Transformer-Based Sleep Staging Using Continuous Photoplethysmography Signals

Joseph A. P. Quino [1], Diego A. C. Cardenas [1], Marcelo A. F. Toledo [1], Felipe M. Dias [1], Estela Ribeiro [1], José E. Krieger [1] and Marco A. Gutierrez [1]

[1] Heart Institute, Clinics Hospital,
University of Sao Paulo Medical School (HCFMUSP)
Sao Paulo – SP – Brazil

{joseph.pena, diego.cardona, marcelo.arruda}@hc.fm.usp.br

{f.dias, estela.ribeiro, j.krieger}@hc.fm.usp.br

marco.gutierrez@incor.usp.br

Abstract. *Sleep disturbances impair health by disrupting the structure of sleep staging. Although polysomnography is the gold standard, its limited availability has driven photoplethysmography (PPG) alternatives. Transformer models have shown strong performance in sleep staging, though mainly with EEG data. In this work, the TransSleepNet is proposed as a hybrid convolutional–transformer architecture that leverages both approaches, along with a flatline detector to identify and manage invalid PPG segments. It achieved Kappa scores of 0.68 (CFS), 0.61 (ABC), and 0.61 (HomePAP). Results show the flatline detector consistently improves performance, while the proposed architecture matches or outperforms baseline models in generalization datasets.*

1. Introduction

Sleep disorders affect human health primarily by disrupting sleep quality. These disturbances often manifest as fragmentation of normal sleep architecture, that is, the physiological sequence of sleep stages occurring throughout the night, including light, deep, and rapid eye movement (REM) sleep stages [Panossian and Avidan 2009, Gottesman et al. 2024].

To diagnose sleep disorders, polysomnography (PSG) is commonly employed. PSG involves the simultaneous recording of multiple physiological signals during sleep, which are subsequently analyzed according to the AASM Manual for the Scoring of Sleep and Associated Events to evaluate architecture and detect abnormalities [Berry et al. 2017]. PSG includes sleep staging, the process of assigning a sleep stage to each 30-second window (epoch) throughout the recording. Although PSG is considered the gold standard for sleep staging, it suffers from certain limitations such as high cost, limited accessibility, and potential alterations in sleep patterns caused but the monitoring setup itself, such as the presence of multiple attached sensors connected and the well-known first-night effect [Byun et al. 2019, Silva et al. 2024].

Wearable devices have emerged as an alternative approach for sleep staging. These devices typically record physiological signals such as photoplethysmography

(PPG) signals [Elgendi 2012]. PPG is an optical sensing technology used to measure changes in blood volume within the microvascular bed of tissue beneath the skin. It operates by illuminating the skin with green, red, or infrared light emitted by a light-emitting diode (LED), while a photodiode-based sensor detects variations in the transmitted or reflected light signal. The recorded signal depends on the relative position of the photodiode with respect to the LED, which determines whether the measurement is based on transmitted or reflected light [Elgendi 2021].

Many methods have been proposed for sleep staging using PPG signals. Early approaches relied on traditional machine learning algorithms that use handcrafted features extracted from PPG signals to classify each epoch independently. More recently, the State Of The Art (SOTA) has shifted toward neural network-based models that process the entire recording as input in order to leverage inter-epoch dependencies.

These methods are predominantly based on convolutional neural network (CNN) architectures, which learn hierarchical representations from continuous PPG signals while capturing temporal patterns across epochs. However, sleep architecture is characterized by cyclical and structured sequences, requiring models capable of capturing both short- and long-term dependencies. Although SOTA methods, such as SleepPPG-Net [Kotzen et al. 2023] and InsightSleepNet [Nam et al. 2024], have demonstrated promising performance in sleep stage classification, their strong reliance on CNN layers may impose architectural limitations. In particular, the inherently local receptive field of CNNs requires either very deep networks or larger dilation factors to capture long-range dependencies, which can lead to unstable training and reduced contextual precision.

Transformers have demonstrated strong capability in capturing both short- and long-term dependencies due to their global receptive field. This property has led to successful applications in several domains, including natural language processes [Vaswani et al. 2023], and more recently in sleep staging using electroencephalogram (EEG) signals rather than PPG [Phan et al. 2022].

Motivated by these advances, this work investigates the use of a transformer-based architecture for sleep staging using PPG signals. The main contributions of this work are (i) the additional preprocessing step that identifies and labels epochs with insufficient physiological information as invalid epochs, and (ii) the incorporation of a transformer module to capture inter-epoch dependencies that explicitly handles these invalid epochs through a masked input strategy.

2. Materials and Methods

This study proposed a deep-learning approach for sleep staging using the PPG signal. The methodology consists of a previously defined dataset and a proposed approach consisting of a preprocessing stage and a neural network architecture. The model is trained on a subset of the available datasets and subsequently evaluated on the remaining datasets to assess its generalization performance.

2.1. Datasets

This work uses five datasets for pretraining, training, validation, and testing. Each of these contains PPG or ECG signals along with sleep stages labeled for each epoch throughout the exam. They are presented in Table 1.

Table 1. Datasets used in this work. The duration column represents values as median and confidence interval at 95%

Dataset	# Subjects	Duration (Hours)	Signal	Usage
SHHS [Quan et al. 1997]	5793	9 (7, 9)	ECG	Pretrain
MESA [Chen et al. 2015]	2056	10 (9, 13)	PPG	Train / Validation / Test
CFS [Redline et al. 1995]	324	10 (8, 11)	PPG	Test
ABC [Bakker et al. 2018]	49	8 (8, 9)	PPG	Test
HomePAP [Rosen et al. 2012]	157	8 (6, 9)	PPG	Test

The first dataset is the *Sleep Heart Health Study* (SHHS) [Zhang et al. 2018, Quan et al. 1997]. Although this dataset contains ECG signals rather than PPG, it was employed for pretraining the proposed model. This choice is justified by the fact that, similar to PPG, ECG signals encode information related to the cardiovascular system. Furthermore, this pretraining strategy has been adopted in prior SOTA approaches [Kotzen et al. 2023].

The second dataset is the *Multi-Ethnic Study of Atherosclerosis* (MESA) [Zhang et al. 2018, Chen et al. 2015]. It is the largest dataset used in this study and is also the most frequently utilized in related works [Carter and Tarassenko 2024, Kotzen et al. 2023, Attia et al. 2025, Yilmaz et al. 2023]. This dataset was partitioned into two subsets: MESA-trainval and MESA-test. To ensure reproducibility and maintain consistency with established SOTA benchmarks, the MESA-test set comprises the 204 subjects defined in the original SleepPPG-Net study [Kotzen et al. 2023]. The MESA-trainval subset includes the remaining 1852 subjects. During model development, this subset was further divided into MESA-train and MESA-val, corresponding to 90% and 10% of the data, respectively, for training and validation.

Additionally, the third, fourth, and fifth datasets are the *Cleveland Family Study* (CFS) [Redline et al. 1995, Zhang et al. 2018], *Apnea, Bariatric Surgery, and Continuous Positive Airway Pressure* (ABC) [Bakker et al. 2018, Zhang et al. 2018], and *Home Positive Airway Pressure* (HomePAP) [Rosen et al. 2012, Zhang et al. 2018] datasets. These test datasets are used exclusively for testing to evaluate the model´s generalization capability. Although smaller than the training datasets, they include PPG recordings, making them suitable for external validation.

2.2. Preprocessing

The preprocessing pipeline applied to both PPG and ECG signals follows the standard approach adopted in SOTA methods [Kotzen et al. 2023, Nam et al. 2024]. It consists of six steps: (i) the PPG signal is first filtered using a low-pass Type II Chebyshev filter with a cutoff frequency of $8\,\mathrm{Hz}$, 8^{th}-order design, and $40\,\mathrm{dB}$ attenuation; (ii) the signal is resampled to $34.13\,\mathrm{Hz}$ to obtain exactly 1024 samples per epoch, as required by the proposed sleep staging method; (iii) the resulting signal is normalized using z-score normalization;

(iv) it is then clipped to three times the standard deviation; (v) each subject's signal is truncated or zero-padded to a uniform duration of 10 hours; and (vi) the signal is segmented into 30-second windows (epochs), yielding one window per label. After preprocessing, windows corresponding to padded segments are labeled as invalid. These windows do not contain meaningful information and are therefore excluded from loss computation during model training.

2.3. Flatline detector

A visual inspection of the PPG signals in the MESA dataset revealed the presence of epochs exhibiting flatline behavior. These segments likely correspond to periods with absent or unreliable physiological information, potentially caused by sensor disconnection, motion artifacts, or signal acquisition failures. One plausible explanation is that the device used to acquire the PPG signal (i.e., the oximeter) was removed before the end of the polysomnography (PSG) recording, resulting in flatline epochs toward the final portion of the signal. Less frequently, such segments are also observed in the middle of the recordings.

Although these PPG epochs lack meaningful information, they are still assigned ground-truth sleep stage labels, as the annotations are derived from electroencephalogram (EEG), electromyogram (EMG), and electrooculogram (EOG) signals rather than from PPG. However, including such epochs in the training set can negatively affect the learning process, as deep learning models may be forced to associate non-informative (flatline) signals with specific sleep stages. Furthermore, their presence may bias evaluation metrics, since these segments would be treated as valid samples during metric aggregation, potentially distorting the reported performance.

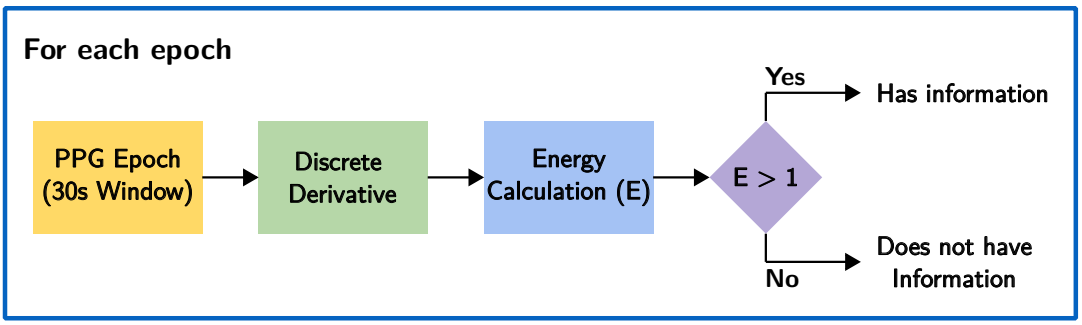
To address this issue, a flatline detection algorithm was developed to identify these segments, remove their associated ground-truth labels, and treat them as invalid epochs, analogous to those introduced during preprocessing via padding. The proposed method, illustrated in Figure 1, aims to detect flatline epochs, defined as signals with near-constant values. The algorithm operates at the epoch level. First, the discrete derivative of the signal is computed to remove constant offsets. Subsequently, the signal energy is calculated. Equation 1 defines the energy computation for the i^{th} PPG epoch x, consisting of 1024 samples. Finally, the computed energy is compared against a predefined threshold (set to 1) to determine whether the epoch contains meaningful information or corresponds to a flatline segment. Since flatline epochs exhibit near-zero energy, epochs with energy below this threshold are classified as flatline and are therefore labeled as invalid.

$$E_i = \sum_{n=0}^{1023} |x_i[n]|^2 \tag{1}$$

2.4. Transformer-based Architecture

The proposed model, depicted in Figure 2, similar to previous approaches, is composed of three steps: an epoch encoder that extracts features from the continuous PPG signal, the sequence encoder to enrich the features per epoch by leveraging the context and by capturing short- and long-term dependencies, and the classifier that uses these enriched

Figure 1. Epoch Filtering procedure to remove flat line windows



Source: Author.

features per epoch and classifies them into a sleep stage: Wakefulness (W), Light (L), Deep (D), and REM [Kotzen et al. 2023, Nam et al. 2024, Attia et al. 2025].

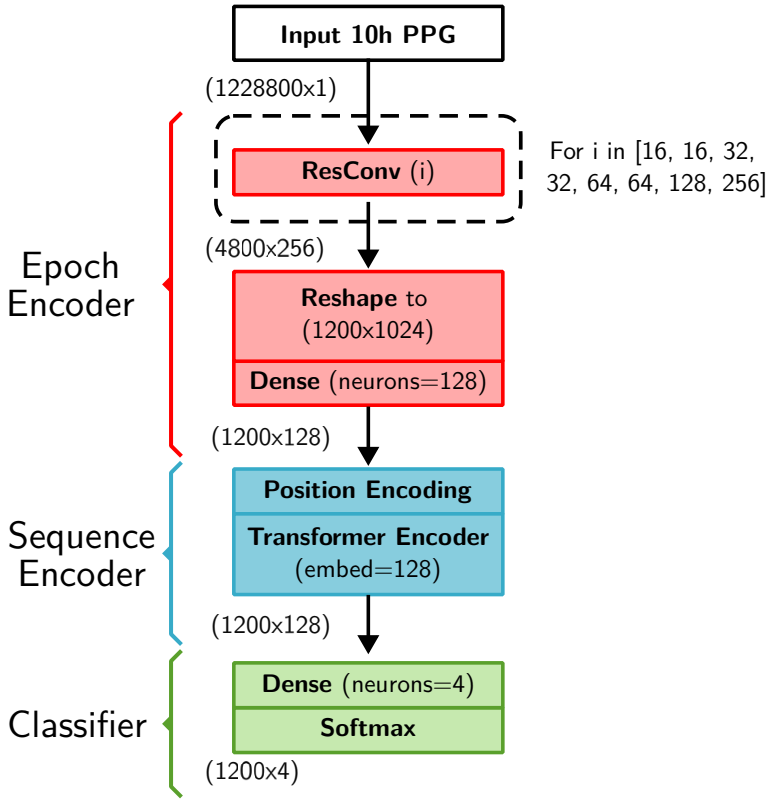
The epoch encoder is based on the architecture proposed in SleepPPG-Net and its subsequent implementation described in [Quino et al. 2025]. This module consists of eight consecutive ResConv blocks which, given an input signal of 10 hours sampled at 34.13 Hz (i.e., 1024 samples per 30-second epoch), progressively downsample the temporal resolution to 4 samples per epoch while increasing the feature dimensionality from 1 to 256. Subsequently, a reshape operation is applied to concatenate every four consecutive samples along the feature dimension, yielding a single sample per epoch with 1024 features. This results in a tensor of shape (1200, 1024). Finally, a fully connected (dense) layer is applied independently to each epoch, reducing the feature dimensionality to 128 and producing an output tensor of shape (1200, 128).

The ResConv block, depicted in Figure 3, consists of three consecutive 1D-CNN layers of kernel 3 and n number of filters each, followed by a residual connection with a 1D-CNN layer of kernel 1 to match the feature dimension, and a max pool layer, which reduces the time dimension by 2. The number of filters n is a variable in the ResConv block, which monotonically increases in the epoch encoder. The ResConv block was initially proposed in [Kotzen et al. 2023], but due to ambiguity in the description, this work follows the implementation from [Quino et al. 2025].

The sequence encoder receives a tensor of shape (1200, 128), which can be interpreted as a sequence of length 1200, where each element is a 128-dimensional embedding. Its objective is to enhance the representation of each epoch by incorporating contextual information from other epochs in the sequence. Given the structured and sequential nature of the hypnogram, this work adopts a Transformer encoder architecture augmented with a positional encoding layer to account for the temporal order of the epochs. Unlike convolutional neural networks, which impose an inductive bias toward local patterns, the Transformer encoder is better suited to capture long-range dependencies across the sequence. The positional encoding layer employs absolute sinusoidal encodings with a maximum sequence length of 1200. The Transformer encoder is configured with an embedding dimension of 128, 4 attention heads, 10 layers, and a feed-forward dimension of 1024.

Furthermore, to mitigate the impact of invalid epochs, a masked input strategy

Figure 2. Proposed Architecture named TransSleepNet



inspired by [Devlin et al. 2019] is employed. Specifically, the attention probabilities associated with invalid epochs are set to zero within the attention mechanism of each Transformer layer, preventing these epochs from contributing to the contextual representation. As a result, epochs identified as invalid, either through flatline detection or artificial padding, do not influence the feature representations of valid epochs.

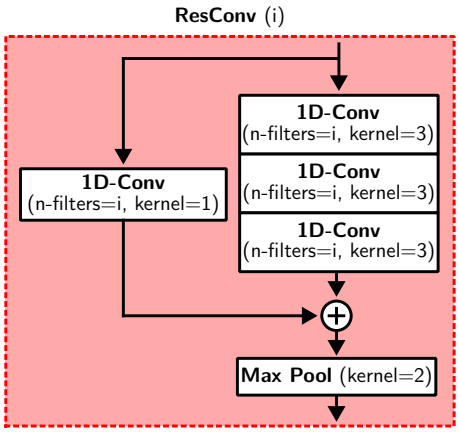
Finally, the classification layer consists of a dense layer executed for each epoch of size 4, followed by a softmax activation function, where the output of each neuron represents the probability of each sleep stage. The resulting sleep stage is selected by using the label with the higher probability.

The hyperparameters used in this architecture, including those of the sequence encoder, were selected after a hyperparameter optimization process using the SHHS dataset split by 90% and 10% for training and validation, respectively. The remaining hyperparameters selected were the activation function – used in the epoch encoder and sequence encoder – as the Gaussian Error Linear Unit (GELU) [Hendrycks and Gimpel 2023], and the dropout of 0.1.

2.5. Training

The model was first pretrained on the SHHS dataset, using 90% of the data for training and 10% for validation. Subsequently, the pretrained model was fine-tuned on the MESA dataset through a transfer learning approach, using the MESA-trainval subset. A seed of 42 and early stopping were used, looking at the accuracy of the validation sub-subset with a patience of 10. Moreover, during training, the model with the per-

Figure 3. ResConv



formance on the validation set, measured by Cohen's kappa, was selected as the final model. The loss function used was the cross-entropy with class weights for data balancing, label smoothing of 0.1, and filtering out those epochs that were invalid during aggregation [Bishop and Bishop 2024]. The ReduceLROPlateau learning scheduler was implemented with a factor of 0.1 and patience of 5 epochs. The AdamW optimizer was used with a learning rate of 0.0001, weight decay of 0.1, and eps of 10^{-7} [Loshchilov and Hutter 2019]. Finally, the experiments were conducted using Python 3.12 and PyTorch 2.6 [Paszke et al. 2019].

3. Results

After applying the flatline epoch detector, the proportions found per dataset are depicted in Table 2. It shows the proportion of flatline epochs found for each class in each dataset. As shown, the class with the greatest proportion of flatline epochs is Wake in all datasets.

Table 2. Percentage of flatline epochs in each class per dataset

Class	MESA (%)	CFS (%)	ABC (%)	HomePAP (%)
Wake	18.09	3.40	6.32	5.15
Light	0.34	0.41	0.07	0.52
Deep	0.23	0.54	0.00	1.93
REM	0.42	0.26	0.03	0.63

The class weights during training on the MESA-trainval were computed based on the valid epochs. Since the number of invalid epochs increased after detecting the flatline epochs, the class weights were affected. These updates are summarized in Table 3. As shown, the largest change occurs in the Wake class, which corresponds to the class with the highest proportion of flatline epochs.

Table 4 presents the results of TransSleepNet in the MESA-test subset, along with the baseline model SleepPPG-Net, which was trained with and without the flatline detector filtering step. The MESA-test was filtered with the flatline detector independently of the model, to avoid result corruption with invalid epochs. As shown, by using the flatline

Table 3. Class weights update

Class	Original Class weight	Filtered Class weight
Wake	0.63	0.72
Light	0.58	0.54
Deep	4.03	3.75
REM	2.29	2.13

epoch detector led to improved performance in SleepPPG-Net. Although TransSleepNet achieved better performance than the original SleepPPG-Net, its performance was lower than that of SleepPPG-Net augmented with the flatline detector. This demonstrates that, when evaluation is conducted on the same dataset used for training, the primary performance gain is due to the flatline detector, rather to the use of the transformer encoder in the sequence modeling stage.

Table 4. MESA-test. Metrics are computed per subject and reported as median and 95% confidence intervals. Statistical significance was assessed using the Wilcoxon signed-rank test, with TransSleepNet as the reference model (paired by subject). Superscripts indicate statistically significant differences with respect to the reference: [a]: p $<$0.05, [b]: p $<$0.01 and [c]: p $<$0.001.

Reference	Kappa	Accuracy	F1 (Wake)	F1 (Light)	F1 (Deep)	F1 (REM)
SleepPPG-Net (Original)	.63 (.36, .81)[c]	.75 (.59, .87)[c]	.88 (.64, .96)	.71 (.54, .83)[c]	.48 (.01, .85)[c]	.78 (.37, .95)[a]
SleepPPG-Net (With flatline filtering)	**.68** (.43, .82)[c]	**.79** (.64, .89)[c]	**.90** (.62, .96)[b]	**.77** (.61, .88)[c]	**.52** (.00, .84)[a]	.79 (.45, .95)
TransSleepNet	.66 (.42, .81)	.78 (.62, .88)	.89 (.64, .97)	.75 (.53, .87)	.51 (.01, .85)	**.80** (.38, .96)

Table 5 presents the results of the SleepPPG-Net implementations and TransSleepNet in the generalization datasets; that is, datasets that were not seen during either training or hyperparameter optimization. As shown, TransSleepNet consistently and statistically significantly outperformed the original SleepPPG-Net in all metrics across the three generalization datasets. Moreover, compared to SleepPPG-Net with a flatline detector, TransSleepNet got a statistically significant improvement in kappa on the CFS dataset but no difference in accuracy. Also, TransSleepNet showed no significant differences on the ABC dataset across any of the evaluated metrics. In contrast, on the HomePAP dataset, TransSleepNet achieved a significant improvement in both metrics. Overall, TransSleepNet outperformed the original SleepPPG-Net across all datasets and achieved performance comparable to, or better than, SleepPPG-Net enhanced with the flatline detector.

4. Discussion

Current sleep staging methods using PPG signal have been developed on the basis of CNN architectures. Though CNNs has demonstrated superior performance in extracting

Table 5. Results in generalization datasets. Metrics are computed per subject and reported as median and 95% confidence intervals. Statistical significance was assessed using the Wilcoxon signed-rank test, with TransSleepNet as the reference model (paired by subject). Superscripts indicate statistically significant differences with respect to the reference: [a]: p <0.05, [b]: p <0.01 and [c]: p <0.001.

	CFS		ABC		HomePAP	
Model	**Kappa**	**Accuracy**	**Kappa**	**Accuracy**	**Kappa**	**Accuracy**
SleepPPG-Net (Original)	.63 (.26, .82) [c]	.74 (.50, .87) [c]	.55 (.15, .74) [c]	.72 (.46, .81) [c]	.54 (.14, .76) [c]	.68 (.40, .84) [c]
SleepPPG-Net (With flatline filtering)	**.68** (.39, .83) [a]	**.79** (.58, .89)	**.62** (.28, .77)	**.78** (.57, .86)	.60 (.08, .80) [b]	.73 (.46, .87) [b]
TransSleepNet	**.68** (.38, .84)	**.79** (.58, .89)	.61 (.28, .80)	.76 (.59, .88)	**.61** (.13, .80)	**.75** (.51, .86)

local pattern information, their fixed local receptive field limits their ability to capture long-range dependencies, often, requiring increased architectural complexity, which can make the learning process more difficult.

In contrast, transformer-based architectures are inherently well-suited to capturing dependencies across long distances, while preserving position information through dedicated encoding mechanisms. Therefore, this work presented two contributions: (i) the first consists of a convolution and transformer-based architecture which leverages convolution layers to extract information intra-epochs, and transformer layers to increase the degree of information by capturing inter-epoch dependencies; and (ii) the second contribution is a flatline epoch detector that marks invalid epochs during the preprocessing step and then, by using masked input in transformer layers, these invalid epochs are handled to not impact the valid ones.

The flatline detection algorithm revealed that the Wake class exhibits the highest prevalence of invalid epochs and was further employed to adjust the class weights used during training. This contribution was found to improve the performance of sleep staging methods independently of the model architecture, including both SleepPPG-Net and TransSleepNet. When evaluated on a subset of the same dataset used for training, TransSleepNet outperformed the original SleepPPG-Net, but achieved similar or lower performance compared to the version of SleepPPG-Net using the flatline detector during training. In contrast, on external generalization datasets, TransSleepNet consistently outperformed or matched the performance of both the original SleepPPG-Net and its flatline-aware variant. These results suggest that incorporating a Transformer-based sequence encoder may enhance generalization capabilities, albeit at the cost of slightly reduced performance when evaluated on the training dataset.

TransSleepNet achieved Cohen's kappa values of 0.68, 0.61, and 0.61 on the CFS, ABC, and HomePAP datasets, respectively. The variability in inter-dataset performance may be attributed to differences in data distributions, including variations in sensors, acquisition protocols, subject populations, and annotation procedures. According to [Landis and Koch 1977], these values correspond to a substantial level of agreement between model predictions and ground-truth labels. For reference, inter-rater Cohen's kappa reported in prior studies is approximately 0.76 [Lee et al. 2022]. Nevertheless, the

confidence intervals are relatively wide, ranging from 0.13 for the HomePAP dataset to 0.84 for the CFS dataset. This variability indicates that, despite promising performance, further work is required to achieve more consistent results across subjects.

5. Conclusion

Two main contributions are presented in this work. First, we propose an algorithm to detect flatline epochs, i.e., signal segments that contain no meaningful physiological information. Second, we introduced a transformer-based module into the sleep staging architecture to capture both short- and long-term dependencies, while handling invalid epochs through a masking strategy. The results indicate that the flatline detector is the primary factor improving sleep stage performance, independently of the underlying model. In contrast, the transformer-based architecture achieved performance comparable to or better than baseline methods when evaluated on external datasets, demonstrating its potential to improve model generalization.

6. Acknowledgments

This work was supported by Foxconn Brazil and Zerbini Foundation as part of the research project "Remote Patient Monitoring System".

References

Attia, S., Hershkovich, R. S., Tabakhov, A., Ang, A., Oksenberg, A., Tauman, R., and Behar, J. A. (2025). SleepPPG-Net2: Deep learning generalization for sleep staging from photoplethysmography. *Physiological Measurement*, page 23.

Bakker, J. P., Tavakkoli, A., Rueschman, M., Wang, W., Andrews, R., Malhotra, A., Owens, R. L., Anand, A., Dudley, K. A., and Patel, S. R. (2018). Gastric Banding Surgery versus Continuous Positive Airway Pressure for Obstructive Sleep Apnea: A Randomized Controlled Trial. *American Journal of Respiratory and Critical Care Medicine*, 197(8):1080–1083.

Berry, R. B., Brooks, R., Gamaldo, C., Harding, S. M., Lloyd, R. M., Quan, S. F., Troester, M. T., and Vaughn, B. V. (2017). AASM Scoring Manual Updates for 2017 (Version 2.4). *Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine*, 13(5):665–666.

Bishop, C. M. and Bishop, H. (2024). *Deep Learning: Foundations and Concepts*. Springer International Publishing, Cham, 1 edition.

Byun, J.-H., Kim, K. T., Moon, H.-j., Motamedi, G. K., and Cho, Y. W. (2019). The first night effect during polysomnography, and patients' estimates of sleep quality. *Psychiatry Research*, 274:27–29.

Carter, J. F. and Tarassenko, L. (2024). Wav2sleep: A Unified Multi-Modal Approach to Sleep Stage Classification from Physiological Signals.

Chen, X., Wang, R., Zee, P., Lutsey, P. L., Javaheri, S., Alcántara, C., Jackson, C. L., Williams, M. A., and Redline, S. (2015). Racial/Ethnic Differences in Sleep Disturbances: The Multi-Ethnic Study of Atherosclerosis (MESA). *Sleep*, 38(6):877–888.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Elgendi, M. (2012). On the Analysis of Fingertip Photoplethysmogram Signals. *Current Cardiology Reviews*, 8(1):12.

Elgendi, M. (2021). *PPG Signal Analysis; An Introduction Using MATLAB*. Taylor & Francis, London, 1 edition.

Gottesman, R. F., Lutsey, P. L., Benveniste, H., Brown, D. L., Full, K. M., Lee, J.-M., Osorio, R. S., Pase, M. P., Redeker, N. S., Redline, S., Spira, A. P., and on behalf of the American Heart Association Stroke Council; Council on Cardiovascular and Stroke Nursing; and Council on Hypertension (2024). Impact of Sleep Disorders and Disturbed Sleep on Brain Health: A Scientific Statement From the American Heart Association. *Stroke*, 55(3).

Hendrycks, D. and Gimpel, K. (2023). Gaussian Error Linear Units (GELUs).

Kotzen, K., Charlton, P. H., Salabi, S., Amar, L., Landesberg, A., and Behar, J. A. (2023). SleepPPG-Net: A Deep Learning Algorithm for Robust Sleep Staging From Continuous Photoplethysmography. *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS*, 27(2):9.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Lee, Y. J., Lee, J. Y., Cho, J. H., and Choi, J. H. (2022). Interrater reliability of sleep stage scoring: A meta-analysis. *Journal of Clinical Sleep Medicine*, 18(1):193–202.

Loshchilov, I. and Hutter, F. (2019). Decoupled Weight Decay Regularization.

Nam, B., Bark, B., Lee, J., and Kim, I. Y. (2024). InsightSleepNet: The interpretable and uncertainty-aware deep learning network for sleep staging using continuous Photoplethysmography. *BMC Medical Informatics and Decision Making*, 24(1):15.

Panossian, L. A. and Avidan, A. Y. (2009). Review of sleep disorders. *The Medical Clinics of North America*, 93(2):407–425, ix.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., dAlché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, New York. Curran Associates, Inc.

Phan, H., Mikkelsen, K., Chén, O. Y., Koch, P., Mertins, A., and De Vos, M. (2022). SleepTransformer: Automatic Sleep Staging with Interpretability and Uncertainty Quantification. *IEEE Transactions on Biomedical Engineering*, 69(8):2456–2467.

Quan, S. F., Howard, B. V., Iber, C., Kiley, J. P., Nieto, F. J., O'Connor, G. T., Rapoport, D. M., Redline, S., Robbins, J., Samet, J. M., and Wahl, P. W. (1997). The Sleep Heart Health Study: Design, rationale, and methods. *Sleep*, 20(12):1077–1085.

Quino, J. A. P., Cardenas, D. A. C., Toledo, M. A. F., Dias, F. M., Ribeiro, E., Krieger, J. E., and Gutierrez, M. A. (2025). Enhancing Photoplethysmography-Based Sleep Staging Models Through Temporal Context Optimization. In *MEDINFO*, Studies in Health Technology and Informatics, page 5, Bristol. IOP Publishing.

Redline, S., Tishler, P. V., Tosteson, T. D., Williamson, J., Kump, K., Browner, I., Ferrette, V., and Krejci, P. (1995). The familial aggregation of obstructive sleep apnea. *American Journal of Respiratory and Critical Care Medicine*, 151(3 Pt 1):682–687.

Rosen, C. L., Auckley, D., Benca, R., Foldvary-Schaefer, N., Iber, C., Kapur, V., Rueschman, M., Zee, P., and Redline, S. (2012). A Multisite Randomized Trial of Portable Sleep Studies and Positive Airway Pressure Autotitration Versus Laboratory-Based Polysomnography for the Diagnosis and Treatment of Obstructive Sleep Apnea: The HomePAP Study. *Sleep*, 35(6):757–767.

Silva, D. I. C. D., Corrêa, C. D. C., Barros, J. L. D., Marão, A. C., and Weber, S. A. T. (2024). Accessibility to manage the obstructive sleep apnea within the Brazilian Unified Health System. *Brazilian Journal of Otorhinolaryngology*, 90(1):101338.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention Is All You Need.

Yilmaz, G., Ong, J. L., Ling, L.-H., and Chee, M. W. L. (2023). Insights into vascular physiology from sleep photoplethysmography. *Sleep*, 46(10):11.

Zhang, G.-Q., Cui, L., Mueller, R., Tao, S., Kim, M., Rueschman, M., Mariani, S., Mobley, D., and Redline, S. (2018). The National Sleep Research Resource: Towards a sleep data commons. *Journal of the American Medical Informatics Association*, 25(10):1351–1358.