

EEG aided boosting of single-lead ECG based sleep staging with Deep Knowledge Distillation

Vaibhav Joshi

*Department of Electrical Engineering
IIT Madras
Chennai, India
ee19s039@smail.iitm.ac.in*

Preejith SP

*Healthcare Technology Innovation
Innovation Center, IIT Madras
Chennai, India
preejith@htic.iitm.ac.in*

Sricharan V

*Healthcare Technology Innovation
Innovation Center, IIT Madras
Chennai, India
sricharany@htic.iitm.ac.in*

Mohanasankar Sivaprakasam

*Department of Electrical Engineering
IIT Madras
Chennai, India
mohan@ee.iitm.ac.in*

Abstract—An electroencephalogram (EEG) signal is currently accepted as a standard for automatic sleep staging. Lately, Near-human accuracy in automated sleep staging has been achievable by Deep Learning (DL) based approaches, enabling multi-fold progress in this area. However, An extensive and expensive clinical setup is required for EEG based sleep staging. Additionally, the EEG setup being obtrusive in nature and requiring an expert for setup adds to the inconvenience of the subject under study, making it adverse in the point of care setting. An unobtrusive and more suitable alternative to EEG is Electrocardiogram (ECG). Unsurprisingly, compared to EEG in sleep staging, its performance remains sub-par. In order to take advantage of both the modalities, transferring knowledge from EEG to ECG is a reasonable approach, ultimately boosting the performance of ECG based sleep staging. Knowledge Distillation (KD) is a promising notion in DL that shares knowledge from a superior performing but usually more complex teacher model to an inferior but compact student model. Building upon this concept, a cross-modality KD framework assisting features learned through models trained on EEG to improve ECG-based sleep staging performance is proposed. Additionally, to better understand the distillation approach, extensive experimentation on the independent modules of the proposed model was conducted. Montreal Archive of Sleep Studies (MASS) dataset consisting of 200 subjects was utilized for this study. The results from the proposed model for weighted-F1-score in 3-class and 4-class sleep staging showed a 13.40 % and 14.30 % improvement, respectively. This study demonstrates the feasibility of KD for single-channel ECG based sleep staging’s performance enhancement in 3-class (W-R-N) and 4-class (W-R-L-D) classification.

Index Terms—Sleep Staging, Deep Learning, Knowledge Distillation, EEG, ECG.

I. INTRODUCTION

Sleep is an intricate dynamic physiological process that occurs in multi-cyclical stages. In sleep medicine, sleep is typically studied by acquiring multiple bio-signals during sleep by conducting a polysomnography (PSG) study. The primary reference for sleep studies is accepted to be the Electroencephalogram (EEG) signal, considering its interpretability with brain activation, the pivot of sleep mechanism. Generally,

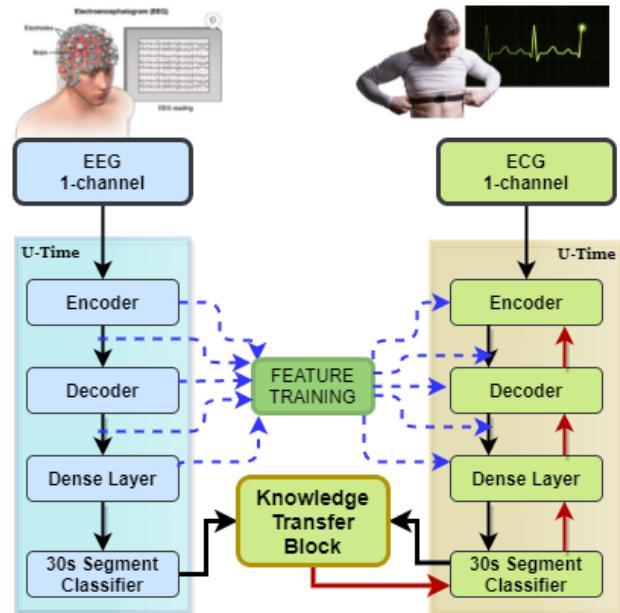


Fig. 1: Proposed Knowledge Distillation framework

experts manually perform sleep stage classification for 30/20 s epochs following the sleep staging guidelines and rules by Rechtschaffen and Kales (1968) (the ‘R and K rules’) [1] or AASM (American Academy of Sleep Medicine) [2]. As manual sleep staging is cumbersome and time-consuming, many automated sleep staging algorithms, including neural network-based approaches [3], have been developed lately with notable performance on par with human accuracy. Traditionally, sleep is divided into five stages: W, wakefulness; REM Rapid Eye Movement stage; N1, a light sleep period in Non-REM stages; N2, an intermediate stage; N3, a deep sleep stage. Different frequencies and patterns observed in the EEG signal during sleep characterize different sleep stages. The Color

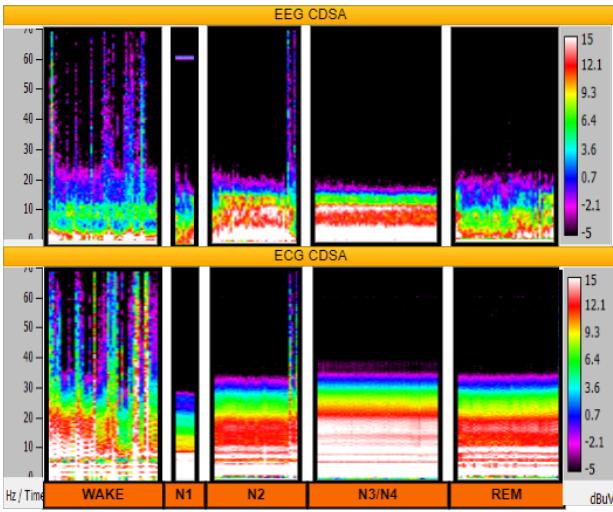


Fig. 2: Color Density Spectral Array-CDSA of EEG and ECG

Density Spectral Array (CDSA), as shown in Fig.2 shows EEG and ECG frequency components for different sleep stages. The patterns match in both signals, with the EEG pattern being more distinctive for sleep stages. However, in a non-clinical setup, the obtrusiveness of EEG renders it impractical. Furthermore, during sleep, the brain-body interaction implies that the stages of sleep can be captured by other physiological signals as well [4].

An electrocardiogram (ECG), not only being readily integrable into wearable devices but also less obtrusive, is an appropriate alternative to the EEG. As seen in Fig.2, the ECG pattern is not distinctive, So Deep Learning approach is a plausible choice. Q.Li *et al.* [5] extracted Respiratory Sinus Arrhythmia (RSA) and ECG Derived Respiration (EDR) based features from a vast PSG dataset (SHHS, CinC, SLPDB) for cross-spectral spectrogram, which helped to generalize the model better and achieved significant results for sleep staging using ECG signal. Their work employed an SVM, which was given high-level features extracted from the input using a Convolutional Neural Network(CNN). An accuracy of 65.90 % on SHHS and 75.40 % on SLPDB, respectively, were obtained for 4-class sleep staging. Radha *et al.* [6] used a temporal model approach, LSTM for sleep staging on 132 HRV features explicitly extracted from ECG signal and achieved an accuracy of 77.0 %. In another study, Fonseca *et al.* [7] extracted 80 expert features from Respiratory Inductance Plethysmography (RIP)and ECG eventually used by Linear Discriminant (LD), which obtained an accuracy of 80.0 % and 69.0 % for 3-class and 4-class sleep staging, respectively. In a separate study by Sridhar *et al.*, [8], 2 stage CNNs achieved an accuracy of 77.0 % from ECG-derived heart rate on an extensive dataset comprising MESA, CinC and SHHS. Despite the burgeoning potential of ECG, EEG based sleep classification performance remains vastly superior for sleep staging [3] [9]. It would be exceedingly beneficial to integrate the advantages of improved accuracy with unobtrusive monitoring. While multi-modal fusion methods utilize features of multiple signals and have

obtained improved accuracy [7] [9], multiple signal acquisition is required during inference which is an additional overhead. Thus, using the primary modality, EEG, to impart information to ECG opens up a lot of potential applications.

Knowledge Distillation (KD) has recently gained traction in deep networks to effectively transfer relevant information to more compact networks from extensive networks. A response based KD method to transfer and distil information to a student model from the softmax layer of a larger size teacher model to achieve better generalization was proposed by Hinton *et al.* [10]. Another approach that implemented feature-based distillation through knowledge distillation from the intermediate feature maps instead of the softmax layers was proposed by Romero *et al.* [11]. The concept to transfer attention maps from the intermediate layers to improvise the feature distillation process was introduced by Zagoruyko *et al.* [12]. Cross-modal KD [13] explored the above idea by applying it across modalities. Inspired by these approaches, a cross-modal KD approach, as depicted in Fig.1 is proposed. The proposed method combines feature-based and response based distillation facilitating multi-modal training of ECG and EEG while enabling uni-modal testing. This study is designed to evaluate the potency of KD in the performance enhancement of ECG based sleep staging. Moreover, we compare individual elements of the KD framework and the proposed model, subsequently exhibiting its potential.

II. METHODS

A. Problem Formulation

Let $s_{eeg} \in \mathbb{R}^{tf}$ and $s_{ecg} \in \mathbb{R}^{tf}$ be the EEG and ECG waveforms, with a sampling rate f for t seconds, respectively. Let $M(eeg; \phi_{eeg})$ and $M(ecg; \phi_{ecg})$ be the models which take T connected segments, each having length i , from s_{eeg} and s_{ecg} respectively. Let e be the frequency at which the signal is split, where the aim is to independently match s_{eeg} and s_{ecg} to $[e * t]$ labels, where each label is based on $i = f/e$ sampled points. The model we adapt is able to handle varying frequencies during inference. To be more specific, the model $M(eeg; \phi_{eeg})$ and $M(ecg; \phi_{ecg})$ maps s_{eeg} and s_{ecg} to ground truth for predicting C classes in all T segments.

Weighted Cross Entropy(WCE) is the chosen loss function that is to be optimized by the teacher model, $M(eeg; \phi_{eeg})$ and is defined as:

$$L(s_{eeg}, y) = \sum_{i=1}^T 1 / \sum_{i=1}^T (-w_{yi}) * l_{eeg}^i \quad (1)$$

where w_{yi} is the class weight which is proportional to the number of data points in the training set from a particular class and

$$l_{eeg}^i = (\log(\exp(s_{eeg}^{(i,y_i)}) * w_{yi})) / \sum_{k=1}^C \exp(s_{eeg}^{(i,k)}) \quad (2)$$

This is followed by Feature Based (FB) distillation of the aggregated attention maps obtained from the features of the pretrained teacher model, $M(eeg; \phi_{eeg})$, to the untrained student

model, $M(ecg; \phi_{ecg})$. To obtain optimal FB distillation, we adopt the following Attention Transfer (AT) loss [12]:

$$L_{FB} = \sum_{j \in I} \left\| \frac{Q_{eeg}^j}{\|Q_{eeg}^j\|_2} - \frac{Q_{ecg}^j}{\|Q_{ecg}^j\|_2} \right\|_2 \quad (3)$$

where the j -th pair of teacher and student attention maps is represented by $Q_{eeg}^j = \text{vec}(F_{eeg}(A_{eeg}^j))$ and $Q_{ecg}^j = \text{vec}(F_{ecg}(A_{ecg}^j))$ in a vectorized form, I denote the set of teacher-student convolution layers which is selected for Attention Transfer. In our framework, j distils the attention maps from all the layers by iterating through the features of the whole architecture. In AT, the main motive is to train a student network that, while being accurate, will also have attention maps that are similar to those of the teacher.

Following this, the pretrained student model $M(ecg; \phi_{ecgpre})$, optimizes the sum of the Response Based (RB) distillation and WCE loss defined as:

$$L(s_{ecg}, y) = (1 - \beta) * l_{ecg}^i * \sum_{i=1}^T 1 / \sum_{i=1}^T (-w_{yi}) + \beta T_d^2 * l_{dist}^i \quad (4)$$

where beta is the weight between RB and FB losses, T_s is the temperature parameter of softmax and

$$l_{ecg}^i = w_{yi} * \log(\exp(s_{ecg}^{(i,y_i)}) / \sum_{k=1}^C \exp(s_{ecg}^{(i,k)})) \quad (5)$$

$$l_{dist}^i = KLD(p(s_{ecg}, T_s), p(s_{eeg}, T_s)) \quad (6)$$

$$p(x, T_s) = \log(\exp(x^{(i,y_i)} / T_s) / \sum_{k=1}^C \exp(x^{(i,k)} / T_s)) \quad (7)$$

Where KLD denotes Kulback Leibler Divergence.

B. Architectural Details

The proposed model adopts the three module structure as proposed in the U-Time architecture [3] which includes an encoder, decoder and segment classifier. The encoder compresses the raw EEG or ECG signal into a group of subsampled feature maps, with the last layer acting as the bottleneck. It consists of five blocks where each block consists of two convolution subblocks followed by a max-pooling layer that subsamples the input by a factor of two. Finally, the bottleneck layer consists of two convolutions that retain the spatial size. The decoder is entasked to learn a mapping from the bottleneck layer back to the input signal domain giving out a dense segmentation map of the same size as the input. The resulting feature maps are concatenated with the corresponding feature maps computed by the encoder at the same scale. The five blocks in the decoder consist of two convolution subblocks, similar to the blocks in the encoder, followed by an upsampling layer. The segment classifier is fed the output segmentation map to predict the final sleep stages at the desired resolution.

The above architecture serves as the base for both the teacher, $M(eeg; \phi_{eeg})$ and the student, $M(ecg; \phi_{ecg})$. This architecture was chosen considering following major reasons:

- 1) **Fully convolutional architecture** - U-Time can be applied across any dataset without much architecture or hyperparameter tuning as it is fully convolutional.
- 2) **Customized for EEG** - U-Time was optimized to improve sleep staging from EEG. Picking this architecture will ensure that the best teacher model is chosen for feature transfer from EEG to ECG.
- 3) **Inference time variable-length segmentation** - A U-Time model may be used to stage sleep at any frequency, i.e., every 20 s or 40 s, at inference time.

C. Dataset Description

This study utilizes the Montreal Archive of Sleep Studies (MASS) [14] dataset containing sleep recordings obtained from 200 participants [103 females (aged 38.3 ± 18.9 years) and 97 males (aged 42.9 ± 19.8 years); age range: 18–76 years]; organized into five sets of PSG records, SS1-SS5. The dataset was acquired online from the Centre for Advanced Research in Sleep Medicine (CARSM) on providing the project proposal approved by the local ethics board. All participants were part of a healthy control group, except for 15 of the SS1 subset who suffered from Mild Cognitive Impairment (MCI). The study utilizes data from all the 200 subjects. Among the many EEG electrodes positioned as per the international 10-20 system, the C3-A2/C4-A1 electrodes were used. Lead 1 was the electrode of choice from the ECG. The data was undersampled to 200 Hz from the initial sampling rate to optimize runtime and uniformity while comfortably satisfying the Nyquist criterion. A window width of thirty seconds was uniformly considered for training and inference. Subsets with annotation for every twenty seconds were converted into thirty-second segments by including five seconds of data before and after the annotation. The sleep stages N1 and N2 were combined into Light Sleep(L) and N3, and N4 into Deep Sleep(D) for the four-class (W-L-D-R) classification problem. N1, N2, N3, and N4 were combined into a single NREM(N) stage for the three-class (W-N-R) classification problem.

D. Experimental Procedures

All the data was split subject-wise into train, eval and test sets in 80:10:10 ratio for both the classification problems. This ensured zero overlaps of data between splits from the same subject. The best model in all the runs was identified based on metrics tracked on the validation set during training. This was then validated on the holdout test-set post-training. The metrics used to validate the model's performance included accuracy and weighted F1 score [3]. The weighted-F1 score is calculated by computing the f1 score for each class separately and averaging the scores, weighted by individual class support (True positive + False Negative) to tackle the intrinsic imbalance in sleep staging. Experiments for both three and four class classification tasks were conducted identically as per the framework shown in Fig.3. Baselines were trained through the optimization of WCE loss as given in Eq.1 whereas distillation was carried out through the optimization of the loss given in Eq.4. Furthermore, we conducted two experiments to

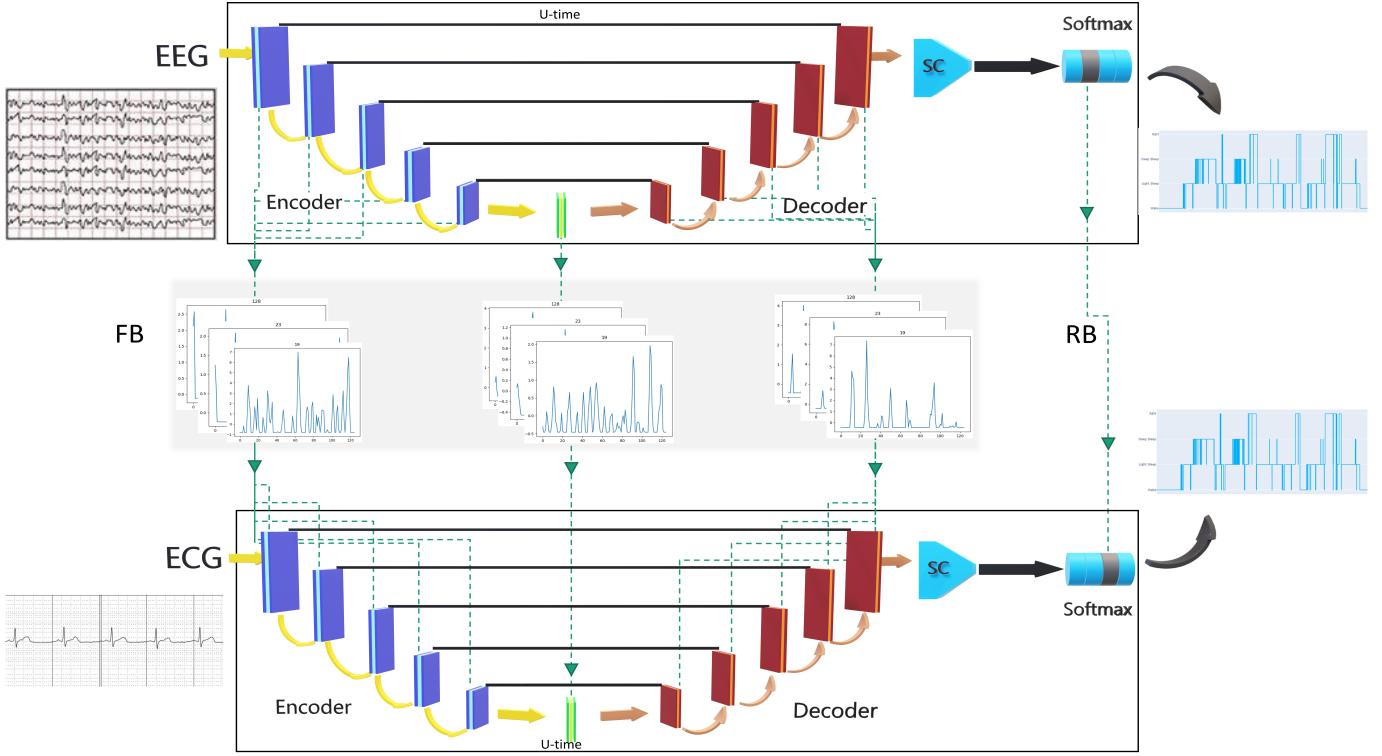


Fig. 3: Knowledge transfer in action from EEG sleep staging model to ECG sleep staging model. SC: Segment Classifier from U-Time. FB: Feature based (Attention Transfer distillation) feature learning. RB: Response based (Softmax distillation) feature learning.

investigate the individual modules in our proposed framework. The primary component in all the experiments involve two significant steps as given below:

- 1) Feature Training (Step 1): The loss (*Eq.3*) between ECG and EEG features is optimized by $M(ecg; \phi_{ecg})$ after freezing the EEG weights. This trains the ECG model to train towards imitating the feature maps of the EEG model.
- 2) Final Training (Step 2): Subsequently, $M(ecg; \phi_{ecgpre})$ is trained further on RB loss(*Eq.4*) for optimizing the ECG model weights. The T parameter was chosen as 1, as given in [10] and beta was chosen empirically to provide standard weights to classification and distillation loss.

The implementation procedure of the proposed distillation method and ablation methods is as follows:

- 1) **FB+RB+WCE(proposed method)**: Step 1 is executed to train features followed by training in step 2 on the loss *Eq.4* with $\beta = 0.5$.
- 2) **FB+WCE(ablation method)**: Step 1 is executed to train features followed by training in step 2 on the loss *Eq.4* with $\beta = 0$, which consequently trains independently on the classification loss in *Eq.1*.
- 3) **RB+WCE(ablation method)**: Only step 2 is executed training on the loss *Eq.4* with $\beta = 0.5$.

¹ Each configuration of distillation was trained with learning rate(LR) of 10^{-3} for 150 epochs. Nvidia GTX3090Ti 24GB GPU was used for the code; Pytorch Lightning framework was used for the algorithm development. ²

III. RESULTS

We evaluated the results from the distillation model against their respective baseline model, given that the principal intention of this work is to demonstrate the potency of KD. The performance of our distillation models on the holdout test data is as shown in Table.I. Both weighted-F1-score and accuracy metrics displayed increments in performance for the proposed distillation method FB+RB+WCE as well as ablation methods FB+WCE and RB+WCE for both 4-class sleep staging and 3-class sleep staging. The best performing model for 4-class was the RB+WCE model, where the weighted-F1 score improved from **0.451** of ECG baseline to **0.512** (weighted-F1 showed improvement by **14.30 %**, Accuracy improved by 15.6 %). FB+WCE was the best performing model for 3-class, with weighted-F1 improved from **0.583** of ECG baseline to **0.661** (weighted-F1 showed improved by **13.41 %**, Accuracy improved by 18.1 %). However, the ECG baseline model was outperformed by all the distillation models, which corroborates the incorporation of KD in ECG based sleep staging.

¹FB: Feature based; RB: Response based; WCE: Weighted Cross Entropy classification Loss

²Code is available at https://github.com/Acrophase/Sleep_Staging_KD

TABLE I: Performance of KD and its components

Sleep Stages	Model	F1-weighted	Accuracy
W-R-L-D	EEG Base	0.85	0.85
	ECG Base	0.45	0.44
	RB + WCE	0.51	0.51
	FB + WCE	0.50	0.50
W-R-N	FB + RB + WCE	0.50	0.49
	EEG Base	0.90	0.90
	ECG Base	0.58	0.56
	RB + WCE	0.61	0.60
	FB + WCE	0.66	0.66
	FB + RB + WCE	0.64	0.63

W:Wake, R:REM, L:Light Sleep, D:Deep Sleep, N:Non-REM Sleep

Insight into the sleep stage-wise performances of distillation approaches is given in Table.II. In 4-class staging, Light sleep(L) showed the best improvement, where RB+WCE ablation method improved weighted-F1 to **0.611** from **0.473**. The best improvement was observed for the NREM class in 3-class staging, with weighted-F1 improved from **0.771** to **0.652** by the FB+WCE method. Noticeably, other classes have underachieved marginally. This is potentially owing to class imbalance resulting in imprecise feature training. However, improvement across all classes in both 3-class and 4-class was observed in the proposed FB+RB+WCE distillation, thus exhibiting relatively robust feature learning against the class imbalance.

IV. DISCUSSION

The above-indicated results demonstrate that the potency of the knowledge distillation as the proposed distillation model for ECG-based sleep staging remarkably outperformed the ECG baseline. Nevertheless, combining Feature-Based (FB) distillation and Response Based (RB) distillation need not necessarily increase the performance over independent usage of these distillation frameworks. This emphasizes the intricacy of the interaction between these two components of KD and suggests the requirement of independent optimization of the two modes of KD.

Figure.4 illustrates the feature learning in the bottleneck layer of the architecture, which represents the most compressed form of features. Although the bottleneck features themselves are not illustratable, a comparative approach assists in analyzing the functioning of KD. The figure shows two scenarios of a sleep stage prediction for 4-class sleep staging, comparing the ECG base model, the KD model and the EEG base model;

- case 1: ECG baseline mispredicts the sleep stage, but the proposed KD model predicts the sleep stage accurately
- case 2: ECG baseline predicts the sleep stage accurately, but the KD model mispredicts the sleep stage.

It is evident that distilled model's bottleneck features are comparable to that of the EEG base model's feature in case 1, which shows distinguished feature learning resulting in performance improvement. However, in case 2, misguided

TABLE II: KD methods class wise Results

	4 class F1 score				3 class F1 score		
	W	L	D	R	W	N	R
EEG Base	0.89	0.86	0.81	0.82	0.89	0.93	0.80
ECG Base	0.57	0.47	0.30	0.40	0.51	0.65	0.40
RB + WCE	0.54	0.61	0.31	0.35	0.53	0.70	0.34
FB + WCE	0.54	0.57	0.34	0.40	0.52	0.77	0.37
FB + RB + WCE	0.57	0.56	0.29	0.41	0.52	0.73	0.39

W:Wake, R:REM, L:Light Sleep, D:Deep Sleep, N:Non-REM Sleep

feature learning can be observed from the difference between the distilled model's feature and both EEG and the ECG baseline features, which were predicted accurately. This could be, to some extent, attributed to asymmetric distillation as a result of the class imbalance in the data, as observed in Table II.

The methods presented here work towards producing a less-invasive alternative to sleep studies by removing cumbersome EEG electrodes, which can be prone to reduced SNR through a patient's restless sleep, and replacing them with ECG electrodes through the help of KD. While the potential to exchange EEG for ECG signals is novel, diagnostic sleep studies still require other sensor measures per AASM guidelines, such as airflow, breathing effort, EMG from upper/lower limbs, and a pulse oximeter for oxygen saturation content. Our future efforts will be directed towards the number of sensors placed on a patient for sleep studies to improve patient comfort and their quality of sleep during the overnight study, thus optimizing for a better trade-off between comfort and accuracy.

In spite of the promising improvement in performance brought about by KD, we identified a few limitations in this study. Firstly, using temporal models like Long Short Term Memory networks (LSTM) can improve the baseline ECG model utilized in this paper because of their ability to identify sparsely distributed features over time. Furthermore, using additional unobtrusive or less obtrusive modalities like respiratory signal and ECG have improved sleep staging performance. Previous works [8] [5] have been trained on an extensively large dataset (>4000 records) which achieved noteworthy results on ECG based sleep staging, whereas our study used a relatively compact dataset. This shows that the choice of the dataset used in sleep staging studies cannot be undermined. Future work would involve exploring the components of KD to optimize for ECG signals as well as incorporating the benefits of KD to more optimized DL architectures, ultimately boosting the overall performance.

V. CONCLUSION

This study expands the present knowledge in sleep staging from ECG by making the following contributions.

- Proposed usage of single modality, single-lead ECG signal, for sleep staging, minimizing the obtrusiveness and making it suitable for point of care setting.
- Demonstration of the viability of a KD framework for two different morphological signals, resulting in improved

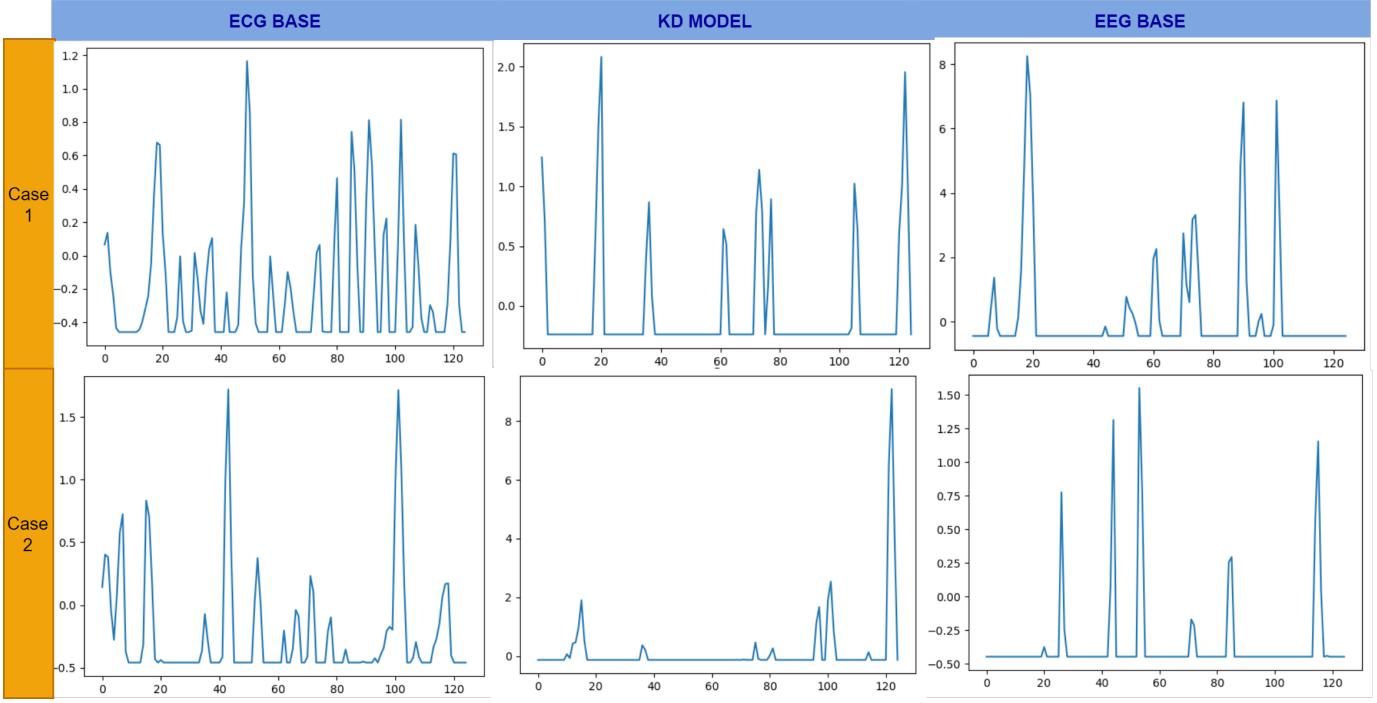


Fig. 4: Comparison of bottleneck layer features for Knowledge Distillation. case1: KD model correct, ECG base incorrect; case2: ECG baseline correct, KD model incorrect.

performance of ECG-based sleep staging with knowledge assistance via EEG features for the same task.

- Analysis of the individual components of the KD by providing comparative analysis from the bottleneck layer features gave insights into the rationale for the performance improvement.

REFERENCES

- [1] A. Rechtschaffen, “A manual for standardized terminology, techniques and scoring system for sleep stages in human subjects,” *Brain information service*, 1968.
- [2] C. Iber, “The aasm manual for the scoring of sleep and associated events: Rules,” *Terminology and Technical Specification*, 2007.
- [3] M. Perslev, M. Jensen, S. Darkner, P. J. Jennum, and C. Igel, “U-time: A fully convolutional network for time series segmentation applied to sleep staging,” *Advances in Neural Information Processing Systems*, vol. 32, pp. 4415–4426, 2019.
- [4] H. Abdullah, G. Holland, I. Cosic, and D. Cvetkovic, “Correlation of sleep eeg frequency bands and heart rate variability,” in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5014–5017, IEEE, 2009.
- [5] Q. Li, Q. Li, C. Liu, S. P. Shashikumar, S. Nemati, and G. D. Clifford, “Deep learning in the cross-time frequency domain for sleep staging from a single-lead electrocardiogram,” *Physiological measurement*, vol. 39, no. 12, p. 124005, 2018.
- [6] M. Radha, P. Fonseca, A. Moreau, M. Ross, A. Cerny, P. Anderer, X. Long, and R. M. Aarts, “Sleep stage classification from heart-rate variability using long short-term memory neural networks,” *Scientific reports*, vol. 9, no. 1, pp. 1–11, 2019.
- [7] P. Fonseca, X. Long, M. Radha, R. Haakma, R. M. Aarts, and J. Rolink, “Sleep stage classification with ecg and respiratory effort,” *Physiological measurement*, vol. 36, no. 10, p. 2027, 2015.
- [8] N. Sridhar, A. Shoeb, P. Stephens, A. Kharbouch, D. B. Shimol, J. Burkart, A. Ghoreyshi, and L. Myers, “Deep learning for automated sleep staging using instantaneous heart rate,” *NPJ digital medicine*, vol. 3, no. 1, pp. 1–10, 2020.
- [9] H. Sun, W. Ganglberger, E. Panneerselvam, M. J. Leone, S. A. Quadri, B. Goparaju, R. A. Tesh, O. Akeju, R. J. Thomas, and M. B. Westover, “Sleep staging from electrocardiography and respiration with deep learning,” *Sleep*, vol. 43, no. 7, p. zsz306, 2020.
- [10] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.
- [11] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, “Fitnets: Hints for thin deep nets,” *arXiv preprint arXiv:1412.6550*, 2014.
- [12] N. Komodakis and S. Zagoruyko, “Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer,” in *ICLR*, 2017.
- [13] J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge distillation: A survey,” *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.
- [14] C. O’reilly, N. Gosselin, J. Carrier, and T. Nielsen, “Montreal archive of sleep studies: an open-access resource for instrument benchmarking and exploratory research,” *Journal of sleep research*, vol. 23, no. 6, pp. 628–635, 2014.