Medical & Biological Engineering & Computing (2024) 62:2769–2783
https://doi.org/10.1007/s11517-024-03096-x

ORIGINAL ARTICLE

SleepBoost: a multi-level tree-based ensemble model for automatic sleep stage classification

Akib Zaman[1] · Shiu Kumar[2] · Swakkhar Shatabda[3] · Iman Dehzangi[4,5] · Alok Sharma[6,7]

Received: 12 December 2023 / Accepted: 14 April 2024 / Published online: 3 May 2024

Abstract

Neurodegenerative diseases often exhibit a strong link with sleep disruption, highlighting the importance of effective sleep stage monitoring. In this light, automatic sleep stage classification (ASSC) plays a pivotal role, now more streamlined than ever due to the advancements in deep learning (DL). However, the opaque nature of DL models can be a barrier in their clinical adoption, due to trust concerns among medical practitioners. To bridge this gap, we introduce SleepBoost, a transparent multi-level tree-based ensemble model specifically designed for ASSC. Our approach includes a crafted feature engineering block (FEB) that extracts 41 time and frequency domain features, out of which 23 are selected based on their high mutual information score (>0.23). Uniquely, SleepBoost integrates three fundamental linear models into a cohesive multi-level tree structure, further enhanced by a novel reward-based adaptive weight allocation mechanism. Tested on the Sleep-EDF-20 dataset, SleepBoost demonstrates superior performance with an accuracy of 86.3%, F1-score of 80.9%, and Cohen kappa score of 0.807, outperforming leading DL models in ASSC. An ablation study underscores the critical role of our selective feature extraction in enhancing model accuracy and interpretability, crucial for clinical settings. This innovative approach not only offers a more transparent alternative to traditional DL models but also extends potential implications for monitoring and understanding sleep patterns in the context of neurodegenerative disorders. The open-source availability of SleepBoost's implementation at https://github.com/akibzaman/SleepBoost can further facilitate its accessibility and potential for widespread clinical adoption.

Keywords Sleep stage classification · Sleep disruption · Ensemble learning · Feature engineering · Deep learning

1 Introduction

Sleep is fundamental to human well-being, occupying about a third of our lives and profoundly influencing our mental and physical health [1, 21]. Unfortunately, a myriad of neurodegenerative diseases linked to sleep disorders often go undetected due to their subtle nature, sometimes taking years before severe symptoms emerge [34, 38, 44]. The intersection between disrupted sleep patterns and neurodegenerative processes necessitates a deeper understanding and more accurate monitoring of sleep stages. Defined via overnight polysomnograms (PSG) with multiple 30-s epochs, sleep stages serve as critical tools for screening, assessing, and diagnosing sleep disturbances, which may precede or exacerbate neurodegenerative conditions [32]. These stages, categorized by the American Academy of Sleep Medicine Manual (AASM), include waking (W), rapid eye movement (REM), and three non-rapid eye movement (NREM) stages: N1, N2, and N3 [3].

✉ Shiu Kumar
shiu.kumar@fnu.ac.fj

1 Computer Science and Artificial Intelligence Laboratory (CSAIL), Electrical Engineering and Computer Science Department, Massachusetts Institute of Technology, Cambridge, MA, USA

2 School of Electrical & Electronics Engineering, Fiji National University, Suva, Fiji

3 Centre for Artificial Intelligence and Robotics (CAIR), United International University, Dhaka, Bangladesh

4 Department of Computer Science, Rutgers University, Camden, NJ, USA

5 Center for Computational and Integrative Biology, Rutgers University, Camden, USA

6 Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan

7 Institute for Integrated and Intelligent Systems, Griffith University, Nathan, Brisbane, QLD, Australia

However, the current manual classification following AASM regulations is labor-intensive and time-consuming. This system struggles to meet the needs of the vast population suffering from sleep disorders, given the high costs associated with each PSG recording and the resultant limited data collection. These constraints underscore the urgency for automating sleep staging, which promises both efficiency and broader accessibility for sleep diagnostics, potentially aiding in early detection and management of neurodegenerative diseases.

In the quest for solutions, electroencephalography (EEG) signals have emerged as a promising avenue. Beyond sleep-stage classification, EEG's applications extend to areas like emotion recognition [15, 45, 47, 48], seizure detection [11, 33, 39], and motor imagery classification [13, 14, 17]. These applications are particularly relevant in the context of neurodegenerative diseases, where early detection and ongoing monitoring are crucial.

The advent of generalized PSG datasets has catalyzed significant advancements in the industry [6, 23, 31, 49]. From these datasets, handcrafted features spanning time, frequency, and non-linear domains are extracted and employed in conventional machine learning techniques. Tools like support vector machine (SVM) [2] and random forest (RF) [10, 36, 50, 51] have been instrumental in pioneering automated sleep scoring systems. A standout advantage of these conventional models lies in their relative simplicity, yielding satisfactory results. This simplicity facilitates seamless integration into monitoring systems for practical applications, including those focused on neurodegenerative disorders. Delving deeper, a detailed analysis of these extracted features can elucidate the significance of different attributes. Notably, in one study, a substantial 85% of 11 selected features were derived from the frequency domain [51].

However, these models are not without limitations. They often overlook the time-series nature of raw EEG data and the intricate contextual relationships therein, presenting a notable shortcoming in their design.

Deep learning (DL) models stand in contrast to traditional methods, offering a robust solution to manage expansive datasets. As public sleep data sets burgeon, with participant counts often scaling from hundreds to thousands [23, 49], DL models are uniquely poised to handle this surge. They excel by continuously refining their learning from incremental data batches until they converge to the optimal model. One salient advantage of DL is its capability to autonomously identify features from rudimentary signals, eliminating the tedious manual crafting of intricate features. The potential of these models in understanding complex neurogenerative pathways linked to sleep disorders is immense.

Several architectures have been floated for autonomous sleep scoring using DL. These include convolutional neural networks (CNNs) [20, 37, 43, 46, 52], recurrent neural networks (RNNs) [4, 18], and hybrid models that leverage both CNNs and RNNs [22, 35, 40, 41]. A recent trend in the field has seen the adoption of attention mechanisms, with many studies [5, 19, 26, 28, 53] presenting it as an apt alternative to RNNs, especially for integrating contextual nuances. Additionally, the concept of transfer learning has been explored in numerous works [7, 24, 27]. This approach leverages vast source datasets to transfer domain-specific knowledge, decision rules, and architectural nuances to augment models trained on smaller target datasets.

The integration of artificial intelligence (AI) into healthcare, particularly deep learning, is met with enthusiasm and skepticism. Despite AI's advances, medical professionals are wary due to the "black box" nature of DL models, spread across interconnected hidden layers. This lack of transparency is especially concerning in areas like automated sleep stage classification (ASSC). Several studies have highlighted the challenges this poses [42, 53]. Concerns persist about the accuracy of DL models, especially when confidently classifying conflicting stages [16, 28]. The trend towards increasingly complex DL solutions sometimes yields only marginal benefits and demands more data and intensive training.

We believe that reverting to a judiciously crafted traditional machine learning model, emphasizing the most pertinent features, might offer a viable solution. Such an approach has the potential to match, if not surpass, the results of DL models, but with reduced complexity and fewer parameters. In our research, we introduce SleepBoost, a novel multi-level tree-based ensemble model tailored for ASSC. We rigorously evaluated its efficacy on the Sleep-EDF-20 dataset, emphasizing its comparative performance against prevailing deep learning models.

Our approach began with the establishment of a feature engineering block (FEB). From this, we extracted 41 time and frequency domain features. Subsequent refinement using a mutual information score threshold (>0.23) identified the most pertinent 23 features. These features underpin SleepBoost's core architecture—a sophisticated multi-level tree structure blending three fundamental tree-based ensemble models: categorical boost (CatBoost), light gradient boost (LGBoost), and random forest (RF). Additionally, we integrated a reward-based adaptive weight allocation mechanism, optimized using the FEB-filtered features. Empirical results indicate our model's superior performance: 86.3% accuracy, 80.9% F1-score, and Cohen's kappa score of 0.807, besting state-of-the-art deep learning alternatives. A salient advantage of SleepBoost is its efficiency: it possesses 1000 to 10,000 times fewer parameters than DL models, and its inference time, inclusive of processing and feature extraction, averages under 110 s for a typical Sleep-EDF dataset night. We further underscored the pivotal role of feature selection through an ablation study, examining four SleepBoost variants. Our aim is to not only offer a fresh

perspective on ASSC but also encourage fellow researchers to concurrently evaluate the potential of both traditional ML and DL models. Such synergy promises enhanced model portability and real-world applicability, especially in clinical settings and compact devices. For those interested, the SleepBoost source code is accessible at https://github.com/akibzaman/SleepBoost.

The remainder of this paper unfolds as follows: In Sect. 2, we delve into the methodology and elaborate on the architecture of SleepBoost. Section 3 showcases the comparative experimental results of SleepBoost vis-a-vis existing models and emphasizes an ablation study on the significance of feature selection. Finally, Sect. 4 wraps up our findings, discussing the constraints of our research and potential avenues for future exploration.

2 Methods

The materials used in this study together with a detailed description of the research methodology are presented in this section. In this study, we utilized EEG signals of the Fpz–Cz channel of the Sleep-EDF dataset. Later, we developed a FEB, extracting 41 features from the time and frequency domains and selecting 23 features using a mutual information score threshold value. Finally, we developed SleepBoost, a multi-level tree-based ensemble model tuned by a reward-based adaptive weight allocation that uses the filtered features of FEB to classify the sleep stages.

2.1 Description of the dataset

The Sleep-EDF dataset, which is accessible via the PhysioNet website, has been utilized for evaluating the performance of the proposed method in this study. EEG data in the Sleep-EDF dataset were obtained from 20 healthy participants aged 25 to 101 years. Each PSG recording in this dataset has two EEG signals acquired at 100 Hz from the Fpz–Cz and Pz–Oz channels. One electrooculogram (EOG) signal, one electromyography (EMG) signal, and one oro-nasal respiration signal were also acquired alongside the EEG signals. Following earlier research [5, 37, 40, 41], we utilized the data from the Sleep Cassette study and used a single Fpz–Cz channel as the input in this study. Experts annotated each EEG recording of the dataset every 30 s separately in the hypnogram files in eight sleep stages: awake (W), N1, N2, N3, N4, REM, movement time (M), and unknown condition (U). Note that the epochs associated with the M and U states are excluded from the experimental investigation. The Rechtschaffen and Kales (R&K) standard [32] classifies sleep cycles into six distinct stages of sleep, which is therapeutically essential [8], and the AASM standard also proposes combining the N3 and N4 stages of sleep into a single class [3]. Therefore, a five-class categorization has been considered for classifying sleep stages in several works [5, 22, 40]. In this study, we also treated N3 and N4 as stage 3 (N3) as per the AASM standards and considered the EEG signal collected from the Fpz–Cz channel following previous works [4, 5, 26, 28, 40, 41]. Table 1 highlights the number of EEG epochs corresponding to various sleep stages. In this work, we split the dataset into three non-overlapping segments: (a) feature selection (14,018 epochs), (b) training (25,708 epochs), and test (9,719 epochs). We utilize feature selection to select the important features from extracted features, training data to train the developed model and finally, test data to run the generalization and ablation evaluation of the trained models.

2.2 Development of feature engineering block (FEB)

EEG signal detection can capture unrelated signals like electrocardiogram (ECG), electromyography (EMG), electrooculogram (EOG), and respiration sounds due to the movements of muscles and bones, creating physiological signal artifacts. To mitigate these effects, the EEG signal is filtered using a second-order Butterworth Bandpass Filter with a 0.5 Hz and 45 Hz cutoff frequency. It is partitioned into seven frequency sub-bands and distributed in 30-s epochs (see Table 2). For accurate sleep staging, 41 features from the time and frequency domain have been obtained.

2.2.1 Time domain features

Empirical mode decomposition (EMD) is a robust time–frequency signal analysis technique suitable for getting the instantaneous properties of non-linear and non-stationary EEG signals by retaining the data attributes during the decomposition process. EMD splits each EEG signal into a set of 7 intrinsic mode functions (IMF) and forms a mathematical expression as shown in (1) in which $X(t)$ is the tth

Table 1 Number of epochs considered for each sleep condition from the Sleep-EDF database. FS denotes feature selection data, TR denotes training data, and TS denotes test data

Sleep condition	FS	TR	TS	Total number of epochs	% of total data
Wake (W)	2891	5303	2005	10,199	20.63
N1	923	1692	640	3255	6.58
N2	5877	10,777	4074	20,728	41.92
N3	1848	3390	1282	6520	13.19
REM	2479	4546	1718	8743	17.68
Total	14,018	25,708	9719	49,445	

Table 2 Frequency ranges of the partitioned sub-bands

Name of the sub-band	Frequency range (Hz)
Total band power	0.50–45.00
Delta-Low (low-δ)	0.50–2.00
Delta-High (high-δ)	2.01–4.00
Theta (θ)	4.01–8.00
Alpha (α)	8.01–12.00
Beta-Low (low-β)	12.01–20.00
Beta-High (high-β)	20.01–30.00
Gamma-Low (low-γ)	30.01–45.00

EEG epoch, Y_i is the ith IMF of total n number of functions, and $r(t)$ is the residue generated in the process.

$$X(t) = \sum_{i=1}^{n} Y_i + r(t) \tag{1}$$

This approach is chosen rather than selecting a single IMF based on a specific criterion because it allows for a more detailed analysis of the data. By leveraging all 14 features from each IMF, we utilize the full spectrum of information available in the data and capture a broader range of intrinsic oscillatory modes inherent in the data. After the extraction of IMFs, Welch's method is applied to each IMF to estimate their Power Spectral Density (PSD), enabling a detailed analysis of the signal's frequency content. A total of 14 features were extracted using the obtained IMFs denoted T1 to T14, which include mean (X_{mean}), standard deviation (S^2), variance, minimum, maximum, argMinimum, argMaximum, root mean square (RMS), median, maximum-minimum distance (MMD), skewness, kurtosis, Hjorth mobility (HM), and Hjorth complexity (HC) (refer to Table S1 of the supplementary material for the formulas).

Frequency domain features The features of EEG signals across the various frequency ranges of the different sleep stages are vital for accurately identifying the unique attributes of a specific sleep stage and classifying it accordingly. Eight direct features and 19 derived features are extracted from each epoch of the EEG signal resulting in 27 frequency domain features. The N1 stage demonstrates mostly the presence of Theta (θ) and Alpha (α) sub-bands. During the N2 stage, the amplitude of the EEG signal increases, and the θ sub-band becomes more prominent. Subsequently, during N3, the θ and Delta-High (high-δ) sub-bands are more propagated. During N4, the frequency of the EEG signal drops further and fluctuates in the frequency range of the Delta-Low (low-δ) sub-band. On the contrary, Beta-Low (low-β), Beta-High (high-β), and Gamma-Low (low-γ) sub-bands are predominant in the REM stage. These facts point out the importance of extracting features related to the frequency domain. Several related works [2, 10, 36, 50, 51] also utilized frequency domain features to develop the sleep stage classification model. Imtiaz et al. [9] developed a REM detection algorithm using spectral edge frequency in the 8–16 Hz frequency band, together with the absolute power and the relative power of the signal. In a broader extent, Zhao et al. [50] calculated the energy of the individual sub-bands to derive the frequency domain features. We utilized the time-domain signals of each sub-band segment and converted them to frequency-domain signals using a fast Fourier transform. The band energy can be obtained according to the different frequency ranges, which have been acquired as the Direct Frequency Domain (DFD) features and were labelled E1 to E8 (refer to Table S2 of the supplementary material for the formulas). Subsequently, inspired by previous work [51], we utilize the DFD features to calculate 19 Derived Frequency Domain (DeFD) features denoted D1 to D19. The equations used for obtaining the DFD and DeFD features can be found in Table S2 and S3 of the supplementary material, respectively.

2.3 Selection of important features

To simplify the computing procedure and enhance the algorithm's portability, we ran feature screening on the extracted features from Sect. 2.2. Moreover, it was assumed that some of the features are more relevant than the other features. Among several feature selection techniques, we utilized the mutual information (MI)-based feature selection technique for selecting the most relevant features. The rationale for choosing MI over other feature selection methods was due to its simplicity and its capability to consider the non-linear dependence of the features. Since the dataset is not massively large in terms of dimensionality, challenges like computational complexity were not an issue in this situation.

MI [25] is a numerical score between two (potentially multidimensional) random variables X and Y that measures the amount of knowledge received about one random variable through the other. Equation (2) shows the calculation of MI:

$$I(X, Y) = \int_x^{\cdot} \int_y^{\cdot} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dxdy \tag{2}$$

where $p(x, y)$ represents the joint probability density function of X and Y and $p(x)$ and $p(y)$ represent the marginal density functions. MI determines the similarity between the joint distribution and the products of the marginal distributions that have been factored. If X and Y are fully irrelevant (and thus independent), then $p(x, y)$ equals $p(x)$ $p(y)$, and this integral is zero. We calculated the mutual information of each of the extracted features (X) and the sleep stage of that sample (Y). In this study, features with feature weight

coefficients greater than 0.23 (which is the mean MI value for all the extracted features) were chosen as the final set of selected features as the outcome of the FEB. We carried out experiments to determine this value by evaluating and analyzing values in the range of 0.1 to 0.3 at an interval of 0.025 using the feature selection dataset (see Sect. 2, Table 1). We acquire the best result for the value of 0.225 (rounded as 0.23), and thus, it has been used as the threshold for the feature weight coefficient during feature selection. A total of 23 features, namely T2–T5, T8, T10, T13, E1, E2, E4, E6–E8, D3–D7, D11, D12, D15, D16, and D18, were chosen based on the MI threshold score (>0.23), while considering all of the classes. This result is also consistent when we consider a single class at a time. We also further evaluate which features were more important and closely related for any specific class (see Sect. 3.3).

2.4 Architecture of SleepBoost

Using the selected features, we developed a multi-level tree-based ensemble model by fusing the prediction of three individual prediction models: categorical boost (CatBoost), light gradient boost (LGBoost), and RF. The general architecture of the ensemble model utilizes a weighted majority voting–based architecture, highlighted in Fig. 1. The rationale to employ an additional ensemble layer atop existing tree ensembles is rooted in empirical evidence suggesting that such meta-ensembles can achieve superior performance by harnessing the strengths of different models while mitigating their individual weaknesses. Individual tree ensembles, though powerful, may develop biases towards particular patterns within the training data. By integrating multiple tree ensembles, we effectively diversify the model's perspective, reducing the likelihood of overfitting and enhancing its generalization capabilities on unseen data. Moreover, the meta-ensemble approach was able to improve model robustness by capturing a wider array of complex patterns and interactions, which might be overlooked by single-layer ensembles.

Bayesian optimization has been utilized for optimizing the hyperparameters of RF, CatBoost, and LGBoost, delivering better overall performance. Thus, specific parameter tuning using random search and grid search is not required. For CatBoost, the learning rate is set to 0.8 from the hyperparameter tweaking along with 100 estimators. Additionally, the number of leaves is set to 100 with a minimum of two data in each leaf, and the bagging percentage is set to 0.65, meaning that 65% of rows are used per tree-building iteration. In addition, the objective was multiclass with a weighted average for each class. In the case of RF, we employ 200 trees, which aids in avoiding overconfidence with a minimum leaf size of 1 and a minimum sample split of 9. We utilize the gradient-boosted decision tree (GBDT) for the LGB model to avoid the excessive computational burden of dart gradient boosting (dart). Moreover, the learning rate is set to 0.8 based on the hyperparameter tuning along with 100 estimators, the number of leaves is set to 100 with a minimum of two data in each leaf, and the bagging fraction is set to 0.65. In addition, the objective was multiclass with a weighted average for each class.

Firstly, we split the dataset into three segments named training (60%), validation (20%), and test (20%) sets. Then, we use the training set to train the individual classifiers and then the validation set to evaluate dependent features. We compare the prediction on the validation dataset with the actual classes. Then, the weights of the classifiers are adaptively calculated by offering a reward for the right prediction and normalizing the weight of each classifier using the cumulative sum of the weights of all three classifiers (see Algorithm S1 in Supplementary Material). Then based on these weights, we combine the output of the training dataset by considering the weights generated by the validation dataset.

3 Results and discussion

In this section, the evaluation procedure along with the performance metrics and results derived from SleepBoost are presented. Firstly, we describe the performance metrics utilized in this evaluation. We then compare SleepBoost with other conventional machine learning models. Later, we show the importance of feature selection by comparing four different variants of SleepBoost. Finally, we compare the performance of SleepBoost with other recognized models using the same SleepEDF-20 dataset.

3.1 Performance metrics

The tenfold cross-validation [12] method has been used to assess the performance of the classifiers under consideration. Furthermore, we use accuracy (ACC), precision (PR), recall (RE), F1-score (F1), and Cohen's kappa coefficient (K) to evaluate the model's performance on each epoch. ACC is the proportion of accurate predictions relative to the total number of predictions made by the model. PR is the ratio of correctly anticipated positives to the total number of positives. RE refers to the percentage of true positives to all class predictions. The weighted mean of PR and RE is F1. K quantifies the degree of agreement between actual and anticipated labels. A high value of K can indicate the effectiveness of the model.

3.2 Comparison of conventional models and SleepBoost

We calculated average performance metrics over tenfold cross-validation from the Sleep-EDF dataset for baseline models and SleepBoost. Figure 2 provides the mean CM of

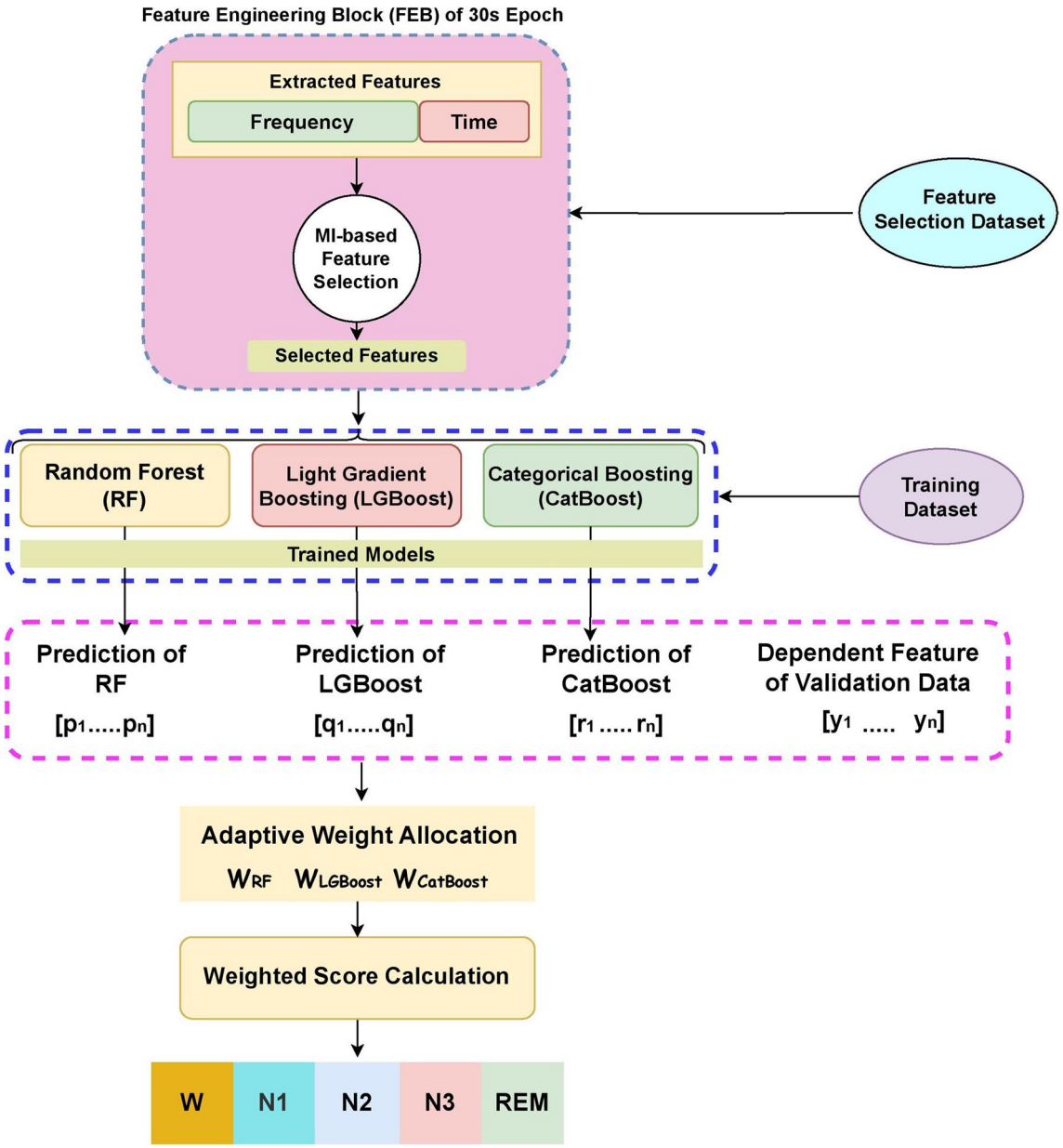


Fig. 1 General architecture of SleepBoost. We trained RF, LGBoost, and CatBoost as a unit block model for SleepBoost using the training dataset. Adaptive weight calculation is initiated using the prediction of the unit block models. Finally, a weighted score is calculated to predict the sleep stage

tenfold cross-validation from the Sleep-EDF dataset (using the selected extracted features) while Table 3 provides a summary of these metrics for both the baseline ensemble models and SleepBoost, using all features and only the selected ones. We also evaluate the variability of our method with respect to the data by visualizing the performance of our model across these ten folds (see Fig. 3), which shows consistent performance over the folds. SleepBoost achieved an accuracy of 84.8% with all the features and 86.3% with the selected features, outperforming baseline models. Notably, performance improved for all models when using the selected features, highlighting the importance of feature selection. More elaboration on this discussion is presented in detail in the next subsection. Using the selected features, SleepBoost demonstrates a Macro–PR of 82.8%, a Macro-RE of 80.3%, and a Macro-F1 of 80.9%, which outperforms the other models with a notable mean improvement of 1.42% (Macro-PR), 2.37% (Macro-RE), and 2.2% (Macro-F1) compared to the unit models (LGBoost, RF, and CatBoost) of SleepBoost. The developed model also demonstrates a high value of K (0.807), outperforming other models with a mean improvement of 1.26%. In terms of the area under the

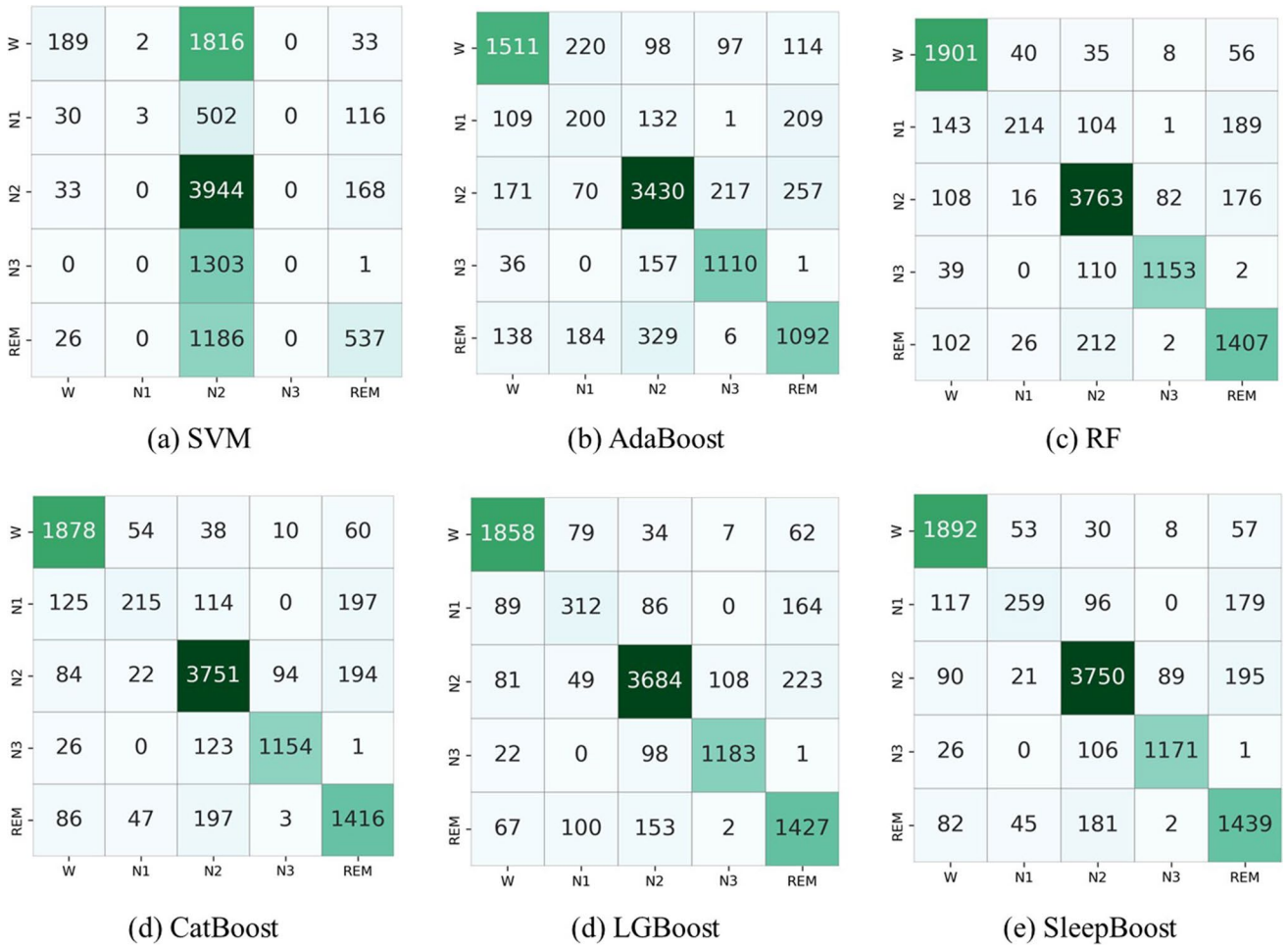


Fig. 2 Confusion matrices of all the models: **a** support vector machine (SVM), **b** adaptive boosting (AdaBoost), **c** random forest (RF), **d** categorical boosting (CatBoost), **e** light gradient boosting (LGBoost), and **f** SleepBoost

Table 3 Comparison of performance metrics among the base conventional models and SleepBoost

Model	Dataset	ACC	Macro PR	Macro RE	MF1	*K*	AUC
SVM	All features	51.4	41.2	40.5	39.1	0.307	0.703
	Selected features	52.1	41.3	40.9	39.8	0.333	0.670
RF	All features	84.3	81.2	75.9	77.1	0.783	0.902
	Selected features	85.1	**82.5**	76.7	77.9	0.793	0.908
CatBoost	All features	84.0	79.2	75.8	76.6	0.779	0.900
	Selected features	85.1	81.0	77.1	78.0	0.792	0.907
AdaBoost	All features	74.1	66.2	66.8	66.3	0.623	0.842
	Selected features	74.3	66.4	67.1	66.7	0.647	0.839
LGBoost	All features	84.0	78.4	77.8	78.1	0.781	0.910
	Selected features	**85.5**	80.7	**80.0**	**80.2**	**0.798**	0.920
SleepBoost	All features	84.8	80.3	77.3	78.5	0.792	**0.928**
	Selected features	**86.3**	**82.8**	**80.3**	**80.9**	**0.807**	**0.936**

Bold - indicates the best value

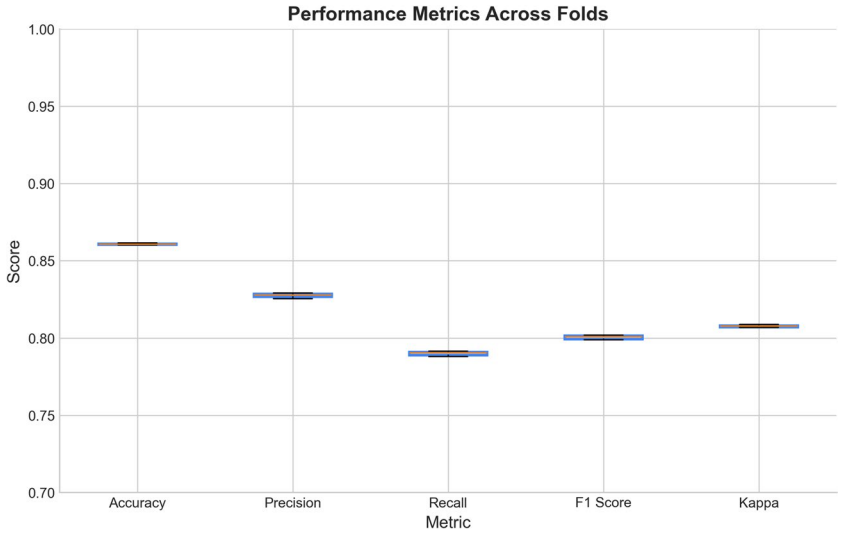


Fig. 3 Performance metrics of SleepBoost (selected features with adaptive weights) over ten folds

receiver operating curve (AUC), SleepBoost demonstrated a score of 0.936, i.e., it can successfully predict the sleep stage of 93.6% of the test epochs (see Fig. 2a).

In our study, the classification accuracy was high for stages W and N3, whereas the recognition accuracy was lower for stage N1, which has been summarized in Table 4. The proposed model shows the best ability to detect the W stage with an ACC of 92.7% and PR of 93.2%. Similarly, N3 also demonstrates a significantly close performance compared to W with an F1 of 91.4% and PR of 92%, which outperforms other classes in terms of PR and F1. By contrast, the performance of stage N1 classification is the worst, which is consistent with the results of existing related works. To be specific, 39.8% of N1 epochs are being recognized correctly, showing an F1 of 51%. On a similar note, REM also shows a below-average performance with an F1 of 79.4% and ACC of 82.3%.

The uniqueness of the EEG characteristics in different sleep stages may cause this phenomenon. Awareness is still relatively intact at the Wake (W) stage, and a blend of Alpha and Beta sub-bands characterize the EEG signal. The N1 is the transition phase of the brain from the aware state to the sleep state, during which the Alpha wave share steadily declines, and the Theta wave begins to develop and replace Alpha waves, indicating that the EEG signal changes dramatically during this period. So, with its steady characteristics, the W phase is more uncomplicated to recognize than the N1 phase with its more varied EEG signals. It should be noted that prior research [5] has demonstrated that an imbalance in the number of categories used during staging will impact the final accuracy, and the class with fewer sample epochs will produce lower classification performance. We observe that the test dataset of N1 is significantly lower than the other classes, which aligns with the poor performance demonstrated by this class. Since most of the real-life sleep datasets tend to be imbalanced, i.e., the amount of data in the N1 class will always be significantly smaller than the other classes. We have not generated synthetic samples of the N1 class to try and balance the data for each class. However, in future works, we will explore and evaluate the use of synthetic data. In some of the previous works, 5 classes have been transformed into 4 classes where N1 has been merged with N2. To evaluate the robustness of our model, we kept N1 separate to analyze the performance of this particular class. Additionally, the reported performance metrics of the SleepBoost are the worst-case or minimum possible metrics which also facilitated the comparison of the

Table 4 Class-wise performance of SleepBoost on test data

Sleep stage	Test samples	Correctly predicted samples	ACC	PR	RE	F1
W	2005	1858	92.7	86.3	93.2	89.5
N1	640	255	39.8	69.7	41.3	51.0
N2	4074	3687	90.5	90.1	90.3	90.2
N3	1282	1121	87.4	92.0	90.6	91.4
REM	1718	1414	82.3	77.2	82.5	79.4

prevailing models (explained in Sect. 3.4) in an unbiased and standardized way. We further reveal the hypnogram comparison labeled by experts and the model's prediction for one subject of Sleep-EDF datasets in Fig. 4.

3.3 Importance of selecting relevant features

The proposed FEB in this study supports a multi-domain integration of time and frequency domain features for each unique epoch, extracting 41 features, and selecting relevant features using the MI threshold score. Features with an MI score greater than 0.23 are selected and fed into SleepBoost for the classification of each epoch into their respective sleep stages. There are 7 time-domain features and 16 frequency-domain features (6 DFD and 10 DeFD) among the 23 selected features. The frequency domain data accounted for a more significant proportion of the automated sleep staging, followed by the time domain features. This may be probably because different stages of sleep exhibited distinct frequency and energy characteristics. In addition, the ratio of DeFD features is the highest among the filtered features of FEB, with a ratio of 8:17. Table 3 shows that SleepBoost, along with all the base models, had superior performance when trained and tested using the filtered features of FEB. We discovered an average improvement of 0.97 in ACC, 1.37 in Macro-PR, 1.33 in Macro-RE, 1.30 in MF1, and 0.02 in K while employing the selected features of FEB instead of using all the features. Figure 4 displays the AUC-ROC, which corroborates the conclusion that the AUC value for SleepBoost improved by 0.008 when trained using the selected features (Fig. 4b) in comparison to when all the features were used (Fig. 4a). A similar statement is also true for most of the other baseline models utilized in the evaluation study, as they also demonstrate better performance with the selected features (refer to Table 3) Fig. 5.

In addition, we developed four variants of SleepBoost to conduct a detailed ablation study to validate the importance of proper feature selection and performance improvement by utilizing the adaptive weight allocation during multi-level ensemble. The SleepBoost variants are as follows:

- M1: All features with trivial (balanced) weight allocation during ensemble.
- M2: Selected features with trivial (balanced) weight allocation during ensemble.
- M3: All features with adaptive weight allocation during ensemble.
- M4: Selected features with adaptive weight allocation during ensemble.

Figure 6 highlights the observed performance metrics from all four variants while evaluating on same test epochs. The radar curve in Fig. 6 demonstrates that in both the case of balanced and adaptive weight allocation during the ensemble, variants (M2 and M4) trained with selected features outperformed variants trained with all extracted features (M1 and M3). Alternatively, in both the case of selected and all features, adaptive weight allocation outperformed trivial weight allocation. In variants with all extracted features (M1 and M3), M3 with adaptive weight allocation outperformed M1 with balanced weight allocation. Similarly, in the case of variants with selected features (M2 and M4), M4 with adaptive weight allocation outperformed M2 with balanced weight allocation. In summary, feature extraction is critical along with selecting suitable features, which has been achieved in this study by incorporating

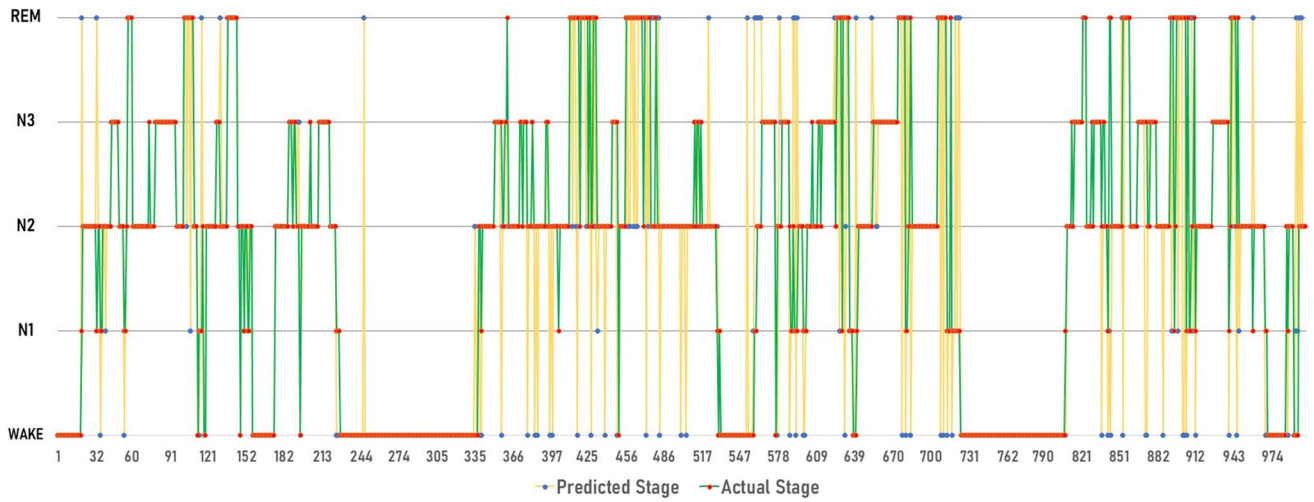


Fig. 4 Comparison of manual sleep stage labelling with SleepBoost model's prediction. The actual label represents the result of manual staging by experts, while the predicted label is the result of the SleepBoost model

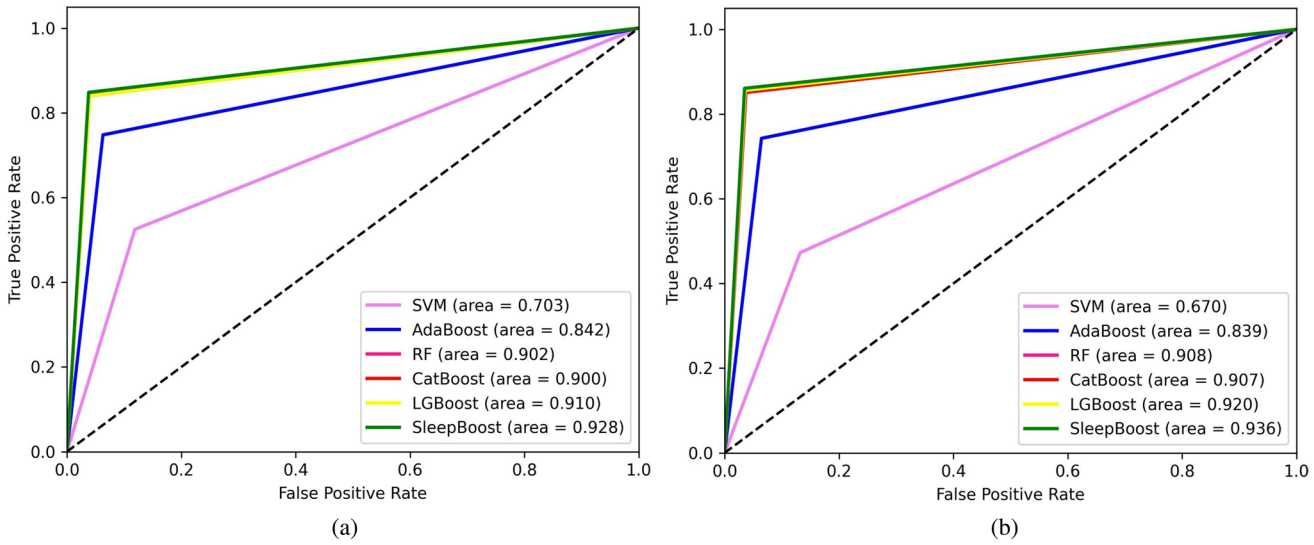


Fig. 5 Comparison of area under the receiver operating curve (AUC-ROC) among the conventional models and SleepBoost. **a** Using all extracted features; **b** using selected features of FEB

the MI score–based feature selection method. Additionally, adaptive weight allocation performs better than the trivial balanced weight allocation in tuning SleepBoost for the classification.

We further evaluate whether our feature selection procedure using MI has a negative impact on the performance of the classes with a small number of samples such as N1.

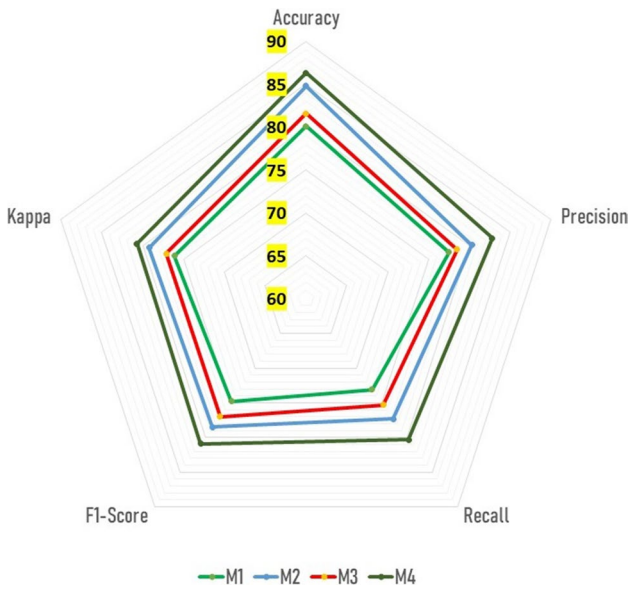


Fig. 6 Comparison of performance in different variants of SleepBoost. M1 (all features + balanced weight), M2 (selected features + balanced weight), M3 (all features + adaptive weight), and M4 (selected features + adaptive weight) represent four variants of SleepBoost with the combinations of features and weight allocation

Two important questions were raised in this process: (Q1) Whether some of the features were important for Class i (for example, REM) classification while not important for Class j (for example, N1)? (Q2) While choosing the important features based on all the classes, have we neglected some of the features that were important for an individual class?

We investigate these questions by calculating feature importance when only considering a single class, termed as Local MI Score (LMIS), versus when considering all classes, termed as Global MI Score (GMIS). Then, we compare the LMIS and GMIS of each feature in the case of all of the five classes and present the result in Fig. 7. Note that to eliminate the bias from sample inadequacy among some classes (for example, N1), both scores are scaled within the range of the minimum and maximum values of MI when all classes are considered, which ensures a fair comparison. The green dots indicate that both LMIS and GMIS are higher than the threshold for that feature, meaning the locally significant feature has also been successfully considered significant on a global scale. In contrast, the red dots indicate situations where LMIS is higher, while GMIS is lower than the threshold for that feature, suggesting that the locally significant feature has been overlooked in the global set of significant features.

This evaluation answers the earlier questions and adds several interesting findings. Firstly, there is a difference in feature importance for a specific class which has been illustrated in Fig. 7. For example, in N1 (6) and N2 (7) classes, a higher number of time domain features are important unlike N3 (2) and REM (3). There is a feature set that is important for all the classes. However, there are also features which are only dependent for a particular class but independent for other

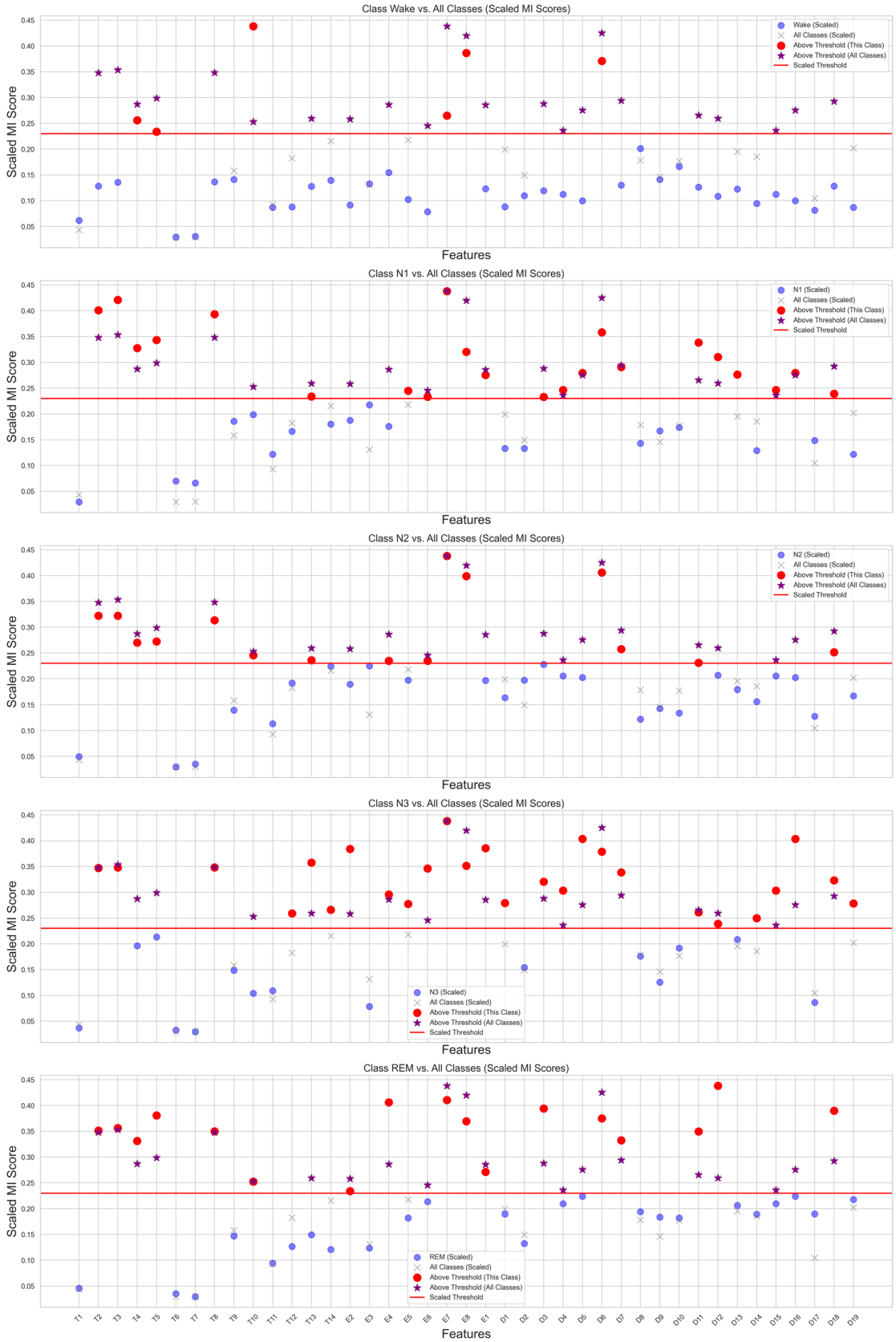


Fig. 7 Comparison of feature importance when considering a single class, termed as Local MI Score (LMIS), versus when considering all classes, termed as Global MI Score (GMIS)

classes. For example, feature E5 is dependent only for N1 and N3 but independent for all other classes. Similarly, T10 is important for Wake, N2, and REM but not for others. Moreover, the score of importance also varies from class to class for a particular feature. Secondly, while choosing the important features based on the GMIS, we have neglected a few of the features that are important for some of the classes like N1 (2) and N3 (6). For our future direction, we aim to try and mitigate this by exploring feature selection techniques that account for class imbalance or by employing alternative measures that can better capture the unique dependencies of features on minority classes. Techniques such as weighted MI, where weights are adjusted based on class proportions, or conditional feature selection methods might provide a more nuanced approach to retaining important features for all classes, including those with fewer samples. However, we were able to capture all the locally important features for Wake, N2, and REM classes, and most of the locally important features for N1 (13 out of 15) and N3 (16 out of 22) classes. These findings demonstrate that the feature selection based on GMIS is able to provide a better model performance despite undermining a few of the locally important features for N1 and N3 classes.

This investigation shows that feature engineering and selection is not only important to reduce the complexity of the model but also helps improve performance by feeding the right data to the model. These answers also open doors for further exploration in the domain of feature selection while classifying sleep stages and demonstrate the need for the growth of transparency to bridge the gap between medical practitioners and computer scientists. While the deep learning model does not provide this transparency, linear models can be useful to achieve this attribute.

3.4 Comparison of SleepBoost with the state-of-the-art deep learning models

Deep learning (DL) is a machine learning technique that transforms a specific input into a representation learned by the model's hidden layers instead of being developed beforehand, resulting in the consumption of scaled data. In contrast, most traditional machine learning (ML) models, including linear and tree-based models, do not create internal representations from the raw data. These models rely on learning relations on the supplied representations of the data fed into the model. Consequently, such models required extracted features from an EEG epoch to classify the sleep stage. We argue that classical ML models are capable of state-of-the-art performance or are on par with the DL models when deployed using relevant features. To validate this argument, we compared the performance of the recognized DL models with SleepBoost on the Sleep-EDF-20 dataset and summarized the result in Table 5. SleepBoost outperforms the DL models in terms of all the performance metrics and shows a promising future scope of exploring more conventional ML models in the domain of ASSC.

Table 5 Comparison of performance with the state-of-the-art methods evaluated on the Sleep EDF-20 dataset. Performance of SleepBoost variants has been scored on the test dataset (as described in Table 1)

Year	Model	Accuracy	F1-Score	K
2020	FKSVM-SRN [2]	82.22	-	-
2021	1D-CNN + Hidden Markov Model [46]	83.23	75.03	0.76
2021	CNN-TCN-CRF [50]	82.46	-	0.78
2018	CNN 1-Max Pool [26]	82.60	74.20	0.76
2021	C–C/R-R SleepNet [22]	84.29	79.81	0.78
2021	TinySleepNet [41]	83.10	78.10	0.77
2017	DeepSleepNet [40]	82.00	76.90	0.76
2019	IITNet [35]	84.00	78.00	0.78
2021	AttnSleep [5]	84.40	78.10	**0.79**
2020	CNN + Attention [53]	82.80	77.80	-
2019	SleepEEGNet [19]	82.83	77.02	0.77
2018	ARNN-SVM [29]	82.50	72.01	0.76
2021	RobustSleepNet [7]	-	78.60	-
2020	SeqSleepNet +, DeepSleepNet + [Expanded with Transfer Learning][30]	81.00	77.50	0.73
2019	SeqSleepNet [27]	**85.50**	**80.00**	**0.79**
Proposed	SleepBoost (M1)	80.10	74.81	0.76
	SleepBoost (M2)	84.83	78.53	**0.79**
	SleepBoost (M3)	81.60	77.01	0.77
	SleepBoost (M4)	**86.30**	**80.90**	**0.81**

Bold - indicates the best value

One of the recent works [42] also showed results aligning with the findings of our study, where a CatBoost-based linear model was developed, and decent results were achieved compared to the recognized DL models. Here we provide a more detailed comparison of our work with [42] which has achieved promising results for sleep stage classification. In [42], the authors used a linear model based on categorical boosting with 131 features extracted from a time and frequency domain from multiple windows resulting in 1048 features in total. In general, they used a combination of EEG, EOF, and EMG data. In our work, we extracted a total of only 41 time and frequency domain features from a single 30-s window and used only EEG data. Moreover, we employed the feature engineering block for feature selection using mutual information to select the important features. The selected features are then fed into the multiple tree-based ensemble model to predict the sleep stages. In terms of method, the feature engineering of both works is different. Unlike [42], we evaluate the effectiveness of feature selection through a detailed ablation study to demonstrate which features are important for a specific class (LMIS) as well as for all of the classes (GMIS). Furthermore, both [42] and our work demonstrate that linear models should be explored more in terms of providing further transparency in classification which is lacking in the deep learning models.

4 Conclusion and future directions

Sleep and/or wake cycle alterations are common in neurodegenerative diseases (ND), posing significant challenges in their early detection and management. Our research diverges from the prevalent deep learning (DL) focus, showcasing the efficacy of a traditional approach for automatic sleep stage classification (ASSC). Using the Sleep-EDF dataset's EEG signals, our feature engineering block (FEB) extracted 41 features, refining down to 23 through mutual information. SleepBoost, our main contribution, outperformed both baseline ensemble and leading DL models. Its lightweight design offers ease of training and deployment, and its transparency bolsters trust—a paramount consideration in medical applications, especially crucial in the sensitive domain of neurodegeneration, where understanding and interpreting sleep patterns can provide valuable insights into disease progression and patient well-being.

Challenges arose, notably in distinguishing between the W and N1 stages and between REM and N1 EEG waves due to their similarities. These challenges are particularly pertinent in neurodegenerative contexts, where sleep disturbances may present subtly but have profound implications. Our results highlight the potential of traditional machine learning in sleep scoring and advocate for its parallel consideration with DL models.

Future research can explore simpler models in ASSC for optimal portability and efficiency. Such models would be particularly valuable in the context of neurodegenerative diseases, where ease of use and interpretability are key. Additionally, investigating the specific impacts of various sleep stages on neurodegenerative disease progression can further enhance our understanding and management of these conditions.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11517-024-03096-x.

Acknowledgements Special thanks to the editors and anonymous reviewers who have provided their positive and constructive comments and suggestions to help improve our manuscript.

Data availability For the benefit of the research community, the SleepBoost algorithm (python scripts) is available at https://github.com/akibzaman/SleepBoost.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Aminoff Michael J, Boller François, Swaab Dick F (2011) We spend about one-third of our life either sleeping or attempting to do so. Handb Clin Neurol 98:vii–vii
2. Basha AJ, Saravana Balaji B, Poornima S et al (2021) Support vector machine and simple recurrent network based automatic sleep stage classification of fuzzy kernel. J Ambient Intell Humaniz Comput 12(6): 6189–6197
3. Berry RB, Brooks R, Gamaldo CE, Harding SM, Marcus C, Vaughn BV et al (2017) The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications. American Academy of Sleep Medicine, Version 2.4
4. Dong H, Supratak A, Pan W et al (2017) Mixed neural network approach for temporal sleep stage classification. IEEE Trans Neural Syst Rehabil Eng 26(2):324–333
5. Eldele E, Chen Z, Liu C et al (2021) An attention-based deep learning approach for sleep stage classification with single-channel EEG. IEEE Trans Neural Syst Rehabil Eng 29:809–818
6. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark R et al (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101(23)
7. Guillot A, Thorey V (2021) RobustSleepNet: transfer learning for automated sleep staging at scale. IEEE Trans Neural Syst Rehabil Eng 29:1441–1451
8. Iber C, Ancoli-Israel S, Chesson AL et al (2007) The new sleep scoring manual–the evidence behind the rules. J Clin Sleep Med 3(02):107–107
9. Imtiaz SA, Rodriguez-Villegas E (2015) An open-source toolbox for standardized use of PhysioNet sleep EDF expanded database. In: Paper presented at the 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC), Milan
10. Jadhav P, Datta D, Mukhopadhyay S (2021) Sleep stage classification based on ensemble decision tree technique using single-channel EEG. In: Paper presented at the international conference on big data, machine learning and applications, Allahabad
11. Jana R, Mukherjee I (2021) Deep learning based efficient epileptic seizure prediction with EEG channel optimization. Biomed Signal Process Control 68:102767–102767
12. Kohavi R et al (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. Paper presented at the international joint conference on AI (IJCAI-95), Canada

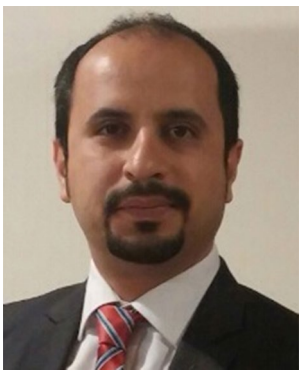
13. Kumar S, Sharma A (2018) A new parameter tuning approach for enhanced motor imagery EEG signal classification. Med Biol Eng Comput 56(10):1861–1874
14. Kumar S, Sharma A, Tsunoda T (2019) Subject-specific-frequency-band for motor imagery EEG signal recognition based on common spatial spectral pattern. In: Paper presented at the 16th Pacific rim international conference on artificial intelligence (PRICAI 2019), Cuvu, Yanuka Island, Fiji
15. Kumar S, Tsunoda T, Sharma A (2021) SPECTRA: a tool for enhanced brain wave signal recognition. BMC Bioinformatics 22(6):1–20
16. Liang S-F, Shih Y-H, Chen P-Y, Kuo C-E (2019) Development of a human-computer collaborative sleep scoring system for polysomnography recordings. PLoS One 14(7)
17. Miah MO, Muhammod R, Mamun KAA, Farid DM, Kumar S, Sharma A, Dehzangi A (2021) CluSem: accurate clustering-based ensemble method to predict motor imagery tasks from multi-channel EEG data. J Neurosci Methods:364
18. Michielli N, Rajendra Acharya U, Molinari F (2019) Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals. Comput Biol Med 106:71–81
19. Mousavi S, Afghah F, Acharya UR (2019) SleepEEGNet: automated sleep stage scoring with sequence to sequence deep learning approach. PLoS One 14(5)
20. Mousavi Z, Rezaii TY, Sheykhivand S, Farzamnia A, Razavi SN (2019) Deep convolutional neural network for classification of sleep stages from single-channel EEG signals. J Neurosci Methods 324
21. Muzet A (2007) Environmental noise, sleep and health. Sleep Med Rev 11(2):135–142
22. Neng W, Lu J, Xu L (2021) Ccrrsleepnet: a hybrid relational inductive biases network for automatic sleep stage classification on raw single-channel eeg. Brain Sciences 11(4)
23. O'Reilly Christian, Gosselin N, Carrier J et al (2014) Montreal Archive of Sleep Studies: an open-access resource for instrument benchmarking and exploratory research. J Sleep Res 23(6):628–635
24. Olesen AN, Jennum P, Mignot E, Sorensen HBD (2020) Deep transfer learning for improving single-EEG arousal detection. In: Paper presented at the 2020 42nd annual international conference of the IEEE engineering in Medicine & Biology Society (EMBC), Canada
25. Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238
26. Phan H, Andreotti F, Cooray N et al (2018) DNN filter bank improves 1-max pooling CNN for single-channel EEG automatic sleep stage classification. Annu Int Conf IEEE Eng Med Biol Soc 2018:453–456
27. Phan H, Chén OY, Koch P, Mertins A, Vos MD (2019) Deep transfer learning for single-channel automatic sleep staging with channel mismatch. In: Paper presented at the 27th European signal processing conference (EUSIPCO), Spain
28. Phan H, Mikkelsen K, Chén OY et al (2022) SleepTransformer: automatic sleep staging with interpretability and uncertainty quantification. IEEE Trans Biomed Eng 69(8):2456–2467
29. Phan H, Andreotti F, Cooray N, Chen OY, Vos MD (2018) Automatic sleep stage classification using single-channel EEG: learning sequential features with attention-based recurrent neural networks. In: Paper presented at the annual international conference of the IEEE engineering in medicine and biology society (EMBS), Honolulu, Hawaii
30. Phan Huy, Chen Oliver Y, Koch Philipp et al (2021) Towards more accurate automatic sleep staging via deep transfer learning. IEEE Trans Biomed Eng 68(6):1787–1798
31. Quan SF, Howard BV, Iber C et al (1997) The sleep heart health study: design, rationale, and methods. Sleep 20(12):1077–1085
32. Rechtschaffen A (1969) A manual for standardized terminology, techniques and scoring system for sleep stages in human subjects. Brain Information Service 20(2):246–247
33. Savadkoohi M, Oladunni T, Thompson L (2020) A machine learning approach to epileptic seizure prediction using electroencephalogram (EEG) signal. Biocybern Biomed Eng 40(3):1328–1341
34. Senaratna CV, Perret JL, Lodge CJ et al (2017) Prevalence of obstructive sleep apnea in the general population: a systematic review. Sleep Med Rev 34:70–81
35. Seo H, Back S, Lee S et al (2020) Intra-and inter-epoch temporal context network (IITNet) using sub-epoch features for automatic sleep scoring on raw single-channel. EEG Biomed Signal Process Control 61:102037–102037
36. Sharma R, Bilas Pachori R, Upadhyay A (2017) Automatic sleep stages classification based on iterative filtering of electroencephalogram signals. Neural Comput Appl 28(10):2959–2978
37. Sors A, Bonnet S, Mirek S et al (2018) A convolutional neural network for sleep stage scoring from raw single-channel. EEG Biomed Signal Process Control 42:107–114
38. Stephansen JB, Olesen AN, Olsen M et al (2018) Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. Nat Commun 9(1):1–15
39. Sukaria W, Malasa J, Kumar S et al (2022) Epileptic seizure detection using convolution neural networks. In: 2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA), pp. 1-5
40. Supratak Akara, Dong Hao, Chao Wu et al (2017) DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG. IEEE Trans Neural Syst Rehabil Eng 25(11):1998–2008
41. Olesen AN, Jennum P, Mignot E, Sorensen HBD (2020) Deep transfer learning for improving single-EEG arousal detection. In: Paper presented at the 42nd annual international conference of the IEEE engineering in Medicine & Biology Society (EMBC)
42. Van Der Donckt J, Van Der Donckt J, Deprost E et al (2023) Do not sleep on traditional machine learning: simple and interpretable techniques are competitive to deep learning for sleep scoring. Biomed Signal Process Control 81:104429
43. Vilamala A, Madsen KH, Hansen LK (2017) Deep convolutional neural networks for interpretable analysis of EEG sleep stage scoring. In: Paper presented at the 2017 IEEE 27th international workshop on machine learning for signal processing (MLSP), Japan
44. Wu H, Dunnett S, Ho Y-S et al (2019) The role of sleep deprivation and circadian rhythm disruption as risk factors of Alzheimer's disease. Front Neuroendocrin 54:100764–100764
45. Xu G, Guo W, Wang Y (2023) Subject-independent EEG emotion recognition with hybrid spatio-temporal GRU-Conv architecture. Med Biol Eng Compu 61(1):61–73
46. Yang B, Zhu X, Liu Y, Liu H (2021) A single-channel EEG based automatic sleep stage classification method leveraging deep one-dimensional convolutional neural network and hidden Markov model. Biomed Signal Process Control 68
47. Zaman A, Tahsin A, Rahman M, Akhter R, Rahman H, Mustary S, Farid DM (2022) Emotion detection for children on the autism Spectrum using BCI and web technology. In: Paper presented at the IEEE/WIC/ACM international joint conference on web intelligence and intelligent agent technology (WI-IAT), Canada
48. Zaman A, Khan RT, Karim N, Nazrul Islam M, Uddin MS, Hasan MM (2022) Intelli-helmet: an early prototype of a stress monitoring system for military operations. In: Garg L et al (eds) Information systems and management science. ISMS 2020. Lecture notes in networks and systems, vol 303
49. Zhang G-Q, Cui L, Mueller R et al (2018) The National Sleep Research Resource: towards a sleep data commons. J Am Med Inform Assoc 25(10):1351–1358
50. Zhao C, Neng W (2021) A sleep stage classification method via combination of time and frequency domain features based on Single-Channel EEG. In: 2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), New York City, pp 1102–1109

51. Zhao S, Long F, Wei X, Ni X, Wang H, Wei B (2022) Evaluation of a Single-Channel EEG-based sleep staging algorithm. Int J Environ Res Public Health 19(5):2845
52. Zhou D, Wang J, Hu G, Zhang J, Li F, Yan R, Kettunen L, Chang Z, Xu Q, Cong F (2022) Singlechannelnet: a model for automatic sleep stage classification with raw single-channel EEG. Biomed Signal Process Control:75
53. Zhu T, Luo W, Feng Yu (2020) Convolution-and attention-based neural network for automated sleep stage classification. Int J Environ Res Public Health 17(11):1–13

Akib Zaman is a PhD student at EECS, MIT, and affiliated with the Algorithmic Design Group (ADG) in the Computer Science and Artificial Intelligence Lab (CSAIL). His research interests include Robotics, Computer vision, Computer Graphics, and Signal Processing. His current research focuses on algorithmic and optimization techniques at the intersection of Robotics and Geometry Processing.

Shiu Kumar received his Ph.D. degree in signal processing and pattern recognition from the University of the South Pacific, Fiji. He is an Assistant Professor at Fiji National University. His research interests include brain–computer interface, bio-medical signal processing, pattern recognition, machine learning, and data mining.

Swakkhar Shatabda is a Professor in the Department of Computer Science and Engineering at United International University (UIU), Bangladesh. He achieved his Ph.D degree from the Institute for Integrated and Intelligent Systems (IIIS), Griffith University in 2014. His research interest includes bioinformatics, optimization, search and meta-heuristics, data mining, constraint programming, approximation algorithms, and graph theory.



Iman Dehzangi received his Ph.D. degree in Bioinformatics and computational biology from Griffith University, Brisbane, Australia. He is currently an Assistant Professor at the department of computer science at the Rutgers University, Camden, NJ, USA. His research interests include Bioinformatics, Genomics, Data mining, Statistical learning theory, Pattern Recognition, and Machine Learning.

Alok Sharma received his Ph.D. degree from Griffith University, Australia, in 2006. He is a Senior Scientist at RIKEN Center for Integrative Medical Sciences, Japan. He is also an Adjunct Professor at Griffith University, Australia. His research interests include pattern recognition, human cancer classification, and proteomics. He has over 150 publications in these areas with the h-index of 44.