Biomedical Signal Processing and Control 92 (2024) 106020

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

SleepEGAN: A GAN-enhanced ensemble deep learning model for imbalanced classification of sleep stages

Xuewei Cheng [a,b], Ke Huang [c], Yi Zou [d], Shujie Ma [c,*]

[a] *MOE-LCSM, School of Mathematics and Statistics, Hunan Normal University, Changsha, China*
[b] *Key Laboratory of Applied Statistics and Data Science, College of Hunan Province, Hunan Normal University, Changsha, China*
[c] *Department of Statistics, University of California, Riverside, USA*
[d] *School of Mathematics and Statistics, Central South University, Changsha, China*

ARTICLE INFO

Keywords:
Class imbalance
EGAN
Ensemble learning
Individual heterogeneity
Sleep stage classification

ABSTRACT

Deep neural networks have played an important role in the automatic classification of sleep stages due to their strong representation and in-model feature transformation abilities. However, class imbalance and individual heterogeneity which typically exist in raw EEG signals of sleep data can significantly affect the classification performance of any machine learning algorithms. To solve these two problems, this paper develops a generative adversarial network (GAN)-powered ensemble deep learning model, named SleepEGAN, for the imbalanced classification of sleep stages. To alleviate class imbalance, we propose a new GAN (called EGAN) architecture adapted to the features of EEG signals for data augmentation. The generated samples for minority classes are used in the training process. In addition, we design a cost-free ensemble learning strategy to reduce the model estimation variance caused by the heterogeneity between the validation and test sets, to enhance the accuracy and robustness of prediction performance. We show that the proposed method improves classification accuracy compared to several existing state-of-the-art methods. The overall classification accuracy and macro F1-score obtained by our SleepEGAN method on three public sleep datasets are: Sleep-EDF-39: 86.8% and 81.9%; Sleep-EDF-153: 83.8% and 78.7%; SHHS: 88.0% and 82.1%.

1. Introduction

Sleep plays a vital role in mental and physical well-being throughout an individual's life [1,2]. According to research in [3–5] and the American Sleep Association, about 35.7% of people in the world and 50–70 million adults in the United States have a sleep disorder. The lack of sleep can cause negative cognitive, emotional, and physical effects [6]. In recent years, the classification of sleep stages has gained a lot of attention in the machine learning community [2,7], as it is crucial to understanding the quality and quantity of sleep and for diagnosing and treating various sleep disorders [8,9].

Sleep stage scoring is generally performed based on polysomnogram (PSG), which is considered the gold standard for evaluating human sleep [10]. PSG monitors many body functions during sleep, including brain activity (electroencephalogram, EEG), eye movements (electrooculogram, EOG), muscle activity (electromyogram, EMG), and heart rhythm (electrocardiogram, ECG). Single-channel EEG signals have been popularly used for sleep stage scoring because they are convenient and less expensive to monitor and collect [11]. Specifically, EEG recordings are typically segmented into epochs of 30 s, and each epoch is manually labeled by sleep specialists and then classified into one of five stages: Wake (W), Rapid eye movement (REM), and three non-REM stages (N1, N2, N3), following the AASM (American Academy of Sleep Medicine) guidance [12]. The task of the manual classification process is labor-intensive and prone to experts' subjective perception [13]. To this end, an automatic classification system for sleep stages can alleviate these problems and assist sleep specialists [14]. In recent years, the deep convolutional neural networks (CNNs) [15] together with recurrent neural networks (RNNs) [16,17], long short-term memory (LSTM) networks [18–22], or attention-based neural networks [23–27] have been successfully applied to sleep stage classification, as they can effectively learn frequency and time domain signals [10,28] from raw EEG epochs.

However, the *class imbalance* and *individual heterogeneity* of EEG signals, which are two common problems in sleep data, have not been well-addressed in the literature. To be specific, the sleep duration at all stages is not evenly distributed, resulting in significant differences in the sample sizes across sleep stages. Stage N2 generally occupies most of the sleep time (40.3%) and stage N1 only accounts for 6.3%

* Corresponding author.
E-mail addresses: xwcheng@hunnu.edu.cn (X. Cheng), khuan049@ucr.edu (K. Huang), zy6868@csu.edu.cn (Y. Zou), shujie.ma@ucr.edu (S. Ma).

https://doi.org/10.1016/j.bspc.2024.106020
Received 9 August 2023; Received in revised form 4 January 2024; Accepted 29 January 2024
Available online 8 February 2024

in the sleep-EDF-v1 dataset [29]. In general, the imbalance of data can seriously affect the classification performance [6]. Individual heterogeneity refers to the presence of certain differences in the EEG signals between individuals, rather than following the same distribution [30]. It is another challenge emerging from the raw data when they originate from different examining environments [2], channel layouts, recording setups [31], or emotional and physiological differences in patients [30]. As a result, we may not have a good generalization ability on the test set based on the model parameters selected by the validation set.

To solve the aforementioned problems, this paper develops a generative adversarial network (GAN)-enhanced ensemble deep learning model, named SleepEGAN, for the imbalanced classification of sleep stages. To alleviate class imbalance, we propose a new GAN (called EGAN) architecture adapted to the features of EEG signals for data augmentation. The generator and discriminator models in our GAN are motivated by, but different from, the deep neural networks called TinysleepNet proposed in [19]. TinysleepNet was originally used to classify sleep stages and was shown to have a great capability to extract features from raw EEG signals and learn their temporal transition rules using only a few convolutional layers and a single LSTM layer. We take advantage of its parsimonious model structure and design a modified model used for the generator and discriminator in our EGAN architecture, specially tailored for EEG signal augmentation. We show that our EGAN achieves a good balance between generalization and parsimony while having a great ability to learn the structure of EEG signals and generate high-quality samples for minority classes.

Our proposed EGAN model is shown to be an effective and efficient tool for generating EEG signals for small classes of sleep stages, such as stage N1 to match the number of samples in the large classes. It is worth noting that in the literature, a few works [32,33] directly employ the existing GAN methods to generate EEG signals, such as the naive GAN and the Wasserstein GAN [34] originally proposed for image generation, which may heavily rely on the convolutional layers and therefore possibly neglect the temporal and transitional features of EEG signals. Next, we design a new classification network structure called SleepEGAN to classify the sleep stages with the augmented data. The synthetic samples generated for the minority classes are used in the training process of classification. Our neural network model in the classification step has more filters than Tinysleepnet [19] to improve the network's representation ability of EEG signals and has fewer convolutional layers than VGG16 [35] for model parsimony, to achieve efficiency and computational convenience.

To tackle the problem of individual heterogeneity of EEG signals, we design a cost-free ensemble algorithm. Ensemble learning is a proven favorable and effective strategy to handle heterogeneous data [36]. It uses multiple diverse classifiers to achieve better generalization performance than a single learner to reduce prediction variance [37]. Throughout the training process, we retain the model parameters obtained in the epochs from the top 10 models chosen based on the prediction accuracy on the validation set instead of keeping only one set of model parameters having the best prediction. The stage prediction on the test set is based on the ensemble result of these 10 models, as the model parameters in different updated epochs are heterogeneous, and they can perform well in the validation set, satisfying two sufficient conditions for a nice ensemble: accurate and diverse [38]. It is worth noting that we only save the model parameters in each epoch during the training process, and then build an ensemble model using the retained parameters from the chosen models evaluated in the validation set. As a result, our ensemble learning procedure does not increase training costs, as it does not require training any additional models compared to the conventional deep neural network algorithms without ensemble learning.

The rest of the paper is organized as follows. Section 2 introduces the proposed GAN-enhanced ensemble deep learning model. The proposed method is illustrated on three real sleep datasets with the numerical results reported in Section 3. Concluding remarks are given in Section 4.

2. Propose method

In this section, we introduce the proposed GAN-based ensemble deep learning model (SleepEGAN) for the imbalanced classification of sleep stages. Our method contains three steps:

- we design a new GAN (EGAN) to generate samples for the minority classes so that the sample size of each class is balanced on the training set;
- we elaborately build a classification network architecture based on convolutional and LSTM layers;
- we develop an ensemble learning strategy without additional computational cost to reduce the variance of the model prediction caused by heterogeneity.

2.1. Data augmentation with EGAN

Signal and sequence augmentation is considered as a primary technique to synthesize new training data from the original data for each training epoch [19] to alleviate the class imbalance problem. We first use this technique to synthesize new signal patterns for each training epoch by signal augmentation and generate new batches of multiple sequences of EEG epochs in the mini-batch gradient descent by sequence augmentation. In addition, the weighted cross-entropy loss function is also introduced to mitigate the class imbalanced problem by setting the weight for the N1 stage to 1.5 and others to 1. However, these strategies of data augmentation cannot completely solve the imbalanced problem, and the learned deep model still prioritizes the majority class.

To solve the above problem, we propose a new generative adversarial network (EGAN) to learn the probability distribution for generating raw EEG epochs, as an advanced strategy for data augmentation. The generator of GAN can generate more samples from the estimated probability distribution. Different from parametric and nonparametric density estimators, where the density function is explicitly defined in a relatively low dimension, GAN can be viewed as an implicit density estimator in a much higher dimension [39].

According to the invariance structure of data, data augmentation by GAN implicitly enlarges the training dataset by sampling original data and generating new data, which usually regularizes the model effectively [40]. However, the naive GAN and other extensions [34,41], which are originally designed for image generation, may not work well for EEG signals as they do not consider the temporal and transitional features of EEG data. Therefore, we design a new GAN architecture (EGAN) tailored for EEG signal generation. The proposed EGAN has two features: it can extract the representative temporal components from the high-dimensional features, and then automatically learn the transition rules of the EEG signals.

For the generator and discriminator models of our proposed EGAN, we design a 4+1 learning framework, where the first four convolutional layers are used to extract frequency signals, followed by an LSTM layer to learn temporal information. The main structure of EGAN is presented in Fig. 1. The network structure was motivated by Tinysleepnet which was originally proposed for the classification purpose instead of sample generation. We propose a modified architecture used in our generator and discriminator models to make our EGAN possible to achieve the high-quality task of sample generation. Specifically, we make the following contributions:

- for the generator, we add an Upsample layer to expand the 100-dimensional (or 125-dimensional) noise to 3000 (or 3750) dimensions as inputs;
- for the generator, after the LSTM layer processing, we add a fully connected layer so that it transforms the learned features into 3000-dimensional (or 3750-dimensional) EEG epochs with Tanh nonlinear activation functions;

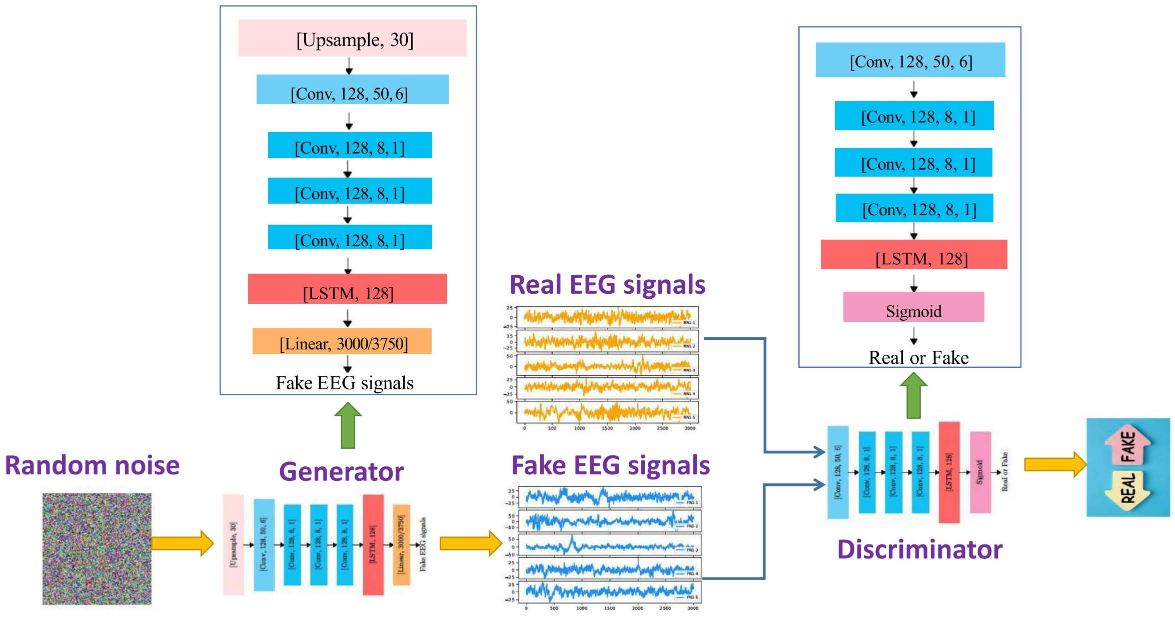

Fig. 1. The main structure of EGAN. For simplicity, the pooling and dropout layers are ignored here.

- for the discriminator, it only needs to discriminate between true and false, not its specific sleep stage, so we modify the activation function of the output layer from Softmax to Sigmoid;
- we change the activation function of ReLU in Tinysleepnet to leaky ReLU in EGAN to make sure the gradient can flow through the entire architecture.

The proposed EGAN model plays a vital role in balancing training samples among different classes, resulting in a superior prediction performance on the N1 stage in sleep data.

2.2. Classification with SleepEGAN

With the augmented data obtained from the previous step, next we design a new classification network SleepEGAN to process a sequence of single-channel EEG epochs and perform the classification of sleep stages (see Fig. 2). Our SleepEGAN uses a 1+2+2 convolutional network to extract frequency features, followed by an LSTM layer to extract time domain features, and it achieves a good balance between model parsimony and efficiency for classification. Our SleepEGAN is motivated by but different from VGG16 [35] and Tinysleepnet [19] such that it has more filters than Tinysleepnet to enhance the network's representation ability of EEG signals and less network layers than VGG16 for model parsimony.

To be specific, we segment the EEG signals into n epochs $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ of E_s seconds, where $\mathbf{x}_i \in \mathbb{R}^{E_s \times F_s}$ and F_s is the sample rating for each second EEG. We obtain the predicted sleep stage $\hat{y}_i$ in the test set using the epoch of $\mathbf{x}_i$ with the network parameters trained using the training data set, where $\hat{y}_i \in \{0, 1, 2, 3, 4\}$ corresponds to the five sleep stages W, N1, N2, N3 and REM, respectively.

The CNNs block CNN_{θ_r} that consists of five convolutional layers (1+2+2), interleaved with three max-pooling and two dropout layers, is firstly employed to learn time-invariant features from single-channel signals $\mathbf{x}_i$. Then, the LSTM layer LSTM_{θ_s} followed by a dropout layer is used to extract time-dependent information from processed features $\widetilde{\mathbf{x}}_i$ by CNNs block. The final out y_i is activated by Softmax function $\sigma(\cdot)$ with parameter vector $\mathbf{v}$. In the procedure (1), θ_r and θ_s are the learnable parameters of the CNNs and LSTM, respectively, where $\mathbf{h}_i$ and $\mathbf{c}_i$ are output vectors of hidden and cell states of the LSTM layer after processing the features $\widetilde{\mathbf{x}}_i$.

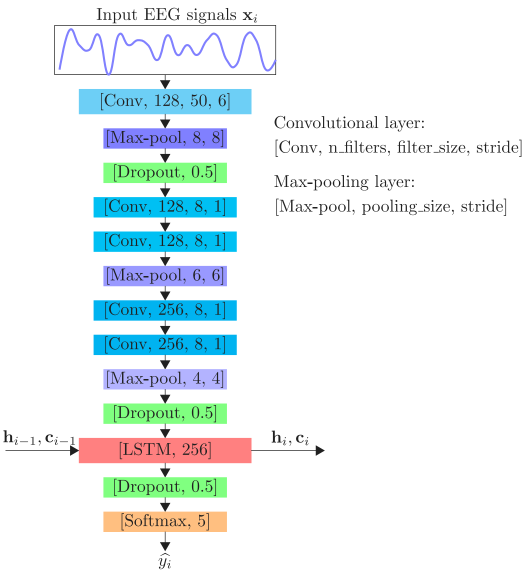

Fig. 2. An overview classification network architecture of SleepEGAN. Each rectangular box represents one layer in the model, and the arrows indicate the flow of data from raw single-channel EEG epochs $\mathbf{x}_i$ to sleep stages $\hat{y}_i$.

$$\begin{aligned} \widetilde{\mathbf{x}}_i &= \mathrm{CNN}_{\theta_r}(\mathbf{x}_i), \\ \mathbf{h}_i, \mathbf{c}_i &= \mathrm{LSTM}_{\theta_s}(\mathbf{h}_{i-1}, \mathbf{c}_{i-1}, \widetilde{\mathbf{x}}_i), \\ y_i &= \sigma(\mathbf{v}\mathbf{h}_i). \end{aligned} \tag{1}$$

We develop the SleepEGAN architecture to achieve a good balance between generalization and parsimony while preserving its ability to learn the structure of EEG signals. The strategies of bidirectional LSTM in Deepsleepnet [18], multi-head attention in AttnSleep [23], dual-stream structure [42] in SalientSleepNet and multi-scale extraction in [6] may enhance the representative ability of deep learning, but huge computational resources may incur. Our contribution is not to develop a deeper and more complicated neural network model with superior

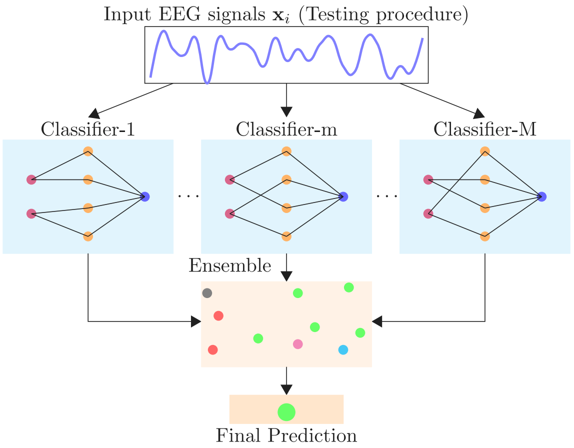

Fig. 3. Ensemble illustration of predicted stage class. Different marks in the ensemble rectangle imply different predicted stage classes. The final stage prediction is obtained by classifier voting.

generalization ability. Instead, we would like to use limited resources to tackle the problems of individual heterogeneity and class imbalance.

2.3. Ensemble learning

For the classification of sleep stages, typically researchers use EEG signals from some individuals as the test set. In addition, they split the individuals in the training data set into the training and validation sets. In this way, if there exists individual heterogeneity in the training and test sets, the model parameters selected by the validation set may not perform well in the test set. This heterogeneity can be inherent to the raw EEG data. To solve this problem, we introduce a cost-free ensemble learning strategy to improve the accuracy and stability of the prediction performance.

Ensemble learning is a machine learning paradigm in which multiple learners are trained and combined for a specific task, achieving better generalization performance than single learners [37]. However, the computational cost of ensemble learning can be much higher than that of a single classifier, especially when the ensemble is performed on deep neural networks. To address this issue, we develop a cost-free algorithm for ensemble learning. Specifically, we record the validation accuracy and F1-score for each training epoch, and we select the model parameters of the top M classifiers ranked by their prediction performance evaluated in the test set. Each model makes a separate prediction for each sample, and eventually, all classifiers take a majority vote on the final prediction class (see Fig. 3). By this method, we reduce the prediction variance and improve the stability of the prediction performance without paying any additional computational cost.

3. Experimental results

In this section, we first introduce the sleep datasets and experimental settings. Then, we show the classification performance of our proposed SleepEGAN, and compare it with several existing classification methods.

3.1. EEG datasets

We evaluate our method using three popular and public real-life sleep datasets; namely, Sleep-EDF-39, Sleep-EDF-153, and SHHS (Sleep Heart Health Study) as shown in Table 1.

The dataset of Sleep-EDF [29] has two versions. One is Sleep-EDF-39 published in 2013 before data augmentation, in which there were 39 PSG recordings from the study of age effects in healthy subjects (SC), collected from 20 subjects. The other one is Sleep-EDF-153, which expands the number of recordings from 39 to 153, including 78 subjects aged between 25–101 years (37 males and 41 females). These recordings are segmented into 30 s epochs and manually labeled by sleep experts in light of R & K (Rechtschaffen and Kales) manual [43]. We evaluated our model using the Fpz-Cz EEG channel provided in these PSG recordings with a sampling rate of 100 Hz.

We also use a larger sleep dataset named SHHS [44,45] which is a multicenter cohort study on cardiovascular and other sleep-disordered breathing diseases to evaluate the performance of the proposed method. The subjects of this dataset suffer from a wide range of diseases, such as lung diseases, cardiovascular diseases, and coronary diseases. Following the study of [23,46], we select 329 subjects with regular sleep from 6441 subjects for our experiments to reduce the effect of other diseases. In addition, we select the C4-A1 channel with a sampling rate of 125 Hz.

3.2. Experiment settings

The 20-fold cross-validation (CV) scheme is used to evaluate the prediction performance on the three datasets. In each fold, we further allocate 10% of the training set to a validation set to evaluate the training model in case of overfitting. The models that achieve the top M overall accuracies are kept for evaluation with the test set. We use Adam optimizer with 200 epochs to train the classification model, where the learning rate, Adam's beta1, and beta2 are 10^{-4}, 0.9 and 0.999, respectively. The mini-batch size is set as 8, 32, and 128 for Sleep-EDF-39, Sleep-EDF-153, and SHHS, respectively. The sequence length is 20. The number M of learners is 10, which works well in the trade-off between diversity and accuracy.

For the generative networks (EGAN), we also use the Adam optimizer to train the generator and discriminator, where the learning rate, Adam's beta1, and beta2 are 2×10^{-4}, 0.5 and 0.999, respectively. For the generated tasks for Sleep-EDF-39 and SHHS, we set up 660 and 843 training epochs. The batch sizes are 16 and 64, respectively. The input to the generator is 100-dimensional (or 125-dimensional) noise for Sleep-EDF-39 (or SHHS).

We use three metrics to evaluate the performance of our proposed method, namely, overall accuracy (ACC), macro-averaged F1-score (MF1), and Cohen's Kappa coefficient (κ). The second one is a common metric to evaluate the performance of imbalanced datasets, and the last one is used to assess the consistency of the prediction results.

We implement the EGAN method to generate EEG signals using Pytorch 2.1.0, and perform sleep stage classification using Tensorflow 1.13.1. Pytorch and Tensorflow are two popular Python libraries for high-performance computation of deep learning models. We train our model using the GPU nodes with 64 GB memory at the High-Performance Computing Center of UC-Riverside. Our source code is publicly available at https://github.com/ChengXuewei/SleepEGAN.

3.3. Generation of minority class samples by EGAN

We generate EEG signals for the minority classes using our EGAN method for two sleep datasets Sleep-EDF-39 and SHHS, respectively, and the results of the generated samples are shown in Table 1. Specifically, for Sleep-EDF-39 and SHHS, N1 is the minority class, so we generate samples in the N1 stage to balance the proportion of each class. For Sleep-EDF-153, the sample size is balanced across sleep stages, so we dropped out of the generation procedure. We generate samples for the smallest class to match the sample size of the penultimate class.

Fig. 4 shows the EEG epochs in the five sleep stages for the dataset Sleep-EDF-39. Clearly, we can see that the EEG signals follow different patterns in the five sleep stages. Moreover, Figs. 5 and 6 show the real and the generated fake EEG signals in the N1 stage, respectively. We see

Table 1
The details of three datasets before and after data augmentation.

Datasets	Subjects	Channel	Sampling rate	Type	W	N1	N2	N3	REM	Epochs
Sleep-EDF-39	20	Fpz-Cz	100 Hz	Before	10 197 23.1%	2804 6.3%	17 799 40.3%	5703 12.9%	7717 17.5%	44 220
				After	10 197 20.6%	**8120** 16.4%	17 799 35.9%	5703 11.5%	7717 15.6%	49 536
Sleep-EDF-153	78	Fpz-Cz	100 Hz	Before	69 824 35.0%	21 522 10.8%	69 132 34.7%	13 039 6.5%	25 835 13.0%	199 352
SHHS	329	C4-A1	125 Hz	Before	46 319 14.3%	10 304 3.2%	142 125 43.8%	60 153 18.5%	65 953 20.3%	324 854
				After	46 319 12.8%	**46 272** 12.8%	142 125 39.4%	60 153 16.7%	65 953 18.3%	360 822

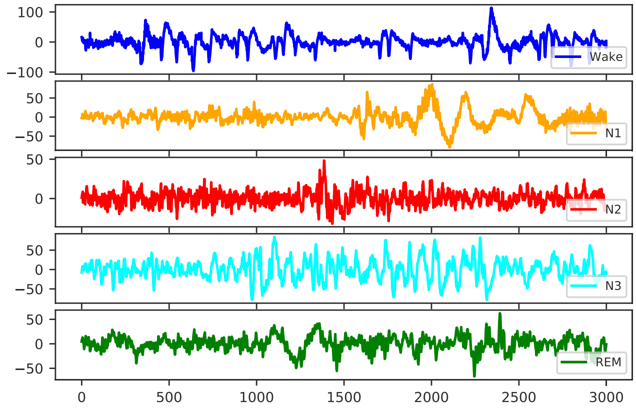

Fig. 4. The real EEG signals in the five stages of Sleep-EDF-39.

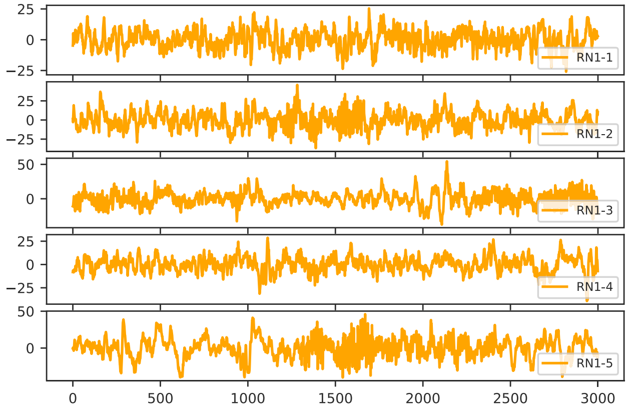

Fig. 5. The real EEG signals in the N1 stage of Sleep-EDF-39.

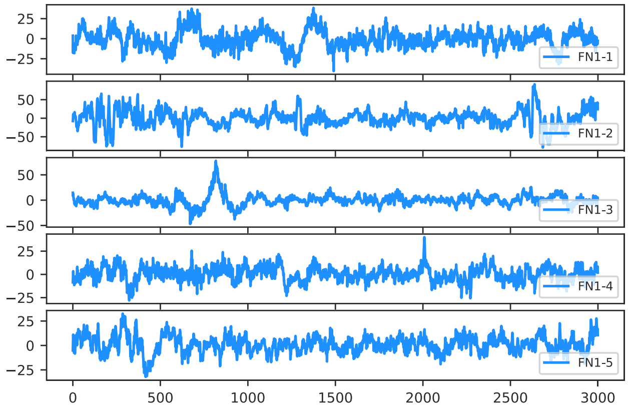

Fig. 6. The generated fake EEG signals in the N1 stage of Sleep-EDF-39.

that the generated samples are generally quite similar to the real ones. The successful training of EGAN can make a good balance of different classes of sleep data for classification to improve prediction accuracy. It also indicates that EGAN has the ability to learn the distribution of temporal data so that it may have potential applications in signal denoising and detection.

3.4. Comparison of classification performance by different methods

We evaluate the prediction performance of our SleepEGAN model against several state-of-the-art approaches. The comparison results among different methods are shown in Table 2. We observed that our sleepEGAN reasonably outperforms the other models in all three real datasets, thanks to the assistance of EGAN and ensemble learning.

Specifically, the more imbalanced the dataset is, the better the performance of our method is after data augmentation. For example, for the dataset SHHS, after applying our EGAN and the ensemble learning, the size of N1 epochs increases from 10,304 to 46,272, and the F1-Score for the N1 class is improved from 40.5% to 54.1% by $13.6\%/40.5\% = 33.6\%$ compared to the second best method. The overall accuracy also improved to 88.0%. In conclusion, the proposed SleepEGAN method has a promising performance for sleep stage prediction and is expected to work well for sleep data with imbalanced classes and individual heterogeneity.

3.5. Ablation study

Our method SleepEGAN contains two strategies to tackle the problems of class imbalance and individual heterogeneity. To analyze the effectiveness of each strategy in our SleepEGAN, we provide an ablation study based on Sleep-EDF-39 as shown in Table 3. To be specific, we develop four model variants as follows.

- Naive: only use main network structure to perform training process for classification.
- Naive + EGAN: only use EGAN to generate naturalistic EEG epochs.
- Naive + Ensemble: only use the ensemble strategy to enhance the prediction performance.
- SleepEGAN: use both strategies to train EEG samples.

Table 3 shows that the prediction performances of SleepEGAN without using EGAN and/or ensemble need to be further improved, especially for the F1-Score of N1. The Naive method has employed the weighted cross-entropy loss function as well as data augmentation without using EGAN. These simple strategies cannot significantly improve the prediction performance for the N1 stage. Then, we use EGAN to generate fake EEG samples in N1, but the performance is still inferior to that of SleepEGAN. Although the distribution of the training data is balanced in this scenario, the optimal model parameter selected by the validation set may not have good generalization ability in the test set due to individual heterogeneity. Therefore, we add a cost-free ensemble learning step, resulting in SleepEGAN, which improves not only the prediction accuracy of N1 but also the overall accuracy.

Table 2
Comparison results between our method and other methods. The best performance on each dataset is highlighted in bold.

Datasets	Methods	Epochs	Overall metrics			Per-class F1-Score (F1)				
			Acc	MF1	κ	W	N1	N2	N3	REM
Sleep-EDF-39	DeepSleepNet [18]	41 950	82.0	76.9	0.76	84.7	46.6	85.9	84.8	82.4
	IITNeT [47]	42 308	84.0	77.0	0.78	87.9	44.7	88.0	85.7	82.1
	SleepEEGNet [16]	42 308	84.3	79.7	0.79	89.2	52.2	86.8	85.1	85.0
	ResnetLSTM [48]	42 308	82.5	73.7	0.76	86.5	28.4	87.7	89.8	76.2
	MultitaskCNN [49]	42 308	83.1	75.0	0.77	87.9	33.5	87.5	85.8	80.3
	AttnSleep [23]	42 308	84.4	78.1	0.79	89.7	42.6	88.8	**90.2**	79.0
	Tinysleepnet [19]	44 220	85.4	80.5	0.80	90.1	51.4	88.5	88.3	84.3
	Our method	44 220	**86.8**	**81.9**	**0.82**	**91.7**	**53.6**	**89.2**	89.1	**86.1**
Sleep-EDF-153	DeepSleepNet [18]	195 479	77.8	71.8	0.70	90.9	45.0	79.2	72.7	71.1
	SleepEEGNet [16]	195 479	74.2	69.6	0.66	89.8	42.1	75.2	70.4	70.6
	ResnetLSTM [48]	195 479	78.9	71.4	0.71	90.7	34.7	83.6	80.9	67.0
	MultitaskCNN [49]	195 479	79.6	72.8	0.72	90.9	39.7	83.2	76.6	73.5
	AttnSleep [23]	195 479	81.3	75.1	0.74	92.0	42.0	85.0	**82.1**	74.2
	Tinysleepnet [19]	199 352	83.1	78.1	0.77	92.8	51.0	85.3	81.1	80.3
	Our method	199 352	**83.8**	**78.7**	**0.82**	**93.1**	**51.7**	**85.8**	81.2	**82.0**
SHHS	DeepSleepNet [18]	324 854	81.0	73.9	0.73	85.4	40.5	82.5	79.3	81.9
	SleepEEGNet [16]	324 854	73.9	68.4	0.65	81.3	34.4	73.4	75.9	77.0
	ResnetLSTM [48]	324 854	83.3	69.4	0.76	85.1	9.4	86.3	87.0	79.1
	MultitaskCNN [49]	324 854	81.4	71.2	0.74	82.2	25.7	83.9	83.3	81.1
	AttnSleep [23]	324 854	84.2	75.3	0.78	86.7	33.2	87.1	87.1	82.1
	Our method	324 854	**88.0**	**82.1**	**0.83**	**89.6**	**54.1**	**89.2**	**87.1**	**90.6**

Table 3
Ablation study conducted on Sleep-EDF-39 dataset.

Methods	Overall metrics			Per-class F1-Score (F1)				
	Acc	MF1	κ	W	N1	N2	N3	REM
Naive	85.1	79.8	0.80	90.1	49.4	87.8	87.5	84.2
Naive + EGAN	85.6	79.8	0.80	91.5	46.6	88.6	88.8	83.7
Naive + Ensemble	86.0	81.1	0.81	90.8	53.0	88.4	**89.2**	84.1
SleepEGAN	**86.8**	**81.9**	**0.82**	**91.7**	**53.6**	**89.2**	89.1	**86.1**

Table 4
Sensitivity analysis conducted on Sleep-EDF-39 dataset.

# Classifiers	Overall metrics			Per-class F1-Score (F1)				
	Acc	MF1	κ	W	N1	N2	N3	REM
M = 5	86.5	81.5	0.82	91.4	51.9	89.1	89.4	85.6
M = 6	86.7	81.7	0.82	91.6	52.8	89.1	89.3	85.6
M = 7	86.9	81.9	0.82	91.6	53.0	89.4	89.5	86.2
M = 8	86.7	81.8	0.82	91.6	53.0	89.2	89.3	85.8
M = 9	86.9	81.9	0.82	91.7	53.2	89.3	89.2	86.3
M = 10	86.8	81.9	0.82	91.7	53.6	89.2	89.1	86.1

3.6. *Sensitivity analysis for the number of classifiers in the ensemble learning*

The number of base learners M is a hyper-parameter in the ensemble procedure and needs to be specified beforehand. We expect this hyper-parameter to be overly insensitive with respect to prediction performance. Therefore, we choose the dataset Sleep-EDF-39 for parameter sensitivity experiments. We fix the other parameters and vary $M \in \{5, 6, 7, 8, 9, 10\}$ to investigate the fluctuation of its prediction result shown in Table 4.

We observe that the parameter M hardly affects the prediction performance of our model. For all values of M, the overall accuracy does not vary by more than 0.4%. Thus, our model is very robust to the hyper-parameter M, and the user can choose it arbitrarily in light of the experiment's purpose.

4. Conclusion

We propose a new GAN-enhanced ensemble deep learning model, called SleepEGAN, for sleep stage classification with imbalanced classes and individual heterogeneity from raw single-channel EEG signals. The proposed SleepEGAN outperforms several existing deep models for sleep stage classification on three popular sleep datasets. The success of SleepEGAN is mainly attributed to two aspects: first, we employ EGAN to generate fake EEG samples for the minority class so that the data become balanced during the training process; second, we develop a cost-free ensemble algorithm to reduce the estimation variance caused by individual heterogeneity, and hence it enhances the robustness of our model. Through ablation experiments, we find that these two strategies are effective to improve classification performance. Finally, we perform a sensitivity analysis on the number of base learners in the procedure of ensemble learning and show that our model works reasonably well using an arbitrary hyperparameter in a given range.

In addition, it is noteworthy that the fake EEG signals generated by our EGAN are quite similar to the real EEG signals. The EGAN method can successfully learn the temporal and transitional structure of the EEG signals, and it has potential applications in signal recognition [50], signal processing [51,52], signal synthesis [53], among others. These can be interesting future research topics to explore.

CRediT authorship contribution statement

Xuewei Cheng: Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Ke Huang:** Data curation, Software, Validation, Visualization. **Yi Zou:** Software. **Shujie Ma:** Conceptualization, Investigation, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We thank the editor and the referees for their helpful and constructive comments and suggestions that have greatly helped us improve the article. Ma and Huang' s research is supported in part by the U.S. NSF grants DMS-17-12558, DMS-20-14221, and DMS-23-10288. Cheng's research is supported by the Research Foundation of Education Department of Hunan Province, China (No. 23B0056), the National Science Foundation of China (No. 52101027) and the Innovation Team Project for Universities in Guangdong Province (No. 2020WCXTD008).

References

[1] E. Estrada, P. Nava, H. Nazeran, K. Behbehani, J. Burk, E. Lucas, Itakura distance: A useful similarity measure between EEG and eog signals in computer-aided classification of sleep stages, in: 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, IEEE, 2006, pp. 1189–1192.

[2] K.A.I. Aboalayon, M. Faezipour, W.S. Almuhammadi, S. Moslehpour, Sleep stage classification using EEG signal analysis: A comprehensive survey and new investigation, Entropy 18 (9) (2016) 1–31.

[3] H. Jahrami, A.S. BaHammam, N.L. Bragazzi, Z. Saif, M. Faris, M.V. Vitiello, Sleep problems during the COVID-19 pandemic by population: a systematic review and meta-analysis, J. Clin. Sleep Med. 17 (2) (2021) 299–313.

[4] E. Estrada, H. Nazeran, F. Ebrahimi, M. Mikaeili, EEG signal features for computer-aided sleep stage detection, in: 2009 4th International IEEE/EMBS Conference on Neural Engineering, IEEE, 2009, pp. 669–672.

[5] E. Estrada, H. Nazeran, EEG and HRV signal features for automatic sleep staging and apnea detection, in: 2010 20th International Conference on Electronics Communications and Computers, CONIELECOMP, IEEE, 2010, pp. 142–147.

[6] Z. Liu, S. Luo, Y. Lu, Y. Zhang, L. Jiang, H. Xiao, Extracting multi-scale and salient features by MSE based U-structure and CBAM for sleep staging, IEEE Trans. Neural Syst. Rehabil. Eng. 31 (2022) 31–38.

[7] R. Boostani, F. Karimzadeh, M. Nami, A comparative review on sleep stage classification methods in patients and healthy individuals, Comput. Methods Programs Biomed. 140 (C) (2017) 77–91.

[8] X. Huang, K. Shirahama, F. Li, M. Grzegorzek, Sleep stage classification for child patients using deconvolutional neural network, Artif. Intell. Med. 110 (1) (2020) 101981.

[9] A. Supratak, P. Haddawy, Quantifying the impact of data characteristics on the transferability of sleep stage scoring models, Artif. Intell. Med. 139 (2023) 102540.

[10] L. Fraiwan, K. Lweesy, N. Khasawneh, H. Wenz, H. Dickhaus, Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier, Comput. Methods Programs Biomed. 108 (1) (2012) 10–19.

[11] S.N. Kundel V, Impact of portable sleep testing, Sleep Med. Clin. 12 (1) (2017) 137–147.

[12] R.B. Berry, R. Brooks, C.E. Gamaldo, S.M. Harding, C. Marcus, B.V. Vaughn, et al., The AASM manual for the scoring of sleep and associated events, Rules Terminol. Tech. Specif. Darien Illinois Am. Acad. Sleep Med. 176 (2012) 1–7.

[13] M. Ronzhina, O. Janoušek, J. Kolářová, M. Nováková, P. Honzík, I. Provazník, Sleep scoring using artificial neural networks, Sleep Med. Rev. 16 (3) (2012) 251–263.

[14] Y. You, X. Zhong, G. Liu, Z. Yang, Automatic sleep stage classification: A light and efficient deep neural network model based on time, frequency and fractional Fourier transform domain features, Artif. Intell. Med. 127 (5) (2022) 102279.

[15] L. Fiorillo, P. Favaro, F.D. Faraci, Deepsleepnet-lite: A simplified automatic sleep stage scoring model with uncertainty estimates, IEEE Trans. Neural Syst. Rehabil. Eng. 29 (2021) 2076–2085.

[16] S. Mousavi, F. Afghah, U.R. Acharya, SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach, PLoS One 14 (5) (2019) 1–15.

[17] C. Zhao, J. Li, Y. Guo, SleepContextNet: A temporal context network for automatic sleep staging based single-channel EEG, Comput. Methods Programs Biomed. 220 (2022) 106806.

[18] A. Supratak, H. Dong, C. Wu, Y. Guo, DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG, IEEE Trans. Neural Syst. Rehabil. Eng. 25 (11) (2017) 1998–2008.

[19] A. Supratak, Y. Guo, TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC, IEEE, 2020, pp. 641–644.

[20] X. Ji, Y. Li, P. Wen, 3DSleepNet: A multi-channel bio-signal based sleep stages classification method using deep learning, IEEE Trans. Neural Syst. Rehabil. Eng. (2023).

[21] R. Li, B. Wang, T. Zhang, T. Sugi, A developed LSTM-ladder-network-based model for sleep stage classification, IEEE Trans. Neural Syst. Rehabil. Eng. 31 (2023) 1418–1428.

[22] H. Phan, K.P. Lorenzen, E. Heremans, O.Y. Chén, M.C. Tran, P. Koch, et al., L-SeqSleepNet: whole-cycle long sequence modelling for automatic sleep staging, IEEE J. Biomed. Health Inf. 27 (10) (2023) 4748–4757.

[23] E. Eldele, Z. Chen, C. Liu, M. Wu, C.-K. Kwoh, X. Li, C. Guan, An attention-based deep learning approach for sleep stage classification with single-channel EEG, IEEE Trans. Neural Syst. Rehabil. Eng. 29 (2021) 809–818.

[24] J. Huang, L. Ren, X. Zhou, K. Yan, An improved neural network based on senet for sleep stage classification, IEEE J. Biomed. Health Inf. 26 (10) (2022) 4948–4956.

[25] J. Phyo, W. Ko, E. Jeon, H.-I. Suk, TransSleep: Transitioning-aware attention-based deep neural network for sleep staging, IEEE Trans. Cybern. (2022) 1–11.

[26] Y. Wei, Y. Zhu, Y. Zhou, X. Yu, Y. Luo, Automatic sleep staging based on contextual scalograms and attention convolution neural network using single-channel EEG, IEEE J. Biomed. Health Inf. (2023).

[27] C. Zhao, J. Li, Y. Guo, Sequence signal reconstruction based multi-task deep learning for sleep staging on single-channel EEG, Biomed. Signal Process. Control 88 (2024) 105615.

[28] A.R. Hassan, M.I.H. Bhuiyan, A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features, J. Neurosci. Methods 271 (2016) 107–118.

[29] A.L. Goldberger, L.A. 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: components of a new research resource for complex physiologic signals, Circulation 101 (23) (2000) e215–e220.

[30] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T.H. Falk, J. Faubert, Deep learning-based electroencephalography analysis: a systematic review, J. Neural Eng. 16 (5) (2019) 051001.

[31] H. Phan, K. Mikkelsen, Automatic sleep staging of EEG signals: recent development, challenges, and future directions, Physiol. Meas. (2022).

[32] D. Zhou, Q. Xu, J. Wang, H. Xu, L. Kettunen, Z. Chang, F. Cong, Alleviating class imbalance problem in automatic sleep stage classification, IEEE Trans. Instrum. Meas. 71 (2022) 1–12.

[33] K.G. Hartmann, R.T. Schirrmeister, T. Ball, EEG-GAN: Generative adversarial networks for electroencephalograhic (EEG) brain signals, 2018, arXiv preprint arXiv:1806.01875.

[34] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 214–223.

[35] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[36] R. Jin, H. Liu, SWITCH: A novel approach to ensemble learning for heterogeneous data, in: Machine Learning: ECML 2004: 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004. Proceedings 15, Springer, 2004, pp. 560–562.

[37] Z.-H. Zhou, J. Feng, Deep forest: towards an alternative to deep neural networks, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017, pp. 3553–3559.

[38] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms, CRC Press, 2012.

[39] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (11) (2020) 139–144.

[40] J. Fan, R. Li, C.-H. Zhang, H. Zou, Statistical Foundations of Data Science, CRC Press, 2020.

[41] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, 2015, arXiv preprint arXiv:1511.06434.

[42] Z. Jia, Y. Lin, J. Wang, X. Wang, P. Xie, Y. Zhang, Salientsleepnet: Multimodal salient wave detection network for sleep staging, 2021, arXiv preprint arXiv:2105.13864.

[43] A. Rechtschaffen, A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects, Public Health Serv. (1968).

[44] S.F. Quan, B.V. Howard, C. Iber, J.P. Kiley, F.J. Nieto, G.T. O'Connor, D.M. Rapoport, S. Redline, J. Robbins, J.M. Samet, et al., The sleep heart health study: design, rationale, and methods, Sleep 20 (12) (1997) 1077–1085.

[45] G.-Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, S. Redline, The national sleep research resource: towards a sleep data commons, J. Am. Med. Inform. Assoc. 25 (10) (2018) 1351–1358.

[46] P. Fonseca, N. Den Teuling, X. Long, R.M. Aarts, Cardiorespiratory sleep stage detection using conditional random fields, IEEE J. Biomed. Health Inf. 21 (4) (2016) 956–966.

[47] H. Seo, S. Back, S. Lee, D. Park, T. Kim, K. Lee, Intra-and inter-epoch temporal context network (iitnet) using sub-epoch features for automatic sleep scoring on raw single-channel EEG, Biomed. Signal Process. Control 61 (2020) 102037.

[48] Y. Sun, B. Wang, J. Jin, X. Wang, Deep convolutional network method for automatic sleep stage classification based on neurophysiological signals, in: 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), IEEE, 2018, pp. 1–5.

[49] H. Phan, F. Andreotti, N. Cooray, O.Y. Chén, M. De Vos, Joint classification and prediction CNN framework for automatic sleep stage classification, IEEE Trans. Biomed. Eng. 66 (5) (2018) 1285–1296.

[50] J. Stallkamp, M. Schlipsing, J. Salmen, C. Igel, Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition, Neural Netw. 32 (2012) 323–332.

[51] M.H. Hayes, Statistical Digital Signal Processing and Modeling, John Wiley & Sons, 1996.

[52] S. Sanei, J.A. Chambers, EEG Signal Processing, John Wiley & Sons, 2013.

[53] H. Smith, Phytochromes and light signal perception by plants—an emerging synthesis, Nature 407 (6804) (2000) 585–591.