A CNN-Transformer Deep Learning Model for Real-time Sleep Stage Classification in an Energy-Constrained Wireless Device*

Zongyan Yao* and Xilin Liu†‡

*Engineering Science (EngSci), †Department of Electrical and Computer Engineering, ‡University of Toronto ‡University Health Network (UHN), Toronto, ON, Canada

Abstract—This paper presents a lightweight deep learning (DL) model for classifying sleep stages based on single-channel EEG. The DL model was designed to run on energy- and memory-constrained devices for real-time operation with all processing on the edge. Four convolutional filter layers are used to extract features and reduce the data dimension, and transformers were utilized to learn the time-variant features of the data. EEG recordings from a publicly available dataset (Sleep-EDF) are used to train and test the model. Subject-specific training was implemented to improve model performance. The testing F1 score was 0.91, 0.37, 0.84, 0.877, and 0.73 for the stages of awake, N1, N2, N3, and rapid eye movement (REM), respectively. The performance of the model was comparable to the state-of-the-art works with significantly greater computational costs. A reduced-size version of the model has been successfully tested on a low-cost Arduino Nano 33 BLE board. This design holds great promise for future integration into a low-power wireless EEG sensor with edge DL for sleep research in pre-clinical and clinical experiments, such as real-time sleep modulation.

I. INTRODUCTION

Sleep quality and health are closely related; therefore, it is important to understand sleep quality to improve health conditions. Sleep consists of five stages, which are awake, N1, N2, N3, and REM. Most of the sleep occurs between stages N1 and N3 [1]. The clinical evaluation of sleep stages is typically performed by polysomnogram (PSG), a procedure that records electroencephalogram (EEG), electrooculogram (EOG), and other physiological features. Currently, clinicians have to manually identify sleep stages over time according to biomarkers in the PSG.

With the development of electronic technology and machine intelligence, wearable devices, such as smartwatches, can measure user biosignals and classify sleep stages. However, the cost of these devices is high and the performance of sleep stage classification is limited. In addition, classification often requires the transmission of data to mobile phones or the cloud, raising concerns about cybersecurity [2], [3]. High-quality real-time sleep classification and sleep modulation still need to be conducted in sleep laboratories with human in the loop. Given the prevalence of sleep disorders, there is a compelling need for low-cost at-home sleep monitoring devices that can perform sleep stage classification on device. These devices can potentially generate intervention signals to modulate sleep stages, treat sleep disorders, and improve sleep benefits [4].

In this work, a lightweight DL model is developed for real-time operation on devices with restricted energy and memory, such as microcontrollers [5]. There are two main constraints to the development of the model for hardware. The first constraint is that the size of the model will be limited by the memory resources available on the hardware, including the non-volatile Flash memory for model storage and the random-access memory (RAM) for model computing [6]. The second constraint is that the computational demand of the model will be limited by the clock rate, bit width, and computational capabilities (such as floating point or fixed point) of the device [7]. Key trade-offs are between model performance and complexity. Here, we developed a DL model that can run on a low-power wireless microcontroller, based on a low-cost Arduino Nano 33 BLE development board. Despite its small size, our model achieved performance comparable to the state-of-the-art during validation using a publicly available sleep dataset (Sleep-EDF [8]).

Our DL model demonstrates a working prototype that can be easily reproduced and used to construct a fully integrated sleep stage classification and modulation device. Fig. 1 shows the envisioned block diagram of such a device featuring the developed DL model. The peripheral modules include EEG amplifiers and wireless communication modules, which are commercially available. The assembled device can be

Fig. 1. An overview of the envisioned wireless device for real-time sleep stage classification using edge DL. This paper focuses on the development and deployment of the DL model.
miniature in size and fully self-contained, which can enable a wide range of sleep research in pre-clinical and clinical studies.

II. METHODS

A. Dataset

Sleep-EDF Expanded Database (version 1, published in 2013) contains 197 whole-night PSG sleep recordings [8], [9]. It contains two subsets, the Sleep Cassette Study (SC) and Sleep Telemetry Study (ST). The SC Study was conducted from 1987 to 1991 to study the effects of age on sleep in healthy Caucasians. Data from the ST Study was obtained in 1994 to study the effects of temazepam on sleep. The experiments of this work were performed on the SC dataset.

Two 20-hour PSG recordings were taken for 77 subjects between the age of 25 and 101. The first nights of subjects 36 and 52, and the second night of subject 13 were lost. The PSG recordings contain three channels of EEG signals, one channel of EOG and chin EMG signals, oronasal airflow, rectal body temperature, and event marker. Only EEG signal from the Fpz-Cz channel is selected as the input of the developed model. The EEG signal was sampled at 100 Hz. Each of the 30-second segments of the signals was labeled by sleep experts. Eight labels are used in the dataset, N1, N2, N3, N4, Wake, REM, MOVEMENT, UNKNOWN). To make our results consistent and comparable with previous studies [10]–[12], we preprocessed the data with the following methods:

1) Discarded the segments with UNKNOWN and MOVEMENT labels.
2) Combined N4 and N3 together as N3 stage.
3) Ignored wake epochs longer than 30 minutes outside of sleep periods.

B. Performance Metrics

We evaluated the performance of the developed model using per-class precision (PR), per-class recall (RE), per-class F1 score (F1) and overall precision (acc). Overall accuracy is the ratio between the number of correct predictions and the population. For a category prediction, there are four outcomes: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The metrics are defined as:

\[ PR = \frac{TP}{TP + FP} \]  \hspace{1cm} (1)
\[ RE = \frac{TP}{TP + FN} \] \hspace{1cm} (2)
\[ F1 = \frac{2TP}{2TP + FN + FP} \] \hspace{1cm} (3)

Table I shows the distribution of the dataset, which is highly imbalanced. Overall accuracy is commonly used to measure classification performance. However, for an imbalanced dataset, precision does not provide adequate information on classifiers, because it hardly reveals performance in minority groups [13]. Therefore, we used the additional metrics, PR, RE, and F1, to comprehensively evaluate the performance of our model.

![Table I: Data Distribution](image)

C. Proposed Model

EEG signal contains time-invariant and time-variant features. With a sampling rate of 100 Hz, each 30-second input data segment results in an input shape of (3000, 1), which is too large for a transformer unit. Fig. 2 shows the architecture and key parameters of the proposed model. A convolutional neural network is used to extract time-invariant data and output smaller data. Four sequential convolutional layers are used to extract features with shape (19, 128). Then, a transformer unit is added to learn time-variant information from the features. Its attention mechanism learned the contexts of all positions of the time series data. The two dense layers inside the transformer unit work as an encoder. The output of the encoder is then added to the input data for additional features. Finally, to correctly classify sleep scores, a dense layer is used with a softmax activation function to obtain the most possible categories. Experimental results showed that this architecture yielded the best performance compared to commonly used recurrent neural networks and auto-encoders.

![Fig. 2: Architecture of the proposed CNN-Transformer DL model for real-time sleep stage classification using single channel EEG signal.](image)

III. EXPERIMENTS

A. Data Preprocessing

Given that the DL model is developed for devices with limited memory and computational resources, complex data preprocessing is prohibitive. As a result, only a simple standardization is used. The performance of DL models is highly dependent on the statistical properties of the input data. If the input data is too small or too large, the models perform poorly. Standardization transforms the EEG signal into a vector with zero mean and a standard deviation of one. For each sample \( X \), we standardized it to \( Z \) using

\[ Z = \frac{X - M}{S} \]  \hspace{1cm} (4)
where M represents the mean of the samples and S represents the standard deviation of the samples. After standardization, the input data was in the range of $10^{-1}$ to $10^1$.

### B. Basic Training

A 5-fold cross-validation was used to train and test the developed model. Since there are 77 subjects in total, each fold contains 16 subjects’ data except the last one which has 13 subjects. In each iteration, four folds were used for training and validation, and one fold was used for testing. Among the four folds of data, we randomly sampled 10% for validation and used the remaining data for training. Adam algorithm was used for optimization. Adam algorithm is a stochastic gradient descent method based on adaptive estimation of first-order and second-order moments, which computes efficiently and requires little memory [14]. Categorical cross-entropy was used as the loss function and fed the model with a batch size of 64 samples.

### C. Subject-Specific Training

There are many unique features embedded inside individual subjects’ EEG signals, so we performed subject-specific training in the testing stage to further adapt the patterns of each subject. After learning from the common features among subjects during the training stage, our model would extract subjects’ unique features from the subject-specific training. During the testing stage, we first randomly selected 10% of the test data and fed them to our trained model. The remaining 90% of the test data were used to test our models after subject-specific training.

### IV. RESULTS

Table II and Table III show the confusion table and the performance of the developed model, respectively. Due to the known data imbalance issue, the per-class performance in N1 and REM was worse than the rest of the classes. The model performed well in the awake and N2 classes, similar to the prior arts. Table VI shows that all models have an F1 score per class less than 0.5 in N1 and less than 0.8 in REM. To improve the performance of the developed model, we performed subject-specific training to further adapt subject-wise patterns. Table IV and Table V show that performance improved after subject-specific training. Firstly, the overall accuracy increased from 0.775 to 0.795. Secondly, per-class precision, recall, and F1 score increased for all classes.

### TABLE II

| Actual/Predict | Wake | N1 | N2 | N3 | REM |
|----------------|------|----|----|----|-----|
| Wake           | 0.90 | 0.04 | 0.01 | 0.01 | 0.04 |
| N1             | 0.20 | 0.26 | 0.31 | 0.01 | 0.21 |
| N2             | 0.01 | 0.04 | 0.82 | 0.07 | 0.06 |
| N3             | 0.05 | 0.19 | 0.30 | 0.12 | 0.04 |

Table VI compares the F1 score of different models from the literature using the same SleepEDF dataset. The difference in model architecture and the use of patient-specific training distinguishes our approach. For the awake stage, we yielded the highest F1-score of 0.91. For other classes, we were close to the highest. Some models performed better in certain classes, but there is no model that could beat our performances in all classes. Despite the fact that the performance of our model does not outperform all prior work, it has advantages in size and computational cost.

### V. DISCUSSION

Our lightweight model was designed for memory-constrained microcontrollers. It had around 300,000 parameters with a total size of 2 MB. We tested the model using an Arduino Nano 33 BLE board, which integrates a wireless microcontroller (nRF52840, Nordic Semiconductor) with a 64 MHz 32-bit ARM CPU, 1 MB of flash memory, and 256 KB of SRAM [19]. Since our current model size is bigger than the on-device flash memory, a reduced size version of the model was used for testing. The inference of the model using the microcontroller was fully successful with a degraded accuracy of 68% due to model reduction, as expected. In the future, we will add more flash memory to the device in order to support the size of the developed CNN-Transformer model.

We randomly selected 10% of the test data set to perform subject-specific training. If a dataset is sufficiently large, we expect that the randomly selected data should follow the distribution of the dataset. In our experiments, the distribution varied because the dataset was not large enough. As a result, the improvement in performance depends on the distribution of...
the selected data. In the future, we will enforce the distribution of subject-specific training data, which means that the number of selected data in each category will be calculated and fixed based on the test dataset. This method would maximize the benefits of subject-specific training.

As indicated by the results, a major limitation of the performance of the model is the imbalanced class distribution, which particularly affects the minority categories. There are different methods to mitigate the issue of class imbalance, including oversampling and undersampling. Oversampling is used to duplicate samples in minority classes, and undersampling is used to remove samples from the majority classes. In the future, we will also try to add weights to the loss function. The loss function can be weighted differently for different classes, so that minority classes are learned more.

VI. CONCLUSION

In this work, we developed a lightweight DL model for real-time sleep stage classification using single-channel EEG data. The DL model features CNN and transformers. We validated the model using the Sleep-EDF dataset and tested it in a low-power microcontroller. The model achieved performance comparable to the state-of-the-art works. In the future, we plan to develop a fully integrated wireless EEG sensor using the model. The developed device can potentially enable a wide range of novel sleep research paradigms.

REFERENCES

[1] A. K. Patel, V. Reddy, and J. F. Araujo, “Physiology, sleep stages,” in StatPearls [Internet]. StatPearls Publishing, 2022.
[2] X. Liu, A. G. Richardson, and J. Van der Spiegel, “An energy-efficient compressed sensing-based encryption scheme for wireless neural recording,” IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 11, no. 2, pp. 405–414, 2021.
[3] X. Liu, H. Zhu, T. Qiu, S. Y. Sritharan, D. Ge, S. Yang, M. Zhang, A. G. Richardson, T. H. Lucas, N. Engheta et al., “A fully integrated sensor-brain–machine interface system for restoring somatosensation,” IEEE Sensors Journal, vol. 21, no. 4, pp. 4764–4775, 2020.
[4] X. Liu and A. G. Richardson, “A system-on-chip for closed-loop opticogenic sleep modulation,” in 2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2021, pp. 5678–5681.
[5] X. Liu, B. Subei, M. Zhang, A. G. Richardson, T. H. Lucas, and J. Van der Spiegel, “The pennmbni: A general purpose wireless brain-machine-brain interface system for unrestrained animals,” in 2014 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2014, pp. 650–653.
[6] X. Liu and A. G. Richardson, “Edge deep learning for neural implants: a case study of seizure detection and prediction,” Journal of Neural Engineering, vol. 18, no. 4, p. 046034, 2021.
[7] X. Liu, J. Sacks, M. Zhang, A. G. Richardson, T. H. Lucas, and J. Van der Spiegel, “The virtual trackpad: An electromyography-based, wireless, real-time, low-power, embedded hand-gesture-recognition system using an event-driven artificial neural network,” IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 64, no. 11, pp. 1257–1261, 2016.
[8] B. Kemp, A. Zwinderman, B. Tuk, H. Kamphuisen, and J. Oberey, “Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity hypothesis,” IEEE Transactions on Biomedical Engineering, vol. 47, no. 9, pp. 1185–1194, 2000.
[9] A. L. Goldberger, L. A. 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.
[10] A. Supratak, H. Dong, C. Wu, and Y. Guo, “Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg,” IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 25, no. 11, pp. 1998–2008, 2017.
[11] W. Qu, Z. Wang, H. Hong, Z. Chi, D. D. Feng, R. Grunstein, and C. Gordon, “A residual based attention model for eeg based sleep staging,” IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 10, pp. 2833–2843, 2020.
[12] N. Goshtasbi, R. Boostani, and S. Sanei, “Sleeppfn: A fully convolutional deep learning framework for sleep stage classification using single-channel electroencephalograms,” IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 30, pp. 2088–2096, 2022.
[13] H. He and E. A. Garcia, “Learning from imbalanced data,” IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009.
[14] D. P. Kingma, “Adam: A method for stochastic optimization/diederik p;” Kingma, Jimmy Ba, URL: https://arxiv.org/abs/1412.6980.
[15] S. Chang, Z. Yang, Y. You, and X. Guo, “Dssnet: A deep sequential sleep network for self-supervised representation learning based on single-channel eeg,” IEEE Signal Processing Letters, vol. 29, pp. 2143–2147, 2022.
[16] H. Phan, F. Andretti, N. Cooray, O. Y. Chen, and M. De Vos, “Dnn filter bank improves 1-max pooling cnn for single-channel eeg automatic sleep stage classification,” in 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2018, pp. 453–456.
[17] S. Mousavi, F. Afghah, and U. R. Acharya, “Sleepprognost: Automated sleep stage scoring with sequence to sequence deep learning approach,” PloS one, vol. 14, no. 5, p. e0216456, 2019.
[18] C. A. Ellis, R. L. Miller, and V. D. Calhoun, “A novel local explainability approach for spectral insight into raw eeg-based deep learning classifiers,” in 2021 IEEE 21st International Conference on Bioinformatics and Biomedicine (BIBM), 2021, pp. 1–6.
[19] A. Kurniawan, “IoT projects with arduino nano 33 ble sense;” Berkeley: Apress, vol. 129, 2021.