

Dual Deep Learning for Sleep Classification

Mohamed Abdelmagid¹, Marwa Yusuf¹, Basem M. ElHalawany^{2,1}, *Senior Member, IEEE*, and Ahmed Fares^{3,1}

Abstract—Sleep stage classification based on EEG is a crucial tool for understanding sleep quality. It enables the identification of various stages of sleep, including REM and non-REM sleep, which exhibit distinct patterns of brain activity and physiological responses. This valuable information can be utilized to diagnose sleep disorders, monitor treatments' impact, and better understand the correlation between sleep and overall health. Over the years, various machine learning models have been proposed to help sleep specialists identify and distinguish different stages of sleep. However, most of these models have relied on multiple EEG signals or additional modalities such as EOG or ECG, which is impractical for real-time sleep stage classification at home or home monitoring. To address this issue, we propose a novel model that leverages the transformer network to classify sleep stages based on a single EEG channel automatically. In addition, we study the impact of combining the input EEG signal with its time-frequency representations on classification accuracy. Our proposed model achieves competitive results with currently available work on highly imbalanced data sets, namely edf-20.

I. INTRODUCTION

Sleep quality is a significant contributing factor to a better quality of life. Studies have shown the significance of sleep for improved mental and physical health [1]. Sleep disorders can be the cause of multiple medical conditions or a symptom of others [2]. As a result, there is an increasing desire to measure sleep quality. Manual Sleep staging requires sleep specialists to examine the Polysomnography PSG records of patients to annotate the different periods of sleep with their respective sleep stages. A PSG consists of Electroencephalogram (EEG), Electrooculogram (EOG), Anelectromyogram (EMG), and Electrocardiogram (ECG) signals recorded for the patient [3]. Electroencephalogram (EEG) signals are electromagnetic waves emitted from the brain as it performs its functions. The characteristics of an EEG signal change in correlation with the cognitive processes taking place. There is an increasing interest in conducting sleep studies using only a single EEG channel. This is for the convenience of patients to enable at-home treatment and monitoring.

Researchers have conducted many studies to automate the sleep staging process using Machine learning ML models [4][5][6]. Most previous work on the task of sleep staging tends to utilize contextual information to raise the performance of their model. Contextual information is available by using concurrently recorded modalities. Multimodal learning is used

in many fields and has the potential to improve the accuracy of brain-representation decoding [7]. However, The results obtained through EEG alone are comparably high, proving the additional modalities redundant. Additionally, using other modalities requires attaching more electrodes to the patient, which is inconvenient for most patients, and impractical for at-home treatment and monitoring. Alternatively, researchers can obtain contextual information by using previous or upcoming epochs to classify the current epoch. This method requires either using a large window size to take multiple epochs at once [6] or the use of sequence-to-sequence approaches [8], especially LSTMs extensively used with EEG signals [9][10]. Those methods, however, require the storage of previous or future epochs, which is unsuitable for real-time sleep staging. Another challenge facing automatic sleep staging is the dataset imbalance problem in all sleep datasets [11]. Previous work addresses the dataset imbalance by using oversampling [12] or a customized loss function [4][5].

To address the above issues, we present a study of the effectiveness of combining raw EEG signal and its visual representation for improved model performance. Additionally, we propose a novel model architecture using only a single-channel EEG for the single-shot automatic sleep classification task. Our proposed model leverages multiple representations of a single EEG signal to extract more features from a single EEG channel instead of relying on additional modalities or previous epochs. Hence, the contributions of this publication are: i) examine the effectiveness of merging time-frequency representations with raw signals for the task of a single shot sleep stage classification; and ii) propose a novel model for dually learning EEG signals and visual features.

II. PROPOSED METHOD

II-A Overview of the proposed model structure : The proposed model consists of two paths. The input of the signal path is the raw signal, while the input to the image path is the power-spectrum image generated from the raw signal. The signal path is heavily inspired by [4]. However, instead of using their causal convolutions block we opted for the original positional encoder. For the image path, we first transform the raw signal into its power spectrum representation, resize it to a 144*144 image, and finally, employ the ViT [13] model to process the generated image. Both paths produce a feature vector later used by a classifier to make the final decision.

II-B Signal path : To process the raw signal, we used the AttnSleep model [4]. It starts with an MRCNN. MRCNNs are often used in signal processing to extract frequency-time features, as different kernel sizes can capture different

¹Electrical Engineering Department, Faculty of Engineering at Shoubra, Benha University, 11629 Cairo, Egypt. {mohamed.abdelmagid, marwayusuf, ahmed.fares}@feng.bu.edu.eg

²Electronics and Communication Engineering Department, Kuwait College of Science and Technology Doha District, Block 4, 93004 Kuwait. basem.mamdoh@feng.bu.edu.eg

³Egypt-Japan University of Science and Technology (E-JUST), Alexandria, 21934, Egypt. ahmed.fares@ejust.edu.eg

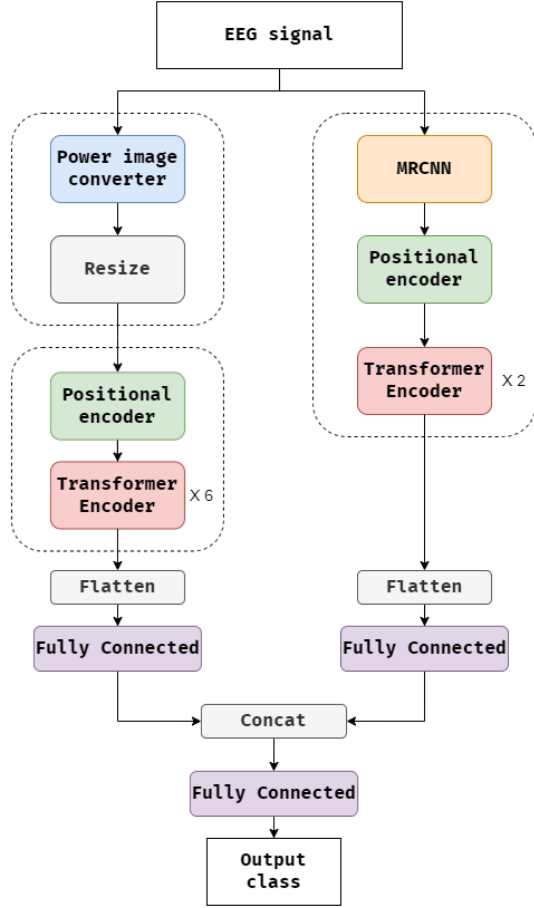


Fig. 1: The proposed model

frequency bands. The kernel sizes are chosen to focus on the significant channels in the EEG signal, followed by a positional encoder to obtain a feature vector that is then passed to a two-layer transformer encoder.

II-C Image path : Many models attempt to take advantage of alternative data representations that can highlight hidden features that are harder to detect with neural networks alone[14]. Frequency domain features are one such example. There are multiple representations for frequency domain features, the most commonly used are spectrogram and power spectrum image. i) *Spectrogram*: a 2D representation of the input signal depicting the magnitude of each frequency band over time. It is computed by dividing the signal into windows of equal size, often overlapping, to capture its temporal evolution. Fast Fourier transform (FFT) is then applied to each window to obtain the magnitude of each frequency component. Choosing suitable parameters for spectrogram generation for a specific task can require a long process of trial and error. ii) *Power spectrum image*: a 2D representation of the input signal that depicts the power of each frequency band over time. It is computed in the same way we compute spectrogram except we compute the square of the magnitude to get the power then we use a logarithmic scale for better visualization. Figure 2 depicts samples of the power spectrum image generated for each class in the dataset.

We experimented with both representations for the image

TABLE I: The distribution of samples across the class Wake, N1, N2, N3, And REM in edf-20 and edf-78 datasets

	W	N1	N2	N3	REM
Edf-20	8285	2804	17799	5703	7717
Edf-78	65951	21522	69132	13039	25835

path. We use the ViT model[13] to process the generated visual representation. The model contains six transformer encoder blocks, each using 4 heads and a dimensionality of 512.

II-D Classification : The output of the two paths is two feature vectors. We flatten them to 2400 and 41472 output features, respectively. We then compress the feature vectors, by Fully-connected layers to 512 features each, and concatenate them to build a feature vector. Finally, we pass this feature vector to a fully connected layer classifier, all using linear output.

III. EXPERIMENTAL RESULTS

III-A Datasets and Evaluation Metrics We used preprocessed versions of edf-20 [15] and edf-78 shared by the authors of [4]. The original edf-20 provided by PhysioBank [11] had 20 subjects with 2 hypnogram files, each recorded in two nights except subject 13, whose second-night record was lost. Each file contains a 20-hour-long record of EEG (Fpz-Cz and Pz-Oz channels), EOG (horizontal), and submental chin EMG were sampled at 100 Hz. The dataset contains two studies: sleep-cassette and sleep-telemetry. We used only the former, which studies the effect of age on sleep. The dataset was later expanded to edf-78 by including more subjects, all older than 50 years, increasing the number of subjects to 78. Since the other was lost, only one record is available for both subjects 36 and 52. Sleep specialists annotated every 30-second period in all provided records with one of the classes W (Wake), N1, N2, N3, N4, REM (Rapid Eye Movement), Movement (M), and UNKNOWN.

Datasets used for automatic sleep stage classification are often highly imbalanced. As shown in Table I, stages W, N2, and REM are the most dominant classes in edf-20 and edf-78, with N1 and N3 being minority classes and often misclassified, especially N1, whose best accuracy across all previous work did not exceed 52%. To evaluate the performance of our model on highly imbalanced datasets, we adopted four metrics, namely the accuracy (ACC), macro-averaged F1-score (MF1), Cohen Kappa (κ), and the macro-averaged G-mean (MGm).

III-B Baselines and Experimental setup Our experiments compared our model to other state-of-the-art models on the task of stage sleep classification on the aforementioned datasets like AttnSleep [4]. i) *DeepSleepNet*: Uses CNN for time-invariant feature extraction and Bi-LSTM to learn inter-sleep-stage transition. ii) *SleepEEGNet*: Uses CNN for feature map encoding and encoder-decoder Bi-LSTM followed by a softmax classifier. iii) *AttnSleep*: Uses MRCNN for feature extraction and the attention mechanism of the transformer encoder blocks' MHA. iv) *MultiTask CNN*: Uses CNNs with multiple resolutions and multi-task softmax classifier.

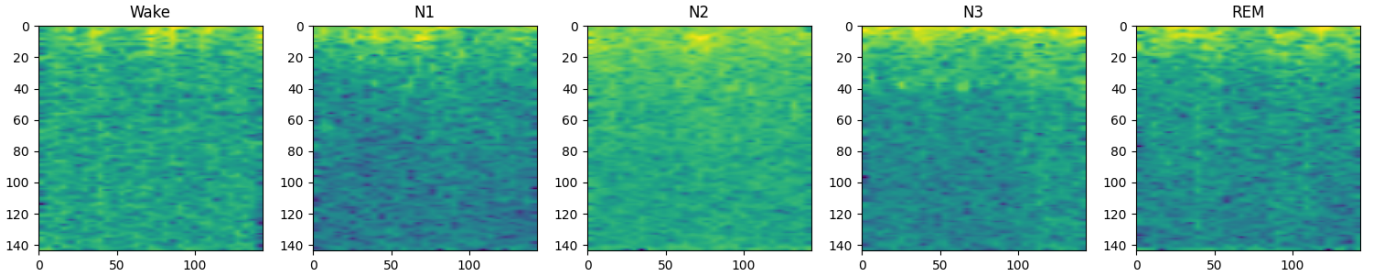


Fig. 2: The power spectrum image generated from raw signals corresponding to one of each class

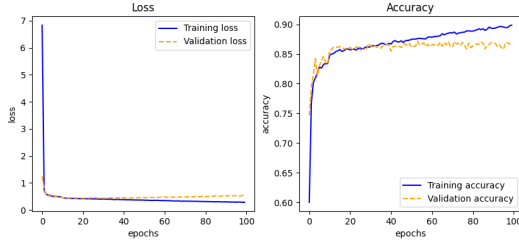


Fig. 3: Training and validation accuracy progression with epoch-wise 80-20 split on the edf-20[11] dataset

We assumed the same experimental setup by the authors of AttnSleep [4]. We split the data on a subject and epoch-wise basis. We used 20-fold cross-validation, so the subjects of each dataset were split into 20 groups, one of which is used for the testing of each fold. The epoch-wise split used the shuffled dataset at an 80-20 split strategy. Our model was trained for 20 trials (subjects-wise and epoch-wise) on each dataset. Each trial was performed for 100 epochs. We used an Adam optimizer [16] with a learning rate of 0.001, which automatically decreases to 0.0001 after 10 epochs for a more fine-tuned optimization, weight decay of 0.001, and AMSGrad enabled.

We used the class-aware loss function proposed by [4] To address the class imbalance. It uses a weighted cross-entropy loss as computed in equation 1.

$$L = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^M w_c y_{i,c} \log(p_{i,c}) \quad (1)$$

Where N is the number of dataset samples, M is the number of classes, $y_{i,c}$ is a binary indicator if the i^{th} sample belongs to class c , and $p_{i,c}$ is the predicted probability that sample i belongs to class c . For each class, the weight is computed to consider the proportion of the class in the dataset and how distinguishable it is.

III-C Choosing a suitable visual representation We experimented with both spectrogram image and power spectrum image representation. For both, we used a 200-sample hamming window with 50% overlap and 256-point Fast Fourier Transform (FFT) similar to [14], which gave the best result among the other configurations we tried. In this experiment, we used visual representations only. We obtained a validation accuracy of 43% using the spectrogram image and an average of 64% for the power spectrum image. Therefore, we chose the power spectrum as the visual representation in our model.

III-D Modality fusion We attempted to merge the signal and image paths on the decision level and feature level. We trained the two paths separately on the edf-20 dataset to test the feasibility of decision-level fusion. Though the image path reached an accuracy of 71% on this trial, it was still superseded by the signal path, managing to correctly predict at most 1% of the dataset samples that the signal path failed to predict. Hence we conclude that merging the models optimally on the decision level will not add any additional value to the model.

As for merging on the feature level. The proposed model shown in Figure 1 reached an average accuracy of 83.93% on the edf-20 dataset split subject-wise. Using an epoch-wise training strategy with an 80-20 train-test split, the model reached an accuracy of 86.22%. Figure 3 shows the changes in the training and validation accuracy of the model. The confusion matrix of epoch-wise training for edf-20 and edf-78, shown in Tables III and IV respectively show that the proposed model reaches higher than 80% on all classes except for N1 due to its under-representation in the datasets.

III-E Ablation Study We hypothesize that using the dual representation of EEG signals in their raw form and power spectrum image would boost the performance of machine learning models operating on EEG data. To examine the effectiveness of each modality independently and when combined we derived Three models from our own components. i) *AttnSleep Model*[4]: Path for processing raw signal only. ii) *ViT*: Path for processing the power spectrum image generated from the input signal. iii) *Merged-model*: ViT + AttnSleep.

The study used the edf-20 dataset with an epoch-wise 80-20 split. The signal path (AttnSleep) reaches an accuracy of 86.22% on the edf-20 dataset. The ViT model using the generated power spectrum image representation reaches an accuracy of 71% at best. The merged model reached 86.23%. The main strength of using power spectrum image is to allow the model to easily capture frequency-domain data which has a notable impact on multiple fields of signal processing. However, the signal path uses MRCNN, which can capture frequency domain features, and two layers of transformer encoders making the signal path more formidable in capturing both temporal and spectral domain representations. This is why the image features could not contribute much to the overall model accuracy.

IV. CONCLUSION

Our research introduced an innovative multi-modal sleep stage classification approach, harnessing the raw EEG signal

TABLE II: Comparison between the performance between our proposed model and other previous work on the datasets edf-20 and edf-78

Dataset	Model	Metrics				Per-Class F1 score				
		Acc	F1	K	MGM	W	N1	N2	N3	REM
Edf-20	DeepSleepNet [12]	82%	76.9	0.76	86.9	84.7	46.6	85.9	84.8	82.4
	SleepEEGNet [5]	84.26%	79.66	0.79	85.3	89.19	52.19	86.77	85.13	85.02
	XSleepNet [14]	83.90%	78.7	0.771	--	--	--	--	--	--
	MultitaskCNN [6]	83.10%	75	0.77	83.1	87.9	33.5	87.5	85.8	80.3
	AttnSleep [4]	84.40%	78.1	0.79	85.5	89.7	42.6	88.8	90.2	79
	Ours - subjectwise	83.93%	78.48	0.78	86.48	90.13	45.84	88.21	89.3	78.92
	Ours - epochwise	86.22%	80.74	0.81	87.82	92.34	48.2	89.92	90.59	82.65
Eddf-78	DeepSleepNet [12]	77.80%	71.8	0.7	81.6	90.9	45	79.2	72.7	71.1
	SleepEEGNet [5]	80.03%	73.55	0.73	82.3	91.72	44.05	82.49	73.45	76.06
	MultitaskCNN [6]	79.60%	72.8	0.72	82.5	90.9	39.7	83.2	76.6	73.5
	AttnSleep [4]	81.30%	75.1	0.74	83.6	92	42	85	82.1	74.2
	XSleepNet [14]	80.30%	76.4	0.726	--	--	--	--	--	--
	Ours - subjectwise	79.92%	74.16	0.72	82.78	91.65	43.06	83.96	81.36	70.8
	Ours - epochwise	82.22%	77.12	0.76	85.07	93.01	47.25	85.12	83.65	76.58

TABLE III: The Confusion matrix of epoch-wise training of the proposed model on the edf-20 dataset

	W	N1	N2	N3	REM
W	92	5	0.86	0.22	1.7
N1	13	51	15	1.7	21
N2	0.62	3.7	88	3.1	4.2
N3	2.9	0.053	8.3	91	0.057
REM	1.6	7.9	6.5	0.19	84

TABLE IV: The Confusion matrix of epoch-wise training of the proposed model on the edf-78 dataset

	W	N1	N2	N3	REM
W	92.18	5.87	0.63	0.06	1.25
N1	13.07	49.04	24.97	0.19	12.73
N2	0.6	7.9	84.63	2.72	4.15
N3	0.21	0.24	16.56	82.82	0.17
REM	2.87	12.19	7.39	0.07	77.49

and its corresponding power spectrum image representation. Our model comprises two distinct paths: one based on the AttnSleep model for processing raw signals and the other on ViT for handling image representations. Our experimental results demonstrate the promising potential of this approach, achieving a commendable accuracy rate of 86% with epoch-wise training and 84% with subject-wise training. This underscores the viability of utilizing multiple data modalities for the brain signal decoding stage. During our ablation study, we explored the signal and image path fusion at the decision level for single-shot sleep stage classification. While our findings indicated that this fusion approach did not yield substantial improvements, it is worth noting that the image representation alone provided valuable insights, Which open areas for future research and improvements.

REFERENCES

- [1] J. Kühnel, H. Zacher, J. De Bloom, and R. Bledow, "Take a break! benefits of sleep and short breaks for daily work engagement," *European Jour. Work Organiz. Psych.*, vol. 26, no. 4, pp. 481–491, 2017.
- [2] C. WORSNOP, M. NAUGHTON, C. BARTER, T. MORGAN, A. ANDERSON, and R. PIERCE, "The prevalence of obstructive sleep apnea in hypertensives," *American Journal of Respiratory and Critical Care Medicine*, vol. 157, no. 1, pp. 111–115, 1998, pMID: 9445287. [Online]. Available: <https://doi.org/10.1164/ajrccm.157.1.9609063>
- [3] A. D. Krystal and J. D. Edinger, "Measuring sleep quality," *Sleep medicine*, vol. 9, pp. S10–S17, 2008.
- [4] E. Eldele, Z. Chen, C. Liu, M. Wu, C.-K. Kwok, X. Li, and C. Guan, "An attention-based deep learning approach for sleep stage classification with single-channel eeg," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 809–818, 2021.
- [5] S. Mousavi, F. Afghah, and U. R. Acharya, "Sleepeegnet: Automated sleep stage scoring with sequence to sequence deep learning approach," *PloS one*, vol. 14, no. 5, p. e0216456, 2019.
- [6] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "Joint classification and prediction cnn framework for automatic sleep stage classification," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 5, pp. 1285–1296, 2018.
- [7] J. Jiang, A. Fares, and S.-H. Zhong, "A context-supported deep learning framework for multimodal brain imaging classification," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 6, pp. 611–622, 2019.
- [8] H. Phan, O. Y. Chén, P. Koch, Z. Lu, I. McLoughlin, A. Mertins, and M. De Vos, "Towards more accurate automatic sleep staging via deep transfer learning," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 6, pp. 1787–1798, 2021.
- [9] S.-h. Zhong, A. Fares, and J. Jiang, "An attentional-1stm for improved classification of brain activities evoked by images," in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1295–1303. [Online]. Available: <https://doi.org/10.1145/3343031.3350886>
- [10] A. Fares, S. Zhong, and J. Jiang, "Region level bi-directional deep learning framework for eeg-based image classification," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2018, pp. 368–373.
- [11] B. Kemp, A. Zwinderman, B. Tuk, H. Kamphuisen, and J. Obery, "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 9, pp. 1185–1194, 2000.
- [12] 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.
- [13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.
- [14] H. Phan, O. Y. Chén, M. C. Tran, P. Koch, A. Mertins, and M. De Vos, "Xsleepnet: Multi-view sequential model for automatic sleep staging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5903–5915, 2022.
- [15] E. Eldele, "Preprocessed Sleep-EDF-20 dataset," 2022. [Online]. Available: <https://doi.org/10.21979/N9/MA1AVG>
- [16] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.