wav2sleep: A Unified Multi-Modal Approach to Sleep Stage Classification from Physiological Signals

Jonathan F. Carter   JCARTER@ROBOTS.OX.AC.UK
Lionel Tarassenko   LIONEL.TARASSENKO@ENG.OX.AC.UK
Institute of Biomedical Engineering, University of Oxford

Abstract

Accurate classification of sleep stages from less obtrusive sensor measurements such as the electrocardiogram (ECG) or photoplethysmogram (PPG) could enable important applications in sleep medicine. Existing approaches to this problem have typically used deep learning models designed and trained to operate on one or more specific input signals. However, the datasets used to develop these models often do not contain the same sets of input signals. Some signals, particularly PPG, are much less prevalent than others, and this has previously been addressed with techniques such as transfer learning. Additionally, only training on one or more fixed modalities precludes cross-modal information transfer from other sources, which has proved valuable in other problem domains. To address this, we introduce wav2sleep, a unified model designed to operate on variable sets of input signals during training and inference. After jointly training on over 10,000 overnight recordings from six publicly available polysomnography datasets, including SHHS and MESA, wav2sleep outperforms existing sleep stage classification models across test-time input combinations including ECG, PPG, and respiratory signals.

Keywords: Sleep, time-series, deep learning

Data and Code Availability All datasets used in this work are available via the National Sleep Research Resource (NSRR, (Zhang et al., 2018)). Training code, model weights, and processing pipelines for all datasets can be found here: https://github.com/joncarter1/wav2sleep.

Institutional Review Board (IRB) The datasets used herein are publicly available via request to the NSRR and have been deidentified or fully anonymised, thus not requiring IRB approval.

1. Introduction

Quantitative analysis of sleep stages is important for applications such as the diagnosis of sleep disorders, the validation of sleep disorder medications, and the discovery of sleep biomarkers. This is typically done through a polysomnography (PSG) study following the American Academy of Sleep Medicine (AASM) guidelines (Iber, 2007), during which a subject will undertake one or more nights of sleep whilst being continuously monitored using a variety of sensors.

After the study, a human expert will review the recording and assign each 30-second window of data (a sleep epoch) to one of 5 discrete sleep states defined by the AASM: Wake, N1 (light), N2 (intermediate), N3 (deep), or rapid eye movement (REM) sleep. This is traditionally done using signals such as the electroencephalogram (EEG), electrooculogram (EOG), and chin electromyogram (EMG), which are measured using electrodes attached to the subject's head.

The goal of sleep stage classification (sleep staging) algorithms is to automate this process using one or more input signals. Accurate sleep staging from inputs such as the photoplethysmogram (PPG), electrocardiogram (ECG), abdominal (ABD) and/or thoracic (THX) respiratory signals (see Figure 1) is of particular interest since acquiring these signals is typically more comfortable for the subject and easier to set up than a full PSG study.

Sleep staging models have commonly been trained and evaluated using one or more fixed input modalities. However, the recorded signals often have high mutual information, meaning that information available from one input modality may be useful for learning to classify sleep stages from another.

To better leverage this information, we introduce wav2sleep, a unified model for sleep stage classification that operates on sets of physiological signals. During training, the model can use all available input signals from each night of data, enabling it to be
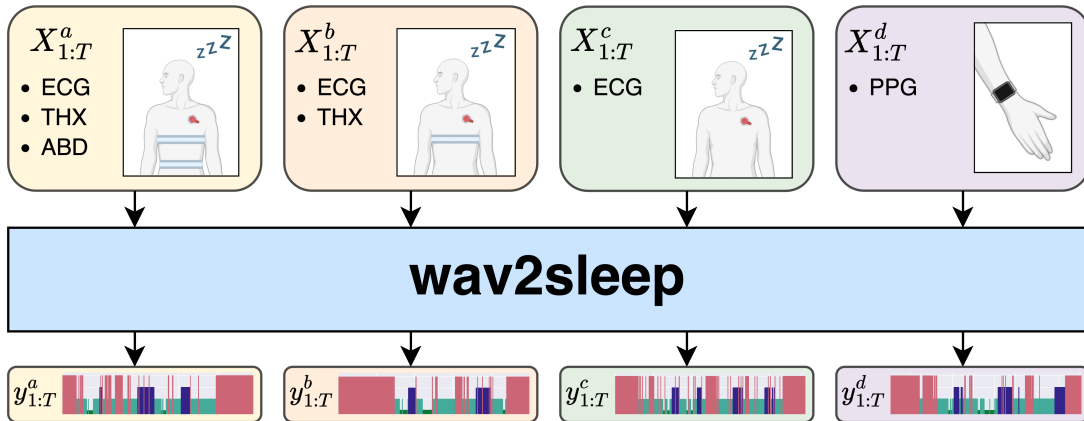
Figure 1: **Overview of wav2sleep**. The model operates on *sets* of time-series signals $\boldsymbol{X}_{1:T}$ to classify sleep stage sequences $y_{1:T}$. This enables it to be jointly trained on heterogeneous datasets, with different available signals, which are especially common in the healthcare domain. At inference time, the same model can be applied to any subset of the signals seen during training.

jointly trained across heterogeneous datasets, where the available signals may differ between recordings. This trait is particularly valuable for our problem domain of sleep staging since the set of recorded signals often varies, and many are often discarded due to poor signal quality.

After jointly training on multiple modalities, our model can use any subset of the original input signals during inference. We show that this unified approach outperforms existing methods across several test-time input modalities and datasets.

To summarise, our contributions are as follows:

- We introduce wav2sleep, a novel deep learning architecture for multi-modal time-series that operates on sets of signals. This enables joint training on heterogeneous datasets, where the availability of signals can vary.

- We introduce a simple but effective stochastic masking procedure during training, which enables generalisation to an arbitrary subset of signals at test time.

- After training, we show that our single, unified model outperforms prior methods for sleep staging across a range of modalities. For example, outperforming existing ECG-based methods when using only the ECG at test time.

2. Background and Motivation

Prior work has shown that deep learning models can classify stages of sleep from brain signals recorded using the EEG with expert-level accuracy (Phan and Mikkelsen, 2022). However, EEG signals are typically measured using electrodes attached to the patient's scalp. To overcome this, other work has investigated classifying sleep stages using alternative, less obtrusive modalities such as the ECG (Sridhar et al., 2020), PPG (Kotzen et al., 2023)), and combinations of cardiac and respiratory signals (Bakker et al., 2021).

Two-step learning Automatic sleep staging from modalities such as the ECG or PPG is possible because these signals encode measures of physiological activity, such as heart rate variability (HRV), and these are known to be predictive of the sleep stage (Shinar et al., 2001). Sleep staging from these signals can therefore be formulated as a two-step learning problem. First, we must find a mapping $f : \boldsymbol{x} \mapsto \boldsymbol{z}$ from the input signals $\boldsymbol{x}$ to relevant physiological features $\boldsymbol{z}$, then a mapping $g : \boldsymbol{z} \mapsto y$ from physiological features $\boldsymbol{z}$ to sleep stages y. For models trained end-to-end to classify sleep stages from input signals, these mappings f and g are jointly and implicitly learnt by a deep neural network. This has been shown to outperform methods where the physiological features $\boldsymbol{z}$ are derived using a manual feature engineering approach (Kotzen et al., 2023), i.e. where the mapping f is explicit and human-designed.

Transfer learning Sleep staging from modalities such as the PPG is of particular interest since they can be measured from ubiquitous wearables such as smartwatches (Charlton et al., 2023). However, PPG signals are less common in historical PSG datasets (Radha et al., 2021), hindering the use of deep learning methods which are notoriously data-hungry. To overcome this issue, transfer learning has been used to improve the performance of PPG-based models, by first pre-training using ECG data (Radha et al., 2021; Kotzen et al., 2023).

Transfer learning approaches from ECG to PPG signals work because these signals have high mutual information and morphological similarity. For example, both signals encode HRV and respiratory rate variability (RRV) information. This fact is explicitly used by Radha et al. (2021), who first pre-train a model using HRV features derived from the ECG, then fine-tune using HRV features derived from PPG data. Using the two-step learning formulation introduced at the start of this section, this approach can be thought of as explicitly mapping each input signal x_i to a modality-agnostic feature space z via a modality-dependent mapping f_i.

Modality-agnostic learning Ideally, modality-agnostic relationships such as the link between HRV and deep sleep (Shinar et al., 2001), or between RRV and REM sleep (Kantelhardt et al., 2003), should not need to be learnt from a specific modality, a fact which is already partially exploited by transfer learning approaches. However, transfer learning is prone to the problem of catastrophic forgetting (Kemker et al., 2018) i.e. where after training on new data, information learnt from the old dataset is 'forgotten'. Hence, even if our end goal is to produce a specialised model for a particular modality, we will show that it is beneficial to use a model that is jointly trained across multiple modalities, rather than directly training or using a transfer learning approach.

Multi-modal learning Multi-modal sleep staging methods have varied in their approaches to combining cross-modal information. For example, the model proposed by Chambon et al. (2018) independently turns each input signal into features before passing them to a classifier (late fusion). In contrast, models such as SeqSleepNet (Phan et al., 2019) concatenate modalities at the input level (early fusion). In the middle ground, separately extracting features from each modality for each 30-second sleep epoch, then fusing and jointly modelling sequential informa-

tion, is also a common approach, e.g. via concatenation (Bakker et al., 2021; Carter et al., 2024) or cross-modal attention (Wang et al., 2024; Pradeepkumar et al., 2024).

For EEG and EOG signals, Kontras et al. (2024) handle heterogeneity by simultaneously training both uni-modal and multi-modal models using an alignment loss, with the uni-modal model used when only one signal is available. This is shown to improve uni-modal performance at test time. Alternatively, MaskSleepNet (Zhu et al., 2023), designed for EEG, EOG and/or EMG signals, handles heterogeneity by zero-padding the EOG and/or EMG input signals, which is done randomly during training.

Objective Our aim is to learn a modality-agnostic intermediate representation of the input signals because we hypothesise this can enable more robust representation learning, improving sleep staging performance. However, we want to maintain the ability to perform end-to-end deep learning from the raw input signals, since this has led to improved performance not just in sleep staging, but across numerous problem domains (Goodfellow et al., 2016). Finally, we desire a model that can be jointly trained across heterogeneous datasets, improving the diversity of input data, and avoiding the problem of catastrophic forgetting. In Sections 3 and 4, we describe the wav2sleep model which addresses these challenges, and discuss its advantages over prior methods. In Section 5, we empirically validate its effectiveness in sleep staging across a range of test-time modalities.

3. Model Architecture

In this section, we describe the architecture of the wav2sleep model, which turns sets of time-series signals spanning multiple hours into sleep stage classifications for each 30-second sleep epoch. The model architecture, illustrated in Figure 2, consists of three high-level components:

1. **Signal Encoders**, which independently extract features for each signal in the input set $X_{1:T}$.

2. **Epoch Mixer**, which fuses cross-modal information into a unified representation z_t for each sleep epoch.

3. **Sequence Mixer**, which mixes information temporally to classify sleep stages $y_{1:T}$.

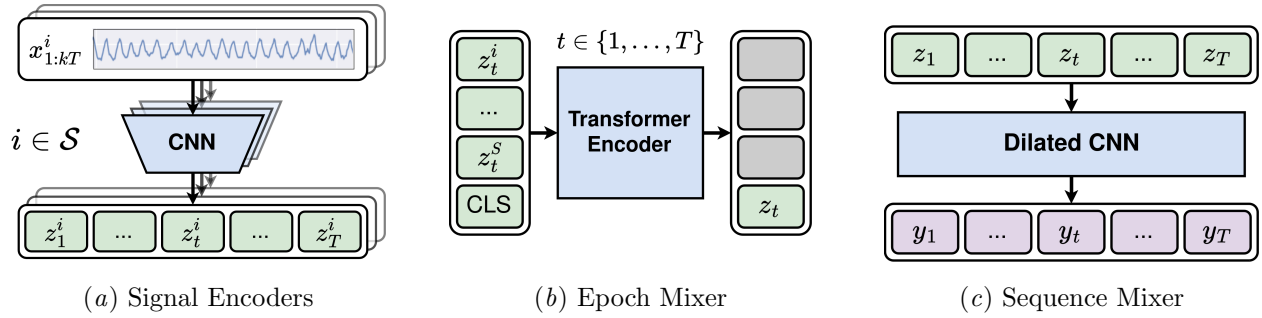(a) Signal Encoders   (b) Epoch Mixer   (c) Sequence Mixer

Figure 2: **wav2sleep architecture for sets of signals.** (a) Each input signal $x^i_{1:kT}$ from modality $i \in \mathcal{S}$ is passed to a CNN to form a sequence of feature vectors $\boldsymbol{z}^i_{1:T}$. (b) For each time-step t, a transformer encoder turns the set of features into a single aggregate feature vector $\boldsymbol{z}_t$ using a CLS token (Devlin et al., 2019). (c) A dilated CNN mixes sequential information to classify sleep stage output sequences $y_{1:T}$

3.1. Signal Encoders

The model first turns the set of continuous 1D input signals $\boldsymbol{X}_{1:T} = \{x^i_{1:kT} \,|\, i \in \mathcal{S}\}$ into a set of feature vector sequences $\boldsymbol{Z}_{1:T} = \{\boldsymbol{z}^i_{1:T} \,|\, i \in \mathcal{S}\}$, where $\boldsymbol{z}^i_t$ denotes the feature vector for modality i for sleep epoch t, k denotes the relative sampling rate of each signal, and $\mathcal{S}$ denotes the set of available modalities e.g. ECG and PPG signals. We use separate CNN encoders for each input modality, which follow the design of the early layers of SleepPPG-Net (Kotzen et al., 2023). These consist of a stack of residual layers (He et al., 2016), each containing three convolutional layers followed by a max pooling layer to downsample the signal by a factor of 2. The residual layers are followed by a reshape operation and a time-distributed dense layer to produce the sequence of feature vectors $\boldsymbol{z}^i_{1:T}$.

3.2. Epoch Mixer

Having independently transformed each modality i into a sequence of feature vectors $\boldsymbol{z}^i_{1:T}$, we next fuse information from the set of modalities to provide a single unified representation $\boldsymbol{z}_t$ for each sleep epoch i.e. to complete the mapping f described in Section 2. We use a transformer encoder (Vaswani et al., 2017) to do this, providing the transformer with an extra learnable vector, i.e. a CLS token (Devlin et al., 2019; Dosovitskiy et al., 2020), and using the output at that position as our unified feature vector. This design straightforwardly handles a varying number of input modalities during training and inference whilst keeping the dimensionality of the fused feature sequence $\boldsymbol{z}_{1:T}$ fixed.

3.3. Sequence Mixer

The feature vectors $\boldsymbol{z}_{1:T}$ are passed to the sequence mixer, which mixes sequential information to produce sleep stage outputs $y_{1:T}$. This is a desirable property since sleep exhibits long-range time-series structures such as sleep cycles (Patel et al., 2022). We use a dilated CNN design as previously used by Sridhar et al. (2020); Kotzen et al. (2023). This consists of multiple blocks of dilated convolutional layers where the dilation doubles at each layer, meaning that the size of the model's receptive field increases exponentially with network depth.

3.4. Advantages

Returning to the two-step learning formulation (f and g) introduced in Section 2, the wav2sleep architecture has two key advantages:

1. Because it operates on *sets* of input signals, the model can be trained on heterogeneous datasets, increasing the variety of data available in terms of both the input modalities (for learning f) *and* physiology (for learning g).

2. By training on all available modalities *jointly*, this should lead to more robust learning in the presence of noise, by avoiding shortcut learning when one or more modalities are corrupted.

Using a unified model also has practical advantages, since only a single model needs to be trained, validated and deployed, reducing operational complexity for real-world applications.

4. Experimental Set-up

4.1. Datasets and Preprocessing

We use 7 PSG datasets available from the National Sleep Research Resource (Zhang et al., 2018): SHHS (Quan et al., 1997), MESA (Chen et al., 2015), CFS (Redline et al., 1995), MROS (Blackwell et al., 2011), CHAT (Marcus et al., 2013), CCSHS (Rosen et al., 2003), and WSC (Young et al., 2009). Demographic information for the datasets used is provided in Table 1. Collectively, these datasets contain over 15,000 pairs of overnight polysomnography recordings and expert-annotated sleep stages. Notably, there is significant variation in patient demographics. For example, the SHHS, MESA and WSC datasets are mostly comprised of recordings from older adults with high apnea-hypopnea indices (sleep-disordered breathing). In contrast, the CCSHS and CHAT datasets both contain PSG recordings from children. Joint training across all datasets exposes the model to a wider variety of contact sensors (makes, models etc.) and individual physiological variations.

There is also variation in the availability of signals between the datasets. For example, recordings from MESA and CCSHS, and some from the CFS and CHAT datasets, contain a PPG signal, but recordings from the other datasets do not. Where available, we used the ABD and THX respiratory signals, the PPG signal, and the ECG signal from each recording.

Dataset splits Although numerous prior works have explored the problem of sleep staging on the datasets used, there are no widely-established fixed training, validation and test partitions. We therefore establish new splits for all data sets, excluding nights that have not been annotated with multiple sleep stages as done in prior work (Phan et al., 2022). For datasets that contain multiple recordings from a single participant, we ensured that no participant appeared in both the test set and either the training or validation sets. No other exclusion criteria–such as signal quality heuristics (Jones et al., 2024)–were explicitly used since one of the key aims of training on multiple modalities jointly is to improve robustness to noise on any particular channel.

The size of our training, validation and test set splits for each dataset are listed in Table 1. These were chosen to be in line with those used in prior work, e.g. (Sridhar et al., 2020). Our splits were carefully constructed to additionally allow evaluation

on the aggregated test set proposed by Jones et al. (2024), which uses multiple PSG datasets to create a test set that approximately matches the 2022 US census demographics. Throughout the remainder of this paper, we refer to this as the 'Census' test set. More detail on the construction of our training, validation and test sets is provided in Appendix B.

Preprocessing We minimally processed all signals using a similar process to that described by Kotzen et al. (2023), padding or truncating each recording to 10 h (i.e. sequence length $T = 1200$), re-sampling each signal to the same frequency across recordings, and applying unit normalisation. The ECG and PPG signals were resampled such that each 30-second sleep epoch consisted of $k = 1024$ data points (≈ 34 Hz), which simplifies temporal alignment during pooling operations within the convolutional layers of the signal encoders. Since respiratory signals are generally sampled at a lower frequency during PSG recordings (e.g. 5-10 Hz in SHHS), the ABD and THX signals were resampled to a lower frequency of $k = 256$ data points per sleep epoch (≈ 8 Hz), reducing the computational and memory requirements of the model during training and inference.

4.2. Model training

All models were trained to minimise the cross-entropy loss between expert-annotated sleep stages and model outputs using the AdamW optimiser (Loshchilov and Hutter, 2019) with a batch size of 16 and weight decay of 10^{-2}. For the learning rate schedule, we used a linear warm-up of 2000 steps to a maximum learning rate $\epsilon = 10^{-3}$ followed by an exponential decay to zero. Training continued until there was no decrease in the loss on the validation set for 5 epochs, which typically required around 30 epochs in total. Further training details can be found in Appendix D. The checkpoint which resulted in the lowest validation loss was restored for evaluation. Model hyper-parameters were tuned using the validation sets before evaluation on the test sets took place.

Augmentation As noted by Jones et al. (2024), signals such as the ECG are sometimes inverted due to electrodes being connected the wrong way around. To improve robustness, all signals were randomly inverted (multiplied by -1) with a 50% probability during training.

Table 1: Demographics, dataset split sizes, and signal availability for the PSG datasets used.

Characteristic	SHHS	MESA	WSC	CHAT	CFS	CCSHS	MROS
Demographics[†]							
Age, mean	65.2	69.6	59.8	7.2	41.4	17.7	78.7
Sex, m:f	0.88:1	0.87:1	1.17:1	0.94:1	0.81:1	1.02:1	1:0
AHI[‡], mean	15.2	20.4	20.0	5.5	13.2	1.5	18.3
Splits, N (%)							
Train	6441 (81%)	1541 (84%)	1380 (65%)	1132 (79%)	452 (75%)	272 (64%)	0
Validation	500 (6%)	100 (5%)	250 (12%)	100 (7%)	50 (8%)	50 (12%)	0
Test	1000 (13%)	200 (11%)	500 (23%)	200 (14%)	100 (17%)	100 (24%)	1000
Signals							
ECG/ABD/THX	7941	1841	2130	1432	602	422	1000
PPG	0	1841	0	1139	284	422	0

[†]Calculated from NSRR harmonized variables (nsrr_age, nsrr_sex, nsrr_ahi_hp3u). [‡]AHI: Apnea-Hypopnoea Index.

4.3. Model hyper-parameters

Hyper-parameters for the wav2sleep model are listed in Table 2. In each signal encoder, the number of residual layers, and the number of channels in each layer, were chosen so that the resulting feature dimension is independent of the relative sampling rate k. For simplicity, we retained this feature dimension ($\dim(z_t^i) = \dim(z_t) = 128$) throughout the remainder of the model. Additional architecture details can be found in Appendix C.

Table 2: wav2sleep model hyper-parameters.

Hyper-parameter	Value
Global	
Feature dimension	128
Activation function	GELU[†]
Dropout	0.1
Signal Encoders	
Kernel size	3
Channels ($k = 256$)	(16,32,64,64,128,128)
Channels ($k = 1024$)	(16,16,32,32,64,64,128,128)
Epoch Mixer	
Transformer layers	2
Hidden dimension	512
Attention heads	8
Sequence Mixer	
Dilated blocks	2
Kernel size	7
Dilations (per block)	(1,2,4,8,16,32)

[†]Hendrycks and Gimpel (2023).

4.4. Stochastic masking

During training, to handle differences in the available modalities within a batch, we padded unavailable signals and added a mask to the attention matrices of the epoch mixer. To aid test-time generalisation to a subset of modalities, we randomly sampled a subset of the available signals for each recording via additional masking of the attention matrix. Where available, the input signals were masked with the following probabilities:

$$p(m_{\text{ABD}}) = 0.7 \quad p(m_{\text{THX}}) = 0.7$$
$$p(m_{\text{ECG}}) = 0.5 \quad p(m_{\text{PPG}}) = 0.1$$

These values were intuitively chosen so that the higher frequency ECG and PPG signals were less likely to be masked, and to increase the prevalence of the scarcer PPG signal.
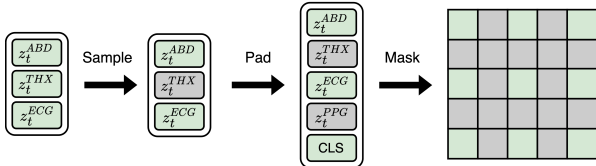

Figure 3: **Stochastic masking.** During training, we sample a random subset of the available modalities for each night of data. To retain a fixed batch shape, we pad unavailable modalities and apply a mask to the self-attention matrices of the epoch mixer.

Our stochastic masking process draws inspiration from masked modeling approaches e.g. (Baevski et al., 2020; He et al., 2021), and is similar to hierarchical channel sampling (HCS, (Bao et al., 2024)) from the visual domain. However, the practical implementation of HCS requires all samples within a batch to have the same available channels, to retain a fixed batch shape during training. Our approach (illustrated in Figure 3) allows heterogeneity *within* batches, which simplifies training on heterogeneous datasets. During inference, we simply pass only the available signals to the transformer i.e. there is no need for any padding or manipulation of the attention mechanism to handle different numbers of input signals at test time.

5. Results and Discussion

In this section, we report the performance of our model in four-class sleep staging, merging N1 and N2 into a single 'Light' sleep class as commonly done in prior work. We report total Cohen's κ (κ_T) and accuracy (Ac_T) calculated over all sleep epochs in each test set. All results are averages over three training runs using different random seeds. Our full set of results for all dataset-modality combinations evaluated can be found in Appendix A.3.

5.1. Cross-modal learning

In Table 3, we compare the performance of three approaches to PPG-based sleep staging:

1. Direct training on (scarce) PPG signals.

2. Transfer learning from ECG to PPG signals.

3. Joint training on all available modalities.

We additionally compare with a re-implementation of SleepPPG-Net (Kotzen et al., 2023), trained using the same splits and learning procedure as our model. Using transfer learning ($\mathcal{S}_{\mathrm{Train}} = \mathrm{ECG} \rightarrow \mathrm{PPG}$), we pre-train using the ECG signal, then fine-tune using the PPG signal, resuming the learning rate schedule. Across datasets, we find that our joint training approach with stochastic masking consistently leads to better performance than either direct training or transfer learning for the scarce PPG modality.

Similarly, Table 4 compares the performance of direct and joint training for sleep staging using the (abundant) ECG signal. For the SHHS and WSC datasets, and for the completely held-out MROS

Table 3: Performance (κ_T) for $\mathcal{S}_{\mathrm{Test}} = \mathrm{PPG}$.

Model	$\mathcal{S}_{\mathrm{Train}}$	Dataset			
		MESA	CHAT	CFS	CCSHS
SleepPPG-Net	PPG	0.713	0.757	0.731	0.803
	ECG→PPG	0.724	0.767	0.746	0.808
wav2sleep	PPG	0.728	0.777	0.751	0.817
	ECG→PPG	0.732	0.779	0.754	0.811
	All[†]	**0.742**	**0.793**	**0.763**	**0.832**

[†]ABD+THX+ECG (+PPG for MESA, CHAT, CFS, CCSHS)

dataset, joint training resulted in the best performance. However, for some datasets, we found that joint training without the PPG signal ($\mathcal{S}_{\mathrm{Train}} = $ No PPG) resulted in better ECG-only performance. This indicates that cross-modal learning from respiratory signals to ECG was able to occur, but that there is a trade-off between learning from ECG and PPG signals for some datasets. This is a limitation of our work which is further discussed in Appendix A.1.

Table 4: Performance (κ_T) for $\mathcal{S}_{\mathrm{Test}} = \mathrm{ECG}$.

Model	$\mathcal{S}_{\mathrm{Train}}$	Dataset				
		SHHS	WSC	CFS	Census	MROS
SleepPPG-Net	ECG	0.722	0.671	0.762	0.765	0.712
wav2sleep	ECG	0.733	0.683	0.785	0.786	0.746
	No PPG	0.738	0.686	**0.789**	**0.792**	0.748
	All	**0.739**	**0.689**	0.784	0.783	**0.750**

[†]ABD+THX+ECG (+PPG for MESA, CHAT, CFS, CCSHS)
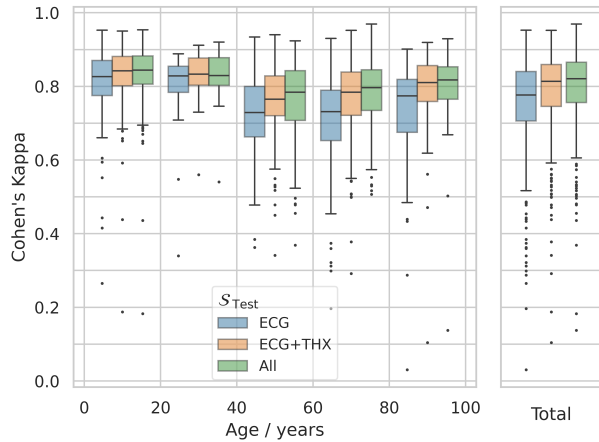
5.2. Varying modalities

Figure 4 shows confusion matrices between expert-annotated sleep stages and model outputs for different test-time modalities using the Census test set, aggregated over all sleep epochs. Here we see that the addition of breathing signals (ABD, THX) is particularly helpful in distinguishing both Wake and REM from Light (N1+N2) sleep. Using just the ECG and THX signals, we obtain a Cohen's κ of 0.812. Whilst caution should be taken when interpreting κ values, notably, using the rule-of-thumb proposed by Landis and Koch (1977) this corresponds to 'almost perfect' agreement with the expert-annotated sleep stages.

In Figure 5, we plot the performance of the wav2sleep model for different age ranges and test-time modalities $\mathcal{S}_{\mathrm{Test}}$. We observe good performance across age ranges, and that using more modalities consistently leads to improved performance, particularly by reducing the quantity and severity of outliers.
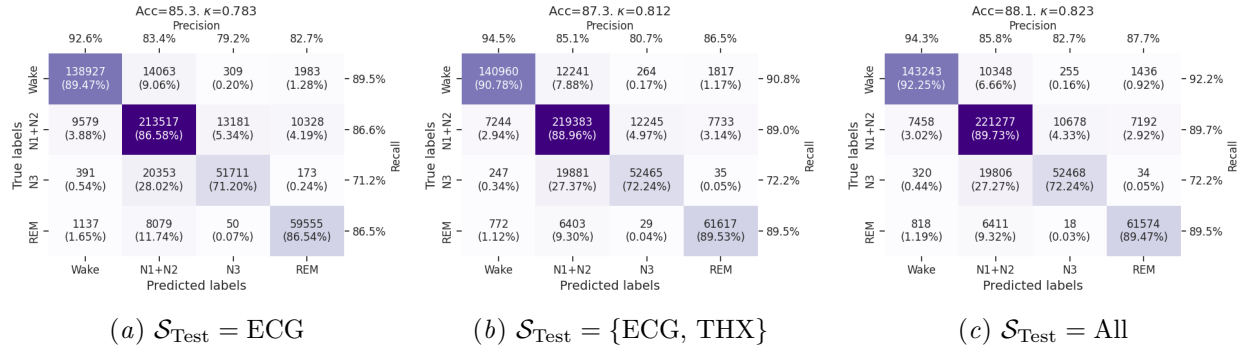
Figure 4: Sleep stage confusion matrices for varying $\mathcal{S}_{\text{Test}}$ on the Census test set.
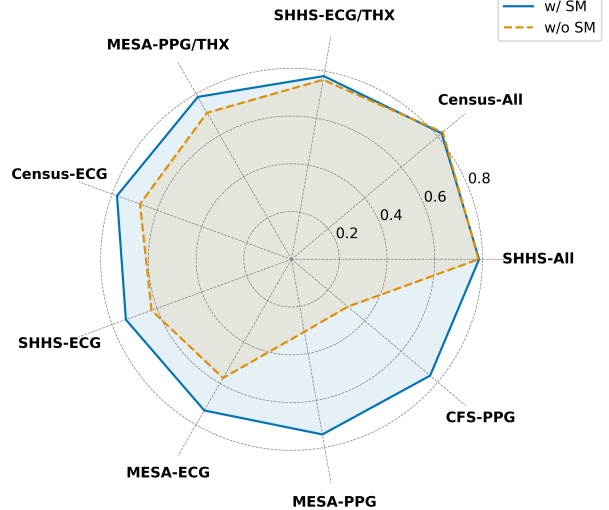
These outliers are often caused by noise on a particular signal (see Appendix A.2), but can also be caused by specific physiological conditions. Notably, we found that when using the ECG as the sole input, performance improved with apnea severity for subjects with cardiac arrhythmia (see Appendix A.4). This is in contrast to the general trend seen in prior work that the performance of sleep staging models tends to decrease with apnea severity (Korkalainen et al., 2020).

5.3. Stochastic masking

Figure 6 shows the performance of wav2sleep for various dataset-modality combinations with and without the use of stochastic masking during training. Here we can see that stochastic masking is essential for generalisation to subsets of modalities at test-time, whilst maintaining equivalent performance when using all modalities.

Figure 5: Performance (κ_T) of wav2sleep against age for varying $\mathcal{S}_{\text{Test}}$ on the Census dataset.

This highlights how, during real-world deployment, the set of input modalities (contact sensors) may need to be chosen in a patient-specific manner to ensure an expected level of accuracy in the presence of physiological confounders. The use of a single, unified model such as wav2sleep can help to simplify such a process.

Figure 6: Performance (κ_T) of wav2sleep for various dataset-modality combinations with and without stochastic masking (SM) during training.

5.4. Comparison with prior work

In Table 5, we compare the performance of wav2sleep after training on all modalities with prior models trained on specific modalities. We follow the exclusion criteria of Jones et al. (2024) and compare against prior work that has explicitly reported the use of distinct training, validation and test sets. Across multiple datasets and combinations of test-time modalities, the wav2sleep model outperforms existing methods for sleep staging from cardio-respiratory signals.

Table 5: Comparison of cardio-respiratory sleep staging methods for different test-time modalities $\mathcal{S}_{\text{Test}}$.

Dataset	$\mathcal{S}_{\text{Test}}$	Method	κ_T	Ac_T
SHHS	ECG, THX	Bakker et al. (2021)[†]	0.64	76.7
		Carter et al. (2024)	0.75	83.0
		wav2sleep	**0.78**	**85.0**
	ECG	Sridhar et al. (2020)	0.66	77.0
		wav2sleep	**0.74**	**82.3**
MESA	PPG, THX	Bakker et al. (2021)[†]	0.68	79.8
		wav2sleep	**0.78**	**86.2**
	ECG, THX	Carter et al. (2024)	0.77	85.2
		wav2sleep	**0.78**	**86.1**
	ECG	Sridhar et al. (2020)	0.69	80.0
		wav2sleep	**0.73**	**82.8**
Census	ECG	Jones et al. (2024)[‡]	0.77	-
		wav2sleep	**0.78**	**84.8**

Additional model inputs: [†]Nasal airflow,[‡]age and sex.

6. Conclusions

In this paper, we have introduced wav2sleep, a deep learning model for automated sleep stage classification that can operate on a variable number of input modalities during training and inference. After joint training on over 10,000 nights of publicly available data from six heterogeneous datasets, this single, unified model leads to improved performance compared to direct training and transfer learning methods across a range of test-time modalities and datasets. Our work further improves the accuracy of sleep staging across a range of important modalities, such as ECG, PPG and respiratory signals, bringing accurate, low-cost sleep monitoring from less obtrusive contact sensors closer to clinical practice.

Future Work We have focused on learning from cardio-respiratory signals since sleep staging from these modalities is of particular interest. However, using additional signals such as the EEG may help to further improve the quality of the learnt representations. Finally, the generalised architecture of wav2sleep, particularly the ability to jointly train it on heterogeneous, multi-modal time-series, means that it could be used to complement unsupervised approaches, e.g. (Thapa et al., 2024).

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization, July 2016. URL http://arxiv.org/abs/1607.06450. arXiv:1607.06450 [cs, stat].

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html.

Jessie P. Bakker, Marco Ross, Ray Vasko, Andreas Cerny, Pedro Fonseca, Jeff Jasko, Edmund Shaw, David P. White, and Peter Anderer. Estimating sleep stages using cardiorespiratory signals: validation of a novel algorithm across a wide range

of sleep-disordered breathing severity. *Journal of Clinical Sleep Medicine*, 17(7):1343–1354, July 2021. ISSN 1550-9389, 1550-9397. doi: 10.5664/jcsm.9192. URL http://jcsm.aasm.org/doi/10.5664/jcsm.9192.

Yujia Bao, Srinivasan Sivanandan, and Theofanis Karaletsos. Channel Vision Transformers: An Image Is Worth 1 x 16 x 16 Words. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=CK5Hfb5hBG.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The Long-Document Transformer, December 2020. URL http://arxiv.org/abs/2004.05150. arXiv:2004.05150 [cs].

Terri Blackwell, Kristine Yaffe, Sonia Ancoli-Israel, Susan Redline, Kristine E. Ensrud, Marcia L. Stefanick, Alison Laffan, Katie L. Stone, and for the Osteoporotic Fractures in Men Study Group. Associations Between Sleep Architecture and Sleep-Disordered Breathing and Cognition in Older Community-Dwelling Men: The Osteoporotic Fractures in Men Sleep Study. *Journal of the American Geriatrics Society*, 59(12):2217–2225, 2011. ISSN 1532-5415. doi: 10.1111/j.1532-5415.2011.03731.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1532-5415.2011.03731.x.

Jonathan F. Carter, João Jorge, Oliver Gibson, and Lionel Tarassenko. SleepVST: Sleep Staging from Near-Infrared Video Signals using Pre-Trained Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12489, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Carter_SleepVST_Sleep_Staging_from_Near-Infrared_Video_Signals_using_Pre-Trained_Transformers_CVPR_2024_paper.html.

Stanislas Chambon, Mathieu N. Galtier, Pierrick J. Arnal, Gilles Wainrib, and Alexandre Gramfort. A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26 (4):758–769, April 2018. ISSN 1558-0210. doi: 10.1109/TNSRE.2018.2813138. Conference Name: IEEE Transactions on Neural Systems and Rehabilitation Engineering.

Peter H. Charlton, John Allen, Raquel Bailón, Stephanie Baker, Joachim A. Behar, Fei Chen, Gari D. Clifford, David A. Clifton, Harry J. Davies, and Cheng Ding. The 2023 wearable photoplethysmography roadmap. *Physiological measurement*, 44(11):111001, 2023. URL https://iopscience.iop.org/article/10.1088/1361-6579/acead2/meta.

Xiaoli Chen, Rui Wang, Phyllis Zee, Pamela L. Lutsey, Sogol Javaheri, Carmela Alcántara, Chandra L. Jackson, Michelle A. Williams, and Susan Redline. Racial/Ethnic Differences in Sleep Disturbances: The Multi-Ethnic Study of Atherosclerosis (MESA). *Sleep*, 38(6):877–888, June 2015. ISSN 1550-9109. doi: 10.5665/sleep.4732.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. URL http://arxiv.org/abs/1810.04805. arXiv:1810.04805 [cs].

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, October 2020. URL https://openreview.net/forum?id=YicbFdNTTy.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners, December 2021. URL http://arxiv.org/abs/2111.06377. arXiv:2111.06377 [cs].

Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs), June 2023. URL http://arxiv.org/abs/1606.08415. arXiv:1606.08415 [cs].

C. Iber. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology, and Technical Specification. 2007. URL https://cir.nii.ac.jp/crid/1370004237604151044.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. URL http://proceedings.mlr.press/v37/ioffe15.html.

Adam M. Jones, Laurent Itti, and Bhavin R. Sheth. Expert-level sleep staging using an electrocardiography-only feed-forward neural network. *Computers in Biology and Medicine*, 176: 108545, June 2024. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2024.108545. URL https://www.sciencedirect.com/science/article/pii/S0010482524006292.

Jan W. Kantelhardt, Thomas Penzel, Sven Rostig, Heinrich F. Becker, Shlomo Havlin, and Armin Bunde. Breathing during REM and non-REM sleep: correlated versus uncorrelated behaviour. *Physica A: Statistical Mechanics and its Applications*, 319:447–457, March 2003. ISSN 0378-4371. doi: 10.1016/S0378-4371(02) 01502-9. URL https://www.sciencedirect.com/science/article/pii/S0378437102015029.

Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. Measuring Catastrophic Forgetting in Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. ISSN 2374-3468. doi: 10.1609/aaai.v32i1.11651. URL https://ojs.aaai.org/index.php/AAAI/article/view/11651. Number: 1.

Konstantinos Kontras, Christos Chatzichristos, Huy Phan, Johan Suykens, and Maarten De Vos. CoRe-Sleep: A Multimodal Fusion Framework for Time Series Robust to Imperfect Modalities. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:840–849, 2024. ISSN 1558-0210. doi: 10.1109/TNSRE.2024.3354388. URL https://ieeexplore.ieee.org/abstract/document/10400520. Conference Name: IEEE Transactions on Neural Systems and Rehabilitation Engineering.

Henri Korkalainen, Juhani Aakko, Sami Nikkonen, Samu Kainulainen, Akseli Leino, Brett Duce, Isaac O. Afara, Sami Myllymaa, Juha Töyräs, and Timo Leppänen. Accurate Deep Learning-Based Sleep Staging in a Clinical Population With Suspected Obstructive Sleep Apnea. *IEEE Journal of Biomedical and Health Informatics*, 24(7):2073–2081, July 2020. ISSN 2168-2208. doi: 10.1109/JBHI.2019.2951346. URL https://ieeexplore.ieee.org/abstract/document/8936942/authors#authors.

Kevin Kotzen, Peter H. Charlton, Sharon Salabi, Lea Amar, Amir Landesberg, and Joachim A. Behar. SleepPPG-Net: A Deep Learning Algorithm for Robust Sleep Staging From Continuous Photoplethysmography. *IEEE Journal of Biomedical and Health Informatics*, 27(2):924–932, February 2023. ISSN 2168-2194, 2168-2208. doi: 10.1109/JBHI.2022.3225363. URL https://ieeexplore.ieee.org/document/9965588/.

J. Richard Landis and Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, 1977. ISSN 0006-341X. doi: 10.2307/2529310. URL https://www.jstor.org/stable/2529310. Publisher: [Wiley, International Biometric Society].

Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, Daniel Haziza, Luca Wehrstedt, Jeremy Reizenstein, and Grigory Sizov. xFormers: A modular and hackable Transformer modelling library, 2022. URL https://github.com/facebookresearch/xformers.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Carole L. Marcus, Reneé H. Moore, Carol L. Rosen, Bruno Giordani, Susan L. Garetz, H. Gerry Taylor, Ron B. Mitchell, Raouf Amin, Eliot S. Katz, Raanan Arens, Shalini Paruthi, Hiren Muzumdar,

David Gozal, Nina Hattiangadi Thomas, Janice Ware, Dean Beebe, Karen Snyder, Lisa Elden, Robert C. Sprecher, Paul Willging, Dwight Jones, John P. Bent, Timothy Hoban, Ronald D. Chervin, Susan S. Ellenberg, Susan Redline, and Childhood Adenotonsillectomy Trial (CHAT). A randomized trial of adenotonsillectomy for childhood sleep apnea. *The New England Journal of Medicine*, 368 (25):2366–2376, June 2013. ISSN 1533-4406. doi: 10.1056/NEJMoa1215881.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Aakash K. Patel, Vamsi Reddy, and John F. Araujo. *Physiology, Sleep Stages*. StatPearls Publishing, April 2022. URL https://www.ncbi.nlm.nih.gov/books/NBK526132/.

T. Penzel. Is heart rate variability the simple solution to diagnose sleep apnoea? *European Respiratory Journal*, 22(6):870–971, December 2003. ISSN 0903-1936, 1399-3003. doi: 10.1183/09031936.03.00102003. URL https://erj.ersjournals.com/content/22/6/870. Publisher: European Respiratory Society Section: Editorials.

Huy Phan and Kaare Mikkelsen. Automatic sleep staging of EEG signals: recent development, challenges, and future directions. *Physiological Measurement*, 43(4):04TR01, April 2022. ISSN 0967-3334. doi: 10.1088/1361-6579/ac6049. URL https://dx.doi.org/10.1088/1361-6579/ac6049.

Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y. Chen, 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, March 2019. ISSN 1534-4320, 1558-0210. doi: 10.1109/TNSRE.2019.2896659. URL https://ieeexplore.ieee.org/document/8631195/.

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, September 2022. ISSN 1939-3539. doi: 10.1109/TPAMI.2021.3070057.

Jathurshan Pradeepkumar, Mithunjha Anandakumar, Vinith Kugathasan, Dhinesh Suntharalingham, Simon L. Kappel, Anjula C. De Silva, and Chamira U. S. Edussooriya. Towards Interpretable Sleep Stage Classification Using Cross-Modal Transformers. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, pages 1–1, 2024. ISSN 1558-0210. doi: 10.1109/TNSRE.2024.3438610. URL https://ieeexplore.ieee.org/abstract/document/10623416.

S. F. Quan, B. V. Howard, C. Iber, J. P. Kiley, F. J. Nieto, G. T. O'Connor, D. M. Rapoport, S. Redline, J. Robbins, J. M. Samet, and P. W. Wahl. The Sleep Heart Health Study: design, rationale, and methods. *Sleep*, 20(12):1077–1085, December 1997. ISSN 0161-8105.

Mustafa Radha, Pedro Fonseca, Arnaud Moreau, Marco Ross, Andreas Cerny, Peter Anderer, Xi Long, and Ronald M. Aarts. A deep transfer learning approach for wearable sleep stage classification with photoplethysmography. *npj Digital Medicine*, 4(1):1–11, September 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00510-8. URL https://www.nature.com/articles/s41746-021-00510-8.

S. Redline, P. V. Tishler, T. D. Tosteson, J. Williamson, K. Kump, I. Browner, V. Ferrette, and P. Krejci. The familial aggregation of obstructive sleep apnea. *American Journal of Respiratory and Critical Care Medicine*, 151(3 Pt 1): 682–687, March 1995. ISSN 1073-449X. doi: 10.1164/ajrccm/151.3_Pt_1.682.

Carol L. Rosen, Emma K. Larkin, H. Lester Kirchner, Judith L. Emancipator, Sarah F. Bivins, Susan A. Surovec, Richard J. Martin, and Susan Redline. Prevalence and risk factors for sleep-disordered breathing in 8- to 11-year-old children:

association with race and prematurity. *The Journal of Pediatrics*, 142(4):383–389, April 2003. ISSN 0022-3476. doi: 10.1067/mpd.2003.28.

Z. Shinar, A. Baharav, Y. Dagan, and S. Akselrod. Automatic detection of slow-wave-sleep using heart rate variability. In *Computers in Cardiology 2001. Vol.28*, pages 593–596, September 2001. doi: 10. 1109/CIC.2001.977725.

Niranjan Sridhar, Ali Shoeb, Philip Stephens, Alaa Kharbouch, David Ben Shimol, Joshua Burkart, Atiyeh Ghoreyshi, and Lance Myers. Deep learning for automated sleep staging using instantaneous heart rate. *npj Digital Medicine*, 3(1):1–10, August 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-0291-x. URL https://www.nature.com/articles/s41746-020-0291-x.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, February 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.127063. URL https://www.sciencedirect.com/science/article/pii/S0925231223011864.

Rahul Thapa, Bryan He, Magnus Ruud Kjaer, Hyatt Moore, Gauri Ganjoo, Emmanuel Mignot, and James Zou. SleepFM: Multi-modal Representation Learning for Sleep Across Brain Activity, ECG and Respiratory Signals, May 2024. URL http://arxiv.org/abs/2405.17766. arXiv:2405.17766 [cs, eess].

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance Normalization: The Missing Ingredient for Fast Stylization, November 2017. URL http://arxiv.org/abs/1607.08022. arXiv:1607.08022 [cs].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv:1706.03762 [cs]*, December 2017. URL http://arxiv.org/abs/1706.03762. arXiv:1706.03762.

Jiquan Wang, Sha Zhao, Haiteng Jiang, Yangxuan Zhou, Zhenghe Yu, Tao Li, Shijian Li, and Gang Pan. CareSleepNet: A Hybrid Deep Learning Network for Automatic Sleep Staging. *IEEE Journal of Biomedical and Health Informatics*, pages 1–14, 2024. ISSN 2168-2208. doi: 10.1109/JBHI.2024.

3426939. URL https://ieeexplore.ieee.org/abstract/document/10595067.

Terry Young, Mari Palta, Jerome Dempsey, Paul E. Peppard, F. Javier Nieto, and K. Mae Hla. Burden of sleep apnea: rationale, design, and major findings of the Wisconsin Sleep Cohort study. *WMJ: official publication of the State Medical Society of Wisconsin*, 108(5):246–249, August 2009. ISSN 1098-1861.

Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The National Sleep Research Resource: towards a sleep data commons. *Journal of the American Medical Informatics Association: JAMIA*, 25(10):1351–1358, October 2018. ISSN 1527-974X. doi: 10.1093/jamia/ocy064.

Hangyu Zhu, Wei Zhou, Cong Fu, Yonglin Wu, Ning Shen, Feng Shu, Huan Yu, Wei Chen, and Chen Chen. MaskSleepNet: A Cross-Modality Adaptation Neural Network for Heterogeneous Signals Processing in Sleep Staging. *IEEE Journal of Biomedical and Health Informatics*, 27(5):2353–2364, May 2023. ISSN 2168-2194, 2168-2208. doi: 10.1109/JBHI.2023.3253728. URL https://ieeexplore.ieee.org/document/10061562/.

Appendix A. Additional results and discussion

A.1. Stochastic masking trade-offs

As shown in Table 4, we observed a small trade-off between ECG and PPG performance using our stochastic masking approach. During training, there are a finite number of optimisation steps before the model begins to overfit. This results in a small performance trade-off between different test-time modality combinations depending on the masking parameters, which determine the relative frequency of modalities observed during training. For example, by increasing the PPG masking probability $p(m_{\text{PPG}})$ to 0.2 we found that the kappa values slightly decreased by 0.01-0.02 across datasets when using only the PPG at test-time, but increased by around the same amount using just the ECG.

A similar effect was noted in the similar approach of Hierarchical Channel Sampling (Bao et al., 2024),

where performance was best for combinations of channels that were most frequently sampled during training. Our stochastic masking procedure means that ECG-only examples are infrequently sampled during training, accounting for less than 1 example per batch on average for datasets that have all four signals available. Improvements to the stochastic masking procedure, stronger regularisation, and/or a larger batch size (see Appendix C.1) may help to address this in future work.

A.2. Signal noise

Figure 7 shows the performance of wav2sleep on the MESA test set for different test-time modalities, grouped by ECG signal quality.[1] Here we can observe how the use of multiple input modalities provides improved redundancy. When the ECG signal is of poor quality, the use of additional signals, e.g. THX, helps to maintain good performance.
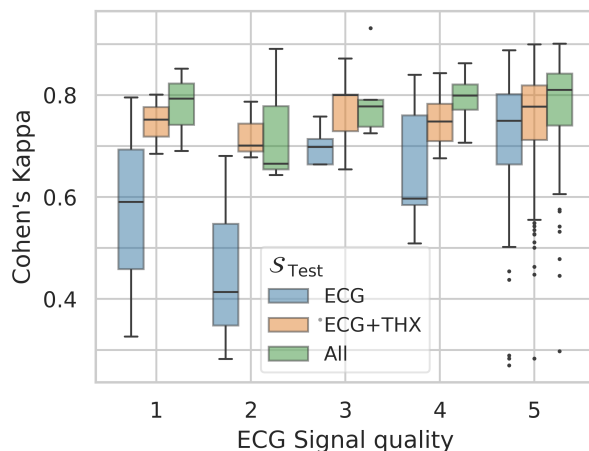
Figure 7: Performance of wav2sleep on the MESA test set, grouped by ECG signal quality index.

A.3. Varying test-time modalities

After joint training on all datasets and input modalities ($\mathcal{S}_{\text{Train}} = \text{All}$), the performance of the wav2sleep model for different test-time modalities $\mathcal{S}_{\text{Test}}$ is listed in Table 6.

---

1. As measured by the 'quecg5' metadata variable.

A.4. Example hypnograms

Figure 8 shows example sleep hypnograms generated by the wav2sleep model using different test-time modalities. This night of data corresponds to an elderly male with diagnosed cardiac arrhythmia and mild sleep apnea. Visually, the ECG is of good quality. However, using the ECG as the sole input to the model results in poor agreement with expert-annotated sleep stages. The addition of the thoracic signal results in a significant performance improvement.

Notably, we found that when using the ECG as the sole input, performance **improved** with apnea severity for subjects with cardiac arrhythmia (see Figure 9). This is in contrast to the general trend seen in prior work that performance tends to decrease with apnea severity e.g. (Korkalainen et al., 2020). We hypothesise that, for subjects with arrhythmia, the model may confuse heart rate variability (HRV) caused by arrhythmia for HRV caused by the more common condition of sleep apnea (Penzel, 2003). In turn, this may confound the learnt mapping between physiological features and sleep stages i.e. the mapping g described in Section 2. Adding a respiratory signal (e.g. using ECG and THX) appears to remedy this issue, enabling the model to correctly distinguish apnea-induced variability.

Appendix B. Dataset processing

B.1. Scoring exclusions

Because of signal quality issues, some of the recordings in each dataset only have binary sleep–wake annotations, rather than full AASM (Wake, N1, N2, N3, REM) sleep stages. For these recordings, *all* sleep stages are typically assigned to the same integer as 'N2' sleep. This means that these labels should not be used for training or evaluation of multi-class sleep staging models. Where available, we used the harmonised 'nsrr_flag_spsw' metadata variable produced by the National Sleep Research Resource to exclude these recordings. Otherwise, we checked for the existence of either N1, N3 or REM sleep labels.

B.2. Construction of test sets

From the CFS and CHAT datasets, we created our validation and test sets using recordings where the PPG signal was available, to enable evaluation across all combinations of modalities. The remaining

Table 6: wav2sleep performance (κ_T) for different test-time modalities.

$\mathcal{S}_{\text{Test}}$	Dataset							
	SHHS	MESA	WSC	CHAT	CFS	CCSHS	MROS	Census
All†	0.786	0.796	0.737	0.836	0.803	0.857	0.805	0.821
ECG,THX	0.779	0.783	0.728	0.827	0.802	0.854	0.796	0.812
ECG	0.739	0.731	0.689	0.800	0.784	0.833	0.750	0.783
PPG	-	0.742	-	0.793	0.763	0.832	-	-

†ABD+THX+ECG (+PPG for MESA, CHAT, CFS and CCSHS)
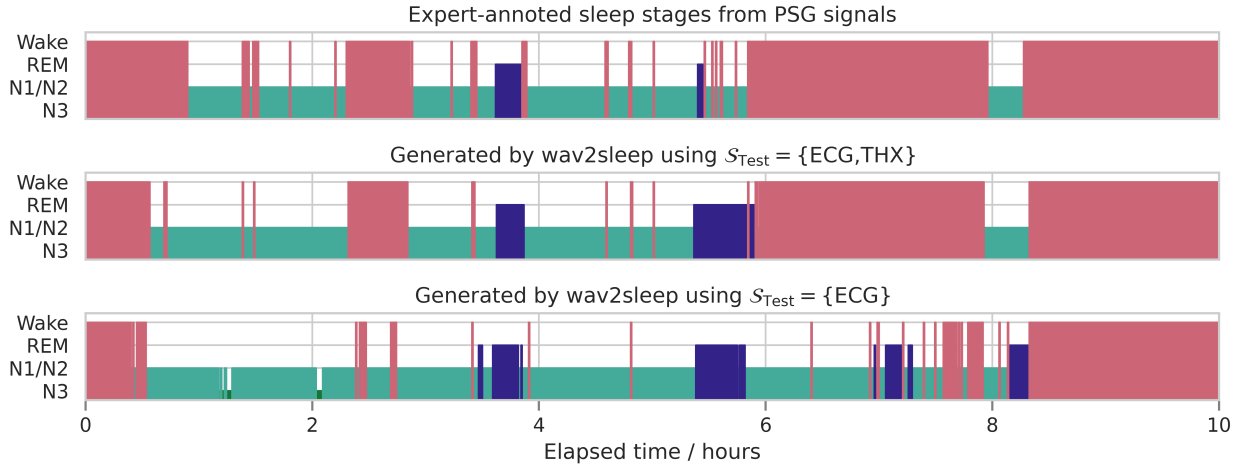
Figure 8: **Example sleep hypnograms for a subject with diagnosed cardiac arrhythmia.** (top) Annotated by a human expert using the PSG recording. (middle) Produced by the wav2sleep model using ECG and THX signals ($\kappa_T = 0.74$). (bottom) Produced by the wav2sleep model using the ECG signal ($\kappa_T = 0.19$).

recordings (with and without PPG available) were used for training. We used recordings from the non-randomised (single night per participant) arm of the CHAT dataset for our test set. Similarly to Carter et al. (2024), we selected 1000 nights for our SHHS test set by randomly choosing 500 participants who participated in both visits. For the WSC dataset, we selected 2 recordings from 250 participants who had undertaken at least 2 visits to form our test set of 500 recordings; additional recordings from participants in the test set were excluded. Our decision to use test sets with two recordings per person for the SHHS and WSC datasets was taken to maximise data usage while avoiding the same participant appearing simultaneously in both the training and test sets. In future work, these sets could be used for additional analysis such as the variation in perfor-

mance with age after controlling for identity. To enable evaluation on the census-balanced test set proposed by Jones et al. (2024)–which uses recordings from CCSHS, CFS, CHAT, MESA and WSC–we excluded their test set recordings from our training and validation sets.

Appendix C. Model design

Here we describe additional experiments and observations that informed the design and hyper-parameters of the wav2sleep model. Hyper-parameter search was informed by the minimum validation loss $\mathcal{L}_{\text{Min}}$ during initial experiments.
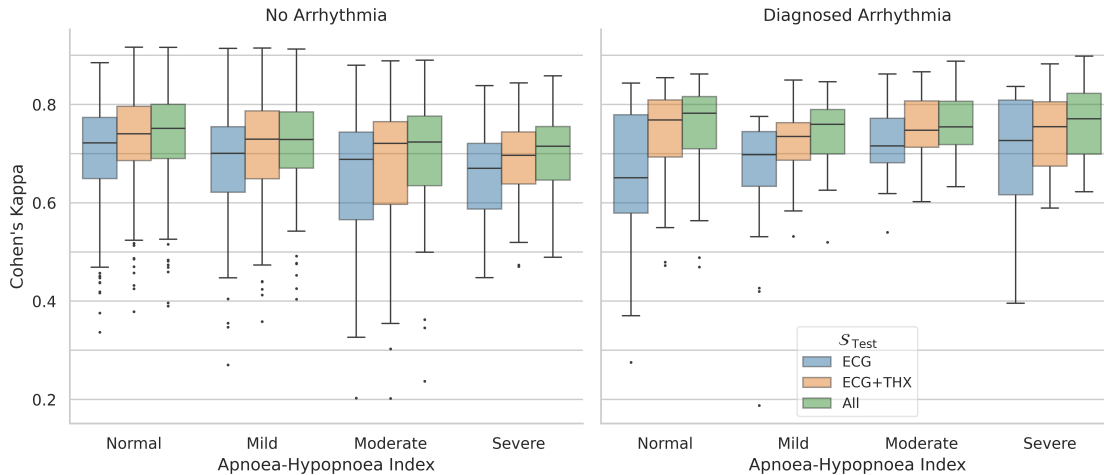
Figure 9: **Performance of wav2sleep on the WSC test set, grouped by apnea severity and arrhythmia**. For subjects with arrhythmia, and using only the ECG at test-time, performance *improves* with apnea severity. This is in contrast to the trend that performance decreases with apnea severity seen using other modalities and for subjects without diagnosed arrhythmia.

C.1. Signal Encoders

We found that using instance normalisation (Ulyanov et al., 2017) within the signal encoders and layer normalisation (Ba et al., 2016) in the sequence mixer improved training stability and performance. Because of our stochastic masking procedure, the number of examples of a signal within a batch will often be much smaller than the actual batch size, increasing the variance of statistics used by the more common approach of batch normalisation (Ioffe and Szegedy, 2015).

C.2. Epoch Mixer

We evaluated two designs for the epoch mixer:

1. A small transformer encoder (TE) i.e. our best-performing approach.

2. A linear concatenation and projection layer, zero-padding unavailable inputs.

The attention-based epoch mixer achieved a lower validation loss and higher Cohen's κ values across multiple datasets and modalities.

C.3. Sequence Mixer

We evaluated two designs for the sequence mixer:

1. A dilated convolutional (DCNN) design, as originally proposed by Sridhar et al. (2020).

Table 7: Validation metrics for different epoch mixer designs.

| Design | $\mathcal{L}_{\text{Min}}$ | Dataset-modality κ_T | | |
		C-ECG	M-PPG	S-ECG
Linear	0.351	0.770	0.741	0.719
TE	0.349	0.773	0.743	0.723

C - Census. M - MESA. S - SHHS

2. A transformer encoder (TE) with sliding window attention (Beltagy et al., 2020) and rotary positional embeddings (Su et al., 2024).

We found that the DCNN design consistently achieved better performance across different modalities, and converged after fewer training epochs. The hyper-parameters of the dilated convolutional design have been carefully tuned through extensive hyper-parameter search in prior work (Sridhar et al., 2020; Kotzen et al., 2023).

Though we did perform a basic search over transformer hyper-parameters, such as the number of encoder layers and the context length, performing extensive tuning was deemed unnecessary given the results achieved using a convolutional design, and outside the scope of this paper. Using a well-tuned mixture of local and global attention (Beltagy et al., 2020) may yet lead to superior performance using a

Table 8: Validation metrics for different sequence mixer designs.

Design	$\mathcal{L}_{\mathrm{Min}}$	Dataset-modality		
		C-ECG	M-PPG	S-ECG
TE	0.355	0.764	0.724	0.706
DCNN	0.349	0.773	0.743	0.723

C - Census. M - MESA. S - SHHS

transformer-based architecture, but is left for future work.

Finally, it is worth noting that our implementations of both stochastic masking and local attention relied on naive masking of the attention matrix using PyTorch (Paszke et al., 2019). Using optimised sparse kernels (e.g. from xformers (Lefaudeux et al., 2022)) could provide significant speed-ups and efficiency gains on modern GPU architectures, making a transformer-based architecture a more attractive option for training on an even larger quantity of data.

Appendix D. Model training

Model parameters θ were found by minimising the unweighted cross-entropy loss between one-hot encoded labels $\boldsymbol{y}_{1:T} \in \mathbb{R}^{C \times T}$ and output probabilities $\boldsymbol{p}_{1:T} \in \mathbb{R}^{C \times T}$. For each night of data, the total cross-entropy loss is given by:

$$\mathcal{L}_\theta(\boldsymbol{y}_{1:T}, \boldsymbol{p}_{1:T}) = -\sum_{i=1}^{C} \sum_{j=1}^{T} \left(\boldsymbol{y}_{1:T} \odot \log(\boldsymbol{p}_{1:T})\right)_{ij} \quad (1)$$

where $\odot$ denotes the Hadamard product.

GPU training Experiments were performed using a computing cluster containing multiple GPU architectures. Gradient accumulation was used to ensure a consistent effective batch size of 16, using the largest (power of 2) batch size that could fit on the particular GPU(s) used in a given experiment. Using a single NVIDIA A100 40 GB, the actual batch size was 4 samples, and each epoch took 21 minutes, resulting in an average training time of around 10 hours.

Appendix E. Additional Acknowledgements