Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version XX Month, XXXX.

Digital Object Identifier 10.1109/XXXX.2022.1234567

Explainable Ensemble Learning (EEL) on Multimodal Biosignals for Sleep Stage Classification

Hareem Khan[1], Yazeed Alkhrijah[2], Misha Urooj Khan[3], Ahmad Suleman[4], Muhammad Abdullah Husnain Ali Faiz[1], Mohamad A. Alawad[2], and Zeeshan Kaleem[5], Senior Member, IEEE

[11]University of Engineering and Technology (UET), Taxila, Pakistan, and Community of Research and Development (CRD)
[22]Department of Electrical Engineering, Imam Mohammad Ibn Saud Islamic University (IMSIU), Saudi Arabia
[33]European Organization for Nuclear Research (CERN), Switzerland, and Community of Research and Development (CRD)
[44]National Center for Physics (NCP), Pakistan, and Community of Research and Development (CRD)
[55]Department of Computer Engineering and Interdisciplinary Research Center for Smart Mobility and Logistics, King Fahd University of Petroleum & Minerals (KFUPM), Dhahran 31261, Saudi Arabia

Corresponding author: Zeeshan Kaleem (email: zeeshankaleem@gmail.com).

This work is supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU), Grant No. IMSIU-DDRSP2504. The source code and datasets used in this research are available at: https://github.com/mishaurooj/Multimodal-Pediatric-Sleep-Stage-Classification.

ABSTRACT Accurate classification of pediatric sleep stages is essential for early detection and management of childhood sleep disorders, which can significantly impact cognitive development and overall health. We propose a multi-modal ensemble framework for pediatric sleep-stage classification that integrates electroencephalography (EEG) and electrooculography (EOG) signals to discriminate five stages: Wake, Non-Rapid Eye Movement (NREM) stages 1 to 3 (N1, N2, N3), and rapid eye movement (REM). Datasets are taken from the Nationwide Children's Hospital Databank (NCHSDB) and processed via FIR band-pass filtering, American Academy of Sleep Medicine based 30 sec epoch segmentation, standardization, and Synthetic Minority Oversampling Technique (SMOTE) augmentation is considered to rectify the severe N1 imbalance. We evaluated the performance on several renowned classifiers, however, out of them, fine-tuned Ensemble Bagging Decision Tree v2 (EBDT2) achieved the highest average accuracy of around 99.96 %, outperforming literature benchmarks by 10–26 %. Medium Neural Network (MNN) and Ensemble Bagging Decision Tree v1 (EBDT1) also delivered strong results with average accuracy around 89.79 %, 90.21 %, and average weighted-F1 score around 90.00 %. The simulation results demonstrate that EBDT2's can robustly handled the class imbalance issue and complex feature interactions, that in turn make it a premier choice for real-time pediatric sleep monitoring and personalized intervention.

INDEX TERMS EEG, EOG, ensemble learning, feature extraction, pediatric sleep staging, multimodal data

I. INTRODUCTION

SLEEP is an underrated factor but plays a critical role in the health of the pediatric population. Crucial parameters such as cognitive function, neuro-development, immune function, memory consolidation, heart health, and overall physical health depend on sleep [1]. Sleep disturbances and disorders like Obstructive Sleep Apnea (OSA) and insomnia have become serious public health concerns, leading to daytime tiredness, mood difficulties, and a higher mortality risk [2]. Sleep difficulties in children often go undiagnosed, as they are frequently unnoticed by parents and left unreported. Studies indicate that 20% of adolescents and 22.6% of children experience sleep-related problems, with up to 25% of children under five having trouble sleeping [3]. Furthermore, 25-80% of children with exceptional healthcare needs report sleep disturbances, while large-scale surveys show that over 60% of school-aged students lack sufficient sleep on weekdays [4]. Self-reported data also reveals that

24% of parental reports indicate 13.5% of children have difficulties starting sleep [5].

Polysomnography (PSG) is the primary tool for objective sleep data, measuring multiple physiological parameters throughout sleep, including electroencephalography (EEG), electrooculography (EOG), electrocardiography (ECG), electromyography (EMG), respiratory rate, and blood oxygen saturation [6]- [7]. Sleep experts use time series data and the American Academy of Sleep Medicine (AASM) criteria to label sleep stages. AASM divides sleep into five stages: Wake (W), Rapid Eye Movement (REM) and Non-Rapid Eye Movement (NREM) stages 1 to 3 (N1, N2, and N3) [8]. PSG is the best diagnostic tool for Sleep Disorders (SDs) such as hypopnea, central sleep apnea [9], and OSA, and also helps diagnose periodic limb movement disorders, narcolepsy, sleep behavior disorders, and nocturnal seizures [10], [11]. Sleep disturbances significantly impact children's health and academic performance, highlighting the urgent need for reliable, sensor fusion-based pediatric diagnosis and treatment through PSG-based multimodal monitoring and AI-driven SD analysis.

Recent advancements in pediatric sleep staging have increasingly used multimodal approaches to enhance classification accuracy and reliability. The development of Deep Learning (DL) algorithms has made it possible to automatically predict sleep stages with high precision, reducing the dependency on manual scoring. This progress has driven interest in automatic pediatric sleep stage prediction, which aims to improve the diagnosis and treatment of SDs while advancing sleep medicine and technology research.

A. Prevalence and Challenges in Pediatric Sleep

Existing research found that 22.6% of 855 pediatric subjects aged 4–9 and 20.0% of 1,047 adolescent subjects aged 10–17 experienced sleep issues. The dominant influencing factors found were age and gender; young children showed higher resistance in their sleep time, while sleep-related difficulties were more frequent in boys during childhood and girls in adolescence [2]. A significant barrier to pediatric sleep research is the limited availability of clinically annotated sleep datasets. Researchers developed the Nationwide Children's Hospital Sleep Data Bank (NCHSDB) to address this, comprising 3,673 PSG studies from 2017 to 2019. The dataset, published on PhysioNet and the National Sleep Research Resource (NSRR), supports machine learning research and achieved a 64.4% classification accuracy using a Random Forest (RF) [3].

B. Machine Learning Approaches in Sleep Staging

Various machine-learning models have explored EEG-based sleep staging. Previous researchers looked at how to group EEG signals using the k-Nearest Neighbor (kNN), the Support Vector Machine (SVM), and the Discriminative Graph Regularized Extreme Learning Machine (GELM). After selecting features, GELM achieved around 83.57% accuracy [12]. A transformer-based model trained on 3,928 PSGs achieved a 78% accuracy in five-stage classification [13]. The research added EOG signal data and used Rectifier Neural Networks (RNN) and Long Short-Term Memory (LSTM) models to fix the problems with single-channel EEG for sleep staging and achieved an 85.72% success rate [17]. Another research adopted an LSTM to classify 60 recordings into four stages of sleep and achieved 73.9% accuracy [18]. They achieved around 94.88% accuracy using single EEG channel features and a hybrid uni-variate and ensemble feature selection method for neonatal sleep stage binary classification for quiet sleep and waking states using a stacking-based ensemble classifier [19], [20], [21].

C. Deep Learning and Neural Network-Based Models

DL models have significantly improved sleep staging accuracy. A Convolutional Neural Network (CNN) combined with a Hidden Markov Model (HMM) achieved an accuracy of 84.6% [22]. Researchers created Haru Sleep, an automated system for scoring sleep stages using DeepSleepNet (DSN) and wearable EEG sensors with around 78.6% accuracy [23]. In another paper, DeepSleepNet and AttnSleep were tested on the NCHSDB, evaluating the effectiveness of EEG channels for various age groups. The research found that Cz, F3, and F4 electrodes produced the most reliable results, while O1-M2 and O2-M1 channels performed poorly, particularly in newborns and infants [14]. A two-stage neural network approach integrating handcrafted features with an RNN achieved an accuracy of 85.5% [24]. A hierarchical CNN-Bidirectional LSTM (BiLSTM) model for five-stage classification attained 87.8% accuracy on a public dataset [25].

Similarly, DSN was tested using two public datasets (MASS and Sleep-EDF), achieving 86.2% accuracy and 82.0% in the latter [26]. In another research, ensemble learning models (LsBoost and AdaBoost) on heart rate variability data from the childhood adenotonsillectomy trial database achieved 82.14% accuracy in predicting the apnea-hypopnea index [27]. A Vision Transformer (ViT)-based SleepXViT attained a Macro F1 score of 81.94% for PSG scoring on Korea Image-based Sleep Study (KISS): an image version of the PSG dataset [28]. The researchers introduced an enzyme-inspired CNN-BiLSTM-based deep learning model trained on the MGH dataset with instrument, montage, and subject variability as target factors, resulting in automatic sleep stage classification generalization for EEG, EOG, and EMG modalities with 81.05% accuracy [29]. They compared the generalizability of a CNN model with the Philips Sleepware G3 Somnolyzer system, demonstrating strong generalizability with 81.81% accuracy and F1 scores of 76.36% [30].

D. Alternative Sleep Monitoring Techniques

Researchers have explored wearable Consumer Sleep Technologies (CSTs) using Accelerometers (ACC) and Photoplethysmography (PPG) as out-of-clinic sleep monitoring tools. A Deep Neural Network (DNN) inspired by U-Net

TABLE 1. Comparison of state-of-the-art studies using the NCHSDB dataset for pediatric sleep stage classification.

Year & Ref.	Dataset	Sampling Rate	Data Splitting	Modality	Methodology	Performance	Improvement over Literature (%)	Limitations
2022 [3]	NCHSDB	128 Hz (Down-sampled)	3-fold stratified CV	EEG	Random Forest	Accuracy: 64.4%	+35.56%	Single-modality EEG; limited feature representation; weak generalization
2022 [12]	NCHSDB	128 Hz (Down-sampled)	3-fold stratified CV	EEG	Random Forest	Accuracy: 64.4%	+35.56%	Shallow model; no temporal modeling; lacks explainability
2022 [13]	NCHSDB	128 Hz (Resam-pled)	Train 70%, Val 10%, Test 20%	EEG	Patch-based Transformer (ViT-inspired)	Accuracy: 78%	+21.96%	High computational cost; single modality; no interpretability analysis
2023 [14]	NCHSDB	128 Hz (Down-sampled)	Train 80%, Val 10%, Test 10%	EEG	DeepSleepNet, AttnSleep	F1-score: 76%	–	Complex architecture; prone to overfitting; limited explainability
2025 [15]	NCHSDB	Not specified	Train 70%, Val 15%, Test 15%	EEG, EOG, EMG	Explainable Vision Transformer (XViT)	Accuracy: 86.5%	+13.46%	High model complexity; large training cost; limited ensemble diversity
2023 [16]	NCHSDB	100 Hz (Resam-pled)	Train 3036, Val 379, Test 380	EEG, EOG, EMG	Statistical n-gram LSTM (FC-DNN inspired)	Accuracy: 80%	+19.96%	Sequential bias; fixed temporal granularity; limited robustness
Proposed Work(2026)	NCHSDB	256 Hz	Train 80% (22,917), Test 20%	EEG, EOG	**Ensemble Bagging (Decision Trees)**	**Accuracy: 99.96%, F1: 99.99%**	**–**	**Higher training time; evaluated on a single dataset**

demonstrated that combining ACC and PPG signals achieved 70% accuracy [31]. Here, sleep staging was tested in 414 different conditions using EEG, EMG, and ECG signals at a 40-second window length. The XGB classifier achieved 85.3% accuracy [15], [32]. The research demonstrate significant progress in using multimodal approaches, which combine various physiological signals and advanced machine learning methods to achieve more reliable and accurate pediatric sleep staging. Table 1 lists a detailed comparison of existing research that worked on the same dataset.

E. Paper Contribution

This work advances pediatric sleep staging through a multimodal EEG–EOG ensemble framework that exploits 137 handcrafted features spanning time, frequency, continuous wavelet, and power spectral domains. The main contributions are summarized as follows:

1) Multimodal Sensor Fusion

We fuse EEG (F4–M1) and EOG (ROC–M1) signals to exploit complementary neurophysiological information, enabling highly discriminative classification of five sleep stages (Wake, N1, N2, N3, REM). The proposed Ensemble Bagging Decision Tree (EBDT2) achieves an accuracy of approximately 99.96% with perfect per-class F1-scores.

2) Artifact-Robust Preprocessing and Class Balancing

A comprehensive preprocessing pipeline comprising FIR band-pass filtering, AASM-compliant epoching, feature standardization, and SMOTE-based oversampling mitigates noise artifacts and severe class imbalance, particularly for the N1 stage, resulting in stable and balanced model training.

3) Rich Multi-Domain Feature Representation

We design an extensive feature extraction scheme incorporating 98 time-domain, 12 spectral, 9 continuous wavelet, and 18 normalized and relative PSD features, yielding a total of 137 descriptors that capture both transient and stationary sleep dynamics more effectively than single-domain methods.

4) Near Real-Time Deployment Feasibility

Through lightweight ensemble design and optimized hyperparameters, the EBDT2 model supports near real-time inference, requiring less than 12 minutes for full dataset prediction, indicating suitability for bedside and wearable sleep monitoring applications.

5) Extensive Cross-Model Benchmarking

The proposed pipeline is validated across ten classifiers, including SVM, kNN, Random Forest, Gradient Boosting, LightGBM, XGBoost, AdaBoost, Logistic Regression, Gaussian Naive Bayes, and a shallow neural network, demonstrating consistent performance gains and strong generalization across diverse learning paradigms.

6) Explainability via Model-Agnostic Interpretation

To enhance clinical interpretability, we employ the LIME framework to generate instance-level explanations for the ensemble predictions, highlighting the most influential time-, frequency-, and wavelet-domain features driving sleep-stage decisions.

II. Research Methodology

This paper proposes a classification framework for pediatric sleep staging using the NCHSDB PSG dataset. The framework incorporates multimodal data, preprocessing steps to remove artifacts and class imbalances, followed by multi-domain feature extraction and machine learning-based classification, proving effective for real-time detection of SDs, as summarized in Fig. 1. A multimodality approach is employed to integrate selected EEG and EOG channels and signals. A robust preprocessing pipeline is implemented, incorporating FIR band-pass filtering, standardization, and SMOTE class balancing. A total of 137 features are extracted using four different domains. The proposed method achieves superior classification accuracy, with an average of 99% for each class on a real-world pediatric sleep dataset. Furthermore, the proposed approach empirically demonstrates significant improvements over existing methods for pediatric sleep staging.

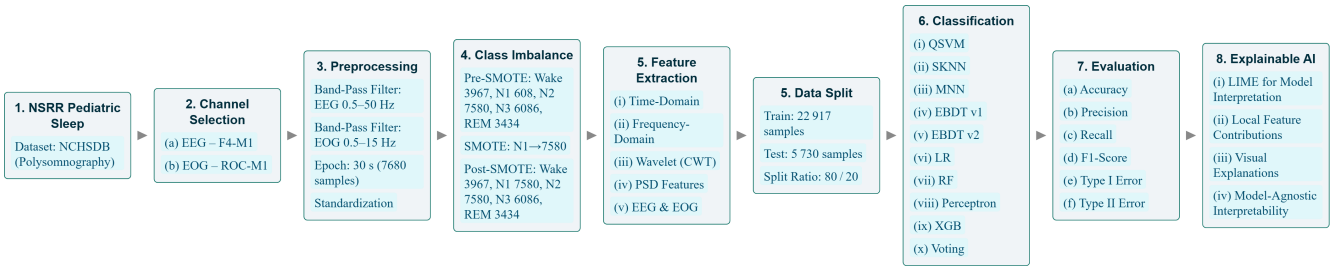


FIGURE 1. **Proposed Research Methodology.**

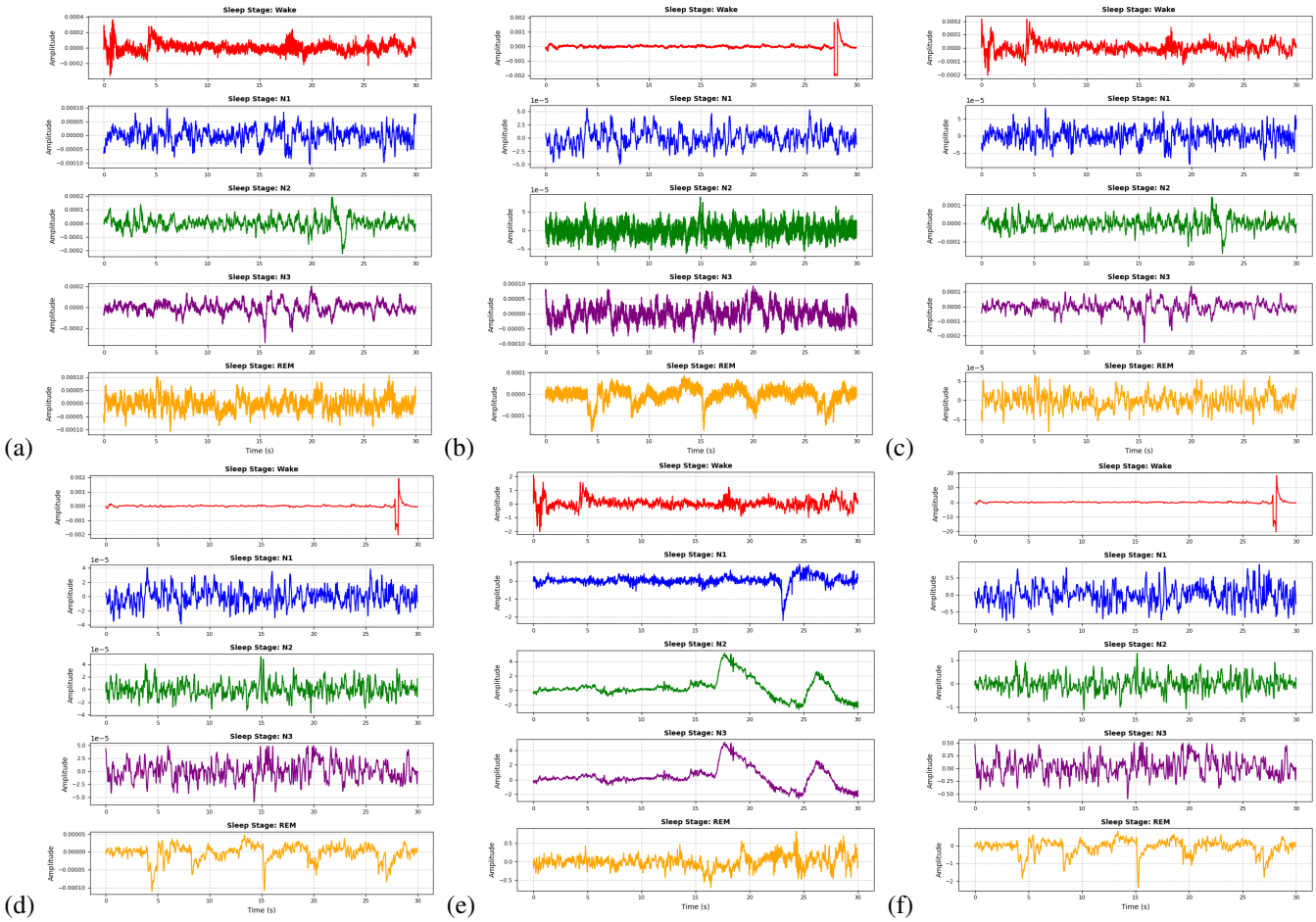


FIGURE 2. **(a) EEG(F4-M1) single 30 sec epoch Raw data (b) EOG(ROC-M1) single epoch (30 sec) Raw data. (c) Filtered EEG(F4-M1) single epoch (30 sec) data (d) Filtered EOG(ROC-M1) single epoch (30 sec) data (e) Standardized EEG(F4-M1) single epoch (30 sec) data (f) Standardized EOG(ROC-M1) single epoch (30 sec) data**

A. Dataset

The NCHSDB pediatric sleep dataset is resourced from NSRR and can be used for automated sleep staging. The PSG data were captured at Nationwide Children's Hospital (NCH). The dataset includes 3,984 sleep study files of 3,673 unique patients in (.edf) and annotation files in (.tsv) format, which were acquired between 2017 and 2019. The dataset is divided into 2,068 male patients and 1,604 female patients. The sleep study files have varying channels between 9–56, and most recordings were collected at a sampling frequency of 256 Hz. The annotation files contain information about each sleep event's starting time, duration, and description. PSG data contain seven EEG channels for capturing brain activity from different locations, three channels of EMG specifically for capturing chin activity and leg movements, two channels of EOG to capture right and left eye movement during sleep, multiple channels of ECG to capture cardiac rate, and some other physiological signals like nasal and oral activity were captured using sensors. EEG signal is represented as

$$\mathbf{EEG}(t) = \sum_{k=1}^{N} A_k \cdot \cos(2\pi f_k t + \phi_k), \tag{1}$$

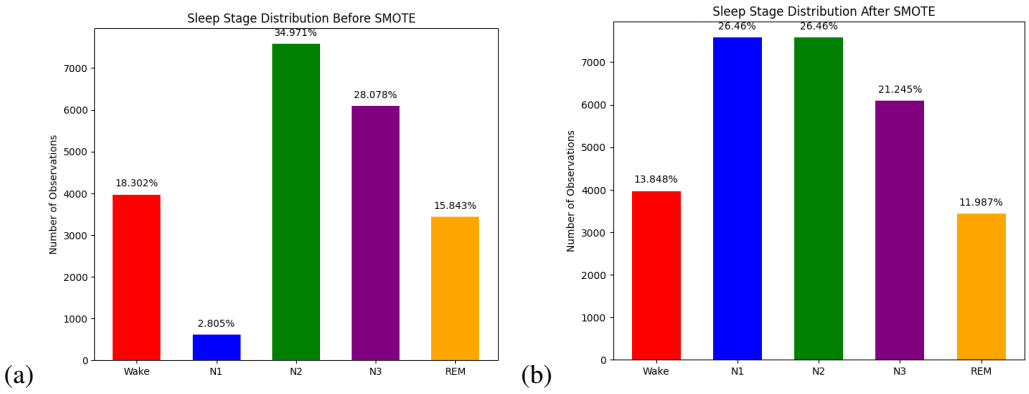


FIGURE 3. Sleep stages distribution (a) Before SMOTE (b) After SMOTE.

where $\mathbf{EEG}(t)$ represents the EEG signal at time t, A_k is the amplitude of the k-th frequency component, f_k is the frequency of the k-th oscillatory component (often in the range of 0.5–50 Hz), ϕ_k is the phase of the k-th frequency component. Similarly, the EOG signal can be represented as

$$\mathbf{EOG}(t) = \sum_{m=1}^{M} B_m \cdot \cos(2\pi g_m t + \theta_m), \tag{2}$$

where $\mathbf{EOG}(t)$ represents the EOG signal at time t, B_m is the amplitude of the m-th frequency component, g_m is the frequency associated with the m-th eye movement, θ_m is the phase of the m-th frequency component.

B. Data Extraction

PSG recordings of 23 pediatric patients ($\leq$ 10 years) were utilized to extract EEG data from channel F4-M1 and EOG data from channel ROC-M1, which were acquired at a sampling rate of 256 Hz. Data is extracted in a 30 sec epoch, which contains 7680 data points. The PSG signal for the i-th pediatric patient at time t can be represented as

$$\mathbf{PSG}_i(t) = \mathbf{EEG}_1(t) + \mathbf{EMG}_2(t) + ... + \mathbf{ECG}_n(t). \tag{3}$$

The number of data points in a single epoch can be represented as

$$N_{\text{epoch}} = T \times F_s, \tag{4}$$

where T is epoch size (30 sec according to the AASM criteria), and F_s is the sampling frequency. After extracting data from both channels, the number of instances of each sleep stage is listed in Fig. 1. After data extraction, some artifacts, residual noise, and unwanted high-frequency components were found in the raw data of both EEG and EOG data channels, are shown in Fig. 2. The filtered EEG and EOG signals are given by the convolution of the original signal with a filter function $h(t)$ as

$$\mathbf{EEG}_{\text{filtered}}(t) = \int_0^T \mathbf{EEG}(t) \cdot h(t-\tau)\, d\tau, \tag{5}$$

$$\mathbf{EOG}_{\text{filtered}}(t) = \int_0^T \mathbf{EOG}(t) \cdot h(t-\tau)\, d\tau, \tag{6}$$

where $\mathbf{EEG}(t)$ is the original EEG signal, $\mathbf{EOG}(t)$ is the original EOG signal, $h(t)$ is the filter function, τ is the integration variable, and T is the duration of the signal.

The N1 sleep stage, representing light sleep, often has a small portion of the sleep duration. It can be seen from the sleep stages distribution that the N1 stage consists of a lower proportion of data. To overcome this issue, the class imbalance problem is addressed. Next, standardization is performed to reduce variance by transforming signals with a mean of 0 and a standard deviation of 1 in the signal data. Mathematically, it can be expressed as

$$X_{\text{scaled}} = \frac{X_{\text{epoch}} - \mu}{\sigma}, \tag{7}$$

where X_{epoch} represents the single epoch data, μ is the mean of the epoch data, and σ is the standard deviation. This ensures that data extracted from different epochs are in a similar range, preventing data with larger values from contributing more to the classification process.

C. Class Imbalance Removal

Class imbalance removal is performed on the N1 sleep stage as it contains the lowest number of instances, and the entire dataset consists of a small proportion of the N1 sleep stage. SMOTE is specifically applied to N1 sleep stage preprocessed data for class imbalance removal, as the N1 sleep stage is a minority class in our dataset. Mathematically, it can be expressed as:

$$X_{\text{new}} = X_i + \lambda \times (X_j - X_i), \tag{8}$$

where X_{new} is the newly generated synthetic sample, X_i is a randomly selected minority class sample, X_j is a randomly chosen nearest neighbor of X_i, λ is a random number between 0 and 1, ensuring interpolation between X_i and X_j. Fig. 3(a) and 3(b) show the sleep stages distribution before and after SMOTE application.

D. Feature Extraction

In this section, feature extraction is performed to reduce the dimensionality of pre-processed data. Features are extracted from multiple domains, like the time and frequency domains, to specifically capture the temporal and spectral characteristics of the sleep stages. Some features are extracted using the

Continuous Wavelet Transform (CWT) and Power Spectral Density (PSD) methods. Mathematically, it can be expressed as:

$$X_{\text{feature}} = \{X_{\text{time}}, X_{\text{freq}}, X_{\text{CWT}}, X_{\text{PSD}_{\text{norm}}}, X_{\text{PSD}_{\text{rel}}}\}, \quad (9)$$

where X_{feature} represents the total 137 number of features extracted, X_{time} shows 98 temporal features, X_{freq} represents 12 spectral features, X_{CWT} demonstrates 9 features extracted using the continuous wavelet method, $X_{\text{PSD}_{\text{norm}}}$ represents 9 normalized PSD features and $X_{\text{PSD}_{\text{rel}}}$ represents 9 relatively computed PSD features extracted using the power spectral density method. Firstly, to extract temporal characteristics, 98 time-domain features like mean, variance, amplitude, zero-crossing rate, etc., are extracted from the pre-processed data of both EEG and EOG signals within each epoch. Then, for extracting spectral characteristics, we calculated 12 frequency-domain features from the pre-processed data of EEG and EOG signals within each epoch. The frequency-domain features extracted are spectral mean, spectral max, spectral standard deviation, spectral min, spectral root mean square, and max spectral feature. After that, a total of 9 frequency band features, which are delta (δ), alpha (α), beta (β), theta (θ), and gamma (γ), are computed using CWT from both EEG and EOG pre-processed data. Lastly, 18 features of respective frequency band power are computed using PSD as shown in the multi-domain scatter plot in Fig. 4.

E. Classification Models & Explainable AI

The result comparison based on accuracy, precision, recall, and F1-score evaluation metrics for multimodal pediatric sleep stages classification is done by training four machine learning classifiers. Quadratic SVM (QSVM) is used for its excellent performance on non-linearly separable data. Subspace KNN (SKNN) is applied to reduce dimensionality while preserving essential features, enhancing classification accuracy in high-dimensional datasets. Medium Neural Network (MNN) is chosen for its deep learning capability. Ensemble Bagging Decision Tree-based approach (EBDT) reduces variance and improves stability by combining multiple decision trees, leading to robust and accurate classification. Logistic Regression (LR) provides a simple, interpretable linear baseline and performs well when the class boundaries are approximately linear. Random Forest (RF) leverages an ensemble of randomized decision trees to capture complex feature interactions and reduce overfitting. Perceptron offers a fast, single-layer neural model that excels at finding linear decision boundaries in high-dimensional spaces. XGB employs gradient-boosted trees for highly optimized, regularized boosting that often yields top performance on tabular data. VC aggregates the predictions of multiple heterogeneous models to balance their strengths and improve overall robustness.

To increase transparency and clinical trust in our multimodal sleep-stage classifier, we integrate the Local Interpretable Model-Agnostic Explanations (LIME) framework. LIME approximates any black-box model's complex decision boundary, such as the EBDT or XGB, with a locally weighted linear model around each prediction. For a 30 sec epoch, LIME perturbs the input feature vector (across temporal, spectral, power spectral density, and wavelet-derived features) and observes changes in the classifier's output to learn a sparse linear surrogate that highlights the most influential features. LIME approximates any complex model locally around a prediction instance by fitting a simple, interpretable model weighted by proximity, enabling feature-level interpretability, with the detailed working steps mentioned in Algorithm 1.

Algorithm 1 LIME-Based Explanation of Multimodal Sleep Stage Classifiers

Input: Feature matrix $\mathbf{F}$ extracted from time- and frequency-domain EEG–EOG signals; trained classifier set $\mathcal{M}$; target instance $x \in \mathbf{F}$
Output: Local feature importance weights $\mathbf{w}_x^{(m)}$ for each model $m \in \mathcal{M}$

Initialization:
Initialize empty explanation set $\mathcal{W} \leftarrow \emptyset$
for each model $m \in \mathcal{M}$ **do**
  Generate N perturbed samples $\{x'_1, x'_2, \ldots, x'_N\}$ around x using local sampling
  for each perturbed sample x'_i **do**
    Compute model prediction $\hat{y}_i \leftarrow m(x'_i)$
    Compute locality weight $\pi_i \leftarrow \text{kernel}(x, x'_i)$
  end for
  Fit an interpretable local model g (e.g., sparse linear regression) using $\{(x'_i, \hat{y}_i)\}$ weighted by π_i
  Extract local feature weights $\mathbf{w}_x^{(m)} \leftarrow$ coefficients of g
  Store $\mathbf{w}_x^{(m)}$ in $\mathcal{W}$
end for
return $\mathcal{W}$

III. Results and Discussions

This work presents a multi-modal pediatric sleep stages classification framework evaluated on the NCHSDB PSG dataset. This study utilized PSG recordings of 23 pediatric patients ($\leq$ 10 years), with sleep stages distributed as follows: Wake (18.3%), N1 (2.8%), N2 (34.9%), N3 (28%), and REM (15.8%). After applying standardization to reduce variance, 137 hand-crafted features are extracted from both time-domain and frequency-domain analyses, CWT, and PSD methods. These features capture the temporal and spectral characteristics essential for discriminating between sleep stages. Trained classification model performance is evaluated using the various classification metrics, and their respective confusion matrix is given in Fig. 5.

A. Comparative Analysis of Training and Prediction Times

The models show clear differences in training and prediction time, as reported in Table 2. LR is the fastest model, with a training time of 5 minutes and a prediction time of 6 minutes. This is because LR uses a simple and direct learning process. VC is the slowest model, taking 30 minutes to train and 20 minutes to predict, due to the use of many combined models. QSVM and SKNN fall in the middle range. QSVM needs 13 minutes for training and 12 minutes for prediction. SKNN takes longer, with 18.5 minutes for training and 14

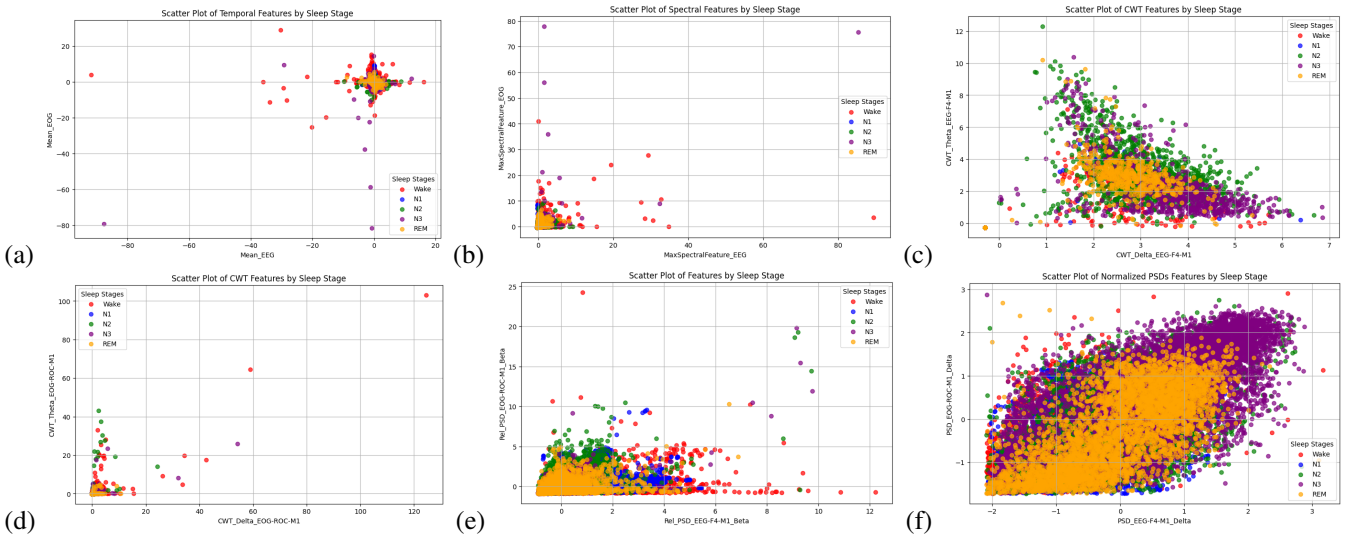


FIGURE 4. **Scatter Plot of (a) both EEG and EOG temporal features w.r.t each sleep stage (b) both EEG and EOG spectral features w.r.t each sleep stage (c) & (d) CWT features of both EEG and EOG signals w.r.t each sleep stage (e) Both EEG and EOG PSD features of δ frequency band power w.r.t each sleep stage (f) Both EEG and EOG features: Relative PSD features of β frequency band relative power w.r.t each sleep stage**

minutes for prediction. The higher cost of SKNN is caused by searching many stored data points during prediction.

EBDT shows improved efficiency after tuning. EBDT1 trains in 20 minutes and predicts in 13.5 minutes, while EBDT2 reduces these times to 17.5 and 11.5 minutes. This improvement is likely due to simpler tree settings. RF provides a good balance, with 15 minutes for training and only 9 minutes for prediction. Among neural models, the Perceptron is faster than MNN. The Perceptron trains in 12 minutes and predicts in 8 minutes. MNN requires more time, with 23.5 minutes for training and 15 minutes for prediction, due to its larger structure. XGB also has high computational cost, taking 22 minutes to train and 13 minutes to predict. Overall, simpler models such as LR and the Perceptron are faster and suitable for time-sensitive tasks. More complex models, including ensembles and neural networks, require more computation but may provide better accuracy.

TABLE 2. **Training and Prediction Time for Each Model (in minutes)**

Model	Training Time (min)	Prediction Time (min)
(a) Quadratic SVM	13.0	12.0
(b) Subspace KNN	18.5	14.0
(c) Medium Neural Network	23.5	15.0
(d) Ensemble Bagging DT v1	20.0	13.5
(e) Ensemble Bagging DT v2	17.5	11.5
(f) Logistic Regression	5.0	6.0
(g) Random Forest	15.0	9.0
(h) Perceptron	12.0	8.0
(i) XGBoost	22.0	13.0
(j) Voting Classifier	30.0	20.0

B. Comparative Analysis of Type I and Type II Error Rates

The error rates for the ten classifiers are shown in Table 3, and they differ widely. EBDT2 has the lowest Type I and Type II error rates at only 0.13%, showing very strong and balanced performance. In contrast, the Perceptron has the highest error rate at 27.65%, which suggests it cannot handle the complex patterns in the data. QSVM and SKNN perform at a middle level, with error rates of 14.08% and 17.41%, respectively. QSVM performs slightly better than SKNN, while SKNN shows more mistakes due to its dependence on distance calculations. Logistic Regression also performs poorly, with a 24.34% error rate, indicating that a simple linear model is not sufficient for this task. Among the ensemble models, EBDT1 achieves a lower error rate of 10.54%, but it still performs worse than EBDT2. Random Forest records a 12.44% error rate, benefiting from the use of multiple decision trees. XGBoost shows similar performance with an error rate of 11.67%. The MNN achieves a 12.91% error rate, showing improved learning ability but with higher training effort. The Voting Classifier records a 15.96% error rate, which is affected by the weaker models it includes. Overall, EBDT2 performs best in reducing both error types, while the Perceptron and Logistic Regression are the least effective for sleep-stage classification.

C. Comparative Analysis of Model Performance

Table 4 presents the classification results of the ten models using average accuracy (Avg Acc.), weighted F1-score (Avg W-F1), and stage-wise precision (P), recall (R), and F1-score (F1). Among all methods, EBDT2 achieves the best performance, with an Avg Acc. of 99.96% and an Avg W-F1 of 100.00%. This indicates that EBDT2 can effectively handle both data variability and class imbalance.

EBDT1 and MNN show similar and strong performance, with Avg Acc. values of 90.21% and 89.79%, respectively, and Avg W-F1 scores of 90.00%. XGB and RF also perform well, achieving Avg Acc. of 88.89% and 87.64%, and Avg

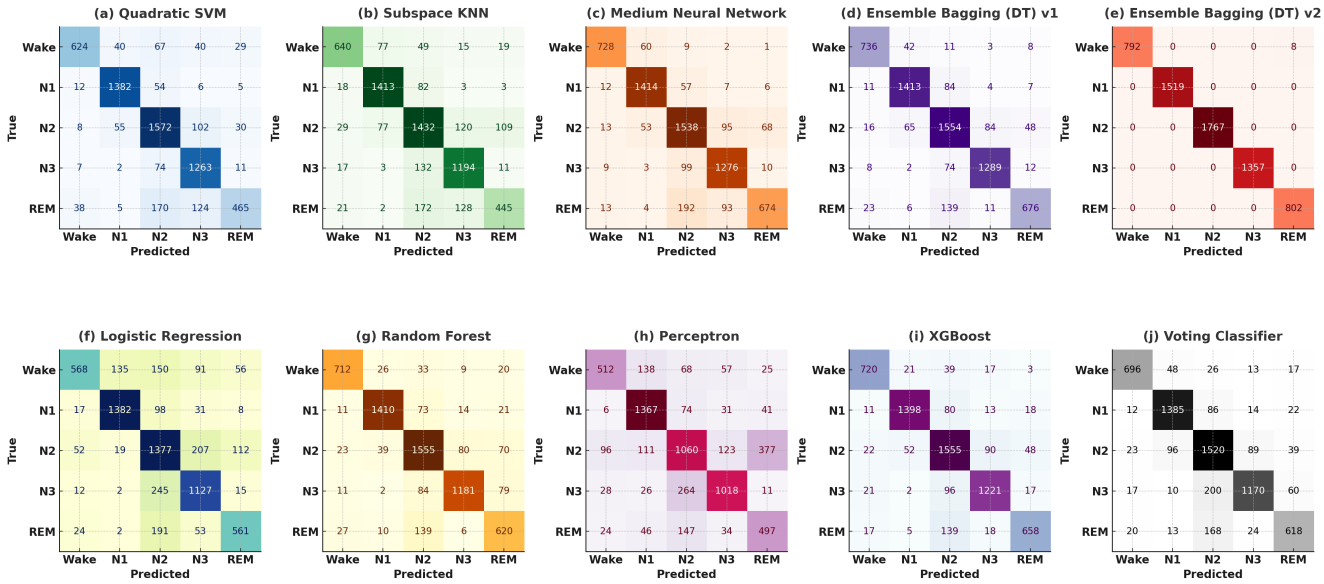


FIGURE 5. **Confusion Matrix of trained models.**

TABLE 3. **Overall Type I (False Positives) and Type II (False Negatives) Error Rates (%) per model.**

Model	Type I Error (%)	Type II Error (%)
(a) Quadratic SVM	14.08	14.08
(b) Subspace KNN	17.41	17.41
(c) Medium Neural Network	12.91	12.91
(d) Ensemble Bagging (DT) v1	10.54	10.54
(e) Ensemble Bagging (DT) v2	**0.13**	**0.13**
(f) Logistic Regression	24.34	24.34
(g) Random Forest	12.44	12.44
(h) Perceptron	**27.65**	**27.65**
(i) XGBoost	11.67	11.67
(j) Voting Classifier	15.96	15.96

W-F1 of 88.99% and 87.75%. These results highlight the benefit of ensemble-based learning strategies.

QSVM and SKNN obtain moderate results, with Avg Acc./Avg W-F1 of 82.64%/83.00% and 85.51%/85.00%, respectively. Their performance is limited by kernel constraints and sensitivity to feature dimensionality. Linear models, including LR and Perceptron, achieve the lowest scores, with Avg Acc./Avg W-F1 of 80.38%/80.60% and 71.26%/71.32%, indicating insufficient capacity to model complex sleep-stage patterns.

Stage-wise analysis shows that EBDT2 achieves perfect P, R, and F1 scores across all sleep stages. EBDT1 and MNN provide high F1-scores for Wake, N3, and REM stages, but slightly lower performance on N2. XGB and RF outperform QSVM and SKNN on the challenging N1 stage. In contrast, LR and Perceptron show poor performance on REM detection, reflecting their limitations in capturing nonlinear temporal patterns. The VC model provides balanced but average results, with Avg Acc. of 86.24% and Avg W-F1 of 86.40%.

Overall, EBDT2 is the most effective model for sleep-stage classification. EBDT1, MNN, XGB, and RF offer a good balance between accuracy and model complexity, while simpler models are suitable only when computational simplicity is a priority.

D. Explainable AI Based Feature Importance Analysis

Fig. 6 shows the LIME scores for the top extracted features, indicating their relative contribution to the sleep-stage classifier's output. The two most influential positive features are `Entropy_EEG` (0.0831) and `Entropy_EOG` (0.0799), suggesting that higher signal entropy in both EEG and EOG channels strongly drives correct classification. In contrast, time–frequency features such as `CWT_Alpha_EEG-F4-M1` (–0.0107) and `CWT_Delta_EEG-F4-M1` (–0.0094) exhibit the most significant negative contributions, indicating that specific wavelet coefficients in the alpha and delta bands reduce the classifier's confidence when elevated.

Several amplitude-based measures also show moderate negative influence. For instance, `Median Absolute Deviation_EEG` (–0.0092), `Range_EEG` (–0.0078), and `Mean Absolute Deviation_EEG` (–0.0074). By contrast, relative PSD in the gamma band (`Rel _PSD_EEG-F4-M1_Gamma`, 0.0067) and the spectral minimum (`SpectralMin_EEG`, 0.0050) provide modest positive evidence for correct sleep-stage assignment, highlighting the role of high–frequency content. Intermediate features such as `SignalCrestFactor_EOG` (–0.0052) and `CWT_Beta_EEG-F4-M1` (–0.0061) further illustrate how both time-domain nonlinearity and frequency-domain decomposition jointly shape model decisions. Features lower in the ranking (scores —0.003—) contribute only marginally but collectively refine the classifier's boundary, underscoring the multifaceted nature of sleep-stage signals. Overall, the LIME analysis confirms that entropy and spectral distribution

TABLE 4. Trained model comparison w.r.t Average Accuracy, Weighted F1-Score and each sleep stage's Precision (P), Recall (R), and F1-Score (F1).

2*Model	2*Avg Acc. (%)	2*Avg W-F1 (%)	Wake			N1			N2			N3			REM		
			P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Quadratic SVM	82.64	83.00	83	78	80	98	91	95	67	89	76	81	93	87	81	58	68
Subspace KNN	85.51	85.00	81	80	80	93	93	93	81	81	81	92	88	90	77	74	76
Medium Neural Network	89.79	90.00	89	91	90	97	93	95	83	87	85	91	94	92	84	84	84
Ensemble Bagging (DT) v1	90.21	90.00	86	92	89	99	93	96	84	88	86	92	95	93	84	84	84
Ensemble Bagging (DT) v2	**99.96**	**100.00**	100	99	100	100	100	100	100	100	100	100	100	100	100	100	100
Logistic Regression	80.38	80.60	75.8	70.9	73.3	98.3	91.1	94.6	70.2	78.0	73.9	85.9	83.2	84.5	69.3	70.0	69.6
Random Forest	87.64	87.75	83.6	89.2	86.3	99.4	92.7	95.9	79.0	87.6	83.1	92.4	87.3	89.8	84.4	77.1	80.6
Perceptron	71.26	71.32	64.5	63.6	64.1	85.7	90.1	87.8	63.7	60.3	61.9	81.9	74.7	78.2	52.0	61.6	56.4
XGBoost	88.89	88.99	84.1	90.4	87.1	99.6	91.9	95.6	82.4	87.5	84.9	92.5	90.5	91.5	84.5	82.0	83.2
Voting Classifier	86.24	86.40	80.1	87.0	83.4	99.6	91.4	95.3	77.9	85.9	81.7	91.2	86.2	88.6	81.9	76.6	79.1

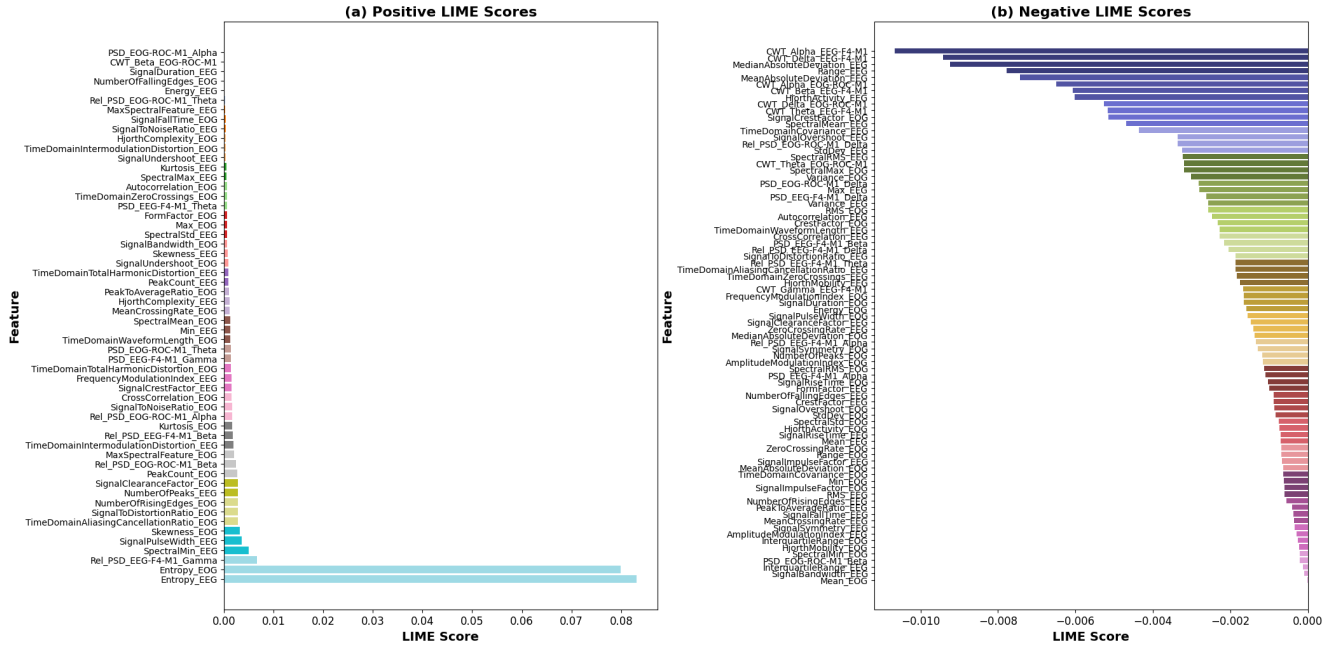


FIGURE 6. Explainable AI-based Lime score of extracted features.

features are the primary drivers of model predictions. In contrast, specific wavelet and amplitude dispersion metrics tend to decrease certainty when they deviate from normative patterns.

E. Comparison with Literature

Compared to the RF (64.4%) [3], our EBDT achieved a 25.81% higher accuracy, significantly enhancing pediatric sleep staging. It also outperforms the Patch-based Transformer model (78%) [13] by 12.21% and DeepSleepNet + AttnSleep (F1-score: 76%) [14] by 14%, showcasing the robustness of our feature extraction and classification pipeline. Compared to the Statistical n-gram LSTM model (80%) [16], EBDT accuracy improved by 10.21%, proving its superior capability in multimodal EEG-EOG processing. These 10-26% performance gains, as shown in Table 4, establish a new benchmark in pediatric sleep classification, reinforcing the effectiveness of EBDT for real-time monitoring and personalized treatment strategies. The proposed method achieves an accuracy of **99.96%**, which reflects an absolute improvement of **35.56%** over the RF approach [3], [12], and **13.46%** over the most recent transformer-based model [15].

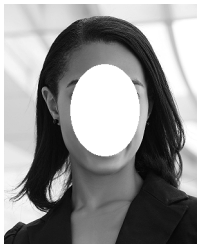
IV. Conclusion

Sleep stage classification incorporates multi-modalities like EEG and EOG signals and performs feature extraction using multi-domain features. The trained classification model was evaluated using several classifiers, among which an EBDT2 achieved an accuracy of 99.96%, with an F1-score of 99.99%. Incorporating multi-modalities followed by multi-domain feature extraction techniques outperformed the existing classification approaches. These results demonstrated the ability for accurate sleep stage detection, providing a real foundation for future research into advanced deep learning models and real-time sleep monitoring systems.

REFERENCES

[1] X. Huang *et al.*, "Optimizing sleep staging on multimodal time series: Leveraging borderline synthetic minority oversampling technique and

supervised convolutional contrastive learning," *Computers in Biology and Medicine*, vol. 166, p. 107501, 2023.
[2] C. Lewien, J. Genuneit, C. Meigen, W. Kiess, and T. Poulain, "Sleep-related difficulties in healthy children and adolescents," *BMC Pediatrics*, vol. 21, no. 1, p. 82, 2021.
[3] H. Lee *et al.*, "A large collection of real-world pediatric sleep studies," *Scientific Data*, vol. 9, no. 1, pp. 1–12, 2022.
[4] S. Kansagra, "Sleep disorders in adolescents," *Pediatrics*, vol. 145, no. Supplement2, pp. S204–S209, 2020.
[5] L. Fricke-Oerkermann *et al.*, "Prevalence and course of sleep problems in childhood," *Sleep*, vol. 30, no. 10, pp. 1371–1377, 2007.
[6] Z. Cui, X. Zheng, X. Shao, and L. Cui, "Automatic sleep stage classification based on convolutional neural network and fine-grained segments," *Complexity*, vol. 2018, 2018.
[7] M. Samaee, M. Yazdi, and D. Massicotte, "Multi-modal signal integration for enhanced sleep stage classification: Leveraging EOG and 2-channel EEG data with advanced feature extraction," *Artificial Intelligence in Medicine*, p. 103152, 2025.
[8] M. M. Grigg-Damberger, "The aasm scoring manual: a critical appraisal," *Curr Opin Pulm Med*, vol. 15, no. 6, pp. 540–549, 2009.
[9] A. F. Babaei, J. Tanha, M. A. Balafar, and S. Roshan, "A novel multimodal deep learning approach with loss function for detection of sleep apnea events," *IEEE Access*, 2025.
[10] J. V. Rundo and R. Downey, "Chapter 25 - polysomnography," in *Handbook of Clinical Neurology*, K. H. Levin and P. Chauvel, Eds. Elsevier, 2019, vol. 160, pp. 381–392.
[11] E. Blok *et al.*, "Sleep and mental health in childhood: a multi-method study in the general pediatric population," *Child and Adolescent Psychiatry and Mental Health*, vol. 16, no. 1, p. 11, 2022.
[12] L.-L. Wang *et al.*, "Measuring sleep quality from eeg with machine learning approaches," in *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2016, pp. 905–912.
[13] H. Lee and A. Saeed, "Automatic sleep scoring from large-scale multi-channel pediatric eeg," in *NeurIPS 2022 Workshop on Learning from Time Series for Health*, 2022.
[14] W. Nazih *et al.*, "Influence of channel selection and subject's age on the performance of the single channel eeg-based automatic sleep staging algorithms," *Sensors*, vol. 23, no. 2, p. 899, 2023.
[15] J. Choi *et al.*, "Validation of the influence of biosignals on performance of machine learning algorithms for sleep stage classification," *DIGITAL HEALTH*, vol. 9, p. 20552076231163783, 2023.
[16] I. Choi and W. Sung, "Sleep model–a sequence model for predicting the next sleep stage," *arXiv preprint arXiv:2302.12709*, 2023.
[17] H. Dong *et al.*, "Mixed neural network approach for temporal sleep stage classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 2, pp. 324–333, 2017.
[18] S. H. Choi *et al.*, "Long short-term memory networks for unconstrained sleep stage classification using polyvinylidene fluoride film sensor," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 12, pp. 3606–3615, 2020.
[19] M. Irfan *et al.*, "Multidomain selective feature fusion and stacking based ensemble framework for eeg-based neonatal sleep stratification," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–10, 2025.
[20] S. H. Mostafaei, J. Tanha, and A. Sharafkhaneh, "EnsembleSleepNet: a novel ensemble deep learning model based on transformers and attention mechanisms using multimodal data for sleep stages classification," *Applied Intelligence*, vol. 55, no. 7, pp. 1–21, 2025.
[21] M. U. Khan, S. Samer, M. D. Alshehri, N. K. Baloch, H. Khan, F. Hussain, S. W. Kim, and Y. B. Zikria, "Artificial neural network-based cardiovascular disease prediction using spectral features," *Computers and Electrical Engineering*, vol. 101, p. 108094, 2022.
[22] J. Huang *et al.*, "An improved neural network based on senet for sleep stage classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 10, pp. 4948–4956, 2022.
[23] S. Matsumori *et al.*, "Haru sleep: A deep learning-based sleep scoring system with wearable sheet-type frontal eeg sensors," *IEEE Access*, vol. 10, pp. 13 624–13 632, 2022.
[24] C. Sun *et al.*, "A two-stage neural network for sleep stage classification based on feature learning, sequence learning, and data augmentation," *IEEE Access*, vol. 7, pp. 109 386–109 397, 2019.
[25] H. Phan *et al.*, "Joint classification and prediction cnn framework for automatic sleep stage classification," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 5, pp. 1285–1296, 2018.
[26] A. Supratak *et al.*, "Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 1998–2008, 2017.
[27] A. Martín-Montero *et al.*, "Pediatric sleep apnea: Characterization of apneic events and sleep stages using heart rate variability," *Computers in Biology and Medicine*, p. 106549, 2023.
[28] H. Lee *et al.*, "Explainable vision transformer for automatic visual sleep staging on multimodal psg signals," *npj Digital Medicine*, 2025.
[29] N. Jirakittayakorn *et al.*, "An enzyme-inspired specificity in deep learning model for sleep stage classification using multi-channel psg signals input: Separating training approach and its performance on cross-dataset validation for generalizability," *Computers in Biology and Medicine*, vol. 182, p. 109138, 2024.
[30] H. Cheng *et al.*, "Comparison of automated deep neural network against manual sleep stage scoring in clinical data," *Computers in Biology and Medicine*, vol. 179, p. 108855, 2024.
[31] M. Olsen *et al.*, "A flexible deep learning architecture for temporal sleep stage classification using accelerometry and photoplethysmography," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 1, pp. 228–237, 2022.
[32] M. U. Khan, M. A. Kamran, W. R. Khan, M. M. Ibrahim, M. U. Ali, and S. W. Lee, "Multi-sensor fusion for remote sensing of metallic and non-metallic object classification in complex soil environments and at different depths," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024.



FIRST A. AUTHOR (Fellow, IEEE)