Alleviating Class Imbalance Problem in Automatic Sleep Stage Classification

Dongdong Zhou, Qi Xu*, Jian Wang, Hongming Xu, Lauri Kettunen, Zheng Chang, *Senior Member, IEEE*, and Fengyu Cong, *Senior Member, IEEE*

***Abstract*—For real-world automatic sleep stage classification tasks, various existing deep learning based models are biased towards the majority with high proportion. Because of the unique sleep structure, most of the current polysomnography datasets suffer an inherent class imbalance problem (CIP), in which the number of each sleep stage is severely unequal. In this study, we first define the class imbalance factor (CIF) to describe the level of CIP quantitatively. Afterwards, we propose two balancing methods to alleviate this problem from the dataset quantity and the relationship between the class distribution and the applied model respectively. The first one is to employ the data augmentation (DA) with the generative adversarial network (GAN) model and different intensities Gaussian white noise to balance samples, thereinto, Gaussian white noise addition is specifically tailored to deep learning based models, which can work on raw electroencephalogram (EEG) data while preserving their properties. In addition, we try to balance the relationship between the imbalanced class and biased network model to achieve a balanced state with the help of class distribution and neuroscience principles. We further propose an effective deep convolutional neural network (CNN) model utilizing bidirectional Long Short-Term Memory (Bi-LSTM) with single-channel EEG as the Baseline. It is used for evaluating the efficiency of two balancing approaches on three imbalanced polysomnography datasets (CCSHS, Sleep-EDF and Sleep-EDF-V1). The qualitative and quantitative evaluation of experimental results demonstrates that the proposed methods could not only show the superiority of class balancing through the confusion matrix and class-wise metrics, but also get better N1 stage and whole stages classification accuracies compared to other state-of-the-art approaches.**

***Index Terms*—Sleep stage classification, Class imbalance problem, Deep neural network, Data augmentation, Generative adversarial network, Network connection.**

This work was support by National Key R&D Program of China (No.2021ZD0109803), National Natural Science Foundation of China (No.91748105), Youth Fund of National Natural Science Foundation of China (No.82102135), National Foundation in China (No. JCKY2019110B009, 2020-JCJQ-JJ-252), Fundamental Research Funds for Central Universities [DUT2019, DUT20LAB303, DUT21RC(3)091] in Dalian University of Technology in China, Open Project of Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University (No.MMC202104), Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ, No. GML-KF-22-11), CAAI-Huawei Mindspore Open Fund (CAAIXSJLJJ-2021-003A) and the Scholarships from China Scholarship Council (No.201806060164, No.202006060226).

D. Zhou is with School of Biomedical Engineering, Faculty of Electronic and Electrical Engineering, Dalian University of Technology, 116024, Dalian, China & Faculty of Information Technology, University of Jyväskylä, 40014, Jyväskylä, Finland & Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) (e-mail: dongdong.w.zhou@student.jyu.fi)

Q. Xu is with School of Artificial Intelligence, Faculty of Electronic and Electrical Engineering, Dalian University of Technology, 116024, Dalian, China & Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) (corresponding author: xuqi@dlut.edu.cn).

J. Wang and F. Cong are with School of Biomedical Engineering, Faculty of Electronic and Electrical Engineering, Dalian University of Technology, 116024, Dalian, China & Faculty of Information Technology, University of Jyväskylä, 40014, Jyväskylä, Finland (wangjian009@mail.dlut.edu.cn, cong@dlut.edu.cn).

H. Xu is with School of Biomedical Engineering, Faculty of Electronic and Electrical Engineering, Dalian University of Technology, 116024, Dalian, China (e-mail: mxu@dlut.edu.cn).

L. Kettunen and Z. Chang are with Faculty of Information Technology, University of Jyväskylä, 40014, Jyväskylä, Finland (e-mail: lauri.y.o.kettunen@jyu.fi, zheng.chang@jyu.fi).

I. Introduction

CORRECT sleep stage classification with overnight polysomnography (PSG) recordings plays an essential role in diagnosing and treating sleep-related disorders [1]–[3]. The PSG data consist of the EEG, electromyogram (EMG), electrocardiogram (ECG), electrooculogram (EOG), etc [4]. Clinically, the PSG data are divided into sequential 30-second (30s) epochs and then each epoch is labeled as one of the sleep stages by clinicians manually following the guidelines of the Rechtschaffen and Kales (R&K) [5] or the American Academy of Sleep Medicine (AASM) [6]. Regarding the AASM manual, the sleep stages can be defined as Wake (W), Rapid Eye Movement (REM), Non-REM1 (N1), Non-REM2 (N2) and Non-REM3 (N3).

However, it is cumbersome, time-consuming and prone to be subjective errors for the manual approach with visual inspection of PSG recordings [3]. Hence a large body of automatic sleep stages classification methods including the conventional machine learning [7]–[9] and the deep networks [10]–[15] have been proposed. Although these methodologies achieve promising performance in terms of overall accuracy, the inherent class imbalance problem (CIP) of PSG datasets have been barely explored. The class distribution of PSG databases is highly imbalanced on account of the specific sleep architecture. Additionally, the structure of whole-night sleep is greatly related to the subject's physiological and psychological condition and data acquisition environment. Hereinto, the stage N1 is the most challenging to be recognized and regarded as a representative of minority groups which usually accounts for 2%-5% of total sleep time, and the N1 stage plays the role of indicator in some sleep disorders. Typically, stage N1 would start within minutes of going to sleep, whereas insomnia may delay the beginning of the N1 stage. Moreover, people who have insomnia show a higher proportion of the N1 stage [16]. Besides, the sufferer with apnea may experience abnormal breathing during sleep, which would awaken the brain from deeper sleep. This could lead to an increase in stage N1 [17]. The N1 stage is also highly related to narcolepsy

[18]. Considering the importance of stage N1 recognition, the high misclassification rate of N1 has tremendously limited the practical application of automatic sleep stages classification approaches.

Only a few literature attempt to address the CIP in the sleep stage classification task. Sun *et al*. [19] introduced a DA approach employing the synthetic minority oversampling technique algorithm. A range of 8–14 dB white noise were added to enable the equal number of each sleep stage. It can be more appropriate to apply the DA to stage N1 rather than all sleep stages to maintain original structure of whole-night sleep maximally. In addition, the scope of signal noise ratio (SNR) of white noise could be extended to investigate the efficiency of different intensities noise. Tsinalis *et al*. [20] used class-balanced random sampling across sleep stages to avoid biased performance on the side of the most representative sleep stages and significantly improve the recall of the stage N1. But the overall accuracy achieved was 78%, which is not good enough compared to other state-of-the-art methodologies. One important reason is that the class-balanced random sampling diminished the importance of major classes making the primary contribution to classification performance. It should be noted that keeping a sensible equilibrium among different distribution classes. Fan *et al*. [21] investigated the efficiency of five DA approaches for sleep EEG signals. New training datasets were created with each class equals in number by means of DA algorithms. The overall classification performance was improved, nevertheless, the stage N1 showed a slight drop in terms of F1 scores. Apart from applying DA methods to balance the class distribution of PSG datasets, the correlation between categories and the trained model should not be ignored. CIP poses a big challenge for the prediction model as most machine learning or deep learning algorithms for classification were designed based on the assumption of the same number of samples in each category. The loss weight of each class is equal, which may lead to discrimination against the minority class.

We initially introduce CIP and define the class imbalance factor (CIF) in sleep PSG datasets systematically. To tackle the CIP in the field of automatic sleep stage classification, two solutions are introduced. The first one is to balance the database quantity by means of the DA approaches using the generative adversarial network (GAN) model and Gaussian white noise (GWN) addition, which increases the number of the N1 stage in the training set. The second method is to balance the relationship between the trained model and the original imbalanced dataset through setting different class weights (CW) in the loss function. To assess the efficiency of DA and CW methods, we further propose an efficient deep model that implements Bi-LSTM and CNNs to extract features across temporal and spatial scales with single-channel EEG simultaneously. In this paper, the proposed model is regarded as the Baseline, the proposed framework with the DA of GAN model, the DA of Gaussian white noise and CW are named the Baseline + GAN , the Baseline + GWN and the Baseline + CW, respectively. The main contributions of this work are summarized as follows:

i) We systematically analyze the class imbalance problem in PSG datasets. Furthermore, we propose two solutions to tackle the CIP from the database quantity and the correlation between classes and the applied model.
ii) We explore the GAN model and the method with Gaussian white noise addition to balance the PSG dataset samples. We further search for the balanced network connection from the perspectives of class distribution and neurology.
iii) We develop a novel model that utilizes one convolution block and two multi-convolution (MC) blocks with different filter sizes as the spatial feature extractor. Another temporal feature extractor consisting of one CNN and Bi-LSTM can learn the information of sleep stage transition rules.
iv) The overall performance and recognition of the N1 stage could be improved to different extents by proposed methods on three public datasets.

The rest of this paper is organized as follows. We demonstrate the experimental datasets and methodologies in Sec. II. In Sec. III, the experimental results are represented. The final discussion and conclusion are included in Sec. VI and Sec. V.

II. Materials and Methods

A. Data Description

We employ three public PSG datasets in this study: Cleveland Children's Sleep and Health Study (CCSHS) [22], [23], Sleep-EDF Database (Sleep-EDF-V1, version 2013) and Sleep-EDF Database Expanded (Sleep-EDF, version 2018) [24]. As the recommendation of the AASM manual, the central and frontal lobes are used. More specifically, C4/A1 and Fpz-Cz EEG channels are selected from the CCSHS and Sleep-EDF datasets respectively.

The CCSHS database is one of the largest pediatric cohorts, including 515 children whose ages range from 16-19 years. In our experiments, C4/A1 channel EEG signals sampled at 128 Hz are used. Each 30s epoch was labeled by trained-well sleep experts.

There are two subsets: sleep-cassette (SC) and sleep-telemetry (ST) in the Sleep-EDF dataset (Sleep-EDF-V1). We use 39 whole-night PSG recordings from 20 subjects aged 25 to 34 years in the SC cohort. Each subject has two full night PSG recordings except for subject 13. The number of individuals in SC subset is increased to 78 with 153 over-night sleep recordings in Sleep-EDF Database Expanded (Sleep-EDF). The oldest subject is 101 years. In our study, we employ Fpz-Cz EEG signals with a sampling rate (i.e., f_s) of 100Hz. It is worthy that the resampling method is not applied to restrict the sampling rate, which means our model can be adaptable to different input lengths. Besides, we only adopt 30 minutes of W epochs before and after sleep stages, as there are long W stages at the start and end of the whole-night sleep in Sleep-EDF and Sleep-EDF-V1 datasets. Considering the correlation and dependency between surrounding epochs, we use the many-to-one scheme described in our prior study that combines one 30s epoch with its neighboring epochs (i.e., three sequential 30s epochs) as the 90s epoch [25]. There is 60s overlap between the adjacent 90s epochs and the label

TABLE I
THE NUMBER OF 90S EPOCHS FOR EACH SLEEP STAGE FROM EXPERIMENTAL DATASETS

Stage	CCSHS	Sleep-EDF	Sleep-EDF-V1
W	211030 (30.6%)	69518(34.9%)	10197 (23.1%)
N1	**19211 (2.8%)**	21522 (10.8%)	**2804 (6.3%)**
N2	249681 (36.2%)	69132 (34.7%)	17799 (40.3%)
N3	110188 (16.0%)	**13039 (6.6%)**	5703 (13.0%)
REM	100252 (14.5%)	25835 (13.0%)	7717 (17.5%)
Total	690372	199046	44220
CIF	3.6	1.5	1.6

TABLE II
THE NUMBER AND PROPORTION OF N1 STAGE BEFORE AND AFTER DATA AUGMENTATION (GAN MODEL) IN THE TRAINING SET

Status	CCSHS	Sleep-EDF	Sleep-EDF-V1
Before	15721 (2.8%)	19284 (11.2%)	2024 (5.4%)
After	31442 (5.5%)	38568 (20.1%)	4048 (10.3%)

of the 90s epoch is the same as the label from the middle 30s epoch. We show in Table I the number and percentage of 90s epochs for each sleep stage from three datasets in our experiments, the class with the smallest number of samples is labeled in bold. The N1 stage occupies the smallest percentage, which equals 2.8% and 6.3% respectively in CCSHS and Sleep-EDF-V1 datasets. While the proportion of N1 in the Sleep-EDF dataset is 10.8% and the N3 stage has the smallest number of samples. Sleep architecture changes with ages [26], sleep efficiency would decline with the increase of age due to frequent arousals from sleep, these changes result in an increment of N1 stage.

B. Class Imbalance Problem

In computer vision (CV), the equal number of each category of some image datasets (e.g., CIFAR-10 database) can be guaranteed. However, the sleep pattern differs from ages, genders and physical conditions of individuals [26], [27], the sleep PSG database suffers severe CIP with imbalanced class distribution. In other words, some sleep stages occupy the dominant proportion, whereas the other stages become the minority classes. For instance, the number of the N2 stage is several times that of the N1 stage. When training a model, the majority class contributes the leading weight updating and therefore the performance of minority classes is biased with a higher misclassification rate. The severity of CIP is described using the class imbalance factor (CIF), which is calculated as follow:

$$CIF = \frac{N}{2 \cdot c \cdot \min\{N_i\}} \quad i \in \{1, 2, \ldots, c\} \tag{1}$$

Where c is the number of classes, N represents the number of all epochs, N_i refers to the number of epochs of class i. We argue that the dataset suffers CIP when CIF is greater than or equal to 1. The greater the CIF is, the more imbalanced the database is. In this study, the CIF of CCSHS, Sleep-EDF, and Sleep-EDF-V1 datasets are 3.6, 1.5 and 1.6, respectively.

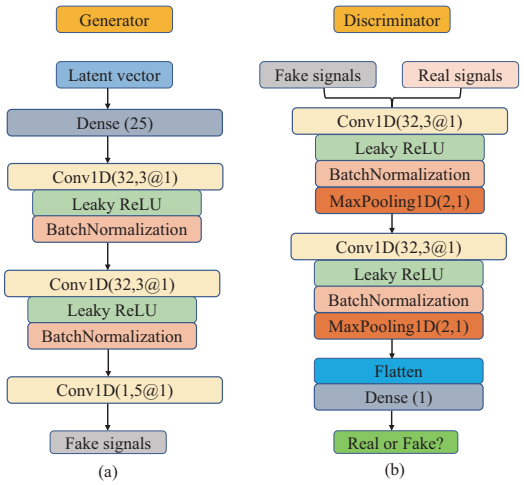


Fig. 1. The framework of the GAN model. Demonstrate the structure: (a) Generator, (b) Discriminator.

In order to alleviate the negative effect of CIP on classification performance, we propose two balancing solutions. The first one is to raise the number of N1 stage with the DA method, which could improve the severity of imbalance to some extent. Another one is to find out the inner network connection between classes and the trained model while maintaining the original dataset quantity, that is to say, setting different class weights for each category depend on the specific class distribution and the neuroscience rule.

C. Balance the Dataset Samples

The imbalanced class distribution has negative effect on the training procedure, which means the applied model could not be trained efficiently. Hence, it is natural and straightforward to increase the number of minority classes to achieve the same proportion, whereas this would break the original architecture of whole-night sleep. By contrast, we choose to produce new epochs of the N1 stage in the training set to maintain the physiologic sleep structure maximally, but the test set is kept independent without balancing sample operation.

The generative adversarial network model has attained significant achievement in the CV field, however, this technology is barely adopted to augment synthetic EEG signals. We use the GAN model as the first method to generate artificial EEG signals of the N1 stage in this study. The GAN is generally comprised of two opposing networks (i.e., generator (G) and discriminator (D)) as shown in Fig. 1. The generator mainly includes three one-dimensional convolutional (Conv1D) layers, thereinto, the first two Conv1D layers are assembled with LeakyReLU (the activation function) and the batch normalization and the last one is used to generate the demanded length signals. In addition, the padding is set as casual to keep the length unchanged. In terms of the discriminator, the Conv1D layer is followed by the LeakyReLU, batch normalization and MaxPooling1D sequentially. The final dense layer makes the prediction for the inputting signal. Given a latent vector

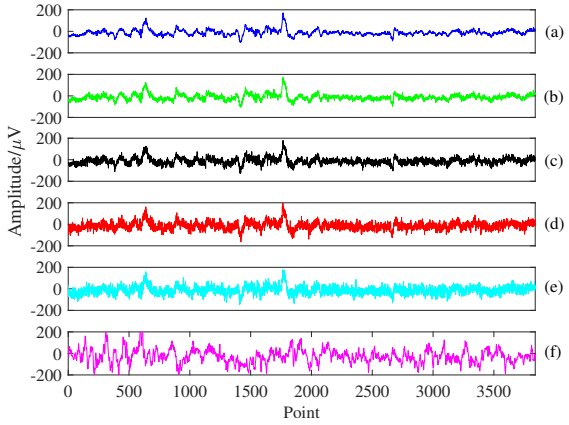


Fig. 2. Raw EEG signal (N1 stage) and Gaussian white noise addition with four SNR. (a) Raw EEG. (b) Gaussian white noise addition with 10 dB. (c) Gaussian white noise addition with 5 dB. (d) Gaussian white noise addition with 2 dB. (e) Gaussian white noise addition with 1 dB. (f) Artificial signal by the proposed GAN model.

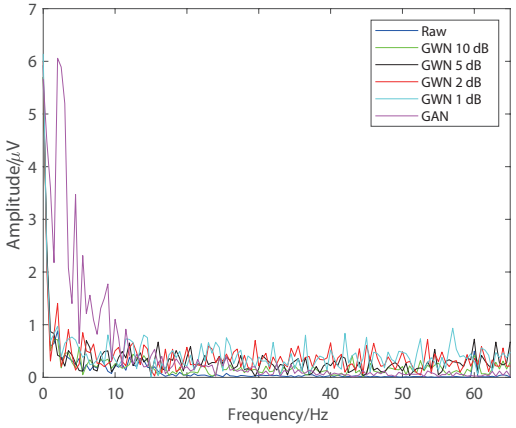


Fig. 3. Spectrogram of raw EEG signal and artificial signals generated by the Gaussian white noise addition with 10 dB, 5 dB, 2 dB, 1 dB and the GAN model.

z following the standard normal distribution (N(0,1)), the generator maps it to the input space and learns a distribution $\mathbb{P}_g$ to approach the distribution $\mathbb{P}_{data}$. The discriminator is designed for distinguishing the fake signals generated by the generator and real signals by estimating the correspondence between $\mathbb{P}_g$ and $\mathbb{P}_{data}$. It can be defined as the minimax objective:

$$\begin{aligned}\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] \\ +\mathbb{E}_{\boldsymbol{z}\sim\mathbb{P}_{\boldsymbol{g}}(\boldsymbol{z})}[\log(1-D(G(\boldsymbol{z})))]\end{aligned} \tag{2}$$

where $D(\boldsymbol{x})$ means the probability of $\boldsymbol{x}$ sampled from the real samples $\mathbb{P}_{data}$. $G(\boldsymbol{z})$ stands for the artificial signals produced by the generator. Additionally, we adopt the loss function presented by Gulrajani *et al.* [28]:

$$\begin{aligned}L\left(\mathbb{P}_{data},\mathbb{P}_g\right) = E_{x_r\sim\mathbb{P}_{data}}[D(x)] \\ -E_{x_g\sim\mathbb{P}_g}\left[D\left(x_g\right)\right] + P(\tilde{x})\end{aligned} \tag{3}$$

$$P(\tilde{x}) = \lambda \cdot E_{\hat{x}\sim\tilde{X}}\left[\max\left(0, \|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1\right)^2\right] \tag{4}$$

where $P(\tilde{x})$ is defined as the one-sided gradient penalty, λ denotes the the penalty coefficient and $\tilde{X}$ includes points sampling alone the straight line between $\mathbb{P}_{data}$ and $\mathbb{P}_g$. We employ the Adam optimizer to update the model parameters and choose five iterations to train the generator for each iteration of the discriminator. We demonstrate the number and the proportion of N1 in the training set before and after the data augmentation with the GAN model in Table II.

The second method for balancing the dataset samples is the noise addition. Unlike repeating samples of the minority stage directly [11], the data augmentation method with Gaussian white noise addition is implemented in this work for two important reasons. On the one hand, the acquisition of EEG signals always accompanies with noise, a Gaussian noise that imitates the line-related noise that is commonly found in electrophysiology recordings, hence the data generated by noise addition can be more real-like sleep EEG signals. On the other hand, generated data with noise addition can provide the trained model with new features and enhance the generalization. To be specific, we investigate the efficiency of the DA algorithm with four different intensities Gaussian white noise ranging from 1-10 dB. Fig. 2 and Fig. 3 show an example of this DA procedure with different intensities and the DA with GAN model in terms of the amplitude and spectrogram, we can find that these implementations with Gaussian white noise addition retain wave properties of the raw EEG signal. We further explore the effectiveness of various times noise addition. Specifically, once obtaining the optimal intensity (x dB), the intensities of three and five times noise addition are defined as (x - 0.2, x, x + 0.2) dB and (x - 0.2, x - 0.1, x, x + 0.1, x + 0.2) dB with a type of arithmetic progression, respectively. Compared with the way of repeating corresponding times noise addition with x dB, this could provide with the trained model with additional information.

D. Balance Relationship Between the Imbalanced Dataset and Trained Model

The CIP is not only the imbalance of class distribution but also the imbalanced network connection. Although the DA method could mitigate the imbalance of PSG datasets, whether DA with the GAN model or DA with noise addition, the generated data are still fake. More importantly, we could not ignore the corresponding physiological information behind the PSG dataset for real-world application. In other words, it would be more meaningful to achieve the performance improvement without changing the distribution of class. Therefore, another alternative is to balance the network connection between the sample distribution and the trained model with the original imbalance PSG dataset. By default, the weight of each class is the same. As a consequence, the majority class occupies the dominant weight updating with a more considerable length of the gradient component. Furthermore, the performance of the minority classes is prejudiced by the trained model. To eliminate the discrimination, we reassign

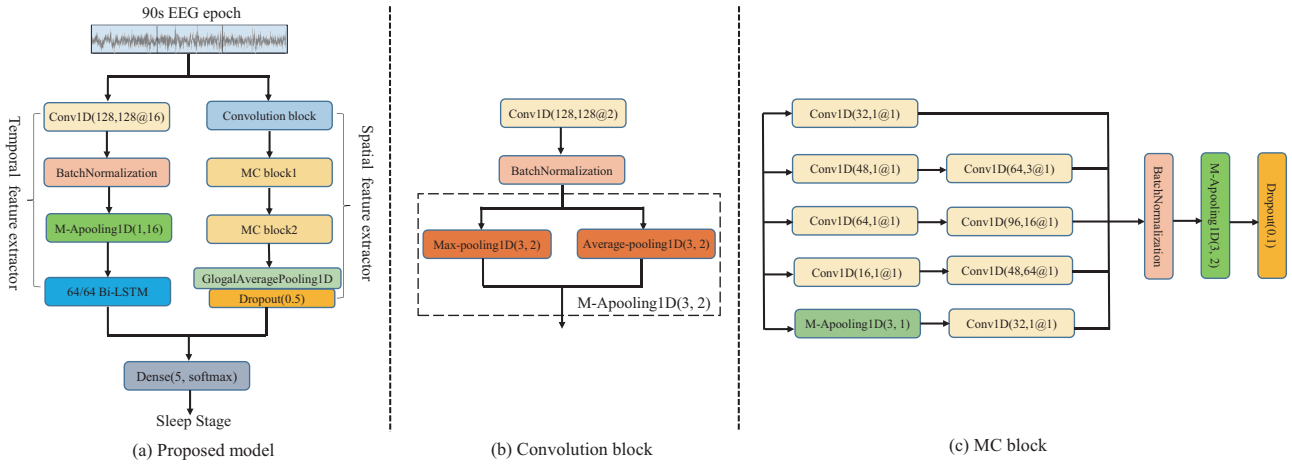


Fig. 4. The schematic diagram of the proposed model.

TABLE III
PARAMETERS OF THE PROPOSED MODEL

Layer	Layer Type	Filters	Size	Stride	Activation	Output dimension
SFE1	Input	-	-	-	-	$(90 \times f_s, 1)$
SFE2	Convolution block	128	128	2	relu	$(\lceil 45 \times f_s/2 \rceil, 256)$
SFE3	MC block1	-	-	-	-	$(\lceil 45 \times f_s/4 \rceil, 544)$
SFE4	MC block2	-	-	-	relu	$(\lceil 45 \times f_s/8 \rceil, 544)$
SFE5	GAP	-	-	-	-	544
SFE6	Dropout (0.5)	-	-	-	-	544
TFE1	Input	-	-	-		$(90 \times f_s, 1)$
TFE2	Conv1D	128	128	16	relu	$(\lceil (90 \times f_s - 128)/16 \rceil, 128)$
TFE3	BatchNormalization	-	-	-	-	$(\lceil (90 \times f_s - 128)/16 \rceil, 128)$
TFE4	M-Apooling1D	-	1	16	-	$(\lceil (90 \times f_s - 128)/256 \rceil, 256)$
TFE5	Bi-LSTM	64	-	-	tanh	128
Decision	Dense	-	-	-	softmax	5

the weight (W) of each class based on the class distribution and the brain-inspired rule, namely CW (Ratio), CW (Log_R) and CW (E_I), respectively. The W_i, W_j using CW (Ratio), CW (Log_R) methods are shown in equations (5), (6):

$$W_i = \frac{N}{N_i} \quad i \in \{1, 2, \ldots, 5\} \tag{5}$$

$$W_j = \ln \frac{N}{N_j} \quad j \in \{1, 2, \ldots, 5\} \tag{6}$$

Where N, N_i and N_j are respectively the numbers of whole classes, class i and j samples. The CW (Ratio) is a direct way to get the W_i by calculating the ratio of the numbers of all samples and each class. Additionally, we attempt a more moderate and sensible approach, the CW (Log_R), to attain the natural logarithm of W_i of the CW (Ratio) method. The CW (E_I) algorithm considers the allocation of neurons during information processing procedures in the human brain [29], [30], namely the ratio of excitatory neurons to inhibitory neurons. Zeng *et al*. [29] investigated the effect of the proportion of inhibitory neurons on the spiking neural networks. As a result, the 15% of inhibitory neurons are the optimal for good performance. Inspired by this brain-inspired rule, we regard the samples of N1 stage as the excitatory neurons, other stages as the inhibitory neurons. To be specific, we set the weight of N1 stage with the value of 8.5, other stages with the weight of 1.5. Three CW methods adopted in this study aim to strengthen the contribution of the minority class and ultimately mitigate the bias towards the majority class.

E. Proposed Model

To evaluate the efficiency of two balancing methods used in this study, we propose a CNN based model for automatic sleep stage classification. The proposed framework is composed of two key parts as illustrated in Fig. 4. The first part is the temporal feature extractor (TFE), which could learn the temporal information (e.g., transition rules between stages). Another part is the spatial feature extractor (SFE) for extracting spatial features. The concatenation of feature maps extracted from the

temporal and spatial feature extractors is fed into the dense layer with the activation function of softmax to make the final decision.

The temporal feature extractor consists of a one-dimensional convolutional (Conv1D) layer, batch normalization, M-Apooling layer and Bi-LSTM layer. The main function of the Conv1D is to attain the feature map from the raw EEG signal. Then the Bi-LSTM is responsible for leaning the temporal information, such as the transition rule between successive stages. Practically, the clinicians decide the next probable stage based on the prior stage on some occasions.

The spatial feature extractor includes four components: a convolution block, two multi-convolution (MC) blocks (inspired by the inception module [31]), a GlobalAveragePooling (GAP) layer and a dropout layer. The convolution block is followed by a Conv1D layer with 128 filters of size 128 and a stride of 2, batch normalization and M-Apooling layer in sequence. Analogously, the MC block comprises different sizes of filters, batch normalization, M-Apooling layer and dropout layer. The purpose of different filter sizes is to capture feature representations in multi-scales. We optimize the filter sizes with small (3, 5 and 7), medium (16 and 32) and large (64, 128 and 256) sizes to adapt to the long input length. In addition, the filter size of 1 is applied to enhance the nonlinearity of the network. The filter sizes are selected with 1, 3, 16 and 64 as they provide the optimal results in our testing. We use the M-Apooling layer, the concatenation of the average-pooling and max-pooling layer, to replace the conventional max-pooling layer in our model. The GAP layer plays the role of the traditional fully connected layer to flat the previous output without introducing extra trainable parameters, which can prevent the overfitting problem efficiently [32]. Table III shows the detailed information of the proposed model, the length of input is $90 \times f_s$, which is related to the sampling rate.

F. Experimental Setup

We divide the whole dataset into the training and test sets randomly based on the subject-wise scheme (i.e., 80% subjects for training, 20% subjects for test). Only recordings from the CCSHS dataset are employed to tune the hyper-parameters of the proposed model. Besides, we choose the Adam as the model optimizer with the algorithm of learning rate (LR) reducing, and the LR would decrease to half of it when the accuracy of test set shows no improvement within three epochs. The value of LR ranges from 10^{-7} to 10^{-3}. In addition, the size of mini-batch is set to 64 chosen from four batch sizes (32, 64, 128, and 256). We select the categorical cross entropy as the loss function, which is always employed for the multi-class model. The number of iteration is 40 as the proposed model could achieve the convergence state within 40 epochs. Furthermore, we save the model with the best test accuracy in all iterations.

To prevent the overfitting problem, we adopt two regularization strategies in this study. The first strategy is the L2 regularization, which adds a squared magnitude of coefficient as penalty term to the loss function. Then we test four regularization rates (10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}), and 10^{-3} is adopted finally. The second technology is the dropout that drops units from the model with a probability from 0-1. In the MC block and dropout layer, the probabilities are set to 0.1 and 0.5 respectively.

In our cases, we conduct the experiments on a workstation with two Inter Xeon E5-2640 V4 CPUs and four Nvidia Tesla P100 GPUs with 16 Gbytes memory.

III. Experimental Results

A. Performance Metrics

We use class-wise recall (RE), overall accuracy (ACC) and Cohen's kappa coefficient (K) to evaluate the performance. Similar to the binary classification, we regard each class as a positive class, other classes as a negative class to compute the class-wise metrics. The calculation of RE, ACC and K are shown as follows:

$$RE = \frac{TP}{TP + FN}. \tag{7}$$

$$ACC = \frac{\sum_{i=1}^{n} x_{ii}}{N} \tag{8}$$

$$K = \frac{\frac{\sum_{i=1}^{n} x_{ii}}{N} - \frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{n} x_{ij} \sum_{j=1}^{n} x_{ji}\right)}{N^2}}{1 - \frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{n} x_{ij} \sum_{j=1}^{n} x_{ji}\right)}{N^2}}. \tag{9}$$

where TP and FN, respectively, stand for the true positive and false negative, N is the total number of test epochs, c represents the number of classes. In this study, c equals 5, x_{ii} ($1 \leq i \leq 5$) refers to the diagonal value of the confusion matrix.

B. Efficiency of Balancing the Dataset Samples

Table IV illustrates the performance of DA methods with proposed GAN model and different intensities and times Gaussian white noise addition, the bold format stands for the best performance of each index. Compared to the Baseline model, the proposed GAN model can improve the overall accuracy, however, show an slight decrease in terms of the RE of N1 stage (RE_N1) on the experimental datasets. By contrast, the ACC, K and RE_N1 have been enhanced to a different extent on three datasets with the GWN method. Specifically, the RE_N1 has an increase of 9.7%, 16.2%, 12.0% with systems of Baseline + GWN (1 dB), Baseline + GWN (1 dB) and Baseline + GWN (10 dB) on the CCSHS, Sleep-EDF, Sleep-EDF-V1 databases, respectively. In addition, ACC and K are also improved with a range of 0.1% to 2.2%. The improvement of N1 performance (RE_N1) is the priority thing to be considered in the situation of comparable ACC and K. Besides, the enhancement of N1 recognition should not sacrifice the overall performance. Considering the overall and N1 performance, the optimal intensity of Gaussian white noise addition is set as 1 dB. Hence, the intensities of GWN methods with three and five times are respectively set to (0.8, 1.0, 1.2) dB and (0.8, 0.9, 1.0, 1.1, 1.2) dB. Generating more samples of N1 stage could not achieve better overall (ACC and K) and N1 (RE_N1) performance simultaneously compared to the

TABLE IV
PERFORMANCE COMPARISON OF THE PROPOSED GAN MODEL AND DIFFERENT INTENSITIES AND TIMES GAUSSIAN WHITE NOISE ADDITION IN THIS WORK

	CCSHS			Sleep-EDF			Sleep-EDF-V1		
	ACC (%)	K (%)	RE_N1 (%)	ACC (%)	K (%)	RE_N1 (%)	ACC (%)	K (%)	RE_N1 (%)
Baseline	88.2	83.8	23.0	86.4	81.1	24.7	85.4	79.9	33.6
Baseline + GAN	88.5	84.3	21.4	86.9	82.0	20.4	86.5	81.5	32.1
Baseline + GWN (1 dB)	88.3	84.0	**32.7**	86.5	81.5	**40.9**	86.3	81.4	44.6
Baseline + GWN (2 dB)	88.3	84.0	30.0	86.7	81.7	28.3	**86.8**	**82.1**	35.9
Baseline + GWN (5 dB)	88.4	84.1	28.4	86.6	81.5	26.7	86.1	80.9	34.6
Baseline + GWN (10 dB)	88.4	84.0	29.0	**87.0**	**82.2**	32.4	86.0	80.9	45.6
Baseline + GWN (three times)	**88.6**	**84.3**	28.2	86.2	80.9	38.2	85.8	80.8	**49.0**
Baseline + GWN (five times)	88.4	83.9	31.0	86.5	81.6	30.2	85.9	80.9	47.4

TABLE V
THE WEIGHT OF EACH CLASS WITH DIFFERENT CW METHODS

	CCSHS			Sleep-EDF			Sleep-EDF-V1		
	CW (Ratio)	CW (Log_R)	CW (E_I)	CW (Ratio)	CW (Log_R)	CW (E_I)	CW (Ratio)	CW (Log_R)	CW (E_I)
W	3.3	1.2	1.5	2.5	0.9	1.5	3.7	1.3	1.5
N1	34.9	3.6	8.5	9.0	2.2	8.5	18.4	2.9	8.5
N2	2.8	1.0	1.5	3.1	1.1	1.5	2.6	1.0	1.5
N3	6.3	1.8	1.5	19.4	3.0	1.5	8.0	2.1	1.5
REM	6.9	1.9	1.5	8.6	2.1	1.5	5.9	1.8	1.5

TABLE VI
PERFORMANCE COMPARISON OF DIFFERENT CW METHODS IN THIS WORK

	CCSHS			Sleep-EDF			Sleep-EDF-V1		
	ACC (%)	K (%)	RE_N1 (%)	ACC (%)	K (%)	RE_N1 (%)	ACC (%)	K (%)	RE_N1 (%)
Baseline	**88.2**	**83.8**	23.0	86.4	81.1	24.7	85.4	79.9	33.6
Baseline + CW (Ratio)	85.3	80.3	**75.0**	**86.5**	**81.5**	30.4	**87.3**	**82.7**	**42.6**
Baseline + CW (Log_R)	87.8	83.4	51.3	86.3	81.1	33.9	85.9	80.8	36.4
Baseline + CW (E_I)	86.6	81.8	67.7	85.9	80.4	**35.7**	85.8	80.8	34.5

Baseline + GWN (1 dB). It is noteworthy that we do not apply the DA operation to the test set, which means the sleep structure of the test set is not destroyed. Employing more times noise addition stands for the worse consistency of training and test sets, which may hinder the classifier from achieving better performance.

C. Efficiency of Balancing the Network Connection

Table V shows the class weight of the training set using three CW methods from the experimental datasets. To further demonstrate how different CW methods may affect the performance, we make a performance comparison in Table VI. The performance obtained by CW methods differs significantly on three datasets. It can be seen that the RE_N1 shows a dramatic increase by all CW approaches, corresponding to 52.0%, 28.3%, and 44.7% (by CW (Ration), CW (Log_R) and CW (E_I) respectively) on the CCSHS dataset. Nevertheless, ACC and K decrease slightly instead. By contrast, on the Sleep-EDF and Sleep-EDF-V1 databases, ACC and K attain slight improvements except by the CW (Log_R) and CW (E_I) methods on the Sleep-EDF dataset. Additionally, the improvements of RE_N1 is relatively lower than those on the CCSHS dataset.

We show in Fig. 5 the confusion metrics of three datasets utilizing four systems (the Baseline, the Baseline + GAN, the Baseline + GWN, and the Baseline + CW). For both CCSHS and Sleep-EDF datasets, the Baseline + GWN (1 dB) and the Baseline + CW (Log_R) are selected as the optimal decision considering the overall performance and the accuracy rate of N1 stage. Whereas, we choose the Baseline + GWN (1 dB) and the Baseline + CW (Ratio) based on experimental results of the Sleep-EDF-V1 database. We further in Fig. 6 reveal the hypnogram comparison labeled by experts and the predictions of four systems for one subject (ccshs-trec-1800905) of the CCSHS dataset. Fig. 7 demonstrates the distribution of weights in the layer with the largest number of parameters (without and with the CW method). We also calculate the kurtosis and skewness of two weight distributions, the kurtosis and skewness of the weight distribution without and with the

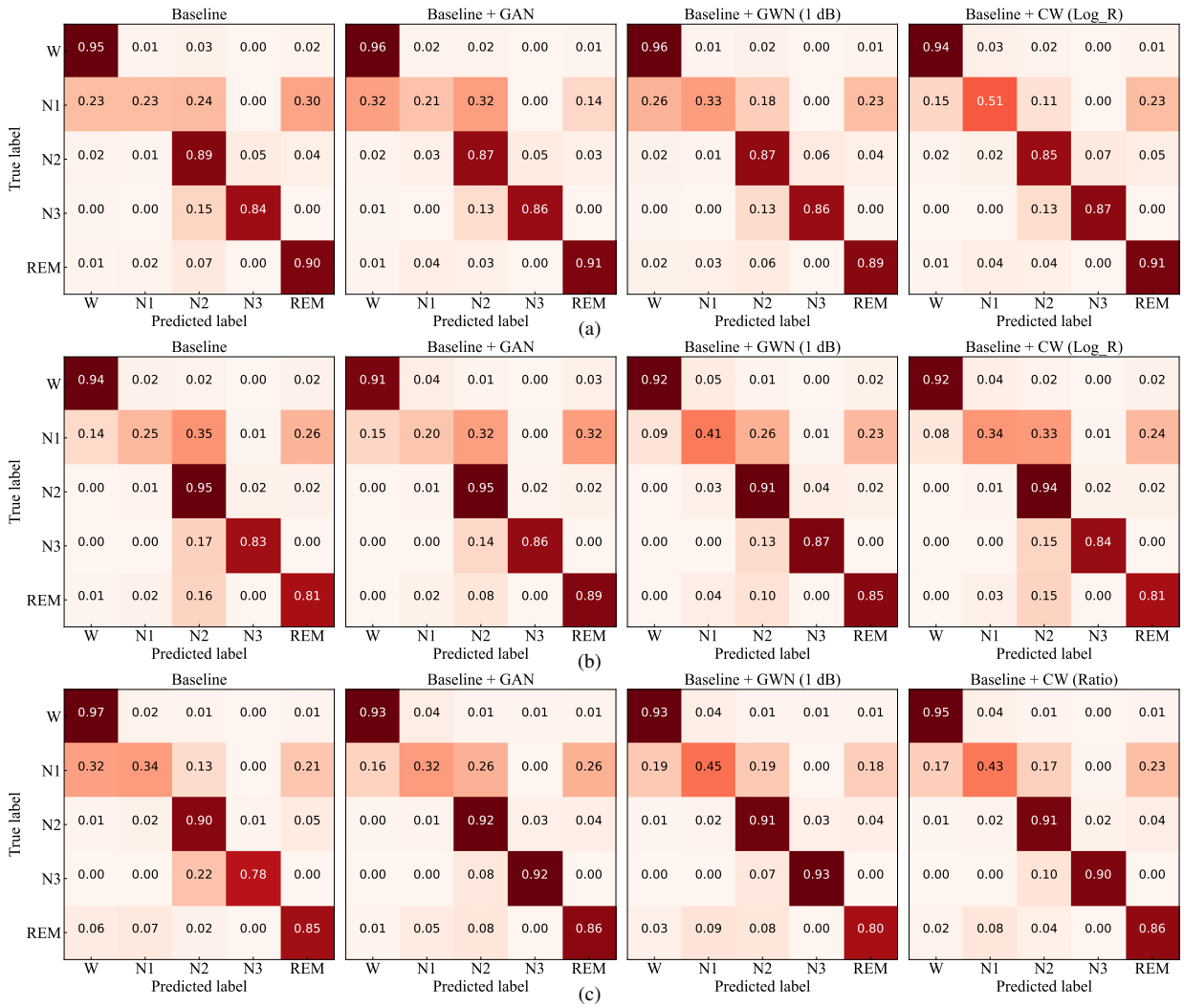


Fig. 5. The confusion matrices of three datasets with four systems. (a) the CCSHS dataset. (b) the Sleep-EDF dataset and (c) the Sleep-EDF-V1 dataset.

CW (Log_R) method are (0.034, -0.414) and (0.094, -0.090), respectively. The CW method more closely resembles a normal distribution (i.e., (0,0)). In such way the network convergence velocity becomes faster [33] and achieving more efficient training procedure for the minority class.

D. Performance Comparison

To see an overall picture, we demonstrate the performance comparison with previous works on the three datasets in Tables VII and VIII. Only a few studies employ the CCSHS dataset, the proposed systems (the Baseline, the Baseline + GWN (1 dB)) could achieve better performance compared to [34], [35]. Similarly, we compare the performance of the Baseline, the Baseline + GWN (1 dB) with [25], [36]–[39] on the Sleep-EDF database, the best ACC, K and RE_N1 are obtained by the Baseline + GWN (1 dB). Those literature [21], [38], [40], [41] utilize the Sleep-EDF-V1 dataset to develop automatic sleep stage classification model, the Baseline + CW (Ratio) framework shows a better ACC, K and a more favorable RE_N1 compared with them.

IV. Discussion

Class imbalance problem is one of the critical factors in real-world automatic sleep stage classification tasks especially using deep learning based models. Here in this paper, we introduce the CIP and define the CIF in the currently common PSG datasets. Correspondingly, this paper introduces two balancing methods to alleviate its negative effect from the dataset quantity and the relationship between the class distribution and the applied model respectively. One is to balance the dataset quantity through increasing the number of samples in the N1 stage, the other aims to balance the relationship between the original imbalanced datasets and deep neural networks while keeping the original dataset quantity. Embedding with two introduced methods, this paper propose a deep convolution neural network based model with Bi-LSTM units for automatic sleep stage classification tasks with single-channel EEG.

In order to enhance the ability of feature extraction, we use the MC block with four sizes of filters to capture spatial features from different scales. The small and large filters are responsible for capturing local features and big context, respectively [10]. In addition, the Bi-LSTM is designed as the

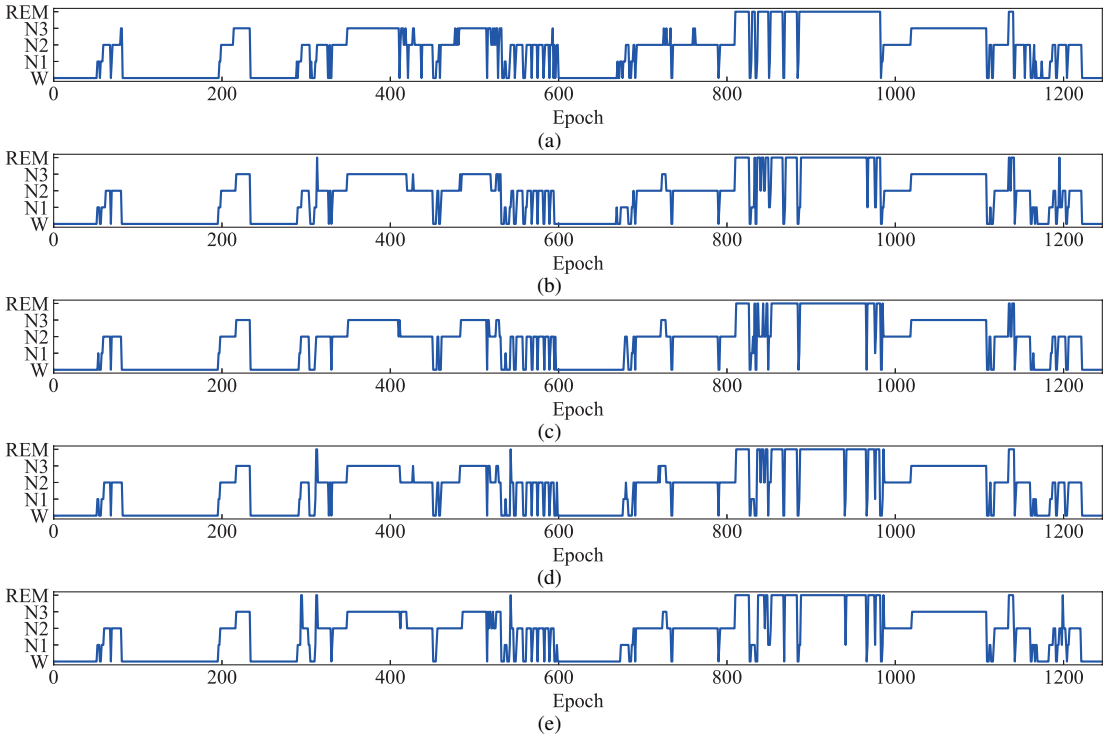
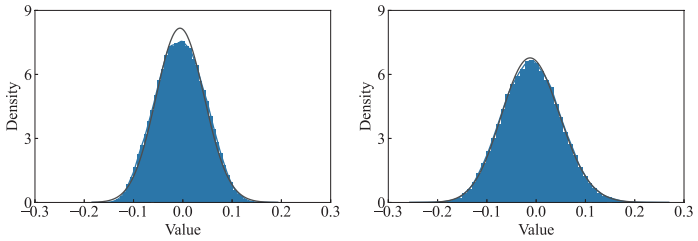


Fig. 6. Hypnogram of one subject from the test set (ccshs-trec-1800905). (a) the ground truth. (b) the prediction of the Baseline. (c) the prediction of the Baseline + GAN. (d) the prediction of the Baseline + GWN (1 dB). (e) the prediction of the Baseline + CW (Log_R).



(a) imbalanced network connection (b) balanced network connection

Fig. 7. The distribution of weights, black lines represent the curve of normal distribution, Y-axis refers to the probability density. (a) without the CW method, (b) with the CW (Log_R) method.

temporal feature extractor to learn the information of sleep stage transition rules. It further enriches features learned from the proposed model. The principle of the DA method here is quite different from previous studies [19], [21], in which the number of each category is designed as the same proportion. However, doing so, the original sleep structure is seriously destroyed. We argue that the physiological correlation between successive sleep stages should not be ignored. That is to say, the initial architecture of whole-night sleep needs to be intact maximally for clinical significance. By contrast, we only increase the number of N1 stage as it is typically considered as the archetype of minority classes with the highest misclassification rate. Different from duplicating selected samples from minority classes in [11], this paper adopts two DA methods with the proposed GAN model and Gaussian white noise addition to generate EEG signals. Although the under-sampling method can also improve the proportion of the minority class and does not need to generate new samples, the evaluation model may suffer from the underfitting problem with the decrease in the training samples. Employing the proposed DA methods, we could not only achieve the goal of increasing the samples of the minority class, but also introduce additional features to enhance the generalization of the applied model. As can be seen from Table IV, the applied GWN method could obtain different degrees of improvement of overall accuracy and recall of N1 stage simultaneously on three datasets compared to those of the baseline model. Nevertheless, the performance of N1 stage showed a slight decrease in [21]. Unlike the image database with independent classes, it is not necessary to keep the equal percentage of each class for mitigating the CIP in PSG datasets. More importantly, we should take into consideration in maintaining inherent characteristics of PSG datasets when employing DA methods. On the other hand, we should develop tailor made DA methods (e.g., different intensities and times noise addition) to deal with the diversity of subjects in different PSG datasets. For instance, both the macro-level (including the sleep stages and duration) and micro-level (such as the quality and quantity of sleep oscillations) structure of sleep would change with the older age [42] and sleep disorders.

Nevertheless, the generated EEG signals by the DA approaches are still artificial. Apart from balancing the class distribution of datasets, another method is to discover the balanced network connection with the original imbalanced dataset. Compared to the DA method, this method could enable the original architecture of sleep and handle general imbalanced PSG datasets. More specifically, we try to balance the relationship between the class and the trained model from the data distribution and the brain-inspired rule. According

TABLE VII
PERFORMANCE COMPARISON BETWEEN THE PROPOSED SYSTEMS AND PREVIOUS METHODS ON THE CCSHS DATASET

Study	Method	Input channel	Input type	Subjects	$ACC(\%)$	$K(\%)$	RE_N1
Ref. [34]	HMM	C4/A1 + C3/A2	Spectrogram	515	-	73.0	-
Ref. [35]	Random Forest	C4/A1	Features	116	86.0	80.5	7.3
Baseline	**CNN + LSTM**	**C4/A1**	**Time series**	**515**	**88.2**	**83.8**	**23.0**
Baseline + GWN (1 dB)	**CNN + LSTM**	**C4/A1**	**Time series**	**515**	**88.3**	**84.0**	**32.7**

TABLE VIII
PERFORMANCE COMPARISON BETWEEN THE PROPOSED SYSTEMS AND PREVIOUS METHODS ON THE SLEEP-EDF AND SLEEP-EDF-V1 DATASETS

Study	Database	Method	Input channel	Input type	Subjects	$ACC(\%)$	$K(\%)$	RE_N1
Ref. [25]	Sleep-EDF	CNN	Fpz-Cz	Time series	78	83.9	77.8	-
Ref. [36]	Sleep-EDF	CNN + LSTM	Fpz-Cz	Time series	78	80.0	73	-
Ref. [37]	Sleep-EDF	CNN + LSTM	Fpz-Cz	Time series	78	83.1	77	-
Ref. [38]	Sleep-EDF	RNN	Fpz-Cz	Time series	78	84.0	77.8	-
Ref. [39]	Sleep-EDF	CNN	Fpz-Cz	Spectrogram	78	83.4	76.7	-
Baseline	**Sleep-EDF**	**CNN + LSTM**	**Fpz-Cz**	**Time series**	**78**	**86.4**	**81.1**	**24.7**
Baseline + GWN (1 dB)	**Sleep-EDF**	**CNN + LSTM**	**Fpz-Cz**	**Time series**	**78**	**86.5**	**81.5**	**40.9**
Ref. [21]	Sleep-EDF-V1	Deep CNN	Fpz-Cz	Time series	20	74.8	66.0	-
Ref. [38]	Sleep-EDF-V1	RNN	Fpz-Cz	Time series	20	83.9	77.1	-
Ref. [40]	Sleep-EDF-V1	CNN + LSTM	Fpz-Cz	Time series	20	83.9	78.0	40.0
Ref. [41]	Sleep-EDF-V1	1-max CNN	Fpz-Cz	Time-frequency image	20	82.6	76	29.9
Baseline	**Sleep-EDF-V1**	**CNN + LSTM**	**Fpz-Cz**	**Time series**	**20**	**85.4**	**79.9**	**33.6**
Baseline + CW (Ratio)	**Sleep-EDF-V1**	**CNN + LSTM**	**Fpz-Cz**	**Time series**	**20**	**87.3**	**82.7**	**42.6**

to the experimental results demonstrated in Table VI, we conclude some important findings. Firstly, it is essential to keep a sensible equilibrium between minority and majority classes, there is a trade-off between the overall accuracy and the recognition of the N1 stage on the CCSHS dataset, the RE improvement of the N1 stage is accompanied by the sacrifice of ACC and K. Secondly, even the same rule of relationship may result in different results on experimental datasets. The overall and N1 performance could be improved simultaneously on the Sleep-EDF and Sleep-EDF-V1 databases, but much lower enhancement of N1 stage than that on the CCSHS dataset. As mentioned in Sec. II. A, three experimental datasets comprise of subjects from different age groups (CCSHS: 16-19 years, Sleep-EDF: 25-101 years, Sleep-EDF-V1: 25-34 years).

In summary, the CW method is suitable for avoiding generating new EEG samples and keeping the dataset intact for retaining overnight sleep structure. In addition, when recognizing the N1 stage for diagnosing some related sleep disorders, the CW method is prone to show better performance (CCSHS dataset). If we prefer to enhance the performance of all stages and N1 simultaneously, the GWN method can improve the accuracy of the N1 stage without the sacrifice of overall accuracy. In this study, although the GAN model can enhance the overall accuracy, the stage N1 shows a slight drop in recall on three datasets.

V. CONCLUSION

In this study, we aim to deal with the widely existing class imbalance problem in the field of automatic sleep stage classification through balancing the dataset quantity and network connection. The attained results suggest that the proposed methods could make positive contribution to the improvement of biased performance. In most cases, the accuracies of N1 and whole stages are enhanced simultaneously on three public PSG datasets. In addition, our frameworks could outperform the state-of-the-art studies on the same dataset. This study paves new avenues for enhancing the sleep stage classification performance with class imbalance and monitoring the sleep equality and disorders. However, there are some aspects worthy of further exploration in future works. Firstly, more DA methods for balancing the dataset quantity could be investigated, such as the Variational Auto-Encoding network (VAE), which has obtained significant achievements in CV field. In terms of the imbalanced network connection, we will take into consideration of the activation function simulating the operation of neural's synapse for the duration of information processing procedures.

ACKNOWLEDGMENT

This study is to memorize Prof. Tapani Ristaniemi from University of Jyväskylä for his great help to the authors and Prof. Tapani Ristaniemi has supervised this study very much.

REFERENCES

[1] S. J. Redmond and C. Heneghan, "Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 3, pp. 485–496, 2006.

[2] G. Zhu, Y. Li, and P. Wen, "Analysis and classification of sleep stages based on difference visibility graphs from a single-channel eeg signal," *IEEE J. Biomed. Health. Inf.*, vol. 18, no. 6, pp. 1813–1821, 2014.

[3] H. Phan *et al.*, "Joint classification and prediction cnn framework for automatic sleep stage classification," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 5, pp. 1285–1296, 2018.

[4] H. Dong *et al.*, "Mixed neural network approach for temporal sleep stage classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 324–333, 2017.

[5] A. Rechtschaffen, "A manual of standardized terminology and scoring system for sleep stages of human subjects," *Electroencephalogr. Clin. Neurophysiol.*, vol. 26, no. 6, p. 644, 1969.

[6] C. Iber *et al.*, *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*, vol. 1. Westchester, IL, USA: Amer. Acad. Sleep Med., 2007.

[7] E. Alickovic and A. Subasi, "Ensemble svm method for automatic sleep stage classification," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 6, pp. 1258–1265, 2018.

[8] D. Silveira *et al.*, "Single-channel eeg sleep stage classification based on a streamlined set of statistical features in wavelet domain," *Med. Biol. Eng. Comput.*, vol. 55, no. 2, pp. 343–352, 2017.

[9] P. Memar and F. Faradji, "A novel multi-class eeg-based sleep stage classification system," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 1, pp. 84–95, 2017.

[10] R. Yan *et al.*, "Automatic sleep scoring: A deep learning architecture for multi-modality time series," *J. Neurosci. Methods.*, vol. 348, p. 108971, 2021.

[11] A. Supratak *et al.*, "Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 11, pp. 1998–2008, 2017.

[12] S. Chambon *et al.*, "A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 4, pp. 758–769, 2018.

[13] Q. Wei *et al.*, "A residual based attention model for eeg based sleep staging," *IEEE J. Biomed. Health. Inf.*, vol. 24, no. 10, pp. 2833–2843, 2020.

[14] R. Yan *et al.*, "A deep learning model for automatic sleep scoring using multimodality time series," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, pp. 1090–1094, 2021.

[15] B. Yang *et al.*, "A novel sleep stage contextual refinement algorithm leveraging conditional random fields," *IEEE Trans. Instrum. Meas.*, 2022.

[16] Y. Wei *et al.*, "Sleep stage transition dynamics reveal specific stage 2 vulnerability in insomnia," *Sleep.*, vol. 40, no. 9, 2017.

[17] D. Shrivastava *et al.*, "How to interpret the results of a sleep study," *J. Community Hosp. Intern. Med. Perspect.*, vol. 4, no. 5, p. 24983, 2014.

[18] T. Nakamura *et al.*, "Automatic detection of drowsiness using in-ear eeg," in *Proc Int Jt Conf Neural Netw (IJCNN).*, pp. 1–6, IEEE, 2018.

[19] C. Sun *et al.*, "A two-stage neural network for sleep stage classification based on feature learning, sequence learning, and data augmentation," *IEEE Access.*, vol. 7, pp. 109386–109397, 2019.

[20] O. Tsinalis, P. M. Matthews, and Y. Guo, "Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders," *Ann Biomed Eng.*, vol. 44, no. 5, pp. 1587–1597, 2016.

[21] J. Fan *et al.*, "Eeg data augmentation: towards class imbalance problem in sleep staging tasks," *J. Neural Eng.*, vol. 17, no. 5, p. 056017, 2020.

[22] G. Zhang *et al.*, "The national sleep research resource: towards a sleep data commons," *J. Am. Med. Inform. Assoc.*, vol. 25, no. 10, pp. 1351–1358, 2018.

[23] C. L. Rosen *et al.*, "Prevalence and risk factors for sleep-disordered breathing in 8-to 11-year-old children: association with race and prematurity," *J. Pediatr.*, vol. 142, no. 4, pp. 383–389, 2003.

[24] B. Kemp *et al.*, "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 9, pp. 1185–1194, 2000.

[25] D. Zhou *et al.*, "Singlechannelnet: A model for automatic sleep stage classification with raw single-channel eeg," *Biomed. Signal Process. Control.*, vol. 75, p. 103592, 2022.

[26] B. A. Edwards *et al.*, "Aging and sleep: physiology and pathophysiology," in *Semin Respir Crit Care Med.*, vol. 31, pp. 618–633, 2010.

[27] V. Krishnan and N. A. Collop, "Gender differences in sleep disorders," *Curr Opin Pulm Med.*, vol. 12, no. 6, pp. 383–389, 2006.

[28] I. Gulrajani *et al.*, "Improved training of wasserstein gans," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[29] Y. Zeng, T. Zhang, and B. Xu, "Improving multi-layer spiking neural networks by incorporating brain-inspired rules," *Sci. China Inf. Sci.*, vol. 60, no. 5, pp. 1–11, 2017.

[30] D. J. Heeger and D. Ress, "What does fmri tell us about neuronal activity?," *Nat. Rev. Neurosci.*, vol. 3, no. 2, pp. 142–151, 2002.

[31] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit (CVPR).*, pp. 1–9, IEEE, 2015.

[32] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv Prepr. arXiv:1312.4400.*, 2013.

[33] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Conf. Artif. Intell. Stat (AISTATS)*, pp. 249–256, 2010.

[34] T. Nakamura, H. J. Davies, and D. P. Mandic, "Scalable automatic sleep staging in the era of big data," in *Proc. IEEE Eng. Med. Biol. Soc (EMBC).*, pp. 2265–2268, 2019.

[35] X. Li *et al.*, "Hyclasss: a hybrid classifier for automatic sleep stage scoring," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 2, pp. 375–385, 2017.

[36] S. Mousavi, F. Afghah, and U. R. Acharya, "Sleepeegnet: Automated sleep stage scoring with sequence to sequence deep learning approach," *PLOS ONE.*, vol. 14, no. 5, pp. 1–15, 2019.

[37] A. Supratak and Y. Guo, "TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel eeg," in *Proc. IEEE Eng. Med. Biol. Soc (EMBC).*, pp. 641–644, 2020.

[38] H. Phan *et al.*, "Xsleepnet: Multi-view sequential model for automatic sleep staging," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[39] D. Zhou *et al.*, "Lightsleepnet: A lightweight deep model for rapid sleep stage classification with spectrograms," in *Proc. IEEE Eng. Med. Biol. Soc (EMBC).*, pp. 43–46, 2021.

[40] H. Seo *et al.*, "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.*, vol. 61, p. 102037, 2020.

[41] H. Phan *et al.*, "Dnn filter bank improves 1-max pooling cnn for single-channel eeg automatic sleep stage classification," in *Proc. IEEE Eng. Med. Biol. Soc (EMBC).*, pp. 453–456, 2018.

[42] B. A. Mander, J. R. Winer, and M. P. Walker, "Sleep and human aging," *Neuron.*, vol. 94, no. 1, pp. 19–36, 2017.

Dongdong Zhou received the B.S. and M.S. degrees in biomedical engineering from Dalian University of Technology, Dalian, China, in 2015, and 2018, respectively. Supported by China Scholarship Council through the Dalian University of Technology, he is currently pursuing the Ph.D. degree in software and communications engineering with the University of Jyväskylä, Jyväskylä, Finland.

His research interests include biomedical signal processing, sleep analysis, deep learning.

Qi Xu received the B.S. degree from the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China, in 2015, and the Ph.D. degree from the College of Computer Science and Technology, Zhejiang University, Hangzhou, China, in 2021. He was ever granted as the honorary Visiting Fellow of the Centre for Systems Neuroscience, University of Leicester, Leicester, U.K., in 2019.

He is curently a tenure-track Associate Professor with the School of Artificial Intelligence, Dalian University of Technology, Dalian, China. His research interests include brain-inspired computing, neuromorphic computing, neural computation, computational neuroscience, biomedical signal processing, sleep analysis and cyborg intelligence.

Jian Wang received the B.S. degree in Department of Japanese from Dalian Neusoft University of Information, Dalian, China in 2011 and the M.S. degree in School of Kinesiology and Health Promotion from Dalian University of Technology, Dalian, China, in 2017. She is currently pursuing the Ph.D. degree in biomedical engineering from Dalian University of Technology, Dalian, China and the visiting doctor of Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland.

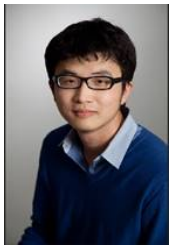
Her research interests include biomedical signal processing, rehabilitation.

Hongming Xu received B.S. and M.S. degrees from College of Inormation Engineering at Northwest A&F University, Yangling, China, in 2009 and 2012, respectively. He received his Ph.D. degree in Department of Electrical and Computer Engineering at University of Alberta in Edmonton, Canada, in 2017.

He is currently an Associate Professor in the School of Biomedical Engineering at Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, China. His research focuses on artificial intelligence in biomedical imaging fields (particularly Pathology AI), medical image computing, imaging informatics, and machine learning.

Lauri Kettunen received the Ph.D. degree in electrical engineering from the Tampere University of Technology, Tampere, Finland, in 1992.

He is currently a Professor of Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland. His research interests include mathematical physics, category theory, boundary value problems.



Zheng Chang (Senior Member, IEEE) received Ph.D. degree from the University of Jyväskylä, Jyväskylä, Finland in 2013. He has published over 130 papers in journals and conferences, and received best paper awards from IEEE TCGCC and APCC in 2017 and has been awarded as the 2018 IEEE Best Young Research Professional for EMEA and 2021 IEEE MMTC Outstanding Young Researcher. He has been served as symposium chair, publicity chair and workshop chair and also participated in organizing workshops and special sessions for many IEEE flagship conferences, such as Infocom, ICC and Globecom. He is an editor of Springer Wireless Networks, International Journal of Distributed Sensor Networks, and IEEE Wireless Communications Letters. He was the exemplary reviewer of IEEE Wireless Communication Letters in 2018. He also acts as a guest editor of IEEE Communications Magazine, IEEE Wireless Communications, IEEE Networks, IEEE Internet of Things Journal and IEEE Transactions on Industrial Informatics. His research interests include IoT, cloud/edge computing, security and privacy, vehicular networks, and green communications.

Fengyu Cong (Senior Member, IEEE) received the B.S. degree in power and thermal dynamic engineering and the Ph.D. degree in mechanical design and theory from Shanghai Jiao Tong University, Shanghai, China, in 2002 and 2007, respectively, and the Ph.D. degree in mathematical information technology from the University of Jyväskylä, Jyväskylä, Finland, in 2010.

He is currently a Professor with the School of Biomedical Engineering, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, China and the visiting Professor with Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland. His current research interests include brain signal processing, acoustic signal processing, independent component analysis, tensor decomposition, and pattern recognition/machine learning/data mining.