Automatic Sleep Stage Classification Using Nasal Pressure Decoding Based on a Multi-Kernel Convolutional BiLSTM Network

Minji Lee, *Member, IEEE*, Hyeokmook Kang, Seong-Hyun Yu, *Student Member, IEEE*, Heeseung Cho, Junhyoung Oh, Glenn van der Lande, Olivia Gosseries, and Ji-Hoon Jeong, *Associate Member, IEEE*

***Abstract*— Sleep quality is an essential parameter of a healthy human life, while sleep disorders such as sleep apnea are abundant. In the investigation of sleep and its malfunction, the gold-standard is polysomnography, which utilizes an extensive range of variables for sleep stage classification. However, undergoing full polysomnography, which requires many sensors that are directly connected to the heaviness of the setup and the discomfort of sleep, brings a significant burden. In this study, sleep stage classification was performed using the single dimension of nasal pressure, dramatically decreasing the complexity of the process. In turn, such improvements could increase the much needed clinical applicability. Specifically, we propose a deep learning structure consisting of multi-kernel convolutional neural networks and bidirectional long short-term memory for sleep stage classification. Sleep stages of 25 healthy subjects were classified into 3-class (wake, rapid eye movement (REM), and non-REM) and 4-class (wake, REM, light, and deep sleep) based on nasal pressure. Following a leave-one-subject-out cross-validation, in the 3-class the accuracy was 0.704, the F1-score was 0.490, and the kappa value was 0.283 for the overall metrics. In the 4-class, the accuracy was 0.604, the F1-score was 0.349, and the kappa value was 0.217 for the overall metrics. This was higher than the four comparative models, including the class-wise F1-score. This result demonstrates the possibility of a sleep stage classification model only using easily applicable and highly practical nasal pressure recordings. This is also likely to be used with interventions that could help treat sleep-related diseases.**

***Index Terms*— Sleep stage classification, nasal pressure, deep learning, biomedical application, healthcare.**

Manuscript received 21 November 2023; revised 22 April 2024 and 18 June 2024; accepted 22 June 2024. Date of publication 28 June 2024; date of current version 17 July 2024. This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by Korean Government [Ministry of Science and ICT (MSIT)] (Artificial Intelligence Innovation Hub) under Grant RS-2021-II212068; in part by the National Research Foundation of Korea (NRF) grant funded by Korean Government (MSIT) under Grant RS-2023-00252624 and Grant RS-2024-00336880; and in part by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry (IPET) through the Agriculture and Food Convergence Technologies Program for Research Manpower Development Program funded by the Ministry of Agriculture, Food and Rural Affairs (MAFRA) under Grant RS-2024-00398561. *(Minji Lee and Hyeokmook Kang contributed equally to this work.) (Corresponding authors: Junhyoung Oh; Ji-Hoon Jeong.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of Samsung Medical Center under IRB No. 2021-04-133, and performed in line with the Declaration of Helsinki.

Minji Lee is with the Department of Biomedical Software Engineering, The Catholic University of Korea, Bucheon, Gyeonggi 14662, South Korea (e-mail: minjilee@catholic.ac.kr).

Hyeokmook Kang is with the Soldier Combat Research and Development Team and the Land Combat System Center, Hanwha Systems, Seoul, Gyeonggi 13524, South Korea (e-mail: khm812@hanwha.com).

Seong-Hyun Yu and Ji-Hoon Jeong are with the Department of Computer Science, Chungbuk National University, Cheongju, Chungcheongbuk 28644, South Korea (e-mail: sh.yu@chungbuk.ac.kr; jh.jeong@chungbuk.ac.kr).

Heeseung Cho is with the Department of Artificial Intelligence, Korea University, Seoul 02841, South Korea (e-mail: hscho9384@korea.ac.kr).

Junhyoung Oh is with the Division of Information Security, Seoul Women's University, Seoul 01797, South Korea (e-mail: ohjun02@gmail.com).

Glenn van der Lande and Olivia Gosseries are with the Coma Science Group, GIGA Consciousness, University of Liège, 4000 Liège, Belgium (e-mail: Glenn.vanderLande@uliege.be; ogosseries@uliege.be).

Digital Object Identifier 10.1109/TNSRE.2024.3420715

I. Introduction

SLEEP is one of the most fundamental human processes and plays an essential role in maintaining physical and cognitive functions in everyday life [1]. However, in recent years, the prevalence of sleep disorders, which pose challenges to public health, has increased annually [2], [3]. Sleep disorders (e.g., obstructive sleep apnea and insomnia) increase the risk of medical complications, including cardiovascular disease, diabetes, and depression [4], [5]. To effectively address the issues related to poor sleep, it is essential to establish a comprehensive understanding of the underlying biology of sleep. The classification of the sleep stage is usually performed using polysomnography (PSG) signals, including electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), and electrocardiography (ECG), as well as respiratory effort and other physiological signals [6], [7]. However, such an extensive protocol can be challenging, only achievable in laboratory or hospital settings, making it impractical as a widespread diagnostic tool.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

Sleep stages are classified mainly according to the American Academy of Sleep Medicine (AASM) standard [6]. This divides sleep into five stages: wakefulness (wake), rapid eye movement (REM), and three sleep stages (N1–N3) with non-rapid eye movement (NREM) [8]. The characteristics of the EEG signals in each sleep stage include, for example, sleep spindles for N2 and slow waves for N3. Outside of these dominant EEG changes, other signals such as EOG and EMG can also be used, prominent examples being horizontal eye movements and reduced muscle tension, respectively, that occur during REM sleep. In addition, various signals related to breathing are also associated with sleep stages. In general, breathing patterns can closely follow sleep stages in healthy participants, with light sleep associated with an irregular frequency and a moderate decrease in ventilation, which decreases further in deep sleep, but the frequency is stable, while during REM sleep breathing is erratic and shallow [9]. Compared to wakefulness, the rate of reversal of the nasal cycle, which indicates alternating decongestion and congestion of the nasal airways that produce a resistance change, is low during sleep [10]. In particular, the amplitude of the airflow signal affects the difference between the upper and lower envelopes of a nasal pressure signal during sleep [11]. In this sense, nasal pressure signals are also one of the important candidates for effectively classifying sleep stages.

Given that each sleep stage is defined based on unique EEG characteristics, most studies using EEG signals use the 5-class system (i.e., wake, N1–N3, and REM) [8], [12]. However, sleep stage classification studies using other biosignals employ a strategy that reduces the number of classes for two main reasons. First, changes in autonomic nervous activity are slower than brain cortical changes [13]. Second, N1 is much shorter than other sleep stages, making physiological changes outside of the EEG rarely noticeable, causing unnecessarily poor performance [13]. Some studies focus on the 3-class, that is, wake-NREM-REM, sleep stage classification [14]. However, within NREM, N3 has received special attention, as the most restorative period of sleep for metabolic function, associated with sleep maintenance and sleep quality, which makes it meaningful to divide NREM into a sleep classification system [15]. In this sense, most approaches focus on the 4-class sleep stage in which N1 and N2 are combined [16], resulting in sleep stages classified into four classes: wake, light sleep (N1–N2), deep sleep (N3), and REM [17].

Conventional manual classification of sleep stages can be time consuming and subjective because trained professionals must visually examine and classify neurophysiological signals [18]. Automated methods can decrease the subjective nature, and be less time-consuming. Consequently, this technological automation approach to sleep stage classification can improve the accuracy and efficiency of sleep analysis. Recently, automated sleep stage classification frameworks using deep learning, such as convolutional neural networks (CNNs) and bidirectional long short-term memory (biLSTM), have been proposed [19], [20]. The strength of CNNs is the excellent feature extraction, while the class of LSTM techniques utilizes the temporal context of a signal to optimize performance, which is potentially crucial in prediction based on time series [21]. Using deep learning instead of the appropriate hand-made features extracted from biosignals can increase classification performance because the characteristics of each sleep stage are inferred without restrictions from the training data [22]. In general, machine learning approaches using hand-crafted features that rely on expert experience and prior knowledge can be limited by hidden features, unknown to be important. Moreover, with the risk of error accumulation, they are not always guaranteed to be optimal for classification tasks. On the contrary, deep learning approaches are data-driven systems that can train feature representations for the sleep stage directly from raw data [16], [23].

In this study, we propose a deep learning framework consisting of a multi-kernel CNN and biLSTM, that extracts relevant characteristics of nasal pressure, for automatic sleep stage scoring, as depicted in Fig. 1. Merging at least N1 and N2 deals with the bigger issue of class imbalances in sleep classification tasks. However, these likely still persist, which is why, after pre-processing nasal pressure data, we used the synthetic minority oversampling technique (SMOTE) to resolve the class imbalance for sleep stages [24]. The sleep stage consisted of wake, NREM, and REM in the 3-class stage [14], and wake, light sleep, deep sleep, and REM in the 4-class stage [13], [25], [26]. Furthermore, the leave-one-subject-out cross-validation (LOO-CV) method was employed to demonstrate the generalizability of the proposed model [27]. The proposed framework using nasal pressure could demonstrate the potential for healthcare, in the sense that automatic sleep stage classification in a simplified recording procedure can aid in the diagnosis and treatment of sleep disorders.

II. Related Works

A. Types and Number of Biosignals

Several studies have been conducted to develop automated methods for sleep stage classification using multi-channel biosignals. Phan et al. [28] designed XSleepNet, which uses three neurophysiological signals: EEG, EOG, and EMG. They propose a sequence-to-sequence sleep staging model that can learn a joint representation from both raw signals and time-frequency features. Using the Montreal Archive of Sleep Studies (MASS) dataset, 87.6% overall accuracy was achieved in the 5-class sleep stage using EEG, EOG, and EMG signals, which is higher than the result obtained using only EEG signals (i.e., 85.2%). Cui et al. [7] presented an automatic sleep stage classification method based on CNNs combined with a fine-grained segment with multiscale entropy in EEG signals. Consequently, the authors achieved an average accuracy of 92.2% in another open-access sleep dataset, ISRUC-Sleep, for the 5-class classification using multi-channel EEG signals.

The classification often relies on the use of multi-channel biosignals, while the increased accuracy may not parallel the increased impractical acquisition and analysis. Therefore, the field has been turning to single-channel classification, recognizing that classification performance is hardly affected as it readily distinguishes transitions in the brain state [29]. As an example, Supratak et al. [30] proposed a deep learning model called DeepSleepNet for automatic sleep stage scoring

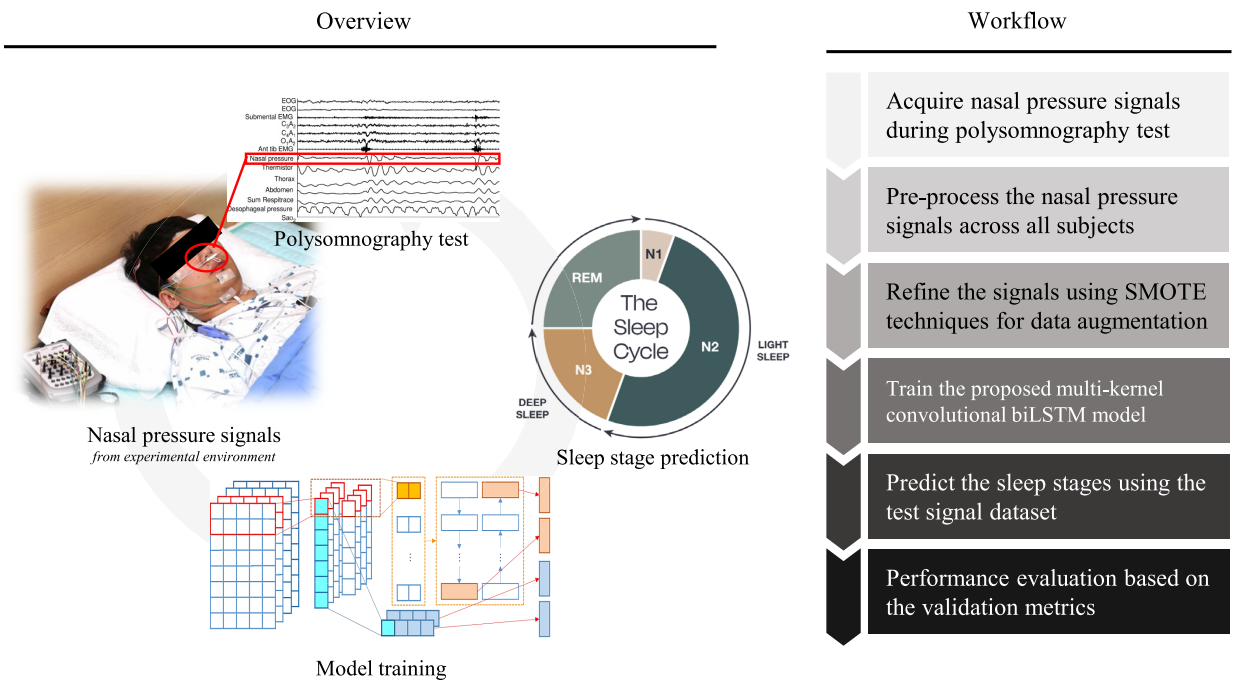

Fig. 1. Overview of the automatic classification of sleep stages using nasal pressure. The left side shows the overall framework for predicting sleep stages from data collection to model training based on a multi-kernel convolutional biLSTM network. Nasal pressure signals were acquired during the polysomnography test in an experimental environment. The sleep stage includes wake, NREM (N1+N2+N3), and REM for 3-class. In the 4-class, the sleep stages are wake, light sleep (N1+N2), deep sleep (N3), and REM. The right side briefly summarizes the entire process of the proposed sleep stage classification system.

based on raw single-channel EEG signals. DeepSleepNet utilizes CNNs to extract time-invariant features and biLSTM to learn the transition rules among sleep stages from the EEG. Using the Sleep-EDF dataset, the results indicated an overall accuracy of 82.0% from the bipolar Fpz-Cz EEG channel. Eldele et al. [31] introduced, which includes a feature extraction module that employs a multi-resolution CNN and adaptive feature recalibration to extract low- and high-frequency features of a single-channel EEG signal. Using the Sleep Heart Health Study (SHHS) dataset, an overall accuracy of 84.2% applied on C4-A1 Channel was achieved for the classification of 5 class sleep stages. Recalling the multi-channel XSleepNet's accuracy of 87.6%, these approaches are on par with the state-of-the-art performance.

EEG signals are measured by attaching electrodes to the scalp. The AASM recommends six electrodes and two references [6]. Placing electrodes on the head is time-consuming, can only be done after some training, and can interfere with the comfort required for good quality sleep [32]. Therefore, EEG signal measurement was attempted with other sensors. In automatic overnight sleep monitoring using a standardized in-ear EEG sensor, Nakamura et al. [33] observed abrupt electrode noise caused by participants' movements such as jaw. In addition, the recorded ear-EEG signal included physiological noise from respiration, which was overlaid by a slow oscillation of large amplitude. The around-ear EEG sensor, on the other hand, may be less discrete than in-ear EEG, but is more practical for detecting a wider range of brain signals because it covers a larger area [34]. However, there are few sleep stage classification studies using around-ear EEG sensors.

Changes occur throughout the body, not only within the brain, during sleep, providing a wide array of opportunities in the utilization of biosignals that are easier to acquire without compromising performance. In one such attempt, Fan et al. [35] used a single EOG signal, added oversampling techniques to address class imbalances, and performed sequence learning using a bidirectional gated recurrent unit. For the 4-class sleep stage, an overall accuracy of 82.1 % was achieved on the Sleep-EDF dataset. Furthermore, Sridhar et al. [25] predicted the sleep stage using heart rate measured by ECG signals. Using a CNN model, 77.0% accuracy was achieved on the SHHS dataset for the 4-class sleep stage. However, measurement of these biosignals can interfere with sleep, for example, by friction with clothing or skin [36].

Our study focused on an unprecedented approach to the classification of sleep stages through a single channel with nasal pressure signal. Despite its proven efficacy as a signal, nasal pressure signals have not been explored for classification of sleep stages. The use of nasal pressure signals for classification represents a novel contribution, opening novel pathways for sleep research and potential diagnostic methodologies.

B. State-of-the-Art Deep Learning Methods

Recently, various deep learning models have been developed to classify sleep stages. Many factors affect performance, from the input feature of the model to the detailed architecture of the deep neural network [29]. Simply because the model structure is complex does not necessarily result in high performance. Rather, if there are too many modules in the deep

neural network, it may require a long training time, reducing practicality.

When a single-channel input is used, it mainly focuses on the characteristics of the signals. The EEG signal is the most widely used feature in the sleep stage classification. There are also various types of inputs used for deep learning, and they are used in the form of time-series [12] by utilizing their characteristics, or in the form of time-frequency representation [37] by utilizing spectral information. When using a single EOG signal, only the sequence input was utilized as input of the model [35]. When using multi-channel, it focuses on the relationship between channels or signals. For example, the covariance feature matrix based on multivariate phase space reconstruction was used to utilize geometric properties and spatial information in multiple biosignals [38].

The input form of the signal eventually affects the structure of the CNN. Eldele et al. [31] used a 1D CNN model that selects local features by utilizing a multi-head attention technique from single-channel EEG signals. Mousavi et al. [39] also developed 1D CNN for sequence-to-sequence learning for automated sleep stage scoring. On the other hand, ElMoaqet et al. [37] utilized 2D CNN using time-frequency 2D image data as single-channel EEG signals for sleep stage classification. Even in studies using multi-channel signals, 2D CNNs are widely used if multi-channel data are used as it is [7]. However, 1D CNN is sometimes used to divide signals and put them in the model as a single channel [40].

Finally, it is also important to use deep transfer learning to remove the computational overhead required to set up and properly learn a deep learning scoring system from scratch [29]. A commonly used approach in transfer learning is the cross-dataset experiment. It takes a model that has already been trained on large sleep datasets and transfers it to the current model. In He et al. [41], a baseline model was trained using the SHHS dataset including a total of 5,793 subjects, and the proposed framework for sleep stage scoring was expanded by Sleep-EDF data consisting of 20 subjects. As a result, the performance increased compared to when the sleep stages were classified simply using EDF data. On the other hand, there are cases where models learned from completely different data are imported rather than the same classification model. However, in some cases, models learned from completely different data are used. In ElMoaqet et al. [37], the authors utilized GoogLeNet as a pre-trained CNN to transfer knowledge from natural images in time-frequency EEG data for automatic sleep stage scoring. However, using transfer learning does not necessarily increase performance [41], so much research is required.

III. Methods

A. Dataset

The 25 subjects (aged 35–61 years) were included in this study. Table I presents detailed information on the study participants. The exclusion criteria of the experiment were as follows: (i) Individuals diagnosed with cognitive decline, progressive mental or neurological diseases, lung disease, severe snoring, narcolepsy, REM sleep disorders, or clinically uncontrolled severe internal diseases (e.g., diabetes and hypertension), (ii) shift workers, pregnant and lactating women, and (iii) individuals diagnosed with insomnia.

TABLE I
Subjects' Demographic Characteristics

Description	Value
Age (years)	50.92 ± 9.57
Sex (Male:Female)	8:17
Height (cm)	164.32 ± 11.13
Weight (kg)	62.94 ± 13.75
Body Mass Index (BMI)	23.04 ± 2.41

Mean ± Standard Deviation

The PSGs were performed using Embla RemLogic 4.0 at Samsung Medical Center, Seoul, South Korea. The key feature of this study, nasal pressure, was also measured using the Embla N7000 device with MDrive. Each recording was scored by a skilled PSG technician according to the AASM manual. The ground-truth sleep stage was determined based on EEG signals, eye movement, muscle activity, and respiratory activity. The respiratory signal used here was measured by chest movements and, thus, separate from the nasal pressure measurements. The study protocol was approved by the Institutional Review Board of Samsung Medical Center (IRB No. 2021-04-133) and the study was conducted in accordance with the ethical standards described in the Declaration of Helsinki. Written informed consent was obtained from all participants and all data was de-identified.

B. Nasal Signal Pre-Processing

The initial signals were pre-processed to transform the input signal for model training. The Savitzky-Golay filter [42], [43] was used to smooth and minimize data noise, as shown in Fig. 2. The Savitzky-Golay filtering technique involves fitting a polynomial to a narrow window of consecutive data points. Subsequently, the coefficients derived from this polynomial are used to compute a smoothed value for the central data point positioned within the window. Data consist of a set of points x_j, y_j, $j = 1, \ldots, n$, where x_j is an independent variable and y_j is an observed value. The set of m convolution coefficients C_i is expressed as

$$Y_j = \sum_{i=\frac{1-m}{2}}^{\frac{m-1}{2}} C_i y_{j+i}, \ \left(\frac{m+1}{2} \le j \le n - \frac{m-1}{2}\right) \quad (1)$$

This iterative process was applied to each individual data point within the signal, ultimately generating a filtered signal from the original data. The resulting filtered data were then resampled, producing data with 2,000 timestamps (approximately corresponding to 66.6 Hz) within a 30-sec interval. Furthermore, normalization was performed by adjusting the data range to span from 0 to 1.

Next, we used SMOTE, a widely used machine learning method, to rectify class imbalances in the datasets [44], [45].

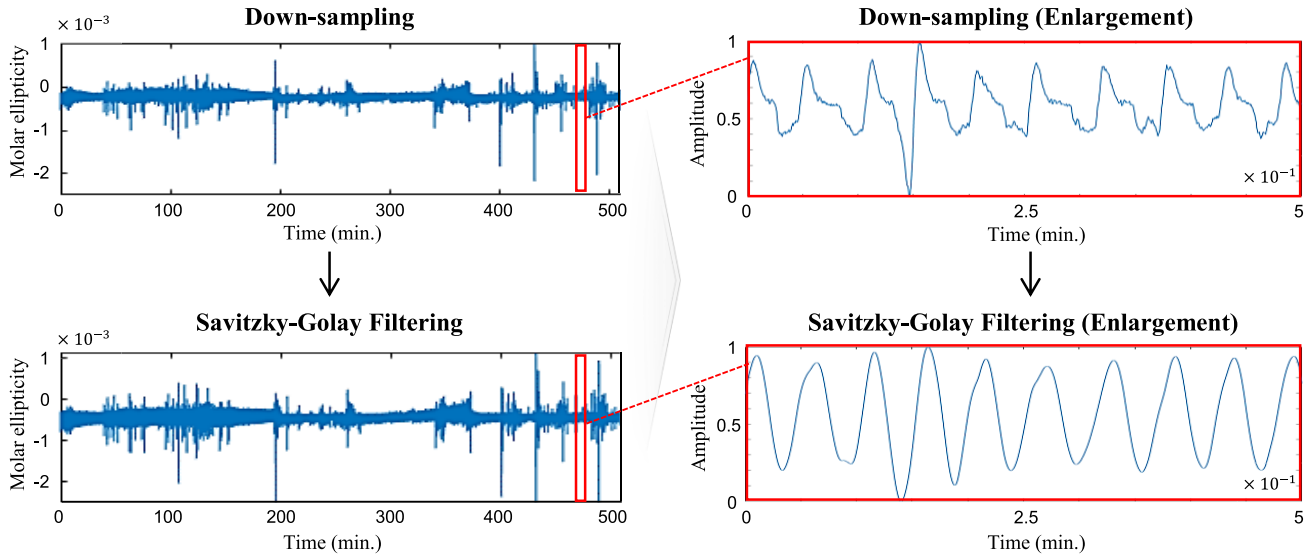

Fig. 2. Pre-processing phase for the nasal pressure signal. The downsampled data were computed to smooth the signal through the Savitzky-Golay filter. The figure on the left shows the signal at the entire sleep time of a representative subject, and the figure on the right shows the enlarged area corresponding to the red box in the figure on the left.

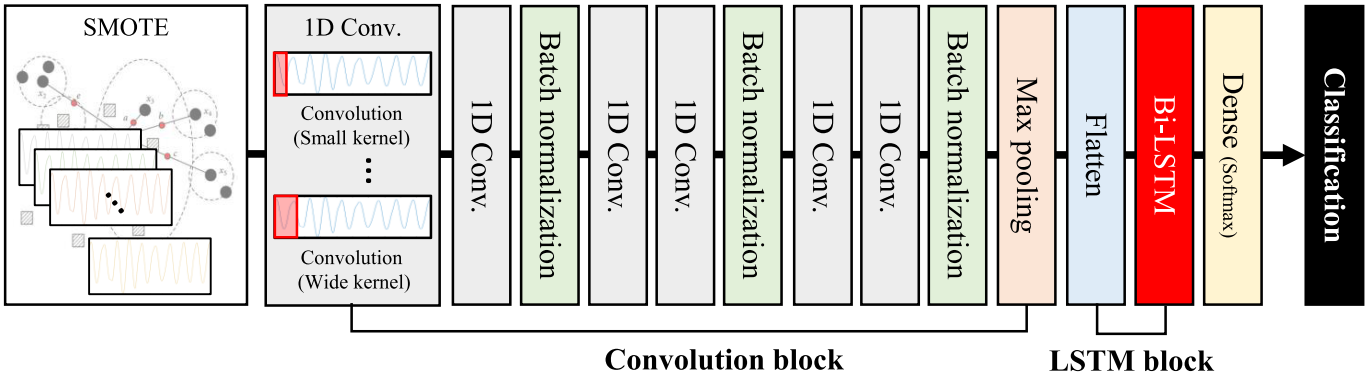

Fig. 3. The overall framework of the proposed multi-kernel convolutional biLSTM network. The proposed framework involves a synthetic minority over-sampling technique (SMOTE) to rectify the imbalanced class in the pre-processing step. The deep learning architecture comprises three main phases: three sets of convolution blocks for representation learning, an LSTM block for sequential learning, and a classification block for prediction. Conv. = convolutional layer.

This technique was selected due to its unbalanced class distribution, as a night's sleep is usually dominated by NREM sleep, as opposed to wake and REM sleep. The deep learning framework was deemed suitable to fully understand the data through SMOTE. SMOTE created synthetic samples for the minority class, thereby enhancing the model's ability to identify and categorize these instances. This involved computing the difference vector between a minority class sample and its nearest neighbor samples, followed by generating new data points by scaling the difference vector using a random factor. The x_0 denotes a candidate for integration as a minority class instance. $I_{B(x_0,r)}$ refers to the coverage of the minority class within a radius range r centered on x_0.

$$I_{B(x_0,r)} = \int_{B(x_0,r)} pX(x)dx \tag{2}$$

where $pX(x)$ denotes the original probability density of the minority class. The newly generated point z is obtained by adding a uniform random variable w multiplied by the vector difference between x_k (a neighboring point) and x_0.

$$\mathrm{z} = (1 - \mathrm{w})x_0 + \mathrm{w}x_k \tag{3}$$

The density function of point z is expressed as

$$\begin{aligned} &p_Z(\mathrm{z}) \\ &= (N-K)\binom{N-1}{K}\int_x p_X(x)\int_{r=\|\mathrm{z}-x\|}^{\infty} p_X\left(x+\frac{(z-x)r}{\|\mathrm{z}-x\|}\right) \\ &\times\left(\frac{r^{d-2}}{\|\mathrm{z}-x\|^{d-1}}\right) B\left(1-I_{B(x,r)};N-K-1,K\right)drdx \end{aligned} \tag{4}$$

where N and K represent the numbers of samples of the minority class and neighboring samples, respectively.

C. Proposed Model

The proposed multi-kernel convolutional biLSTM network architecture was designed to classify sleep stages using nasal signals, as illustrated in Fig. 3. The architecture comprised

three main phases: (i) three sets of convolutional blocks for representation learning, (ii) an LSTM block for sequential learning, and (iii) a classification block.

Each convolution block, which was responsible for capturing features, comprises two primary components: convolution and batch normalization. Convolution employed a 1D format, applying convolutions along the time dimension of the input data to capture the patterns and relationships within the sequence. This involved narrow and wide kernels (i.e., kernel sizes 30, 50, 60, 70, 80, and 100) that were utilized as the input sequence. The kernel multiplied element-wise with the corresponding segments of the input, and the results were aggregated to generate individual output values. This process was repeated throughout the input sequence, resulting in a reduced-length output sequence. The adoption of a multi-kernel approach in our sleep stage classification study is driven by its ability to enhance feature representation, capture temporal dynamics, provide robustness against signal variability, and improve model generalization, making it highly effective for analyzing the complex patterns inherent in nasal pressure signals.

The convolutional kernel size was set to 1×512 using a stride size of 1×1. The rectified linear unit (ReLU) served as the non-linear activation function during this convolution step and enhanced its ability to capture complex patterns in the input data. The ReLU activation function was as follows:

$$f(x) = \begin{cases} 0 & \text{for} \quad x < 0 \\ x & \text{for} \quad x \geq 0 \end{cases} \tag{5}$$

$$f(x) = max(0, x) \tag{6}$$

In addition, a max pooling technique was employed to select the highest value within a small window to represent a segment of a convolved sequence. The output of the convolutional and pooling layers was flattened to a 1D vector. Following the convolutional stages, an LSTM block featuring a biLSTM setup was incorporated. With a hidden unit size of 256, this block effectively captures temporal dependencies and complex patterns in sequential data. Particularly in the case of cyclical data, such as the sleep-wake cycle or the cycle across the sleep stages, such an element could be considered crucial. Subsequently, a fully connected (dense) layer learns the intricate associations between the various features extracted by the convolutional and LSTM layers.

A softmax layer was used in the final classification step. This layer transformed the outputs of the preceding layers into probabilities for different sleep stages. Through the softmax layer, the model generated a probability distribution across distinct classes, thus facilitating accurate categorization of the input data. The specific parameters of the proposed model are listed in Table II.

TABLE II
DETAILS OF PARAMETERS AND LAYER IMPLEMENTATIONS FOR THE PROPOSED MULTI-KERNEL CONVOLUTIONAL BILSTM NETWORK

Blocks	Layers	Parameters
Convolution block I	Convolution	Filter size: 1×512 Stride size: 1×1 Activation: ReLU
	Convolution	Filter size: 1×512 Stride size: 1×1 Activation: ReLU
	Batch Normalization	–
Convolution block II	Convolution	Filter size: 1×256 Stride size: 1×1 Activation: ReLU
	Convolution	Filter size: 1×256 Stride size: 1×1 Activation: ReLU
	Batch Normalization	–
Convolution block III	Convolution	Filter size: 1×128 Stride size: 1×1 Activation: ReLU
	Convolution	Filter size: 1×128 Stride size: 1×1 Activation: ReLU
	Batch Normalization	–
	Max pooling	–
LSTM block	Bidirectional LSTM	Hidden units: 256
Classification block	Softmax	–

TABLE III
DETAILS OF THE DATASET USED IN OUR EXPERIMENTS

Sleep Stage	**Wake**	**N1**	**N2**	**N3**	**REM**	**Total**
Sample Size	2,674	2,704	9,126	1,716	3,554	19,774
Proportion(%)	13.52	13.67	46.15	8.68	17.97	100

IV. EXPERIMENTS

A. Experimental Setup

Many studies that use a single biosignal typically classify sleep stages by reducing them to 3- or 4-class because the slower signal adaptation makes this more feasible [13], [46]. Therefore, we combined the sleep stages by dividing them into three and four stages. Table III details each sample in a 30-sec epoch according to the AASM manual. We adjusted it to 3-class (i.e., wake, NREM, and REM) or 4-class (i.e., wake, light sleep, deep sleep, and REM) to suit our experiments.

To evaluate the various models, we used LOO-CV as the evaluation method. This method can be viewed as a cross-validation of each subject as a transfer learning method to explore the generalizability of the model [27]. For example, we had 25 participants, so we used the training dataset for the remaining 24 subjects, excluding one subject from the test data set, and evaluated the data of the unseen subject who was not included in the training process [31]. For each iteration of the LOO-CV, we further split the training set (which includes data from 24 patients) into a smaller training subset and a

TABLE IV
SPECIFICATIONS OF HYPERPARAMETERS FOR COMPARATIVE MODELS

Model	Hyperparameter	Type	Search space
RF	*n_estimator*	Discrete	100
	min_sample_spilt	Discrete	2
	min_samples_leaf	Discrete	1
	criterion	Catergorical	Entropy
NBC	*alpha*	Discrete	1
	kernel	Catergorical	Gaussian
LDA	*solver*	Catergorical	Singular value decomposition
	kernel	Catergorical	Shrinkage
SVM	*C*	Discrete	1.2
	kernel	Catergorical	Radial basis function

validation subset. Specifically, we used a stratified split to ensure that the validation subset is representative of the overall sleep stage distribution. We randomly allocated 80% of the data for the training subset and 20% for validation subset within each leave-one-out iteration. This validation subset was used to monitor the model's performance during training and to tune the hyperparameters. The final model was evaluated in the test patient (left out during the LOO-CV iteration) to assess its generalizability. This process was repeated for each patient in the dataset to evaluation. SMOTE was applied only to training datasets.

B. Comparative Methods

We further experimented with the following four models to compare the sleep stage classification performance using the proposed model. The same training and test procedure was performed for the comparison model as used for the proposed model. The hyperparameter tuning was applied using the default setting provided by scikit-learn library (https://scikit-learn.org/). The selection of default parameters was based on an unbiased comparison of the baseline performance of each machine learning model against our proposed model. That means, by using the standard hyperparameter tuning (Table IV), we aimed to mitigate any potential biases that might arise from custom-tailored hyperparameter tuning.

1) Random Forest (RF): This type of ensemble learning method learns from several decision trees constructed during training [47]. A tree comprises a set of nodes and edges that form a hierarchical structure. Multiple input variables can be handled without erasing them, and the generalization performance is high through randomization. Due to this randomness, the trees have slightly different characteristics that improve generalization performance by making the predictions of each tree uncorrelated.

2) Naïve Bayes Classifier (NBC): This is a type of probability classifier that applies Bayes' theorem, which assumes the independence between classes. There is the advantage that the amount of training data to estimate the parameters required for classification is very small [48]. Although it is known for its simplicity, the Naïve Bayes can outperform other state-of-the-art classification methods in many applications such as automatic medical diagnosis [49].

3) Linear Discriminant Analysis (LDA): LDA classifies data by learning the data distribution and creating decision boundaries. It aims to find a straight line that can effectively distinguish between two classes after projecting them onto a particular axis. This straight line indicates that the centers of the multiple classes are far from each other and that their variance must be small after the projection [49]. Therefore, it reduces dimensions by projecting the input data set into a low-dimensional space.

4) Support Vector Machine (SVM): SVM classified the data by determining the optimal separation hyperplane from the training data [50]. It determines a hyperplane that maximizes the distance between the data points of each class closest to the hyperplane. SVM is capable of non-linear and linear classification. To achieve this, it is necessary to consider the given data as a high-dimensional feature space. In this study, a stochastic gradient descent algorithm was used for optimization.

C. Evaluation Metrics

For overall performance, three metrics used mainly in the classification of the sleep stage were adopted: accuracy (ACC), macro F1-score, and kappa value [18]. In particular, the F1-score refers to the harmonic mean of precision and recall [12] and is an essential performance metric for class imbalance datasets [31]. When true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) are provided for each class, the following three metrics can be calculated:

$$ACC = \frac{TP}{TP + FP + TN + FN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{10}$$

Kappa values were calculated as follows:

$$Kappa\ value = \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_0}{1 - p_e} \tag{11}$$

where p_0 and p_e denote the accuracy and probability of the chance rate, respectively.

Finally, classification performance was measured for each class. In this study, the F1-score was calculated as a class-wise performance metric.

V. EXPERIMENTAL RESULTS & DISCUSSION

A. Classification Performance

Fig. 4 presents a confusion matrix that delineates the outcomes of the proposed model. In Fig. 4(a), the classification accuracy for the 3-class configuration was 0.7040 (± 0.0550) in a data set comprising 25 subjects. In particular, the highest prediction rate was recorded for NREM, with a value of 0.7896 for TP. In contrast, for the wake, the TP value was lower (0.2838), indicating a pronounced confusion between the predictions for the wake and NREM.

Fig. 4(b) shows the classification performance for the 4-class stages, showing an accuracy of 0.6044 (± 0.0606). Among TP, the preponderance of predictive instances belonged to light sleep with a substantial likelihood of 0.6726. The REM

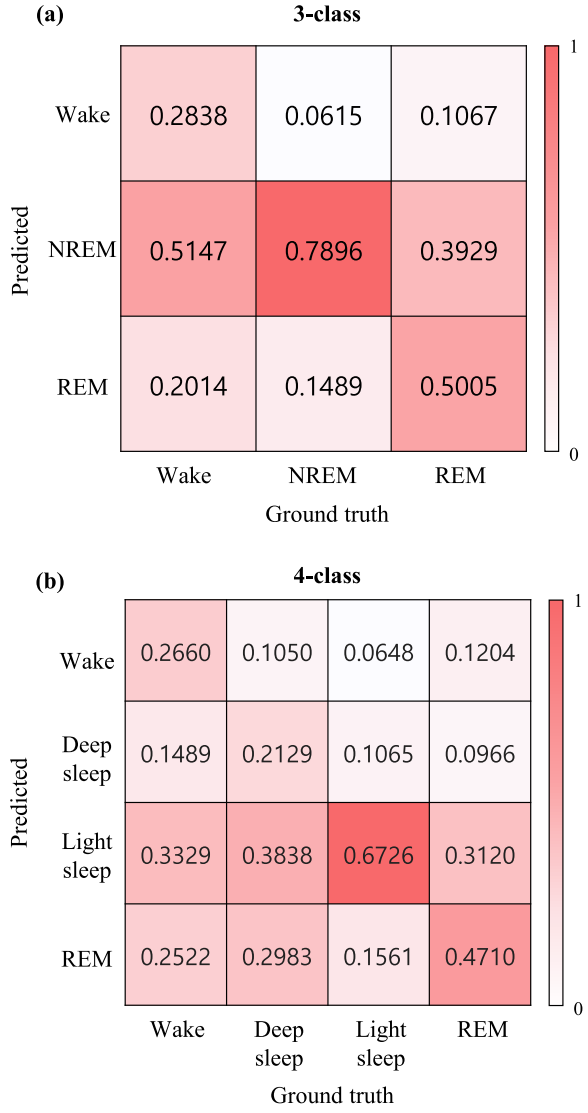

Fig. 4. Confusion matrix for the sleep stage classification. (a) Classification of the three-class sleep stages: wake, NREM, and REM. (b) Classification of the four-class sleep stages: wake, deep sleep, light sleep, and REM.

had a probability of 0.4710. On the contrary, the lowest probability quantified at 0.2129, was associated with the prediction of deep sleep. Regarding the probability of TN, the most confusing results were obtained when predicting light sleep, producing a confusion probability of 0.3838, specifically when deep sleep was the actual class. On average, the most accurate predictions were related to light and deep sleep conditions.

This evaluation underscores the potential efficacy of REM predictions, as well as the model's ability to accurately predict instances aligned with the light- and deep-sleep categories.

B. Representation of Sleep Stage Classification

A hypnogram is a graphical representation of the sleep stages through which an individual progresses during the night. In Fig. 5, time is represented on the horizontal axis for approximately 650 min, and the stages of sleep are indicated on the vertical axis. Different sleep stages were color-coded or labeled using different symbols. The hypnogram provides a visual summary of the amount of time an individual spends in each stage of sleep during the night, providing information on sleep quality and patterns using the proposed model. The pre-processed nasal signal for sleep time is shown in Fig. 5(a). In the 3- and 4-class classifications, NREM and light sleep exhibited the highest predicted probabilities (Fig. 5(b), (c)). Furthermore, when considering the 4-class classification, the highest probability of prediction was observed for light sleep and REM (Fig. 5(c)).

C. Evaluation for Comparative Methods and Proposed Model

Our proposed model exhibited superiority in both 3-class and 4-class sleep stage classifications compared with the four methods: RF, NBC, LDA, and SVM. Table V presents the complete comparison results for the 3-class sleep stage classifications. Our proposed model achieved an accuracy of 0.7040 (± 0.0550), a macro F1-score of 0.4904, and a kappa value of 0.2831, thus outperforming all other comparative models. A particularly remarkable result was observed for the class-wise F1-score, where the proposed model demonstrated the highest performance for the NREM, yielding an score of 0.8216. In contrast, the comparative models yielded the following results: RF achieved an accuracy of 0.3368 (± 0.0454), NBC achieved 0.5720 (± 0.1231), LDA achieved 0.3992 (± 0.0681), and SVM achieved 0.3560 (± 0.1809). These performances converged around chance-level accuracy (0.3342). SVM showed the lowest predicted classification performance with an accuracy of 0.3560, a macro F1-score of 0.2320, and a kappa value of 0.0149. SVM has a difficult time determining an optimal decision of the decision boundaries [51]. On the contrary, RF exhibited robust predictive performance for NREM, substantiated by a class-wise F1-score of 0.4236.

In Table VI, for the 4-class sleep stage classification, the proposed model achieved the highest performance, with an accuracy of 0.6044 (± 0.0606), a macro F1-score of 0.3496, and a kappa value of 0.2174. Among the class-wise F1-scores, the predictive performance for light sleep was good, attaining a score of 0.7616. The NBC recorded an accuracy of 0.5068 (± 0.1143) and macro F1-score and kappa values of 0.3292 and 0.1828, respectively, showing a high performance compared to other models. The class-wise F1-score for NBC was 0.6680, which means its strength in predicting light sleep. On the contrary, other comparative models showed predicted performances; for example, RF achieved an accuracy of 0.3564 (± 0.0774), LDA achieved 0.2908 (± 0.0494), and SVM recorded 0.2020 (± 0.0852), similar to chance-level accuracy.

In particular, in the 4-class sleep classification, the ratio of deep sleep (N3) is so low at 8.68%, that overall classification performance appears to be lower than that of the 3-class, which has a relatively good class ratio. However, in the proposed framework, SMOTE was used to solve the class imbalance problem, so performance improvement was particularly noticeable in the 4-class compared to other models.

In addition, even though both the class-wise F1-score and the macro F1-score are higher than the comparative models,

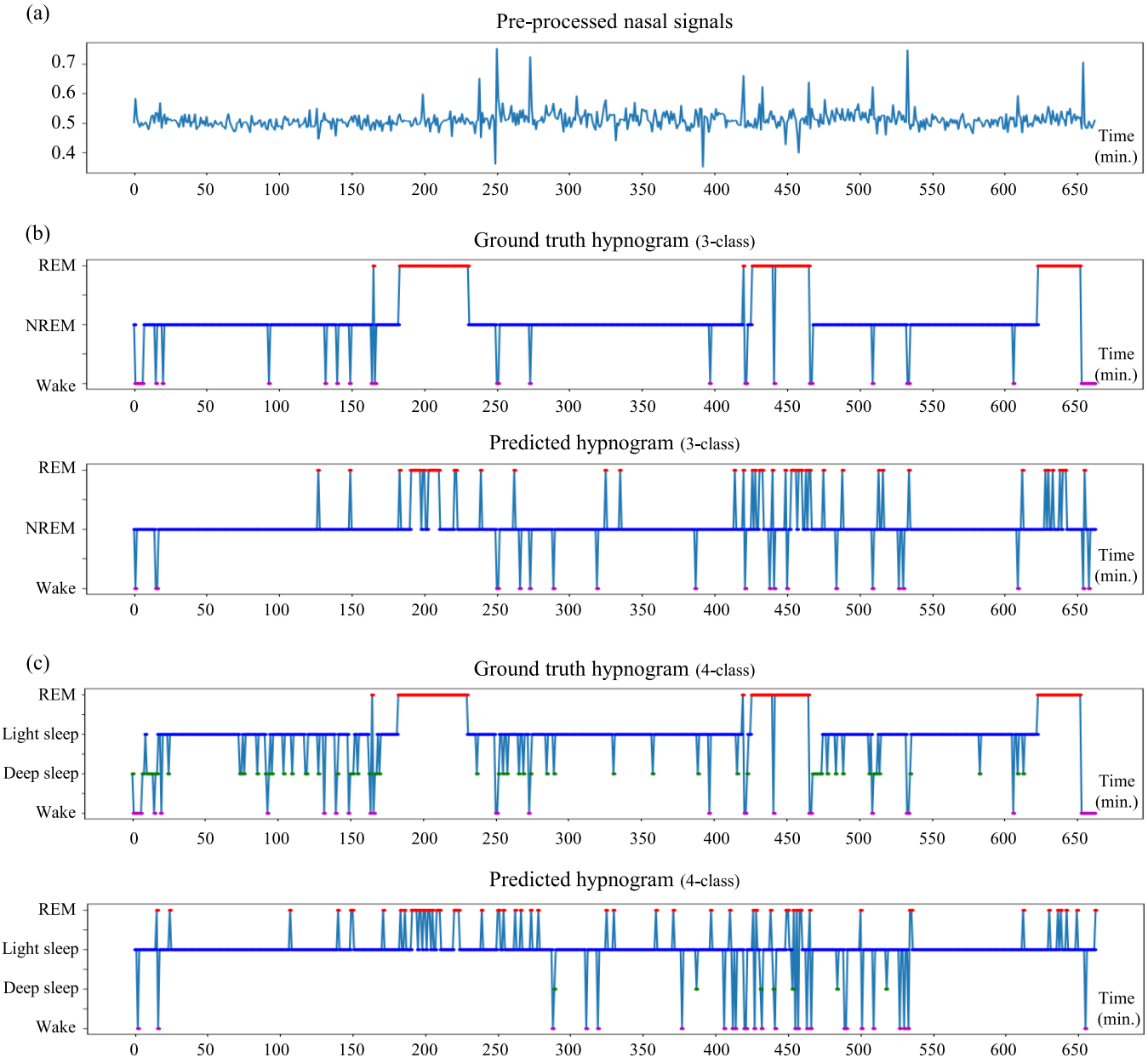

Fig. 5. Example of hypnogram according to inputs, labels, and predictions from a representative subject who has shown the highest performance. (a) The pre-processed nasal signals were normalized from the range of 0 to 1. (b) Predicted stages follow the ground truth for 3-class sleep stages (wake, NREM, and REM). The wake refers to purple, NREM refers to blue, and REM refers to red. (c) The predicted stages track the ground truth for 4-class sleep stages (wake, light sleep, deep sleep, and REM). Wake, light sleep, deep sleep, and REM refer to purple, blue, green, and red, respectively. The ground truth is visually indicated by an expert.

the F1-score seems very low in both the 3-class and 4-class sleep stage classification. In the 3-class sleep stage classification, when the proposed model was used, a low class-wise F1-score was shown at 0.0756 in wake and 0.3704 in REM, while a high performance was shown at 0.8216 in NREM. It is believed that wake and REM have a relatively low class ratio in the entire sleep stage, and both have movements in the body, so their performance is relatively lower than that of NREM. Likewise, in all other comparative models, the F1-score of NREM was higher than that of wake and REM. In the 4-class sleep stage classification, the proposed model showed relatively high performance with light sleep of 0.7616 and REM of 0.5724, and low performance with wake of 0.2432 and deep sleep of 0.1881. In the wake, it seems that misclassification appeared because it was relatively confused with REM where movement occurs and light sleep with a large number of data. In addition, deep sleep seems to have had difficulty learning because the number of data was 8.68%. In other comparative models, the F1-score of light sleep was also higher than in other stages of sleep. In particular, the proposed model showed a relatively higher F1-score of REM than other models. Therefore, based on the results of this classification, the macro F1-score, a performance indicator suitable for class imbalance, was 0.4904 in the 3-class sleep stage and 0.3496 in the 4-class sleep stage, which was somewhat lower than the accuracy.

TABLE V
CLASSIFICATION PERFORMANCE OF THREE-CLASS FOR SLEEP STAGE SCORING USING EVALUATION METRICS

Overall Metrics			
Methods	Accuracy	Macro F1-score	Kappa value
RF	0.3368 (±0.0454)	0.2624	0.1547
NBC	0.5720 (±0.1231)	0.4060	0.2174
LDA	0.3992 (±0.0681)	0.2992	0.0097
SVM	0.3560 (±0.1809)	0.2320	0.0149
Proposed.	0.7040 (±0.0550)	0.4904	0.2831
Class-wise Macro F1-score			
Methods	Wake	NREM	REM
RF	0.2608	0.4236	0.3356
NBC	0.2428	0.6948	0.2752
LDA	0.1632	0.5460	0.1888
SVM	0.1088	0.3868	0.2020
Proposed.	0.2756	0.8216	0.3704
Class-wise Sensitivity			
Methods	Wake	NREM	REM
RF	0.3064	0.8679	0.2886
NBC	0.4476	0.6632	0.2977
LDA	0.4250	0.4520	0.2054
SVM	0.2983	0.3452	0.4000
Proposed.	0.2838	0.7896	0.5005
Class-wise Specificity			
Methods	Wake	NREM	REM
RF	0.9263	0.4014	0.9238
NBC	0.8086	0.5749	0.8379
LDA	0.6659	0.5559	0.7898
SVM	0.7349	0.6678	0.6097
Proposed.	0.9321	0.5545	0.8452

TABLE VI
CLASSIFICATION PERFORMANCE OF FOUR-CLASS FOR SLEEP STAGE SCORING USING EVALUATION METRICS

Overall Metrics			
Methods	Accuracy	Macro F1-score	Kappa value
RF	0.3564 (±0.0774)	0.1256	0.0912
NBC	0.5068 (±0.1143)	0.3292	0.1828
LDA	0.2908 (±0.0494)	0.2180	0.0061
SVM	0.2020 (±0.0852)	0.1460	0.0033
Proposed.	0.6044 (±0.0606)	0.3496	0.2174

Class-wise Macro F1-score				
Methods	Wake	Deep sleep	Light sleep	REM
RF	0.1380	0.1576	0.2524	0.3520
NBC	0.2400	0.1840	0.6680	0.2244
LDA	0.1640	0.1008	0.4236	0.1864
SVM	0.0928	0.1400	0.2164	0.1368
Proposed.	0.2432	0.1036	0.7616	0.2884
Class-wise Sensitivity				
Methods	Wake	Deep sleep	Light sleep	REM
RF	0.3416	0.1654	0.7961	0.3149
NBC	0.3825	0.2098	0.6765	0.2209
LDA	0.4265	0.1244	0.3348	0.1959
SVM	0.2708	0.4001	0.1618	0.1829
Proposed.	0.2876	0.0792	0.8636	0.2460
Class-wise Specificity				
Methods	Wake	Deep sleep	Light sleep	REM
RF	0.9149	0.9100	0.5326	0.9136
NBC	0.8549	0.8916	0.5979	0.8712
LDA	0.6788	0.8633	0.6698	0.7949
SVM	0.7486	0.5955	0.8440	0.8173
Proposed.	0.9032	0.8826	0.6571	0.7644

D. Limitations & Future Works

Although this study shows promising results, it has several limitations. First, the 5-class sleep stage classification was not performed, which would have corresponded to the AASM guidelines. We focused on 4-class sleep stage classification similar to other studies using a single biosignal because it was based on scenarios for digital healthcare applications using simple equipment, such as wearable devices.

Second, the performance of our proposed model was decent, but the number of subjects is relatively small, limiting generalizability. In addition, since this experimental performance is somewhat inferior to other studies using a single biosignal, the difference in the number of subjects is believed to be a major factor. For example, 561 subjects were used in the study by Sridhar et al. [25], which achieved 77% accuracy in the 4-class sleep stage using a single ECG signal. However, the performance of the proposed model with limited training data highlights its potential. In addition, we conducted the experiment heuristically considering the more complex deep learning scenario. For the first time, when we designed our model initially, the convolution block and the LSTM block were designed to be shorter and deeper. However, it was observed that the classification performance was rather saturated around average classification performance (± 5%). We could not expect a dramatic performance improvement even if we stacked the layers more deeply. Rather, it showed similar performance, and more computational cost occurred. Furthermore, after the entire LOO-CV training, the proposed method took about 22 minutes on average to train, but other models took more than average 46 minutes. In particular, as the LSTM layer increased, more training time was required. Therefore, in this study, we reported the most appropriate architectural configuration as the proposed model. In future work, after collecting the large amount of data, we will adopt AutoML techniques [52] to obtain an optimized deep learning architecture for classification.

Third, we focused on sleep stage classification, not sleep apnea, even though the nasal pressure signals may be more favorable for the diagnosis of sleep apnea. The nasal pressure signals can perform well in apnea events and have historically shown excellent ability to detect apnea events [23].

However, our data was measured to confirm the possibility of nasal pressure in the sleep stage classification in healthy individuals, and unfortunately, the apnea-hypopnea index for the diagnosis of sleep apnea was not measured. In other words, the method was used in healthy participants to detect the sleep stage. Future work should test this method on patients with apnea (as this might be more difficult to assess sleep stage in the presence of apnea) and in doing so, we could integrate an apnea detection module that operates on the same signal in our model (to counteract the potential "artifact" of apnea). In this regard, the proposed framework itself could be applicable to the diagnosis of sleep apnea.

Furthermore, we will develop an advanced model to conduct a further evaluation on a test-bed that can also distinguish classes for sleep classification up to N1 and N2 for patients and individuals with sleep disorders.

VI. Conclusion

In this study, a deep learning model was proposed to classify sleep stages based on nasal pressure. After preprocessing the nasal pressure signal, SMOTE was used to resolve the class imbalance of the sleep stage classification task. In addition, a multi-kernel technique was used to extract sleep-stage features from the pre-processed data through convolution blocks, which were predicted through an LSTM block. Our results indicate that our model achieved superior performance compared to existing baseline models and is meaningful in that it can reduce sleeper inconvenience by wearing the low impact measuring equipment used to record nasal pressure. The well-documented rise in sleep disorders [4], [5] requires matching technology that can match the need for diagnosis and, potentially, help with treatment. In particular, devices with a low burden such as those used here could allow for easy training, home-environment use, and even longitudinal tracking of disease and treatment progression. Furthermore, the current device offers a unique opportunity to be used in unity with some interventions, which could both track and facilitate sleep through nasal equipment. In the future, testing the proposed methodology in clinical contexts may broaden its applicability even further, where in particular respiratory issues could be easily and sufficiently captured.

References

[1] S. Sarasso et al., "Local sleep-like cortical reactivity in the awake brain after focal injury," *Brain*, vol. 143, no. 12, pp. 3672–3684, Dec. 2020.

[2] L. A. Panossian and A. Y. Avidan, "Review of sleep disorders," *Med. Clin. N. Am.*, vol. 93, no. 2, pp. 407–425, 2009.

[3] A. Jegou et al., "Cortical reactivations during sleep spindles following declarative learning," *NeuroImage*, vol. 195, pp. 104–112, Jul. 2019.

[4] B. He, B. Baxter, B. J. Edelman, C. C. Cline, and W. W. Ye, "Noninvasive brain–computer interfaces based on sensorimotor rhythms," *Proc. IEEE*, vol. 103, no. 6, pp. 907–925, Jun. 2015.

[5] M. Darracq et al., "Evoked alpha power is reduced in disconnected consciousness during sleep and anesthesia," *Sci. Rep.*, vol. 8, no. 1, p. 16664, Nov. 2018.

[6] M. M. Grigg-Damberger, "The AASM scoring manual: A critical appraisal," *Current Opinion Pulmonary Med.*, vol. 15, no. 6, pp. 540–549, 2009.

[7] 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, pp. 1–13, Oct. 2018.

[8] M. Fu et al., "Deep learning in automatic sleep staging with a single channel electroencephalography," *Frontiers Physiol.*, vol. 12, Mar. 2021, Art. no. 628502.

[9] J. Krieger, "Respiratory physiology: Breathing in normal subjects," in *Principles and Practice of Sleep Medicine*. Philadelphia, PA, USA: Saunders, 2005, pp. 232–244.

[10] A. Kimura et al., "Phase of nasal cycle during sleep tends to be associated with sleep stage," *Laryngoscope*, vol. 123, no. 8, pp. 2050–2055, Aug. 2013.

[11] M. Varis, T. Karhu, T. Leppänen, and S. Nikkonen, "Utilizing envelope analysis of a nasal pressure signal for sleep apnea severity estimation," *Diagnostics*, vol. 13, no. 10, p. 1776, May 2023.

[12] M. Lee, H.-G. Kwak, H.-J. Kim, D.-O. Won, and S.-W. Lee, "SeriesSleepNet: An EEG time series model with partial data augmentation for automatic sleep stage scoring," *Frontiers Physiol.*, vol. 14, Aug. 2023, Art. no. 1188678.

[13] M. Radha et al., "A deep transfer learning approach for wearable sleep stage classification with photoplethysmography," *npj Digit. Med.*, vol. 4, no. 1, p. 135, Sep. 2021.

[14] U. Erdenebayar, Y. Kim, J.-U. Park, S. Lee, and K.-J. Lee, "Automatic classification of sleep stage from an ECG signal using a gated-recurrent unit," *Int. J. FUZZY Log. Intell. Syst.*, vol. 20, no. 3, pp. 181–187, Sep. 2020.

[15] M. Radha et al., "Sleep stage classification from heart-rate variability using long short-term memory neural networks," *Sci. Rep.*, vol. 9, no. 1, p. 14149, Oct. 2019.

[16] M. Olsen et al., "A flexible deep learning architecture for temporal sleep stage classification using accelerometry and photoplethysmography," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 1, pp. 228–237, Jan. 2023.

[17] P. Memar and F. Faradji, "A novel multi-class EEG-based sleep stage classification system," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 1, pp. 84–95, Jan. 2018.

[18] L. Fiorillo, P. Favaro, and F. D. Faraci, "DeepSleepNet-lite: A simplified automatic sleep stage scoring model with uncertainty estimates," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 2076–2085, 2021.

[19] N. Goshtasbi, R. Boostani, and S. Sanei, "SleepFCN: A fully convolutional deep learning framework for sleep stage classification using single-channel electroencephalograms," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 2088–2096, 2022.

[20] J. Jeong, K. Shim, D. Kim, and S. Lee, "Brain-controlled robotic arm system based on multi-directional CNN-BiLSTM network using EEG signals," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 5, pp. 1226–1238, May 2020.

[21] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 3, pp. 400–410, Mar. 2019.

[22] S. J. Redmond and C. Heneghan, "Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 3, pp. 485–496, Mar. 2006.

[23] H. ElMoaqet, M. Eid, M. Glos, M. Ryalat, and T. Penzel, "Deep recurrent neural networks for automatic detection of sleep apnea from single channel respiration signals," *Sensors*, vol. 20, no. 18, p. 5037, Sep. 2020.

[24] Y. K. Kim, M. Lee, H. S. Song, and S. Lee, "Automatic cardiac arrhythmia classification using residual network combined with long short-term memory," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–17, 2022, doi: 10.1109/TIM.2022.3181276.

[25] N. Sridhar et al., "Deep learning for automated sleep staging using instantaneous heart rate," *Npj Digit. Med.*, vol. 3, no. 1, p. 106, Aug. 2020.

[26] Q. Li et al., "Transfer learning from ECG to PPG for improved sleep staging from wrist-worn wearables," *Physiological Meas.*, vol. 42, no. 4, Apr. 2021, Art. no. 044004.

[27] F. Fahimi, Z. Zhang, W. B. Goh, T.-S. Lee, K. K. Ang, and C. Guan, "Inter-subject transfer learning with an end-to-end deep convolutional neural network for EEG-based BCI," *J. Neural Eng.*, vol. 16, no. 2, Apr. 2019, Art. no. 026007.

[28] H. Phan, O. Y. Chen, M. C. Tran, P. Koch, A. Mertins, and M. De Vos, "XSleepNet: Multi-view sequential model for automatic sleep staging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5903–5915, Sep. 2021.

[29] M. Lee et al., "Quantifying arousal and awareness in altered states of consciousness using interpretable deep learning," *Nature Commun.*, vol. 13, no. 1, pp. 1–14, Feb. 2022.

[30] A. Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 11, pp. 1998–2008, Nov. 2017.

[31] E. Eldele et al., "An attention-based deep learning approach for sleep stage classification with single-channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 809–818, 2021.

[32] P. Tripathi et al., "Ensemble computational intelligent for insomnia sleep stage detection via the sleep ECG signal," *IEEE Access*, vol. 10, pp. 108710–108721, 2022.

[33] T. Nakamura, Y. D. Alqurashi, M. J. Morrell, and D. P. Mandic, "Hearables: Automatic overnight sleep monitoring with standardized in-ear EEG sensor," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 1, pp. 203–212, Jan. 2020.

[34] N. Kaongoen et al., "The future of wearable EEG: A review of EAR-EEG technology and its applications," *J. Neural Eng.*, vol. 1, no. 1, pp. 1–15, Jun. 2023.

[35] J. Fan, C. Sun, M. Long, C. Chen, and W. Chen, "EOGNET: A novel deep learning model for sleep stage classification based on single-channel EOG signal," *Frontiers Neurosci.*, vol. 15, Jul. 2021, Art. no. 573194.

[36] S. Guo, X. Zhao, K. Matsuo, J. Liu, and T. Mukai, "Unconstrained detection of the respiratory motions of chest and abdomen in different lying positions using a flexible tactile sensor array," *IEEE Sensors J.*, vol. 19, no. 21, pp. 10067–10076, Nov. 2019.

[37] H. ElMoaqet, M. Eid, M. Ryalat, and T. Penzel, "A deep transfer learning framework for sleep stage classification with single-channel EEG signals," *Sensors*, vol. 22, no. 22, p. 8826, Nov. 2022.

[38] X. Zhou, B. W.-K. Ling, W. Ahmed, Y. Zhou, Y. Lin, and H. Zhang, "Multivariate phase space reconstruction and Riemannian manifold for sleep stage classification," *Biomed. Signal Process. Control.*, vol. 88, Jul. 2024, Art. no. 105572.

[39] S. Mousavi, F. Afghah, and U. R. Acharya, "SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach," *PLoS One*, vol. 14, no. 5, 2019, Art. no. e0216456.

[40] J. Pradeepkumar et al., "Towards interpretable sleep stage classification using cross-modal transformers," 2022, *arXiv:2208.06991*.

[41] Z. He et al., "Cross-scenario automatic sleep stage classification using transfer learning and single-channel EEG," *Biomed. Signal Process. Control.*, vol. 81, Jul. 2023, Art. no. 104501.

[42] P. Marchand and L. Marmet, "Binomial smoothing filter: A way to avoid some pitfalls of least-squares polynomial smoothing," *Rev. Sci. Instrum.*, vol. 54, no. 8, pp. 1034–1041, 1983.

[43] P. Gans and J. B. Gill, "Examination of the convolution method for numerical smoothing and differentiation of spectroscopic data in theory and in practice," *Appl. Spectrosc.*, vol. 37, no. 6, pp. 515–520, Nov. 1983.

[44] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[45] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Mach. Learn.*, vol. 113, no. 7, pp. 4903–4923, Jul. 2024.

[46] O. K. Utomo, N. Surantha, S. M. Isa, and B. Soewito, "Automatic sleep stage classification using weighted ELM and PSO on imbalanced data from single lead ECG," *Proc. Comput. Sci.*, vol. 157, pp. 321–328, Jan. 2019.

[47] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.

[48] N. Y. Oktavia, A. D. Wibawa, E. S. Pane, and M. H. Purnomo, "Human emotion classification based on EEG signals using naive Bayes method," *Int. Seminar Appl. Technol. Inf. Commun.*, vol. 1, pp. 319–324, Jul. 2019.

[49] M. Lee, J.-H. Jeong, Y.-H. Kim, and S.-W. Lee, "Decoding finger tapping with the affected hand in chronic stroke patients during motor imagery and execution," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1099–1109, 2021.

[50] S.-H. Lee, M. Lee, and S.-W. Lee, "Neural decoding of imagined speech and visual imagery as intuitive paradigms for BCI communication," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 12, pp. 2647–2659, Dec. 2020.

[51] I. Steinwart, D. Hush, and C. Scovel, "Training svms without offset," *J. Mach. Learn. Res.*, vol. 12, no. 1, pp. 1–26, Aug. 2011.

[52] D. Aquino-Brítez et al., "Optimization of deep architectures for EEG signal classification: An automl approach using evolutionary algorithms," *Sensors*, vol. 21, no. 6, p. 2096, 2021.