A Controlled Evaluation of Class Imbalance Handling Techniques in Automated EEG-Based Sleep Stage Classification

Blessing C. Uzo[1], Ponsak S. Bande[1*], Micah Okpara[1], John O. Onyima[1], Nnaemeka V. Ugwu[2]

[1]Department of Computer Science, Faculty of Physical Sciences, University of Nigeria, Nsukka, Enugu, Nigeria

[2]Department of Computer Science, Godfrey Okoye University, Enugu, Nigeria

***Corresponding Author**

DOI: https://doi.org/10.51244/IJRSI.2026.1307000240

Received: 27 July 2026; Accepted: 01 August 2026; Published: 09 August 2026

ABSTRACT

Automated sleep stage classification from electroencephalogram (EEG) signals is a central task in sleep medicine, yet polysomnographic datasets are inherently imbalanced, with the transitional stage N1 typically comprising only a small fraction of epochs. Although class imbalance handling is widely applied, its independent contribution to sleep staging performance remains unclear because balancing techniques are usually embedded within larger deep learning systems and evaluated jointly with them. This study presents a controlled evaluation of four imbalance handling techniques, namely random oversampling, random undersampling, the Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN), against an imbalanced baseline, within an otherwise fixed pipeline comprising single-channel frontal EEG preprocessing, eighteen handcrafted time, frequency and nonlinear features, a Random Forest classifier and five-fold subject-wise cross-validation on the Sleep-EDF Expanded database (20 subjects, 39 recordings, 106,107 epochs, of which N1 comprised 2.64%). Resampling was applied exclusively to training folds. The baseline achieved an accuracy of 0.899 and a Cohen's kappa of 0.795 while correctly identifying only 13.8% of N1 epochs, suggesting that aggregate measures can conceal minority-stage failures. Although several resampling strategies improved N1 recall, random undersampling produced the highest N1 recall of 0.558 while also producing the lowest precision, suggesting a shift in the decision boundary rather than improved class separability. SMOTE achieved the highest observed macro F1-score of 0.693 with the smallest reduction in majority-stage performance, while random oversampling was largely ineffective and ADASYN showed no advantage over SMOTE. Friedman and Wilcoxon tests did not reach statistical significance under the five-fold design, highlighting the limited statistical power of the evaluation. The findings provide evidence-supported, goal-specific recommendations for selecting imbalance handling strategies and reinforce the need for per-class evaluation in automated sleep staging.

Keywords: Sleep stage classification; Class imbalance; Electroencephalography (EEG); SMOTE; Random Forest

INTRODUCTION

Sleep is a fundamental physiological process characterized by cycles of wakefulness, non-rapid eye movement sleep and rapid eye movement sleep. In clinical sleep assessment, sleep specialists use polysomnography (PSG), in which overnight physiological recordings are divided into 30-second segments and scored according to standardized guidelines [1]. Manual scoring of a full-night sleep recording is time-consuming and may vary among specialists. This has led to a growing interest in the development of automated sleep stage classification systems. The availability of publicly accessible databases, such as Sleep-EDF and its expanded version available through PhysioNet, has allowed researchers to begin developing automated systems for sleep analysis [2], [3]. The automation of sleep stage classification has progressed significantly, moving from classical machine learning systems that extract handcrafted EEG features, to current deep learning algorithms that generate representations directly from raw EEG signals. However, there is still a major shortcoming: one can achieve

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

high levels of performance for overall system accuracy, while simultaneously witnessing insufficient recognition of stage N1. A review of current methods of EEG-based sleep stage analysis showed that relying on accuracy may lead to misleading interpretation of results [4]. This issue is particularly important because N1 is a transitional stage and is usually underrepresented in overnight sleep recordings.

A principal cause of this weakness is the natural class imbalance of sleep data. Because N1 occupies only a small fraction of total sleep time, conventional learning algorithms may become biased toward majority classes, since aggregate error can be minimized by correctly predicting frequent categories while neglecting infrequent ones [5]. Several imbalance handling strategies have therefore been adopted in sleep staging, including random oversampling, random undersampling, synthetic oversampling, class weighting and ensemble-based remedies. However, in most published studies, the balancing mechanism is introduced together with a new architecture, augmentation scheme, loss function or temporal model. As a result, the independent contribution of the imbalance handling strategy itself remains difficult to isolate.

This study addresses that gap by isolating class imbalance handling as the primary experimental variable in automated EEG-based sleep stage classification. Using a fixed single-channel EEG pipeline, handcrafted time-domain, frequency-domain and nonlinear features, and a Random Forest classifier [6], five experimental conditions are compared under leakage-free subject-wise cross-validation: no balancing, random oversampling, random undersampling, SMOTE and ADASYN. The study is organised around four research questions: RQ1. How does class imbalance influence the performance of automated EEG-based sleep stage classification? RQ2. Which class imbalance handling technique provides the best overall classification performance? RQ3. Which balancing strategy most effectively improves recognition of the minority stage N1? RQ4. Do improvements in minority-stage recognition translate into meaningful overall model improvement? The contributions of the work are fourfold. Firstly, it introduces a systematic, data-oriented assessment of four common resampling methods in which all pipeline components except the balancing strategy are kept constant. Secondly, it evaluates the influence of each method on minority-stage recognition through the precision-recall balance. Thirdly, it assesses whether synthetic sampling techniques provide meaningful advantages over simpler resampling methods. Fourthly, it provides practical, goal-dependent recommendations for selecting imbalance handling strategies and reinforces per-class evaluation as a methodological requirement for sleep staging research.

The remainder of this paper is organised as follows. Section 2 reviews related work on automated sleep staging, class imbalance and imbalance handling techniques. Section 3 describes the dataset, preprocessing, feature extraction, balancing strategies, classifier, evaluation protocol and statistical analysis. Section 4 presents the experimental results. Section 5 discusses the findings, implications and limitations. Section 6 concludes the paper.

Related Work

Automated EEG-Based Sleep Stage Classification

Sleep staging has traditionally been performed by trained *specialists* using polysomnography (PSG) sessions of thirty seconds according to standardized American Academy of Sleep Medicine (AASM) guidelines [1]. Despite the fact that manual scoring is still regarded as the clinical standard, this method is very labor-intensive and has numerous problems of inter-scorer discrepancies. For this reason, the development of automated methods for sleep stage classification became important. Public databases like Sleep-EDF and Sleep-EDF Expanded opened new possibilities for reproducible development and assessment of sleep stage classification methods [2], [3].

Two main types of methods have been developed. The first approach extracts signal descriptors from EEG epochs and classifies them using classical machine-learning methods. The features used include time-domain statistics, frequency-band powers, Hjorth parameters, entropy-based features, wavelet characteristics, and nonlinear descriptors. For example, Jain and Ganesan [7] applied visibility-graph features, AR parameters, Hjorth parameters, Higuchi fractal dimensions as well as features concerning entropy in order to conduct EEG sleep scoring. Dogaheh and Moradi [8] used features based on time, frequency, and entropy from wearable headband signals and found that Random Forest classifier outperformed SVM and kNN for five-stage sleep

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

classification. Stenwig et al. [9] further demonstrated that feature-based ensemble learning with optimized EEG-channel selection can provide competitive performance while reducing the number of EEG channels required. Similarly, Rathod et al. [10] compared several feature-based machine-learning models on Sleep-EDF data and found that ensemble learning achieved the strongest overall performance.

The second direction uses deep learning to learn representations directly from raw or minimally processed EEG signals. Recent work includes CNN-BiGRU networks with self-attention [11], ensemble CNN-BiLSTM architectures [12], transition-aware sleep classifiers [13], multimodal encoder-decoder networks [14], and explainable transformer-based frameworks [15]. These systems often report strong overall accuracy and Cohen's kappa values on Sleep-EDF-type datasets. However, high aggregate performance does not imply equal performance for all stages. In particular, N1 remains consistently difficult to identify. EnsembleNet, for instance, reported lower N1 F1-scores than those of W, N2, N3, and REM on Sleep-EDF-20 and Sleep-EDF-78 [12]. Salfi et al. [16] similarly reported substantially weaker N1 performance than performance for other stages in a wearable single-channel EEG setting. A recent review by Yang and Wang [4] concluded that sleep-staging research still relies excessively on overall accuracy, which can obscure poor performance on minority stages and difficult sleep segments.

Class Imbalance in Sleep Staging

Class imbalance occurs when the number of samples differs substantially among classes. In such settings, conventional learning algorithms may favour majority classes because minimizing aggregate classification error can be achieved by correctly predicting frequent categories while neglecting infrequent ones [5]. Sleep staging is naturally imbalanced because different sleep stages are not represented equally. Among the sleep stages, N1 is typically underrepresented, while N2 and W often dominate. Zhou et al. [17] conducted an extensive analysis of class imbalance in the automatic sleep staging process to demonstrate the severity of the imbalance with the help of a specific imbalance factor and concluded that the use of imbalanced PSG data affects the correct operation of deep learning models. Xu et al. [18] also stated that class imbalance in the PSG datasets has not attracted nearly the same amount of interest as the development of architectures in the field of sleep staging. For instance, Lee et al. [19] recognized class imbalance as a limitation in single-channel EEG sleep-stage estimation and introduced spectral-band blending augmentation to improve classifier performance. These studies establish that class imbalance is not merely a property of sleep data but a factor that influences classifier behaviour.

The impact of imbalance is also evident in studies focused on minority-stage recognition. Yu et al. [20] used GAN-based signal generation to improve N1 detection under class imbalance. Huang et al. [21] integrated Borderline-SMOTE with supervised contrastive learning and reported improved macro-level performance across multiple sleep datasets. Li et al. [22] noted that class imbalance can undermine pseudo-label quality in semi-supervised pediatric sleep staging. Together, these studies indicate that imbalance handling is increasingly recognized as necessary for reliable automated sleep analysis, especially for difficult classes such as N1.

Imbalance Handling Approaches

Imbalance-handling techniques are generally grouped into data-level, algorithm-level, and ensemble-level approaches [5]. Data-level approaches modify the class distribution of the training data. These include random oversampling, which duplicates minority examples; random undersampling, which removes examples from majority classes; and synthetic oversampling approaches such as Synthetic Minority Oversampling Technique (SMOTE) [23] and Adaptive Synthetic Sampling (ADASYN) [24]. SMOTE creates synthetic minority examples by interpolating between nearby minority samples, whereas ADASYN adaptively generates more samples in regions where minority examples are difficult to learn. These methods are widely available through the imbalanced-learn Python library [25]. Algorithm-level approaches change the learning process rather than directly modifying the training data. Examples include class weighting, cost-sensitive learning, focal-type losses, and weighted cross-entropy functions. In sleep staging, Liu et al. [26] proposed a class-adaptive weighted cross-entropy loss within a deep multi-scale salient-feature network. Zhang et al. [14] used Lovász loss to address class imbalance in a multimodal encoder-decoder framework. NeuroSleepNet also employed a logarithmic class-weighting strategy to improve minority-stage learning [15]. Similarly, MCTSleepNet incorporated an adaptive cross-entropy polynomial loss to increase attention to underrepresented sleep stages [27].

Most published imbalance-aware sleep-staging studies combine a balancing technique with a new deep-learning architecture. For example, Zhou et al. [17] combined GAN-based data generation with Gaussian-noise augmentation and a CNN-BiLSTM model. Lee et al. [19] incorporated spectral-band blending into an EEGNet-BiLSTM framework. Yu et al. [20] used transformer-encoder GAN generation alongside a residual sequence network. Huang et al. [21] combined Borderline-SMOTE with contrastive learning, residual networks, and modified loss functions. Although these studies demonstrate that imbalance-aware designs can improve performance, their experimental designs make it difficult to determine whether the reported improvements arise from the balancing method, network architecture, feature representation, temporal context, loss function, or a combination of these components. Classical machine-learning studies with explicit imbalance handling are comparatively limited. Rahman et al. [28] evaluated Random Under-Sampling Boosting, Random Forest, and SVM methods for single-channel EOG sleep staging and reported improved S1-stage performance using RUSBoost. Sharma et al. [29] compared unbalanced and balanced sleep-stage classification settings using EEG features and an ensemble of bagged trees. Park et al. [30] examined resampling and class-weighting strategies for overnight sleep-stage imbalance. These studies support the relevance of imbalance handling in conventional machine-learning pipelines; however, they do not provide a controlled comparison of random oversampling, random undersampling, SMOTE, and ADASYN under an otherwise fixed EEG-processing and classification framework.

Research Gap and Positioning of the Present Study

The reviewed literature demonstrates that class imbalance is widely recognized in automated sleep staging, particularly for N1 recognition. However, four important gaps remain. First, much imbalance research is embedded within deep-learning studies that simultaneously introduce new architectures, data-augmentation approaches, loss functions or temporal modelling techniques. Thus, the isolated contribution of the balancing procedure is difficult to determine. Second, common resampling techniques such as random oversampling, random undersampling, SMOTE, and ADASYN have rarely been compared quantitatively under standardized sleep-staging conditions. Third, many studies focus on accuracy and Cohen's kappa even though high accuracy may coexist with poor performance on minority stages. Macro F1-score, per-class F1-score, minority-class recall and minority-class precision are necessary to reveal this behaviour, while confusion matrices can provide additional detail on class-specific error patterns. Finally, classical feature-based machine-learning pipelines remain useful because they are computationally efficient, interpretable, and practical for resource-limited biomedical applications. Random Forest, in particular, is robust for structured data and provides a suitable controlled classifier for evaluating data-level imbalance interventions [6].

Thus, this study treats class imbalance handling as the key experimental variable. A standard single-channel EEG processing pipeline, handcrafted time-domain, frequency-domain, and nonlinear features, and the Random Forest classifier allowed a controlled comparison of no resampling, random oversampling, random undersampling, SMOTE, and ADASYN. For evaluation, subject-wise cross-validation, macro F1-score, per-class F1-score, Cohen’s kappa, minority-stage precision and recall, and non-parametric statistical tests were used. This design provides a basis for evaluating the practical effect of class imbalance handling methods on automated EEG-based sleep stage classification.

METHODS

Dataset

The research employed the Sleep-EDF Expanded database (Cassette subset), which is a freely available database that can be accessed at PhysioNet [3]. This subset contains polysomnographic (PSG) recordings from 20 participants aged between 25 and 34 years, comprising 39 night-time recordings because Subject 13 provided only one night. Each recording consists of several channels, including EEG, EOG, and EMG, as well as expert-annotated hypnograms scored according to Rechtschaffen and Kales criteria. In this research, sleep stages were grouped into five stages according to the guidelines of the American Academy of Sleep Medicine: Wake (W), N1, N2, N3 (combining Stages 3 and 4), and Rapid Eye Movement (REM).

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

Signal Preprocessing

Data processing was performed using the MNE-Python library [31], and the same steps were applied uniformly to all recordings. Only the frontal EEG channel (EEG Fpz-Cz) was retained, to maintain consistency with single-channel studies and reduce computational complexity. A 0.3 to 35 Hz bandpass filter removed DC offsets and high-frequency muscle activity while preserving the delta, theta, alpha, sigma and beta bands. Continuous signals were segmented into non-overlapping 30-second epochs aligned with the annotated hypnogram labels, and epochs with peak-to-peak amplitude exceeding 400 μV were excluded to remove movement artifacts while preserving the large-amplitude slow waves characteristic of N3. No additional artifact correction such as independent component analysis was applied, in order to avoid altering the natural signal distribution.

Feature Extraction

Eighteen handcrafted features were extracted from each epoch to capture time-domain, frequency-domain and nonlinear characteristics of the EEG. The time-domain set comprised the mean, standard deviation, variance, Hjorth mobility and Hjorth complexity. For the frequency domain, power spectral density was estimated using Welch's method, and absolute and relative powers were computed for the delta (0.5 to 4.0 Hz), theta (4.0 to 8.0 Hz), alpha (8.0 to 12.0 Hz), sigma (12.0 to 16.0 Hz) and beta (16.0 to 30.0 Hz) bands, with relative power obtained by normalizing each band against the total power across all bands. The nonlinear set comprised spectral entropy, permutation entropy and Higuchi fractal dimension, capturing spectral irregularity, time-series complexity and signal self-similarity respectively. All features were standardized using Z-score normalization fitted solely on the training folds to prevent data leakage during cross-validation.

Class Imbalance Handling Strategies

To address the severe class imbalance inherent in the dataset, in which N1 constitutes approximately 2.6% of epochs, five experimental conditions were evaluated under identical pipeline settings. The baseline condition trained the classifier on the original imbalanced distribution. Random oversampling duplicated minority samples at random until the class distribution was balanced. Random undersampling removed majority samples at random to match the size of the minority class. SMOTE [23] generated synthetic minority samples by interpolating between neighbouring minority instances. ADASYN [24] followed the same principle but concentrated synthesis on harder-to-learn minority samples according to local minority density. Resampling was applied only to the training set within each fold, so that the test set remained representative of the natural, unbalanced distribution.

Within each cross-validation fold, normalization was performed before resampling. Specifically, the StandardScaler was fitted only on the original training features and was then used to transform both the training and test features. The selected resampling strategy was subsequently applied only to the standardized training data. The test fold remained unchanged throughout the process. Random oversampling, random undersampling, SMOTE, and ADASYN were implemented using imbalanced-learn version 0.14.2 with sampling_strategy = 'auto' and random_state = 42. Random oversampling, SMOTE, and ADASYN increased minority classes toward the majority-class size, whereas random undersampling reduced non-minority classes to the minority-class size. SMOTE used five nearest neighbours (k_neighbors = 5), while ADASYN used five nearest neighbours (n_neighbors = 5).

Classification Model

A Random Forest classifier was employed as the primary algorithm because of its robustness to overfitting, its suitability for mixed feature types and its interpretability [6]. The model used 100 trees, unlimited maximum depth, the Gini impurity split criterion, a minimum of one sample per leaf and a fixed random state of 42 for reproducibility. Because Random Forest requires no architectural tuning, it is well suited to isolating the effect of data-level balancing strategies. Parallel processing was enabled through n_jobs = -1, and no class weighting was applied (class_weight = None).

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

Evaluation Protocol

A five-fold subject-wise cross-validation strategy was implemented so that all epochs of a given subject appeared exclusively in either the training or the test set of each fold, ensuring that the model was always evaluated on unseen subjects and preventing data leakage. Performance was assessed using accuracy, the macro F1-score as the primary metric, the weighted F1-score, Cohen's kappa, and per-class precision, recall and F1-score with particular emphasis on the minority N1 stage. Per-class performance was examined through precision, recall and F1-score. Aggregate out-of-fold confusion matrices were not generated under the completed cross-validation procedure.

Statistical Analysis

Non-parametric tests were used to determine whether performance differences between the five strategies were statistically significant. The Friedman test was used to assess overall differences across methods for the macro F1 and N1 F1 metrics, and the Wilcoxon signed-rank test was applied as a post-hoc pairwise comparison with Bonferroni correction to identify specific differences between the baseline and the resampling methods. A significance level of 0.05 was used, and all statistical analyses were performed with the SciPy library.

RESULTS

All results were obtained under five-fold subject-wise cross-validation. Values are reported as mean ± standard deviation across the five folds. The recordings, EEG channel, filtering, epoching, feature set, classifier configuration, fold assignment and random seed were identical across conditions; only the imbalance-handling strategy applied to the training fold varied.

Cohort and Class Imbalance

Feature extraction was completed for all 39 recordings of the 20-subject cohort. Of 106,377 segmented 30-second epochs, 270 (0.25%) exceeded the 400 µV peak-to-peak amplitude threshold and were excluded, leaving 106,107 epochs, each represented by 18 features with no missing or non-finite values. The final class distribution is summarised in Table 1.

Table 1. Class distribution of the 20-subject cohort (39 recordings, 106,107 epochs).

Sleep stage	Number of epochs	Proportion (%)	Ratio relative to N1
W	72,085	67.94	25.71
N1	2,804	2.64	1.00
N2	17,798	16.77	6.35
N3	5,703	5.37	2.03
REM	7,717	7.27	2.75
Total	106,107	100.00	

The distribution confirms severe natural imbalance, with the majority class more than 25 times larger than N1, as illustrated in Figure 1.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

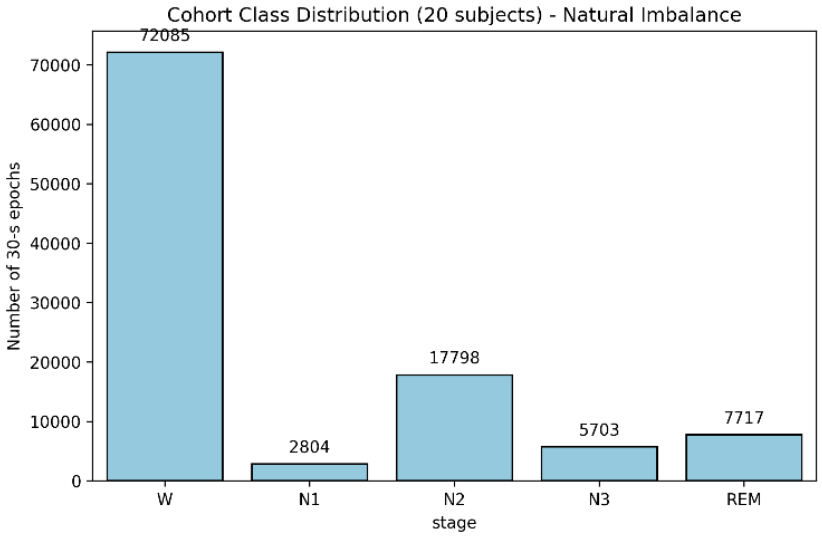

Figure 1: Cohort-level sleep stage distribution for 20 subjects, showing severe under-representation of N1.

The same pattern was present at the level of individual recordings. Figure 2 shows the distribution for a representative night (SC4001E0), in which N1 was again the smallest class, confirming that the cohort-level imbalance reflects the structure of individual nights rather than aggregation across subjects.

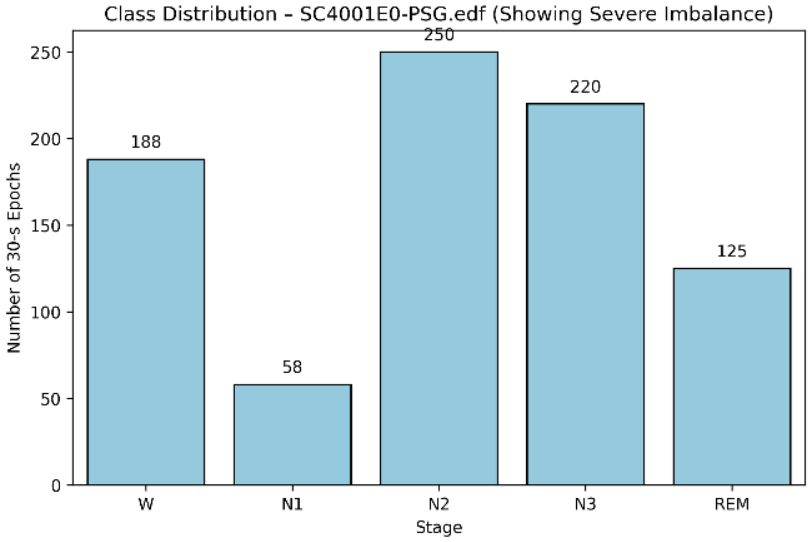

Figure 2: Sleep stage distribution for a single representative recording (SC4001E0).

Baseline Performance Without Imbalance Handling

Trained on the untouched distribution, the Random Forest classifier achieved an accuracy of 0.899 ± 0.008 and a Cohen's kappa of 0.795 ± 0.014. Per-class analysis revealed a different picture. F1-scores were 0.965 ± 0.005 for Wake, 0.841 ± 0.041 for N2 and 0.827 ± 0.041 for N3, but only 0.173 ± 0.113 for N1, with N1 recall of 0.138 ± 0.123. Approximately 86% of true N1 epochs were therefore assigned to other stages. The macro F1-score of 0.688 ± 0.027 lagged the weighted F1-score of 0.893 ± 0.009 by 0.205 and lagged accuracy by 0.211, quantifying the extent to which aggregate metrics concealed minority-stage failure. This dissociation between accuracy and minority recall is the central observation of the study and addresses RQ1.

Overall Performance Across Balancing Strategies

Table 2 reports overall performance under the five conditions, i.e. across imbalance handling strategies (mean ± standard deviation over five folds).

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

Table 2. Overall classification performance by imbalance handling strategy.

Method	Accuracy	Macro F1	Weighted F1	Cohen's kappa
None (baseline)	0.899 ± 0.008	0.688 ± 0.027	0.893 ± 0.009	0.795 ± 0.014
Random oversampling	0.895 ± 0.011	0.686 ± 0.030	0.891 ± 0.010	0.789 ± 0.022
Random undersampling	0.841 ± 0.028	0.674 ± 0.019	0.865 ± 0.020	0.710 ± 0.043
SMOTE	0.880 ± 0.021	0.693 ± 0.029	0.888 ± 0.016	0.767 ± 0.040
ADASYN	0.874 ± 0.027	0.684 ± 0.035	0.882 ± 0.022	0.758 ± 0.049

SMOTE achieved the highest observed mean macro F1-score (0.693 ± 0.029), exceeding the baseline by 0.005, a margin smaller than the between-fold standard deviation of roughly 0.03. The baseline retained the highest accuracy, weighted F1 and kappa. Random undersampling ranked lowest on every aggregate metric, with accuracy reduced by 5.8 percentage points and kappa by 0.085, because it discarded approximately 87% of the training samples in each fold. The fold-level distributions shown in Figure 3 overlap substantially, indicating that between-fold variability was comparable in magnitude to the differences between methods.

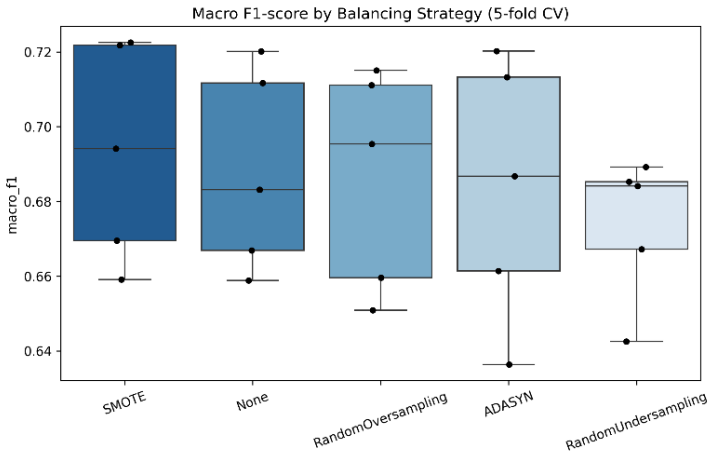

Figure 3: Macro F1-score by imbalance handling strategy across five subject-wise folds, with individual fold results overlaid.

Training cost also differed between conditions. Oversampling, SMOTE and ADASYN expanded each training fold from roughly 85,000 to about 290,000 instances, whereas undersampling produced the smallest and fastest training sets. The complete experiment of 25 models required 64.5 minutes on the CPU runtime, following 12.7 minutes of feature extraction for the full cohort.

Minority-Stage Recognition and the Precision and Recall Trade-off

Table 3 reports performance specific to stage N1 across imbalance handling strategies (mean ± standard deviation over five folds). CV denotes the coefficient of variation of N1 F1; the lowest CV, indicates the most stable strategy.

Table 3. Stage N1 performance by imbalance handling strategy.

Method	N1 recall	N1 precision	N1 F1	CV of N1 F1 (%)
None (baseline)	0.138 ± 0.123	0.327 ± 0.093	0.173 ± 0.113	65.3

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

Random oversampling	0.150 ± 0.126	0.285 ± 0.110	0.177 ± 0.109	61.6
Random undersampling	0.558 ± 0.125	0.172 ± 0.049	0.255 ± 0.048	18.8
SMOTE	0.345 ± 0.143	0.208 ± 0.076	0.244 ± 0.066	27.0
ADASYN	0.326 ± 0.144	0.219 ± 0.077	0.243 ± 0.063	25.9

Random undersampling achieved the highest N1 recall (0.558 ± 0.125), a fourfold increase over the baseline, with SMOTE (0.345 ± 0.143) and ADASYN (0.326 ± 0.144) intermediate and random oversampling only marginally above the baseline (0.150 versus 0.138). In terms of N1 F1, undersampling again ranked highest (0.255 ± 0.048), followed closely by SMOTE (0.244 ± 0.066) and ADASYN (0.243 ± 0.063). Balancing also improved stability: the coefficient of variation of N1 F1 fell from 65.3% under the baseline to 18.8% under undersampling and roughly 26% under the synthetic methods, and Figure 4 shows that the lower tail of N1 performance was raised accordingly.

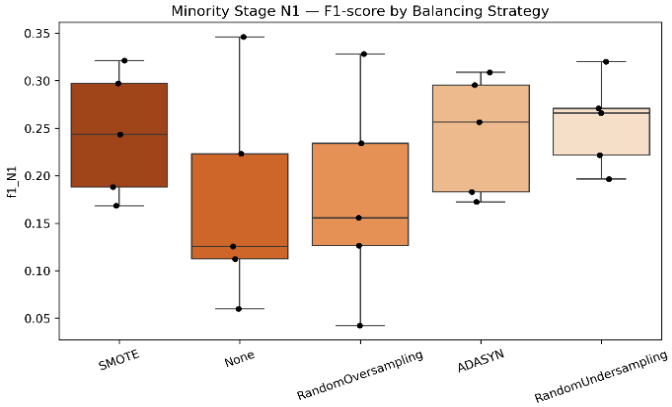

Figure 4: Stage N1 F1-score by imbalance handling strategy across five subject-wise folds.

The recall gains came at a systematic cost in precision. The baseline achieved the highest N1 precision (0.327 ± 0.093), while undersampling, with the highest recall, achieved the lowest precision (0.172 ± 0.049), a relative reduction of 47.4%. This trade-off is represented in Figure 5. The simultaneous fall in precision suggests that the extra N1 predictions were accompanied by more false-positive N1 assignments. Thus, balancing appears to have shifted the decision boundary toward N1 rather than improving its inherent separability in the feature space. The specific source stages of these false-positive predictions could not be quantified because aggregate out-of-fold confusion matrices were not generated. The largest net gain in N1 F1 was 0.082 in absolute terms, yet N1 F1 remained below 0.26 under every method and remained the poorest class throughout.

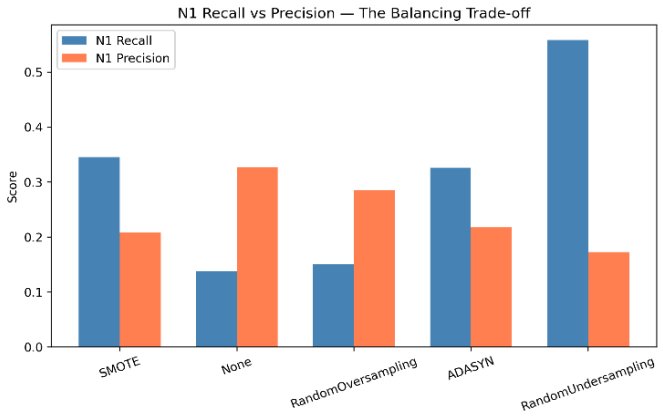

Figure 5: Mean stage N1 recall and precision for each strategy, illustrating the trade-off between the two quantities.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

Effects on Non-Minority Stages

Table 4 reports per-class F1-scores for all five sleep stages across the imbalance handling strategies (mean ± standard deviation over five folds).

Table 4. Per-class F1-scores by imbalance handling strategy.

Method	W	N1	N2	N3	REM
None (baseline)	0.965 ± 0.005	0.173 ± 0.113	0.841 ± 0.041	0.827 ± 0.041	0.634 ± 0.034
Random oversampling	0.964 ± 0.005	0.177 ± 0.109	0.839 ± 0.041	0.823 ± 0.038	0.629 ± 0.043
Random undersampling	0.936 ± 0.025	0.255 ± 0.048	0.815 ± 0.049	0.785 ± 0.062	0.578 ± 0.057
SMOTE	0.960 ± 0.015	0.244 ± 0.066	0.830 ± 0.041	0.826 ± 0.026	0.607 ± 0.048
ADASYN	0.958 ± 0.019	0.243 ± 0.063	0.813 ± 0.054	0.817 ± 0.031	0.588 ± 0.063

Every strategy that improved N1 reduced the F1-scores of the other stages, but the magnitude differed. Under undersampling, the N1 gain of 0.082 was accompanied by losses of 0.029 for Wake, 0.026 for N2, 0.042 for N3 and 0.056 for REM, producing a net negative effect on macro F1. Under SMOTE, a comparable N1 gain of 0.071 came with losses of only 0.005 to 0.027, producing the small net positive effect seen in Table 2. REM was the non-minority stage most affected by all four strategies, as shown in Figure 6.

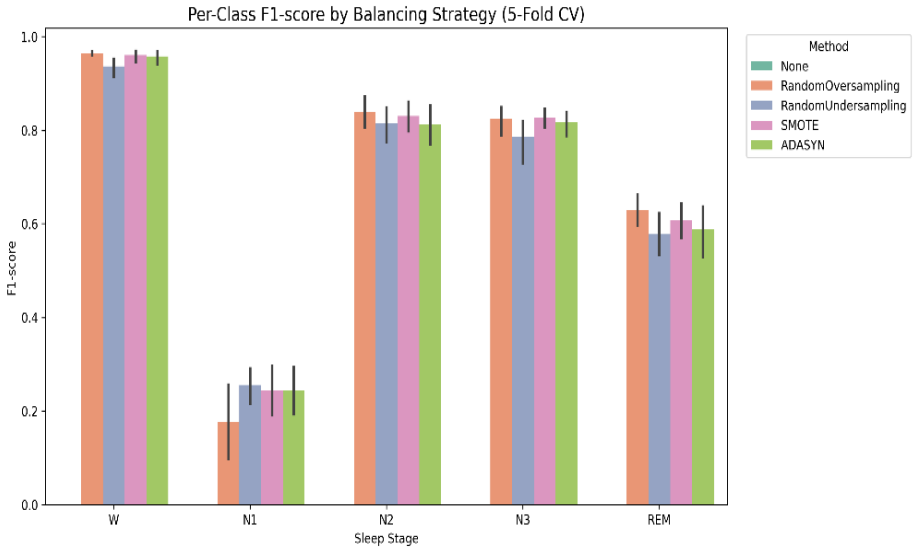

Figure 6: Per-class F1-scores by imbalance handling strategy, with error bars denoting variation across folds.

Fold Consistency and Statistical Analysis

Fold-level examination showed a consistent direction of recall improvement. Random undersampling, SMOTE and ADASYN increased N1 recall relative to the baseline in all five folds, and random oversampling did so in four. Improvement in N1 F1 occurred in four of five folds for the three stronger methods, whereas improvement in macro F1 was inconsistent, with SMOTE exceeding the baseline in only three folds. The benefit of balancing was inversely related to baseline difficulty: in the hardest fold, where baseline N1 F1 was 0.060, undersampling produced roughly a 4.4-fold gain, whereas in the fold where the baseline already achieved an N1 F1 of 0.346, every balancing strategy reduced it.

As reported in Table 5, the Friedman test results across the five imbalance handling strategies (five paired fold-level observations per method) did not detect overall differences between methods.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

Table 5. Friedman test results across imbalance handling strategies.

Metric	Chi-square statistic	Degrees of freedom	p-value
Macro F1	5.120	4	0.2752
Stage N1 F1	5.600	4	0.2311

Pairwise Wilcoxon signed-rank tests for macro F1 with Bonferroni correction likewise produced no significant comparison across the ten method pairs. The two smallest uncorrected p-values were 0.0625, for undersampling versus SMOTE and for SMOTE versus ADASYN, both corresponding to consistent directional differences across all five folds. With only five paired observations the smallest attainable two-sided Wilcoxon p-value is 0.0625, so significance at the 0.05 level is mathematically unattainable under this design. The non-significant results may reflect limited statistical power and should not be interpreted as evidence of equivalence between methods. Accordingly, the statistical analyses were interpreted as exploratory, and the method rankings should be understood as descriptive patterns rather than definitive evidence of superiority.

Summary in Relation to the Research Questions

For RQ1, imbalance strongly influenced minority recognition in a way invisible to accuracy: the baseline reached 0.899 accuracy while recovering only 13.8% of N1 epochs. For RQ2, SMOTE produced the highest observed macro F1 (0.693 ± 0.029), but this advantage was not statistically significant; the baseline remained highest on accuracy, weighted F1 and kappa. For RQ3, random undersampling gave the highest N1 recall (0.558 ± 0.125) and N1 F1 (0.255 ± 0.048), with SMOTE and ADASYN close behind and random oversampling largely ineffective. For RQ4, minority gains did not improve the overall model: every N1 improvement degraded the other stages, and only SMOTE achieved a small net positive effect on macro F1.

DISCUSSION

This study evaluated four class imbalance handling techniques within an otherwise fixed pipeline for EEG-based sleep stage classification, isolating the balancing strategy as the sole experimental variable while holding the signal processing, feature representation, classifier and evaluation protocol constant. Four principal findings emerged. First, the imbalanced baseline achieved high overall agreement with expert scoring while failing to recover the large majority of N1 epochs. Second, all effective balancing strategies increased N1 recall but reduced N1 precision, producing bounded gains in N1 F1. Third, the strategies differed in how they redistributed performance across stages, with SMOTE providing the only observed net positive effect on macro F1 and random undersampling providing the largest minority gain at the largest overall cost. Fourth, no pairwise difference reached statistical significance under the five-fold design, reflecting limited power rather than equivalence. Each finding is examined below in relation to the research questions and the existing literature.

Overall Accuracy Conceals Minority-Stage Failure

The baseline classifier attained an accuracy of 0.899 and a Cohen's kappa of 0.795, values that, reported alone, would indicate a satisfactory system. The same model, however, recovered only 13.8% of N1 epochs, and its macro F1 lagged its weighted F1 by 0.205. This dissociation directly confirms RQ1 and reproduces, in a classical machine-learning setting, the concern raised by Yang and Wang [4] that accuracy-centric evaluation conceals minority-stage weakness. The pattern is consistent with deep-learning studies in which N1 F1 remained between 46% and 55% while all other stages exceeded 83% [11], [12], [13], and with wearable-EEG ensembles in which N1 F1 of 53% contrasted with 86% to 94% for the remaining stages [16].

The absolute N1 performance obtained here, with F1 between 0.17 and 0.26, is lower than the N1 figures reported by sequence-aware deep models [11], [12], [13]. Two factors explain this gap. The present pipeline classifies each 30-second epoch independently from handcrafted features and therefore lacks the temporal transition context that recurrent and attention architectures exploit, whereas N1 is defined partly by its position in the transition from Wake to N2 under AASM rules [1]. The pipeline was also deliberately kept simple so that

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

performance differences could be attributed to the balancing strategy rather than to representation learning. The low absolute N1 scores therefore do not weaken the comparative conclusions of the study; they reinforce its central premise, namely that N1 remains the dominant failure mode of automated staging and that aggregate metrics are insufficient to reveal it.

Balancing Improves Minority Recall at the Cost of Precision

Random undersampling increased N1 recall fourfold relative to the baseline, SMOTE and ADASYN produced intermediate gains, and random oversampling produced almost none. These results address RQ3 and are consistent with prior imbalance-aware sleep studies, in which RUSBoost improved S1 recognition in single-channel staging [28] and resampling raised agreement on the imbalanced CAP database [29]. The ordering of the methods is also interpretable. Random duplication adds no new decision-boundary information and is known to risk overfitting to repeated samples [5], which explains the near-baseline behaviour of random oversampling. Synthetic interpolation in SMOTE and ADASYN smooths the minority boundary [23], [24], and removal of majority samples in undersampling produces the strongest shift in the effective class priors seen by the Random Forest, hence the largest recall gain.

The recall gains, however, were accompanied by systematic precision losses, with N1 precision falling from 0.327 under the baseline to 0.172 under undersampling. This trade-off suggests that balancing shifted the decision boundary toward N1 rather than improving the intrinsic separability of the stage in the feature space. Such behaviour is expected for N1, whose low-amplitude mixed-frequency morphology overlaps with Wake, and whose theta activity and vertex sharps overlap with N2 [1]. The practical consequence, relevant to RQ4, is that N1 F1 gains were bounded: the largest absolute improvement was 0.082, and N1 remained the poorest class under every condition.

A less discussed but practically important benefit was stability. The coefficient of variation of N1 F1 fell from 65.3% under the baseline to 18.8% under undersampling and roughly 26% under the synthetic methods. Balancing therefore made minority-stage performance more predictable across unseen subjects, which matters for deployment even when mean gains are modest.

Why the Strategies Produced Different Overall Outcomes

Although undersampling, SMOTE and ADASYN delivered similar N1 benefits, their overall outcomes diverged because of their differing effects on the majority stages. Undersampling discarded approximately 87% of the training data in each fold, degrading Wake, N2, N3 and REM and producing the lowest accuracy, weighted F1 and kappa. SMOTE preserved the majority data and confined its losses to between 0.005 and 0.027 per stage, which explains why it achieved the highest observed macro F1 and addressed RQ2, albeit by a margin of 0.005 that was smaller than the between-fold standard deviation. ADASYN offered no clear overall advantage over SMOTE despite its additional computational cost and its adaptive focus on difficult minority samples; it also exhibited slightly higher variance across folds. One possible explanation is that adaptively generated synthetic N1 epochs may have concentrated near the Wake and N2 boundaries, adding noisy minority examples that reduced precision without improving recall. Random oversampling, meanwhile, showed limited benefit for this task in the present experiment.

REM was the non-minority stage most affected by all four strategies. This is consistent with the observation that REM epochs occupy a small proportion of the data (7.27%) and share spectral features with Wake and N1, so any shift in the classifier toward the minority stages disproportionately affects REM predictions.

Subject Variability and the Power Limitation

The Friedman test yielded p-values of 0.275 for macro F1 and 0.231 for N1 F1, and no Bonferroni-corrected Wilcoxon comparison was significant. These results must be interpreted against the constraint that five paired observations make a two-sided Wilcoxon p-value below 0.0625 mathematically unattainable. The non-significance therefore reflects insufficient power rather than demonstrated equivalence. Directional consistency suggests a recurring descriptive pattern: undersampling, SMOTE and ADASYN increased N1 recall in all five

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

folds, and the two smallest uncorrected p-values of 0.0625 corresponded to comparisons in which one method outperformed the other in every fold.

The fold analysis further showed that the benefit of balancing was inversely related to baseline difficulty. In the hardest fold, where baseline N1 F1 was 0.060, undersampling produced roughly a 4.4-fold improvement, whereas in the fold where the baseline already reached an N1 F1 of 0.346, every balancing strategy reduced it. This pattern implies that between-subject variability in sleep architecture and EEG characteristics is a major determinant of whether resampling provides benefit, and that balancing is most valuable precisely when the held-out subjects are most difficult, which is the clinically relevant regime. Future studies should power their statistical design accordingly, for example through repeated cross-validation yielding fifty or more paired observations, so that differences of the magnitude observed here can be tested at conventional significance levels.

PRACTICAL RECOMMENDATIONS

The results support goal-dependent selection of imbalance handling in EEG-based sleep staging. Where overall agreement with expert scoring is the priority, the imbalanced baseline retained the highest observed accuracy, weighted F1-score, and Cohen's kappa. SMOTE may be considered when a modest descriptive increase in macro F1 is preferred, although this advantage was not statistically significant. Where detection of N1 is clinically important, for example in studies of sleep fragmentation or transition dynamics, random undersampling and SMOTE may be considered, while acknowledging the associated precision loss and, for undersampling, lower overall performance. Random oversampling did not demonstrate a clear benefit, and ADASYN showed no clear advantage over SMOTE in this experiment, despite its higher computational cost. Regardless of the strategy chosen, evaluation should report macro F1, per-class F1, and minority recall and precision, because accuracy and kappa alone concealed an 86% N1 miss rate in the baseline. Subject-wise data splitting and train-only resampling should be treated as minimum methodological requirements, since epoch-wise splitting and test-set resampling would inflate both overall and minority metrics.

Strengths, Limitations and Future Work

The principal strength of the study is its controlled design, which attributes observed differences to the balancing strategy alone, together with leakage-free subject-wise cross-validation, train-only resampling, reproducible public data and explicit statistical testing. Several limitations qualify the conclusions. The evaluation used a single public dataset of healthy young adults, a single frontal EEG channel, a single classifier and single-epoch handcrafted features without temporal context, so generalisation to patient populations, other montages, other classifiers and sequence-aware models remains to be established. The five-fold design limited statistical power. In addition, aggregate out-of-fold confusion matrices were not generated due to computational constraints. Consequently, the source stages of false-positive N1 predictions could not be directly quantified, and interpretation of the precision-recall trade-off was based on per-class metrics rather than detailed error distributions. Future research should address these limitations with techniques such as repeated cross-validation, validation classifiers like XGBoost and gradient boosting machines, cohorts such as SHHS and MASS, sequence-aware features that include temporal transition context, algorithmic strategies such as cost-sensitive learning and class-weighted loss functions, and probability threshold calibration.

CONCLUSION

This study provided a controlled, data-centric evaluation of class imbalance handling in automated EEG-based sleep stage classification. By treating the resampling method as the only experimental variable in a standard feature-based pipeline, the results showed that the natural imbalance in Sleep-EDF data allowed a Random Forest classifier to achieve an accuracy of 0.899, while only 13.8% of N1 epochs were successfully recovered, indicating that aggregate metrics can mask minority-class underperformance. Standard resampling techniques improved N1 recall, yet the fundamental difficulty of N1 classification remained unresolved: each method that raised recall also reduced N1 precision, which suggests that balancing shifted the classifier operating point rather than improving the inherent separability of N1. SMOTE provided the most favourable descriptive trade-off by achieving the highest observed macro F1-score while maintaining competitive accuracy and kappa, whereas random undersampling provided the highest N1 recall at the expense of overall performance. ADASYN offered

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

no clear advantage over SMOTE, and random oversampling was largely ineffective. Although no pairwise difference reached statistical significance under the five-fold design, the recurring direction of effects across folds suggests that the observed patterns may have practical relevance and warrant validation through stronger evaluation protocols. The results provide evidence-based recommendations for practitioners, highlight the importance of per-class assessment in sleep staging, and indicate that future efforts to address class imbalance should focus on representation learning and algorithm-level remedies aimed at increasing the discriminability of transitional sleep stages.

Declaration Statements

Author Contributions

Blessing C. Uzo conceived the study, designed the methodology, implemented the feature extraction and machine learning pipeline, performed the experiments, analyzed the results, and wrote the original draft.

Ponsak S. Bande contributed to the literature review and statistical analysis, and co-wrote sections of the manuscript.

Micah Okpara assisted with data preprocessing, validation of the experimental setup, and interpretation of the clinical implications.

John O. Onyima provided critical feedback on the study design, helped with the discussion of limitations, and reviewed and edited the manuscript.

Nnaemeka V. Ugwu supervised the project, secured resources, and edited and approved the final manuscript.

All authors read and approved the final version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Ethical review and approval were waived for this study because it used only the publicly available, de-identified Sleep-EDF Expanded database distributed through PhysioNet. The original data collection was conducted with informed consent and ethical approval as described in the original dataset publications [2], [3].

Informed Consent Statement

Not applicable. The study involved no new human participants and used only pre-existing, anonymized public data.

Data Availability Statement

The Sleep-EDF Expanded database analysed in this study is openly available on PhysioNet at https://physionet.org/content/sleep-edfx/1.0.0/ (reference [3]). The feature matrices, per-fold results and analysis scripts supporting the findings are available from the corresponding author upon reasonable request.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the contributors of the Sleep-EDF Expanded database and the PhysioNet platform for making the polysomnographic recordings publicly available, and the developers of the MNE-Python, scikit-learn and imbalanced-learn libraries used in this work.

Conflicts of Interest

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

REFERENCES

1. Silber, M. H., Ancoli-Israel, S., Bonnet, M. L., Chokroverty, S., Grigg-Damberger, M., Hirshkowitz, M., Kapen, S., Keenan, T., Kryger, S. H., Penzel, T., Pressman, M., & Iber, C. (2007). The visual scoring of sleep in adults. Journal of Clinical Sleep Medicine, 3(2), 121–131.
2. Kemp, B., Zwinderman, A. H., Tuk, B., Kamphuisen, H. A. C., & Oberye, J. J. L. (2000). Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG. IEEE Transactions on Biomedical Engineering, 47(9), 1185–1194. https://doi.org/10.1109/10.867928
3. Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation, 101(23), e215–e220. https://doi.org/10.1161/01.CIR.101.23.e215
4. Yang, K., & Wang, H. (2026). Deep learning for EEG-based sleep stage classification: A review. Medical & Biological Engineering & Computing, 64(6), 2017–2043. https://doi.org/10.1007/s11517-026-03583-3
5. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239
6. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
7. Jain, R., & Ganesan, R. A. (2021). An efficient sleep scoring method using visibility graph and temporal features of single-channel EEG. In Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 6306–6309). https://doi.org/10.1109/EMBC46164.2021.9630863
8. Dogaheh, S. B., & Moradi, M. H. (2020). Automatic sleep stage classification based on Dreem headband's signals. In Proceedings of the 27th National and 5th International Iranian Conference on Biomedical Engineering (ICBME) (pp. 259–263). https://doi.org/10.1109/ICBME51989.2020.9319415
9. Stenwig, H., Soler, A., Furuki, J., Suzuki, Y., Abe, T., & Molinas, M. (2022). Automatic sleep stage classification with optimized selection of EEG channels. In Proceedings of the 21st IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1708–1715). https://doi.org/10.1109/ICMLA55696.2022.00262
10. Rathod, T., Satapathy, S. K., Das, N., Mohapatra, R. K., Sahu, S. K., & Shah, V. R. (2024). Automated sleep staging system using EEG signal feature-based classification by machine learning techniques. In Proceedings of the IEEE 8th International Conference on Information and Communication Technology (CICT) (pp. 1–6). https://doi.org/10.1109/CICT64037.2024.10899578
11. Sadık, E. Ş. (2026). A hybrid CNN-BiGRU model with multi-head self-attention for EEG-based sleep stage classification. In Proceedings of the 8th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA) (pp. 1–6). https://doi.org/10.1109/ICHORA69329.2026.11537080
12. Intarawichian, S., Thiennviboon, P., Laothamatas, J., & Sungkarat, W. (2026). EnsembleNet: Single-channel EEG sleep stage classification based on ensemble architecture of deep convolutional neural networks. IEEE Access, 14, 14449–14469. https://doi.org/10.1109/ACCESS.2026.3655811
13. Saowapark, W., & Jirakittayakorn, N. (2026). SleepTNet: Automatic sleep stage classification with transition model using multi-channel EEG. IEEE Access, 14, 63894–63915. https://doi.org/10.1109/ACCESS.2026.3686667
14. Zhang, Z., Lin, B.-S., Peng, C.-W., & Lin, B.-S. (2024). Multi-modal sleep stage classification with two-stream encoder-decoder. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 32, 2096–2105. https://doi.org/10.1109/TNSRE.2024.3394738
15. Dip, M. S. S., Motin, M. A., Karmakar, C., Penzel, T., & Palaniswami, M. (2026). NeuroSleepNet: An explainable multi-head attention-based framework with spatial and multi-scale independent temporal

context learning for automatic sleep stage scoring. IEEE Access, 14, 4341–4355. https://doi.org/10.1109/ACCESS.2025.3640231

16. Salfi, F., Corigliano, D., Amicucci, G., Mombelli, S., D'Atri, A., Axelsson, J., & Ferrara, M. (2026). The potential of ensemble-based automated sleep staging on single-channel EEG signal from a wearable device. Journal of Sleep Research, 35(4), e70282. https://doi.org/10.1111/jsr.70282
17. Zhou, D., Xu, Q., Wang, J., Xu, H., Kettunen, L., Chang, Z., & Cong, F. (2022). Alleviating class imbalance problem in automatic sleep stage classification. IEEE Transactions on Instrumentation and Measurement, 71, 1–12. https://doi.org/10.1109/TIM.2022.3191710
18. Xu, Q., Zhou, D., Wang, J., Shen, J., Kettunen, L., & Cong, F. (2022). Convolutional neural network based sleep stage classification with class imbalance. In Proceedings of the International Joint Conference on Neural Networks (IJCNN) (pp. 1–6). https://doi.org/10.1109/IJCNN55064.2022.9892741
19. Lee, C.-H., Kim, H.-J., Heo, J.-W., Kim, H., & Kim, D.-J. (2021). Improving sleep stage classification performance by single-channel EEG data augmentation via spectral band blending. In Proceedings of the 9th International Winter Conference on Brain-Computer Interface (BCI) (pp. 1–5). https://doi.org/10.1109/BCI51272.2021.9385297
20. Yu, L., Tang, P., Jiang, Z., & Zhang, X. (2023). Denoise enhanced neural network with efficient data generation for automatic sleep stage classification of class imbalance. In Proceedings of the International Joint Conference on Neural Networks (IJCNN) (pp. 1–8). https://doi.org/10.1109/IJCNN54540.2023.10191282
21. Huang, X., Schmelter, F., Irshad, M. T., Piet, A., Nisar, M. A., Sina, C., & Grzegorzek, M. (2023). Optimizing sleep staging on multimodal time series: Leveraging borderline synthetic minority oversampling technique and supervised convolutional contrastive learning. Computers in Biology and Medicine, 166, 107501. https://doi.org/10.1016/j.compbiomed.2023.107501
22. Li, Y., Peng, C., Zhang, Y., Zhang, Y., & Lo, B. (2022). Adversarial learning for semi-supervised pediatric sleep staging with single-EEG channel. Methods, 204, 84–91. https://doi.org/10.1016/j.ymeth.2022.03.013
23. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953
24. He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Proceedings of the International Joint Conference on Neural Networks (IJCNN) (pp. 1322–1328). https://doi.org/10.1109/IJCNN.2008.4633969
25. Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research, 18(17), 1–5.
26. Liu, Z., Luo, S., Lu, Y., Zhang, Y., Jiang, L., & Xiao, H. (2023). Extracting multi-scale and salient features by MSE based U-structure and CBAM for sleep staging. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31, 31–38. https://doi.org/10.1109/TNSRE.2022.3216111
27. Liu, Z., Wu, Y., Tan, K., & Gao, Y. (2026). MCTSleepNet: A multiscale waveform and composite attention network with temporal dependency learning for robust EEG-based sleep staging. Cognitive Neurodynamics, 20(1), 50. https://doi.org/10.1007/s11571-026-10423-5
28. Rahman, M. M., Bhuiyan, M. I. H., & Hassan, A. R. (2018). Sleep stage classification using single-channel EOG. Computers in Biology and Medicine, 102, 211–220. https://doi.org/10.1016/j.compbiomed.2018.08.022
29. Sharma, M., Tiwari, J., & Acharya, U. R. (2021). Automatic sleep-stage scoring in healthy and sleep disorder patients using optimal wavelet filter bank technique with EEG signals. International Journal of Environmental Research and Public Health, 18(6), 3087. https://doi.org/10.3390/ijerph18063087
30. Park, C., Byun, J. I., Choi, S. H., & Shin, W. C. (2025). Machine learning classifier solving the problem of sleep stage imbalance between overnight sleep. Biomedical Engineering Letters, 15(3), 513–523. https://doi.org/10.1007/s13534-025-00466-8
31. Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. Frontiers in Neuroscience, 7, 267. https://doi.org/10.3389/fnins.2013.00267