Exploration of Deep Learning Approaches for Sleep Staging on a Multimodal Wearable Bio-medical Device

Andrew H. Zhang
Department(s) of Computer Science; Physics
University of Toronto
andrewhz.zhang@mail.utoronto.ca

Chunlin Li
Department of Computer Science
University of Toronto
edwardphi.li@mail.utoronto.ca

Yuzhi Tang
Department of Computer Science
University of Toronto
yuzhi.tang@mail.utoronto.ca

Abstract

We present a pilot study exploring deep learning approaches for sleep staging using the ANNE One — a multi-modal wearable device that measures ECG, PPG, tri-axial accelerometry, and core body temperature — with a newly available data set provided by the Sunnybrook research institute. Employing a Convolution then Recurrent Neural Network architecture, the impact of recurrent layer architecture and using different engineered feature groups on performance is investigated. See appendix C for our code repository on GitHub.

1 Introduction

1.1 Background

Sleep staging is the categorization of different stages of sleep based on time series physiological signals collected during sleep. Its clinical significance comes from its pivotal role in the diagnosis of sleep disorders and neurodegenerative diseases such as Alzheimer's disease. The gold standard for sleep staging is through a diagnostic process called Polysomnography (PSG), which records physiological parameters such as brain activity, heart rate, eye movements, muscle tone, and breathing patterns during sleep. Though highly accurate, performing sleep staging via PSG is a laborious task requiring suites of specialized equipment and highly trained professionals, which often makes the cost of performing PSG prohibitively high. Additionally, performing PSG sleep staging requires the patient to be physically present at a sleep clinic for the duration of the process, which makes mass clinical studies involving sleep staging unscalable. If sleep staging can be accurately performed through a wearable bio-medical device, we can overcome the aforementioned constraints imposed by PSG for conducting mass clinical studies. This helps to identify those at risk for neurodegenerative diseases early on, which could help to conserve clinical resources and suggest preventative treatments before the diseases' onset or deterioration.

Identified as a promising alternative to PSG, the ANNE One [10] developed by Sibel Health is an FDA-approved multimodal wearable bio-medical device that records a continuous stream of various physiological data. The ANNE One contains a finger module and a chest module. The chest module measures ECG, tri-axial accelerometry, and core body temperature. The finger module measures PPG. While sleep staging via PSG relies mainly on its EEG channels, the ANNE One has no sensor for EEG, which makes sleep staging via the signals provided by the ANNE One an interesting machine learning problem to solve.

Preprint. Under review.

1.2 Data Set

Since the ANNE One only received its 510(k) FDA clearance in 2021, data sets of ANNE One recordings for sleep staging are scarce. Courtesy of the sleep and brain health laboratory at the Sunnybrook research institute, this study is provided access to a novel data set of 43 simultaneous ANNE One and PSG full night recordings from different individuals, which totals at roughly 294 hours of time series data. The PSG-derived sleep staging ground truth labels are annotated for every consecutive 30 second window. We perform sleep staging using 3 classes: wakefulness (Wake) (30% of the data set), non rapid eye movement sleep (NREM) (56% of the data set), and rapid eye movement sleep (REM) (14% of the data set). This amounts a total of 35383 labeled time windows. All ANNE One channels are provided at a resolution of 25 Hz. In addition to the raw channels of ECG, PPG, tri-axial accelerometry, and core body temperature, the data set also comes with additional engineered time domain channels including ECG-derived heart rate, as well as the z-angle and Euclidean norm minus one (ENMO) of tri-axial accelerometry.

1.3 Related Works

As ANNE One is a relatively new product, we have not found any previous works using deep learning for sleep staging on the ANNE One specifically. Instead, we will survey the most successful deep learning strategies for sleep staging in general.

In 2017, Supratak et al. [9] proposed DeepSleepNet, which used Convolutional Neural Networks (CNNs) to extract time-invariant features from EEG signals. A bidirectional-Long Short-Term Memory (LSTM) model was used to learn transition rules among sleep stages. They achieved 0.82 accuracy and 76.9 macro F1 on the Sleep-EDF dataset. In 2022, Phan et al. [6] proposed a transformer-based architecture for sleep stage classification using EEG signals. They achieved 0.85 accuracy and 78.8 macro F1 on the SleepEDF-78 dataset. It should be noted that the SleepEDF data set has 7 classes including a movement class while decomposing NREM into sub-classes N1 to N4.

As an approach not using EEG, in 2019 Sun et al. [8] used CNNs to extract features from ECG and chest/abdominal respiratory effort. A bi-directional LSTM was used to learn temporal patterns among consecutive time windows. Using 8682 PSGs at the Sleep Laboratory at Massachusetts General Hospital, they achieved 0.681 macro F1 for 5 sleep stages, and 0.842 macro F1 for 3 sleep stages.

2 Approaches and Hypothesis

As seen from section 1.3, some of the most successful strategies employ CNN layers as feature encoders within each labeled time windows and RNN layers such as LSTM to capture the temporal relationship between each labeled time window. We will employ the same strategy to first use CNN to encode features, and then RNN to capture temporal relationship. More specifically, for the RNN portion of our model, we will explore the difference of using a relatively simpler architecture, namely the gated recurrent unit (GRU), and the LSTM architecture.

Given the limited size of our data set, and since the use of human-engineered features is prolific in earlier studies [2,3,4], we will engineer our additional feature groups from the data set as extra inputs into our model instead of using many large CNN layers. This process is outlined in more detail in section 3.1. We will also explore how the addition of different combination of engineered feature groups influence the model's performance. Overall, we expect the LSTM to outperform the GRU as it is a more expressive model, and we also expect the model to perform better the more engineered feature we include as inputs.

3 Method

3.1 Preprocessing and Feature Engineering

Out of all the provided time domain channels, we selected 6 as inputs into our model, namely: ECG, PPG, tri-axial accelerometry ENMO and z-angle, core body temperature, and heart rate.

From the provided time domain channels, we derived two additional feature groups: The frequency spectra feature group and the scalar feature group.

For the frequency spectra feature group, we took the frequency spectra of each 30 second window of ECG, PPG, x, y, z accelerometry, and accelerometry z-angle. For each 30 second labelled time domain series, this is done by first detrending if neccesary, applying a Hanning window, applying FFT, and then truncating the resulting frequency spectra at the Nyquist frequency.

For the scalar feature group, we included for each 30 second labelled window the ECG-derived heart rate variability (HRV), variance of the ECG frequency spectra, skew of the PPG frequency spectra, skew of the accelerometry z-angle, standardized average core body temperature, and the kurtosis of accelerometry ENMO. HRV is calculated as the ratio between the integral of ECG frequency spectra from 0.04 to 0.15 Hz (LF) and the integral of ECG frequency spectra from 0.15 to 0.4 Hz (HF).

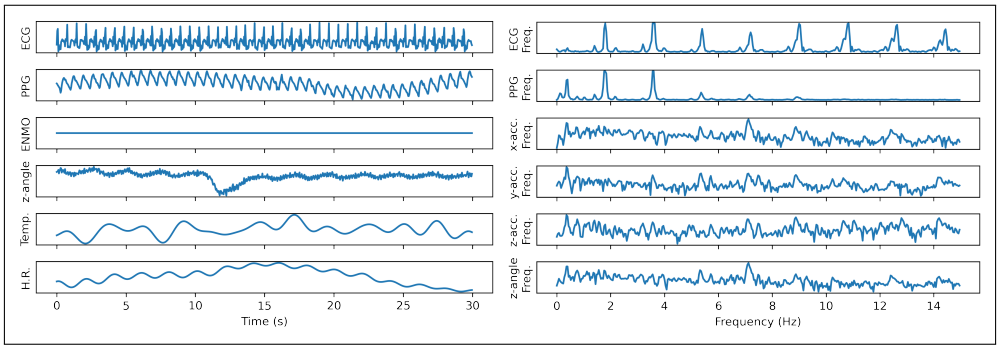


Figure 1: Time and Frequency domain feature groups for a labeled 30 second window. The class label is Wake. The corresponding scalar feature group values are, in the same order that they are reported above: `[1.0014439, 1.6034065, 8.994368, 0.5242344, -0.8302446, -3.]`

3.2 Architecture

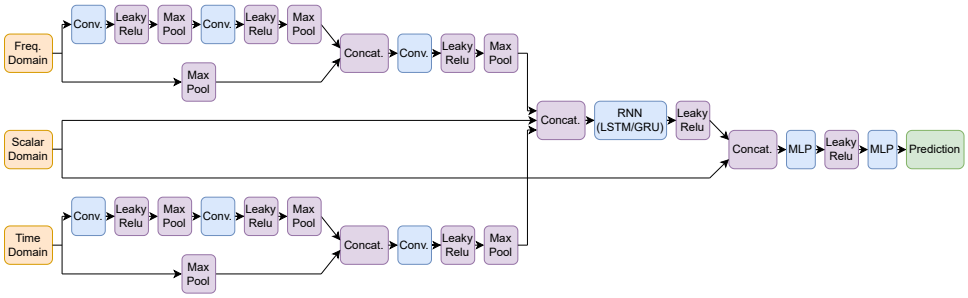


Figure 2: Architecture diagram. The orange boxes are the input data. The blue boxes are layers with trainable weights while the purple boxes don't have any weights. The green box is the three dimensional output vector. Note that our RNN model, LSTM or GRU, is bidirectional

Figure 2 shows the architecture with all three feature groups, time domain, scalar domain, and frequency domain. In our study, we tested four combinations of feature group setups. For each feature group not present in a setup, we remove all inputs, convolution layers, and residual connections from the model as demonstrated in Figure 2 and adjust the dimension of each layer accordingly. For a more detailed view of the model on Figure 2, see appendix A.

Table 1: Number of Trainable Parameters for Each Training Setup

	Time Only	Time+Freq.	Time+Scalar	Time+Freq.+Scalar
GRU	4376	11648	5048	12320
LSTM	5144	14336	6008	15200

3.3 Training Setup

As mentioned in section 1.2, there is a significant class imbalance in the data set that we use. Since we hope to learn temporal relationship between each class through recurrent layers, we could not just artificially re-balance the data set by throwing out excess classes. We thus augmented our loss function by using weighted cross-entropy loss to make it more rewarding to correctly classify one of the rarer classes. Specifically, for a single prediction:

$$\ell_{WCE} = -\sum_{i \in C} \left(\frac{1}{f_i}\right)^{\alpha} t_i \log(p_i)$$

where C is the set of classes, f_i is the frequency at which the class i occurs in the data set, and p_i is the softmax probability of each class, and t_i is the truth label value of the prediction each class. α is a parameter used to further tune the class weight of $\frac{1}{f_i}$. After running small-scale experiments, we we determined that a value of $\alpha = 1.25$ provides the best separation of classes. Out of the 43 subjects, we used a train-validation-test ratio of 37:3:3, drawn at random. We used the Adam optimizer [5] and the cyclical learning rate scheduler (triangular2 policy) [7] with a base learning rate of 0.0001, a max learning rate of 0.01, and a step up size of 200 iterations. We used early return by monitoring the validation loss and a patience of 50 epochs. We used a batch size of 4096. All trainable weights are initialized with the Xavier initialization [1]. Each model is trained for upwards of 1000 epochs if not early-stopped.

4 Results, Conclusion, and Further Studies

Table 2: F1 Scores for Each Training Setup
(see appendix B for confusion matrices)

	Time Only	Time+Freq	Time+Scalar	Time+Freq+Scalar
F1 Score for Wake				
GRU	0.68	0.82	0.56	0.73
LSTM	0.63	0.75	0.81	0.76
F1 Score for NREM				
GRU	0.51	0.47	0.48	0.57
LSTM	0.58	0.43	0.55	0.54
F1 Score for REM				
GRU	0.10	0.47	0.37	0.35
LSTM	0.17	0.21	0.42	0.45
Macro F1 Score				
GRU	0.43	0.59	0.47	0.55
LSTM	0.46	0.47	0.59	0.59

We trained our model using GRU and LSTM for four multimodal setups. Then, we computed F1 score for the three sleep stage and computed macro F1 score for all three classes. From Table 2, using F1 score as a metrics of performance (where higher F1 score means better performance), we conclude that our multi-model approach outperforms the model with one feature group (i.e. time domain only). It points out that the multi-model approach should be further investigated. The LSTM outperforms GRU in general, which supports our hypothesis. Moreover, the table indicates that our idea of incorporating frequency domain and time domain features to the model works and improves overall performance. Nonetheless, our best result for three sleep stages reaches macro F1 score 0.59 and does not approach the macro F1 score 0.842 achieved by Sun et al. [5] There are two reasons. First, they uses 8682 PSGs while we use PSGs and ANNE signals from 43 patients. Second, they use ECG from PSG signals while we used a wearable device, which has poorer signal quality.

Although we have achieved incremental progress, the result indicates there is still great potential for our study. Specifically, the currently performance of our model on predicting REM sleep is not optimal. This might be due to small proportion of REM samples in our training data, and one solution is to collect more data with ANNE. Also, given that LSTM with more parameters outperforms GRU in general, it's likely that increasing training weight would further improve the performance. So, larger models like Transformer should be tested. This work is a preliminary exploration for future large studies. Our dataset currently consists of data of 43 patients, and the next step is to expand the dataset and try larger models to uncover more hidden patterns.

References

[1] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, 249–256. Retrieved April 19, 2023 from https://proceedings.mlr.press/v9/glorot10a.html

[2] Shenda Hong, Yuxi Zhou, Junyuan Shang, Cao Xiao, and Jimeng Sun. 2020. Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. Comput. Biol. Med. 122, (July 2020), 103801. DOI:https://doi.org/10.1016/j.compbiomed.2020.103801

[3] Wu Huang, Bing Guo, Yan Shen, Xiangdong Tang, Tao Zhang, Dan Li, and Zhonghui Jiang. 2020. Sleep staging algorithm based on multichannel data adding and multifeature screening. Comput. Methods Programs Biomed. 187, (April 2020), 105253. DOI:https://doi.org/10.1016/j.cmpb.2019.105253

[4] S. Khalighi, T. Sousa, D. Oliveira, G. Pires, and U. Nunes. 2011. Efficient feature selection for sleep staging based on maximal overlap discrete wavelet transform and SVM. 2011 Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (August 2011), 3306–3309. DOI:https://doi.org/10.1109/IEMBS.2011.6090897

[5] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. Retrieved April 19, 2023 from http://arxiv.org/abs/1412.6980

[6] Huy Phan, Kaare Mikkelsen, Oliver Y. Chén, Philipp Koch, Alfred Mertins, and Maarten De Vos. 2022. SleepTransformer: Automatic Sleep Staging With Interpretability and Uncertainty Quantification. IEEE Trans. Biomed. Eng. 69, 8 (August 2022), 2456–2467. DOI:https://doi.org/10.1109/TBME.2022.3147187

[7] Leslie N. Smith. 2017. Cyclical Learning Rates for Training Neural Networks. DOI:https://doi.org/10.48550/arXiv.1506.01186

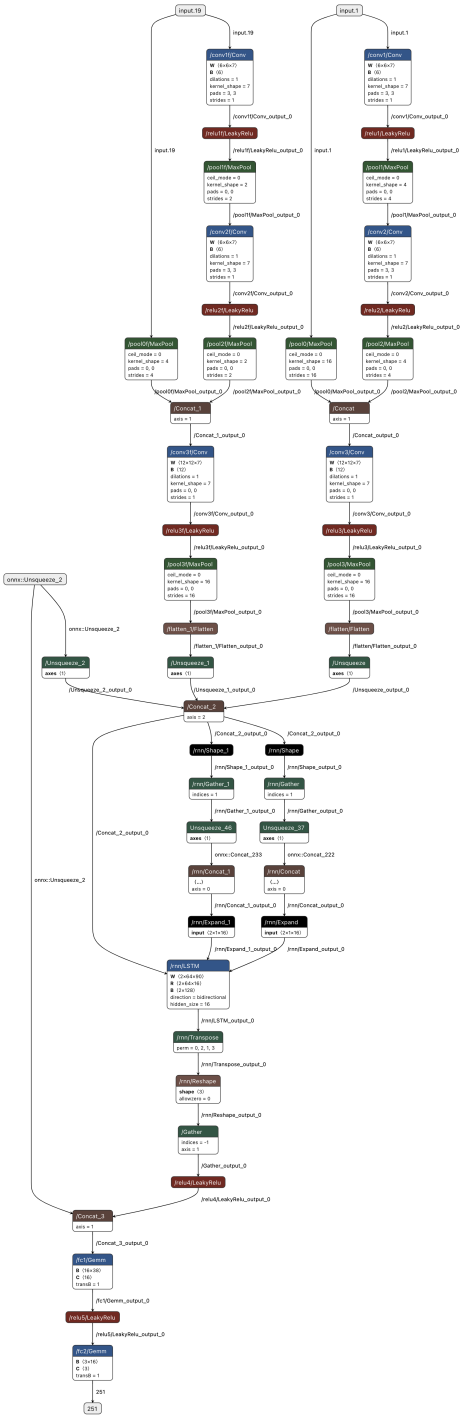
[8] Haoqi Sun, Wolfgang Ganglberger, Ezhil Panneerselvam, Michael J. Leone, Syed A. Quadri, Balaji Goparaju, Ryan A. Tesh, Oluwaseun Akeju, Robert J. Thomas, and M. Brandon Westover. 2020. Sleep staging from electrocardiography and respiration with deep learning. Sleep 43, 7 (July 2020), zsz306. DOI:https://doi.org/10.1093/sleep/zsz306

[9] Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. 2017. DeepSleepNet: A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG. IEEE Trans. Neural Syst. Rehabil. Eng. 25, 11 (November 2017), 1998–2008. DOI:https://doi.org/10.1109/tnsre.2017.2721116

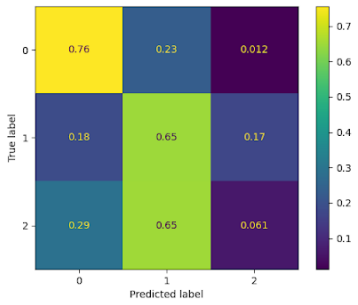
[10] ANNE One — Sibel. Sibel (Copy). Retrieved April 19, 2023 from https://www.sibelhealth.com/anne-one

A Detailed View of Architecture

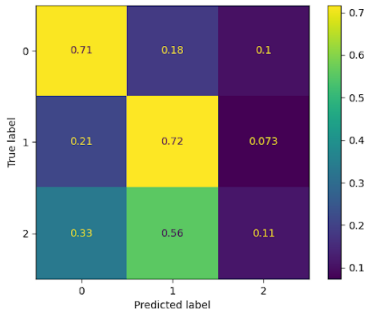
Below is a visualization of our model.



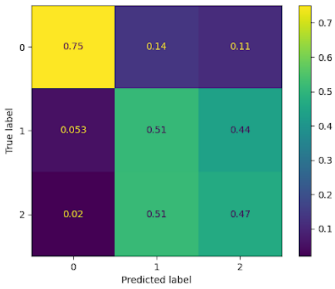
B Confusion matrix for each setup



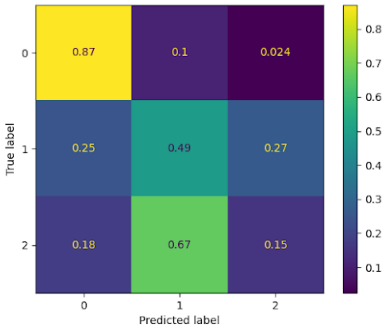
(a) GRU Time Domain Only



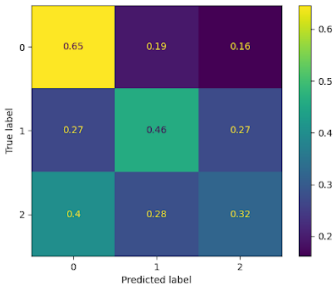
(b) LSTM Time Domain Only



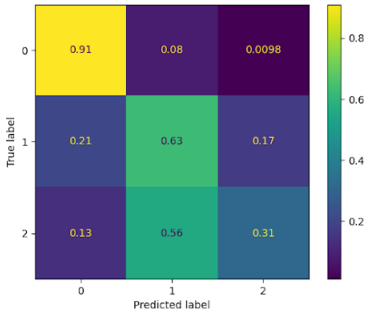
(c) GRU Time + Frequency Domain



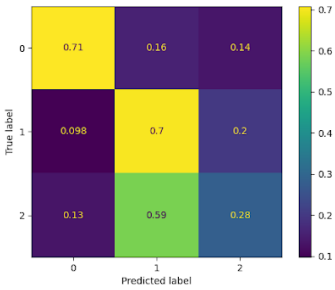
(d) LSTM Time + Frequency Domain



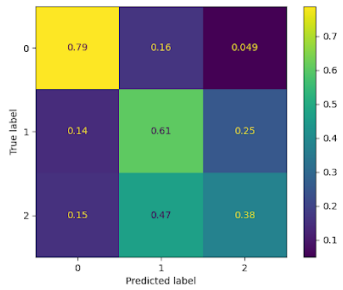
(e) GRU Time + Scalar Domain



(f) LSTM Time + Scalar Domain



(g) GRU Time, Frequency, & Scalar Domain



(h) LSTM Time, Frequency, & Scalar Domain

C GitHub Repository

https://github.com/a663E-36z1120/ANNE_dl/