Adaptive Margin Aware Complement-Cross Entropy Loss for Improving Class Imbalance in Multi-View Sleep Staging Based on EEG Signals

Fahui Miao, Li Yao, and Xiaojie Zhao

Abstract—Sleep data are typically characterized by class imbalance, which can cause the model to be overly biased toward frequent classes, resulting in low accuracy of minority class classification. However, the minority class of sleep staging has important value in diagnosing certain disorders, such as an N1 Stage that is too short indicating possible hypsomnla or narcolepsy. To address this problem, we propose a multi-view CNN model based on adaptive margin-aware loss. A novel margin-aware factor that considers the relative sample sizes of both frequent and minority classes can improve the overfitting of minority classes by increasing the regularization strength of minority classes without changing the sample size to maximize the decision margins of minority classes. On this basis, we propose margin-aware complement entropy and margin-aware cross-entropy loss, respectively. Margin-aware complement entropy can be achieved to increase the regularization for minority classes while neutralizing errors for minority classes, thus improving the classification accuracy for minority classes. Finally, the synergy of margin-aware complement entropy and cross-entropy is realized in an adaptive way to improve the sleep staging classification accuracy. We tested on three sleep datasets and compared them with the state-of-the-art, and the results demonstrate that our proposed algorithm not only improves the accuracy of sleep staging in general, but also improves the minority classes to a greater extent.

Index Terms—Sleep staging, class imbalance, margin awareness, complement-cross entropy, multi-view CNN.

I. INTRODUCTION

Sleep is a very important physiological process for human beings, and the quality of sleep has a great impact on physical and mental health [1]. Deep learning not only enables end-to-end learning, but can also remain free from prior knowledge, so sleep staging research based on deep learning has become a hot topic [2]. The main existing deep learning frameworks used in the field of sleep staging are Deep Belief Networks (DBNs) [3], Auto-encoders [4], Deep Neural Networks (DNNs) [5], Convolutional Neural Networks (CNNs) [6] and Recurrent Neural Networks (RNNs) [7], as well as their hybrid structures such as CNN+RNN [8], CNN+LSTM [9] and RNN+LSTM [10]. Numerous studies [6], [11], [12] have shown that CNNs are still the most important network type in sleep staging. In addition, networks in the sleep domain are mostly built in a serial or parallel manner. They can be divided into two categories. One [8], [9], [13] class uses different network architectures to mine different feature representations of the same data. The other uses different scales of the same network architecture or different types of inputs to mine different scales of information [14], [15], [16]. Compared to two-stream networks using different network types, two-stream networks using the same architecture are simpler. However, most existing neural networks [2], [6], [9], [10] in the field of sleep staging use single-view data from the same data source as input, and lack information mining of data from different views. Usually, feature representations extracted from multi-view data are complementary in nature [18]. Therefore, we automatically constructed a 1D CNN containing the original EEG sequence signals and a 2D CNN architecture with EEG spectral images, respectively, using the previously proposed CNN architecture automatic search algorithm [19].

Although the two-stream network with multiple views can improve the mining ability of the data, it also causes the degradation of the model performance when encountering the data with class imbalance. Sleep datasets are typically classically imbalanced datasets due to the different durations of each stage of sleep. In all stages of sleep, the N2 stage usually occupies 45%-55% of the entire sleep stage, while the N1 stage only accounts for 2%-5%[20]. Only a few sleep staging studies have improved the class imbalance. The available studies fall into three main categories. One category improves...
the class imbalance by means of reweighting. Emamelddeen [21] et al. proposed a class-sensitive weighted loss function for improving the sleep EEG staging class imbalance problem. Re-weighting may make model optimization unstable and difficult [22]. The second category improves the class imbalance problem by oversampling the minority class. Supratak et al. [9] proposed balancing the sample classes by replicating the minority class samples to increase the sample size of that class. Chawla et al. used the synthetic minority oversampling technique (SMOTE) [23] based on synthetic minority classes to improve the class imbalance. Phan et al. [15] trained the network by random sampling so that each batch sample category was balanced. However, by oversampling a minority of classes may cause overfitting of that class [23]. The third class improves the class imbalance problem by augmenting the minority class data, e.g., Sun et al. [24] proposed augmenting the minority class by geometric transformations such as flips and translations. F. Lotte [25] proposed amplification of the minority class samples using segmentation and recombination of the original signal. J. Fan et al. [26] generated the minority class of samples by using adversarial generative networks (GANs). However, sleep EEG is time-series data, and such an expansion may change the original time-dependent information of the signal.

Resampling and data expansion techniques are most commonly used in sleep staging to improve the overfitting problem in the minority class by changing the number of samples. Margin loss [27], [28] does not require changing the number of samples, but improves the final classification accuracy by making the classifier have the maximum classification margin: that is, making each class have a larger interval from the decision boundary, thus improving the overfitting problem. The large-margin softmax [29] and angular softmax [30] expand the class margins by minimizing the intraclass distance and maximizing the interclass distance. However, the classes in these methods are independent of each other, resulting in equal margins for each final class from the decision boundary, which is not more favorable to enhance the overfitting problem of the minority class. Cao et al. [31] proposed a method called label-distribution-aware margin (LDAM) loss to improve the overfitting problem for the minority class. LDAM is a way to improve the generalization error of minority classes from the perspective of applying stronger regularization to minority classes, which improves the classification accuracy of minority classes on the test set without reducing the model fitting ability. The margin of each class in LDAM depends on the derivative of its own sample size, resulting in a larger decision margin for the minority class. However, LDAM ignores the fact that the imbalance rate between minority and frequent classes may reduce the probability of obtaining larger margins for minority classes.

In addition, Chen proposed complement entropy loss [32], which suppresses misclassification and achieves error neutralization by exploiting the complement information of samples in this class, thus improving the confidence level. Because the minority class is richer in error information, Kim et al. [33] used complement entropy to improve the class imbalance problem. However, the existing complement entropy loss only considers the improvement of accuracy from the perspective of neutralizing the error information, and the available training samples are still insufficient for a very small percentage of minority class samples, which can limit the effectiveness of complement entropy loss for models in performing error neutralization of minority classes. Therefore, we introduce the proposed margin-aware factor into the complement entropy loss to further improve the classification accuracy of complement entropy loss for class imbalance data from the perspective of increasing the regularization of the minority class.

To avoid the introduction of unknown noise by generating new simulated samples and the possible instability of model fitting by reweighting techniques, this paper proposes a novel approach to improve the sleep staging class imbalance problem. We enhance the minority class regularization strength and error suppression of complement information by increasing the decision margins of minority classes and by learning the wrong prediction information, thus, ultimately, improving the model's ability to identify sleeping minority classes without adding new samples.

II. PROPOSED METHOD

In this paper, we propose an adaptive margin aware complement-cross entropy loss based multi-view CNN (AMC-MVCNN), which is shown in Fig. 1.

A. Architecture of the Multi-View CNN

To more fully exploit the information contained in the sleep data, we also propose a multi-view CNN (MVCNN) architecture. Data from different views are obtained with complementary feature representations [18], which helps to indirectly improve the learning of the model for the minority class. The 1D CNN and 2D CNN of the architecture in this paper are derived from our previous architecture search algorithm SOSCNN [19]. The SOSCNN algorithm is driven by the sleep EEG data and automatically designs a network architecture suitable for sleep staging. In this study, two CNN network architectures with different dimensions are combined to construct the MVCNN. The MVCNN architecture is divided into 1-dimensional convolutional and 2-dimensional convolutional branches, as shown in Fig. 1. The 1-dimensional convolutional branch of the network contains 7 layers, of which layers 1, 2, 4 and 6 are convolutional layers with the number of neurons and convolutional kernel size being (57@21×1, 211@21×1, 154@58×1, 140@69×1), respectively, and the step size is 3. Both layer 3 and layer 5 are average pooling layers with a pooling window of 3×1 and a step size of 2. Layer 7 is a flatten layer. The network’s 2-dimensional convolutional branch of contains a total of 7 layers, of which layers 1, 2, 4, 5 and 6 are all convolutional layers, and their number of neurons and convolutional kernel size are (46@5×5, 187@6×6, 68@7×7, 68@7×7, 254@7×7), respectively, while the step size is 1. Layer 3 is the average pooling layer with a pooling window of 3×3 and a step size of 2. Layer 7 is the flatten layer. Each convolutional layer is followed by an activation layer (ReLU) and a batch normalization layer.

\[
\mathbf{S} = \left\{ \left( S^{(1)}_1, S^{(2)}_2, y_1 \right), \ldots, \left( S^{(1)}_N, S^{(2)}_N, y_N \right) \right\}
\]

is a dataset with sample size \( N \). \( S^{(1)}_i \) denotes the \( i \)-th sample with the original signal data type. \( S^{(2)}_i \) denotes the \( i \)-th sample with a
time-frequency image data type. \( y_i \) denotes the label of the \( i \)-th sample one-hot type. The output of the CNN feature extraction layer for both branches of MVCNN is shown below.

\[
\begin{align*}
\mathbf{o}_1^{(1)} & = 1D - \text{CNN} \left( F_1 \left( S_1^{(1)} \right), \ldots, F_1 \left( S_N^{(1)} \right) \right) \\
\mathbf{o}_2^{(1)} & = 2D - \text{CNN} \left( F_2 \left( S_1^{(2)} \right), \ldots, F_2 \left( S_N^{(2)} \right) \right)
\end{align*}
\]

where \( F_1 \) and \( F_2 \) denote the feature extraction process after 1-dimensional CNN and 2-dimensional CNN, respectively. The representations extracted from the 1D CNN and 2D CNN in MVCNN are subjected to the flatten operation, and then the outputs of each branch are directly spliced to obtain the joint sleep EEG representations. The process of obtaining the joint representation is shown in Equation (3).

\[
\mathbf{o} = \left( \mathbf{o}_1^{(1)}, \ldots, \mathbf{o}_N^{(1)} \right)
\]

The MVCNN takes the joint representation \( \mathbf{o} \) through the fully connected layer and uses the softmax function to calculate the predicted probability \( y' \) of the input samples. The classification error between the predicted values and the true labels is calculated with the LAMCE loss function proposed in this study. The final task of staging sleep EEG is accomplished by minimizing the classification error and performing backpropagation to update the MVCNN network weights. The weight update can be expressed by the following Equation (4).

\[
\hat{\theta} = \arg \min_{\theta} \mathcal{L}_{\text{AMCE}}(y, y')
\]

B. Margin Aware Factor

Margin loss [29], [30], [36], [37], [38] is a classification method that obtains the maximum margin. In this class of methods, the larger the sample distance from the decision margin is, the higher the confidence of the classification. The core of the margin-aware factor is to improve the model generalization ability by a stronger regularization of the minority class based on the ratio of the sample sizes between the majority class and the sample being classified, thus improving the classification accuracy of the minority class for the test set. The stronger regularization of the minority class is achieved by larger margins, and the process of such adjustment requires the computation of fine-grained upper bounds on the generalization error. The error for each class in the classification task can be defined by the following process[31]. We use \( \mathbb{R}^d \) to denote the input space, \( \{1, \ldots, k\} \) to denote the real category label space, and \( x \) and \( y \) to denote the input data and their corresponding labels, respectively. Assuming that the class conditional probability distributions \( P(x|y) \) of the training and test datasets are the same, we denote the conditional distribution of class \( j \) by \( P_j = P(x|y = j) \). We first define the relevant modeling on the balanced dataset and later extend it to the class imbalance dataset. We denote by \( L_{bal}[\text{MVCNN}] \) the
standard 0-1 test error calculated by MVCNN for the balanced data distribution, which can be expressed by the following equation.

\[
L_{bal}[MVCNN] = \Pr_{(x,y)\sim P_{bal}}[MVCNN(x),y < \max MVCNN(x)]
\] (5)

Similarly, the error of the \(j\)th category can be expressed by the following equation.

\[
L_j[MVCNN] = \Pr_{(x,y)\sim P_j}[MVCNN(x),y < \max MVCNN(x)]
\] (6)

We use \([x_i,y_i]_{i=1}^n\) to denote the training dataset and \(n_j\) to denote the number of samples in class \(j\). The margins of the sample \((x,y)\) can be defined as follows.

\[
\gamma(x,y) = MVCNN(x) - \max_{j \neq y} MVCNN(x)
\] (7)

The training margin of class \(j\) is defined as follows.

\[
\gamma_j = \min_{i \in S_j} \gamma(x_i,y_i)
\] (8)

We denote by \(\mathcal{G}\) the family of hypothetically classes in the dataset, the complexity of class \(\mathcal{G}\) by \(C(\mathcal{G})\), and \(\gamma_{\text{min}} = \min\{\gamma_1, \cdots, \gamma_k\}\) the minimum margin of all classes. The typical generalization error bound when the training set data are codistributed with the test set data is denoted by \(C(\mathcal{G})/\sqrt{n}\). Then, the imbalance generalization error for the training and test sets with the same class imbalance distribution can be expressed as follows.

\[
\text{imbalanced test error} \lesssim \frac{1}{\gamma_{\text{min}}} \sqrt{\frac{C(\mathcal{G})}{n}}
\] (9)

Since the training data have a high probability \((1 - n^{-5})\) of randomness, the fine-grained error for each class, such as the \(j\)th class error, can be reduced to \(L_j\).

\[
L_j[MVCNN] \lesssim \frac{1}{\gamma_{\text{min}}} \sqrt{\frac{C(\mathcal{G})}{n}} + \log n / \sqrt{n}
\] (10)

From the generalization error formula of each class, it is known that the generalization error \(L\) of the minority class is reduced by increasing its margin. The margin modification of the original LDAM [31] only considers the number of samples themselves, and ignores the relative relationship between the numbers of samples in the minority class and the frequent class, although this relative relationship directly affects the performance of the model. We propose a margin aware factor \(\Delta m\), which is considers both the number of samples themselves and the number of most frequent samples. In Fig. 2(a), a schematic diagram of margin modification with a margin aware factor is shown. In Fig. 2(a), \(l_1\) is the original decision margin without any adjustment, which is equidistant between the two classes. \(l_2\) is the decision margin considering only the number of samples of the class itself, \(l_3\) is the decision margin adjusted by the margin-aware factor, and \(\gamma_1, \gamma_2, \gamma_3, \gamma_4\) indicates the distance between the two classes and the different decision margins. Our proposed margin aware factor accounts for the amount of difference between the minority class and the frequent class samples; when the data are more imbalanced, the more the minority class is encouraged to obtain larger margins. The model of \(\Delta m\) is shown in Equation (11).

\[
\Delta m = \frac{C}{n_i \log(n_{\text{max}} - (n_i / n_{\text{max}}) - (n_{\text{max}} - n_i))} \quad \text{for } i \in \{1, \ldots, k\}
\] (11)

where \(n_{\text{max}}\) is the number of frequent class samples, \(n_i\) is the number of class \(i\) samples, and \(k\) is the number of classes in the dataset. \(C\) is a hyperparameter to be tuned. In this paper, \(C\) is set to 1. Fig. 2(b) shows the effect of the numbers of minority and frequent class samples on the margin-aware factor. The variations of the margin-aware factor with the expanding number of samples within the majority class. When the number of minority class samples is fixed (e.g., \(n=291\)), the presence of more samples in the frequent class leads to more imbalanced the sample size in the dataset and greater adjustment of the margin-aware factor of the minority class so that the minority class obtains a larger margin and thus a greater degree of correctness and certainty.

The mathematical model of the modified margin-aware cross entropy loss \(L_{MCE}\) is shown in Equation (12).

\[
L_{MCE} = -\frac{1}{N} \sum\limits_{i=1}^{N} \sum\limits_{j=1}^{C} \log \left( \frac{e^{z_{ij} - \Delta m}}{e^{z_{ij} - \Delta m} + \sum_{k=1,k\neq j}^{C} e^{z_{ik}}} \right)
\] (12)

where \(N\) is the number of samples in a batch and \(C\) is the number of classes. \(z_{ij}\) denotes the output data corresponding
to the real class \( j \) in the network. \( z_k \) denotes the output data corresponding to class \( k \), \( k \neq j \) in the network.

C. Margin-Aware Complement Entropy

The purpose of complement entropy loss is to utilize the valuable complementary information provided by the misclassified information. Complement entropy loss is used to indirectly increase the probability of correct classification through error suppression of the complementary information provided by the error probabilities of all class samples. When classifying a sample with class imbalance, it has a higher probability of being misclassified due to the overfitting of the minority class. Since the error probability value of the minority class is larger than that of the frequent class, the suppression of error information is stronger for the minority class than for the frequent class when error neutralization is performed. Therefore, complement entropy loss [32] is introduced to achieve the utilization of error information. Although complement entropy loss improves the classification accuracy through the use of complementary information, the minority class overfitting problem still occurs in the face of insufficient minority class training samples.

The original complement entropy loss is computed to improve the probability of correct classification by suppressing misinformation only. The overfitting of minority classes in the class imbalance problem can be further improved if the regularization of minority classes is also increased from the perspective of regularization. Therefore, this paper modifies the margins of the samples in the complementary entropy by the margin-aware factor to make the modified loss more favorable to the data with class imbalance. We name the improved complement entropy loss as the margin-aware complement entropy \( \mathcal{L}_{MCOE} \). The process of \( \mathcal{L}_{MCOE} \) is shown in Fig. 3.

The mathematical model of margin aware complement entropy is shown in Eqs. (13-15).

\[
\mathcal{L}_{MCOE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1, j \neq g}^{C} P_i \log(P_i) 
\]

\[
P_i = \frac{e^{z_{ig}}}{e^{z_{ig}} - \Delta m + \sum_{j=1, j \neq g}^{C} e^{z_{ij}}}
\]

\[
\Delta m = \frac{1}{n_i \log((n_{max}/(n_{max} - n_i ) + (n_i / n_{max})))} \text{ for } i \in {1, \ldots, k}
\]

where \( N \) is the number of samples in a batch and \( C \) is the number of classes. \( P_i \) is the complementary information calculated for the \( i \)-th sample using the error probability. \( g \) is the subscript of the correct class. \( z_{ij} \) is the output data of the \( i \)-th sample for class \( j \). \( \Delta m \) is the proposed margin aware factor that relies on both the number of frequent class samples and its own sample size. \( e^{z_{ig}} - \Delta m \) denotes the information that the \( i \)-th sample is of class \( g \) and is corrected by margin aware \( \Delta m \).

D. Adaptive Margin Aware Complement-Cross Entropy

We propose adaptive margin aware complement-cross entropy loss to achieve synergy between margin aware complement entropy and margin aware cross entropy. In margin aware cross entropy (MCE) loss, we focus on learning the correct class information to guide the training of the model, while in margin aware complement entropy (MCOE) loss, we focus more on learning the complementary error information. Therefore, we use an incremental function \( f(x) \in (0, 1) \) to achieve adaptive synergy between MCE and MCOE. \( t \) is the horizontal coordinate translation, which is set to 50. \( \lambda \) is the parameter to be adjusted and is set to 1 in this paper. In the early stages of training, the model focuses more on learning the correct samples to actively improve the prediction probability for the correct class. As the number of training sessions increases, the final prediction probability is indirectly improved by increasing the learning for incorrect information. The mathematical model of the adaptive margin aware complement-cross entropy loss is shown in Eqs. (16-17).

\[
f(x) = \frac{\lambda}{1 + e^{-\frac{x-t}{10}}}
\]

\[
\mathcal{L}_{AMCE} = \mathcal{L}_{MCE} + f(x) \times \mathcal{L}_{MCOE}
\]

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. Datasets

We used three different sleep staging benchmark datasets to validate the performance of the model. The percentage of data in each sleep stage is described in detail in Table I and Table II.

1) DRM-SUB [39]: This dataset consists of 20 full-night sleep recordings from people aged between 20 and 65 years. Twenty subjects comprised 4 males and 16 females. They were all healthy individuals without sleep-related disorders. The data were sampled at a frequency of 200 Hz. Since the sleep EEG from the Cz-A1 channel possesses a better quality of sleep staging [40], the EEG signal from this channel was used in this paper to evaluate the model. On the DRM-SUB dataset, the AAMS [41] staging criteria were used to classify sleep staging into five stages: W, N1, N2, N3 and R.
2) **SleepEDF-20 [42]**: This dataset consists of data from 10 male and 10 female subjects between the ages of 24 and 35. Due to equipment problems, data were collected for only one night for Subject 13; data were collected for two nights for all subjects except other subjects. The collected data were divided into 30-second increments, and then sleep experts manually classified the segments into eight categories (W, N1, N2, N3, N4, R, MOVEMENT, UNKNOWN). N3 and N4 were combined into N3 according to the AAMS staging criteria[43]. In addition, we removed MOVEMENT and UNKNOWN in the preprocessing process.

3) **SleepEDF-ST[44]**: This dataset is a subset of the SleepEDF extended dataset. The sleep telemetry subset is composed of 22 Caucasian subjects, all of whom have mild sleep disorders. These subjects were between 18 and 79 years of age, with 7 males and 15 females. The sampling frequency of this dataset was 100 Hz, and this dataset was labeled according to the R&K sleep staging criterion.

The experiment was divided into three parts. First, the performance of our algorithm was analyzed in comparison with the state-of-the-art for sleep staging with single-channel EEG. To ensure the accuracy of the comparison experiments, the results of the comparison experiments between this paper and other algorithms for DRM-SUB and SleepEDF-20 are derived from the data provided in published literature. In addition, we performed ablation experiments on the proposed algorithm in this paper, which were used to verify the performance of each step of the proposed algorithm improvement. Finally, the classification performance of the AMC-MVCNN was tested under the R&K staging standard. In this paper, the average classification accuracy, Cohen's kappa coefficient (kappa), precision, sensitivity, the per-class F1-score and the macroaveraging F1-score were used to measure the model’s sleep staging performance.

### B. Comparison of AMC-MVCNN and State-of-the-Art

We compared the overall performance of the AMC-MVCNN with other superior algorithms for sleep staging with single-channel EEG. Table III shows the comparison results based on the DRM-SUB dataset. The average classification accuracy of our model is improved by 8.46% when compared to the model that also uses only the CNN structure. Our average accuracy is also improved by 3.91% compared to the improved CNN-HMM. This result indicates that our adjustments improve the accuracy of the algorithm when using the same type of network model. Our model also improved by 2.64%-3.85% compared to the improved version using only LSTM or LSTM. This result indicates that our improved model has stronger sleep staging capability than the LSTM. We also compared our model with the hybrid class model from CNN-LSTM, and our model improved the classification accuracy by 3.39%-3.43% on average. This result indicates that our improved multi-view CNN model achieves higher classification performance than the hybrid model on DRM-SUB. Compared with AMC-MVCNN, EEMD+RUSBoost uses a resampling technique to equalize the training samples to improve the class imbalance, which may change the original EEG timing distribution and increase the memory overhead. In addition, only two CNN models with different dimensions are used in AMC-MVCNN, which is less complex, less difficult to fit, and can effectively improve the recognition of minority classes compared with other methods that improve the classification accuracy based on the structure of mixing multiple models. When the test set includes an imbalanced class, the average classification accuracy may be too biased toward frequent classes to fully reflect the performance of the classification algorithm, so we also compare it from a consistency perspective. Our model has a larger Kappa value, which indicates that our algorithm achieves a stronger classification performance for class imbalanced data.

To validate the classification performance of our model for each sleep stage on the DRM-SUB dataset in more detail, we compared the results of staging for each stage of sleep from both precision and sensitivity perspectives with results from other state-of-the-art approaches, and the detailed data are shown in Table IV. In terms of precision, our model achieves the highest precision in all four sleep stages, W, N1, N3 and R, and the percentages of sleep stages in the DRM-SUB dataset are ranked as N1 < W < N3 < N2. Our model precision improves for all four phases of N1, R, W and N3 when compared with the optimal values produced by the comparison models, and the highest improvement is 9.33% in the N1 phase, which accounts for the least amount of data, and 4.95% in the R phase, which accounts for the second-least amount of data. The improvement in the N3 and W stages was 1.05%-1.54%. In the class with the largest sample size, N2 is 0.62% lower than the optimal value. This indicates that our model effectively improves the precision of the minority class. Based on the analysis of sensitivity, our model has the highest sensitivity for all four phases of W, N1, N2 and R. The maximum improvement is 3.23% in phase W. In phase N3, our model sensitivity is reduced by 0.7% compared with the optimal value from the comparison model. Overall, our model improved in the three sleep stages N1, R and W with smaller sample sizes, both in terms of precision and sensitivity analysis. This indicates that our model has better classification performance for the minority classes in the DRM-SUB dataset.

To more visually compare the performance of our model with the overall state-of-the-art classification for the 5 sleep stages, we compared the F1-score values of the models. The overall comparison results are shown in Fig. 4. Our model’s F1-score achieves the highest values for all 5 sleep stages, indicating that our model offers a superior sleep staging capability.

To verify the generalization ability of the model on different datasets, we also compared AMC-MVCNN with the state of the art on the SleepEDF-20 dataset. Table V presents the model comparison results on the SleepEDF-20 dataset. In terms of average classification accuracy, the average classification accuracy of our model improves by 2.52%-5.52% compared to the model that also uses only the CNN structure. Our average accuracy is also improved by 0.44%-1.62 compared to the improved CNN. This shows that our improvement enhances the average classification accuracy of the model.
when using the same type of network model. Our model also improved by 2.22%-5.32% compared to using only RNN models. This indicates that our improved model has a stronger sleep staging capability than RNN. We also compared a hybrid class of models consisting of multiple network models, and our models improved the average classification accuracy by 0.52%-3.12%. This indicates that the AMC-MVCNN model achieves higher classification performance than the hybrid model on SleepEDF-20. The AMC-MVCNN does not need to generate new simulation samples and does not change the original EEG temporality compared with the methods CNN+Attention, Auto-encoder, Deep auto-encoder,

### Table I

| Dataset   | Total epochs | W (%) | N1 (%) | N2 (%) | N3 (%) | R (%) |
|-----------|--------------|-------|--------|--------|--------|-------|
| DRM-SUB   | 20254        | 17.6  | 7.3    | 40.7   | 19.4   | 14.9  |
| SleepEDF-20 | 38150      | 11.82 | 7.24   | 46.07  | 14.66  | 20.21 |

### Table II

| Dataset   | Total epochs | W (%) | S1 (%) | S2 (%) | S3 (%) | S4 (%) | REM (%) |
|-----------|--------------|-------|--------|--------|--------|--------|---------|
| SleepEDF-ST | 21071       | 10.72 | 9.7    | 45.05  | 8.09   | 6.83   | 19.61   |

### Table III

**Comparison of the Average Classification Accuracy and Kappa Values of Our Model With State-of-the-Art on the DRM-SUB Dataset. The '-' in the Table Indicates That No Relevant Data Are Provided in the Literature**

| DataSet       | Architecture     | Channel | Fold | Overall performance | Acc (%) | Kappa |
|---------------|------------------|---------|------|---------------------|---------|-------|
| DRM-SUB       | EEMD+RUSBoost    | Cz-A1   | 10   | 74.59               | 0.738   |
|               | IMBED+ Bagged Trees | Cz-A1   | 10   | 79.90               | -       |
|               | RF+HMM           | Cz-A1   | 2    | 80.35               | 0.73    |
|               | 1D-CNN-HMM       | Cz-A1   | 20   | 81.68               | 0.74    |
|               | CNN-BiLSTM       | Cz-A1   | 5    | 80.83               | 0.738   |
|               | CNN-LSTM         | Cz-A1   | 5    | 80.87               | 0.736   |
|               | BiLSTM           | Cz-A1   | 5    | 80.41               | 0.733   |
|               | LSTM             | Cz-A1   | 5    | 80.53               | 0.733   |
|               | CNN              | Cz-A1   | 5    | 75.82               | 0.664   |
|               | AT-BiLSTM        | Cz-A1   | 5    | 81.62               | 0.749   |
| **Ours**      |                  | Cz-A1   | 20   | **84.26**           | **0.784** |

### Table IV

The Results of Our Model With Other State-of-the-Art Models on DRM-SUB Dataset in Terms of Precision and Sensitivity. Ours - Comparisons (Best) Indicate the Amount of Difference Between Our Model and the Best Value in the Comparison Model.

| DataSet       | Networks       | Fold | Precision (%) | Sensitivity (%) |
|---------------|----------------|------|---------------|-----------------|
| DRM-SUB       | AT-BiLSTM      | 5    | **84.98**     | **86.28**       |
|               | CNN            | 5    | 83.68         | 85.36           |
|               | LSTM           | 5    | 87.24         | 86.12           |
|               | BiLSTM         | 5    | 86.46         | 85.36           |
|               | CNN-LSTM       | 5    | 84.34         | 84.26           |
|               | CNN-BiLSTM     | 5    | 86.22         | 84.26           |
| **Ours**      | 20             | **89.07** | **91.12** | **85.58**       |
| Ours - Comparisons (best) | 1.05 | 3.23 | 83.24 | 83.24 |
CNN + RNN and 1-max CNN, which use class imbalance processing techniques. In addition, the AMC-MVCNN model is simple and improves the classification performance only through the improvement of the loss function, the model fitting cost is lower compared with the models of other hybrid architectures among the comparison methods, and the AMC-MVCNN can effectively improve the recognition of minority classes in sleep staging. We also compared models in terms of consistency. Our model obtains a larger Kappa value, indicating that AMC-MVCNN achieves stronger classification performance for class imbalanced data.

Similarly, for the SleepEDF-20 dataset, we compared the classification performance of AMC-MVCNN with the state of the art for each sleep stage. Table VI shows the comparison results of the models on the SleepEDF-20 dataset. Based on analysis of the precision values, our model achieved the highest values in three sleep stages N1, N3 and R. In the N1 stage, in which the sample size was the smallest, the precision improved by 9.47% over the optimal values from other state-of-the-art models. This was slightly worse than the comparison model in the W and N2 stages. Analyzed from the perspective of the sensitivity, our model obtained the highest values in the three sleep stages N1, N2 and R. The sensitivity in the N1 stage, where the sample size is the smallest, improved by 7.12% compared to the optimal value in the comparison model. The values of the sensitivity in the W and N3 stages were slightly worse than those of the comparison model. Overall, from both precision and sensitivity perspectives, AMC-MVCNN gains improvement in the N1 phase with the smallest sample size, and the amount of improvement in classification performance during the N1 phase is also much larger than that for the other sleep phases in both evaluation metrics. This result indicates that our model’s performance is better for classes with smaller sample sizes.

We also compared the F1-score values of the models. The overall comparison results are shown in Fig. 5. Our model’s F1-score values are the highest for the N1, N2, N3 and R stages, while in the W phase, our model’s F1-score values are not competitive.

C. Ablation Experiments

To verify the effectiveness of each step of the model improvement, we conducted ablation experiments for three models using only multi-view CNN (MVCNN), simultaneously using margin aware cross entropy loss and multi-view CNN (M-MVCNN), and simultaneously using adaptive margin aware complement-cross entropy and multi-view CNN (AMC-MVCNN).

The comparison results for these three models with respect to the DRM-SUB and SleepEDF-20 datasets are shown in Table VII. In terms of the average classification accuracy, the MVCNN, M-MVCNN and AMC-MVCNN gradually improved on both datasets. In addition, we used the F1-score to evaluate each model’s performance for each stage of sleep. On the DRM-SUB dataset, the F1-score values of the AMC-MVCNN model were higher than those of the MVCNN and M-MVCNN in all five stages of sleep staging. The greatest increase in the AMC-MVCNN compared to the MVCNN was observed in the N1 stage, where the number of sleep samples was the lowest. Similarly, on the SleepEDF-20 dataset, the F1-score values of the AMC-MVCNN model were also higher than those of the MVCNN and M-MVCNN in all five stages of sleep staging. On the SleepEDF-20 dataset, the AMC-MVCNN showed the largest increase compared to the MVCNN in the N1 period, when the number of sleep samples was the lowest. The AMC-MVCNN achieves the best results in each sleep phase on both datasets. In addition, it demonstrated the largest improvement during the N1 stage with values of 2.37% and 9.88%, respectively, which indicates that the AMC-MVCNN can not only improve the overall performance of the model with the addition of improvement strategies but also achieve better improvement for the class with the smallest sample percentage. Overall, the AMC-MVCNN obtained better results in terms of both the stages of sleep and the overall average classification accuracy, indicating that each component of our improvements effectively enhances the performance of the model for sleep staging.

Fig. 6 shows the convergence curves of the average classification accuracies for the MVCNN, M-MVCNN and AMC-MVCNN on the test set. On the DRM-SUB dataset, the AMC-MVCNN essentially reaches the convergence state at 40 iterations and maintains the convergence advantage until the end of the iteration. On the SleepEDF-20 dataset, the average classification accuracy of the AMC-MVCNN was slightly worse than that of the M-MVCNN during the early stage of convergence but surpassed the other two models.
at 14 iterations and gradually improved its classification accuracy as the model was trained, ultimately reaching the convergence state at approximately 80 generations. According to the convergence curves, the AMC-MVCNN achieves better average classification accuracy than the MVCNN and M-MVCNN.

The results of each of the 20-fold cross-validations for the MVCNN, M-MVCNN and AMC-MVCNN are shown in Fig. 7. The AMC-MVCNN can achieve better classification accuracy for more cross-validations. Therefore, when compared with the MVCNN and M-MVCNN, the AMC-MVCNN achieves stronger classification performance.
TABLE VIII
PERFORMANCE EVALUATION OF AMC-MVCNN ON SLEEPEDF-ST DATASET

| DataSet    | Approach       | Fold | Acc (%) | W     | S1    | S2    | S3    | S4    | REM  |
|------------|----------------|------|---------|-------|-------|-------|-------|-------|------|
| SleepEDF-ST| VGG-RW         | 5    | 73.09   | 73.35 | 35.71 | 84.10 | 50.84 | 70.18 | 73.61|
|            | VGG-SMOTE      | 5    | 64.97   | 73.78 | 32.35 | 76.77 | 25.30 | 60.53 | 59.72|
|            | Resnet-RW      | 5    | 70.51   | 67.41 | 36.63 | 83.73 | 34.06 | 40.63 | 72.39|
|            | Resnet-SMOTE   | 5    | 70.44   | 60.20 | 23.15 | 82.90 | 24.58 | 70.61 | 68.47|
|            | MVCNN-RW       | 5    | 75.80   | 76.84 | 38.83 | 84.91 | 46.39 | 75.06 | 79.24|
|            | MVCNN-SMOTE    | 5    | 76.61   | 76.23 | 34.07 | 86.27 | 37.11 | 72.39 | 79.97|
| **Ours**   |                | 5    | **80.54**| **78.63**| **43.12**| **88.64**| **53.12**| **80.32**| **82.56**|

Based on the experimental results in Table VIII, our proposed AMC-MVCNN method achieves the highest overall classification accuracy compared to the RW and SMOTE techniques using the class imbalance technique. In addition, the F1-score values of the six stages of sleep showed that the AMC-MVCNN method improved the classification performance in all stages of sleep, and the greatest improvement was observed in stage S4, which contained the fewest samples. AMC-MVCNN does not need to introduce new simulation samples and has less computational overhead compared to the SMOTE method. AMC-MVCNN more easily fits the model and enables better identification of S4 compared to the reweighting technique.

**E. Discussion**

Existing methods regarding the improvement of the accuracy of the minority classes in sleep staging are still shortcomings. Reweighting not only requires manual adjustment of hyperparameters, but also tends to make model optimization unstable and difficult in extreme data imbalance or large-scale scenarios. Oversampling techniques for a minority of classes may lead to overfitting of that class. The way in which data enhancement techniques enhance the sample size of a minority class of sleep EEG is controversial. There is uncertainty as to whether the EEG signal generated by data enhancement techniques matches the true EEG distribution and whether the temporal correlation of the original signal is retained. And it may even introduce unknown noise and increase the cost of model fitting. The AMC-MVCNN method proposed in this paper avoids the destruction of the original sleep EEG timing structure and the introduction of unknown information through the adjustment of the generalization margin. In addition, a few classes of N1 are highly susceptible to misclassification due to their greater similarity to other sleep stages. The AMC-MVCNN method can reduce the confidence of the incorrect class by reverse suppression of the sample error information, which can further indirectly improve the confidence of the correct classification of the class and the minority class identification. The AMC-MVCNN achieved good classification performance on all three datasets, indicating that the AMC-MVCNN has good generalization ability. The AMC-MVCNN method showed the most significant
performance improvement in the minority class N1 of sleep staging, a result that may have clinical implications because the performance improvement of the AMC-MVCNN for the N1 stage is potentially valuable in the assessment of aiding the diagnosis and evaluation of multiple sleep disorders, such as episodic narcolepsy [45].

The AMC-MVCNN model proposed in this paper learns complementary information from multi-view data and improves the representation mining of the minority class samples, but there is also the possibility of introducing redundant features. Therefore, the introduction of attention mechanism or feature selection mechanism for further learning of the joint representations extracted by AMC-MVCNN can be considered in future work, which will help to improve the recognition ability of minorities classes. In addition, most of the existing proposed algorithms to improve the accuracy of a minority of classes (e.g., stage N1) have been studied from the perspective of increasing the sample size or weight adjustment, and they usually ignore studies that address the unique characteristics of the N1 stage of sleep EEG. Therefore, the learning of special waveforms (e.g., vertex waves) in N1 stage can be increased by introducing a priori knowledge in the future to further improve the recognition of sleep N1 stage.

IV. CONCLUSION

To improve the performance impact of imbalanced sleep data on the sleep staging model, we propose an adaptive margin aware complement-cross entropy loss based multi-view CNN. First, we learn complementary feature representations from the raw data of sleep and the time-frequency images by building two-stream CNNs with different convolutional dimensions to further mine the effective information of the minority classes. Second, we also propose a margin aware factor that considers both the relative relationship between the number of minority and frequent class samples to encourage minority classes to obtain larger margins and thus improve the classification accuracy for minority classes. In addition, we improve the complement entropy by means of a margin aware factor and propose an adaptive margin aware complement-cross entropy to enable the model not only to learn correct information but also to achieve error neutralization by suppressing incorrect predictions. Finally, we tested on three benchmark sleep datasets, and the test results demonstrated that we not only improved the accuracy of sleep staging overall but also improved the classification accuracy for minority classes. The AMC-MVCNN has been considered only from the perspective of performance improvement of the method itself, and the following work can combine new methods such as contrast learning and the characteristics of sleep EEG itself to further improve the accuracy of the minority classes in sleep staging from the perspective of mining the more discriminative representations of sleep staging minority classes.

REFERENCES

[1] S. Taran, P. C. Sharma, and V. Bajaj, “Automatic sleep stages classification using optimize flexible analytic wavelet transform,” Knowl-Based Syst., vol. 192, Mar. 2020, Art. no. 105367.

[2] X. Chen, J. He, X. Wu, W. Yan, and W. Wei, “Sleep staging by bidirectional long short-term memory convolution neural network,” Future Gener. Comput. Syst., vol. 109, pp. 188–196, Aug. 2020, doi: 10.1016/j.future.2020.03.019.

[3] M. Längkvist, L. Karlsson, and A. Loutfi, “Sleep stage classification using unsupervised feature learning,” Adv. Artif. Neural Syst., vol. 2012, pp. 1–9, Jul. 2012.

[4] 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, May 2016.

[5] H. Dong, A. Supratak, W. Pan, C. Wu, P. M. Matthews, and Y. Guo, “Mixed neural network approach for temporal sleep stage classification,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 26, no. 2, pp. 324–333, Feb. 2018.

[6] O. Tsinalis, P. M. Matthews, Y. Guo, and S. Zafeiriou, “Automatic sleep stage scoring with single-channel EEG using convolutional neural networks,” 2016, arXiv:1610.01683.

[7] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen, and M. D. Vos, “Automatic sleep stage classification using single-channel EEG: Learning sequential features with attention-based recurrent neural networks,” in Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Jul. 2018, pp. 1452–1455.

[8] H. Seo, S. Back, S. Lee, D. Park, T. Kim, and K. Lee, “Intra- and inter-epoch temporal context network (ITNet) using sub-epoch features for automatic sleep scoring on raw single-channel EEG,” Biomed. Signal Process. Control, vol. 61, Aug. 2020, Art. no. 102037, doi: 10.1016/j.bspc.2020.102037.

[9] A. Supratak, H. Dong, C. Wu, and Y. Guo, “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, Nov. 2017.

[10] H. Phan, O. Y. Chen, M. C. Tran, P. Koch, A. Mertins, and M. D. Vos, “XSleepNet: Multi-view sequential model for automatic sleep staging,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 9, pp. 5905–5915, Sep. 2021, doi: 10.1109/TPAMI.2021.3070057.

[11] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen, and M. D. Vos, “DNN filter bank improves 1-max pooling CNN for single-channel EEG automatic sleep stage classification,” in Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Jul. 2018, pp. 453–456.

[12] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, “Deep learning-based electroencephalography analysis: A systematic review,” 2019, arXiv:1901.05498.

[13] A. Malafeev et al., “Automatic human sleep stage scoring using deep neural networks,” Frontiers Neurosci., vol. 12, p. 781, Nov. 2018.

[14] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen, and M. De Vos, “SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 27, no. 3, pp. 400–410, Mar. 2019.

[15] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen, and M. De Vos, “Joint classification and prediction CNN framework for automatic sleep stage classification,” IEEE Trans. Biomed. Eng., vol. 66, no. 5, pp. 1285–1296, May 2019, doi: 10.1109/TBME.2018.2872652.

[16] E. Khalili and B. Mohammadzadeh Asl, “Automatic sleep stage classification using temporal convolutional neural network and new data augmentation technique from raw single-channel EEG,” Comput. Methods Programs Biomed., vol. 204, Jun. 2021, Art. no. 106063.

[17] M. Fu et al., “Deep learning in automatic sleep staging with a single channel electroencephalography,” Frontiers Physiol., vol. 12, p. 179, Mar. 2021, doi: 10.3389/fphys.2021.628502.

[18] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, “Multi-view 3D object detection network for autonomous driving,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 1907–1915.

[19] F. Miao, L. Yao, and X. Zhao, “Evolving convolutional neural networks by symbiotic organisms search algorithm for image classification,” Appl. Soft Comput., vol. 109, Sep. 2021, Art. no. 107537.

[20] B. M. Alhe呕吐 and H. R. Colten, “Sleep disorders and sleep deprivation: An unmet public health problem,” Tech. Rep., 2006.

[21] E. Eldele et al., “An attention-based deep learning approach for sleep stage classification with single-channel EEG,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 29, pp. 809–818, 2021.

[22] C. Huang, Y. Li, C. C. Loy, and X. Tang, “Deep imbalanced learning for face recognition and attribute prediction,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 11, pp. 2781–2794, Nov. 2020.

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, no. 1, pp. 321–357, 2002.
[24] C. Sun, J. Fan, C. Chen, W. Li, and W. Chen, “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, doi: 10.1109/ACCESS.2019.2933814.

[25] F. Lotte, “Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain-computer interfaces,” Proc. IEEE, vol. 103, no. 6, pp. 871–890, Jun. 2015, doi: 10.1109/JPROC.2015.2404041.

[26] J. Fan et al., “EEG data augmentation: Towards class imbalance problem in sleep staging tasks,” J. Neural Eng., vol. 17, no. 5, 2020, Art. no. 056017.

[27] C. Wei, J. Lee, Q. Liu, and T. Ma, “On the margin theory of feedforward neural networks,” Tech. Rep., 2018.

[28] J. A. K. Suykens and J. Vandewalle, “Least squares support vector machine classifiers,” Neural Process. Lett., vol. 9, no. 3, pp. 293–300, Jun. 1999.

[29] W. Liu, Y. Wen, Z. Yu, and M. Yang, “Large-margin softmax loss for convolutional neural networks,” in Proc. ICML, vol. 2, no. 3, 2016, pp. 1–10.

[30] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “SphereFace: Deep hypersphere embedding for face recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 212–220.

[31] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, “Learning imbalanced datasets with label-distribution-aware margin loss,” 2019, arXiv:1906.07413.

[32] H.-Y. Chen et al., “Complement objective training,” 2019, arXiv:1903.01182.

[33] Y. Kim, Y. Lee, and M. Jeon, “Imbalanced image classification with complement cross entropy,” 2020, arXiv:2009.02189.

[34] F. Fathimi, Z. Zhang, W. B. Goh, K. K. Ang, and C. Guan, “Towards EEG generation using GANs for BCI applications,” in Proc. IEEE EMBS Int. Conf. Biomed. Health Informat. (BHI), May 2019, pp. 1–4.

[35] A. Jalaly Bidgoly, H. Jalaly Bidgoly, and Z. Arezoumand, “A survey on methods and challenges in EEG based authentication,” Comput. Secur., vol. 93, Jun. 2020, Art. no. 101788.

[36] Y. Li, H. Zaragoza, R. Herbrich, J. Shawe-Taylor, and J. Kandola, “The perceptron algorithm with uneven margins,” in Proc. ICML, vol. 2, pp. 379–386, 2002.

[37] Z. Li, K. Kamitsas, and B. Gloker, “Overfitting of neural nets under class imbalance: Analysis and improvements for segmentation,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent, 2019, pp. 402–410.

[38] F. Wang, J. Cheng, W. Liu, and H. Liu, “Additive margin softmax for face verification,” IEEE Signal Process. Lett., vol. 25, no. 7, pp. 926–930, Jul. 2018.

[39] S. Devuyst. (2005). The DREAMS Databases and Assessment Algorithm [Data Set]. [Online]. Available: http://www.tcts.fpms.ac.be?devuyst/Databases/DatabaseSubjects/

[40] S. Seifpour, H. Niknazar, M. Mikaeili, and A. M. Nasrabadi, “A new automatic sleep staging system based on statistical behavior of local extrema using single channel EEG signal,” Exp. Syst. Appl., vol. 104, pp. 277–293, Aug. 2018.

[41] C. Iber, “The AASM manual for the scoring of sleep and associated events: Rules,” Terminol. Tech. Specification, to be published.

[42] A. L. Goldberger et al., “PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals,” Circulation, vol. 101, no. 23, pp. e215–e220, 2000.

[43] H. Phan et al., “Towards more accurate automatic sleep staging via deep transfer learning,” IEEE Trans. Biomed. Eng., vol. 68, no. 6, pp. 1787–1798, Jun. 2021, doi: 10.1109/TBME.2020.3020381.

[44] B. Kemp, A. Janssen, and M. Roessen, “A digital telemetry system for ambulatory sleep recording,” Sleep-Wake Res. Netherlands, vol. 4, pp. 129–132, 1993.

[45] A. A. S. Medicine, “International classification of sleep disorders,” Diagnostic Coding Manual, pp. 51–55, 2005.