IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 27 March 2023, accepted 17 April 2023, date of publication 2 May 2023, date of current version 9 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3272373

RESEARCH ARTICLE

A Deep Neural Network Based Wake-After-Sleep-Onset Time Aware Sleep Apnea Severity Estimation Scheme Using Single-Lead ECG Data

DAE-WOONG SEO[1], JEEYOUNG KIM[2], (Member, IEEE), HO-WON LEE[3,4], AND YOUNG-KYOON SUH[1], (Member, IEEE)

[1]School of Computer Science and Engineering, Kyungpook National University, Daegu 41566, Republic of Korea
[2]Graduate School of Data Science, Kyungpook National University, Daegu 41566, Republic of Korea
[3]Department of Neurology, School of Medicine, Kyungpook National University, Daegu 41404, Republic of Korea
[4]Brain Science and Engineering Institute, Kyungpook National University, Daegu 41404, Republic of Korea

Corresponding authors: Ho-Won Lee (neuromd@knu.ac.kr) and Young-Kyoon Suh (yksuh@knu.ac.kr)

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2018R1A6A1A03025109.

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 Kyungpook National University Chilgok Hospital, and performed in line with the Declaration of Helsinki.

ABSTRACT Obstructive sleep apnea (OSA) is a prevalent yet potentially severe sleep disorder. Polysomnography (PSG) is most commonly used to assess the severity of OSA. However, there have been numerous studies to find OSA patients more effectively since running a PSG test is expensive and time-consuming. The existing studies, however, raise four major concerns, such as (i) the use of inaccurate sleep time data to calculate the apnea-hypopnea index, (ii) the use of poor preprocessing techniques for real patient clinical datasets, (iii) the lack of multi-stage classification capability, and (iv) the absence of experiments on sufficiently large data sets. To address these concerns, we propose a novel OSA severity classification scheme based on single-lead electrocardiogram (ECG) data, as well as a novel deep learning model, *CLNet*, to perform apnea/hypopnea and sleep stage classification. By identifying apnea/hypopnea events from a patient's ECG data and computing AHI using ''pure'' sleep duration via CLNet, our method improves patient OSA severity degree estimation. CLNet was trained and evaluated using two different real-world datasets containing 286 OSA patient records and a total of 2,155 hours of ECG data. In our experiments, the proposed scheme outperforms existing approaches by up to 10% in total accuracy and AUC on the public PhysioNet dataset. In terms of apnea classification sensitivity, we show that the proposed CLNet model outperforms the state-of-the-art model by up to 41.8% for our clinical dataset. Our scheme can be used as a successful, high-quality pre-screening tool by more effectively prioritizing prospective OSA patients. We will be able to perform PSG on only the most severe patients, saving both time and money. Our algorithms are publicly available on GitHub.

INDEX TERMS Apnea-hypopnea index, classification, deep learning, electrocardiogram, polysomnography, sleep apnea severity classification.

I. INTRODUCTION

Obstructive Sleep Apnea (OSA) is a significant sleep disorder that can cause various complications, including a high risk of cardiovascular diseases such as heart attacks, high blood

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Kafiul Islam.

pressure, or stroke, to name a few. Recently, patients with OSA are rapidly increasing due to widespread obesity [1], [2] and the aging population [3].

Polysomnography (PSG) is an overnight multi-parameter sleep study measuring body functions including brain waves, skeletal muscle activity, blood oxygen levels (SpO_2), heart rate, breathing rate, and eye movement. The severity of OSA is diagnosed by clinical experts using PSG examinations.

Despite being the most reliable OSA severity measurement tool, PSG has a critical drawback. While patients stay at an overnight sleep facility with multiple sensors attached to their bodies, clinical experts must monitor the PSG sensor signals capturing the patient's sleep overnight. Hence, a PSG exam can be costly, cumbersome, and time-consuming [4].

To overcome this limitation, there has been a rich body of existing literature that seeks an alternative to PSG by using easily measurable single-lead bio-signal data such as electrocardiogram (ECG) signals [5], [6], [7], snoring [8], or oxygen saturation [9], [10] (a.k.a SpO_2), referring to the percentage of oxygen in one's blood. By pre-screening the degree of OSA severity before the actual PSG examination is taken, can help prioritize potential patients.

One of the most important issues in automatically determining the degree of OSA severity via single-lead bio-signal data is to *accurately* calculate the apnea-hypopnea index (AHI). Existing studies [11], [12] have two critical problems in calculating the AHI. First, given that ECG signals are divided into 30-second or 1-minute epochs for counting apnea/hypopnea (A/H) events, there may be some disparity in calculating the exact number of A/H events. For example, when modeling the widely used PhysioNet Apnea-ECG data [13], A/H events are detected only when the beginning of the epoch has an abnormal respiratory signal (as shown in Figure 2(a)).

It is essential to conduct accurate labeling on the dataset because there may be a significant discrepancy between the A/H count annotated by the actual expert and the A/H count of the labeled dataset. Second, the exact sleep time is seldom reflected when AHI is calculated. The total sleep time recorded in PSG includes ''wake-after-sleep-onset,'' so it is not equal to ''pure'' sleep time [4]. If the wake time is *not* excluded from the total recorded sleep time, the accuracy of the AHI calculation may substantially drop, leading to significantly underestimating OSA severity.

To address these concerns, we propose a novel OSA severity classification scheme leveraging a deep learning model we devise for respiratory status and sleep stage classification. First, we explore a method of accurately detecting an A/H event per 1-minute epoch. Apnea and hypopnea are defined by the American Academy of Sleep Medicine as the cessation of airflow for at least 10 seconds consecutively [14]. The only difference between the two is that apnea is a complete pause in breathing where airflow is cut down by more than 90%, whereas hypopnea is a period of shallow breathing where the airflow is reduced by 30% or more and SpO_2 shows a decrease of 3% or more [4]. We follow this definition to label our data. Specifically, we define an apnea event when there is a period of over 10 seconds of continuous abnormal respiratory events, including both apnea and hypopnea. Second, we seek another method of deriving pure sleep time by excluding intermediate wake time from the total recorded sleep time via the sleep state classification model. We then combine these methods to form the model, termed *CLNet*, introduced in this paper. The CLNet is a novel deep-learning-based model that estimates OSA severity by dividing the number of identified A/H events by the extracted pure sleep time in hours. Unlike previous studies [11], [15], [16], our proposed scheme is the first to identify and eliminate wakeup time from recorded sleep time for a more accurate OSA classification. It employs new labeling techniques and the classification model.

There are several practical *impacts* that we anticipate from our model. The first is that clinical experts can see the degree of estimated OSA without fully scanning complicated PSG signals. If the pre-screened results are *Severe*, the experts can choose to proceed with the actual PSG test for a fuller examination. Another is that a potential patient may wear a simple ECG device to check their OSA status before deciding whether to undergo the more reliable PSG test based on their pre-screened OSA severity level. By doing so, the patient may save a considerable fee for an actual PSG examination as well. We expect our model to evolve into a useful OSA pre-screening tool, by which we can prioritize OSA patients under limited PSG resources.

Our contributions are briefly summarized below.

- We propose a novel OSA severity classification scheme using deep learning models trained with actual datasets, including rigorous ablation studies, to contribute to establishing and optimizing our classification model.
- We present a novel labeling technique using the closest definition to OSA. Specifically, our technique, termed C-20, shows a mere 2.3% MAPE compared to the human-expert annotation method. This results in outperforming the existing labeling method by reducing the number of measurement errors up to 14×.
- We resolve the data imbalance problem, which is the most challenging in this OSA severity estimation research, by populating more of the underrepresented Apnea and Wake data into the training set and applying the Synthetic Minority Oversampling Technique (SMOTE).
- We introduce a new technique to derive pure sleep time by deducting the nontrivial wake-after-sleep-onset time, which contributes to overcoming underestimating the degree of OSA severity.
- We demonstrate in our experiments using actual clinical datasets that the proposed CLNet scheme outperforms the state-of-the-art work by up to 41.8% in sensitivity and 4.4% in accuracy. In particular, we achieve a 100% recall rate in classifying severe OSA patients.

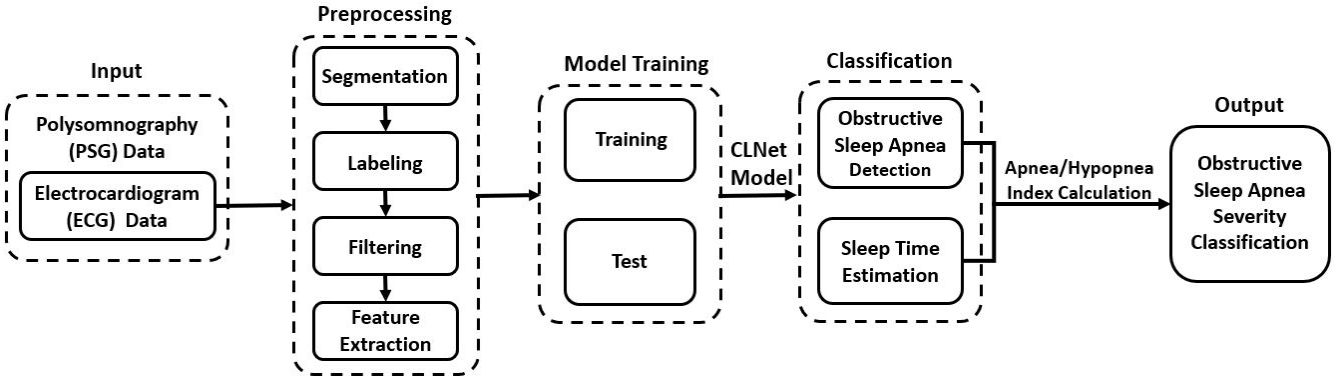


FIGURE 1. The overall process of our OSA severity classification using single-lead ECG signal data.

The rest of this paper is organized as follows. In the following section, we propose our overall methodology for estimating OSA severity. Then, we present the performance evaluation results of the proposed model on two different datasets and discuss significant implications. Subsequently, we review the existing literature. Finally, we conclude our work and suggest future research directions in Section V.

II. PROPOSED METHODOLOGY

In this section, we elaborate on our OSA severity estimation method using deep learning techniques as depicted in Figure 1. The datasets we use are a private clinical dataset collected from Kyungpook National University Chilgok Hospital (hereafter, KNUCH Dataset) and the PhysioNet Apnea-ECG database [13], [17], [18] (hereafter, PhysioNet Dataset). Informed consent was obtained from all subjects involved in the study. The KNUCH Dataset is provided in a full PSG data format, which is a mixture of several different signals in varying sampling frequencies, such as SpO_2 (100Hz), ECG (250Hz) and snoring (250Hz). First, we extract the raw ECG signal which comes in varying sampling frequency depending on the PSG machine in which the data was collected. We have access to 250Hz and 1024Hz ECG signal data from the KNUCH Dataset. We then perform data segmentation and labeling. The PhysioNet Dataset is pre-labeled by the authors of the dataset, hence we have no control over the segmentation and labeling of this data and use the data in its given labeled format. The filtering process is used for the labeled KNUCH Dataset and the pre-labeled PhysioNet Dataset. We then extract important features derived from the ECG signal analysis. Below, we explain each step in detail beginning with preprocessing.

A. PREPROCESSING

This section describes our preprocessing method including segmentation, labeling, filtering, and feature extraction.

1) SEGMENTATION AND LABELING

The PhysioNet Dataset is the most commonly used dataset in sleep apnea detection [6], [7], [11], [15], [16]. It is segmented into 60-second epochs to provide ECG signal data where each epoch has an annotation of Apnea or Non-apnea. Following the example of this well-known dataset, sleep apnea detection studies typically use 60-second long epochs. In the case of determining sleep stage, due to the fact that sleep experts normally annotate sleep stages every 30 seconds, we also segment the data to determine sleep into 30-second epochs. For the KNUCH Dataset, segmentation is performed using 30-second epochs when classifying sleep stage, and 1-minute epochs for sleep apnea classification. The PhysioNet Dataset does not provide sleep stage labels, but only recorded sleep time.

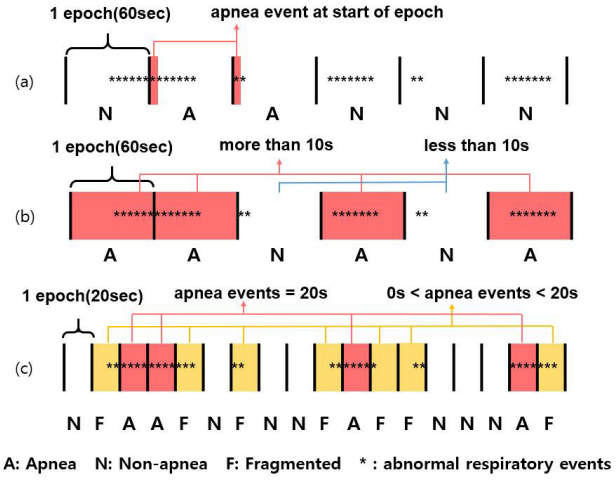


FIGURE 2. Example of ECG data snippet with three different labeling methods applied: (a) P-60 (existing), (b) C-60 (proposed), and (c) C-20 (proposed).

The PhysioNet Dataset is sampled at 100 Hz and has 6,000 sample points within a 1-minute epoch. Among them, the respiratory event of the first data point is designated as the label of the corresponding epoch as shown in Figure 2(a). (This labeling technique is termed P-60.) This is contrary to the medical definition of an apnea event. According to the medical definition, an apnea event occurs when abnormal breathing persists for more than 10 seconds [4].

This discrepancy leads to many errors when comparing the A/H count of the P-60 label to the real-world data, thus making it difficult to measure the exact A/H count in sleep apnea detection.

To address these concerns, we propose two methods of OSA data labeling in this paper. For both methods, we use the expert annotated data as input. The first is C-60 labeling technique. If breathing stops for more than 10 consecutive seconds within a 60-second epoch, the epoch is labeled as '*A*' (Apnea). Everything else is labeled '*N*' (Non-apnea). The C-60 label is a 2-class configuration, consisting of *A* and *N*. In this labeling method, the A/H count is defined by the number of epochs labeled *A*. Therefore, the total number of *A* becomes the actual A/H count as seen in Figure 2(b).

The second labeling technique, C-20 is shown in Algorithm 1. The C-20 label is a *3-class* ('*N*', '*A*', and '*F*') configuration. This labeling method uses 20-second epochs for labeling (Lines 2 and 5-6). If all the data points in the 20-second epoch are *non-apnea*, the epoch is labeled *N*, if all the data points are *apnea*, it is labeled *A* (Lines 7-8). In all other cases where there exists a mixture of non-apnea breathing and apnea data points, thus fragmenting the 20-second epoch, we label '*F*' for fragmented (Line 10). (Refer to Figure 2(c).)

The main idea behind the method of counting OSA events for both C-20 and C-60 is that we count the number of epochs that are not labeled as *N*. However, in C-20, we have an additional label termed *F*, resulting in a slight divergence in the counting method as shown in Algorithm 2. To count the A/H events using this label, we sequentially check the labels until the label is *A* or *F* (Line 5). The A/H count is incremented, *flag* is set to 1, and we continue to look at the next label (Lines 6-7). If the next label turns out to be *A*, we keep on with our search without changing the count or flag since it is still the same Apnea event. If the next label is *F* or *N*, the *flag* is set to 0 (Line 9) since the Apnea has ceased. We continue looking through the list until we run into another label that is *A* or *F*.

Algorithm 1 C-20 Labeling Method

```
Input  : The Expert Annotated List (Label) and
         Sampling Frequency (Hz)
Output: The Labeled List
1  Function Labeling (Label, Hz) :
2      e ← Hz * 20; // e: Epoch
3      l ← len(Label) - mod(len(Label), Hz*60);
       // l: label length
4      L ← an empty list;
5      for i = 0 to l, i = i + e do
6          Find x ∈ [i, i+e) for Label.
7          if x is unique then
8              L[i] ← x; // x: A or N
9          else
10             L[i] ← F; // F: Fragmented Label
11         end
12     end
13     return L;
```

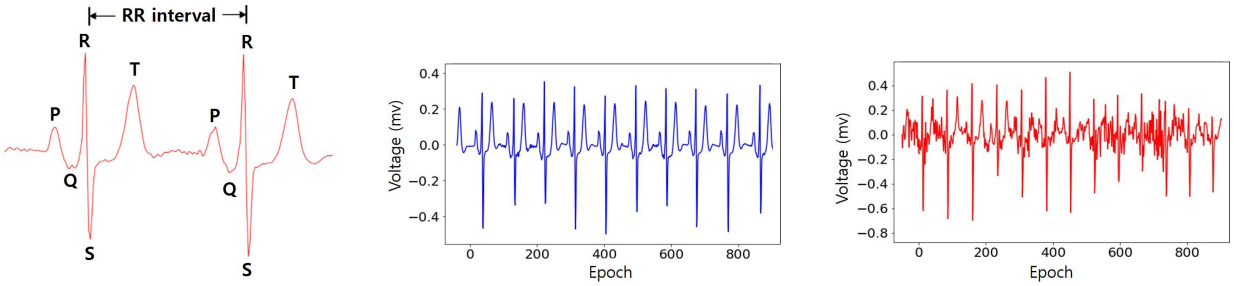
Algorithm 2 C-20 OSA Event Counting Method

```
Input  : The Labeled List (LabeledList)
Output: Apnea/Hypopnea Counts (count)
1  Function Counting (LabeledList ) :
2      count ← 0;
3      flag ← 0;
4      foreach x ∈ LabeledList do
5          if x ≠ N and flag = 0 then
6              count ← count + 1;
7              flag ← 1 ;
8          else
9              if x ≠ A then flag ← 0 ;
10         end
11     end
12     return count;
```

2) FILTERING AND FEATURE EXTRACTION

In the ECG signal, there are a series of waveforms P, Q, R, S, and T and various intervals such as S-T, Q-T, P-R, and R-R (RR) as illustrated in Figure 3(a). Among them, the RR-interval and R-peak amplitude contain significant information about OSA detection operations [8], [15], [19], as can be seen in Figures 3(b) and 3(c). Our ECG signal preprocessing sequence borrows from that of Wang et al. [15]. To remove noise from the ECG signal, we go through a filtering process. The Finite Impulse Response filtering methods built into BioSPPy [20] and the median filtering methods from SciPy [21] are used for this purpose. There may exist epochs in which noise is not completely eliminated despite applying the filtering methods. These epochs are removed by applying certain thresholds to eliminate highly abnormal R-peaks and physiologically impossible heart rates.

Feature extraction is performed using a 5-minute window in which the characteristics of the ECG signal are best revealed for 1-minute epochs [15]. Depending on the ECG signal's sampling frequency, a discrepancy in the R-peak detection may exist. To complement this difference, we use the sampling rate of each ECG signal to determine not only the threshold for R-peak detection but the sampling size of the interpolation. Our work applies the Hamilton algorithm [22] from BioSPPy [20] to extract R-peaks. The R-peak amplitude is then used to calculate the RR-interval by the distance between two adjacent R-peaks. The numbers of R-peak amplitudes and RR-interval differ by epoch, which makes these potential features inadequate input for the model. Thus, cubic interpolation is used to align the number of R-peak amplitude and RR-interval of all epochs to 900 points each.



(a) An example of ECG signal (b) Normal ECG signal (c) Apnea ECG signal

FIGURE 3. Snapshots of sample ECG waveform and signal.

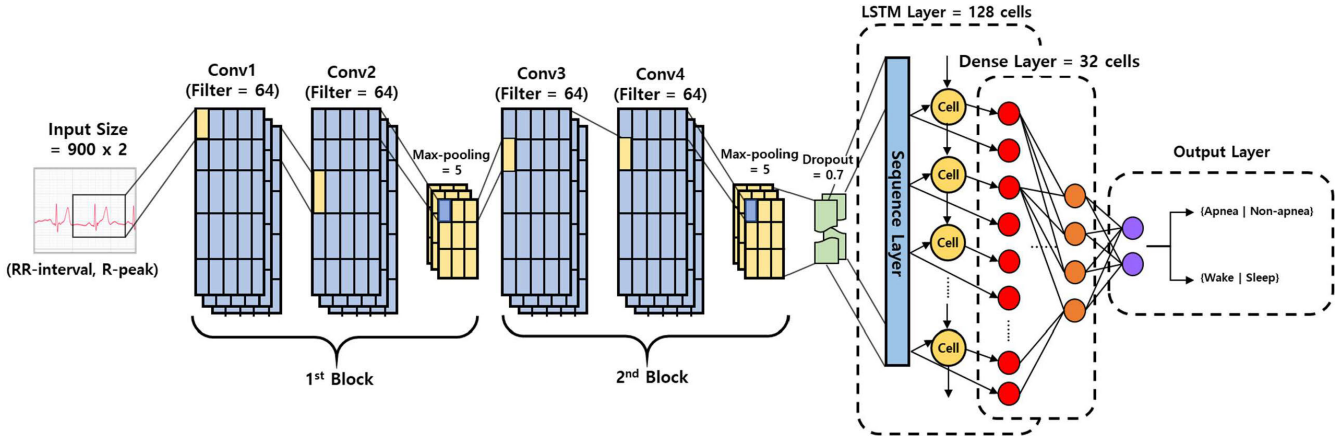


FIGURE 4. Our CLNet model architecture showing the classification part of the overall OSA severity classification scheme process in detail.

B. MODEL TRAINING AND CLASSIFICATION

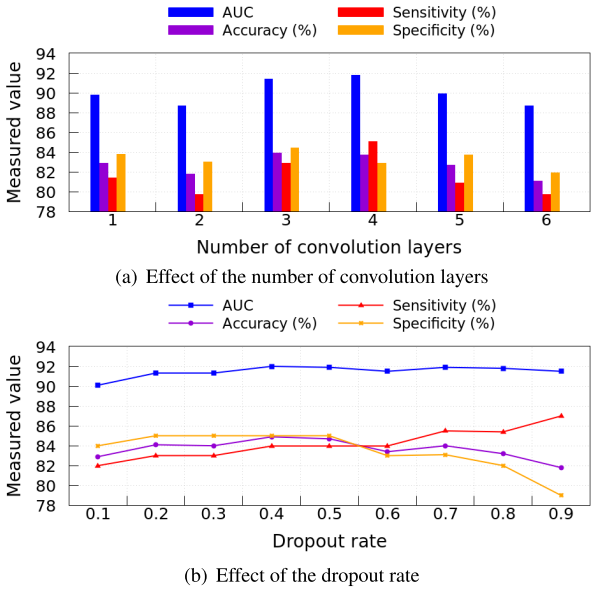
1) DATA IMBALANCE PROBLEM

As prevalent in many real-world data used in classifying problems, PSG data *also* suffers from the data imbalance problem [22]. The data imbalance problem arises when the classes of a given dataset are not represented equally, which leads to unfair disadvantages to the classification result for the minority class. In our study, the Normal to Apnea ratio for the PhysioNet Dataset is 7:4, while that of the KNUCH Dataset is 6:1 for OSA classification. The Sleep to Wake ratio for the KNUCH Dataset is 9:2 for sleep classification. However, the Sleep to Wake ratio for the PhysioNet Dataset is unknown, due to the fact that the PhysioNet Dataset does not provide Sleep and Wake labels. We find the Wake and Apnea classes to be grossly underrepresented. To alleviate this data imbalance problem, we apply SMOTE [23]. But the results did not show a noticeable difference (as will be demonstrated in a later table). Hence, we suggest an additional method of introducing specific patient data representing more of the underrepresented class as the training data. This allows our model to look at a sufficiently balanced binary class, allowing for a fair chance of classification for both classes.

2) RESPIRATORY STATUS AND SLEEP STATE CLASSIFICATION

We propose a hybrid deep learning model, termed *CLNet*, of CNN (Convolutional Neural Networks) and LSTM (Long Short Term Memory), built to classify respiratory states and sleep states using single-lead ECG data. CNN shows high classification performance while LSTM is suitable to classify time series data such as patient ECG signals. Figure 4 is the deep learning classification model structure of our work. Once the ECG data goes through the segmentation, labeling, filtering, and feature extraction processes, it is fed into the CLNet model with a size of 900×2 as input. In other words, the input of our model is a feature map of 900 by 2, in which the first and second rows are the respective sequences of time-series points from the RR-interval and R-peak features extracted from a preprocessed ECG signal with a 30-second epoch.

We have conducted an *ablation study* to understand our model better as we determine the causation of a better metric and apply it to our CLNet model. Here we show the two most significant results of our ablation study, where we investigated the optimal number of convolution layers and the dropout rate. In this study, the filter size is set to 64,



(a) Effect of the number of convolution layers

(b) Effect of the dropout rate

FIGURE 5. Results of the ablation study on our proposed CLNet model.

showing the best performance among several options ranging from 16 through 128 by a factor of two. As exhibited in Figure 5, we chose 4 to be the number of our convolution layers. A dropout rate of 0.7 shows the best sensitivity and AUC results.

Consequently, CLNet has two convolution layers with 64 filters and a max pooling layer in the first block which builds a feature map. The second block is identical to the first block, with a dropout set to 0.7 to prevent overfitting. This results in a 5×64 feature map output. Adding an additional dropout after the first block resulted in an extreme reduction of features. Then, we use LSTM with 128 cells to generate a model that can well reflect the characteristics of time series data. We then use the dense layer to generate a fully connected layer of features from LSTM. Finally, the softmax function is used to classify respiratory and sleep states.

3) AHI CALCULATION AND OSA SEVERITY ESTIMATION

As mentioned, AHI represents how many A/H events occurred during pure sleep hours. The AHI indicates how often a human subject cannot breathe consecutively for more than 10 seconds during sleep. AHI can be derived by dividing the total counts of the A/H events by the genuine sleep hours [2]. Hence, the occurrence of A/H events during the Wake state must be excluded from the total number of A/H events during the entire sleep time. (There was no such case where an A/H event was observed at the Wake state in our dataset.) Also, the true sleep time must be calculated *without* the time classified as Wake.

The degree of OSA severity is measured by AHI, as shown below in Equation 1:

$$AHI = \frac{A/H \ count \ during \ entire \ sleep \ time}{entire \ sleep \ time - wake \ time}. \quad (1)$$

An AHI lower than 5, which can be interpreted as the subject having fewer than 5 A/H events on average in an hour of sleep, is deemed normal. Otherwise, the subject suffers from OSA. We then further classify the different degrees of OSA severity [4] into a 4-class definition where Normal is (AHI $<$ 5), Mild ($5 \leq$ AHI $<$ 15), Moderate ($15 \leq$ AHI $<$ 30) and Severe (AHI $\geq$ 30). We use this 4-class definition for our performance evaluation.

III. EMPIRICAL EVALUATION

A. ENVIRONMENT SETTINGS

The evaluation was run on an Ubuntu 20.04 LTS server with Intel *i*7-9700K, 128 GB RAM, and 1 TB M.2 NVMe SSD. Our models were written in Python. Our code and publicly available datasets, to be described shortly, are released at `https://github.com/lab-paper-code/CLNet`.

For our evaluation, all data were validated using the inter-patient method, which does not use the same patient's data in training and testing. Two types of features were used, with 900 data points extracted per feature per 1-minute epoch. We used CNN and LSTM for our sleep state and respiratory event classification models. As shown in Figure 4 our convolution filter size is 64, Kernel size 5, stride 2, Pooling size 5, Dropout 0.7, LSTM 128, Dense 32, padding valid with a categorical cross-entropy loss function. The Adam optimizer was used with a batch size of 128. Note that these parametric values were all determined by our rigorous ablation studies as exemplified in Figure 5.

In this paper, we use two datasets: a publicly available widely-used data set [13], [17] (termed PhysioNet Dataset) and a private hospital dataset collected from Kyungpook National University Chilgok Hospital (termed KNUCH Dataset) [24]. Next, we describe each dataset and discuss the experimental results of the dataset in detail.

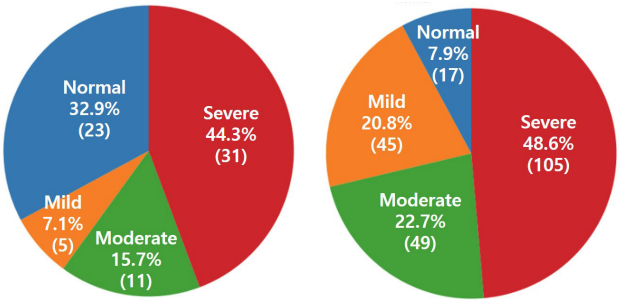
B. PUBLIC PhysioNet DATASET

1) DATA SET INFORMATION

The Public PhysioNet Dataset is the well-known Apnea-ECG database 1.0.0 used at the Computer in Cardiology Challenge 2000 available on PhysioNet [13], [17]. This database consists of 70 ECG records containing 35 in the training set and 35 in the test set. Each ECG signal is 6 to 8 hours long and sampled at 100Hz with sleep apnea states annotated by experts. Segmentation was performed at 1-minute epochs, resulting in 34,428 epochs which translates to 573 hours and 48 minutes worth of data. Non-apnea is labeled as *N* while Apnea is labeled *A*. Apnea and hypopnea are *not* distinguished in this dataset. Additionally, basic patient information such as age, gender, height, and weight, including AHI (0 - 83), is provided for each record. Sleep is not annotated in this particular dataset, and only the total sleep time is recorded. This condition is equal for all research done on this public dataset. A brief summary of the patient demographic is exhibited in Table 1. Figure 6(a) shows the patient distribution of different degrees of OSA severity presented in the data.

TABLE 1. Descriptions of the datasets used in this study: data structure, source, demographic information, number of subjects, and dataset size.

Dataset Name (Structure)	Source	Hertz	Sex	Average						# of Subjects	Total Dataset Size
				Age	Height (cm)	Weight (kg)	BMI	Rec. Hours	AHI		
PhysioNet (Time-series signal)	Apnea-ECG Database [18]	100	M	48.0	176.9	91.6	-	8.2	31.5	57	573 hrs (580.6 MB)
			F	32.8	171.3	65.6	-	8.0	12.8	13	
KNUCH (Time-series signal)	Kyungpook National University Chilgok Hospital [24]	250	M	48.3	171.2	79.7	27.1	7.2	38.7	99	1,580 hrs (70 GB)
			F	41.0	152.8	60.7	24.7	6.9	26.2	6	
		1024	M	47.7	172.8	80.8	27.0	7.4	31.7	77	
			F	61.7	154.8	59.4	25.2	7.5	19.4	34	



(a) PhysioNet Dataset (b) KNUCH Dataset

FIGURE 6. OSA severity distribution of both datasets.

2) RESPIRATORY EVENT CLASSIFICATION RESULTS

Table 2 provides a comparative evaluation of the performance of our proposed model and that of the various earlier studies using the PhysioNet Dataset with the existing label, P-60. As exhibited in Table 2, CLNet, our proposed model, showed 91.2% accuracy, 97.2 AUC, 91.1% sensitivity, and 91.3% specificity. We show a higher level of accuracy compared to the relevant studies, even while using the existing P-60. Especially, AUC increased by up to 10.3% with an accuracy increase of 9.8%. All in all, our CLNet is overall the best and outperforms all other existing work.

C. PRIVATE KNUCH DATASET

1) DATA SET INFORMATION

This dataset is derived from a private database consisting of single-lead ECG among PSG data collected at KNUCH. As illustrated in Table 1, ECG records of on average 7 hours for a total of 216 patients (40 female, 176 male) with sleep problems were collected. This adds up to 94,849 1-minute epochs, translating into 1,580 hours and 49 minutes of overall data. 105 patients were recorded at 250Hz and 111 patients were recorded at 1024Hz, depending on the type of PSG machine utilized. A single expert annotated sleep stages and respiratory events. In the case of sleep stage, the American Academy of Sleep Medicine Guideline [4] was followed. In the case of apnea events, it consisted of Normal, Obstructive Apnea, Hypopnea, Mixed Apnea, and Central Apnea. Note that Mixed Apnea includes Obstructive Apnea and Central Apena. In addition, a PSG report for each patient is provided to check basic patient information including AHI. The subjects are between the ages of 5 and 97, with 3 non-adult subjects whose ages are 5, 17, and 18, respectively. The number of non-adults was not sufficient enough to separate into a different group, so we merged the non-adult subjects as one group. However, with the number of non-adult subjects being minuscule, the effect on the overall data set should be minimal. Subject BMIs (Body Mass Indexes) are measured between 13.3 and 45.9. AHI varies much from 0.2 to 98.3, resulting in a mix of Normal, Mild, Moderate, and Severe OSA cases as seen in Figure 6(b).

2) SEGMENTATION AND LABELING METHOD EVALUATION RESULTS

Table 3 compares the actual expert annotated A/H count and the A/H count resulting from each labeling technique described in Section II-A1. C-20 was the best with the lowest errors. Specifically, MAE (Mean Absolute Error) of 5.7, MAPE (Mean Absolute Percentage Error) of 2.3%, and RMSE (Root Mean Square Error) of 16.6, while P-60 showed the worse with MAE of 84.9, MAPE of 47.7% and RMSE of 103.7. C-20 showed a whopping 20 times improvement of MAPE over P-60. Meanwhile, C-60 fell approximately in the median of the other two methods.

3) RESPIRATORY EVENT CLASSIFICATION RESULTS

Machine learning has been widely used in automatic apnea detection before the introduction of deep learning. In this paper, we have compared the results of several machine-learning techniques with our proposed scheme. In order to build a machine learning model on ECG data, a feature extraction process is needed. HRV (Heart Rate Variability) in the ECG data is a key feature for apnea detection [31]. We use NeuroKit [32] for extracting HRV features, consisting of time-domain, frequency-domain, and nonlinear features.

The time-domain feature is the most widely used, intuitive measure to explain the characteristics of HRV [33]. It uses statistical methods to evaluate the RR-interval fluctuation. Specifically, 14 features, including the average, standard deviation, and median values from RR-interval, are

TABLE 2. Performance comparison on the PhysioNet dataset using the existing labeling scheme (P-60).

Reference	Classifier	Accuracy	AUC	Sensitivity	Specificity	F_1-Score
Faal et al. [25]	LeNet	81.4	-	76.6	84.4	-
Li et al. [11]	Hidden Markov	84.7	86.9	88.9	82.1	-
Sheta et al. [7]	CNN+LSTM	86.3	95.1	88.8	-	-
Wang et al. [15]	LeNet	87.6	95.0	83.1	90.3	-
Bahrami et al. [26]	ZFNet-BiLSTM	88.1	-	81.5	92.3	84.0
Mashrur et al. [16]	SCNN	88.5	-	87.3	89.3	-
Shen et al. [27]	MSDA-1DCNN+WLTD	89.4	96.4	89.8	89.1	-
Almutairi et al. [19]	CNN+LSTM	90.9	-	91.2	90.4	92.8
Sharma et al. [6]	Gaussian SVM	90.9	96.0	92.4	88.3	92.6
Our work	**CLNet (CNN+LSTM)**	**91.2**	**97.2**	91.1	**91.3**	88.1

TABLE 3. Labeling performance comparison where the apnea/hypopnea count for each labeling technique is compared to the ground truth.

Labeling Techniques	MAE	MAPE(%)	RMSE
P-60 (existing)	84.9	47.7	103.7
C-60 (proposed)	40.0	20.1	64.6
C-20 (proposed)	**5.7**	**2.3**	**16.6**

extracted. The frequency-domain feature estimates the power spectrum density to split the RR-interval frequency bandwidth [33]. From the Low-Frequency band (0.04 - 0.15Hz) and High-Frequency band (0.15 - 0.4Hz) we extract 7 features such as the ratio and normalized values. Since the time-domain and frequency-domain features do not reflect the changes in HRV, we use the nonlinear features to quantify the difference [34]. Additionally, we extract 34 features, including the Area Index, Guzik's index [34], Slope Index, and Porta's Index [34]. We apply the 5-minute window method to extract these features.

For classification on the KNUCH Dataset using machine learning, we experiment with well-known algorithms such as Gradient Boosting [28], K-Nearest Neighbor [28], Light Gradient Boosting [29], Logistic Regression [28], Random Forest [28], and XGBoost [30]. Table 4 shows the sleep apnea classification results when applying the existing P-60 and new C-60 labeling methods, respectively, for the various classification algorithms. In all classifiers with the P-60 label, the specificity is very high. Still, the sensitivity is too low, meaning that most cases are classified as Non-apnea, but Apnea can *not* be well classified. However, with the C-60 label that we propose applied, the Apnea sensitivity has considerably grown in most existing machine learning models up to by about 18 times (in Logistic Regression). We confirm that our proposed C-60 is very effective and specialized in classifying sleep apnea correctly.

On one hand, to address the data imbalance problem, we created four variations of our proposed model (CLNet): 'unbalanced' using a randomly configured training set, 'balanced SMOTE' using the SMOTE oversampling technique, 'balanced Severe' which introduces high proportion of minority classes in the training set, and 'balanced Severe+SMOTE' which combines both balanced Severe and SMOTE. When applying the P-60 label, the Non-Apnea to Apnea ratio for the unbalanced group is 4.5:1, the balanced Severe group is 2:1, and the rest is 1:1. With the C-60 labeling technique, the Non-Apnea to Apnea ratio for the unbalanced group is 2.5:1, while the other balanced groups showed 1:1. In effect, the proposed C-60 labeling method allows us to introduce unbiased Non-Apnea and Apnea data into the model's training set, which helps resolve the data imbalance problem.

We evaluated the performance of our variations of the proposed CLNet classification model performance on the KNUCH Dataset labeled by P-60 and by C-60. (Note that any ECG data with respiratory event annotations *can* be labeled by these two labeling techniques.) The PhysioNet Dataset is public, with many known studies available. However, since the KNUCH Dataset is a private dataset that is used for the first time in this paper, for performance comparison, we have implemented Wang et al.'s model [15], termed 'Modified LeNet-5,' on this same dataset, given that their code is publicly available.

As exhibited in Table 4, the proposed CLNet when combined with 'balanced Severe+SMOTE' outperforms the compared model [15] with up to 5.8 difference in the experiment using the P-60 label regarding AUC, and an astonishing 41.8% difference regarding sensitivity. In the competitor, the sensitivity (31.7%) is significantly lower than the specificity (97.9%). Namely, the model is highly *biased* toward the majority class, which shows high performance in the Non-Apnea classification but low performance in the Apnea classification. Overall, our sensitivity results are the highest, showing the best performance on Apnea classification, which is *critical* in our work.

The results of the experiment using the C-60 label are also shown in Table 4. Our model improves the accuracy and AUC by 4.4% and 6.7, respectively, compared to the Modified LeNet-5 [15] model. However, we show a shocking increment of 32.1% in sensitivity, meaning that the Apnea class is highly well classified despite a decrease in Non-apnea classification with 84.2% specificity. When using a well-balanced training set, the classification for the minority class increases while the classification for the majority class may decrease. AUC evaluates classification performance using sensitivity

TABLE 4. Performance comparison on the KNUCH dataset.

Label	Model	Accuracy	AUC	Sensitivity	Specificity	F_1-Score
P-60 (existing)	Gradient Boosting [28]	79.3	76.4	10.4	98.1	17.7
	K-Nearest Neighbor [28]	75.8	61.4	13.8	92.7	19.6
	Light Gradient Boosting [29]	79.1	75.7	11.3	97.5	18.8
	Logistic Regression [28]	78.6	68.5	1.2	99.6	2.4
	Random Forest [28]	78.8	72.6	4.2	99.2	7.9
	XGBoost [30]	78.8	73.3	15.6	96.1	23.9
	Modified LeNet-5 [15]	83.7	83.3	31.7	97.9	45.5
	CNN (of **CLNet**) (baseline 1)	84.3	84.9	44.7	95.1	55.0
	LSTM (of **CLNet**) (baseline 2)	78.5	74.2	2.4	95.1	35.0
	CLNet (unbalanced)	**84.5**	84.7	44.8	95.3	54.3
	CLNet (balanced SMOTE)	82.3	82.5	52.6	90.4	56.0
	CLNet (balanced Severe)	84.3	89.0	73.1	87.9	**69.2**
	CLNet (balanced Severe+SMOTE)	83.8	**89.1**	**73.5**	87.0	68.5
C-60 (proposed)	Gradient Boosting [28]	76.7	81.2	57.9	87.2	64.0
	K-Nearest Neighbor [28]	65.9	66.6	39.6	80.5	45.4
	Light Gradient Boosting [29]	75.9	81.2	59.4	85.1	63.8
	Logistic Regression [28]	68.2	72.3	18.8	95.6	29.6
	Random Forest [28]	74.6	78.6	50.3	88.1	58.5
	XGBoost [30]	75.0	80.3	57.0	84.9	61.9
	Modified LeNet-5 [15]	80.1	85.4	52.8	95.5	65.7
	CNN (of **CLNet**) (baseline 1)	84.2	89.2	68.1	93.2	75.6
	LSTM (of **CLNet**) (baseline 2)	67.0	76.9	11.1	98.4	19.6
	CLNet (unbalanced)	84.2	89.0	69.3	92.6	75.9
	CLNet (balanced SMOTE)	83.5	88.8	74.4	88.7	76.5
	CLNet (balanced Severe)	83.7	91.6	84.5	93.1	79.2
	CLNet (balanced Severe+SMOTE)	**84.5**	**92.1**	**84.9**	84.2	**80.1**

and specificity. The fact that AUC is high indicates that both classes of the model are well-represented and well-classified. For the balanced Severe+SMOTE, AUC is highest at 92.1, indicating that the overall classification of both classes is unparalleled.

We also include the results of running only CNN and LSTM of CLNet each on the KNUCH Dataset labeled with P-60 and C-60 as a baseline for our work. The accuracy of CLNet for P-60 label is slightly higher than CNN, but the sensitivity, which represents the performance of Apnea classification, shows a significant increment of 28.8%. This means that Apnea classification performed poorly for CNN, which can be interpreted as most of the data points being classified as Non-Apnea, leading to high specificity. Meanwhile, the accuracy, AUC, and sensitivity are all higher for C-60; in particular, the sensitivity shows a 16.8% difference. Also, LSTM underperforms CLNet for both P-60 and C-60; in particular, the sensitivity of LSTM offers a critically low 2.4% and 11.1% performance for P-60 and C-60, respectively.

In the case of C-20, which is 3-class, the introduction of the *F* label creates a *harsher* data imbalance attributed to relatively more minor labeling of A classes than with the 2-class labeling. The balanced Severe+SMOTE is applied to show an accuracy of 79.2% while the F_1-scores of *N*, *A*, and *F* class show 87.2%, 51%, and 71%, respectively. Future work needs to improve the classification performance of the *A* label to achieve a more sophisticated respiratory-state classification.

4) SLEEP TIME ESTIMATION RESULTS

Here we discuss the overall sleep time estimation performance measured using the error metrics. To derive true sleep time, our proposed CLNet model performs a binary class classification: Sleep or Wake. As a result, our model achieves an accuracy of 85.3%, an AUC of 85, and a sensitivity of 87%. Pure sleep time is calculated by deducting the classified Wake time from the entire sleep time. As with the respiratory event classification, the minority class was intensively configured in the training set. We compared the estimated pure sleep time with the expert annotated sleep time without any Wake time. Consequently, our model achieves an MAE of 0.73, MAPE of 10.92%, and RMSE of 0.99.

TABLE 5. 4-Class OSA severity classification report.

Class	Precision	Sensitivity	Specificity	F_1
Normal	0.60	0.60	0.96	0.60
Mild	0.85	0.89	0.90	0.87
Moderate	1.00	0.92	1.00	0.96
Severe	1.00	1.00	1.00	1.00

5) OSA SEVERITY ESTIMATION RESULTS

Exhibited in Table 5 are the results of the 4-class OSA severity classification that our model performs. When we detect OSA, the higher the precision, the higher the confidence that the subject estimated as OSA is truly an OSA patient, and the higher the sensitivity, the higher the probability of estimating OSA patients as true OSA. The precision in boundary cases

such as Mild is relatively low, at about 85%, but the estimation for serious OSA patients show perfect precision at 100%, in both Moderate and Severe cases.

Figure 7 compares the performance of OSA severity classification between our proposed model and Wang et al.'s model [15]. Our model's total accuracy for 4-class shows 90%, as displayed in Figure 7(c). In the compared model, the authors used the ''mixed'' sleep time without excluding the Wake time when calculating AHI. Thus, to make a fair comparison between ours and their model, we also compute a value of AHI over the entire sleep time, including the Wake time. As a result, our model wins their model by gaining an increase of 2% (Figure 7(a) vs. Figure 7(b)). When the deduction of the Wake time is reflected, our model still outperforms their model by 12% (Figure 7(a) vs. Figure 7(c)). From these results, we find that the proposed CLNet is more effective than the compared model [15] in OSA severity classification. Furthermore, our model's accuracy spikes when we consider and deduct the Wake time. These empirical results support our decision to use a Wake time-aware sleep state classification model to estimate sleep time for better calculating AHI to classify OSA severity.

6) SIGNIFICANCE AND IMPACT

Our proposed CLNet model yielded a total accuracy of 90% for the 4-class OSA severity classification on the 50 tested human subjects, as visualized in Figure 7(c). We emphasize that our results were obtained across many human subjects never used in training. We could reproduce a similar estimation model [15] using modified LeNet-5 and fit this on our datasets. Our model also performed 12% better total accuracy on the KNUCH dataset in 4-class classification compared to the state-of-the-art model.

Based on these results, our model has revealed a clear potential as a ''pre-screening'' tool to locate patients with Severe OSA more accurately. Again, to the best of our knowledge, this is the first to estimate the degree of OSA severity, reflecting the Wake state in the sleep stage on a large-scale group of actual human subjects reaching a triple-digit number.

7) DISCUSSION

Table 6 summarizes the advantages and drawbacks of our model compared with the closest work, LeNet-5 [15]. Our CLNet and LeNet-5 models use a single-lead ECG signal, getting them to work easily in a wearable device. Our work, though, is superior to the LeNet-5 in many folds. One of the biggest advantages, which makes our work original and novel, is that our model can estimate OSA severity degree more accurately by deducting Wake time from the total sleep hours. In the meantime, the LeNet-5 does not consider this; thus, it may risk underestimating the OSA severity degree. Also, our CLNet resolves the data imbalance problem by intensively learning Wake and Apnea/hypopnea labels when trained, but the compared work does not address this imbalance issue. Moreover, our model outperforms the LeNet-5 up to 12% in accuracy; in particular, it achieves 100% recall on severe OSA cases. In the meantime, the proposed CLNet model spends slightly more training time than the LeNet-5 model, mainly due to its complexity.

On one hand, both works reveal three disadvantages. First, due to an ECG epoch segmentation, an apnea or hypopnea event may occur between two 60-second epochs, and a 60-second epoch may contain several apnea and hypopnea events. In addition, the data used in the models cannot distinguish between apnea and hypopnea events, soliciting a more sophisticated classification model. Lastly, both works do not consider detecting central sleep apnea that occurs less frequently than OSA.

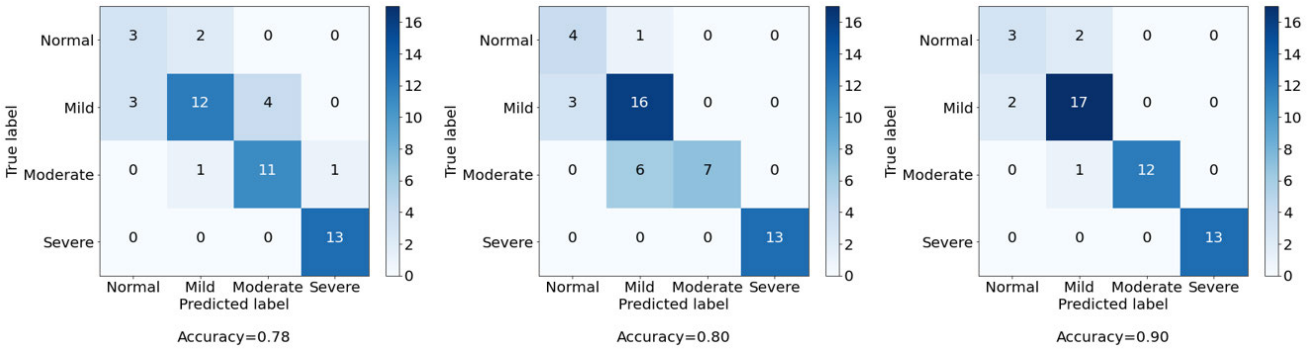
In the following section, we compare our work with other major studies [15], [16], [19], [35], [36] in greater detail.

IV. RELATED WORK

Many researchers have studied sleep apnea and experimented on better detecting and classifying patients with this OSA disease. A wealth of the research employs machine learning techniques, such as Support Vector Machine (SVM) [37] or Gaussian SVM [6] and Hidden Markov [11] to name a few. Kim et al. [35] explored various machine learning models such as Logistic Regression, Random Forest, XGBoost as well as SVM and concluded that SVM showed the best result in their study. Many others imported assorted deep learning techniques such as various adaptations of CNN [7], [15], [16], [19], [38] or other deep learning models [11], [37] in an attempt to classify patients with OSA. Others used different biosignals such as SpO2 [39] or respiration signals such as Oronasal thermal airflow (FlowTh), Nasal pressure (NPRE), and abdominal respiratory inductance plethysmography (ABD) [40]. Even when using single-lead ECG to classify Apnea, previous work used different features such as Instant Heart Rate (IRH) [39], [41].

Schlüter et al. [36] used frequency analysis techniques to extract features from ''quadruple'' bio-signals captured by PSG, classify sleep stages using decision trees, and detect A/H, resulting in a 95.2% accuracy for sleep stage scoring and 94.5% for classifying A/H. However, they used the Rechtschaffen and Kales rule annotation [42] with multi-channel classification. Thus, their work fundamentally differs from our American Academy of Sleep Medicine Guideline [4] on single-lead ECG signal classification. This makes it difficult to perform a fair comparison with our work. Also, there are previous studies to predict the sleep stages by using other biosignals such as Electroencephalography (EEG) [43]. In contrast, we use the ECG signals, making it possible to test the degree of OSA severity in a wearable device.

Almutairi et al. [19] used CNN and LSTM networks to automatically extract the features from the ECG signals. SVM was used to classify OSA and healthy ECG signals, with sensitivity, specificity, and overall accuracy being 91.24%, 90.36%, and 90.92%, respectively.



(a) Wang et al.'s model [15] (b) CLNet - Wake Time Oblivious (c) CLNet - Wake Time Aware

FIGURE 7. 4-class confusion matrix on OSA severity classification.

TABLE 6. Advantages and drawbacks of our model and other studies on the PhysioNet dataset and the KNUCH dataset.

Model	Advantages	Drawbacks
CLNet (ours)	• A single-lead ECG signal • Wake-aware: accurate estimation of OSA • Data imbalance resolved • Better accuracy: perfect recall on severe OSA	• Relatively more training time • Potential counting error on apnea/hypopnea event • Cannot distinguish between apnea and hypopnea • Does not consider central sleep apnea
LeNet-5 [15]	• A single-lead ECG signal • Relatively less training time	• Wake-oblivious: underestimation of OSA • Vulnerable to data imbalance • Potential counting error on apnea/hypopnea event • Cannot distinguish between apnea and hypopnea • Does not consider central sleep apnea

Mashrur et al. [16] explored a Scalogram-based CNN to detect OSA using single-lead ECG data on the PhysioNet Apnea ECG dataset. Their sensitivity, specificity, and overall accuracy were 87.3%, 89.3%, and 88.5%, respectively.

Wang et al. [15] selected the R-peak feature from the ECG signal, which is highly relevant to breathing. By introducing a modified LeNet-5 deep learning model using RR-interval and R-peak amplitude, the accuracy, sensitivity, and specificity were 87.6%, 83.1%, and 90.3%, respectively. We could reproduce this model thanks to the released code provided by the authors. We have used this model as the closest competitor for direct comparison to our model. Our work reveals *better* accuracy, sensitivity, and specificity, as demonstrated in Tables 2 and 4.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel deep learning-based scheme for accurately estimating the degree of sleep apnea based on single-lead ECG data recorded during actual PSG. Unlike previous works, for more accurate AHI estimation results we introduced a method of calculating and excluding the total duration of Wake states from a night of overnight sleep. Moreover, we presented another method of accurately detecting A/H events by applying CNN and LSTM on a single-lead ECG. Using these two methods, along with a novel labeling technique as well as features extracted from the data such as RR-interval and R-peak amplitude, we were able to predict a subject's apnea classification with a total accuracy of 84.5% on the KNUCH dataset of 216 subjects, and 91.2% total accuracy on the PhysioNet Dataset of 70 subjects. Our scheme demonstrates the effectiveness of accurately predicting Severe OSA or Normal cases on real-world clinical datasets.

In the future, we plan to strengthen the proposed classification model from a variety of viewpoints. First, we can use a combination of other bio-signals such as SpO_2, adding more features to extract such as EDR(ECG-Derived Respiration), or utilizing different deep learning techniques to further improve the performance, such as Bidirectional LSTM (BiLSTM) and Gated Recurrent Unit (GRU). In addition, it would be meaningful to investigate the most important features of the model from the clinical perspective. Another research path will include ways to further resolve the data imbalance problem, using different undersampling and oversampling techniques or a mix of both [44]. It is also essential to build a better sleep time estimation model [45] on real patient datasets. We are currently developing a lightweight version of our model with plans to launch an OSA pre-screening service for wearable or mobile devices. A better judgment between the boundary cases at the intermediate level of OSA is to be investigated. Although our work focuses on determining the degree of OSA severity through binary classification of breathing events, it can be extended to classify respiratory status into three categories: Apnea, Hypopnea, and Normal. Future work may include regression analysis to assess Apnea severity, investigate potential

influences of abnormal respiratory events on the estimation of OSA severity degree using ECG signals, and explore feature selection using image interpretation techniques such as Grad-CAM [46] and deepSHAP [47]. Additionally, exploring the effects of appropriately handling AC power and EMG artifacts in raw ECG signals during preprocessing could be interesting in enhancing the accuracy of estimating the severity degree of OSA. Finally, we will extend our model to cover more patients and look into other clinical datasets.

ACKNOWLEDGMENT

(Dae-Woong Seo and Jeeyoung Kim contributed equally to this work.)

REFERENCES

[1] S. Jehan, A. K. Myers, F. Zizi, S. R. Pandi-Perumal, G. Jean-Louis, and S. I. McFarlane, "Obesity, obstructive sleep apnea and type 2 diabetes mellitus: Epidemiology and pathophysiologic insights," *Sleep Med. Disorders, Int. J.*, vol. 2, no. 3, p. 52, Jun. 2018.

[2] J. Kim, K. In, J. Kim, S. You, K. Kang, J. Shim, S. Lee, J. Lee, S. Lee, C. Park, and C. Shin, "Prevalence of sleep-disordered breathing in middle-aged Korean men and women," *Amer. J. Respiratory Crit. Care Med.*, vol. 170, no. 10, pp. 1108–1113, Nov. 2004.

[3] D. Norman and J. S. Loredo, "Obstructive sleep apnea in older adults," *Clinics Geriatric Med.*, vol. 24, no. 1, pp. 151–165, Feb. 2008.

[4] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. Marcus, and B. V. Vaughn, "The AASM manual for the scoring of sleep and associated events," Amer. Acad. Sleep Med., Rules, Terminol. Tech. Specifications, Darien, IL, USA, Tech. Rep., 2012, vol. 176.

[5] H. Sharma and K. K. Sharma, "An algorithm for sleep apnea detection from single-lead ECG using Hermite basis functions," *Comput. Biol. Med.*, vol. 77, pp. 116–124, Oct. 2016.

[6] M. Sharma, M. Raval, and U. R. Acharya, "A new approach to identify obstructive sleep apnea using an optimal orthogonal wavelet filter bank with ECG signals," *Informat. Med. Unlocked*, vol. 16, Jan. 2019, Art. no. 100170.

[7] A. Sheta, H. Turabieh, T. Thaher, J. Too, M. Mafarja, M. S. Hossain, and S. R. Surani, "Diagnosis of obstructive sleep apnea from ECG signals using machine learning and deep learning classifiers," *Appl. Sci.*, vol. 11, no. 14, p. 6622, Jul. 2021.

[8] A. Yadollahi and Z. Moussavi, "Acoustic obstructive sleep apnea detection," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Sep. 2009, pp. 7110–7113.

[9] L. Almazaydeh, M. Faezipour, and K. Elleithy, "A neural network system for detection of obstructive sleep apnea through SpO_2 signal features," *Int. J. Adv. Comput. Sci. Appl.*, vol. 3, no. 5, p. 18, 2012.

[10] S. Nikkonen, I. O. Afara, T. Leppänen, and J. Töyräs, "Artificial neural network analysis of the oxygen saturation signal enables accurate diagnostics of sleep apnea," *Sci. Rep.*, vol. 9, no. 1, pp. 1–9, Sep. 2019.

[11] K. Li, W. Pan, Y. Li, Q. Jiang, and G. Liu, "A method to detect sleep apnea based on deep neural network and hidden Markov model using single-lead ECG signal," *Neurocomputing*, vol. 294, pp. 94–101, Jun. 2018.

[12] A. C. T. de Oliveira, D. Martinez, L. F. T. Vasconcelos, S. C. Gonçalves, M. do Carmo Lenz, S. C. Fuchs, M. Gus, E. O. de Abreu-Silva, L. B. Moreira, and F. D. Fuchs, "Diagnosis of obstructive sleep apnea syndrome and its outcomes with home portable monitoring," *Chest*, vol. 135, no. 2, pp. 330–336, Feb. 2009.

[13] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[14] M. J. Sateia, "International classification of sleep disorders," *Chest*, vol. 146, no. 5, pp. 1387–1394, 2014.

[15] T. Wang, C. Lu, G. Shen, and F. Hong, "Sleep apnea detection from a single-lead ECG signal with automatic feature-extraction through a modified LeNet-5 convolutional neural network," *PeerJ*, vol. 7, p. e7731, Sep. 2019.

[16] F. R. Mashrur, M. S. Islam, D. K. Saha, S. M. R. Islam, and M. A. Moni, "SCNN: Scalogram-based convolutional neural network to detect obstructive sleep apnea using single-lead electrocardiogram signals," *Comput. Biol. Med.*, vol. 134, Jul. 2021, Art. no. 104532.

[17] T. Penzel, G. B. Moody, R. G. Mark, A. L. Goldberger, and J. H. Peter, "The apnea-ECG database," in *Proc. Comput. Cardiol.*, vol. 27, Sep. 2000, pp. 255–258.

[18] PhysioNet. (2023). *The Official Website*. Accessed: Mar. 23, 2023. [Online]. Available: https://physionet.org/content/apnea-ecg/1.0.0/

[19] H. Almutairi, G. M. Hassan, and A. Datta, "Classification of obstructive sleep apnoea from single-lead ECG signals using convolutional neural and long short term memory networks," *Biomed. Signal Process. Control*, vol. 69, Aug. 2021, Art. no. 102906.

[20] C. Carreiras, A. Alves, A. Lourenço, F. Canento, H. Silva, and A. Fred. (2015). *BioSPPy: Biosignal Processing in Python*. [Online]. Available: https://github.com/PIA-Group/BioSPPy/

[21] P. Virtanen et al., "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, 2020.

[22] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[23] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, Apr. 2018.

[24] Kyungpook National University Chilgok Hospital. (2023). *The Official-Website*. Accessed: Mar. 23, 2023. [Online]. Available: https://en.knuch.kr/

[25] M. Faal and F. Almasganj, "Obstructive sleep apnea screening from unprocessed ECG signals using statistical modelling," *Biomed. Signal Process. Control*, vol. 68, Jul. 2021, Art. no. 102685.

[26] M. Bahrami and M. Forouzanfar, "Sleep apnea detection from single-lead ECG: A comprehensive analysis of machine learning and deep learning algorithms," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022.

[27] Q. Shen, H. Qin, K. Wei, and G. Liu, "Multiscale deep neural network for obstructive sleep apnea detection using RR interval from single-lead ECG signal," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[28] M. Mohammed, M. B. Khan, and E. B. M. Bashier, *Machine Learning: Algorithms and Applications*. Boca Raton, FL, USA: CRC Press, 2016.

[29] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 3146–3154.

[30] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[31] V. C. C. Sequeira, P. M. Bandeira, and J. C. M. Azevedo, "Heart rate variability in adults with obstructive sleep apnea: A systematic review," *Sleep Sci.*, vol. 12, no. 3, p. 214, 2019.

[32] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. A. Chen, "NeuroKit2: A Python toolbox for neurophysiological signal processing," *Behav. Res. Methods*, vol. 53, no. 4, pp. 1689–1696, 2021.

[33] P. K. Stein, "Assessing heart rate variability from real-world Holter reports," *Cardiac Electrophysiol. Rev.*, vol. 6, no. 3, p. 239, 2002.

[34] C. Yan, P. Li, L. Ji, L. Yao, C. Karmakar, and C. Liu, "Area asymmetry of heart rate variability signal," *Biomed. Eng. OnLine*, vol. 16, no. 1, pp. 1–14, Dec. 2017.

[35] Y. J. Kim, J. S. Jeon, S.-E. Cho, K. G. Kim, and S.-G. Kang, "Prediction models for obstructive sleep apnea in Korean adults using machine learning techniques," *Diagnostics*, vol. 11, no. 4, p. 612, Mar. 2021.

[36] T. Schluter and S. Conrad, "An approach for automatic sleep stage scoring and apnea-hypopnea detection," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2010, pp. 1007–1012.

[37] H. Singh, R. K. Tripathy, and R. B. Pachori, "Detection of sleep apnea from heart beat interval and ECG derived respiration signals using sliding mode singular spectrum analysis," *Digit. Signal Process.*, vol. 104, Sep. 2020, Art. no. 102796.

[38] A. John, B. Cardiff, and D. John, "A 1D-CNN based deep learning technique for sleep apnea detection in IoT sensors," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2021, pp. 1–5.

[39] R. K. Pathinarupothi, J. D. Prathap, E. S. Rangan, E. A. Gopalakrishnan, R. Vinaykumar, and K. P. Soman, "Single sensor techniques for sleep apnea diagnosis using deep learning," in *Proc. IEEE Int. Conf. Healthcare Informat. (ICHI)*, Aug. 2017, pp. 524–529.

[40] H. El Moaqet, 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.
[41] R. K. Pathinarupothi, R. Vinaykumar, E. Rangan, E. Gopalakrishnan, and K. P. Soman, "Instantaneous heart rate as a robust feature for sleep apnea severity detection using deep learning," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat. (BHI)*, 2017, pp. 293–296.
[42] A. Kales and A. Rechtschaffen, *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. Bethesda, MD, USA: US Department of health, Education and Welfare, Neurological Information Network, 1968.
[43] I. Hussain, M. A. Hossain, R. Jany, M. A. Bari, M. Uddin, A. R. M. Kamal, Y. Ku, and J.-S. Kim, "Quantitative evaluation of EEG-biomarkers for prediction of sleep stages," *Sensors*, vol. 22, no. 8, p. 3079, Apr. 2022.
[44] J. Wang and M.-L. Zhang, "Towards mitigating the class-imbalance problem for partial label learning," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 2427–2436.
[45] S. Paisarnsrisomsuk, M. Sokolovsky, F. Guerrero, C. Ruiz, and S. A. Alvarez, "Deep sleep: Convolutional neural networks for predictive modeling of human sleep time-signals," in *Proc. KDD Deep Learn. Day*, 2018, pp. 1–10.
[46] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[47] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 4765–4774.

DAE-WOONG SEO received the B.S. and M.S. degrees from the Department of Mathematics, School of Computer Science and Engineering, Kyungpook National University, Daegu, Republic of Korea, in 2020 and 2022, respectively. His research interests include data mining and machine learning.

JEEYOUNG KIM (Member, IEEE) received the Ph.D. degree in computer and information science and engineering from the University of Florida, Gainesville, FL, USA, in 2013. From 2013 to 2018, she was a Senior Engineer with Samsung Electronics Mobile Division. From 2018 to 2019, she held the role of Research Faculty with the Center of Self-Organizing Software, Kyungpook National University, Daegu, Republic of Korea. From 2019 to 2022, she was a Teaching Faculty with the School of Computer Science and Engineering, Kyungpook National University. Since September 2022, she has been an Assistant Professor with the Graduate School of Data Science, Kyungpook National University. Her research interests include real-life human generated data, machine learning, and big data systems.

HO-WON LEE is currently a Professor of neurology with the School of Medicine, Kyungpook National University, Daegu, South Korea, and the Director of the Geriatric Health Care Center, Kyungpook National University Chilgok Hospital. After finishing his neurology residency with Kyungpook National University Hospital, in 2003, he completed a fellowship with the Department of Neurology and Alzheimer Disease Center, Baylor College of Medicine, in 2004, and the Center for Movement Disorders and Neurorestoration, University of Florida, in 2014. The primary goal of his research has been the establishment of a scientific basis for the rational early diagnosis and treatment of Alzheimer's disease, Parkinson's disease, and sleep apnea.

YOUNG-KYOON SUH (Member, IEEE) received the Ph.D. degree from the Department of Computer Science, University of Arizona, in 2015. From 2005 to 2017, he was a Senior Researcher with the Korea Institute of Science and Technology Information. Since September 2017, he has been a Faculty Member with the School of Computer Science and Engineering, Kyungpook National University, Daegu, Republic of Korea, where he is currently an Associate Professor. His research interests include databases, data mining, big data systems, machine learning, and storage systems. He is a member of ACM and IEICE. He was a recipient of the Best Poster Award of IEEE/ACM CCGrid 2016 and the Best Paper Award of EDB 2016.