

Comparing Class-Weighted Loss and Focal Loss against SMOTE for Class Imbalance Handling in EEG-Based Sleep Stage Classification

Raka Yudistira^{1*}, Endah Setyowati^{2*}

* Departement Telecommunication Systems, Universitas Pendidikan Indonesia
rakayudistira678@upi.edu¹, endahsetyowati@upi.edu²

Article Info

Article history:

Received 2026-08-11

Revised 2026-09-11

Accepted 2026-10-03

Keyword:

EEG,
Sleep Stage Classification,
Class Imbalance,
SMOTE,
Class-Weighted Loss,
Focal Loss.

ABSTRACT

Class imbalance poses a challenge in EEG-based sleep stage classification, particularly for the N1 class, which has the fewest samples. This study compares three class imbalance handling strategies—SMOTE, Class-Weighted Loss, and Focal Loss—against a baseline using an identical 1D-CNN architecture. The Sleep-EDF Expanded dataset (30 subjects) was used, and evaluation was conducted via 3-fold subject-level cross-validation, ensuring subject-level separation between training, validation, and test sets. Results show that while the baseline achieved the highest overall accuracy (91.0%), it yielded the lowest balanced accuracy (72.9%) and N1 recall (25.1%). SMOTE improved N1 recall to 38.8%, whereas Class-Weighted Loss and Focal Loss achieved the highest N1 recall rates at 56.2% and 50.6%, respectively, with balanced accuracies of 77.6% and 77.8%. These findings indicate that, in this experiment, algorithm-level approaches were more effective than SMOTE at enhancing N1 class sensitivity and balancing performance across classes. However, this advantage did not extend to all metrics, as SMOTE produced higher accuracy and N1 F1-scores compared to the two algorithm-level approaches. Therefore, selecting a strategy to address class imbalance requires weighing the trade-off between minority class sensitivity and overall classification performance.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

I. INTRODUCTION

Sleep disorders are a widespread global health issue; multi-country meta-analyses indicate that conditions such as sleep apnea, poor sleep quality, and insomnia affect a significant portion of the adult population worldwide, with a pooled prevalence observed across 995,544 individuals from 36 countries [1]. To accurately diagnose these disorders, medical professionals employ polysomnography (PSG) an overnight multi-signal recording that includes electroencephalography (EEG), electrooculography (EOG), and electromyography (EMG) [2].

Human sleep stages are universally classified into five categories based on brain activity characteristics recorded via EEG signals [2]. These stages include Wake, referring to the state of being awake either before or after sleep [3]. N1 is the transitional stage from wakefulness to sleep, characterized by slowing brain waves while the individual remains easily aroused; this stage is the shortest and the most difficult to

distinguish or analyze using EEG [3],[4]. Next is N2, a light sleep stage that accounts for the majority of the night's sleep duration; this phase is marked by the appearance of sleep spindles and K-complexes [5],[6]. Then there is N3, the deep sleep stage also known as slow-wave sleep which is crucial for physical recovery and characterized by low-frequency, high-amplitude delta waves [3],[5]. Finally, REM (Rapid Eye Movement) is the stage where brain activity resembles that of the waking state; this phase is characterized by rapid eye movements and is generally associated with dreaming [3]. These five stages cycle repeatedly throughout the night with unequal duration proportions; this characteristic underlies the class imbalance issue that is the focus of this study [4] [7].

This process involves the manual labeling of sleep stages from PSG data by professional technicians a task that is inherently time-consuming and prone to inter-rater inconsistency making artificial intelligence-based automation an urgent necessity to support diagnoses that are more efficient, rapid, and consistent.

The availability of public PSG datasets, such as the sleep-EDF Expanded dataset from PhysioNet [2], [8], has spurred the development of various artificial intelligence approaches including deep learning for automated EEG-based sleep stage classification. Some studies have explored Convolutional Neural Network (CNN) architectures to extract features directly from raw EEG signals, bypassing the need for complex manual feature engineering [9], [10], [11]. Other research has enhanced CNN architectures with additional mechanisms, such as residual attention and channel fusion [9], attention mechanisms to capture long-term temporal dependencies [12], or combinations with Long Short-Term Memory (LSTM) layers to effectively model transitions between sleep stages [13]. Meanwhile, other approaches are specifically designed for computational efficiency on wearable or edge devices such as TinySleepNet [13] and lightweight CNN-Transformer models for low-power wireless devices [14]. Furthermore, hybrid approaches frequently combine multiple feature types or classifiers to improve accuracy by leveraging the combined strengths of the individual components [15].

While various architectural approaches have been proposed, significant challenges consistently appear in the literature regarding this topic, notably class imbalance and cross-subject generalization. The distribution of sleep stages is naturally uneven; the "wake" and N2 stages dominate the total sleep duration, whereas the N1 stage acting as a transition lasts for a relatively short time [15], [16]. Techniques for addressing imbalance via machine learning include the Synthetic Minority Over-Sampling Technique (SMOTE), which generates synthetic samples through interpolation in the feature space [17], and Focal Loss, which modifies the loss function to direct the model's focus toward samples that are difficult to classify [18]. However, the specific effectiveness of these methods within the domain of EEG physiological signals particularly through controlled, direct comparisons has rarely been explored systematically.

Several studies have investigated this topic, including the work of Candrakasih [16], who proposed a hybrid model combining 15 handcrafted features (statistical and spectral) with 128-dimensional features derived from a 1D-CNN; these are then classified using a stacking ensemble approach (employing Random Forest and SVM as base learners and Logistic Regression as the meta-learner). Utilizing 40 subjects from the Sleep-EDF Expanded dataset with a subject-independent split scheme (80/20), this research also applied SMOTE to address extreme class imbalance (Wake: 67.6%, N1: 2.95%; resulting in a macro F1-score of 31.4%

% and Cohen's Kappa of 0.34). The study demonstrated that CNN features alone achieving 72% accuracy outperformed both handcrafted features (approx. 70.4%) and hybrid features (approx. 65.5%), although the minority sleep stages remained difficult to identify. These findings underscore that cross-subject generalization remains a major challenge, making rigorous subject-independent validation essential to avoid over-optimistic results. Subsequent research by Liu et al. [19] proposed a hybrid CNN-BiLSTM-

Transformer architecture featuring multi-scale feature extraction via parallel convolutional branches. Sequential modeling was handled using BiLSTM and a Transformer-based attention mechanism. This model achieved 88.6% accuracy, a macro F1-score of 0.833, and a coefficient of 0.84 on the Sleep-EDF dataset, alongside 85.6% accuracy, a macro F1-score of 0.811, and a Cohen's Kappa of 0.80 on the Sleep-EDF Expanded dataset. The study showed that attention mechanisms and spectral fusion consistently improved performance, with significant gains observed in the N1, N3, and REM stages; however, this approach requires a more complex and computationally intensive architecture compared to simpler CNN methods. Another study by Zhu et al. [9] proposed a Multi-Branch CNN model incorporating stage refinement and residual attention channel fusion for automated sleep stage classification; this work demonstrated that adding a refinement mechanism at the end of the classification process corrected errors in distinguishing between difficult-to-differentiate classes, without the need for additional synthetic data augmentation techniques. These studies reveal a distinct pattern. While consistent performance improvements in minority classes can often be achieved through architectural complexity or SMOTE-based data augmentation, there has been limited systematic research comparing training-level imbalance handling such as class-weighted loss and focal loss against data-level approaches using identical architectures, particularly in the context of single-channel EEG analysis with relatively limited computational resources [20].

The objectives of this study include developing a 1D-CNN model as a baseline for single-channel EEG-based sleep stage classification using the Sleep-EDF Expanded dataset, and subsequently comparing the effectiveness of three class imbalance handling approaches SMOTE [17], Class-Weighted Loss, and Focal Loss [18] using nearly identical model architectures to ensure a fair comparison. Furthermore, the study aims to quantitatively evaluate whether algorithm-level handling of class imbalance outperforms data-level handling within the domain of EEG physiological signals. Finally, it analyzes recurring classification errors per class using a confusion matrix to provide insights into which sleep stages are the most difficult to identify [20].

Methodologically, this research contributes a systematic comparison between data-level and algorithm-level approaches for addressing class imbalance in the domain of EEG sleep staging, thereby serving as a reference and providing valuable insights for future research.

II. METHOD

A. Research Flow

This research was conducted through several key stages—spanning from data collection to the study's conclusion—as illustrated in Figure 1. The process began with the collection of a public EEG dataset, followed by signal preprocessing (involving epoch segmentation and normalization), subject-independent data partitioning, model design and training

across four data imbalance handling schemes, performance evaluation using multiple metrics, analysis and discussion of results, and finally, the study's conclusions.

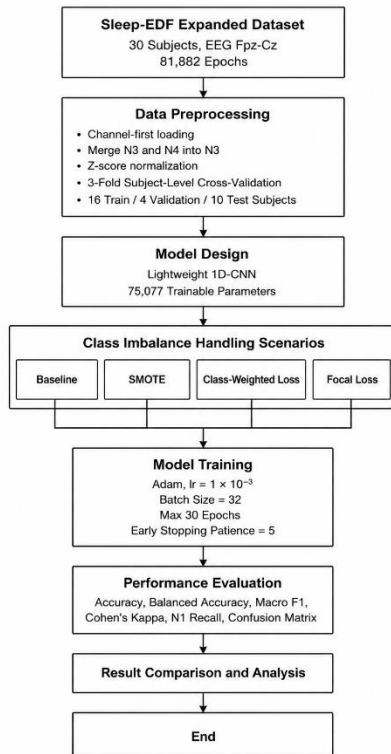


Figure 1. Research Flowchart

B. Data Collection

The data for this study was obtained secondarily from the public Sleep-EDF Database Expanded, a polysomnography (PSG) dataset published via the PhysioNet platform [2], [8]. The data was accessed using the MNE-Python library, which provides automated functions for the "Sleep Cassette" subset of the Sleep-EDF Expanded dataset. This study's subset consists of the first 30 subjects available via the sleep cassette data retrieval function in MNE-Python. For each subject, only the first-night recording was used to ensure that each individual contributed a single recording, maintaining a relatively consistent selection procedure. This selection of 30 subjects represents the upper limit of the experimental scope and does not constitute an evaluation of the entire subject pool in the Sleep-EDF Expanded dataset. Consequently, the generalizability of the study's results is limited to the specific data subset and experimental configuration employed. The Sleep-EDF Expanded dataset offers various physiological signals, including EEG, EOG, and EMG; however, this study specifically utilizes a single primary channel—the Fpz-Cz EEG—sampled at 100 Hz. The single-channel approach was adopted to minimize input complexity and computational requirements, while also allowing for the evaluation of class imbalance handling strategies without the confounding effects of multi-channel or multi-modal fusion. The Fpz-Cz channel was selected because the data is consistently available

across the Sleep-EDF recordings used and has been widely utilized in single-channel EEG-based sleep stage classification research [9], [10], [12], [15]. Thus, this design is also relevant to the development of simple sleep staging systems and sleep monitoring devices with a limited number of sensors. The sleep stage labels were obtained from hypnogram files accompanying each recording—data that had been manually scored by professional technicians following the Rechtschaffen and Kales guidelines. Data from each recording were segmented into 30-second epochs—the standard unit for sleep stage assessment—resulting in a total of 81,882 epochs across 30 subjects.

C. Preprocessing Data

Raw EEG signals are extracted using a "channel-first loading" strategy where channels are selected before the data is fully loaded into memory to ensure computational efficiency. Stages N3 and N4 from the Rechtschaffen and Kales scheme are merged into a single N3 class in accordance with modern AASM standards, resulting in the following target classes: wake, N1, N2, N3, and REM. The signals are normalized using z-score standardization to standardize the signal amplitude distribution; this normalization process is defined in Equation 1.

$$x' = \frac{x - \mu_{\text{train}}}{\sigma_{\text{train}}} \quad (1)$$

Mean and standard deviation parameters were calculated exclusively from the training set to prevent data leakage. Data partitioning was performed using 3-fold subject-level cross-validation to maintain independence between subjects. In each fold, 10 subjects served as the test set, while the remaining 20 subjects constituted the development set; the development set was further split into 16 subjects for training and 4 for validation. Crucially, no subject appeared in more than one of the training, validation, or test sets within a single fold. The validation set was used for early stopping based on the macro F1-score, whereas the test set was used for final evaluation once the best model was obtained. Under this scheme, each subject served as the test set exactly once across the three folds.

D. Model Architecture

The classification model employed in this study is a lightweight 1D-CNN, as illustrated in Figure 2. The model accepts input in the form of 30-second single-channel EEG epochs sampled at 100 Hz, resulting in 3,000 samples per input. The architecture comprises three convolutional blocks. The first block utilizes 16 filters with a kernel size of 7, a stride of 2, and padding of 3, followed by batch normalization, ReLU activation, and max pooling with a size of 4. The second block employs 32 filters with a kernel size of 5, a stride of 1, and padding of 2, followed by batch normalization, ReLU activation, and max pooling with a size of 4. Finally, the third block uses 64 filters with a kernel size of 3, a stride of 1, and padding of 1, followed by batch normalization,

ReLU activation, and adaptive average pooling with an output size of 16.

The resulting feature extraction outputs were flattened and passed to a fully connected layer comprising 64 units—utilizing ReLU activation and a dropout rate of 0.3—followed by a five-unit output layer representing the Wake, N1, N2, N3, and REM classes. This architecture comprises 75,077 parameters and allows for identical training and application across all experimental conditions; consequently, performance differences between conditions were evaluated primarily based on the class imbalance handling strategy, while the model architecture remained constant.

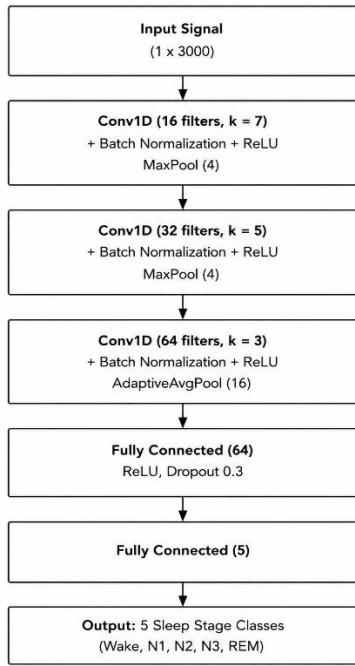


Figure 2. Research model architecture.

E. Experimental Scheme for Handling Class Imbalance

This study compares four experimental conditions using the same architecture. The first is a standard baseline using cross-entropy loss without handling class imbalance. SMOTE is applied exclusively to the training set following subject-independent splitting and normalization. Each 30-second EEG epoch is represented as a one-dimensional signal vector, which serves as the space for the interpolation process. The SMOTE implementation utilizes the `imbalanced-learn` library with `k\_neighbors = 5` and a fixed random seed. Synthetic samples are generated through interpolation between minority class samples and their nearest neighbors within the signal representation space. Oversampling for each minority class is capped at either eight times the class's initial count or the size of the majority class, whichever limit is reached first. SMOTE is applied only during training, ensuring that test data is not used to generate synthetic samples. The second condition employs Class-Weight Loss, where cross-entropy is weighted by the inverse class frequency. Finally, Focal Loss is used in combination with

class weighting, focusing the training process on minority class samples; in this implementation, the focusing parameter is set to $\gamma = 2.0$. Meanwhile, the α parameter is determined based on the inverse-frequency weights of the class distribution in the training set for each fold, assigning relatively higher weights to classes with fewer samples. The γ value remains constant across all folds without tuning against the test set, thereby protecting the evaluation procedure from potential data leakage.

To ensure a fair comparison, all experimental conditions utilized identical EEG inputs, subject comparisons, 1D-CNN architectures, optimizers, learning rates, batch sizes, maximum epoch limits, and evaluation procedures. The conditions differed solely in their mechanisms for handling class imbalance: the baseline employed no specific handling; SMOTE operated at the data level by generating synthetic samples within the training set; and Class-weighted loss and Focal Loss operated at the algorithm level without altering the training samples. Thus, data-level and algorithm-level approaches were compared using the same architecture and training budget. The mathematical equations for the baseline, SMOTE, Class-weighted loss, and Focal Loss are presented in Equations 2, 3, 4, 5, and 6.

$$L_{CE} = - \sum_{c=1}^C y_c \log(\hat{y}_c) \quad (2)$$

$$x_{\text{new}} = x_i + \delta(x_{\text{nn}} - x_i) \quad (3)$$

$$w_c = \frac{N}{C n_c} \quad (4)$$

$$L_{CW} = - \sum_{c=1}^C w_c y_c \log(\hat{y}_c) \quad (5)$$

$$L_{FL} = - \sum_{c=1}^C \alpha_c (1 - \hat{y}_c)^\gamma y_c \log(\hat{y}_c) \quad (6)$$

F. Implementation Environment and Evaluation Scheme

The experiment was implemented using PyTorch on an AMD Ryzen 5 CPU without GPU acceleration, employing the Adam optimizer with a learning rate of 1×10^{-3} and a batch size of 32; a maximum of 30 epochs was set, with early stopping based on the validation macro F1-score. Model performance was evaluated using Accuracy, Macro F1-score, Cohen's Kappa, as well as per-class precision, recall, and F1-score, accompanied by a normalized confusion matrix for qualitative error analysis. The macro F1-score was prioritized as the primary metric because accuracy can be misleading in cases of severe class imbalance [16], [17]. The equations for Accuracy, Precision, Recall, F1-score, macro F1-score, and Cohen's Kappa are presented in equations 7, 8, 9, 10, 11, and 12.

$$\text{Accuracy} = \frac{\sum_{c=1}^C TP_c}{N} \quad (7)$$

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \quad (8)$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \quad (9)$$

$$F1_c = \frac{2 \text{ Precision}_c \text{ Recall}_c}{\text{Precision}_c + \text{Recall}_c} \quad (10)$$

$$F1_{\text{macro}} = \frac{1}{C} \sum_{c=1}^C F1_c \quad (11)$$

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (12)$$

The final evaluation will be conducted using out-of-fold predictions from a three-fold subject-level cross-validation. The evaluation metrics employed will include accuracy, balanced accuracy, macro F1-score, and Cohen's Kappa, as well as precision, recall, and F1-score for each class. A normalized confusion matrix will be used to analyze misclassification patterns across sleep stages. Additionally, performance will be calculated at the subject level to evaluate inter-subject performance. Comparisons between methods regarding Macro F1-score and N1 recall will be analyzed using the Wilcoxon signed-rank test based on paired results from the same subjects. The Holm correction will be applied separately to each metric to account for multiple comparisons across the six pairwise comparisons among the four methods.

III. RESULT AND DISCUSSION

A. Dataset Class Distribution

Prior to preprocessing, the raw annotations from these 30 subjects yielded a total of 86,400 epochs, each with a duration of 30 seconds. The initial distribution comprised the following classes: Wake (56,069 epochs; 64.89%), N1 (2,506 epochs; 2.90%), N2 (14,061 epochs; 16.27%), N3 (2,293 epochs; 2.65%), N4 (1,401 epochs; 1.62%), REM (5,552 epochs; 6.43%), Unknown stage (4,483 epochs; 5.19%), and Movement time (35 epochs; 0.04%). During the preprocessing stage, the 4,483 epochs annotated as "Unknown stage" and the 35 epochs annotated as "Movement time" were excluded. Additionally, the N3 and N4 labels were merged into a single N3 class, in accordance with the five-class scheme adopted for this study. Following preprocessing, a total of 81,882 epochs remained across the 30 subjects. The class distribution revealed a significant imbalance: the Wake class was the majority class with 56,069 epochs (68.48%), followed by N2 (14,061 epochs; 17.17%), REM (5,552 epochs; 6.78%), and N3 (3,694 epochs; 4.51%), while N1 was

the minority class with the fewest samples (2,506 epochs; 3.06%). Based on this distribution, the imbalance ratio between the majority Wake class and the minority N1 class was 22.37:1, indicating that there were 22 times more Wake samples than N1 samples.

Thus, N1 represents the class with the lowest percentage in the dataset. This imbalance demonstrates that model evaluation cannot rely solely on accuracy, as performance on the majority class can disproportionately influence the overall score. Therefore, the Macro F1-score is used as the primary metric to evaluate balanced performance across classes, accompanied by a specific analysis of the N1 class—which, as shown in Figure 3, has a relatively small sample size.

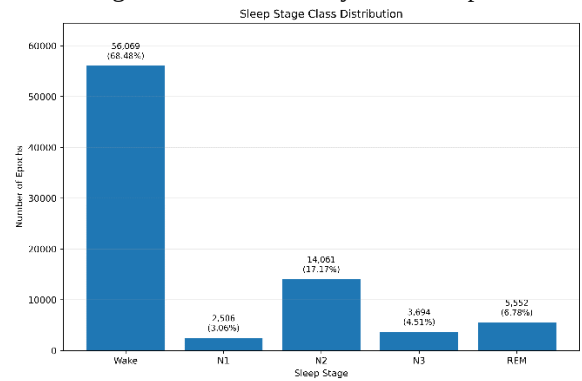


Figure 3. Distribution of sleep stage classes in the dataset used.

B. Overall Performance Comparison

The performance of the four experimental conditions conducted was compared using accuracy, Macro F1-score, Cohen's Kappa, F1-score, and recall for the N1 class. The overall results are presented in Table I.

TABLE I
PERFORMANCE COMPARISON OF FOUR METHODS FOR HANDLING CLASS IMBALANCE

Method	Acc	Bal. Acc.	Macro F1	F1 N1	Recall N1
Baseline	91.0%	72.9%	73.0%	30.2%	25.1%
SMOTE	88.3%	73.7%	70.2%	33.2%	38.8%
Class-Weighted Loss	87.1%	77.6%	71.1%	33.0%	56.2%
Focal Loss	87.6%	77.8%	71.8%	32.6%	50.6%

Based on Table I, the baseline model achieved the highest accuracy of 91.0% and a macro F1-score of 73.0%. Meanwhile, SMOTE—as a data-level approach—yielded a balanced accuracy of 73.7% and an N1 recall of 38.8%. On the other hand, the algorithm-level approach using class-weighted loss demonstrated a balanced accuracy of 77.6% with the highest N1 recall of 56.2%. Focal loss, meanwhile, produced the highest balanced accuracy of 77.8% with an N1 recall of 50.6%. These results indicate that algorithm-level approaches are more effective than SMOTE in improving performance balance across classes and sensitivity toward the

minority class (N1). A comparison of accuracy, macro F1-score, and Cohen's Kappa is presented in Figure 4.



Figure 4. Comparison of Accuracy, Macro F1-score, and Cohen's Kappa across four experimental conditions

Figure 4 illustrates a trade-off between aggregate performance and the implementation of class imbalance handling strategies: the baseline achieved the highest accuracy, macro F1-score, and Cohen's Kappa, whereas SMOTE showed a smaller decline in accuracy and Cohen's Kappa compared to the two algorithm-level approaches. Conversely, Class-weighted loss and Focal loss yielded slightly higher macro F1-scores than SMOTE, although they still fell short of the baseline. These findings indicate that increasing sensitivity toward the minority class through class imbalance handling does not necessarily result in a significant overall performance improvement; consequently, the advantages of the algorithm-level approaches in this study were limited to improvements in balanced accuracy and N1 recall—as shown in Table 1—rather than demonstrating across-the-board superiority in evaluation metrics.

C. Classification performance for each class

A subsequent analysis of the F1-score for each sleep stage was conducted to determine the strategy for handling class imbalance for each class. The comparison of F1-scores by class is shown in Figure 5.

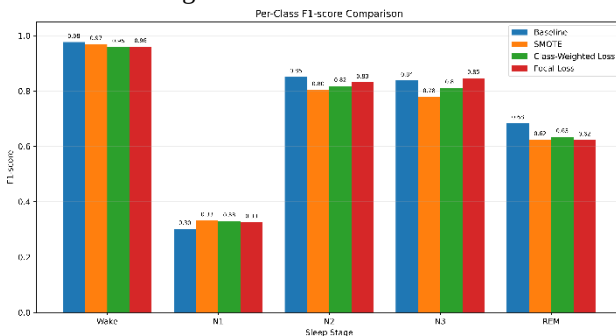


Figure 5. Comparison of F1-scores for each sleep stage across four experimental conditions.

Classification performance revealed varying responses across the different classes following the application of class imbalance handling strategies. The most notable change was observed in class N1, which was designated as the minority

class in the dataset. The baseline achieved an N1 F1-score of 30.2%, whereas SMOTE improved this to 33.2%, marking the highest value for class N1. Class-weighted loss and focal loss yielded N1 F1-scores of 33.0% and 32.6%, respectively. These results indicate that SMOTE provided a slightly better balance between precision and recall for class N1 compared to the baseline and the other two algorithm-level models. However, the differences in N1 F1-scores across the various class imbalance handling methods were relatively small; consequently, their effectiveness regarding the minority class warrants further analysis focusing on N1 recall.

D. Recall Analysis for Minority Class N1

Since N1 represents only about 3.1% of the total data, the recall for the N1 class was specifically analyzed to evaluate the model's performance and its ability to detect minority samples. The results are shown in Figure 6.

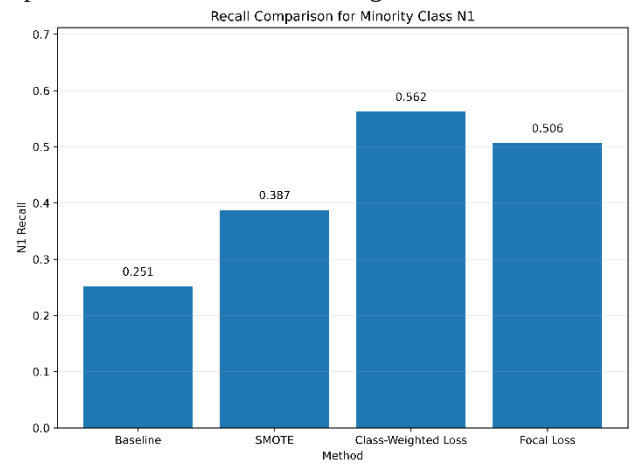


Figure 6. Comparison of Recall for class N1 across four experimental strategies.

Based on Figure 6, the N1 recall increased from 25.1% at baseline to 38.8% following the application of SMOTE. Significant improvements were achieved by the class-weighted loss and focal loss models, reaching 56.2% and 50.6%, respectively. These results confirm that both algorithm-level approaches—particularly class-weighted loss, which yielded the highest recall—were more effective than SMOTE at enhancing model sensitivity toward the N1 minority class. However, the increase in recall did not translate into overall superiority, as the N1 F1-scores across all methods remained relatively similar. This indicates a trade-off between increased sensitivity to N1 and prediction accuracy across classes; consequently, recall requires further analysis alongside precision, F1-score, and the patterns of positive classifications observed in the confusion matrix.

E. Confusion Matrix Analysis

To evaluate error patterns in greater detail, a normalized confusion matrix is used for each of these experimental conditions. Values along the diagonal represent the proportion of accurately classified samples, while off-diagonal values indicate inter-class error patterns. Further

details can be observed in the respective confusion matrices shown in Figures 7, 8, 9, and 10.

As shown in Figure 7, the baseline demonstrates high performance for the 'wake' and N2 classes, with recall rates of 97% and 87%, respectively. However, recall for the N1 class is only 25%, with the most frequent errors involving misclassification as REM (34%) and N2 (29%). This indicates that the baseline still struggles to recognize the N1 class compared to the other classes.

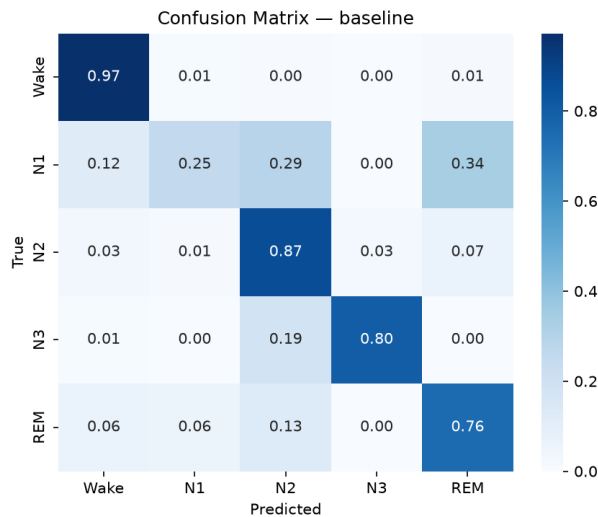


Figure 7. Confusion Matrix Baseline

Figure 8 shows that applying SMOTE increased the recall for the N1 class to 39%, as SMOTE generates synthetic data for the model to learn, thereby enabling it to recognize the minority class. Additionally, the misclassification rate of N1 as N2 decreased to 24%, while the rate of N1 being misclassified as REM dropped to 26%. However, the increased sensitivity for N1 was accompanied by a decline in recall for certain other classes—specifically, N2 fell to 77% and N3 to 86%. These results demonstrate that SMOTE enhances the model's ability to recognize the N1 minority class compared to the baseline; nevertheless, the improvement is relatively modest and entails a trade-off involving reduced classification performance for some other classes.

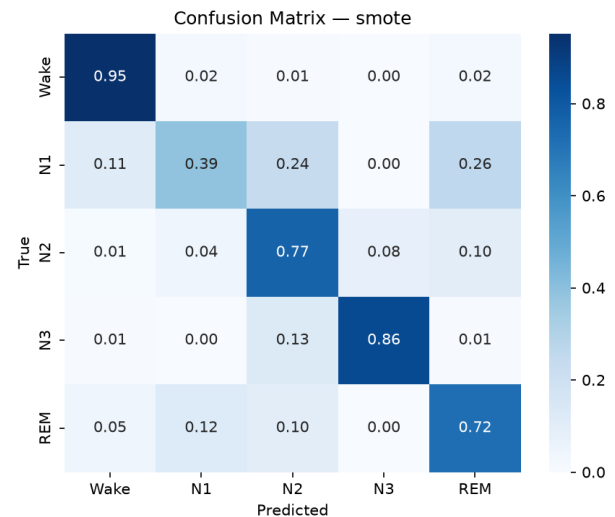


Figure 8. Confusion Matrix SMOTE

In Figure 9, Class-weighted loss yielded the highest recall at 56.2%. Misclassification rates of N1 as N2 and REM decreased to 17% and 23%, respectively. However, the increased sensitivity toward N1 was accompanied by a rise in the proportion of samples from other classes predicted as N1—specifically, 17% for REM and 8% for N2. These results indicate that Class-weighted loss is effective at enhancing the model's sensitivity to the N1 class compared to other models; however, its advantage lies specifically in N1 recall rather than in overall classification performance.

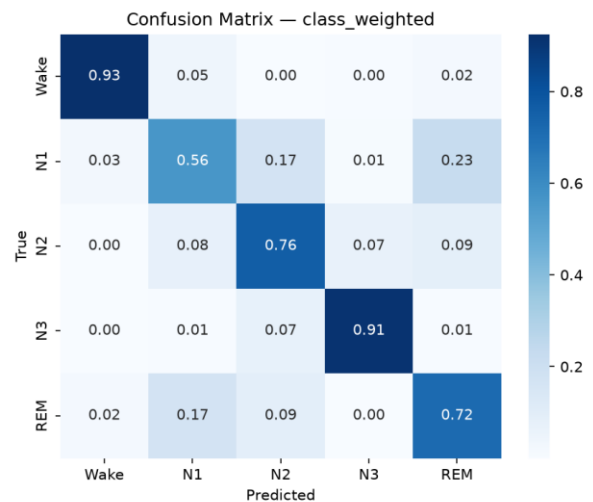


Figure 9. Confusion Matrix Class-Weighted Loss

Finally, Figure 10 shows that focal loss yields an N1 Recall of 50.6%—higher than SMOTE and the baseline, though lower than Class-weighted loss. The misclassification rate of N1 as N2 is 15%, while the rate of misclassifying it as REM is 31%. Additionally, the proportions of REM and N2 predicted as N1 are 15% and 7%, respectively. These results indicate that focal loss improves sensitivity to the N1 class compared to SMOTE and the baseline; however, this improvement is accompanied by false positives where other

classes are misclassified as N1. Thus, while focal loss offers advantages for minority class recognition, it does not demonstrate overall superiority compared to the other methods.

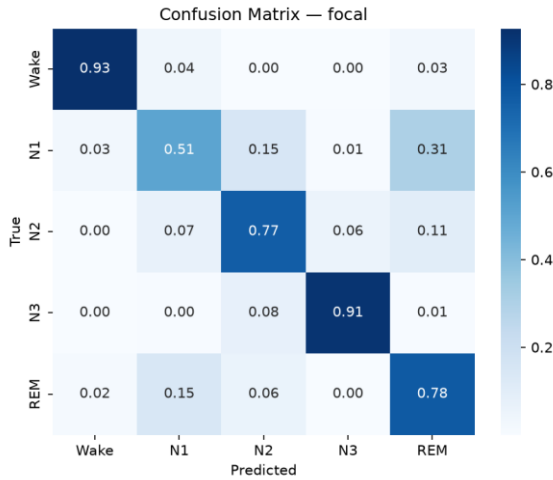


Figure 10. Confusion Matrix Focal Loss

F. Analysis of the effectiveness of class imbalance handling strategies

The overall experimental results reveal a trade-off between aggregate performance and the model's ability to handle the minority class. The baseline achieved the highest accuracy (91.0%) but yielded the lowest balanced accuracy (72.9%) and N1 recall (25.1%). SMOTE, as a data-level approach, increased N1 recall by 38.8%, whereas algorithm-level approaches—specifically class-weighted loss and focal loss—delivered significant improvements of 56.2% and 50.6%, respectively. These gains were accompanied by improvements in balanced accuracy (77.6% for class-weighted loss and 77.8% for focal loss), which slightly exceeded the result achieved with SMOTE (73.7%). These findings indicate that, within the experimental setup used, algorithm-level approaches were more effective than SMOTE at enhancing N1 class sensitivity and balancing performance across classes.

However, this advantage does not extend to overall classification performance. SMOTE achieved an accuracy of 88.3% and an N1 F1-score of 33.2%, outperforming both Class-weighted loss and Focal loss on these metrics. The confusion matrix also reveals that the improvement in N1 recall at the algorithm level is accompanied by a significant number of N1 false positives relative to other classes. Thus, the effectiveness of the algorithm-level approach in this study is limited to improving minority-class sensitivity and balanced accuracy, rather than enhancing overall classification performance.

G. Inter-subject Evaluation and Statistical Analysis

A subject-level evaluation was conducted on 30 subjects using an out-of-fold approach to analyze performance variations across subjects. For N1 recall, the baseline

achieved an average of $23.88\% \pm 18.93\%$, whereas SMOTE achieved $36.78\% \pm 20.14\%$. Class-weighted loss yielded the highest average at $54.67\% \pm 23.06\%$, while focal loss achieved $48.28\% \pm 23.91\%$. These results confirm that the increased sensitivity to the N1 class observed with the algorithm-level approach is also evident in the subject-level evaluation. However, the relatively high standard deviations across all methods indicate performance variability among subjects.

The Wilcoxon signed-rank test was performed on paired results from 30 subjects to evaluate differences among methods in Macro F1-score and N1 recall. After Holm correction, no pairwise differences in Macro F1-score were statistically significant ($p > 0.05$). Conversely, the differences in N1 recall between SMOTE and Class-weighted loss, as well as between SMOTE and focal loss, remained significant, with adjusted p-values of 0.000114 and 0.00678, respectively. These results provide statistical evidence that both algorithm-level approaches yield significant differences in N1 sensitivity compared to SMOTE, yet they do not demonstrate a significant advantage in terms of Macro F1-score.

H. Limitations

This study has limitations that should be highlighted when interpreting the results. First, the experiment utilized only 30 subjects from the total available in the Sleep-EDF Expanded dataset; consequently, the results do not fully represent an evaluation of the entire dataset. Second, the classification relied on a single EEG channel (Fpz-Cz). While using a single EEG channel offers the significant advantage of reduced system complexity, information from other EEG channels or physiological signals remained unutilized. Third, although a 3-fold subject-level cross-validation was employed to mitigate result dependency on a single data split, the study evaluated only one 1D CNN architecture with a consistent hyperparameter configuration across all methods. Therefore, the comparative results regarding SMOTE, class-weighted loss, and focal loss in this study cannot be directly generalized to other model architectures, training configurations, or EEG datasets.

IV. CONCLUSION

This research compares SMOTE as a data level approach with an algorithm level approach, namely Class-weighted loss and focal loss in dealing with class imbalance in the classification of five stages of sleep based on single channel EEG using an identical 1D-CNN architecture. Based on the results of the 3-fold subject-level cross-validation evaluation, the baseline produced the highest accuracy at 91.0% but had the lowest balanced accuracy and N1 recall at 72.9% and 25.1% respectively. SMOTE increased N1 recall to 38.8%, while class weighted loss and focal loss produced relatively high recall, namely 56.2% and 50.6%, and balanced accuracy of 77.6% and 77.8%. These results show that in this experimental configuration, the algorithm level-based approach is more effective than SMOTE in increasing

sensitivity to N1 minority classes and balancing performance between classes. However, this advantage does not apply to all metrics, as SMOTE maintains higher overall accuracy and N1 F1-score than both algorithm-level methods. Apart from that, statistical analysis also shows the difference in N1 recall between SMOTE and Class-weighted loss and focal loss, while the Macro F1-score is not significant after correcting multiple comparisons. Thus, selecting a strategy for handling class imbalance needs to consider the trade-off between minority class sensitivity and overall classification performance.

REFERENCES

- [1] J. B. Canever *et al.*, “Worldwide prevalence of sleep problems in community-dwelling older adults: A systematic review and meta-analysis,” *Sleep Med.*, 2024.
- [2] B. Kemp, A. H. Zwiderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Oberyé, “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.
- [3] I. Hussain *et al.*, “Quantitative evaluation of EEG-biomarkers for prediction of sleep stages,” *Sensors*, vol. 22, no. 8, p. 3079, 2022, doi: 10.3390/s22083079.
- [4] S. Lin *et al.*, “SSC-SleepNet: A Siamese-based automatic sleep staging model with improved N1 sleep detection,” *IEEE J. Biomed. Health Inform.*, vol. 29, no. 9, pp. 6830–6843, 2025, doi: 10.1109/JBHI.2025.3572886.
- [5] H. Huang *et al.*, “EEG-based sleep staging analysis with functional connectivity,” *Sensors*, vol. 21, no. 6, p. 1988, 2021, doi: 10.3390/s21061988.
- [6] A. A. Casciola *et al.*, “A deep learning strategy for automatic sleep staging based on two-channel EEG headband data,” *Sensors*, vol. 21, no. 10, p. 3316, 2021, doi: 10.3390/s21103316.
- [7] W. Li, T. Liu, B. Xu, and A. Song, “SleepFC: Feature pyramid and cross-scale context learning for sleep staging,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 32, pp. 2198–2208, 2024, doi: 10.1109/TNSRE.2024.3406383.
- [8] 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.
- [9] T. Zhu, W. Luo, and F. Yu, “Multi-branch convolutional neural network for automatic sleep stage classification with embedded stage refinement and residual attention channel fusion,” *Sensors*, vol. 20, no. 22, p. 6592, 2020.
- [10] Z. Mousavi, T. Y. Rezaii, S. Sheykhivand, A. Farzamnia, and S. N. Razavi, “Deep convolutional neural network for classification of sleep stages from single-channel EEG signals,” *J. Neurosci. Methods*, vol. 324, p. 108312, 2019.
- [11] C. Lin, C. Lin, and X. Lin, “Automatic sleep stage classification using a Taguchi-based multiscale convolutional compensatory fuzzy neural network,” *Appl. Sci.*, vol. 13, no. 18, p. 10442, 2023.
- [12] E. Eldele *et al.*, “An attention-based deep learning approach for sleep stage classification with single-channel EEG,” *Sensors*, vol. 23, no. 15, p. 6986, 2023.
- [13] A. Supratak and Y. Guo, “TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG,” presented at the Proceedings of the 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2020, pp. 641–644.
- [14] Z. Yao and X. Liu, “A CNN-Transformer deep learning model for real-time sleep stage classification in an energy-constrained wireless device,” *ArXiv Prepr. ArXiv221113005*, 2022.
- [15] M. N. Hasan and I. Koo, “Mixed-input deep learning approach to sleep/wake state classification by using EEG signals,” *Diagnostics*, vol. 13, no. 14, p. 2358, 2023.
- [16] M. A. C. Candrakasih, B. A. Kusuma, and P. Subarkah, “Hybrid CNN for Sleep Stage Classification Based on EEG,” *JAIC*, vol. 10, no. 3, pp. 2723–2729, 2026.
- [17] 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, pp. 321–357, 2002.
- [18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” presented at the Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988.
- [19] C. Liu *et al.*, “MultiScaleSleepNet: A hybrid CNN–BiLSTM–Transformer architecture with multi-scale feature representation for single-channel EEG sleep stage classification,” *Sensors*, vol. 25, no. 20, p. 6328, 2025.
- [20] N. Michielli, U. R. Acharya, and F. Molinari, “Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals,” *Comput. Biol. Med.*, vol. 106, pp. 71–81, 2019.