Pediatric Sleep Scoring In-the-wild from Millions of Multi-channel EEG Signals

Harlin Lee [1] **Aaqib Saeed** [2]

Abstract

Sleep is critical to the health and development of infants, children, and adolescents, but pediatric sleep is severely under-researched compared to adult sleep in the context of machine learning for health and well-being. Here, we present the first automated pediatric sleep scoring results on a recent large-scale sleep study dataset that was collected during standard clinical care. We develop a transformer-based deep neural network model that learns to classify five sleep stages from millions of multi-channel electroencephalogram (EEG) signals with 78% overall accuracy. Further, we conduct an in-depth analysis of the model performance based on patient demographics and EEG channels.

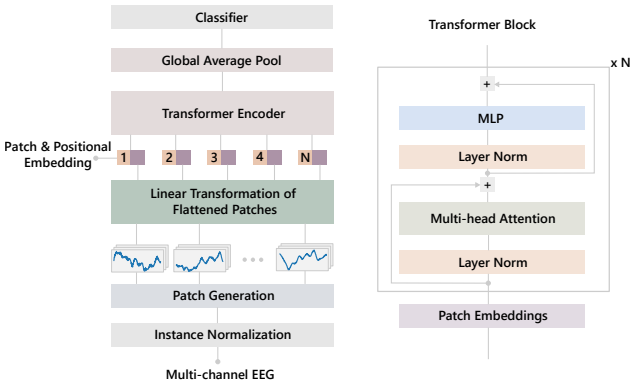


Figure 1. Illustration of our patch-based transformer neural network architecture designed for pediatric sleep scoring from multi-channel EEG signals. MLP stands for multi-layer perceptron.

1. Introduction

Sleep is a necessary physiological process that actively engages multiple organ systems. Sleep disorders or sleep disturbances not only negatively affect one's cognitive and physical functions (Wulff et al., 2010; Dawson & Reid, 1997), but can also lead to serious medical conditions. For example, obstructive sleep apnea (OSA) contributes to increased risk of cardiovascular diseases, such as hypertension (Peppard et al., 2000), stroke (Mohsenin, 2015) and heart failure (Bradley & Floras, 2003) in adults, as well as neurobehavioral issues (Beebe et al., 2004) and even morbidity (Lumeng & Chervin, 2008; Jennum et al., 2013) in infants and children.

Diagnoses of many sleep conditions require polysomnography (PSG), or overnight sleep study, where a patient sleeps in a clinic while their physiological signals are monitored under the supervision of trained technicians (Kushida et al., 2005; Berry et al., 2018; 2017). A PSG dataset may include many modalities, such as electroencephalogram (EEG), electromyelogram (EMG), electrooculogram (EOG), and respiratory airflow. A crucial first step towards diagnosis in PSG data analysis is sleep scoring, or sleep stage classification, which assigns every 30-second segment of sleep into two stages, rapid eye movement (REM), and non-REM, then further divides the latter into shallow sleep (stages N1 and N2) and deep sleep (stage N3). In a typical clinical setting, this process is done manually by a technician, which is highly labor-intensive, time-consuming and prohibitively expensive.

Naturally there have been many attempts to automate sleep scoring, especially in recent years with the help of deep neural networks and freely-available public PSG datasets; see reviews in (Bandyopadhyay & Goldstein, 2022; Fiorillo et al., 2019; Watson & Fernandez, 2021; Phan & Mikkelsen, 2021). Most existing deep-learning-based approaches including (Zhang et al., 2022) rely on traditional convolutional neural network or recurrent neural network architectures, and have yet to be widely adapted in clinical settings. On the other hand, self-attention based transformer models (and convolution-based models that are similar in spirit) have recently gained state-of-the-art performance on a range of tasks involving vision, audio and text modalities (Vaswani et al., 2017; Dosovitskiy et al., 2020; Trockman & Kolter, 2022) due to their capability of modeling long-range dependencies and parallelizability.

Several works (Phan et al., 2022; Yang et al., 2021; Chen et al., 2021; Kim et al., 2021) have attempted to utilize

[1]Department of Mathematics, University of California, Los Angeles, Los Angeles, CA 90095, USA [2]Philips Research, Eindhoven, The Netherlands. Correspondence to: Harlin Lee <harlin@math.ucla.edu>.

transformer-based models for processing EEG signals and achieve superior performance over other classic deep architectures. Given that, we design a simple yet effective neural architecture that can process millions of multi-channel EEG signals (F4-M1, O2-M1, C4-M1, O1-M2, F3-M2, C3-M2, CZ-O1) and learn useful representations for pediatric sleep scoring. Our model is based on the transformer architecture that operates directly on patches as input and maintains the same resolution and representations throughout all layers.

We develop and demonstrate our model on the new Nationwide Children's Hospital (NCH) Sleep DataBank (Lee et al., 2021), which has not been explored in automatic sleep scoring research before. This dataset is of massive scale, containing 3,984 PSG from 3,673 unique patients, allowing us to leverage the full power of machine learning models. It explicitly focuses on pediatric sleep, and the sleep studies were also conducted in a current real-world clinical setting (i.e. in-the-wild in NCH between 2017 and 2019). Hence our model is trained from data that is closest to what it will see in future deployment, which is unlike prior work on pediatric sleep scoring that learn from mostly healthy adult subjects in a clinical trial.

Our transformer-based model achieves an overall pediatric sleep scoring accuracy of 78.2%, and our analysis reveals that the accuracy is above 80% for 6-15 year old patients. Our work is different from previous approaches that use transformers for sleep-stage scoring in several ways:

- Our models directly operate over raw signals as opposed to time-frequency images to further simplify the learning pipeline and improving training efficiency.
- We do not utilize any other modalities except EEG signals as other works employ additional modalities, e.g., EOG.
- We do not require mixing or ensembling of multiple models that can significantly increase model size and inference cost, especially for on-device deployment.
- Our model is trained specifically for pediatric sleep scoring, which is a severely under-researched topic compared to adult sleep in the context of machine learning for health and well-being.

2. Results

We develop a neural network model for predicting sleep stages in a real-world clinical environment from pediatric multi-channel EEG signals. We design a patch-based transformer model that operates over one-second segments of sleep, which provides strong support for long-range modeling dependencies in the input signal to learn discriminative features. We utilize the NCH SleepBank dataset, which comprises approximately 3.6 million fully-annotated EEG examples by domain experts, for training and evaluating models. Only seven-channel EEG signals (F4-M1, O2-M1, C4-M1, O1-M2, F3-M2, C3-M2, and CZ-01) at 128 Hz are used to classify instances into five sleep stages (i.e., wakefulness, non-REM stages 1, 2, 3, REM). Detailed information about the NCH dataset, including patient characteristics and annotation strategy, is provided in supplementary materials Section 4.1. To evaluate model performance, we compute precision, recall, F1-score, and accuracy based on the confusion matrix, and also assess generalization across age groups, races, and gender. Finally, we perform ablation over EEG channels to estimate the contribution of each channel toward sleep scoring.

2.1. Neural network model learns which parts of the raw EEG signals are important

Our transformer-based model is inspired by the ViT (Dosovitskiy et al., 2020) network, which we adapt here to multi-channel time-series signals. The high-level illustration of the model architecture is shown in Figure 1. The model accepts inputs of the shape `(Sampling frequency in Hz` $\times$ `# of seconds)` $\times$ `(# of EEG channels)` $= 3,840 \times 7$, after which the instance normalization layer normalizes each EEG signal channel-wise independently. The patch generation layer then splits the sleep epoch by every second, creating 30 patches of input with shape 128×7. This is analogous to tokenization in natural language processing (NLP), where a piece of text is converted into smaller units (i.e. tokens) such as words or characters. This helps the transformer learn which seconds of the EEG signals are important for sleep scoring.

After the patches are generated, they are embedded into 64-dimensional vectors via a linear patch encoder layer. This is then added to 64-dimensional positional vectors to create images that encode both positional and waveform shape information of the input patches. The rest of the model is similar to a classic transformer encoder with 8 blocks with 4 attention heads, which is explained in more detail in Section 4.2. Each block has a normalization layer, a multi-head attention layer, another normalization layer, and a two-layer multi-layer perceptron (MLP) with 128 and 64 units. For feature aggregation, we use global average pooling followed by a classification layer with units equal to the number of sleep stages, i.e. 5.

2.2. Data preparation

We use 3,928 PSGs from 3,631 unique patients for model training and evaluation. In particular, we split the patients into 70%, 10%, and 20% for training, validation, and testing, respectively, so that the three splits have no overlap in pa-

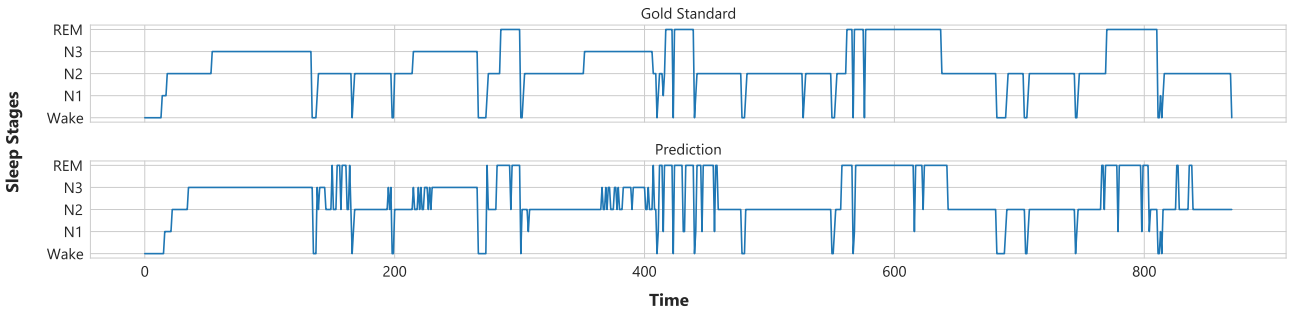
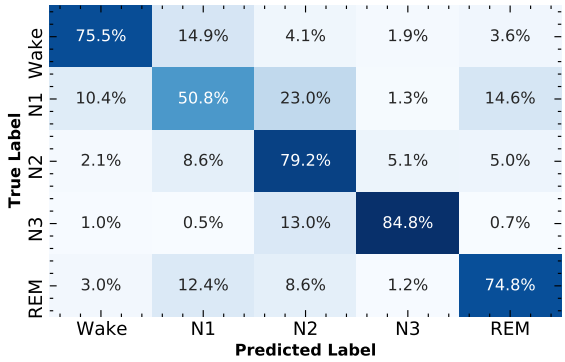
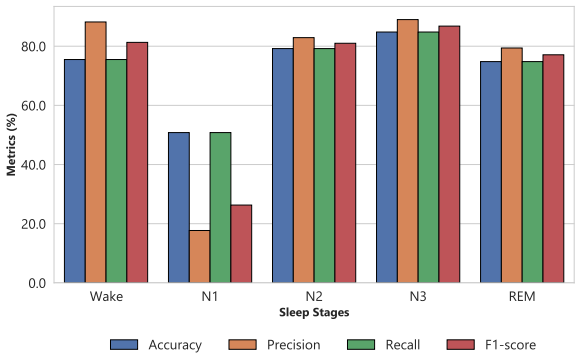


Figure 2. Hypnogram of the gold standard (manual scoring) versus prediction from our transformer-based model for five sleeping stages annotations on a randomly selected subject from the test set. The horizontal axis is in units of sleep epochs, and the entire length corresponds to one overnight sleep study.



(a) Confusion Maatrix



(b) Performance Metrics

Figure 3. a) Normalized confusion matrix for sleep scoring on the entire test set. The number in ith row and jth column indicates the percentage (%) of samples in stage i (according to manual scoring) that were predicted to be in stage j by our classifier. Each row adds to 100%. Overall accuracy of our model across all sleep stages is 78.2%. b) Model performances on the entire test set, as evaluated by accuracy, precision, recall, and F1-score (weighted).

tients. During the learning phase, we monitor the validation set performance for model checkpointing, and report results on the test set. Our training set consists of 2.5+ million instances, and the test set has 730K+ instances, as shown in Table 1. To the best of our knowledge, we, for the first time, report results on a large-scale pediatric sleep stage scoring dataset that is collected in the wild. We provide the rest of the data pre-processing and related information in Section 4.1, including patient demographic characteristics in Table 4.

Sleep Stage	Train	Validation	Test
All	2,611,845	301,116	731,344
Wake	469,473	56,327	135,845
N1	92,615	9,768	25,219
N2	990,299	112,188	273,191
N3	623,164	71,728	176,308
REM	436,294	51,105	120,781

Table 1. Number of samples in train, validation, and test sets. One sample is a 30-second sleep epoch.

2.3. Model demonstrates strong pediatric sleep scoring performance

Across all sleep stages in the test set, the transformer model achieves 78.2% accuracy, F1-score (macro) of 70.5%, F1-score (weighted) of 79.9%, and Cohen's Kappa score of 71.0%. See Figure 2 for a randomly selected subject's hypnogram that is predicted by our transformer model. Model performance for each sleep stage is presented in Figure 3a as a normalized confusion matrix, and in Figure 3b in terms of accuracy, precision, recall, and F1-score. The model demonstrates strong predictive power (near 80%) for Wake, N2, N3, and REM, but not as much in predicting N1, which has the smallest sample size. The model has lower precision for sleep epochs in N1, often inaccurately labeling them as N2 or REM. Nonetheless, this is a huge improve-

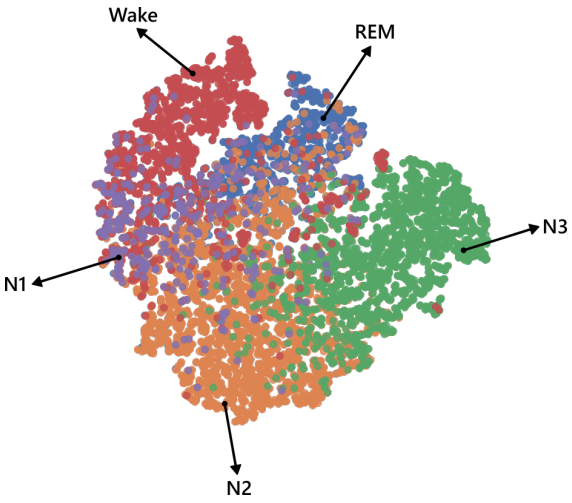


Figure 4. t-SNE embedding of features learned with the transformer network. We project 128-dimensional representations from the model's penultimate layer to 2 dimensions for a random subset of test set instances. Each point in the plot represents one 30-second sleep epoch. Note that t-SNE does not utilize class labels. The colors are added during post-hoc analysis for better interpretability.

ment over the wavelet-based baseline classifier in (Lee et al., 2021), which had 64.4% accuracy across all sleep stages and only 0.9% with N1.

We also visualize the features learned by the transformer model in Figure 4, projecting them from 128-dimensional to 2-dimensional space via t-SNE (Van der Maaten & Hinton, 2008). The clusters that naturally form for each sleep stage suggest that the transformer model learns meaningful features from the raw EEG signals before entering the final classifier layer. Furthermore, we note that N3 samples seem to be most well-separated, while N1 samples seem to overlap with other stages the most, which aligns with the classification accuracy results in Figure 3a.

2.4. Model sleep scores better on 6 to 15 year olds and children of Asian, Others and Unknown race with over 80% accuracy

Figure 5 and Table 2 report the transformer model performance on different subsets of the patients. Figure 5 shows that the model achieves the highest accuracy (85%) on the 8 to 9 year olds, and the lowest (70%) on infants less than 1 year old. From 6 to 15 year old age groups, the classification accuracy is above 80%, and subsequently higher than the model's average accuracy across age groups. In terms of race, the model achieves the highest accuracy (about 81%) on Others and Unknown, and lowest accuracy (about 76%) on Black or African Americans. Finally, we observed slightly better performance on female patients.

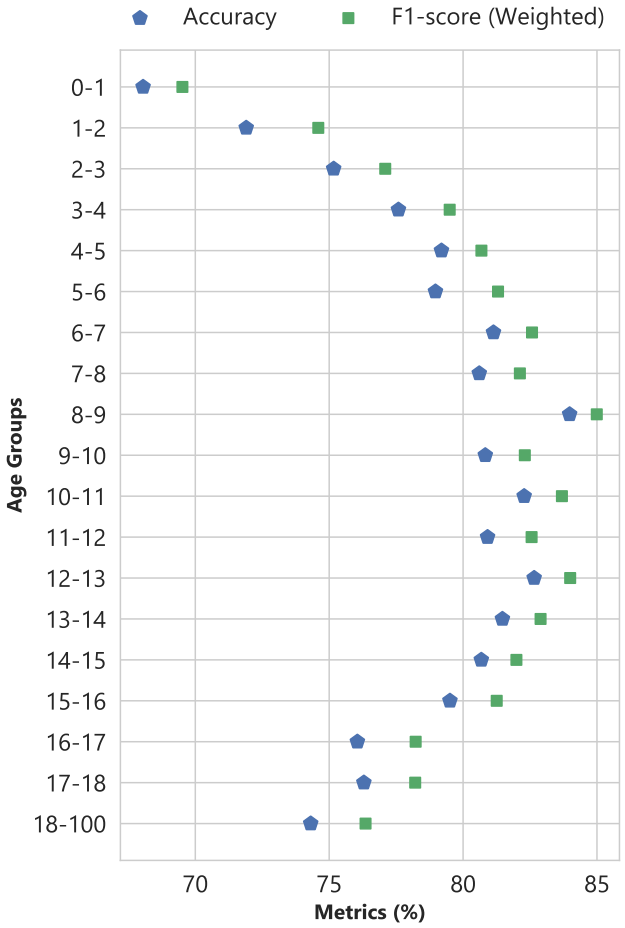


Figure 5. Performance comparison of transformer model on different age groups in the test set, as measured by accuracy and weighted F1-score.

2.5. The predictive power is not from a single EEG channel

Next, we perform an experiment to determine the individual contributions of the EEG channels towards sleep scoring. Seven identical transformer models are created according to the structure described in Figure 1. Then, each model is trained using only one of the seven EEG channels. For example, the first row of Table 3 shows the classification accuracy of a transformer model that only had access to the F4-M1 channel EEG signals during both training and testing. None of the seven models is able to achieve the results of the original transformer model, lending support to the use of multi-channel EEG signals. However, the model trained on F3-M2 channel achieves highest accuracy in classifying sleep stages Wake, N1, and N3, while the C3-M2 channel model does so for N2. Finally, the F4-M1 channel model demonstrates a markedly improved performance in identifying REM stages.

	Accuracy (%)	F1-score (%)
Race		
White	78.6	80.3
Black or African American	76.4	78.3
Multiple Races	78.0	79.6
Asian	78.7	80.3
Others and Unknown	80.6	82.6
Sex		
Male	77.9	79.6
Female	78.5	80.4

Table 2. Transformer model performance on different racial and groups and sex in the test set. Others and Unknown race is defined identically to Table 4. F1 refers to weighted F1-score.

Channel	**Sleep Stage**					
	Wake	N1	N2	N3	REM	All
F4-M1	69.1	39.8	75.0	83.1	**78.0**	75.1
O2-M1	70.1	31.5	74.1	81.4	60.7	71.4
C4-M1	68.2	39.9	77.0	83.5	70.8	74.6
O1-M2	67.7	35.0	73.8	78.7	67.4	71.4
F3-M2	**72.4**	**41.8**	75.5	**84.2**	71.1	75.1
C3-M2	72.2	34.7	**78.3**	83.8	70.8	**75.7**
CZ-O1	69.1	34.6	76.5	80.1	66.5	72.9

Table 3. Classification accuracy (%) on test set for transformer models trained on single EEG channels. The highest accuracy for each sleep stage (column) is bolded.

3. Discussion

We developed and trained a transformer model on more than 3,900 recent pediatric sleep studies collected during standard hospital care. The model predicted 5 sleep stages (Wake, N1, N2, N3, REM) from 7 raw EEG channels (F4-M1, O2-M1, C4-M1, O1-M2, F3-M2, C3-M2, CZ-O1) with 78.2% accuracy, which is the highest accuracy reported for automatic sleep scoring on such a large-scale pediatric dataset to the best of our knowledge.

We believe this work sheds light on many future research ideas for pediatric sleep. First, the challenge in predicting the infrequent N1 stages, while consistent with previous literature, remains an open problem. Prediction performance for infants less than 1 year old also has room for improvement. Finally, as the NCH Sleep DataBank also provides the patients' electronic health records, we plan to build on this work to develop diagnostic models for sleep disorders.

	PSGs, N (Unique Patients, N)		
	Train	Validation	Test
	2812 (2613)	321 (291)	795 (727)
Age			
0-1	157 (132)	26 (21)	59 (43)
1-2	140 (134)	15 (14)	37 (36)
2-3	211 (206)	31 (30)	53 (53)
3-4	189 (187)	31 (31)	57 (55)
4-5	197 (193)	16 (15)	44 (43)
5-6	178 (177)	16 (16)	43 (42)
6-7	178 (176)	16 (16)	48 (46)
7-8	165 (164)	17 (17)	54 (51)
8-9	157 (154)	14 (14)	43 (42)
9-10	136 (134)	18 (18)	38 (35)
10-11	142 (138)	7 (7)	40 (39)
11-12	131 (128)	8 (7)	43 (43)
12-13	136 (130)	9 (9)	34 (34)
13-14	111 (110)	17 (17)	35 (34)
14-15	101 (96)	15 (14)	28 (26)
15-16	132 (123)	13 (12)	23 (22)
16-17	118 (113)	13 (12)	32 (31)
17-18	93 (89)	12 (12)	29 (28)
18-100	140 (131)	27 (25)	55 (46)
Race			
White	1855 (1735)	211 (190)	531 (481)
Black	581 (536)	55 (51)	154 (144)
Multiple race	198 (30)	33 (30)	57 (54)
Asian	71 (60)	11 (9)	29 (24)
Others and Unknown	107 (97)	11 (11)	24 (24)
Sex			
Male	1600 (1471)	185 (166)	450 (408)
Female or Unknown	1212 (1142)	136 (125)	345 (319)

Table 4. Demographic characteristics of the 3,928 PSGs from NCH Sleep DataBank that were used to train, validate, and test our sleep scoring model. N refers to counts; Age is in years; Others and Unknown races include Unknown, Refuse to answer, Native Hawaiian or Other Pacific Islander, and American Indian or Alaska Native, which are aggregated for patient privacy. Note that patients who have gone through multiple sleep studies over the years could have been counted multiple times in different age groups.

4. Experimental Procedures

4.1. Data Description

The NCH Sleep DataBank holds 3,984 pediatric PSG from 3,673 unique patients that were collected between 2017 and 2019 at NCH, Cleveland, Ohio, USA. In this paper, we used 3,928 PSG from 3,631 unique patients that had seven EEG channels of interest (F4-M1, O2-M1, C4-M1, O1-M2, F3-M2, C3-M2, CZ-O1), which is about 98.5% of the dataset. Demographic information is summarized in Table 4, and the distributions of sleep study length are visualized in Figure 6. The PSGs were conducted in standard care at NCH, and all sleep stages were manually scored by a technician and verified by a physician board certified in sleep medicine. Since the EEG signals in this dataset have varying sampling frequency, they were resampled to 128 Hz before training the model. Please see (Lee et al., 2021) for a much more detailed description of the dataset including the de-identification and validation process.

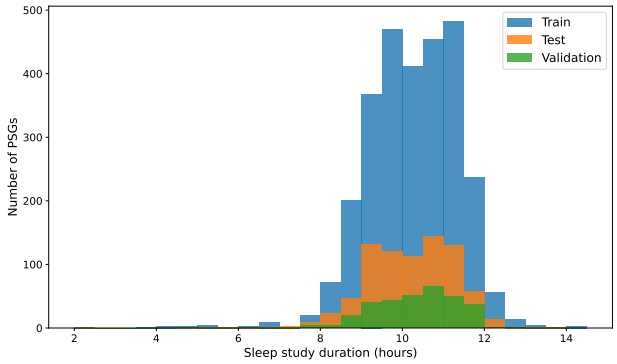


Figure 6. Distributions of sleep study duration in train, test, and validation sets. For all three sets of EDFs, the means were between 10.21 and 10.31 hours, and standard deviations were between 1.05 and 1.09 hours.

4.2. Self-Attention in Transformers

We briefly describe the self-attention mechanism (Vaswani et al., 2017), which is a central building block of the transformer architecture. Self-attention computes a weighted average of tokens (or their representation's) with similarity score being equivalent to weights calculated from pairs of tokens. Given an input sequence with multiple channels $X \in \mathbb{R}^{T\times C}$ of length T and channels C, it is first reshaped into n patches (or tokens) of fixed size, i.e. $X_p \in \mathbb{R}^{n\times(P\cdot C)}$. Once X_p is projected to $X_t \in \mathbb{R}^{n\times d}$ along with the positional information, it is ready to be inputted into the self-attention module in transformers. The normalized importance matrix is computed using three matrices $W_Q \in \mathbb{R}^{d\times d_q}$, $W_K \in \mathbb{R}^{d\times d_k}$, and $W_V \in \mathbb{R}^{d\times d_v}$, which extract query $Q = X_t W_Q$, key $K = X_t W_K$, and value $V = X_t W_V$. The self-attention is then formulated as:

$$\mathcal{F}(Q,K,V) = \texttt{Softmax}\big(\frac{QK^\top}{\sqrt{d_q}}\big)V, \tag{1}$$

where the softmax operation is applied row-wise, and thus each element in the output matrix depends on all other elements in the same row. Building on top of this, the multi-head self-attention layer comprises H independent self-attention layers. Specifically, each head produce a set of query, key and value matrices and compute attention output as: $h_i = \mathcal{F}(Q_i, K_i, V_i)$ for $i = 1, \ldots, H$. Lastly, the fused output is generated by concatenation and linear transformation with learnable weights W_O:

$$\mathcal{M}(Q,K,V) = \texttt{Concat}(h_1, h_2, \ldots, h_H)W_O. \tag{2}$$

For a detailed treatment of how multi-head self-attention and transformers work, we refer the reader to (Park & Kim, 2022). In our model, the parameters $T = 3840, C = 7, n = 30, P = 128, d = d_q = d_k = d_v = 64$, and $H = 4$. T is the signal length or temporal size of the instance, C represents the number of channels, P is the patch size, d is the key (including query and value) dimension, and H denotes the number of heads in the multi-head attention layer.

4.3. Loss Function

We use weighted cross-entropy loss function to train our model as NCH data is slightly imbalanced towards N1 class, i.e., there are fewer samples belonging to N1 sleep stage as compared to rest of the classes. Formally, the objective function we optimize is:

$$\mathcal{L}(\theta) = \frac{1}{M}\sum_{m=1}^{M}\big[w_m \times \mathbb{H}(y_m, f_\theta(y_m|X_m)\big] \tag{3}$$

where M denotes the number of training samples, X_m is the m-th EEG instance in the train set, y_m is the m-th label in the train set, f_θ is the neural network function with learnable parameters θ, and w_m is an instance weight representing the importance that should be given to a particular example. In the case of an imbalanced dataset, the w_m is higher for instances from the minority class while lower or one for the rest. For the N1 class, we found the value of 5 to be optimal, while for the rest of the classes, we used a value of 0.9 as a weighting factor in the loss function. $\mathbb{H}$ is the standard cross-entropy loss. The loss function $\mathcal{L}$ is then optimized with respect to the neural network parameters θ during model training.

4.4. Model Training and Evaluation

We use an Adam optimizer with a default learning rate of 0.001 and batch size of $1,024$ to perform model train-

ing on a single NVIDIA T4 GPU for 25 thousand iterations, iterating over more than 2.5 million multi-channel EEG examples. Our transformer-based model has $775,237$ learnable parameters. We save the model checkpoint at every epoch based on validation set performance to avoid overfitting, and report model performance on the test set. We also experimented with training longer and with an Adam optimizer with weight decay, but we did not notice any improvement in generalization. Finally, we evaluate model performance with four metrics: accuracy, precision, recall, F1-score (macro and weighted averaged variants), and confusion matrix as implemented in the scikit-learn package (Pedregosa et al., 2011). Specifically, the F1-score is the harmonic mean of $\texttt{precision} = \frac{TP}{(TP+FP)}$ and $\texttt{recall} = \frac{TP}{(TP+FN)}$, where TP is True Positive, FP is False Positive, and FN is False Negative. In a multi-class classification setting, the macro average is computed as an unweighted mean of per-class F1-scores. In contrast, the weighted average takes each class's support (i.e., number of samples belonging to a particular class) into consideration.

5. Resource availability

This paper analyzes existing, publicly available dataset for sleep research. The NCH Sleep DataBank can be requested from the National Sleep Research Resource (NSRR) (https://sleepdata.org/datasets/nchsdb) or Physionet (https://physionet.org/content/nch-sleep). The Python code and trained models will be made publicly available on Github upon publication of the manuscript. Further information and requests for resources should be directed to and will be fulfilled by the lead contact: Harlin Lee.

6. Acknowledgements

The work of Harlin Lee was supported by the grant NSF DMS-1952339. The authors declare no competing interests, and thank Yuanting Pan and Lei Xu for their help in data analysis.

References

Bandyopadhyay, A. and Goldstein, C. Clinical applications of artificial intelligence in sleep medicine: a sleep clinician's perspective. *Sleep and Breathing*, pp. 1–17, 2022.

Beebe, D. W., Wells, C. T., Jeffries, J., Chini, B., Kalra, M., and Amin, R. Neuropsychological effects of pediatric obstructive sleep apnea. *J. Int. Neuropsychol. Soc.*, 10(7): 962, 2004.

Berry, R. B., Brooks, R., Gamaldo, C., Harding, S. M., Lloyd, R. M., Quan, S. F., Troester, M. T., and Vaughn, B. V. *The AASM Manual for the scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications. Version 2.4.* American Academy of Sleep Medicine, Darien, IL, 2017.

Berry, R. B., Albertario, C. L., Harding, S. M., Lloyd, R. M., Plante, D. T., Quan, S. F., Troester, M. M., and Vaughn, B. V. *The AASM Manual for the scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications. Version 2.5.* American Academy of Sleep Medicine, Darien, IL, 2018.

Bradley, T. D. and Floras, J. S. Sleep apnea and heart failure: Part i: obstructive sleep apnea. *Circulation*, 107 (12):1671–1678, 2003.

Chen, Z., Yang, Z., Wang, D., Huang, M., Ono, N., Altaf-Ul-Amin, M., and Kanaya, S. An end-to-end sleep staging simulator based on mixed deep neural networks. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 848–853. IEEE, 2021.

Dawson, D. and Reid, K. Fatigue, alcohol and performance impairment. *Nature*, 388(6639):235–235, 1997.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Fiorillo, L., Puiatti, A., Papandrea, M., Ratti, P.-L., Favaro, P., Roth, C., Bargiotas, P., Bassetti, C. L., and Faraci, F. D. Automated sleep scoring: A review of the latest approaches. *Sleep medicine reviews*, 48:101204, 2019.

Jennum, P., Ibsen, R., and Kjellberg, J. Morbidity and mortality in children with obstructive sleep apnoea: a controlled national study. *Thorax*, 68(10):949–954, 2013.

Kim, D., Woo, Y., Jeong, J., Kim, D.-K., and Lee, J.-G. Sleep stage classification for inter-institutional transfer learning. In *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 1797–1800. IEEE, 2021.

Kushida, C. A., Littner, M. R., Morgenthaler, T., Alessi, C. A., Bailey, D., Coleman Jr, J., Friedman, L., Hirshkowitz, M., Kapen, S., Kramer, M., and others. Practice parameters for the indications for polysomnography and related procedures: an update for 2005. *Sleep*, 28(4): 499–523, 2005.

Lee, H., Li, B., DeForte, S., Splaingard, M., Huang, Y., Chi, Y., and Linwood, S. Nch sleep databank: A large collection of real-world pediatric sleep studies. *arXiv preprint arXiv:2102.13284*, 2021.

Lumeng, J. C. and Chervin, R. D. Epidemiology of pediatric obstructive sleep apnea. *Proc. Am. Thorac. Soc.*, 5(2): 242–252, 2008.

Mohsenin, V. Obstructive sleep apnea: a new preventive and therapeutic target for stroke: a new kid on the block. *The American Journal of Medicine*, 128(8):811–816, 2015.

Park, N. and Kim, S. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

Peppard, P. E., Young, T., Palta, M., and Skatrud, J. Prospective study of the association between sleep-disordered breathing and hypertension. *New England Journal of Medicine*, 342(19):1378–1384, 2000.

Phan, H. and Mikkelsen, K. Automatic sleep staging of eeg signals: Recent development, challenges, and future directions. *arXiv preprint arXiv:2111.08446*, 2021.

Phan, H., Mikkelsen, K. B., Chen, O., Koch, P., Mertins, A., and De Vos, M. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Transactions on Biomedical Engineering*, 2022.

Trockman, A. and Kolter, J. Z. Patches are all you need? *arXiv preprint arXiv:2201.09792*, 2022.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Watson, N. F. and Fernandez, C. R. Artificial intelligence and sleep: Advancing sleep medicine. *Sleep medicine reviews*, 59:101512, 2021.

Wulff, K., Gatti, S., Wettstein, J. G., and Foster, R. G. Sleep and circadian rhythm disruption in psychiatric and neurodegenerative disease. *Nature Reviews Neuroscience*, 11 (8):589, 2010.

Yang, Z., Wang, D., Chen, Z., Huang, M., Ono, N., Altaf-Ul-Amin, M., and Kanaya, S. Exploring feasibility of truth-involved automatic sleep staging combined with transformer. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2920–2923. IEEE, 2021.

Zhang, H., Wang, X., Li, H., Mehendale, S., and Guan, Y. Auto-annotating sleep stages based on polysomnographic data. *Patterns*, 3(1):100371, 2022.