The Eurasia Proceedings of Health, Environment and Life Sciences (EPHELS)
ISSN: 2791-8033

The Eurasia Proceedings of Health, Environment and Life Sciences (EPHELS), 2026

Volume 21, Pages 1-10

ICVALS 2026: International Conference on Veterinary, Agriculture and Life Sciences

Comparison of SMOTE and GAN Data Augmentation Methods in Sleep Stage Classification

Sena Ceper-Ozcelik
Konya Technical University

Gulay Tezel
Konya Technical University

Abstract: Sleep is a vital process that occupies a significant portion of human life. Poor sleep quality can lead to the emergence of neurological and cardiovascular diseases. Therefore, sleep staging plays a crucial role in the diagnosis of sleep-related disorders. Traditionally, sleep staging is performed manually by experts in clinical settings by evaluating biomedical signals recorded via polysomnography (PSG) in accordance with the American Academy of Sleep Medicine (AASM) standards. However, since this process is both labor-intensive and prone to human error, its automation is of great importance. Excluding the wake stage, sleep is divided into two main phases: Rapid Eye Movement (REM) and Non-Rapid Eye Movement (NREM). The distribution of these stages throughout the sleep period provides essential information regarding sleep quality. However, these stages are not distributed evenly, which leads to the Class Imbalance Problem (CIP) in automatic classification and can negatively affect model performance. In this study, the MIT-BIH Polysomnographic Database—a challenging dataset due to significant class imbalance and signal variability—was utilized to automatically determine sleep stages. The effects of data augmentation approaches, specifically the classical SMOTE method and the deep learning-based GAN, were investigated using a proposed end-to-end CNN - BiLSTM architecture. Performance evaluations were conducted for both 3-class and 5-class staging scenarios. The experimental results are as follows: The accuracy values achieved for 3-class and 5-class staging were 87.58% and 75.85% without data balancing, 85.91% and 71.82% with SMOTE, and 86.60% and 75.31% with GAN-based augmentation, respectively. The results indicate that data augmentation methods do not always yield performance improvements in sleep staging tasks, particularly due to the complex temporal structure of EEG signals. Nevertheless, the 87.58% accuracy achieved without any data balancing demonstrates that the proposed CNN–BiLSTM architecture possesses a high capacity for learning sleep staging patterns.

Keywords: EEG, Sleep stages, Deep learning, Data augmentation

Introduction

Sleep is very important for human life because if there is any obstacle to regular sleep, bodily functions tend to deteriorate. Sleep quality affects academic performance, and according to research conducted by the National Sleep Foundation, sleep disorders are observed in 40% of individuals with various diseases and disorders, as indicated in the literature (Wara et al., 2025). Sleep disorders are closely associated with many neurological and psychiatric diseases, such as epilepsy and Alzheimer's disease (Gashti & Farjamnia, 2025). Biophysiological signals such as Electroencephalography (EEG), Electrooculography (EOG), Electromyography (EMG), and Electrocardiography (ECG), recorded using a Polysomnography (PSG) device throughout the night, are used to detect sleep-related disorders. This process is referred to as sleep staging or sleep scoring (Phan & Mikkelsen, 2022). In clinical settings, sleep stages are manually scored by sleep specialists using overnight polysomnography (PSG) according to the rules defined by the American Academy of Sleep Medicine (AASM) (Ruehland & Rochford, 2009). In a healthy individual, sleep is divided into two main stages, excluding the

- This is an Open Access article distributed under the terms of the Creative Commons Attribution-Noncommercial 4.0 Unported License, permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
- Selection and peer-review under responsibility of the Organizing Committee of the Conference
©2026 Published by ISRES Publishing: www.isres.org

wakefulness stage: Non-Rapid Eye Movement (NREM) sleep and Rapid Eye Movement (REM) sleep. NREM sleep usually occurs in the early stages of sleep and consists of three sub-stages: NREM1, NREM2, and NREM3. This manual scoring process is time-consuming, costly, and inconsistent (Yu et al., 2025). For this reason, automating sleep staging is of great importance in terms of reducing the workload of specialists and contributing to the prevention of possible misdiagnoses. There are studies in the literature on sleep staging to automate this process can be examined in two categories: those performed using machine learning algorithms (An et al., 2021; Ben Hamouda et al., 2024; Fatimah et al., 2022; Huang et al., 2020; Li et al., 2022a; Liu et al., 2021; Permana & Iramina, 2024; Taran et al., 2020) and those performed using deep learning methods (Chen et al., 2023; Efe & Ozsen, 2023; Goshtasbi et al., 2022; Huttunen et al., 2023; Li et al., 2022b; Mostafaei et al., 2024; Perslev et al., 2021; Ren et al., 2025; Xu et al., 2025).

In recent years, numerous studies have made significant contributions to the problem of automatic sleep stage classification (Fatimah et al., 2022; Mostafaei et al., 2024; Taran et al., 2020; Zan & Yildiz, 2023; Zhao et al., 2022). Despite achieving promising performance, these approaches still suffer from the Class Imbalance Problem (CIP). In a typical sleep recording, the duration of each sleep stage is not uniformly distributed. In particular, N2 stage constitutes the majority class, accounting for approximately 45–55% of total sleep duration, whereas N1 stage represents only about 2–5% (Fan et al., 2020). When deep neural networks are trained with an imbalanced dataset, the network encounters the majority class samples more frequently; therefore, it performs better in recognizing the majority class. In contrast, it shows almost no success in learning the minority class samples. This situation negatively affects both the overall performance and the network's ability to correctly recognize different classes (Xu et al., 2022). Therefore, numerous studies in the literature have focused on data balancing techniques to address the issue of class imbalance (Cheng et al., 2024; Fan et al., 2020; Fan & Wang, 2023; Lee et al., 2021; Permana & Iramina, 2024; Sun et al., 2019; Supratak et al., 2017). Existing approaches to the CIP problem can be broadly categorized into two main groups: undersampling the majority class and oversampling the minority class (Mishra & Jaiswal, 2024).

With the aim of reducing the class imbalance problem, Supratak et al. (2017), duplicated data belonging to minority classes in the training set to equalize the number of examples in the majority class. In another study, Sun et al. (2019) first increased the number of examples belonging to minority classes in the training data using the Synthetic Minority Over-sampling Technique (SMOTE) method, then diversified the data morphology by adding noise to the generated synthetic data. Permana and Iramina (2024) conducted a study utilizing SMOTE to augment the minority class. However, the application of SMOTE to the entire dataset prior to the train–test split may introduce data leakage, thereby potentially resulting in overly optimistic performance evaluations. Fan et al. (2020) proposed five different data augmentation approaches to improve the classification performance of the N1 phase in the sleep staging process and identified SMOTE and Adaptive Synthetic Sampling (ADASYN) as the most commonly used data augmentation techniques in the literature. However, noting that such methods may not be suitable for time-series data such as EEG signals, they focused on GAN-based data generation. Although this method improved overall performance, a decrease in the classification accuracy of the N1 phase was observed. Similarly, in the study conducted by Cheng et al. (2024), GAN—a deep learning–based data augmentation technique—was employed to enhance the classification performance of the N1 stage by increasing the number of N1 samples. However, the performance of the N1 stage remained relatively low. As an innovation in data augmentation methods, Lee et al. (2021) aimed to fill signal interruptions in EEG—caused by sensor detachment, motion artifacts, or technical failures—using the GAN method while preserving the context of the signal. Fan and Wang (2023) filtered EEG signals using a band-pass filter and enabled the GAN architecture to generate data only within specific frequency bands for sleep stages. The authors argued that, for the GAN method to be effective, the process should be conducted in the frequency domain. However, when the results were examined, it was observed that this approach did not provide a significant improvement in overall performance at the expected level. These studies in literature constituted the primary motivation for the present research. However, unlike these studies, an end-to-end approach was adopted in this work, since sleep experts do not apply any preprocessing steps during manual sleep staging. In this context, the aim of this study is to perform a comparative analysis of the traditional SMOTE method and the relatively newer GAN approach on the MIT-BIH dataset, which is an imbalanced dataset and has been rarely used for sleep staging in literature.

Method

In this study, an end-to-end CNN-BiLSTM-based deep learning architecture was proposed for both three-class and five-class sleep staging. Furthermore, the class imbalance problem in sleep staging data was addressed by applying SMOTE and GAN-based data balancing methods only to the training dataset and the effect of these data balancing techniques were comparatively analyzed. Firstly, the dataset was divided into three parts:

training, validation, and test sets. The training data was used in three different versions: the original (imbalanced) data, data balanced using SMOTE, and data balanced using GAN. No balancing procedure was applied to the validation and test datasets. These three different training datasets were fed into the model separately, and the model was trained and tested three times using each training dataset. The block diagram of the study is presented in Figure 1.

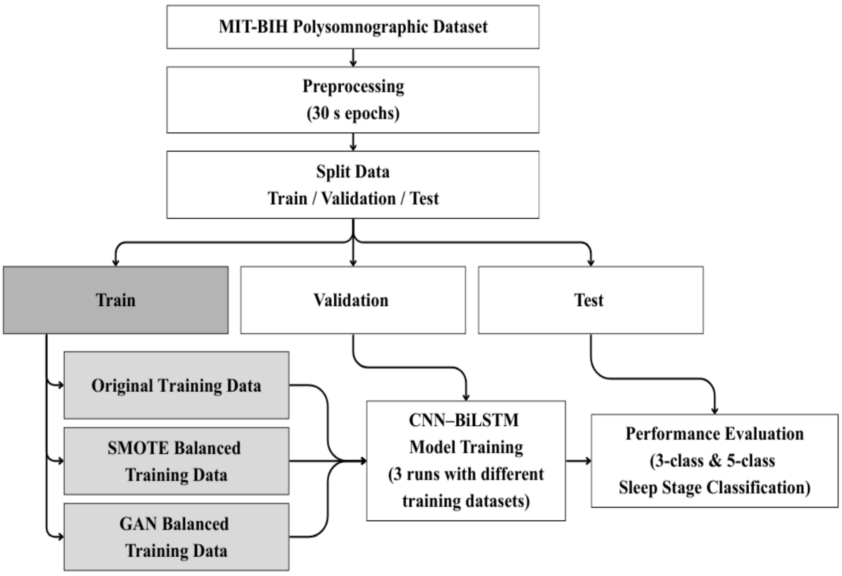


Figure 1. General structure of this study

Data and Preprocessing

The MIT-BIH Polysomnography dataset is a publicly available dataset accessible via the PhysioNet website and contains signals obtained from PSG devices belonging to 18 patients. However, in this dataset, the EEG signal was not recorded from the same channel for each patient. For the 18 patients included in the study, EEG signals were recorded from the C3O1 channel for 11 patients, the C4A1 channel for 5 patients, and the O2A1 channel for 2 patients. In the MIT-BIH dataset, EEG signals are sampled at 250 Hz and segmented into 30-second epochs. Each epoch is annotated with one of the following labels: W (Wake), N1 (NREM1), N2 (NREM2), N3 (NREM3), N4 (NREM4), R (REM), and MT (Movement Time). As a preprocessing procedure, while preparing the dataset for sleep staging, the NREM3 and NREM4 stages were merged and labeled as NREM3 (Berry et al., 2017) and epochs with MT label were removed from the dataset. The resulting number of sleep stages was presented in Table 1. As shown in Table 1, while the NREM2 stage has the highest number of epochs with 3887 samples, the NREM3 stage has the lowest number with 664 epochs.

Table 1. Total number of stages

	Wake	NREM1	NREM2	NREM3	REM	Total
5 class	3119	1815	3887	664	700	10185
3 class	3119	6366			700	10185

Data Augmentation

After dividing the datasets into training, validation, and test sets, the data augmentation methods SMOTE and GAN were applied only to the training data. Within the scope of the study, the performances of the SMOTE and GAN methods in sleep staging were evaluated by using the MIT-BIH dataset.

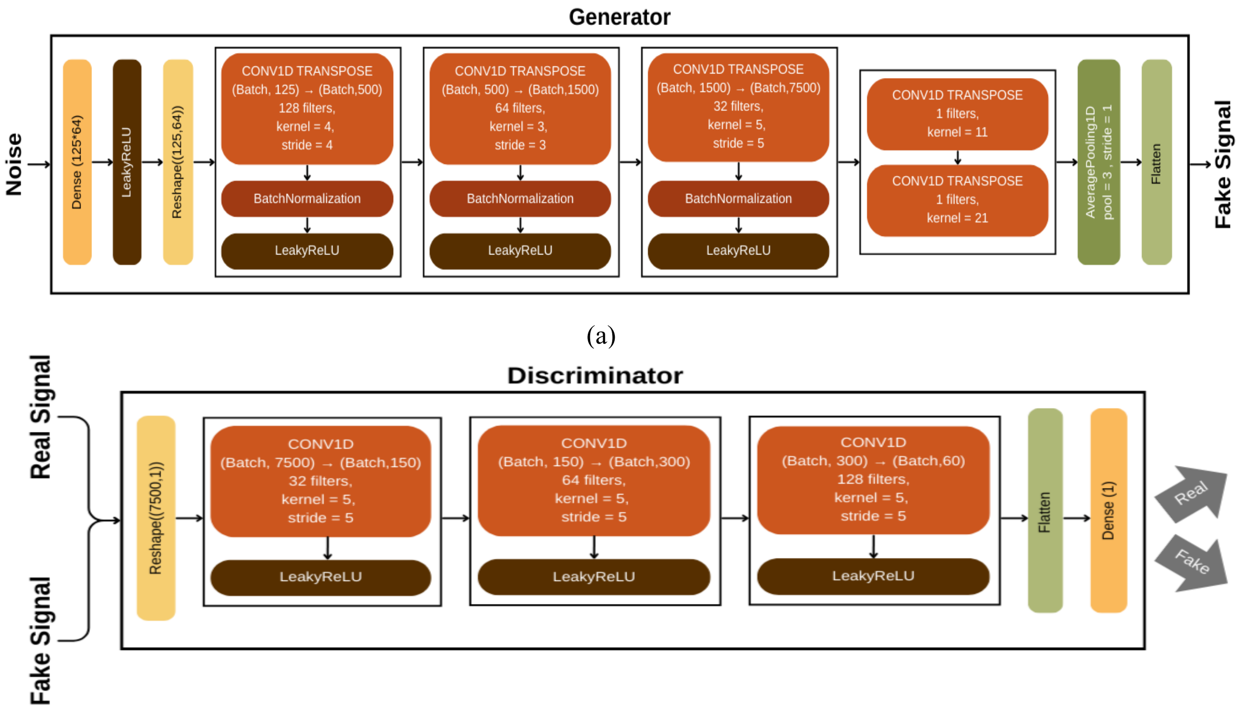
Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE is an oversampling method that generates synthetic data for the minority class in order to balance class distribution. Unlike traditional oversampling methods that increase data by duplicating existing minority class samples, it creates new synthetic samples by interpolating between instances within the minority class (Dokare

& Gupta, 2025). First, a random sample is selected from the minority class, and its k nearest neighbors are identified. In this study, the value of k was set to 5. Then, the difference vector between the selected sample and one of its neighbors is computed, and this difference is multiplied by a randomly chosen coefficient between 0 and 1 and added to the original sample. This process is repeated until the number of minority class samples reaches that of the majority class. As shown in Table 1, when the dataset is split into 20% test and 10% validation for 5 classes, the test set contains 624 Wake, 363 NREM1, 777 NREM2, 133 NREM3, and 140 REM stages. In the training set of raw datasets, there are 2245 epochs for Wake, 1307 for NREM1, 2799 for NREM2, 478 for NREM3, and 504 for REM. When SMOTE was applied only to the training data, all classes were balanced to match the class with the highest number of epochs, which is NREM2 (2799 epochs). As a result, the total number of epochs in the training set increased to 13995. Similarly, under the three-class staging scheme, the dataset is divided into training, validation, and test subsets, where the raw test set includes 624 Wake, 1273 NREM, and 140 REM epochs. In this case, the remaining 2245 epochs for Wake, 4584 for NREM, and 504 for REM formed training subset. After applying SMOTE, the number of epochs in each class within the training set was balanced to match the class with the highest number of epochs (i.e., the NREM stage; 4584 epochs); as a result, a total of 13752 epochs was reached.

Generative Adversarial Networks (GAN)

GAN consists of two competing deep neural networks that try to outperform each other. The first network, discriminator, is trained to distinguish between real data and fake data generated by the generator. The second network, generator, takes random noise signals as input and transforms them into synthetic data that resembles real data. The objective of generator is to minimize the loss function by producing data that is realistic enough to deceive the discriminator (Hartmann et al., 2018). As previously stated, while the training subset of 5-class configuration consists of 2245 Wake, 1307 NREM1, 2799 NREM2, 478 NREM3, and 504 REM epochs, the training subset of 3-class configuration includes 2245 Wake, 4584 NREM, and 504 REM epochs. During the GAN implementation, separate GAN models were trained for the NREM3 and REM stages in the 5-class setting, and 1000 synthetic epochs were generated from each model. In the 3-class setting, since only the REM stage was underrepresented, 2000 REM epochs were generated from the model. After data generation, the total number of epochs in the training subset increased from 7333 to 9333 in both the 5-class and 3-class configurations, following the generation of 2000 synthetic samples. The GAN architecture used in this study is presented in Figure 2.



(a)

(b)

Figure 2. a. The generator structure used in the GAN model b. the discriminator structure used in the GAN model.

Deep Learning Architecture

The deep learning architecture used in this study was constructed by combining Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) algorithms (Figure 3). The convolutional neural network model consists of three fundamental layers: convolutional layers, pooling layers, and fully connected layers. CNNs are composed of sequentially arranged blocks, where each block extracts new features from the information obtained from the previous one (Efe, 2022). The feature maps obtained after the convolutional layers were then fed into the BiLSTM layer to model temporal dependencies. The BiLSTM structure processes the data in both forward and backward directions, enabling more effective learning of long-term dependencies. In this way, the model can achieve higher classification accuracy by considering not only spatial features but also temporal relationships.

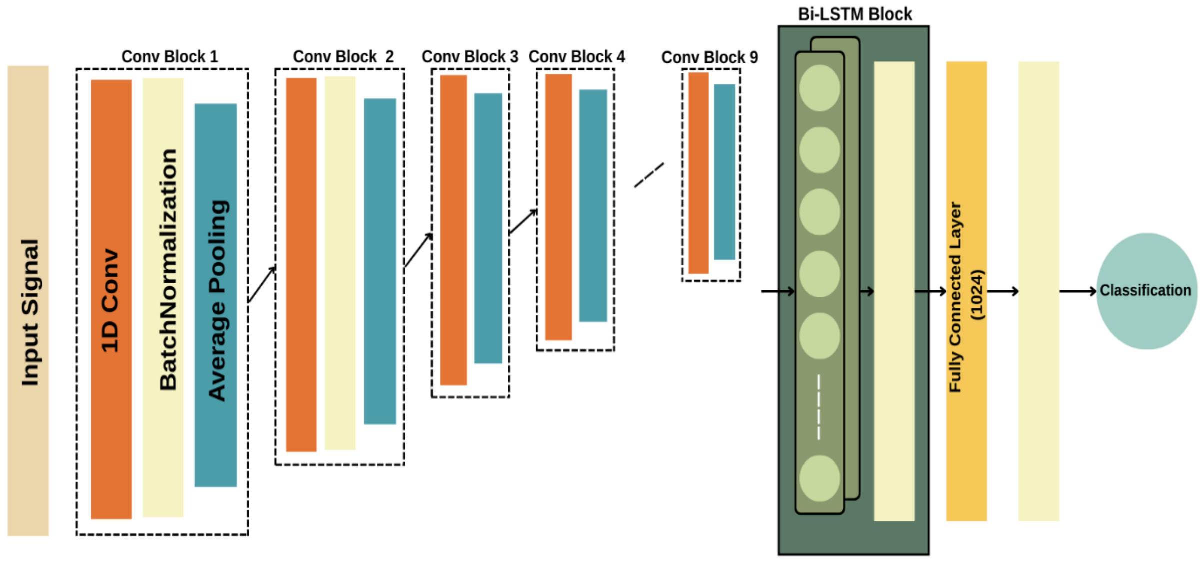


Figure 3. The deep learning architecture used in this study

The deep learning architecture used in this study consists of nine convolutional blocks. Batch normalization was applied in the first two convolutional blocks. Following the convolutional blocks, two BiLSTM layers were incorporated into the model. The blocks between Conv Block 3 and Conv Block 9 consist of a convolutional layer followed by a pooling layer. The number of filters in the convolutional blocks are 16, 32, 64, 128, 256, 512, 512, 512, and 512, respectively, with a filter size of 1×3. The number of units in the BiLSTM layers are 256 and 128, respectively. To evaluate the performance of the study on a class-wise basis, Accuracy, Recall, Precision, and F1-Score metrics were employed (Mostafaei et al., 2024; Satapathy et al., 2023). Accuracy measures the overall correctness of the model, while Recall indicates the ability to correctly identify positive instances, Precision reflects the proportion of correctly predicted positive instances among all predicted positives, and F1-Score provides a balanced measure by combining Precision and Recall.

Results and Discussion

Firstly, as shown in Figure 1, the dataset was divided into 70% training, 10% validation, and 20% test subsets. Using this split, both 3-class and 5-class sleep staging tasks were performed with the proposed deep learning model. In the first phase, the model was trained and tested using the raw (imbalanced) training data without applying any balancing technique. Subsequently, class imbalance was addressed by applying the SMOTE method only to the training dataset, and both 3-class and 5-class sleep staging processes were repeated. Finally, the GAN model was trained separately for the NREM3 and REM classes using only the training dataset, and 1000 synthetic samples were generated for each class. These generated synthetic data were then incorporated into the training dataset, and the model was retrained for the 5-class sleep staging task. For the 3-class staging, in order to perform GAN-based balancing, 2000 synthetic samples belonging to the REM stage were generated and added only to the training dataset. Following this process, the model was retrained for the 3-class sleep staging task, and sleep stage classification was performed.

3-Class Sleep Staging Results

The performance metrics obtained for 3-class staging using the raw, SMOTE-balanced, and GAN-balanced datasets are presented in Table 2, while the corresponding confusion matrices are shown in Table 3. The reported results correspond to the test dataset.

Table 2. Performance metrics of 3 class

	Accuracy	Precision	Recall	F1 score
Raw	87.58%	0.82	0.81	0.81
SMOTE	85.91%	0.78	0.80	0.79
GAN	86.60%	0.80	0.79	0.80

When the results presented in Table 2 are examined, it is observed that the raw dataset achieved the highest accuracy (87.58%) compared to the SMOTE- and GAN-based balancing methods. A slight decrease in overall accuracy was observed in the GAN-balanced dataset (86.60%), while the accuracy obtained with the SMOTE-balanced dataset was approximately 1% lower than the GAN-balanced dataset and nearly 2% lower than the accuracy achieved with the raw dataset. When the F1-score values are examined, scores of 0.81, 0.79, and 0.80 were obtained for the raw, SMOTE-balanced, and GAN-balanced datasets, respectively. These results indicate that the model demonstrates a similar and consistent performance across all experimental groups in recognizing the classes.

Table 3. Confusion matrices of all experiments for 3 class

			Predicted Label		
			Wake	NREM	REM
True Label	Raw	Wake	513	98	13
		NREM	66	1174	33
		REM	9	34	97
	SMOTE	Wake	526	85	13
		NREM	99	1131	43
		REM	16	31	93
	GAN	Wake	504	106	14
		NREM	70	1169	34
		REM	6	43	91

The analysis of the confusion matrices presented in Table 3 reveals the effects of data balancing methods on classification performance. When the model was trained using the dataset balanced with the SMOTE algorithm, it correctly predicted 526 out of 624 epochs belonging to the Wake stage, achieving an accuracy of 84%. In contrast, the highest performance for the NREM and REM stages was obtained using the original (raw) dataset. For the NREM stage, 1174 out of 1273 epochs were correctly classified, resulting in an accuracy of 92%. Similarly, for the REM stage, 97 out of 140 epochs were correctly classified, corresponding to a classification accuracy of 69%. With the dataset balanced using the GAN-based method, the model achieved its second-best performance in the NREM stage, with 1169 correctly classified epochs and an accuracy of 91.83%. However, the performance for the Wake and REM stages was lower compared to the other experimental scenarios. Another notable finding is that, across all three experimental scenarios, the model struggled to distinguish the REM stage and frequently confused it with the NREM stage.

5-Class Sleep Staging Results

For the 5-class staging task, the dataset was first divided into training, validation, and test sets, and the model was initially trained using the raw data. Subsequently, the model was retrained by applying the SMOTE method only to the training dataset. In the final stage, in order to improve the recognition of the NREM3 and REM stages in the training dataset, a GAN model was trained, and 1000 synthetic samples were generated for each class. These synthetic data were then incorporated into the training dataset, and the model was retrained. The performance metrics of the three experimental setups are presented in Table 4, while the corresponding confusion matrices are provided in Table 5.

Table 4. Performance metrics of 5 class

	Accuracy	Precision	Recall	F1 score
Raw	75.85%	0.82	0.81	0.81
SMOTE	71.82%	0.78	0.80	0.79
GAN	75.31%	0.71	0.72	0.72

When Table 4 is examined, it is observed that the highest accuracy and F1 score are obtained from the raw dataset. Although the GAN-based balanced dataset produces results that are quite close to the raw data in terms of accuracy, the lower precision, recall, and especially F1 score indicate that the model's ability to distinguish between classes has weakened. On the other hand, although the SMOTE method has the lowest accuracy, its higher F1 score compared to GAN suggests that it preserves the balance between classes better and provides a more balanced classification performance.

Table 5. Confusion matrices of all experiments for 5 class

			Predicted Label				
			Wake	NREM1	NREM2	NREM3	REM
True Label	Raw	Wake	530	44	34	0	16
		NREM1	61	190	91	0	21
		NREM2	29	79	622	29	18
		NREM3	0	2	34	97	0
		REM	8	10	16	0	106
	SMOTE	Wake	506	72	27	1	18
		NREM1	61	208	71	0	23
		NREM2	30	112	560	56	19
		NREM3	3	0	32	98	0
		REM	11	15	22	1	91
	GAN	Wake	528	44	36	0	16
		NREM1	69	190	88	0	16
		NREM2	31	67	617	41	21
		NREM3	0	0	29	104	0
		REM	12	13	20	0	95

The class-based analysis of the confusion matrices presented in Table 5 indicates statistically significant differences between the datasets in terms of the number of correctly classified samples. When examining the total of 624 test epochs belonging to the Wake stage, the highest performance was achieved using the original (raw) dataset, where 530 epochs were correctly predicted, resulting in an accuracy of 84%. For the 363 epochs in the NREM1 stage, the best performance was observed with the SMOTE-balanced dataset, achieving 208 correct predictions and an accuracy of 52%. In the classification of the 777 epochs belonging to the NREM2 stage, the raw dataset demonstrated the highest performance with 622 correct predictions and an accuracy of 80%, while the GAN-based balancing method closely followed with 617 correct predictions and an accuracy of 79%. For the NREM3 stage, which consists of 133 epochs, the highest classification performance was achieved using the GAN-based dataset, with 104 correct predictions and an accuracy of 78%. For the 140 epochs in the REM stage, the raw dataset achieved the highest performance with 106 correct predictions and an accuracy of 75%, while the SMOTE and GAN methods exhibited similar performance levels for this stage. When examining the general error trends, it was observed that the model frequently confused the NREM2 stage with NREM1 and the REM stage with NREM2. In contrast, the model was able to clearly distinguish between the REM and NREM3 stages, and no significant confusion between these two classes was observed.

Conclusion

Sleep stage classification was carried out in this study using the MIT-BIH Polysomnographic dataset, which exhibits a high degree of class imbalance. In this context, the class imbalance problem in the dataset was addressed, and the data augmentation potentials of the traditional SMOTE method and the more innovative GAN approach based on deep neural networks were investigated for complex and non-stationary EEG signals such as sleep data. In addition, a novel deep learning model, not previously used in the literature and expected to yield competitive performance compared to existing studies, was proposed. When the obtained results are examined, the accuracy values of 87.58% for the 3-class classification and 75.85% for the 5-class classification, achieved without applying any data balancing method, present promising outcomes. When compared with studies in the literature conducted using the MIT-BIH Polysomnographic dataset, Zhao et al. (2021) achieved an accuracy of 80.40% for 4-class sleep staging using the same dataset. Similarly, Bhusal et al. (2022) and Zhang et al. (2020) reported accuracy rates of 88.1% and 87.6%, respectively, for 5-class sleep staging. However, unlike the proposed study, these works applied extensive preprocessing and feature extraction steps to the dataset, which contributed to achieving higher performance levels (Bhusal et al., 2022; Zhang et al., 2020; Zhao et al., 2021). As is well known, the feature extraction process helps focus on the most informative parts of the signal, thereby improving classifier performance. However, the aim of this study was to achieve comparable performance using raw signals; therefore, an end-to-end approach was adopted, and raw time-series signals were used directly without applying preprocessing steps such as filtering or feature extraction methods.

When the balancing methods are compared, it is observed that the SMOTE method performs worse than the GAN-based approach. On the other hand, the GAN model adapted in this study achieves performance results close to those obtained with the raw dataset, demonstrating that GAN is a more effective approach than SMOTE for generating complex signals such as EEG. In addition, as reported by Fan et al., the findings indicate that data augmentation methods do not necessarily improve performance for EEG signals. Furthermore, since the performance obtained from the raw dataset and the proposed model is promising, the model's performance can be further improved through hyperparameter optimization.

Recommendations

Based on the findings and performance of the proposed end-to-end model, several directions for future research are recommended: First, since the model achieved promising results directly from raw signals, systematic hyperparameter tuning should be conducted to further improve classification performance. Second, although the deep learning-based data augmentation approach proved more effective than traditional statistical methods in addressing class imbalance in non-stationary signals, future studies should explore more advanced generative architectures (e.g., GANs, VAEs, or diffusion models) to produce more realistic synthetic data. Third, integrating automated denoising mechanisms within the neural network architecture is recommended. This could help preserve the end-to-end framework while reducing the negative impact of signal artifacts.

Scientific Ethics Declaration

* The authors declare that the scientific ethical and legal responsibility of this article published in EPHELS journal belongs to the authors.

Conflict of Interest

* The authors declare that they have no conflicts of interest

Funding

* The authors received no financial support for the research, authorship, and/or publication of this article.

Acknowledgements or Notes

* This article was presented as an oral presentation at International Conference on Veterinary, Agriculture and Life Sciences (www.icvals.net) held in Konya/Türkiye on April 02-05, 2026.

References

An, P., Yuan, Z., Zhao, J., Jiang, X., & Du, B. (2021). An effective multi-model fusion method for EEG-based sleep stage classification. *Knowledge-Based Systems, 219*, 106890.

Ben Hamouda, G., Rejeb, L., & Ben Said, L. (2024). Ensemble learning for multi-channel sleep stage classification. *Biomedical Signal Processing and Control, 93*, 106184.

Berry, R. B., Brooks, R., Gamaldo, C., Harding, S. M., Lloyd, R. M., Quan, S. F., ... Vaughn, B. V. (2017). AASM scoring manual updates for 2017 (Version 2.4). *Journal of Clinical Sleep Medicine, 13*(5), 665–666.

Bhusal, A., Alsadon, A., Prasad, P. W. C., Alsalami, N., & Rashid, T. A. (2022). Deep learning for sleep stages classification: Modified rectified linear unit activation function and modified orthogonal weight initialisation. *Multimedia Tools and Applications, 81*, 21105–21127.

Chen, J., Han, Z., Qiao, H., Li, C., & Peng, H. (2023). EEG-based sleep staging via self-attention based capsule network with Bi-LSTM model. *Biomedical Signal Processing and Control, 86*, 105351.

Cheng, X., Huang, K., Zou, Y., & Ma, S. (2024). SleepEGAN: A GAN-enhanced ensemble deep learning model for imbalanced classification of sleep stages. *Biomedical Signal Processing and Control, 92*, 106020.

Dokare, I., & Gupta, S. (2025). Brain-region specific epileptic seizure detection through EEG dynamics: Integrating spectral features, SMOTE and long short-term memory networks. *Cognitive Neurodynamics, 19*(1), 67–85.

Efe, E. (2022). *Derin öğrenme ağ yapılarının uyku evreleme problemlerine uygulanması* (Tez No. 704299). Ulusal Tez Merkezi.

Efe, E., & Ozsen, S. (2023). CoSleepNet: Automated sleep staging using a hybrid CNN-LSTM network on imbalanced EEG-EOG datasets. *Biomedical Signal Processing and Control, 80*, 104299.

Fan, J., Sun, C., Chen, C., Jiang, X., Liu, X., Zhao, X., ... Chen, W. (2020). EEG data augmentation: Towards class imbalance problem in sleep staging tasks. *Journal of Neural Engineering, 17*(5), 056017.

Fan, Y., & Wang, B. (2023). A filter-based GAN to enhance learning frequency information of EEG for sleep staging. In *2023 8th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)* (pp. 236–241).

Fatimah, B., Singhal, A., & Singh, P. (2022). A multi-modal assessment of sleep stages using adaptive Fourier decomposition and machine learning. *Computers in Biology and Medicine, 148*, 105877.

Gashti, M. Z., & Farjamnia, G. (2025). EEG sleep stage classification with continuous wavelet transform and deep learning. *Mbeya University of Science and Technology Journal of Research and Development, 6*(3), 428–437.

Goshtasbi, N., Boostani, R., & Sanei, S. (2022). SleepFCN: A fully convolutional deep learning framework for sleep stage classification using single-channel electroencephalograms. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 30*, 2088–2096.

Hartmann, K. G., Schirrmeister, R. T., & Ball, T. (2018). EEG-GAN: Generative adversarial networks for electroencephalographic (EEG) brain signals. *arXiv preprint arXiv:1806.01875*.

Huang, W., Guo, B., Shen, Y., Tang, X., Zhang, T., Li, D., & Jiang, Z. (2020). Sleep staging algorithm based on multichannel data adding and multifeature screening. *Computer Methods and Programs in Biomedicine, 187*, 105253.

Huttunen, R., Leppänen, T., Duce, B., Arnardottir, E. S., Nikkonen, S., Myllymaa, S., ... Korkalainen, H. (2023). A comparison of signal combinations for deep learning-based simultaneous sleep staging and respiratory event detection. *IEEE Transactions on Biomedical Engineering, 70*(5), 1704–1714.

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 *2021 9th International Winter Conference on Brain-Computer Interface (BCI)* (pp. 1–5).

Li, Y., Xu, Z., Zhang, Y., Cao, Z., & Chen, H. (2022). Automatic sleep stage classification based on a two-channel electrooculogram and one-channel electromyogram. *Physiological Measurement, 43*(7), 07NT02.

Liu, C., Tan, B., Fu, M., Li, J., Wang, J., Hou, F., & Yang, A. (2021). Automatic sleep staging with a single-channel EEG based on ensemble empirical mode decomposition. *Physica A: Statistical Mechanics and its Applications, 567*, 125685.

Mishra, S., & Jaiswal, U. C. (2024). Ensemble methods with statistics and machine learning on the class imbalance problems of EEG data. *Journal of Electrical Systems, 20*(5s), 453–462.

Mostafaei, S. H., Tanha, J., & Sharafkhaneh, A. (2024). A novel deep learning model based on transformer and cross modality attention for classification of sleep stages. *Journal of Biomedical Informatics, 157*, 104689.

Permana, K. E., & Iramina, K. (2024). Single channel sleep stage classification using random forest and feature extraction with SMOTE balancing on Fpz-Cz EEG data. *Mathematical Modelling of Engineering Problems, 11*(12), 3243–3250.

Perslev, M., Darkner, S., Kempfner, L., Nikolic, M., Jennum, P. J., & Igel, C. (2021). U-Sleep: Resilient high-frequency sleep staging. *NPJ Digital Medicine, 4*(1), 72.

Phan, H., & Mikkelsen, K. (2022). Automatic sleep staging of EEG signals: Recent development, challenges, and future directions. *Physiological Measurement, 43*(4), 04TR01.

Ren, Z., Ma, J., & Ding, Y. (2025). FlexibleSleepNet: A model for automatic sleep stage classification based on multi-channel polysomnography. *IEEE Journal of Biomedical and Health Informatics, 29*(5), 3488–3501.

Ruehland, W. R., & Rochford, P. D. (2009). The new AASM criteria for scoring hypopneas: Impact on the apnea hypopnea index. *Journal of Clinical Sleep Medicine, 5*(2), 168–174.

Satapathy, S. K., Kondaveeti, H. K., Sreeja, S. R., Madhani, H., Rajput, N., & Swain, D. (2023). A deep learning approach to automated sleep stages classification using multi-modal signals. *Procedia Computer Science, 218*, 867–876.

Sun, C., Fan, J., Chen, C., Li, W., & Chen, W. (2019). A two-stage neural network for sleep stage classification based on feature learning, sequence learning, and data augmentation. *IEEE Access, 7*, 109386–109397.

Supratak, A., Dong, H., Wu, C., & Guo, Y. (2017). DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 25*(11), 1998–2008.

Taran, S., Sharma, P. C., & Bajaj, V. (2020). Automatic sleep stages classification using optimize flexible analytic wavelet transform. *Knowledge-Based Systems, 192*, 105367.

Wara, T. U., Fahad, A. H., Das, A. S., & Shawon, M. M. H. (2025). A systematic review on sleep stage classification and sleep disorder detection using artificial intelligence. *Heliyon, 11*(12), e43576.

Xu, Q., Zhou, D., Wang, J., Shen, J., Kettunen, L., & Cong, F. (2022). Convolutional neural network based sleep stage classification with class imbalance. In *2022 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–6).

Xu, X., Cong, F., Chen, Y., & Chen, J. (2025). Sleep stage classification with multi-modal fusion and denoising diffusion model. *IEEE Journal of Biomedical and Health Informatics, 29*(12), 9233–9244.

Yu, T., Hu, X., He, Y., Wu, W., Gu, Z., Yu, Z., ... Xiao, J. (2025). Multi-view self-supervised learning enhances automatic sleep staging from EEG signals. *IEEE Transactions on Biomedical Engineering, 72*(10), 3056–3070.

Zan, H., & Yildiz, A. (2023). Multi-task learning for arousal and sleep stage detection using fully convolutional networks. *Journal of Neural Engineering, 20*(5), 056034.

Zhang, J., Yao, R., Ge, W., & Gao, J. (2020). Orthogonal convolutional neural networks for automatic sleep stage classification based on single-channel EEG. *Computer Methods and Programs in Biomedicine, 183*, 105089.

Zhao, D., Jiang, R., Feng, M., Yang, J., Wang, Y., Hou, X., & Wang, X. (2022). A deep learning algorithm based on 1D CNN-LSTM for automatic sleep staging. *Technology and Health Care, 30*(2), 323–336.

Zhao, R., Xia, Y., & Wang, Q. (2021). Dual-modal and multi-scale deep neural networks for sleep staging using EEG and ECG signals. *Biomedical Signal Processing and Control, 66*, 102455.

Author(s) Information

Sena Ceper-Ozcelik
Konya Technical University
Konya, Türkiye
ORCID iD: https://orcid.org/0000-0001-9441-5962
*Contact e-mail: *sceper@ktun.edu.tr*

Gulay Tezel
Konya Technical University
Konya, Türkiye
ORCID iD: https://orcid.org/0000-0003-1698-0106

*Corresponding author(s) contact email

To cite this article:

Ceper-Ozcelik, S., & Tezel, G. (2026). Comparison of SMOTE and GAN data augmentation methods in sleep stage classification. *The Eurasia Proceedings of Health, Environment and Life Sciences (EPHELS), 21,* 1-10.