Evaluation of OSA Patient Sleep Stage Classification Performance Using a Multi-Channel PSG Dataset

Younghoon Na[1], Dongyoung Kim[2] Dong-Kyu Kim[3,4,†] and Jeong-Gun Lee[1,2,†]
[1]Division of Software, Hallym University, Chuncheon, Republic of Korea,
[2]Dept. of Computer Engineering, Hallym University, Chuncheon, Republic of Korea,
[3]Department of Otorhinolaryngology-Head and Neck Surgery, Chuncheon Sacred Heart Hospital,
Hallym University College of Medicine, Chuncheon, Republic of Korea,
[4]Institute of New Frontier Research and Division of Big Data and Artificial Intelligence,
Chuncheon Sacred Heart Hospital, Chuncheon, Republic of Korea,
† Equally contributed as co-corresponding authors.
Email: 20185124@hallym.ac.kr, doctordk@naver.com, jeonggun.lee@hallym.ac.kr

***Abstract*—In this paper, we conduct a comparative analysis of sleep stage classification for patients having different levels of obstructive sleep apnea (OSA). For the analysis, we use 10 bio-signal channels: 4 EEG (Electroencephalogram) channels (F3-M2, F4-M1, C3-M2, and C4-M1), 2 EOG (Electrooculogram) channels (E1-M2 and E2-M1), and 4 other bio-channels (EMG, ECG, Flow, and Abdomen).**

In this work, in particular, we consider OSA severity for training and testing. Then, we investigated the detailed impacts of the OSA severity on the accuracy performance of a modern deep learning model with single channel and multiple channels.

***Index Terms*—Sleep Stage Classification, Deep Learning, Multi-channel Sleep Study, Sleep Stage Scoring, OSA**

I. Introduction

Deep learning has steadily been studied in various application fields and it is receiving a lot of attention continuously from industries and academia. Specifically, high interest in deep learning approaches has been increasing in various medical applications.The interest in deep learning has increased not only the image based medical diagnosis but also in the part of medical data analysis with bio-signal data.

Among the medical application exploiting bio-signal data for medical diagnoses, these days, ”sleep stage classification” based on various bio-signal data is getting much attention continuously in the diagnoses for human well-being in modern life. Evaluating the quality of a human/patient through the classification with high reliability is a very important issue because it directly can affect a doctor's medical diagnosis and following treatment. However, scoring sleep quality by staging is highly time-consuming and labor-intensive. In consequence, there have been many studies on automatic sleep stage classification. [1].

Although previous sleep stage classification has been performed with modern deep learning approaches [2]–[4], there hasn't been much work considering "Obstructive Sleep Apnea" (OSA) severity extensively based on multi-channel PSG (polysomnographic) data [5]. On the other hand, the public PSG datasets have some problems when evaluating the model accuracy performance with various channels combination because the public datasets have fewer channels than real-world PSG data.

In this paper, we use 142 real-world PSG patient data gathered in a medical institution using 'Noxturnal' sleep study software to evaluate the model performance when utilizing various channels combination.

II. BACKGROUND

A. Sleep Stage Classification

PSG is a medical test for figuring out the cause of a disease or a disorder that occurs during a patient's sleep from various signals such as EEG, EOG, EMG (Electromyography) and etc that are collected during the test. Using these bio-signals, sleep experts score the sleep stage into 5 classes, Wake, N1 (None-REM1), N2 (None-REM2), N3 (None-REM3), and REM, according to the criterion that is suggested by AASM (American Academy Sleep Medicine) [6].

TABLE I
CLASS DISTRIBUTIONS OF OUR PSG DATASET

Channels	OSA	Wake	N1	N2	N3	REM
Train	Normal	2732 (26.9%)	1048 (10.3%)	4759 (46.8%)	592 (5.8%)	1043 (10.3%)
	Mild	2695 (26.1%)	1124 (10.9%)	4431 (42.8%)	892 (8.6%)	1199 (11.6%)
	Moderate	2536 (25.1%)	1618 (16.0%)	4170 (41.3%)	497 (4.9%)	1279 (12.7%)
	Severe	2680 (30.5%)	2866 (32.6%)	2101 (23.9%)	173 (2.0%)	977 (11.1%)
Validation	Normal	703 (27.9%)	236 (9.4%)	1150 (45.6%)	161 (6.4%)	272 (10.8%)
	Mild	655 (26.0%)	274 (10.9%)	1088 (43.2%)	230 (9.1%)	272 (10.8%)
	Moderate	614 (23.8%)	382 (14.8%)	1119 (43.3%)	124 (4.8%)	343 (13.3%)
	Severe	663 (29.7%)	706 (31.7%)	544 (24.4%)	51 (2.3%)	266 (12.0%)
Test	Normal	2395 (20.6%)	1242 (10.7%)	5445 (46.9%)	1085 (9.3%)	1451 (12.5%)
	Mild	2857 (24.1%)	1174 (9.9%)	5169 (43.6%)	1213 (10.2%)	1434 (12.1%)
	Moderate	2249 (20.0%)	1896 (16.9%)	4647 (41.4%)	859 (7.6%)	1583 (14.1%)
	Severe	2842 (24.5%)	3103 (26.8%)	3548 (30.6%)	416 (3.6%)	1676 (14.5%)

B. Obstructive Sleep Apnea

In a sleep study, OSA is one of key indicators for measuring sleep disorder severity. The OSA severity is determined by AHI (Apnea Hypopnea Index) and the OSA can be classified into 4 categories as follows:

- Normal: AHI is less than 5
- Mild: AHI is between 5 and 15
- Moderate: AHI is between 15 and 30
- Severe: AHI is greater than 30

The AHI is calculated by hourly averaging the sum of the number of apneas (pauses in breathing) and the number of hypopneas that a patient experience during a test.

III. METHODS

A. Dataset

For experiment, we use raw signal data sampled at 200Hz and the dataset includes 10 signal channels : F3-M2, F4-M1, C3-M2, C4-M1, E1-M2, E2-M1, EMG, ECG, Flow, and Abdomen. For deep learning generalization, the train / test ratio is set to around 0.5. Table I shows the detailed class distribution of our dataset.

B. Model Architecture

We evaluate the accuracy performance using DeepSleepNet [2] which consists of two branches including four convolution layers as shown in Fig. 1. Then, the two branches are merged and two bidirectional LSTMs are followed to detect sequential features in the data. In the CNN part, per sample, local features are extracted using two CNNs which are differently configured with different convolution windows. Then, the bidirectional LSTM extracts temporal information between samples. In addition, the model includes a shortcut connection and through the shortcut, the CNN features and temporal features are added together finally to feed the last classifier.

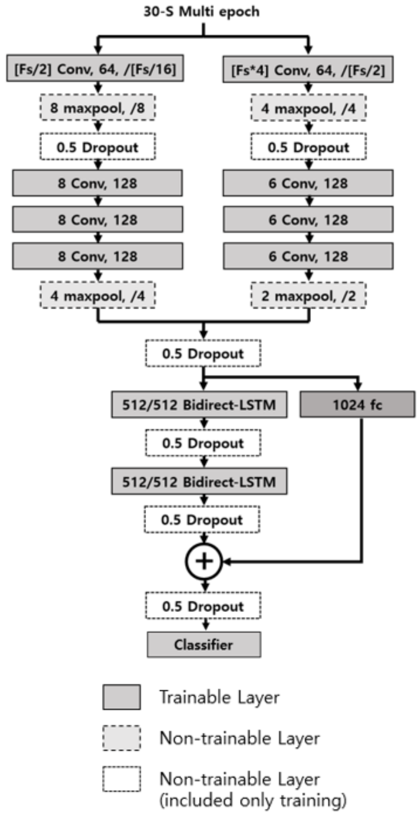

Fig. 1. DeepSleepNet Architecture [2]

C. Training Parameters

The cross-entropy is used as a loss function. An Adam optimizer is used and its parameters are set to 10^{-4}, 0.9, and 0.999 for lr, $beta1$, and $beta2$, respectively. 'CosineAnnealingLR' is employed as a scheduler during training.

IV. EXPERIMENT AND RESULTS

The detailed model performances are presented in Table II. When only a single channel (C3-M2) is used,

TABLE II
MODEL PERFORMANCE ACCORDING TO VARIOUS CHANNELS COMBINATIONS

Channels	Per-class metrics	Wake	N1	N2	N3	REM
C3-M2	Precision	0.831	0.543	0.855	0.773	0.680
	Recall	0.841	0.541	0.833	0.771	0.726
	F1-Score	0.836	0.542	0.843	0.772	0.702
	Accuracy	76.96%				
C3-M2, E1-M2	Precision	0.860	0.612	0.829	0.859	0.693
	Recall	0.830	0.564	0.858	0.731	0.825
	F1-Score	0.845	0.587	0.843	0.790	0.753
	Accuracy	78.53%				
F3-M2, C3-M2, E1-M2	Precision	0.861	0.524	0.868	0.807	0.757
	Recall	0.830	0.614	0.838	0.760	0.780
	F1-Score	0.846	0.565	0.853	0.783	0.768
	Accuracy	79.18%				
ALL	Precision	0.835	0.595	0.863	0.772	0.760
	Recall	0.866	0.593	0.842	0.769	0.778
	F1-Score	0.850	0.594	0.852	0.771	0.769
	Accuracy	79.31%				

TABLE III
MODEL PERFORMANCE ACCORDING TO VARIOUS CHANNELS COMBINATIONS FOR OSA SEVERITY

Channels	OSA	Wake	N1	N2	N3	REM	ACC
C3-M2	Normal	87.1	40.6	87.7	79.6	70.0	79.2
	Mild	85.3	45.5	88.0	84.8	68.8	79.9
	Moderate	85.8	55.9	81.1	69.8	76.9	76.6
	Severe	79.2	62.7	73.6	62.4	75.1	72.0
C3-M2, E1-M2	Normal	84.8	45.4	88.7	75.0	80.8	80.6
	Mild	84.1	49.1	90.2	80.7	74.6	81.1
	Moderate	86.8	56.7	84.9	68.2	89.2	78.7
	Severe	77.8	63.6	76.8	59.5	87.0	73.6
F3-M2, C3-M2, E1-M2	Normal	85.7	48.5	87.3	76.5	77.4	81.5
	Mild	85.6	58.5	87.5	84.3	68.6	81.8
	Moderate	84.4	60.0	82.1	79.0	85.0	79.3
	Severe	77.6	67.4	75.6	53.8	82.7	74.0
ALL	Normal	84.6	44.7	87.2	79.5	74.0	80.5
	Mild	87.9	56.1	87.8	82.3	70.4	82.1
	Moderate	90.4	62.0	82.1	78.0	84.6	80.2
	Severe	84.3	63.0	76.8	53.1	83.1	74.4

76.96% accuracy performance is obtained. When multiple channels (F3-M2, C3-M2, and E1-M2) are used, the model obtains 79.18% accuracy. This performance is similar to the performance (79.31%) we obtained from the use of the whole 10 channels in our PSG dataset. We expect the use of three channels (F3-M2, C3-M2, and E1-M2) can provide quite saturated accuracy performance while minimizing the number of bio-sensors attached to patients during a test.

Table III shows the performance when evaluated according to OSA severity. When the single channel (C3-M2) is used, the model achieves 79.2% and 72.0% accuracy in normal severity and severe severity, respectively. Also, when multiple channels (F3-M2, C3-M2, and E1-M2) are used, the model achieves higher 2.3% and 2.0% accuracies than those of the single channel case in normal severity and severe severity, respectively. From the results, we expect an accuracy increase by adopting more channels is seriously limited by the patient's OSA severity.

It is noteworthy that the performance of N1 increases, and the performance of N3 decreases when the severity is severe. In general, the performance of N3 is higher than N1. However, as shown in Table III, the class accuracies of these N1 and N3 stages are observed in a different pattern when the target of a test dataset is limited to the case of severe OSA patients.

The reason for this abnormal observation comes from the distribution of our dataset used for training a model. As shown in Table I, in the training dataset, the ratios of N1 and N3 are 10.3% and 5.8% in normal severity, respectively. However, the ratios of N1 and N3 are significantly changed to 32.6% and 2.0% in severe severity.

V. CONCLUSION

Sleep stage classification is an essential task for diagnosing sleep disorders but it is time-consuming and labor-intensive. To solve this issue, automatic sleep stage classification is required and modern deep learning can be effectively applied to solve it. In this work, we used a deep learning technique for automatic sleep staging with multiple channels while considering OSA severity. We investigated the detailed impacts of the OSA severity on the accuracy performance of a modern deep learning model with single channel and multiple channels. In future work, based on the current results, we are trying to build a deep learning model and develop generalization techniques that can improve overall accuracy performance even in the case of OSA patients.

ACKNOWLEDGE

This work has been supported by the National Research Foundation through the Basic Science Research Program under Grant 2021R1F1A104796311 and was also supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health Welfare, Republic of Korea (grant number: HR21C0198040022).

REFERENCES

[1] Junming Zhang and Yan Wu. A new method for automatic sleep stage classification. *IEEE transactions on biomedical circuits and systems*, 11(5):1097–1110, 2017.

[2] Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, 2017.

[3] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. Seqsleepnet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(3):400–410, 2019.

[4] Emadeldeen Eldele, Zhenghua Chen, Chengyu Liu, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. An attention-based deep learning approach for sleep stage classification with single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:809–818, 2021.

[5] Alexandra-Maria Tăutan, Alessandro C Rossi, Ruben de Francisco, and Bogdan Ionescu. Automatic sleep stage detection: a study on the influence of various psg input signals. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 5330–5334. IEEE, 2020.

[6] Richard B Berry, Claude L Albertario, Susan M Harding, Robin M Lloyd, David T Plante, Stuart F Quan, Matthew M Troester, and Bradley V Vaughn. *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*. American Academy of Sleep Medicine, 2018.