA should pre-define the kernel, which may be not suitable for the complex data, while Isomap, t-SNE, UMAP can only capture the local relation structures among samples by neighbor embedding. In the results, superiority of DBM_transient over the compared methods on the 2D representations of EEG signals is evaluated by the two evaluation metrics: Fisher's discriminant function and classification measures from a subsequent linear SVM.

### E. Evaluations

**(1)** Here, Fisher's discriminant function [28], as a measure of class separation, is maximized to select a proper projection for linear classification, evaluating the clustering performance of a 2D representation. The projection of the 2D feature $x_n$ of the $n$-th data sample $X_n$ down to one dimension $y_n$ can be given as: $y_n = \omega^T x_n$, where $\omega$ is called a weight vector. Fisher's discriminant function is defined to be the ratio of the between-class variance to the within-class variance and can be given by:

$$J(\omega) = \frac{\omega^T S_B \omega}{\omega^T S_W \omega}, \tag{15}$$

where the between-class covariance matrix $S_B$ and the within-class covariance matrix $S_W$

are given as:

$$S_B = (m_2 - m_1)(m_2 - m_1)^T, \tag{16}$$

$$S_W = \frac{1}{N_1}\sum_{n\in C_1}(x_n - m_1)(x_n - m_1)^T + \frac{1}{N_2}\sum_{n\in C_2}(x_n - m_2)(x_n - m_2)^T, \tag{17}$$

with mean vectors: $m_1 = \frac{1}{N_1}\sum_{n\in C_1} x_n$ and $m_2 = \frac{1}{N_2}\sum_{n\in C_2} x_n$ The larger value of Fisher's discriminant function $J(\omega)$ indicates the better separation of two classes, and thus the better clustering performance.

**(2)** SVM with a linear kernel [28], [42] is a linear classification method for evaluations, which is used to search the best support vector in the data hyperplane to maximize the margins of different classes.

The binary classification performance of SVM can be defined as the standard statistical measures: specificity (SPE), recall (REC), precision (PRE), accuracy (ACC), and F1 Score, defined as:

$$\text{SPE} = \frac{\text{TN}}{\text{TN}+\text{FP}},$$
$$\text{REC} = \frac{\text{TP}}{\text{TP}+\text{FN}},$$
$$\text{PRE} = \frac{\text{TP}}{\text{TP}+\text{FP}},$$
$$\text{ACC} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}},$$
$$\text{F1 Score} = \frac{2\cdot\text{PRE}\cdot\text{REC}}{\text{PRE}+\text{REC}},$$

where TP (true positive) is the number of samples correctly classified as seizure class, TN (true negative) represents the number of correctly as non-seizure class, FP (false positive) indicates the number of samples incorrectly detected as seizure class, and FN (false negative) means the number of samples incorrectly detected as non-seizure class. ACC is the ratio of correctly predicted samples in the testing dataset. F1 Score measures the geometric average of PRE and REC.

## IV. Experimental Results

Section IV-A presents the complexity of feature distributions, indicating the difficulty of isolating seizure events from non-seizure ones. By the DBM, 14 features of each sample can be represented and visualized in a 2D space for clearly clustering seizure and non-seizure events of both the small-size Bonn dataset (Sec. IV-B) and a much more complex large-size raw clinical C301 dataset (Sec. IV-C). DBM_transient is superior over other methods on the performance of representation and visualization, as demonstrated with a higher value of Fisher's discriminant function and a higher classification accuracy by a subsequent linear SVM in Sec. IV-B and IV-C.

### A. Distributions of Input Features

The 14 spectral-temporal features of EEG signals extracted by DWT are regarded as the input to the DBM. These are real-valued and normalized into the range $[0, 1]$. The visible units in $v \leftrightarrow h^{(1)}$ are regarded as logistic units rather than linear units with Gaussian noise. Thus, $v \leftrightarrow h^{(1)}$ is trained as a special Bernoulli-Bernoulli RBM, as depicted in Fig. 6(a).

The scheme is reasonable since the features of both datasets are non-trivially distributed, as shown in Fig. 7. Here only $\text{E}[|D3(t)|]$, $\text{E}[|D3(t)|^2]$, $\text{Std}[D3(t)]$ and $\frac{\text{E}[|D4(t)|]}{\text{E}[|D3(t)|]}$ are taken as examples to show the complicated distribution of features. The main points discovered in Fig. 7 are:

- Non-seizure features are narrowly distributed bounded to 0, while the seizure ones are widely and irregularly distributed over the range [0, 1];
- The distributions of both seizure and non-seizure data features are overlapped;
- Almost all distributions of either seizure or non-seizure features are non-Gaussian, with multiple peaks;
- Bonn dataset has a wide distributed ratio feature $\frac{\text{E}[|D4(t)|]}{\text{E}[|D3(t)|]}$ of both seizure and non-seizure events, while that of C301 dataset is mostly located around 0;
- The ratio feature $\frac{\text{E}[|D4(t)|]}{\text{E}[|D3(t)|]}$ has a high overlap of seizure and non-seizure data for Bonn dataset, while the overlap for C301 dataset is greatly reduced.

The non-Gaussian variability of input features suggests keeping their complicated and intrinsic data structure rather than introducing a simplified Gaussian noise. Therefore, it is justified to train $v \leftrightarrow h^{(1)}$ as a special Bernoulli-Bernoulli RBM (Fig. 6(a)) in order to extract the essential latent properties for a better low-dimensional representation.

### B. Feature Representation for Bonn Dataset

Here, DBM is employed to reduce the dimensions of 14 features nonlinearly and represent them in a 2D feature space, where one can intuitively identify seizure events with no further complicated computation. This approach effectively clusters seizure and non-seizure events in a 2D feature representation space, as shown in Fig. 8 for all 7 cases of Bonn dataset. The data class is colored according to the data label, which is only used for visualization and evaluation of clustering performance rather than for training DBM. Seizure and non-seizure samples gather into linearly separable clusters, although the task is increasingly complex with more and more diverse non-seizure data from Case 1 to Case 7. The scatter points, representing samples of each class, form into a Gaussian shape distribution due to the Gaussian noise, which is introduced into the output layer $\boldsymbol{o}$ of Bernoulli-Gaussian RBM $\boldsymbol{h}^{(3)} \leftrightarrow \boldsymbol{o}$. Seizure and non-seizure events can thus be spatially separated in a 2D feature space, with each class dominating a respective area and easily identified by empirical judgment without label information.

Figure 8 also shows the 2D feature presentations of 5 other dimensionality reducing methods: DBM_converged, KPCA, Isomap, t-SNE, and UMAP for comparison. The clustering performance of DBM_converged is much weakened as compared with DBM_transient (Fig. 8). Understanding the underlying mechanism remains elusive, although some recent work shed a bit of light on it [43], [44], [45]. The superiority of clustering performance by DBM_transient is demonstrated with much higher values of Fisher's discriminant function

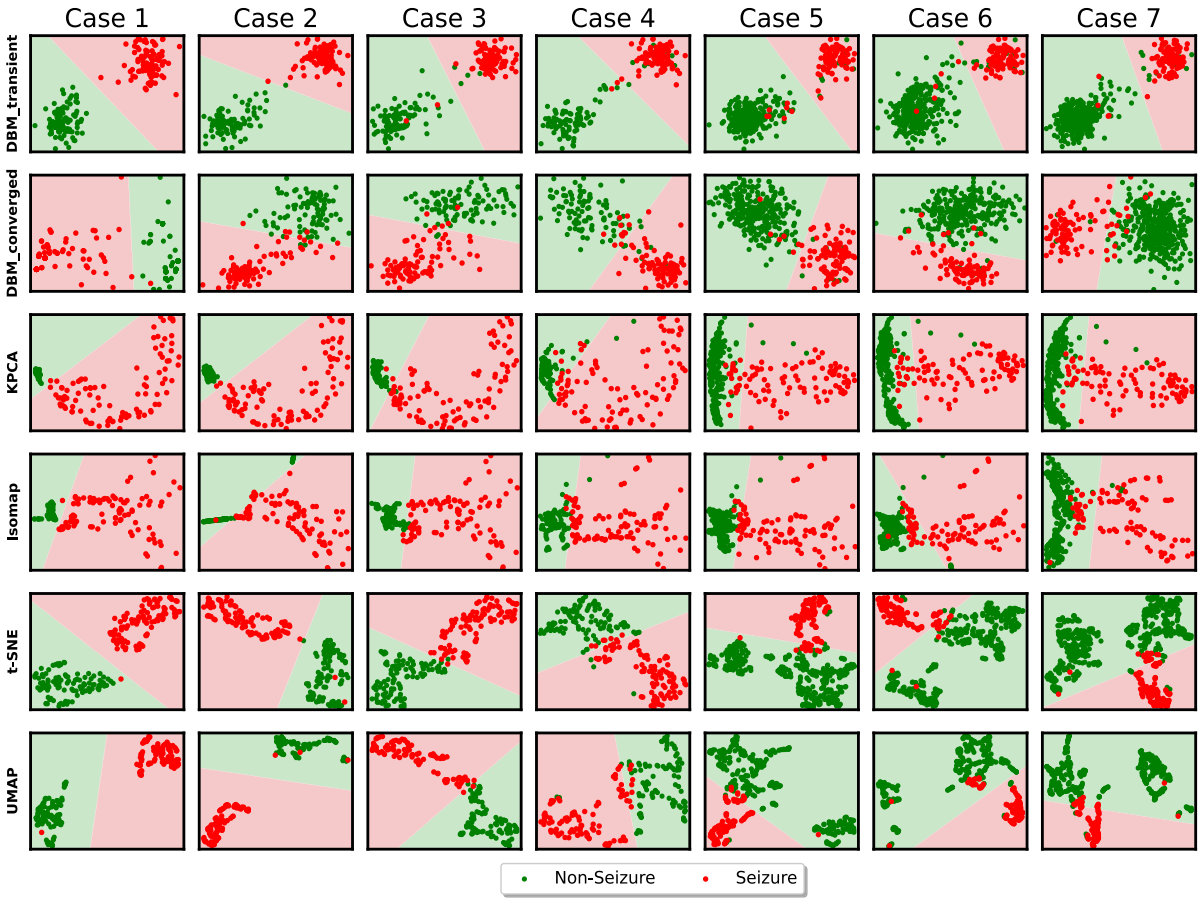


Fig. 8. 2D feature representations of seizure (red dots) and non-seizure (green dots) events in each case of Bonn dataset using 6 different unsupervised learning methods. The data class is colored according to the data label, which is only used for visualization and clustering performance evaluation rather than training. The background colors are obtained by a subsequent linear SVM, with the color boundary as the separating line.

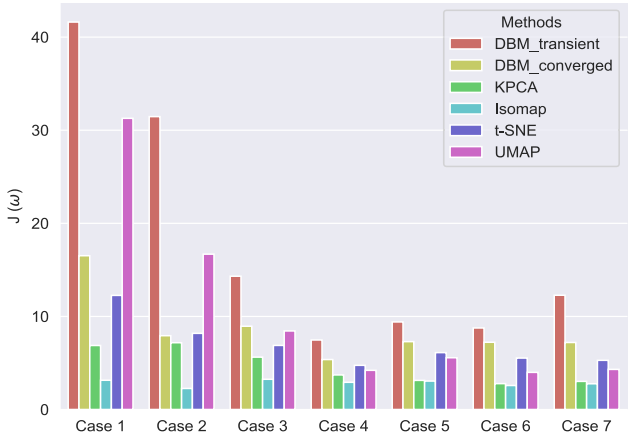


Fig. 9. Fisher's discriminant function J($\omega$) for 6 different unsupervised learning methods in each case of Bonn dataset.

$J(\omega)$ (as shown in Fig. 9) and higher classification accuracy of a subsequent linear SVM (boundary line indicated by the edge of different background colors in Fig. 8) on the testing data, as presented in Table IV. In Fig. 9, $J(\omega)$ consistently achieves the highest values by DBM_transient for all 7 cases, compared to the other methods: DBM_converged, KPCA, Isomap, t-SNE, and UMAP, despite an increase in task complexity from Case 1 to Case 7. The 2D low-dimensional feature representations by Isomap, t-SNE, and UMAP are similar shown in Fig. 8, with the corresponding $J(\omega)$ much smaller than that of DBM_transient (Fig. 9).

As seizure and non-seizure events are clustered and separated with a large between-class distance, seizure activities can be easily identified by physicians or training a subsequent linear SVM. The classification by the linear SVM following DBM_transient achieves an average accuracy of 0.968 for all the cases, which is higher than those achieved by the SVM following DBM_converged (0.951), KPCA (0.937), Isomap (0.863), t-SNE (0.956), and UMAP (0.946). Additionally, DBM_transient achieves the highest values (REC, and F1 Score) in almost all the cases as listed in Table IV. Specifically, REC of DBM_transient shows much higher scores than the other methods, especially in more complex and unbalanced training groups (Case 5 to Case 7).

Furthermore, the results from the proposed DBM_transient followed by the supervised SVM are also compared with the unsupervised K-means and multiscale K-means (MSK-means) methods for Bonn dataset (See Table 8.1 in [3], where K-means and MSK-means are specially treated as classifiers). The results verify the effectiveness of DBM_transient in extracting the essential latent properties of EEG signals for representation and visualization.

### C. Feature Representation for C301 Dataset

Bonn dataset is much cleaner than the acquired raw clinical data. However, the small size of Bonn dataset with only

TABLE IV
SVM BASED CLASSIFICATION PERFORMANCE ON BONN DATASET USING 6 DIFFERENT METHODS

| Criteria | Method | Case 1 | Case 2 | Case 3 | Case 4 | Case 5 | Case 6 | Case 7 | Average |
|---|---|---|---|---|---|---|---|---|---|
| SPE | DBM_transient | **1** | 0.990 | 0.958 | 0.911 | 0.966 | 0.962 | 0.969 | 0.965 |
| | DBM_converged | 0.990 | 0.954 | 0.974 | 0.916 | 0.970 | 0.983 | 0.986 | 0.967 |
| | KPCA | **1** | **1** | **1** | 0.980 | 0.989 | 0.988 | 0.992 | **0.992** |
| | Isomap | **1** | 0.873 | **1** | **0.986** | **0.994** | 0.965 | 0.966 | 0.969 |
| | t-SNE | **1** | 0.993 | 0.987 | 0.964 | 0.993 | **0.990** | **0.993** | 0.988 |
| | UMAP | **1** | **1** | 0.993 | 0.944 | 0.992 | 0.981 | 0.992 | 0.986 |
| REC | DBM_transient | **1** | 0.982 | **0.976** | **0.973** | **0.930** | **0.950** | **0.959** | **0.967** |
| | DBM_converged | 0.982 | 0.931 | 0.954 | 0.883 | 0.925 | 0.902 | 0.848 | 0.917 |
| | KPCA | 0.968 | 0.968 | 0.863 | 0.759 | 0.766 | 0.724 | 0.738 | 0.825 |
| | Isomap | 0.706 | 0.698 | 0.724 | 0.697 | 0.689 | 0.666 | 0.692 | 0.696 |
| | t-SNE | 0.987 | **0.983** | 0.975 | 0.885 | 0.827 | 0.831 | 0.779 | 0.895 |
| | UMAP | 0.991 | 0.972 | 0.967 | 0.779 | 0.796 | 0.752 | 0.721 | 0.854 |
| ACC | DBM_transient | **1** | 0.986 | 0.967 | **0.942** | **0.959** | **0.959** | **0.967** | **0.968** |
| | DBM_converged | 0.986 | 0.942 | 0.964 | 0.899 | **0.959** | 0.953 | 0.958 | 0.951 |
| | KPCA | 0.984 | 0.984 | 0.931 | 0.869 | 0.933 | 0.922 | 0.941 | 0.937 |
| | Isomap | 0.853 | 0.785 | 0.862 | 0.841 | 0.917 | 0.879 | 0.910 | 0.863 |
| | t-SNE | 0.993 | **0.988** | 0.924 | 0.938 | 0.951 | 0.950 | 0.950 | 0.956 |
| | UMAP | 0.995 | 0.986 | **0.980** | 0.861 | 0.943 | 0.924 | 0.938 | 0.946 |
| F1 Score | DBM_transient | **1** | **0.988** | 0.967 | **0.944** | 0.915 | **0.924** | **0.921** | **0.951** |
| | DBM_converged | 0.986 | 0.941 | 0.963 | 0.897 | **0.919** | 0.921 | 0.891 | 0.931 |
| | KPCA | 0.984 | 0.983 | 0.925 | 0.851 | 0.851 | 0.820 | 0.832 | 0.892 |
| | Isomap | 0.826 | 0.763 | 0.838 | 0.813 | 0.805 | 0.733 | 0.754 | 0.790 |
| | t-SNE | 0.994 | **0.988** | 0.920 | 0.937 | 0.893 | 0.893 | 0.861 | 0.926 |
| | UMAP | 0.996 | 0.985 | **0.979** | 0.847 | 0.874 | 0.830 | 0.822 | 0.904 |

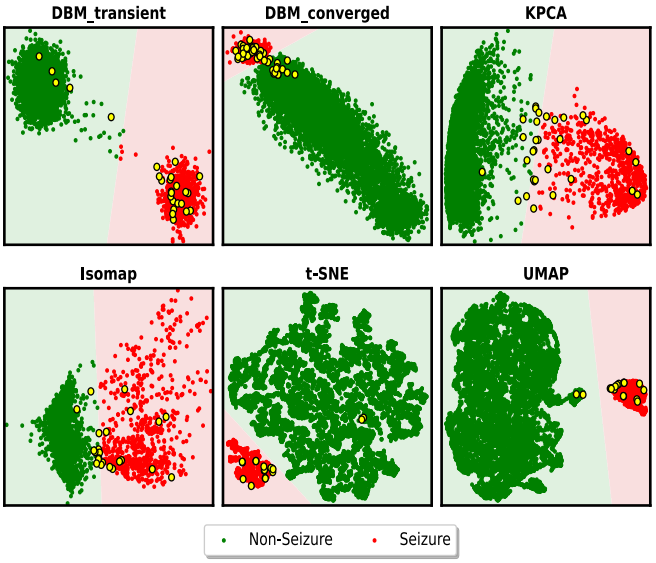


Fig. 10. 2D feature representations of seizure (red dots) and non-seizure (green dots) events for C301 dataset using 6 different unsupervised learning methods, with the description similar to Fig. 8. Additionally, the highlighted yellow dots represents the misclassified samples, eight of which are selected as examples with their original signals plotted in Fig. 11 for further discussions.

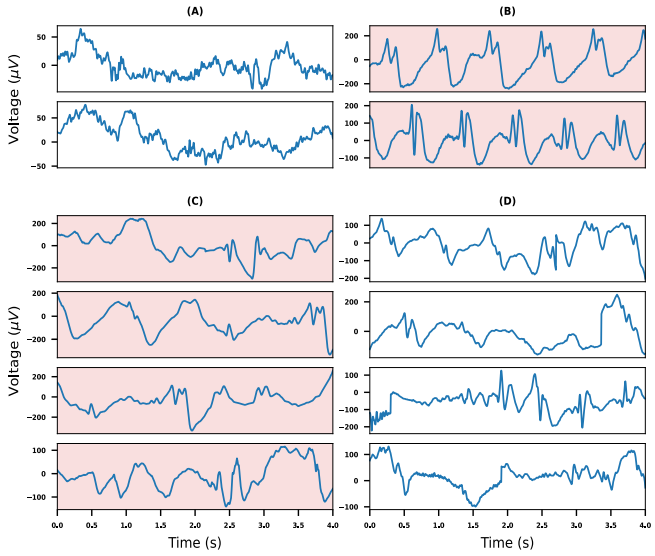


Fig. 11. Examples of original EEG signals with seizure parts red shadowed according to physician's labels. (A) Correctly classified non-seizure samples; (B) Correctly classified seizure samples; (C) Non-seizure samples misclassified as seizure samples; (D) Seizure samples misclassified as non-seizure samples.

500 samples completely hinders the commonly used unsupervised learning methods (e.g. DBM_converged, Isomap, t-SNE and UMAP) from achieving a good low-dimensional feature representation, since the data size is not enough for these methods to converge to a good representation. Surprisingly, DBM_transient can extract the intrinsic latent properties and successfully represent the data in a low-dimensional space for clustering seizure and non-seizure events.

C301 dataset is much more complicated and unbalanced than Bonn dataset. The EEG signals of C301 dataset are stochastic, non-stationary and diverse with non-stereotyped EEG patterns (Fig. 11). Moreover, these raw clinical EEG signals are often polluted by physiological or non-physiological artifacts [16]. However, C301 dataset has a 20 times larger size (757 seizure samples and 9992 non-seizure samples) than Bonn dataset, which may benefit the unsupervised learning methods for clustering. As expected, the large size of C301 dataset can improve the representation performance (Fig. 10) of almost all the methods employed, with a much higher Fisher's discriminant function $J(\omega)$ as listed in Table V.

TABLE V
SVM BASED CLASSIFICATION PERFORMANCE ON C301 DATASET USING 6 DIFFERENT METHODS

| Method | SPE | REC | ACC | F1 Score | J ($\omega$) |
|---|---|---|---|---|---|
| DBM_transient | **0.998** | 0.990 | 0.997 | 0.981 | **70.547** |
| DBM_converged | 0.997 | 0.976 | 0.995 | 0.970 | 2.4544 |
| KPCA | **0.998** | 0.990 | 0.997 | 0.985 | 23.016 |
| Isomap | **0.998** | 0.990 | 0.997 | 0.985 | 12.484 |
| t-SNE | **0.998** | **0.996** | **0.998** | **0.988** | 6.1234 |
| UMAP | **0.998** | **0.996** | **0.998** | **0.988** | 15.796 |

Interestingly, our results verify the effectiveness of DBM_transient on C301 dataset similar to that on Bonn dataset. By DBM_transient, seizure and non-seizure samples are effectively separated into two clusters (Fig. 10). In contrast, by DBM_converged, KPCA, Isomap, t-SNE, and UMAP, non-seizure and seizure samples are widely distributed in the 2D representation space (Fig. 10), due to the diversity of non-seizure activities (Fig. 11). The representations by t-SNE and UMAP are quite similar. These methods separate seizure and non-seizure samples; however, both have a smaller between-class distance in comparison with DBM_transient. The result shows that DBM_transient achieves the best clustering performance with the highest $J(\omega)$, while the classification performance by a linear SVM subsequent to DBM_transient are closely comparable to the linear SVM following the other methods, evaluated by the measures: SPE, REC, ACC and F1 Score, as listed in Table V. The results demonstrate that DBM_transient is competent to cluster seizure and non-seizure events robustly on the small-size Bonn dataset as well as the large-size C301 dataset, with great potential in clinical applications.

It is still challenging to fully identify seizure events of the raw clinical data, e.g., C301 dataset. Some seizure samples are misjudged as non-seizure events by the DBM_transient and the other methods (marked as yellow in Fig. 10), eight of which are selected as examples shown in Fig. 11(C) and (D). For clarification, two correctly classified non-seizure and seizure signals are also presented in Fig. 11(A) and Fig. 11(B), respectively. In Fig. 11, the seizure sections of EEG signals are red-shaded by physicians. It can be observed that the EEG patterns presented in Fig. 11(C) and (D) are non-stereotyped and similar to each other. These samples complicate the task of classification, leading to misclassification.

## V. CONCLUSION

This work leverages the benefit of both the time-frequency domain features and the novel unsupervised learning approach DBM_transient to represent EEG signals in a 2D feature space. The proposed method DBM_transient effectively extracted the essential latent properties of EEG signals for visually clustering seizure and non-seizure events and allowed a linear readout for classification.

The spectral-temporal features of EEG signals by DWT are employed as input for training to reduce dimensions and representation in a 2D feature space for clustering and classification. We found that input features are distributed into a non-Gaussian shape with multiple peaks, indicating the intrinsic complex data structure. DBM trained to a transient state, namely DBM_transient, can capture these intrinsic features and represent them in a 2D feature space with seizure and non-seizure events separated into respective Gaussian-shaped clusters. Such powerful representation ability is beyond that of DBM_converged, KPCA, Isomap, t-SNE, and UMAP. The embedding methods such as Isomap, t-SNE and UMAP are the nonlinear dimensionality reduction methods that evaluate the similarity between two data points without the ability to capture the hidden data structure, which can be learned by training DBM to a transient state. The two evaluation metrics: Fisher's discriminant function $J(\omega)$ and classification measures of a subsequent linear SVM, demonstrated the superior clustering performance of DBM_transient.

DBM_transient worked very well on the small-size Bonn dataset with only 500 samples in total, despite an increase in task complexity from Case 1 to Case 7, beyond the abilities of the other methods employed here. Importantly, it is also effective on the large-size raw clinical C301 dataset, which is much more complicated and unbalanced with a 20 times larger size, and the EEG signals of which are stochastic, non-stationary and diverse with non-stereotyped EEG patterns. The clustering results in the 2D feature space (Fig. 10) are well separated. However, some samples are misclassified due to complicated and non-stereotyped EEG patterns (Figs. 11C and D). More importantly, the network structure and the training procedure of DBM_transient are the same for all 7 cases of Bonn dataset and C301 dataset. Therefore, the feature representation and visualization by DBM_transient can robustly and reliably cluster seizure and non-seizure events, adjustable to different datasets with various data sizes.

In conclusion, our results show that DBM_transient performed well in extracting the essential latent data structure and presenting them in a low-dimensional feature space with visually clear clustering, facilitating its future usage in clinical applications. In future work, DBM_transient should be evaluated with different feature vectors of EEG signals [17] or more generic feature set library, e.g. YASA [46], and further applied to the more complex and larger datasets, e.g. CHB-MIT dataset from Boston Children's Hospital [47], to investigate its power as well as limitations.

## REFERENCES

[1] E. Krook-Magnuson and I. Soltesz, "Beyond the hammer and the scalpel: Selective circuit control for the epilepsies," *Nature Neurosci.*, vol. 18, no. 3, pp. 331–338, Mar. 2015.

[2] L. Kuhlmann, K. Lehnertz, M. P. Richardson, B. Schelter, and H. P. Zaveri, "Seizure prediction—Ready for a new era," *Nature Rev. Neurol.*, vol. 14, no. 10, pp. 618–630, 2018.

[3] G. Zhu, Y. Li, P. P. Wen, and S. Wang, "Classifying epileptic EEG signals with delay permutation entropy and multi-scale K-means," in *Signal and Image Analysis for Biomedical and Life Sciences*. Cham, Switzerland: Springer, 2015, pp. 143–157.

[4] N. Decat et al., "Beyond traditional sleep scoring: Massive feature extraction and data-driven clustering of sleep time series," *Sleep Med.*, vol. 98, pp. 39–52, Oct. 2022.

[5] R. Gautam and M. Sharma, "Prevalence and diagnosis of neurological disorders using different deep learning techniques: A meta-analysis," *J. Med. Syst.*, vol. 44, no. 2, pp. 1–24, Feb. 2020.

[6] H. Guo, S. Huang, C. Huang, F. Shi, M. Zhang, and Z. Pan, "Binary file's visualization and entropy features analysis combined with multiple deep learning networks for malware classification," *Secur. Commun. Netw.*, vol. 2020, pp. 1–19, Dec. 2020.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[8] G. Manjunath, "Evolving network model that almost regenerates epileptic data," *Neural Comput.*, vol. 29, no. 4, pp. 937–967, Apr. 2017.

[9] Y. Bahri, J. Kadmon, J. Pennington, S. S. Schoenholz, J. Sohl-Dickstein, and S. Ganguli, "Statistical mechanics of deep learning," *Annu. Rev. Condens. Matter Phys.*, vol. 11, pp. 501–528, Mar. 2020.

[10] J. Van Der Donckt, J. Van Der Donckt, M. Rademaker, G. Vandewiele, and S. Van Hoecke, "Do not sleep on linear models: Simple and interpretable techniques outperform deep learning for sleep scoring," *Biomed. Signal Process. Control*, vol. 8, Jan. 2023, Art. no. 104429.

[11] O. A. Abdullah, M. I. Aal-Nouman, and A. K. AlJoudi, "Compliance framework for seizure detection via Gaussian deep Boltzmann machine using EEG data signal," in *Proc. IEEE Conf. Sustain. Utilization Develop. Eng. Technol. (CSUDET)*, Penang, Malaysia, Nov. 2019, pp. 1–5.

[12] H. Ocak, "Automatic detection of epileptic seizures in EEG using discrete wavelet transform and approximate entropy," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 2027–2036, Mar. 2009.

[13] K. Rasheed et al., "Machine learning for predicting epileptic seizures using EEG signals: A review," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 139–155, 2021.

[14] N. D. Truong et al., "Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram," *Neural Netw.*, vol. 105, pp. 104–111, Sep. 2018.

[15] A. Subasi, "EEG signal classification using wavelet feature extraction and a mixture of expert model," *Expert Syst. Appl.*, vol. 32, no. 4, pp. 1084–1093, May 2007.

[16] X. Yan, D. Yang, Z. Lin, and B. Vucetic, "Significant low-dimensional spectral-temporal features for seizure detection," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 668–677, 2022.

[17] B. D. Fulcher, M. A. Little, and N. S. Jones, "Highly comparative time-series analysis: The empirical structure of time series and their methods," *J. Roy. Soc. Interface*, vol. 10, no. 83, Jun. 2013, Art. no. 20130048.

[18] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[19] S. Khanmohammadi and C.-A. Chou, "Adaptive seizure onset detection framework using a hybrid PCA–CSP approach," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 1, pp. 154–160, Jan. 2018.

[20] J. Birjandtalab, M. B. Pouyan, and M. Nourani, "Unsupervised EEG analysis for automated epileptic seizure detection," in *Proc. 1st Int. Workshop Pattern Recognit.*, vol. 10011, X. Jiang, G. Chen, G. Capi, and C. Ishll, Eds. Bellingham, WA, USA: SPIE, 2016, p. 100110M.

[21] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[22] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2018, *arXiv:1802.03426*.

[23] F. Movahedi, J. L. Coyle, and E. Sejdić, "Deep belief networks for electroencephalography: A review of recent contributions and future outlooks," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 3, pp. 642–652, May 2018.

[24] L. Le, J. Hao, Y. Xie, and J. Priestley, "Deep kernel: Learning kernel function from data using deep neural network," in *Proc. IEEE/ACM 3rd Int. Conf. Big Data Comput. Appl. Technol. (BDCAT)*, Shanghai, China, Dec. 2016, pp. 1–7.

[25] S. You et al., "Unsupervised automatic seizure detection for focal-onset seizures recorded with behind-the-ear EEG using an anomaly-detecting generative adversarial network," *Comput. Methods Programs Biomed.*, vol. 193, Sep. 2020, Art. no. 105472.

[26] A. M. Al-kaysi, A. Al-Ani, and T. W. Boonstra, "A multichannel deep belief network for the classification of EEG data," in *Proc. Int. Conf. Neural Inf. Process.*, Istanbul, Turkey, 2015, pp. 38–45.

[27] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. E. Elger, "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 64, no. 6, 2001, Art. no. 061907.

[28] C. Bishop, *Pattern Recognition and Machine Learning* (Information Science and Statistics). Cham, Switzerland: Springer, 2006.

[29] M. Sharma, A. A. Bhurane, and U. R. Acharya, "MMSFL-OWFB: A novel class of orthogonal wavelet filters for epileptic seizure detection," *Knowl.-Based Syst.*, vol. 160, pp. 265–277, Nov. 2018.

[30] A. R. Hassan, A. Subasi, and Y. Zhang, "Epilepsy seizure detection using complete ensemble empirical mode decomposition with adaptive noise," *Knowl.-Based Syst.*, vol. 191, Mar. 2020, Art. no. 105333.

[31] M. Sazgar and M. G. Young, *Absolute Epilepsy and EEG Rotation Review: Essentials for Trainees*. Cham, Switzerland: Springer, 2019.

[32] F. C. Viola, J. Thorne, B. Edmonds, T. Schneider, T. Eichele, and S. Debener, "Semi-automatic identification of independent components representing EEG artifact," *Clin. Neurophysiol.*, vol. 120, no. 5, pp. 868–877, May 2009.

[33] F.-G. Tang, Y. Liu, Y. Li, and Z.-W. Peng, "A unified multi-level spectral–temporal feature learning framework for patient-specific seizure onset detection in EEG signals," *Knowl.-Based Syst.*, vol. 205, Oct. 2020, Art. no. 106152.

[34] H. Adeli, Z. Zhou, and N. Dadmehr, "Analysis of EEG records in an epileptic patient using wavelet transform," *J. Neurosci. Methods*, vol. 123, no. 1, pp. 69–87, Feb. 2003.

[35] A. Subasi and M. I. Gursoy, "EEG signal classification using PCA, ICA, LDA and support vector machines," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8659–8666, 2010.

[36] R. Salakhutdinov and G. Hinton, "Deep Boltzmann machines," in *Proc. 12th Int. Conf. Artif. Intell. Statist.* in Proceedings of Machine Learning Research, vol. 5, D. van Dyk and M. Welling, Eds. Clearwater Beach, FL, USA, Apr. 2009, pp. 448–455.

[37] L. T. Xuyen, L. T. Thanh, D. V. Viet, T. Q. Long, N. L. Trung, and N. D. Thuan, "Deep learning for epileptic spike detection," *VNU J. Sci., Comput. Sci. Commun. Eng.*, vol. 33, no. 2, pp. 1–13, Mar. 2018.

[38] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1771–1800, Aug. 2002.

[39] A. M. Sheri, A. Rafique, W. Pedrycz, and M. Jeon, "Contrastive divergence for memristor-based restricted Boltzmann machine," *Eng. Appl. Artif. Intell.*, vol. 37, pp. 336–342, Jan. 2015.

[40] G. E. Hinton, "A practical guide to training restricted Boltzmann machines," in *Neural Networks: Tricks of the Trade*. Cham, Switzerland: Springer, 2012, pp. 599–619.

[41] S. Nie, Z. Wang, and Q. Ji, "A generative restricted Boltzmann machine based method for high-dimensional motion data modeling," *Comput. Vis. Image Understand.*, vol. 136, pp. 14–22, Jul. 2015.

[42] A. S. Dhas et al., "An improved cad system for abnormal mammogram image classification using SVM with linear kernel," *Biomed. Res.*, vol. 28, no. 12, pp. 5499–5505, 2017.

[43] C. Keup, T. Kühn, D. Dahmen, and M. Helias, "Transient chaotic dimensionality expansion by recurrent networks," *Phys. Rev. X*, vol. 11, no. 2, Jun. 2021, Art. no. 021064.

[44] A. Decelle and C. Furtlehner, "Restricted Boltzmann machine: Recent advances and mean-field theory," *Chin. Phys. B*, vol. 30, no. 4, Apr. 2021, Art. no. 040202.

[45] L. Dabelow and M. Ueda, "Three learning stages and accuracy–efficiency tradeoff of restricted Boltzmann machines," *Nature Commun.*, vol. 13, no. 1, p. 5474, Sep. 2022.

[46] R. Vallat and M. P. Walker, "An open-source, high-performance tool for automated sleep staging," *eLife*, vol. 10, Oct. 2021, Art. no. e70092.

[47] A. H. Shoeb, "Application of machine learning to epileptic seizure onset detection and treatment," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 2009.