SLEEPFM: MULTI-MODAL REPRESENTATION LEARNING FOR SLEEP ACROSS BRAIN ACTIVITY, ECG AND RESPIRATORY SIGNALS

Rahul Thapa
Department of Biomedical Data Science
Stanford University
rthapa84@stanford.edu

Bryan He
Department of Computer Science
Stanford University

Magnus Ruud Kjær, Gauri Ganjoo, Hyatt Moore & Emmanuel Mignot
Department of Psychiatry and Behavioral Sciences, Stanford University

James Zou
Department of Biomedical Data Science, Stanford University

ABSTRACT

Sleep is a complex physiological process evaluated through various modalities recording electrical brain, cardiac, and respiratory activities. We curate a large polysomnography dataset from over 14,000 participants comprising over 100,000 hours of sleep recordings. Leveraging this extensive dataset, we developed *SleepFM*, the first multi-modal foundation model for sleep analysis. We show that a novel leave-one-out contrastive learning significantly improves downstream task performance compared to standard pairwise contrastive learning. A logistic regression model trained on *SleepFM*'s learned embeddings outperforms an end-to-end trained convolutional neural network (CNN) on sleep stage classification (macro AUROC 0.88 vs 0.72 and macro AUPRC 0.72 vs 0.48) and sleep disordered breathing detection (AUROC 0.85 vs 0.69 and AUPRC 0.77 vs 0.61). Notably, the learned embeddings achieve 48% top-1 average accuracy in retrieving modality clip pairs from 90,000 candidates. This work demonstrates the value of holistic multi-modal sleep modeling to fully capture the richness of sleep recordings.

1 INTRODUCTION

Sleep monitoring is critical to evaluate sleep disorders and as a proxy to assess overall brain, pulmonary, and cardiac health Worley (2018); Brink-Kjaer et al. (2022); Leary et al. (2021). Polysomnography (PSG) is the current gold standard for studying sleep by recording diverse physiological signals such as electroencephalogram (EEG), electroocculograms (EOG), electromyography (EMG), electrocardiogram (ECG) and respiratory channels Kryger et al. (2010). EOG and EMG are often combined with EEG recordings to determine sleep stages, referred to as Brain Activity Signals (BAS).

Traditionally, sleep data analysis involved manual visual inspection, a labor-intensive and time-consuming process prone to errors Boashash & Ouelha (2016); Hassan & Bhuiyan (2017). Recent advancements in supervised deep learning have shown promise in automating sleep staging and classification of disorders like SDB Nassi et al. (2021); Perslev et al. (2021); Stephansen et al. (2018). However, most methods rely on labeled data from a narrow task. They rarely leverage the full breadth of unlabeled physiological dynamics within and across diverse PSG sensors.

In parallel, contrastive learning (CL) techniques have emerged to enable comprehensive representation learning, with major computer vision frameworks like SimCLR Chen et al. (2020), and CLIP Radford et al. (2021) focused primarily on images. While some works have explored extending CL to medical images and time series like ECG signals, multi-modal contrastive representation learning across diverse physiological modalities remains relatively uncharted. Two prior studies have investigated

contrastive multi-modal clinical time series analysis: one employing SimCLR-style pre-training on ECG data and structured records Raghu et al. (2022), and another deriving ECG representations by contrasting ECGs with electronic health records and clinical notes Lalam et al. (2023). More related works included in Appendix A.1.

Our Contribution We introduce *SleepFM*, a foundation model for sleep analysis trained using CL on a multi-modal PSG dataset comprising over 100,000 hours of sleep monitoring data from over 14,000 participants at [anonymized] sleep clinic collected between 1999 and 2020. By combining BAS, ECG, and respiratory modalities from PSG, *SleepFM* exhibits superior performance on tasks such as demographic attributes, sleep stage, and SDB event classifications, outperforming end-to-end trained CNN models. Additionally, we introduce a novel leave-one-out approach for CL, which significantly outperforms the standard pairwise CL on all of our downstream tasks. To our knowledge, this is the first attempt to build and evaluate a foundation model for sleep analysis.

2 METHOD

2.1 DATASET AND PREPROCESSING

Our dataset encompasses PSG records from a sleep clinic from 1999-2020. Comprising 14,068 recordings, this dataset features diverse waveforms collected over approximately 8 hours per individual. Our preprocessing strategy aimed to make minimal alterations to preserve raw signals crucial for nuanced pattern recognition. Each recording consists of three modalities: BAS, ECG, and respiratory, encompassing 10, 2, and 7 channels, respectively. BAS includes brain activity from various brain regions, as well as EOG for eye movement and EMG for chin muscle activation. ECG contains channels that measure electrical cardiac function. Respiratory includes channels measuring chest and abdomen movements, pulse readings, nasal and oral flow measurements. The selection of these channels was guided by sleep experts and relevant sleep analysis literature Berry et al. (2012).

Subsequently, we segmented the total sleep duration into 30-second clips, following the standard Berry et al. (2012). We resampled the dataset to 256 Hz to standardize the sampling rate. Expert sleep technicians labeled each clip for both sleep stage and SDB. Sleep stage is categorized into Wake, Stage 1, Stage 2, Stage 3, REM, and SDB is a binary label. To prevent data leakage, the dataset is split into participant-level pretrain/train/validation/test sets consisting of 11,261, 1,265, 141, and 1,401 participants respectively. The pretrain dataset is only used to pretrain our foundation model. The remaining set serves to train and test models for downstream applications as explained in Section 3. The validation set is used to optimize the hyperparameters. Demographic statistics for different splits are presented in Table 6. An illustrative snapshot of our data can be found in Figure 3.

2.2 MULTI-MODAL CONTRASTIVE LEARNING

We trained three separate 1D CNNs to generate embeddings separately for each modalities. The architecture is based on CNN developed for classifying ECG measurements Ouyang et al. (2022). We explore two CL frameworks: pairwise CL and leave-one-out CL (Figure 1). The key idea is to bring positive pairs of embeddings from different modalities closer in the latent space while pushing apart negative pairs. The positive pairs are derived from temporally aligned 30-second clips across modalities. All other non-matching instances a training batch are negative pairs. In pairwise CL, we construct prediction tasks between all pairs of modalities. For modalities i and j and sample k in a batch, we have an embedding x_k^i from modality i and an embedding x_k^j from modality j. The loss is:

$$l_{i,j,k}^{\text{pair}} = -\log \frac{\exp(\text{sim}(x_k^i, x_k^j) * \exp(\tau))}{\sum_{m=1}^{N} \exp(\text{sim}(x_k^i, x_m^j) * \exp(\tau))}, \tag{1}$$

where N is the number of samples in a batch, and τ is a trainable temperature parameter. We sum this loss over all the samples in a batch and repeat the process for all pairs of modalities i, j.

In leave-one-out CL, for each modality i, we construct an embedding $\bar{x}^{\neq i}$ by averaging over embeddings from all other modalities, excluding modality i. The loss is:

$$l_{i,k}^{\text{LOO}} = -\log \frac{\exp(\text{sim}(x_k^i, \bar{x}_k^{\neq i}) * \exp(\tau))}{\sum_{m=1}^{N} \exp(\text{sim}(x_k^i, \bar{x}_m^{\neq i}) * \exp(\tau))} \tag{2}$$

This is the loss for a sample k from modality i in a given batch. The motivation behind the leave-one-out method is to encourage each embedding to capture semantics aligned with all other modalities. Our pretraining/training details are available in Appendix A.3.

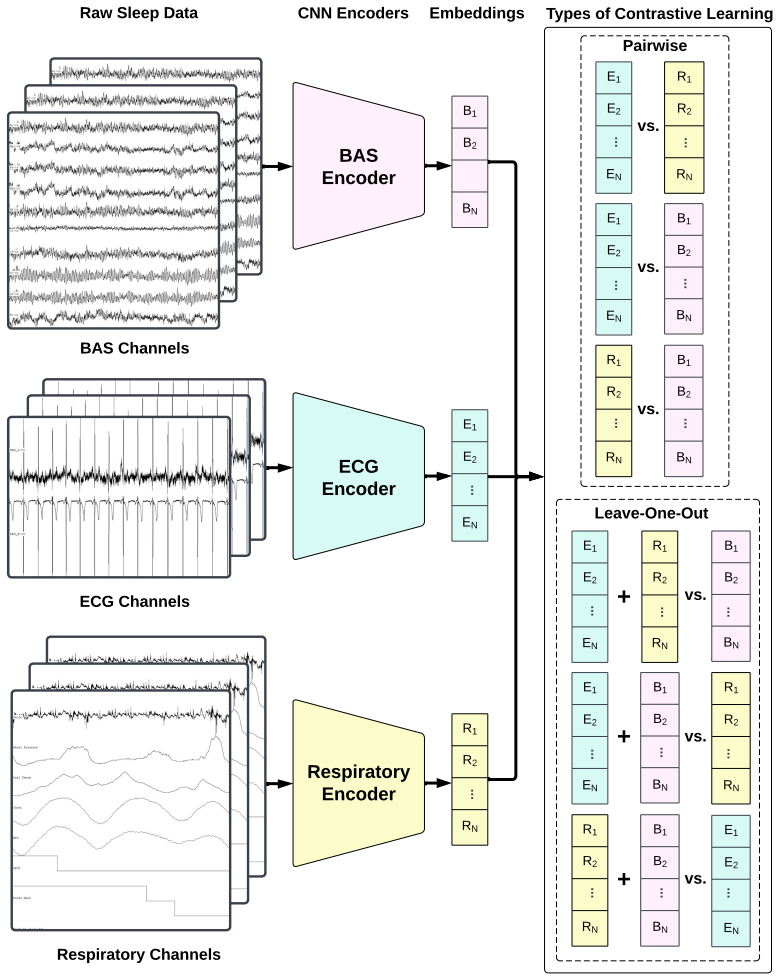


Figure 1: Overview of *SleepFM* pre-training with CL. We experiment with two types of pre-training: standard pairwise CL where we contrast embeddings from each pair of modalities separately, and our novel leave-one-out CL where we contrast the embedding of each modality against the average embedding of all other modalities. BAS measures Brain Activity Signals, ECG measures heart activity, and Respiratory channels measure chest, abdomen movements, pusle, nasal, and oral flow.

3 Experiments and Results

3.1 Retrieval Analysis

Table 1: Retrieval on the test set for our pretrained model. Random baseline for Recall@10 = 0.0001

	Leave-one-out			**Pairwise**		
	BAS	ECG	Resp	BAS	ECG	Resp
BAS	-	0.58	0.05	-	0.74	0.58
ECG	0.46	-	0.39	0.82	-	0.81
Resp	0.05	0.38	-	0.60	0.82	-

We assessed *SleepFM*'s retrieval capabilities by retrieving one modality's closest embeddings from the test set based on another modality. Evaluation was measured using recall@10 rank metrics, which measures the true paired item's appearance within the top 10 recommendations. We measured the performance using 90,000 randomly selected clips from the test set. We uniformly selected clips from various event types within the test set. The Recall@10 for random retrievals is $10/90000 = 0.0001$.

SleepFM achieved 500x-8000x higher Recall@10 than the random chance as shown in Table 1. Pairwise CL yields better overall retrieval performance than leave-one-out, likely because the retrieval evaluation directly maps the training procedure of pairwise. Retrieval performance between respiratory and other modalities is comparatively worse. This discrepancy may stem from the higher variablilty of the respiratory measurements. While BAS and ECG are directly measured via electrical activity from brain and heart respectively, the respiratory channels indirectly measure breathing through the movement of the participant, which can be influenced by body position and other motion.

3.2 Downstream Classification Tasks

Table 2: Sleep stage classification. LOO stands for leave-one-out. ± represents 95% CI.

Macro AUROC			Macro AUPRC		
LOO	**Pairwise**	**Baseline**	**LOO**	**Pairwise**	**Baseline**
0.906	0.876	0.842	**0.685**	0.608	0.579

Table 3: SDB classification. LOO stands for leave-one-out. ± represents 95% CI.

AUROC			AUPRC		
LOO	**Pairwise**	**Baseline**	**LOO**	**Pairwise**	**Baseline**
0.941	0.902	0.843	**0.711**	0.586	0.555

Table 4: Age classification. ± represents 95% CI.

AUROC			AUPRC		
LOO	**Pairwise**	**Baseline**	**LOO**	**Pairwise**	**Baseline**
0.883	0.851	0.724	**0.716**	0.664	0.481

We now evaluate performance on clinically useful downstream tasks: sleep stage and SDB classification. We used the embeddings learned by *SleepFM* to train a logistic regression model and classify sleep stages and SDB events on a held-out test dataset. Sleep stage classification is a multi-class classification task, with 5 classes: Wake, Stage 1, Stage 2, Stage 3, and REM. Prevalence of these groups are 0.21, 0.07, 0.51, 0.09, and 0.12 respectively. SDB classification is a binary classification task, with a prevalence of 0.017. We compared *SleepFM* performance with end-to-end CNN trained on all three modalities, for sleep stage and SDB event classification.

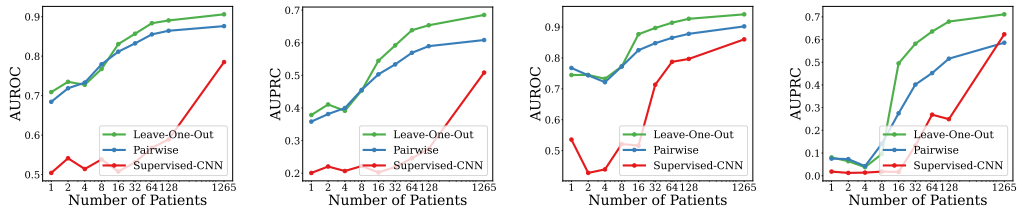
The results for sleep stage classification are presented in Table 2. Notably, across both AUROC and AUPRC metrics, the logistic regression model trained using representations from *SleepFM* outperforms the CNN trained end-to-end in a supervised manner. This superiority holds true across all sleep stage classes as shown in Table 9. Similarly, the SDB classification metrics, displayed in Table 3, underscore our approach's superiority over supervised CNN models. We find that the model pretrained with leave-one-out CL significantly outperforms the model pretrained with pairwise.

Additionally, we also evaluated our model's performance on age and gender classification. We grouped ages into categories of 0-18, 18-35, 35-50, and 50+ and considered male vs female for gender. Our model significantly outperforms the CNN baseline on both tasks, demonstrating it captures salient demographic information effectively (Table 4 and Table 5). The model trained with leave-one-out CL performed best. Analyzing per modality, the BAS signals showed the most distinctive features, indicating they provide useful demographic cues (Table 16, Table 17).

Table 5: Gender classification. Prevalence of female is 0.41. ± represents 95% Confidence Intervals.

AUROC			AUPRC		
LOO	**Pairwise**	**Baseline**	**LOO**	**Pairwise**	**Baseline**
0.850	0.810	0.690	**0.774**	0.731	0.614

3.3 Few-Shot Evaluation



(a) Sleep Stages AUROC (b) Sleep Stages AUPRC (c) SDB AUROC (d) SDB AUPRC

Figure 2: Few Shot Evaluation. Shot 1265 is the total size of our training dataset. Testing is done on the entire test set. Performance average across 3 replicates.

We performed a few-shot performance evaluation to understand how our model performs in low data setting. We steadily increased the number of participants k that each model sees from $k = 1$ to the full training dataset, and recorded the model's AUROC and AUPRC at each k. For the supervised CNN, few-shot examples are the only training examples seen by the model. For the pretrained models, we use embeddings of these few-shot examples to train a logistic regression model.

We see that across all training set sizes, *SleepFM* significantly outperforms baseline supervised CNN model for both sleep stage and SDB classification (Figure 2). The leave-one-out model significantly outperforms pairwise model across all training set sizes, especially for SDB classification.

4 Discussion and Conclusion

Our study develops and evaluates a multi-modal contrastive learning model for sleep analysis, using polysomnography data across over 100,000 hours of sleep from 14,000 patients. The model exhibited strong performance across demographic attributes classification, retrieval analysis, sleep stage classification, and SDB event detection, surpassing end-to-end trained CNNs. The methodology centers on two CL approaches, leave-one-out and pairwise, which both effectively unified BAS, ECG, and respiratory signal representations and demonstrated efficacy in limited data scenarios. Interestingly, we find that pairwise CL is better suited for cross-modality retrieval, while leave-one-out CL is best for learning representations for downstream sleep stage and SDB classification. This might be due to leave-one-modality-out training encourages the model to learn a more unified representation integrating different modalities.

Future Work. Limitations of this study include a reliance on a single institution's data, highlighting opportunities for expanded model evaluation and pretraining across diverse sleep data. Priorities for future includes multi-site pretraining and application to additional clinically meaningful tasks like narcolepsy to enable comprehensive sleep evaluation.

Acknowledgments

RT gratefully acknowledges funding from a Knight-Hennessy graduate fellowship.

References

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to

exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15016–15027, 2023.

Richard B Berry, Rita Brooks, Charlene E Gamaldo, Susan M Harding, C Marcus, Bradley V Vaughn, et al. The AASM manual for the scoring of sleep and associated events. *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine*, 176:2012, 2012.

Boualem Boashash and Samir Ouelha. Automatic signal abnormality detection using time-frequency features and machine learning: A newborn EEG seizure case study. *Knowledge-Based Systems*, 106:38–50, 2016.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pp. 1–21. Springer, 2022.

Andreas Brink-Kjaer, Eileen B Leary, Haoqi Sun, M Brandon Westover, Katie L Stone, Paul E Peppard, Nancy E Lane, Peggy M Cawthon, Susan Redline, Poul Jennum, et al. Age estimation from sleep studies using deep learning predicts life expectancy. *NPJ digital medicine*, 5(1):103, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Bryan Gopal, Ryan Han, Gautham Raghupathi, Andrew Ng, Geoff Tison, and Pranav Rajpurkar. 3kg: Contrastive learning of 12-lead electrocardiograms using physiologically-inspired augmentations. In *Machine Learning for Health*, pp. 156–167. PMLR, 2021.

Rim Haidar, Stephen McCloskey, Irena Koprinska, and Bryn Jeffries. Convolutional neural networks on multiple respiratory channels to detect hypopnea and obstructive apnea events. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, 2018.

Ahnaf Rashik Hassan and Mohammed Imamul Hassan Bhuiyan. Automated identification of sleep states from EEG signals by means of ensemble empirical mode decomposition and random under sampling boosting. *Computer methods and programs in biomedicine*, 140:201–210, 2017.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3942–3951, 2021.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, pp. 5606–5615. PMLR, 2021.

Meir H Kryger, Thomas Roth, and William C Dement. Principles and practice of sleep medicine fifth edition, 2010.

Sravan Kumar Lalam, Hari Krishna Kunderu, Shayan Ghosh, Harish Kumar, Samir Awasthi, Ashim Prasad, Francisco Lopez-Jimenez, Zachi I Attia, Samuel Asirvatham, Paul Friedman, et al. ECG representation learning with multi-modal EHR data. *Transactions on Machine Learning Research*, 2023.

Eileen B Leary, Katie L Stone, and Emmanuel Mignot. Living to dream—reply. *JAMA neurology*, 78 (4):495–496, 2021.

Ming Y Lu, Bowen Chen, Andrew Zhang, Drew FK Williamson, Richard J Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19764–19775, 2023.

Nicola Michielli, U Rajendra Acharya, and Filippo Molinari. Cascaded lstm recurrent neural network for automated sleep stage classification using single-channel eeg signals. *Computers in biology and medicine*, 106:71–81, 2019.

Sheikh Shanawaz Mostafa, Fabio Mendonca, Antonio G Ravelo-Garcia, Gabriel Gabriel Juliá-Serdá, and Fernando Morgado-Dias. Multi-objective hyperparameter optimization of convolutional neural network for obstructive sleep apnea detection. *IEEE Access*, 8:129586–129599, 2020.

Sajad Mousavi, Fatemeh Afghah, and U Rajendra Acharya. Sleepeegnet: Automated sleep stage scoring with sequence to sequence deep learning approach. *PloS one*, 14(5):e0216456, 2019.

Thijs E Nassi, Wolfgang Ganglberger, Haoqi Sun, Abigail A Bucklin, Siddharth Biswal, Michel JAM van Putten, Robert J Thomas, and M Brandon Westover. Automated scoring of respiratory events in sleep with a single effort belt and deep neural networks. *IEEE transactions on biomedical engineering*, 69(6):2094–2104, 2021.

Anders Vestergaard Nørskov, Alexander Neergaard Zahid, and Morten Mørup. CSLP-AE: A contrastive split-latent permutation autoencoder framework for zero-shot electroencephalography signal conversion. *arXiv preprint arXiv:2311.07788*, 2023.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

David Ouyang, John Theurer, Nathan R Stein, J Weston Hughes, Pierre Elias, Bryan He, Neal Yuan, Grant Duffy, Roopinder K Sandhu, Joseph Ebinger, et al. Electrocardiographic deep learning for predicting post-procedural mortality. *arXiv preprint arXiv:2205.03242*, 2022.

Mathias Perslev, Sune Darkner, Lykke Kempfner, Miki Nikolic, Poul Jørgen Jennum, and Christian Igel. U-sleep: Resilient high-frequency sleep staging. NPJ digital medicine, 4 (1), 72, 2021.

Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. Seqsleepnet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(3):400–410, 2019.

Huy Phan, Oliver Y Chén, Minh C Tran, Philipp Koch, Alfred Mertins, and Maarten De Vos. Xsleepnet: Multi-view sequential model for automatic sleep staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5903–5915, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Aniruddh Raghu, Payal Chandak, Ridwan Alam, John Guttag, and Collin Stultz. Contrastive pretraining for multimodal medical time series. In *NeurIPS 2022 Workshop on Learning from Time Series for Health*, 2022.

Hogeon Seo, Seunghyeok Back, Seongju Lee, Deokhwan Park, Tae Kim, and Kyoobin Lee. Intra-and inter-epoch temporal context network (iitnet) using sub-epoch features for automatic sleep scoring on raw single-channel eeg. *Biomedical signal processing and control*, 61:102037, 2020.

Arnaud Sors, Stéphane Bonnet, Sébastien Mirek, Laurent Vercueil, and Jean-François Payen. A convolutional neural network for sleep stage scoring from raw single-channel EEG. *Biomedical Signal Processing and Control*, 42:107–114, 2018.

Jens B Stephansen, Alexander N Olesen, Mads Olsen, Aditya Ambati, Eileen B Leary, Hyatt E Moore, Oscar Carrillo, Ling Lin, Fang Han, Han Yan, et al. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nature communications*, 9(1):5229, 2018.

Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, 2017.

RK Tripathy, Pranjali Gajbhiye, and U Rajendra Acharya. Automated sleep apnea detection from cardio-pulmonary signal using bivariate fast and adaptive emd coupled with cross time–frequency analysis. *Computers in Biology and Medicine*, 120:103769, 2020.

O Tsinalis, PM Matthews, Y Guo, and S Zafeiriou. Automatic sleep stage scoring with single-channel EEG using convolutional neural networks. arxiv 2016. *arXiv preprint arXiv:1610.01683*.

Orestis Tsinalis, Paul M Matthews, and Yike Guo. Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders. *Annals of biomedical engineering*, 44:1587–1597, 2016.

Erdenebayar Urtnasan, Jong-Uk Park, and Kyoung-Joung Lee. Automatic detection of sleep-disordered breathing events using recurrent neural networks from an electrocardiogram signal. *Neural computing and applications*, 32:4733–4742, 2020.

Susan L Worley. The extraordinary importance of sleep: the detrimental effects of inadequate sleep on health and public safety drive an explosion of sleep research. *Pharmacy and Therapeutics*, 43 (12):758, 2018.

Minsoo Yeo, Hoonsuk Byun, Jiyeon Lee, Jungick Byun, Hak-Young Rhee, Wonchul Shin, and Heenam Yoon. Respiratory event detection during sleep using electrocardiogram and respiratory related signals: Using polysomnogram and patch-type wearable device data. *IEEE Journal of Biomedical and Health Informatics*, 26(2):550–560, 2021.

Ozal Yildirim, Ulas Baran Baloglu, and U Rajendra Acharya. A deep learning model for automated sleep stages classification using PSG signals. *International journal of environmental research and public health*, 16(4):599, 2019.

Hui Yu, Dongyi Liu, Jing Zhao, Zhen Chen, Chengxiang Gou, Xueying Huang, Jinglai Sun, and Xiaoyun Zhao. A sleep apnea-hypopnea syndrome automatic detection and subtype classification method based on LSTM-CNN. *Biomedical Signal Processing and Control*, 71:103240, 2022.

Alexander Neergaard Zahid, Poul Jennum, Emmanuel Mignot, and Helge BD Sorensen. MSED: A multi-modal sleep event detection model for clinical sleep analysis. *IEEE Transactions on Biomedical Engineering*, 2023.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pp. 2–25. PMLR, 2022.

Xiaoyun Zhao, Xiaohong Wang, Tianshun Yang, Siyu Ji, Huiquan Wang, Jinhai Wang, Yao Wang, and Qi Wu. Classification of sleep apnea based on EEG sub-band signal characteristics. *Scientific Reports*, 11(1):5824, 2021.

A Appendix

A.1 Related Work

A.1.1 Machine Learning for Analyzing Sleep Data

The application of machine learning (ML) in sleep studies has garnered significant recent attention, promising to streamline and expedite the sleep scoring process as well as detecting respiratory events such as SDB. Models including autoencoders Tsinalis et al. (2016), convolutional neural networks (CNNs) Tsinalis et al.; Sors et al. (2018); Yildirim et al. (2019), recurrent neural networks (RNNs) Michielli et al. (2019); Phan et al. (2019), and multiple other variations of deep neural networks (DNNs) Supratak et al. (2017); Mousavi et al. (2019); Seo et al. (2020); Phan et al. (2021); Perslev et al. (2021) have been proposed for sleep scoring tasks.

Moreover, in the domain of respiratory event classification, automatic detection of SDB using ECG Urtnasan et al. (2020); Tripathy et al. (2020), EEG Zhao et al. (2021), and PSG with its respiratory channels Mostafa et al. (2020); Yu et al. (2022); Haidar et al. (2018); Yeo et al. (2021); Nassi et al. (2021); Stephansen et al. (2018) has been explored extensively. A recent study introduced a multi-task learning approach, training a supervised deep learning model to predict diverse sleep events (e.g., sleep stages, arousal, leg movements, and sleep-disordered breathing) using multiple sleep modalities like EEG, EOG, and EMG Zahid et al. (2023). These studies predominantly utilize supervised learning, often based on relatively small datasets comprising only a few hundred subjects.

A.1.2 Contrastive Learning

A major development in self-supervised learning techniques is the rise of contrastive methods for comprehensive data representation learning. In computer vision, influential frameworks have emerged including: InfoNCE Oord et al. (2018), SimCLR Chen et al. (2020), MoCo He et al. (2020), and SupCon Khosla et al. (2020). These uni-modal contrastive approaches focus primarily on single data modalities like images. A notable multi-modal exception is the Contrastive Language-Image Pretraining (CLIP) model Radford et al. (2021), which aligns image and text embeddings. In medicine, ConVIRT Zhang et al. (2022) pioneered multi-modal CL between chest radiographs and reports. Other works have explored similar directions for medical images Huang et al. (2021); Boecking et al. (2022); Bannur et al. (2023); Lu et al. (2023).

Outside of computer vision, uni-modal contrastive methods have been applied to time series data like ECG signals Kiyasseh et al. (2021); Gopal et al. (2021). CL has also enabled signal conversion tasks Nørskov et al. (2023). However, contrastive representation learning across diverse physiological modalities remains relatively uncharted. Two prior studies have investigated contrastive multi-modal clinical time series analysis. One work employed SimCLR-style pre-training on data encompassing ECG and structured records Raghu et al. (2022). Another derived ECG representations by contrasting ECGs, structured EHRs, and clinical notes Lalam et al. (2023).

A.2 Data Description

In Figure 3, we see a 30 second clip of our raw data for all 19 channels across 3 modalities. Figure 4 shows the distribution of various events across the entire sleep duration for a participant. To ensure the protection of participants' Protected Health Information (PHI), all data has been de-identified.

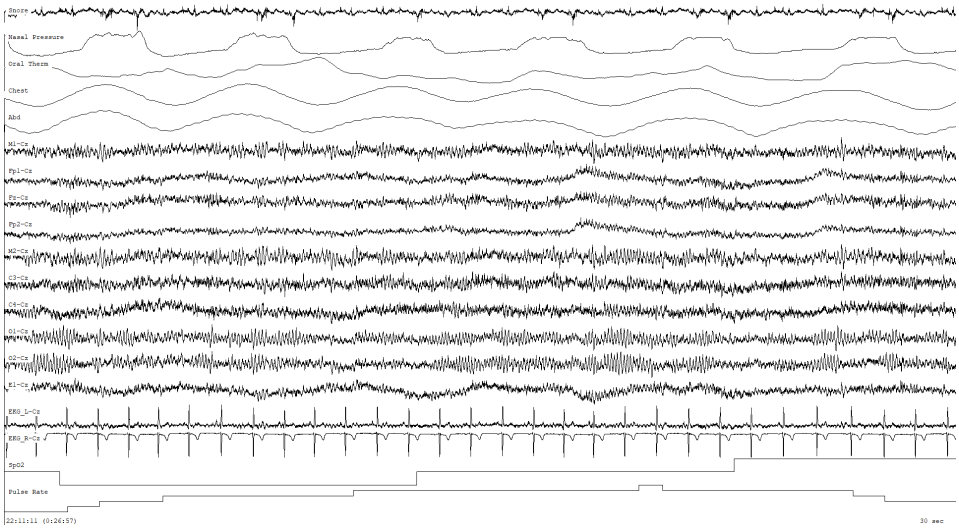


Figure 3: 30-second clip of raw patient data. The x-axis is time and y-axis is different channels across all three modalities: BAS, ECG, and Respiratory.

Table 6: Demographics table. REM: Rapid Eye Movement; AHI: Apnea-Hypopnea Index, a measure used in sleep medicine to assess the severity of sleep apnea; WASO: Wake After Sleep Onset, the total time spent awake after initially falling asleep; SL: Sleep Latency, the time it takes to transition from wakefulness to sleep; REML: REM Sleep Latency, the time it takes to enter REM sleep after falling asleep; TSD: Total Sleep Duration, the overall duration of sleep. $\pm$ represents upper and lower bound.

	pretrain	train	valid	test
Participants (count)	11,261	1,265	141	1,401
Events (count)	10,611,314	1,190,392	130,380	1,314,267
Duration (hours)	88,427	9,920	1,086	10,952
Male (%)	49.9	50.2	47.1	53.0
Female (%)	43.8	44.0	48.1	41.8
Unknown (%)	6.3	5.9	4.8	5.2
Age (years)	42.2 $\pm$ 19.6	43.0 $\pm$ 20.3	40.4 $\pm$ 20.0	41.9 $\pm$ 19.9
TSD (mins)	376.7 $\pm$ 90.8	376.4 $\pm$ 90.6	371.2 $\pm$ 84.9	374.3 $\pm$ 87.5
WASO (mins)	79.4 $\pm$ 60.5	79.7 $\pm$ 62.3	78.8 $\pm$ 57.3	81.5 $\pm$ 62.8
SL (mins)	22.2 $\pm$ 32.8	21.2 $\pm$ 31.6	29.0 $\pm$ 87.8	22.5 $\pm$ 32.6
REML (mins)	151.9 $\pm$ 102.6	149.4 $\pm$ 97.7	148.6 $\pm$ 99.9	154.8 $\pm$ 103.5
Stage 1 (%)	9.4 $\pm$ 9.2	9.3 $\pm$ 8.8	8.2 $\pm$ 7.7	9.0 $\pm$ 8.9
Stage 2 (%)	65.0 $\pm$ 14.7	64.8 $\pm$ 14.7	64.8 $\pm$ 14.7	65.0 $\pm$ 14.7
Stage 3 (%)	10.2 $\pm$ 13.2	10.2 $\pm$ 13.2	10.9 $\pm$ 12.7	10.3 $\pm$ 13.6
REM (%)	15.5 $\pm$ 7.9	15.7 $\pm$ 8.0	16.2 $\pm$ 6.8	15.7 $\pm$ 7.9
AHI (h^{-1})	22.2 $\pm$ 79.3	22.8 $\pm$ 19.1	22.2 $\pm$ 18.5	20.9 $\pm$ 17.0

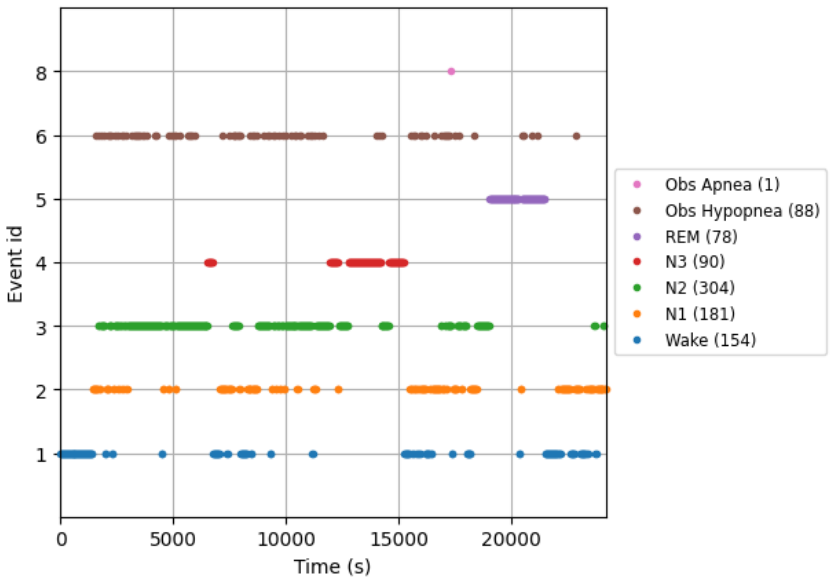


Figure 4: Distribution of events across an entire patient sleep. The x-axis represents approximately 8 hours in seconds, and y-axis is distribution of different sleep events during the entire duration of sleep. N1, N2, N3 refers to Sleep Stage 1, 2, and 3 respectively. Obs Hypopnea and Obs SDB are types of SDBs.

A.3 TRAINING DETAILS

Our model pretraining involves minimizing the contrastive loss with stochastic gradient descent (SGD) using an initial learning rate set to 0.001 and a momentum of 0.9. The learning rate is decayed by a factor of 10 every 5 epochs. The trainable temperature parameter is initialized to 0. Training spans a maximum of 20 epochs with early stopping based on validation loss, employing a batch size of 32 and validating checkpoints at each epoch to ensure robust regularization.

Upon pretraining completion via this self-supervised approach, we generate embeddings for the training, validation, and test sets, utilizing the learned modality encoders. Subsequently, these training embeddings drive the training of a logistic regression classifier. The classifier's performance undergoes evaluation on the test set for both sleep stage and SDB event detection tasks, as outlined in Section 3.2.

In our experiments, we additionally compare against training a supervised CNN without contrastive learning as a baseline. The supervised CNN uses an 1D EfficientNet architecture akin to our pretrained model encoder but is solely trained via supervised learning on the entire (pretraining + training) dataset for classification tasks. This architecture uses a series of 1D convolutions encoding all three modalities into an embedding space, followed by a dropout layer for regularization and a fully-connected layer predicting scores across different classes. This model is trained end-to-end from scratch using cross-entropy loss between the predicted and true labels, optimized by SGD. Mirroring the pretraining phase, this model undergoes training for 20 epochs with a batch size of 32, aligning hyperparameters with our model pretraining strategy.

All model training was executed on a single NVIDIA Tesla V100S GPU with 32GB of memory. Each pretraining epoch consumed approximately 4 hours, while baseline supervised training required roughly 2 hours on the same GPU. Table Table 7 and Table 8 lists the hyperparameters we used in our training runs.

Table 7: Hyperparameters for Pretraining and end-to-end CNN training

Hyperparameter	**Value**
Learning Rate	0.01
Batch Size	32
lr step period	5
epochs	20
momentum	0.9
Temperature (init)	0.0

Table 8: Hyperparameters for logistic regression training during downstream classifications.

Hyperparameter	**Value**
penalty	L2
max iter	10000
class weight	balanced
solver	lbfgs

A.4 ADDITIONAL RESULTS

A.5 DEMOGRAPHIC ATTRIBUTES CLASSIFICATION

We evaluated our *SleepFM*'s embedding quality by training a logistic regression classifier on top of the combined multimodal embeddings to predict common demographic attributes such as age and gender. Our classification task directly used the 30-second clip-level embeddings generated by *SleepFM*. For age prediction, we grouped ages into the following categories: 0-18, 18-35, 35-50, and 50+. The prevalence of these age groups in our dataset is 0.17, 0.18, 0.28, and 0.37, respectively. For gender classification, we considered male vs. female, with the prevalence of females being 0.41 in

Table 9: Sleep stage classification metrics for models trained using different types of contrastive learning (CL). Baseline here is an end-to-end CNN trained on the entire (pretraining + training) dataset to classify sleep stages. The leave-one-out (LOO) and pairwise models are logistic regression models trained on the embeddings generated from only the training dataset. Therefore end-to-end CNN saw 11,261 patient data while pretrained model saw 1,265 training data for sleep stage classification. Prevalence of Wake, Stage 1, Stage 2, Stage 3, and REM are 0.21, 0.07, 0.51, 0.09, and 0.12 respectively. $\pm$ represents 95% Confidence Intervals.

	AUROC			**AUPRC**		
	LOO	**Pairwise**	**Baseline**	**LOO**	**Pairwise**	**Baseline**
Wake	$0.945_{\pm.001}$	$0.930_{\pm.001}$	$0.869_{\pm.001}$	$0.862_{\pm.002}$	$0.827_{\pm.002}$	$0.711_{\pm.002}$
Stage 1	$0.814_{\pm.002}$	$0.782_{\pm.002}$	$0.706_{\pm.002}$	$0.233_{\pm.003}$	$0.186_{\pm.002}$	$0.130_{\pm.002}$
Stage 2	$0.891_{\pm.001}$	$0.861_{\pm.001}$	$0.840_{\pm.001}$	$0.876_{\pm.001}$	$0.849_{\pm.001}$	$0.822_{\pm.001}$
Stage 3	$0.928_{\pm.001}$	$0.918_{\pm.001}$	$0.918_{\pm.001}$	$0.676_{\pm.003}$	$0.615_{\pm.003}$	$0.695_{\pm.002}$
REM	$0.951_{\pm.001}$	$0.891_{\pm.001}$	$0.878_{\pm.001}$	$0.778_{\pm.003}$	$0.565_{\pm.002}$	$0.540_{\pm.003}$
Avg	**0.906**	0.876	0.842	**0.685**	0.608	0.579

Table 10: Age classification metrics for models trained using different types of contrastive learning (CL). The supervised CNN is trained on the entire (pretraining + training) dataset to classify age groups. The leave-one-out (LOO) and pairwise models are logistic regression models trained on the embeddings generated from only the training dataset. Therefore end-to-end CNN saw all data 11,261 participants while pretrained model saw data from 1,265 participants for sleep stage classification. Prevalence of 0-18, 18-35, 35-50, and 50+ are 0.17, 0.18, 0.28, and 0.37 respectively. $\pm$ represents 95% Confidence Intervals.

	AUROC			**AUPRC**		
	LOO	**Pairwise**	**Baseline**	**LOO**	**Pairwise**	**Baseline**
0-18	$0.982_{\pm.001}$	$0.977_{\pm.001}$	$0.864_{\pm.001}$	$0.937_{\pm.002}$	$0.929_{\pm.004}$	$0.628_{\pm.003}$
18-35	$0.852_{\pm.001}$	$0.809_{\pm.002}$	$0.683_{\pm.002}$	$0.549_{\pm.003}$	$0.458_{\pm.002}$	$0.308_{\pm.002}$
35-50	$0.784_{\pm.001}$	$0.740_{\pm.001}$	$0.606_{\pm.003}$	$0.524_{\pm.001}$	$0.476_{\pm.002}$	$0.371_{\pm.002}$
50+	$0.915_{\pm.001}$	$0.880_{\pm.001}$	$0.745_{\pm.002}$	$0.856_{\pm.002}$	$0.796_{\pm.002}$	$0.619_{\pm.002}$
Avg	**0.883**	0.851	0.724	**0.716**	0.664	0.481

our dataset. We evaluated the performance based on AUROC (Area Under the Receiver Operating Characteristic curve) and AUPRC (Area Under the Precision-Recall Curve). As a baseline, we trained a CNN end-to-end to perform age and gender classification given the combined multimodal raw input data.

We find that *SleepFM* can predict age and gender with high accuracy from just 30-second clips of physiological data (Table 10 and Table 11). Both our pre-trained models significantly outperform the end-to-end CNN baseline across all evaluation metrics and tasks. Note that the end-to-end supervised CNN used the full (pretraining + training) dataset during training, while the embeddings from *SleepFM* were only trained on the training set. Notably, the model pre-trained with leave-one-out CL achieves the best performance. The strong clip-level performance indicates *SleepFM*'s embeddings effectively capture salient demographic information. Analyzing the performance per modality, we find that the BAS signals contain the most distinctive features for these tasks as shown in Table 16 and Table 17.

Table 11: Gender classification metrics for models trained using different types of CL. The supervised CNN is trained on the entire (pretraining + training) dataset to classify gender. The leave-one-out and pairwise models are logistic regression models trained on the embeddings generated from only the training dataset. Therefore end-to-end CNN saw 11,261 patient data while pretrained model saw 1,265 training data for SDB classification. Prevalence of female gender is 0.41. ± represents 95% Confidence Intervals.

	AUROC	**AUPRC**
Leave-One-Out CL	$\mathbf{0.850_{\pm.001}}$	$\mathbf{0.774_{\pm.002}}$
Pairwise CL	$0.810_{\pm.001}$	$0.731_{\pm.002}$
Supervised CNN	$0.690_{\pm.002}$	$0.614_{\pm.002}$

Table 12: Sleep stage classification metrics for model trained with leave-one-out CL. After having trained the model with all three modalities, we extract embeddings for each modality separately and train a logistic regression with each modality to identify sleep stages. ± represents 95% confidence intervals.

	AUROC			**AUPRC**		
	ECG	**Respiratory**	**BAS**	**ECG**	**Respiratory**	**BAS**
Wake	$0.934_{\pm.001}$	$0.846_{\pm.001}$	$0.942_{\pm.001}$	$0.829_{\pm.004}$	$0.652_{\pm.003}$	$0.857_{\pm.002}$
Stage 1	$0.786_{\pm.002}$	$0.676_{\pm.002}$	$0.801_{\pm.002}$	$0.193_{\pm.002}$	$0.127_{\pm.001}$	$0.211_{\pm.003}$
Stage 2	$0.874_{\pm.001}$	$0.728_{\pm.001}$	$0.888_{\pm.001}$	$0.860_{\pm.001}$	$0.708_{\pm.001}$	$0.873_{\pm.001}$
Stage 3	$0.919_{\pm.001}$	$0.788_{\pm.001}$	$0.927_{\pm.001}$	$0.638_{\pm.003}$	$0.307_{\pm.002}$	$0.679_{\pm.002}$
REM	$0.939_{\pm.001}$	$0.789_{\pm.001}$	$0.944_{\pm.001}$	$0.745_{\pm.003}$	$0.388_{\pm.003}$	$0.724_{\pm.003}$
Macro Avg	0.891	0.765	0.900	0.436	0.484	0.669

Table 13: SDB classification metrics for model trained with leave-one-out CL. After having trained the model with all three modalities, we extract embeddings for each modality separately and train a logistic regression with each modality to identify SDB. ± represents 95% confidence intervals.

	ECG	**Respiratory**	**BAS**
AUROC	$0.735_{\pm.004}$	$0.925_{\pm.002}$	$0.735_{\pm.004}$
AUPRC	$0.040_{\pm.001}$	$0.697_{\pm.006}$	$0.040_{\pm.001}$

Table 14: Sleep stage classification metrics for model trained with pairwise CL. After having trained the model with all three modalities, we extract embeddings for each modality separately and train a logistic regression with each modality to identify sleep stages. ± represents 95% confidence intervals.

	AUROC			**AUPRC**		
	ECG	**Respiratory**	**BAS**	**ECG**	**Respiratory**	**BAS**
Wake	$0.917_{\pm.001}$	$0.821_{\pm.001}$	$0.925_{\pm.001}$	$0.782_{\pm.002}$	$0.621_{\pm.002}$	$0.816_{\pm.001}$
Stage 1	$0.766_{\pm.002}$	$0.661_{\pm.002}$	$0.772_{\pm.002}$	$0.167_{\pm.002}$	$0.116_{\pm.001}$	$0.174_{\pm.002}$
Stage 2	$0.848_{\pm.001}$	$0.695_{\pm.001}$	$0.857_{\pm.001}$	$0.841_{\pm.001}$	$0.675_{\pm.001}$	$0.845_{\pm.001}$
Stage 3	$0.911_{\pm.001}$	$0.777_{\pm.001}$	$0.917_{\pm.001}$	$0.601_{\pm.002}$	$0.296_{\pm.003}$	$0.614_{\pm.003}$
REM	$0.872_{\pm.001}$	$0.649_{\pm.001}$	$0.880_{\pm.001}$	$0.526_{\pm.003}$	$0.200_{\pm.003}$	$0.522_{\pm.002}$
Macro Avg	0.862	0.720	0.870	0.583	0.381	0.594

Table 15: SDB classification metrics for model trained with pairwise CL. After having trained the model with all three modalities, we extract embeddings for each modality separately and train a logistic regression with each modality to identify SDB. ± represents 95% confidence intervals.

	ECG	**Respiratory**	**BAS**
AUROC	$0.698_{\pm.003}$	$0.893_{\pm.003}$	$0.706_{\pm.004}$
AUPRC	$0.029_{\pm.001}$	$0.601_{\pm.006}$	$0.030_{\pm.001}$

Table 16: Age classification metrics for model trained with leave-one-out CL. After having trained the model with all three modalities, we extract embeddings for each modality separately and train a logistic regression with each modality to identify age groups. ± represents 95% confidence intervals.

	AUROC			**AUPRC**		
	ECG	**Respiratory**	**BAS**	**ECG**	**Respiratory**	**BAS**
0-18	$0.977_{\pm.001}$	$0.965_{\pm.001}$	$0.969_{\pm.001}$	$0.921_{\pm.001}$	$0.883_{\pm.003}$	$0.911_{\pm.001}$
18-35	$0.833_{\pm.001}$	$0.789_{\pm.001}$	$0.755_{\pm.002}$	$0.493_{\pm.003}$	$0.455_{\pm.003}$	$0.380_{\pm.003}$
35-50	$0.774_{\pm.001}$	$0.722_{\pm.001}$	$0.686_{\pm.001}$	$0.516_{\pm.002}$	$0.458_{\pm.003}$	$0.424_{\pm.002}$
50+	$0.905_{\pm.001}$	$0.873_{\pm.001}$	$0.813_{\pm.001}$	$0.843_{\pm.001}$	$0.780_{\pm.001}$	$0.685_{\pm.002}$
Macro Avg	0.872	0.837	0.805	0.693	0.644	0.600

Table 17: Gender classification metrics for model trained with leave-one-out CL. After having trained the model with all three modalities, we extract embeddings for each modality separately and train a logistic regression with each modality to identify gender. ± represents 95% confidence intervals.

	ECG	**Respiratory**	**BAS**
AUROC	$0.829_{\pm.001}$	$0.790_{\pm.002}$	$0.778_{\pm.001}$
AUPRC	$0.754_{\pm.001}$	$0.710_{\pm.003}$	$0.713_{\pm.002}$

Table 18: Age classification metrics for model trained with pairwise CL. After having trained the model with all three modalities, we extract embeddings for each modality separately and train a logistic regression with each modality to identify age groups. ± represents 95% confidence intervals.

	AUROC			**AUPRC**		
	ECG	**Respiratory**	**BAS**	**ECG**	**Respiratory**	**BAS**
0-18	$0.969_{\pm.001}$	$0.962_{\pm.001}$	$0.963_{\pm.001}$	$0.908_{\pm.001}$	$0.883_{\pm.001}$	$0.897_{\pm.001}$
18-35	$0.786_{\pm.001}$	$0.769_{\pm.001}$	$0.767_{\pm.001}$	$0.422_{\pm.002}$	$0.455_{\pm.003}$	$0.389_{\pm.002}$
35-50	$0.712_{\pm.002}$	$0.702_{\pm.001}$	$0.706_{\pm.002}$	$0.441_{\pm.002}$	$0.458_{\pm.003}$	$0.436_{\pm.002}$
50+	$0.865_{\pm.001}$	$0.841_{\pm.001}$	$0.840_{\pm.001}$	$0.722_{\pm.002}$	$0.780_{\pm.001}$	$0.742_{\pm.001}$
Macro Avg	0.832	0.818	0.818	0.634	0.617	0.615

Table 19: Gender classification metrics for model trained with pairwise CL. After having trained the model with all three modalities, we extract embeddings for each modality separately and train a logistic regression with each modality to identify gender. ± represents 95% confidence intervals.

	ECG	**Respiratory**	**BAS**
AUROC	$0.795_{\pm.001}$	$0.746_{\pm.001}$	$0.765_{\pm.001}$
AUPRC	$0.722_{\pm.001}$	$0.676_{\pm.002}$	$0.702_{\pm.002}$

Table 20: Sleep Stage Classification stratified by age group.

	Macro AUROC		Macro AUPRC	
	Leave-One-Out	**Pairwise**	**Leave-One-Out**	**Pairwise**
0-18	0.890	0.849	0.665	0.594
18-35	0.911	0.883	0.702	0.624
35-50	0.897	0.867	0.630	0.559
50+	0.895	0.861	0.616	0.530

Table 21: Sleep Stage Classification stratified by gender.

	Macro AUROC		Macro AUPRC	
	Leave-One-Out	**Pairwise**	**Leave-One-Out**	**Pairwise**
Male	0.899	0.869	0.674	0.594
Female	0.910	0.880	0.693	0.621

Table 22: SDB classification metrics stratified by age group.

	AUROC		AUPRC	
	Leave-One-Out	**Pairwise**	**Leave-One-Out**	**Pairwise**
0-18	$0.93_{\pm 0.01}$	$0.86_{\pm 0.03}$	$0.56_{\pm 0.04}$	$0.35_{\pm 0.04}$
18-35	$0.94_{\pm 0.01}$	$0.90_{\pm 0.01}$	$0.69_{\pm 0.02}$	$0.61_{\pm 0.03}$
35-50	$0.94_{\pm 0.01}$	$0.89_{\pm 0.01}$	$0.73_{\pm 0.01}$	$0.63_{\pm 0.02}$
50+	$0.94_{\pm 0.01}$	$0.90_{\pm 0.01}$	$0.73_{\pm 0.01}$	$0.60_{\pm 0.01}$

Table 23: SDB classification metrics stratified by gender.

	AUROC		AUPRC	
	Leave-One-Out	**Pairwise**	**Leave-One-Out**	**Pairwise**
Male	$0.94_{\pm 0.01}$	$0.90_{\pm 0.01}$	$0.73_{\pm 0.01}$	$0.61_{\pm 0.01}$
Female	$0.95_{\pm 0.01}$	$0.91_{\pm 0.01}$	$0.70_{\pm 0.01}$	$0.59_{\pm 0.01}$

Table 24: AUROC metrics for sleep stage classification with leave-one-out CL, stratified by different age groups.

	0-18	**18-35**	**35-50**	**50+**
Wake	$0.937_{\pm 0.002}$	$0.939_{\pm 0.001}$	$0.938_{\pm 0.001}$	$0.944_{\pm 0.001}$
Stage 1	$0.805_{\pm 0.006}$	$0.831_{\pm 0.003}$	$0.808_{\pm 0.003}$	$0.793_{\pm 0.002}$
Stage 2	$0.861_{\pm 0.002}$	$0.900_{\pm 0.001}$	$0.888_{\pm 0.002}$	$0.889_{\pm 0.001}$
Stage 3	$0.906_{\pm 0.001}$	$0.932_{\pm 0.002}$	$0.902_{\pm 0.002}$	$0.902_{\pm 0.002}$
REM	$0.941_{\pm 0.002}$	$0.956_{\pm 0.001}$	$0.950_{\pm 0.001}$	$0.949_{\pm 0.001}$
Avg	0.890	0.911	0.897	0.895

Table 25: AUPRC metrics for sleep stage classification with leave-one-out CL, stratified by different age groups.

	0-18	**18-35**	**35-50**	**50+**
Wake	$0.809_{\pm 0.005}$	$0.859_{\pm 0.004}$	$0.843_{\pm 0.003}$	$0.872_{\pm 0.002}$
Stage 1	$0.163_{\pm 0.008}$	$0.29_{\pm 0.006}$	$0.236_{\pm 0.005}$	$0.235_{\pm 0.004}$
Stage 2	$0.812_{\pm 0.003}$	$0.890_{\pm 0.002}$	$0.879_{\pm 0.001}$	$0.863_{\pm 0.002}$
Stage 3	$0.818_{\pm 0.004}$	$0.696_{\pm 0.004}$	$0.406_{\pm 0.007}$	$0.325_{\pm 0.005}$
REM	$0.725_{\pm 0.007}$	$0.775_{\pm 0.006}$	$0.787_{\pm 0.004}$	$0.786_{\pm 0.004}$
Avg	0.665	0.702	0.630	0.616

Table 26: Sleep stage classification metrics for model trained with leave-one-out CL. The performance is stratified by different gender groups.

	AUROC		**AUPRC**	
	Male	**Female**	**Male**	**Female**
Wake	$0.937_{\pm 0.001}$	$0.949_{\pm 0.001}$	$0.844_{\pm 0.002}$	$0.872_{\pm 0.002}$
Stage 1	$0.805_{\pm 0.002}$	$0.824_{\pm 0.002}$	$0.251_{\pm 0.004}$	$0.225_{\pm 0.004}$
Stage 2	$0.887_{\pm 0.001}$	$0.890_{\pm 0.001}$	$0.867_{\pm 0.001}$	$0.870_{\pm 0.001}$
Stage 3	$0.919_{\pm 0.001}$	$0.934_{\pm 0.001}$	$0.635_{\pm 0.005}$	$0.729_{\pm 0.004}$
REM	$0.944_{\pm 0.001}$	$0.955_{\pm 0.001}$	$0.771_{\pm 0.004}$	$0.767_{\pm 0.002}$
Avg	0.899	0.910	0.674	0.693

Table 27: AUROC metrics for sleep stage classification with pairwise CL, stratified by different age groups.

	0-18	**18-35**	**35-50**	**50+**
Wake	$0.919_{\pm 0.002}$	$0.928_{\pm 0.002}$	$0.926_{\pm 0.001}$	$0.926_{\pm 0.001}$
Stage 1	$0.712_{\pm 0.009}$	$0.804_{\pm 0.004}$	$0.775_{\pm 0.003}$	$0.758_{\pm 0.003}$
Stage 2	$0.827_{\pm 0.002}$	$0.870_{\pm 0.002}$	$0.863_{\pm 0.002}$	$0.861_{\pm 0.002}$
Stage 3	$0.891_{\pm 0.002}$	$0.911_{\pm 0.002}$	$0.881_{\pm 0.003}$	$0.891_{\pm 0.002}$
REM	$0.894_{\pm 0.002}$	$0.901_{\pm 0.002}$	$0.891_{\pm 0.002}$	$0.868_{\pm 0.002}$
Avg	0.849	0.883	0.867	0.861

Table 28: AUPRC metrics for sleep stage classification with pairwise CL, stratified by different age groups.

	0-18	**18-35**	**35-50**	**50+**
Wake	$0.771_{\pm 0.005}$	$0.828_{\pm 0.003}$	$0.813_{\pm 0.003}$	$0.838_{\pm 0.003}$
Stage 1	$0.103_{\pm 0.006}$	$0.218_{\pm 0.007}$	$0.191_{\pm 0.004}$	$0.198_{\pm 0.004}$
Stage 2	$0.780_{\pm 0.003}$	$0.861_{\pm 0.003}$	$0.857_{\pm 0.002}$	$0.833_{\pm 0.002}$
Stage 3	$0.775_{\pm 0.004}$	$0.617_{\pm 0.003}$	$0.340_{\pm 0.009}$	$0.267_{\pm 0.007}$
REM	$0.539_{\pm 0.009}$	$0.597_{\pm 0.006}$	$0.591_{\pm 0.006}$	$0.516_{\pm 0.005}$
Avg	0.594	0.624	0.559	0.530

Table 29: Sleep stage classification metrics for model trained with pairwise CL. The performance is stratified by different gender groups.

	AUROC		**AUPRC**	
	Male	**Female**	**Male**	**Female**
Wake	$0.924_{\pm 0.001}$	$0.932_{\pm 0.001}$	$0.813_{\pm 0.002}$	$0.834_{\pm 0.002}$
Stage 1	$0.769_{\pm 0.002}$	$0.791_{\pm 0.002}$	$0.194_{\pm 0.003}$	$0.192_{\pm 0.004}$
Stage 2	$0.859_{\pm 0.001}$	$0.861_{\pm 0.001}$	$0.840_{\pm 0.001}$	$0.840_{\pm 0.002}$
Stage 3	$0.910_{\pm 0.001}$	$0.922_{\pm 0.001}$	$0.559_{\pm 0.002}$	$0.687_{\pm 0.004}$
REM	$0.882_{\pm 0.001}$	$0.892_{\pm 0.001}$	$0.561_{\pm 0.002}$	$0.554_{\pm 0.005}$
Avg	0.869	0.880	0.594	0.621