A. Additional Quantitative Results

TABLE S.I
HOLD-OUT TEST SET RESULTS FOR ALL MODELS. '#' REFERS TO THE NUMBER OF RECORDINGS IN THE TEST SET WHERE THE SIGNAL WAS AVAILABLE. AVERAGE ACCURACY IN TERMS OF 5-,4-,3-, AND 2-CLASS SLEEP STAGING IS SHOWN HERE.

		Accuracy [%]			
Signal(s)	**#**	W/N1/N2/N3/REM	W/N1-N2/N3/REM	W/NREM/REM	Wake/Sleep
All PSG electrodes	497	85.9	89.8	93.4	96.4
All odd EEG electrodes	500	85.5	89.5	93.3	96.4
F3-M2 (EEG)	500	85.6	89.5	93.1	96.2
C3-M2 (EEG)	500	85.3	89.4	93.2	96.2
O1-M2 (EEG)	500	83.4	87.9	92.6	96.1
Recommended PSG electrodes	497	86.3	90.2	93.6	96.5
All even EEG electrodes	500	85.8	89.8	93.3	96.5
F4-M1 (EEG)	500	85.5	89.4	93.0	96.1
C4-M1 (EEG)	500	85.5	89.6	93.2	96.3
O2-M1 (EEG)	500	83.8	88.2	92.7	96.2
E2-M2 (EOG)	497	85.0	89.2	93.2	96.2
E1-M2 (EOG)	500	84.3	88.8	92.9	95.9
Chin1-Chin3 (EMG)	500	74.5	80.6	87.8	92.5
Chin1-Chin2 (EMG)	500	74.9	80.9	88.1	92.7
Chin2-Chin3 (EMG)	500	74.9	80.9	88.0	92.6
HSAT expanded	434	79.0	84.3	90.8	94.4
HSAT reduced	434	78.3	83.6	90.2	94.1
Nasal cannula	434	76.5	81.9	89.0	93.4
Finger PPG	500	75.1	80.8	88.0	92.4
Thoracic belt	500	76.3	82.7	89.7	93.6
HSAT reduced	434	77.6	82.9	89.6	93.8
Nasal cannula	434	76.5	81.9	89.0	93.4
IHR from finger PPG	500	71.8	77.6	84.9	90.0
HSAT expanded	434	78.5	84.0	90.6	94.1
HSAT reduced	434	77.7	83.1	89.9	93.5
thermistor	434	72.9	79.2	87.0	91.6
ECG	500	76.9	82.2	89.1	93.2
Thoracic belt	500	76.3	82.7	89.7	93.6
HSAT expanded	66	78.2	83.5	90.5	93.7
HSAT reduced	66	76.4	81.2	88.5	92.1
PAP flow	66	69.5	74.4	83.1	87.8
Finger PPG	500	75.1	80.8	88.0	92.4
Thoracic belt	500	76.3	82.7	89.7	93.6
HSAT reduced	65	74.9	80.1	87.0	92.0
IBR from PAP flow	65	69.2	74.8	83.0	89.3
IHR from finger PPG	500	71.8	77.6	84.9	90.0
Left Leg and SCM	33	71.0	77.0	85.1	90.7
Left Leg (EMG)	500	66.9	72.2	81.2	88.5
Left SCM (EMG)	33	66.2	72.4	80.5	89.5
Right Leg and SCM	33	70.3	76.3	84.4	90.2
Right Leg (EMG)	500	66.7	72.0	80.8	88.2
Right SCM (EMG)	33	67.0	73.1	81.7	89.7
Left Leg and FDS	60	67.4	72.8	81.8	88.7
Left Leg (EMG)	500	66.9	72.2	81.2	88.5
Left FDS (EMG)	60	63.7	69.5	78.8	88.2
Right Leg and FDS	60	68.0	73.1	82.0	88.6
Right Leg (EMG)	500	66.7	72.0	80.8	88.2
Right FDS (EMG)	60	63.2	69.1	78.2	87.9
Abdominal belt	500	76.4	83.0	89.9	93.6
Snore microphone	500	72.1	78.0	85.9	91.9
IHR from ECG	500	70.1	75.8	83.8	89.1
IBR from RIP thorax	500	70.0	76.1	83.9	89.7
IBR from RIP abdomen	500	69.9	76.1	84.0	89.7
SpO2	500	68.7	74.7	82.8	89.1
IBR from nasal cannula	434	66.9	73.3	81.5	87.5
IBR from Thermistor	434	65.4	72.0	80.8	87.4
Suprasternal notch	72	63.1	69.5	78.4	86.0
IBR from esophageal pressure	24	59.0	67.9	76.7	83.2
IBR from Suprasternal notch	72	58.5	65.3	74.5	82.2
Esophageal pressure	24	55.5	62.1	72.5	80.7

TABLE S.II
HOLD-OUT TEST SET RESULTS FOR ALL MODELS. '#' REFERS TO THE NUMBER OF RECORDINGS IN THE HOLD-OUT TEST SET WHERE THE SIGNAL WAS AVAILABLE. AVERAGE COHEN'S KAPPA IN TERMS OF 5-,4-,3-, AND 2-CLASS SLEEP STAGING IS SHOWN HERE.

		Cohen's Kappa			
Signal(s)	**#**	W/N1/N2/N3/REM	W/N1-N2/N3/REM	W/NREM/REM	Wake/Sleep
All PSG electrodes	497	0.793	0.826	0.853	0.858
All odd EEG electrodes	500	0.789	0.822	0.850	0.863
F3-M2 (EEG)	500	0.791	0.823	0.846	0.855
C3-M2 (EEG)	500	0.787	0.821	0.848	0.857
O1-M2 (EEG)	500	0.760	0.795	0.835	0.850
Recommended PSG electrodes	497	0.799	0.832	0.856	0.864
All even EEG electrodes	500	0.794	0.827	0.851	0.863
F4-M1 (EEG)	500	0.790	0.822	0.844	0.850
C4-M1 (EEG)	500	0.791	0.824	0.848	0.857
O2-M1 (EEG)	500	0.764	0.800	0.837	0.851
E2-M2 (EOG)	497	0.784	0.820	0.850	0.858
E1-M2 (EOG)	500	0.776	0.815	0.845	0.852
Chin1-Chin3 (EMG)	500	0.624	0.672	0.731	0.709
Chin1-Chin2 (EMG)	500	0.630	0.677	0.737	0.716
Chin2-Chin3 (EMG)	500	0.631	0.678	0.735	0.716
HSAT expanded	434	0.697	0.740	0.801	0.793
HSAT reduced	434	0.686	0.731	0.791	0.783
Nasal cannula	434	0.661	0.705	0.767	0.762
Finger PPG	500	0.640	0.685	0.746	0.735
Thoracic belt	500	0.657	0.708	0.775	0.765
HSAT reduced	434	0.676	0.719	0.777	0.774
Nasal cannula	434	0.661	0.705	0.767	0.762
IHR from finger PPG	500	0.597	0.637	0.693	0.688
HSAT expanded	434	0.687	0.733	0.797	0.787
HSAT reduced	434	0.674	0.720	0.785	0.774
thermistor	434	0.603	0.650	0.717	0.702
ECG	500	0.669	0.711	0.773	0.772
Thoracic belt	500	0.657	0.708	0.775	0.765
HSAT expanded	66	0.678	0.722	0.797	0.778
HSAT reduced	66	0.652	0.690	0.759	0.735
PAP flow	66	0.562	0.594	0.661	0.634
Finger PPG	500	0.640	0.685	0.746	0.735
Thoracic belt	500	0.657	0.708	0.775	0.765
HSAT reduced	65	0.626	0.666	0.719	0.699
IBR from PAP flow	65	0.538	0.580	0.634	0.579
IHR from finger PPG	500	0.597	0.637	0.693	0.688
Left Leg and SCM	33	0.575	0.624	0.699	0.716
Left Leg (EMG)	500	0.526	0.558	0.621	0.632
Left SCM (EMG)	33	0.502	0.546	0.594	0.681
Right Leg and SCM	33	0.565	0.613	0.689	0.700
Right Leg (EMG)	500	0.523	0.556	0.616	0.629
Right SCM (EMG)	33	0.517	0.561	0.627	0.679
Left Leg and FDS	60	0.532	0.566	0.633	0.647
Left Leg (EMG)	500	0.526	0.558	0.621	0.632
Left FDS (EMG)	60	0.482	0.510	0.555	0.615
Right Leg and FDS	60	0.540	0.569	0.639	0.647
Right Leg (EMG)	500	0.523	0.556	0.616	0.629
Right FDS (EMG)	60	0.476	0.503	0.544	0.609
Abdominal belt	500	0.661	0.715	0.779	0.770
Snore microphone	500	0.597	0.639	0.692	0.714
IHR from ECG	500	0.571	0.609	0.670	0.659
IBR from RIP thorax	500	0.564	0.611	0.662	0.626
IBR from RIP abdomen	500	0.563	0.612	0.664	0.625
SpO2	500	0.542	0.589	0.642	0.624
IBR from nasal cannula	434	0.521	0.568	0.616	0.563
IBR from Thermistor	434	0.497	0.545	0.597	0.536
Suprasternal notch	72	0.464	0.499	0.553	0.573
IBR from esophageal pressure	24	0.419	0.473	0.523	0.474
IBR from Suprasternal notch	72	0.394	0.433	0.473	0.431
Esophageal pressure	24	0.373	0.410	0.466	0.483

TABLE S.III
HOLD-OUT TEST SET RESULTS FOR ALL MODELS. '#' REFERS TO THE NUMBER OF RECORDINGS IN THE HOLD-OUT TEST SET WHERE THE SIGNAL WAS AVAILABLE. AVERAGE F1-SCORES PER SLEEP STAGE ARE SHOWN HERE.

		F1 scores				
Signal(s)	#	Wake	N1	N2	N3	REM
All PSG electrodes	497	0.885	0.588	0.881	0.834	0.882
All odd EEG electrodes	500	0.888	0.599	0.876	0.821	0.871
F3-M2 (EEG)	500	0.882	0.593	0.878	0.833	0.869
C3-M2 (EEG)	500	0.883	0.596	0.874	0.826	0.872
O1-M2 (EEG)	500	0.878	0.574	0.857	0.767	0.857
Recommended PSG electrodes	497	0.890	0.598	0.883	0.845	0.881
All even EEG electrodes	500	0.886	0.599	0.879	0.833	0.871
F4-M1 (EEG)	500	0.876	0.591	0.879	0.834	0.868
C4-M1 (EEG)	500	0.882	0.598	0.877	0.828	0.872
O2-M1 (EEG)	500	0.878	0.572	0.859	0.787	0.858
E2-M2 (EOG)	497	0.887	0.577	0.865	0.828	0.876
E1-M2 (EOG)	500	0.881	0.582	0.855	0.821	0.874
Chin1-Chin3 (EMG)	500	0.760	0.342	0.775	0.677	0.807
Chin1-Chin2 (EMG)	500	0.766	0.352	0.778	0.677	0.813
Chin2-Chin3 (EMG)	500	0.767	0.353	0.778	0.683	0.810
HSAT expanded	434	0.833	0.434	0.811	0.724	0.845
HSAT reduced	434	0.824	0.393	0.804	0.718	0.832
Nasal cannula	434	0.806	0.375	0.787	0.702	0.811
Finger PPG	500	0.787	0.366	0.775	0.689	0.798
Thoracic belt	500	0.813	0.433	0.783	0.689	0.828
HSAT reduced	434	0.818	0.398	0.799	0.712	0.819
Nasal cannula	434	0.806	0.375	0.787	0.702	0.811
IHR from finger PPG	500	0.755	0.351	0.744	0.665	0.757
HSAT expanded	434	0.832	0.429	0.807	0.700	0.844
HSAT reduced	434	0.821	0.386	0.800	0.694	0.833
thermistor	434	0.759	0.331	0.762	0.640	0.774
ECG	500	0.821	0.399	0.786	0.699	0.829
Thoracic belt	500	0.813	0.433	0.783	0.689	0.828
HSAT expanded	66	0.820	0.373	0.814	0.669	0.853
HSAT reduced	66	0.788	0.320	0.798	0.670	0.834
PAP flow	66	0.709	0.239	0.733	0.591	0.784
Finger PPG	500	0.787	0.366	0.775	0.689	0.798
Thoracic belt	500	0.813	0.433	0.783	0.689	0.828
HSAT reduced	65	0.748	0.267	0.793	0.662	0.798
IBR from PAP flow	65	0.641	0.155	0.747	0.626	0.742
IHR from finger PPG	500	0.755	0.351	0.744	0.665	0.757
Left Leg and SCM	33	0.784	0.239	0.721	0.595	0.727
Left Leg (EMG)	500	0.709	0.226	0.698	0.598	0.689
Left SCM (EMG)	33	0.757	0.237	0.668	0.574	0.614
Right Leg and SCM	33	0.774	0.234	0.713	0.585	0.723
Right Leg (EMG)	500	0.705	0.220	0.696	0.597	0.680
Right SCM (EMG)	33	0.754	0.232	0.673	0.543	0.644
Left Leg and FDS	60	0.727	0.235	0.698	0.614	0.712
Left Leg (EMG)	500	0.709	0.226	0.698	0.598	0.689
Left FDS (EMG)	60	0.696	0.257	0.672	0.620	0.578
Right Leg and FDS	60	0.727	0.242	0.706	0.604	0.714
Right Leg (EMG)	500	0.705	0.220	0.696	0.597	0.680
Right FDS (EMG)	60	0.693	0.248	0.663	0.617	0.580
Abdominal belt	500	0.816	0.442	0.781	0.694	0.831
Snore microphone	500	0.771	0.362	0.749	0.668	0.729
IHR from ECG	500	0.732	0.346	0.729	0.637	0.740
IBR from RIP thorax	500	0.691	0.199	0.737	0.666	0.736
IBR from RIP abdomen	500	0.688	0.199	0.737	0.664	0.738
SpO2	500	0.697	0.181	0.722	0.634	0.712
IBR from nasal cannula	434	0.639	0.179	0.712	0.651	0.709
IBR from Thermistor	434	0.610	0.157	0.701	0.610	0.698
Suprasternal notch	72	0.671	0.232	0.662	0.561	0.596
IBR from esophageal pressure	24	0.568	0.129	0.657	0.579	0.607
IBR from Suprasternal notch	72	0.541	0.112	0.639	0.566	0.587
Esophageal pressure	24	0.614	0.117	0.569	0.522	0.529

B. Additional Qualitative Results

We here show a qualitative example for each of the signal(s) as shown in tables S.III and S.I. We show the most typical example for each, defined as the recording where it achieved median performance in terms of accuracy.

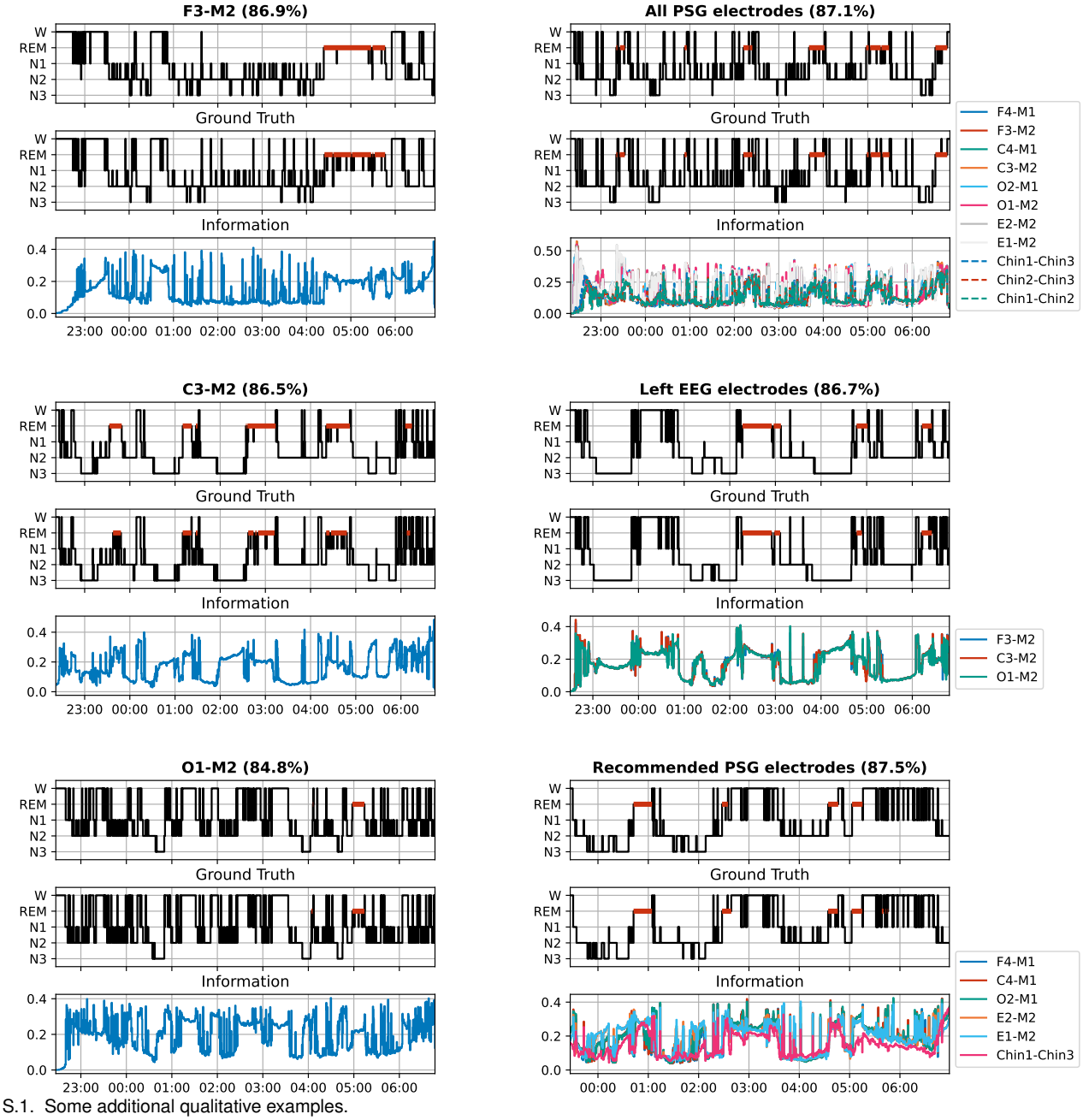


Fig. S.1. Some additional qualitative examples.

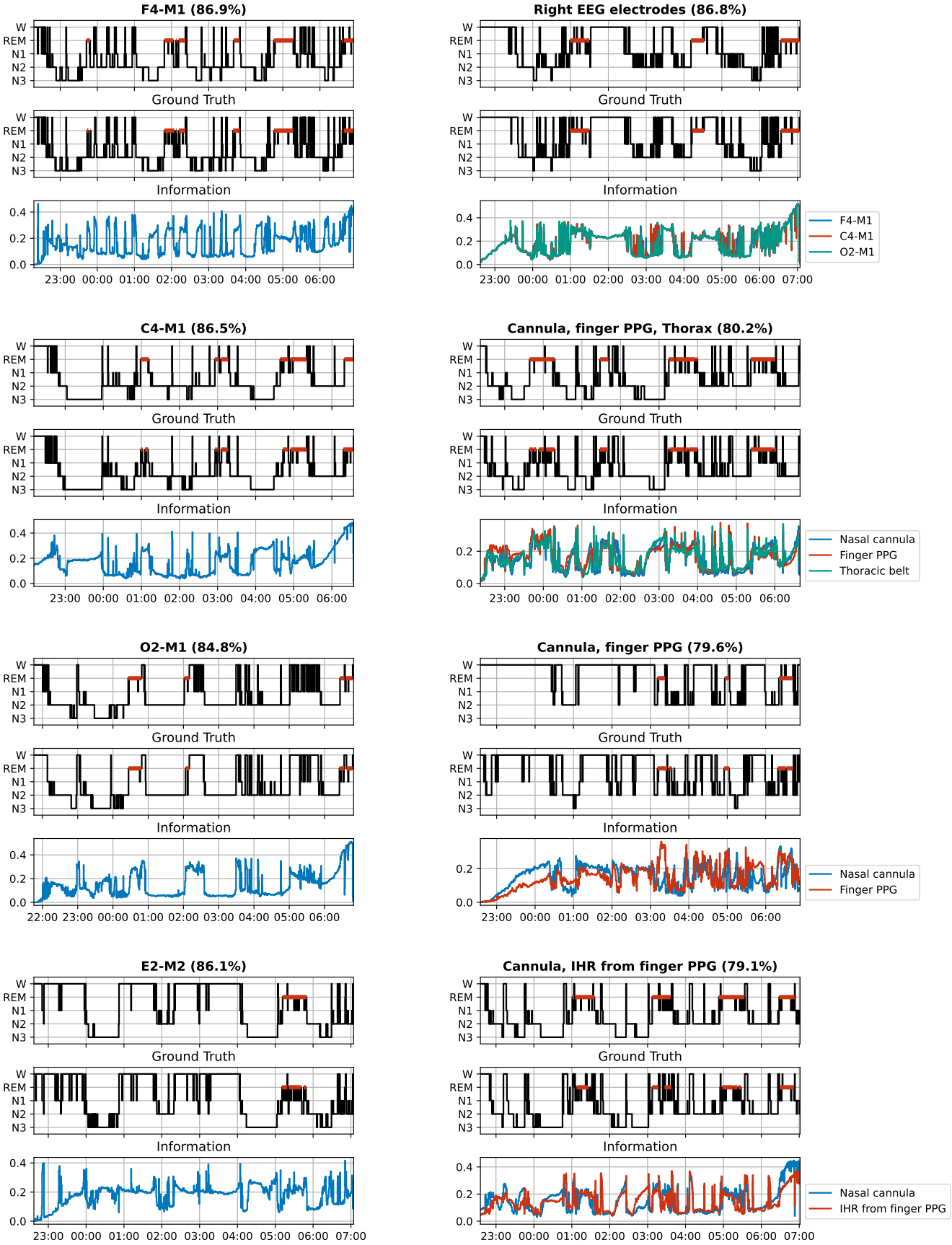


Fig. S.2. Some additional qualitative examples.

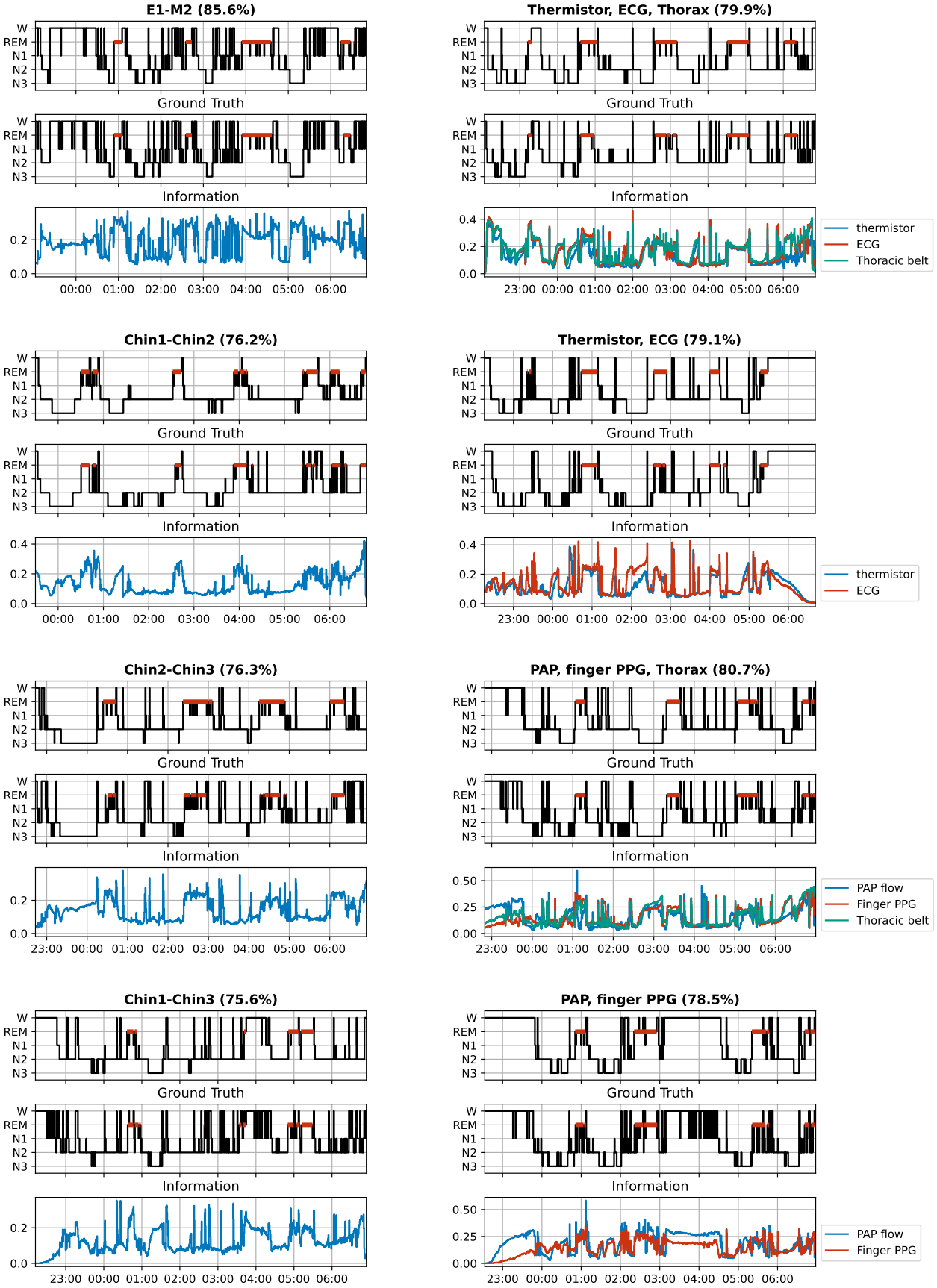


Fig. S.3. Some additional qualitative examples.

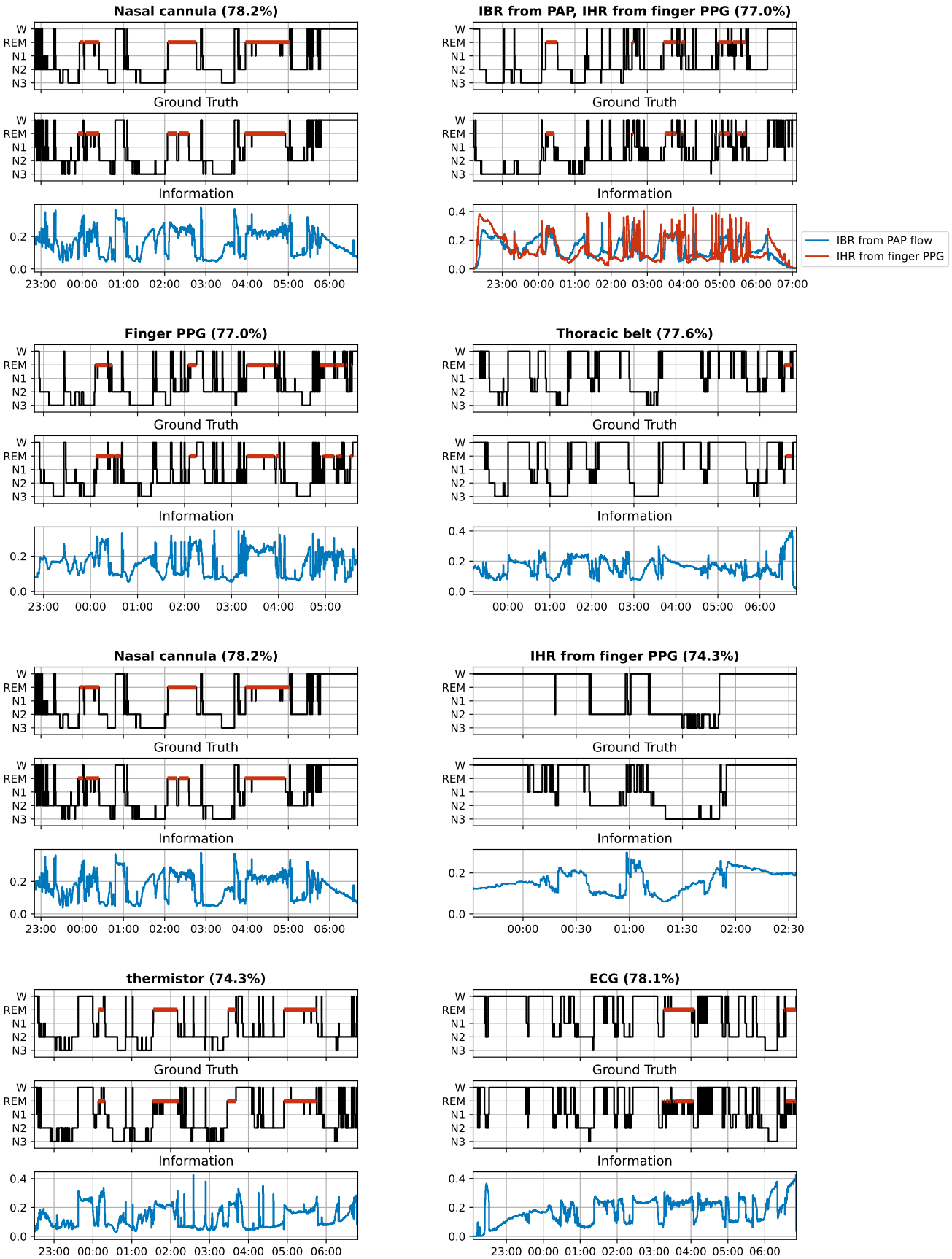


Fig. S.4. Some additional qualitative examples.

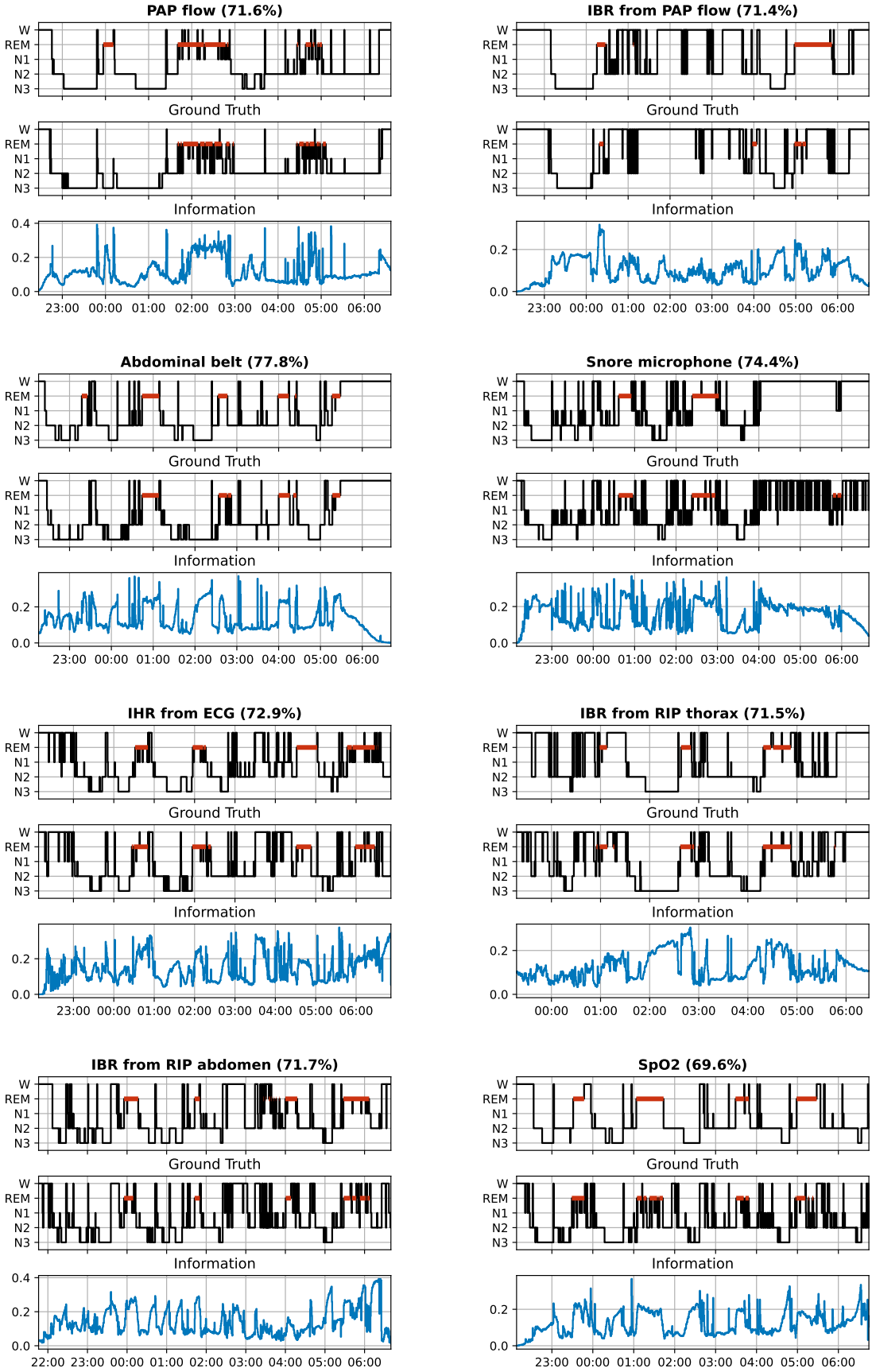


Fig. S.5. Some additional qualitative examples.

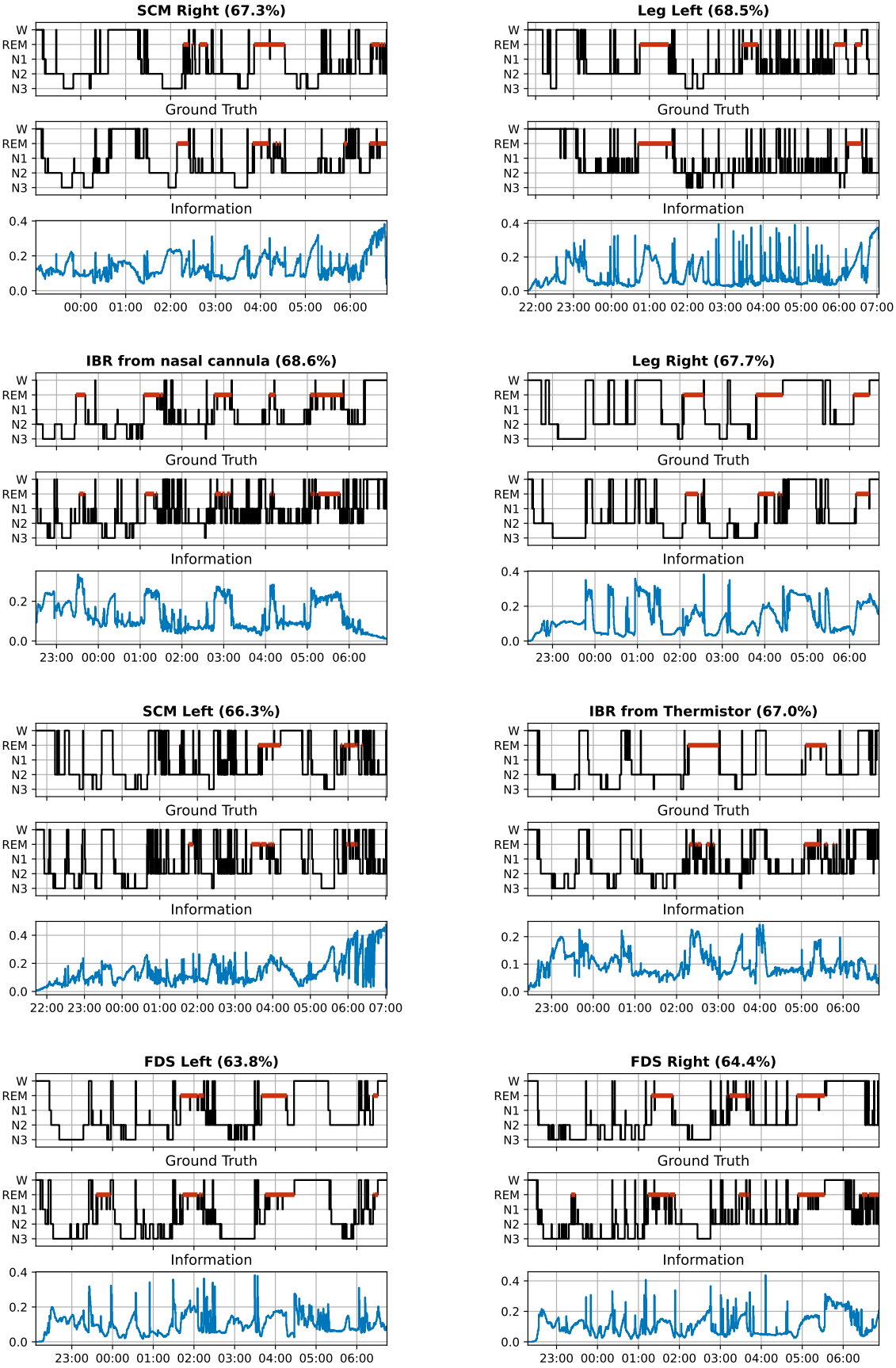


Fig. S.6. Some additional qualitative examples.

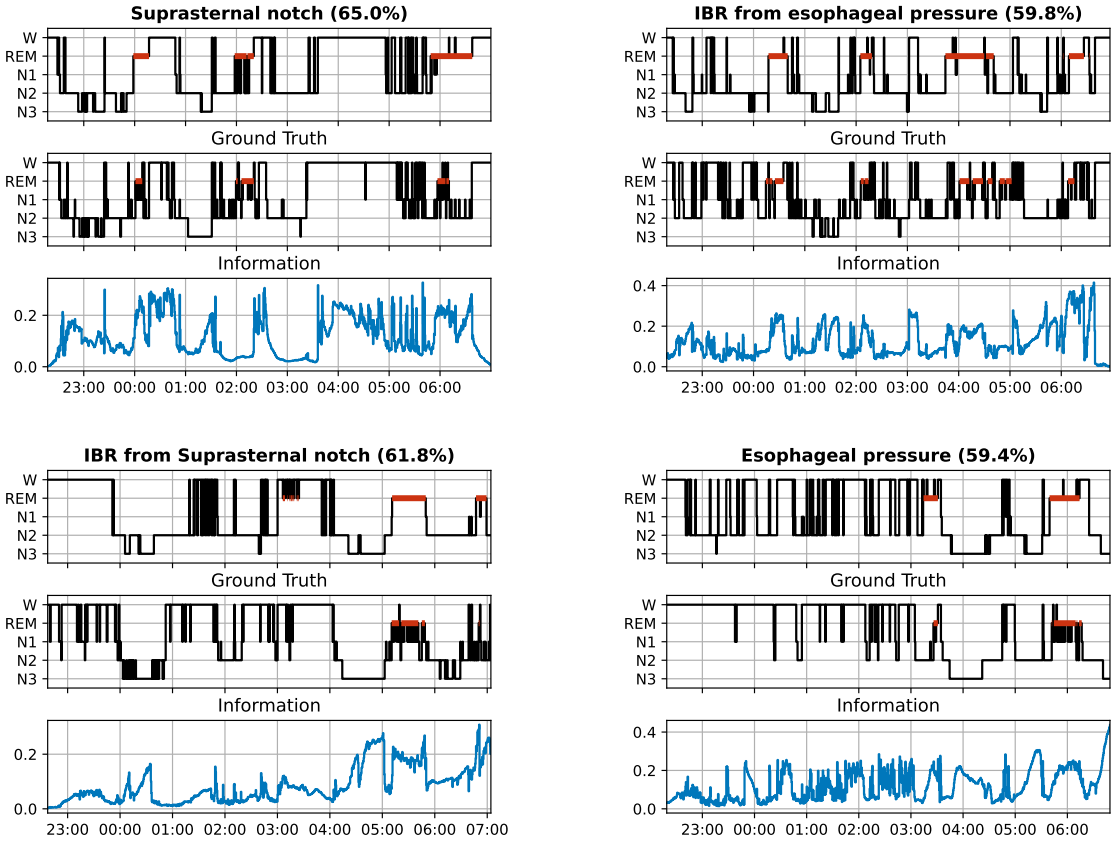


Fig. S.7. Some additional qualitative examples.

C. SAMPLES FROM THE PRIOR

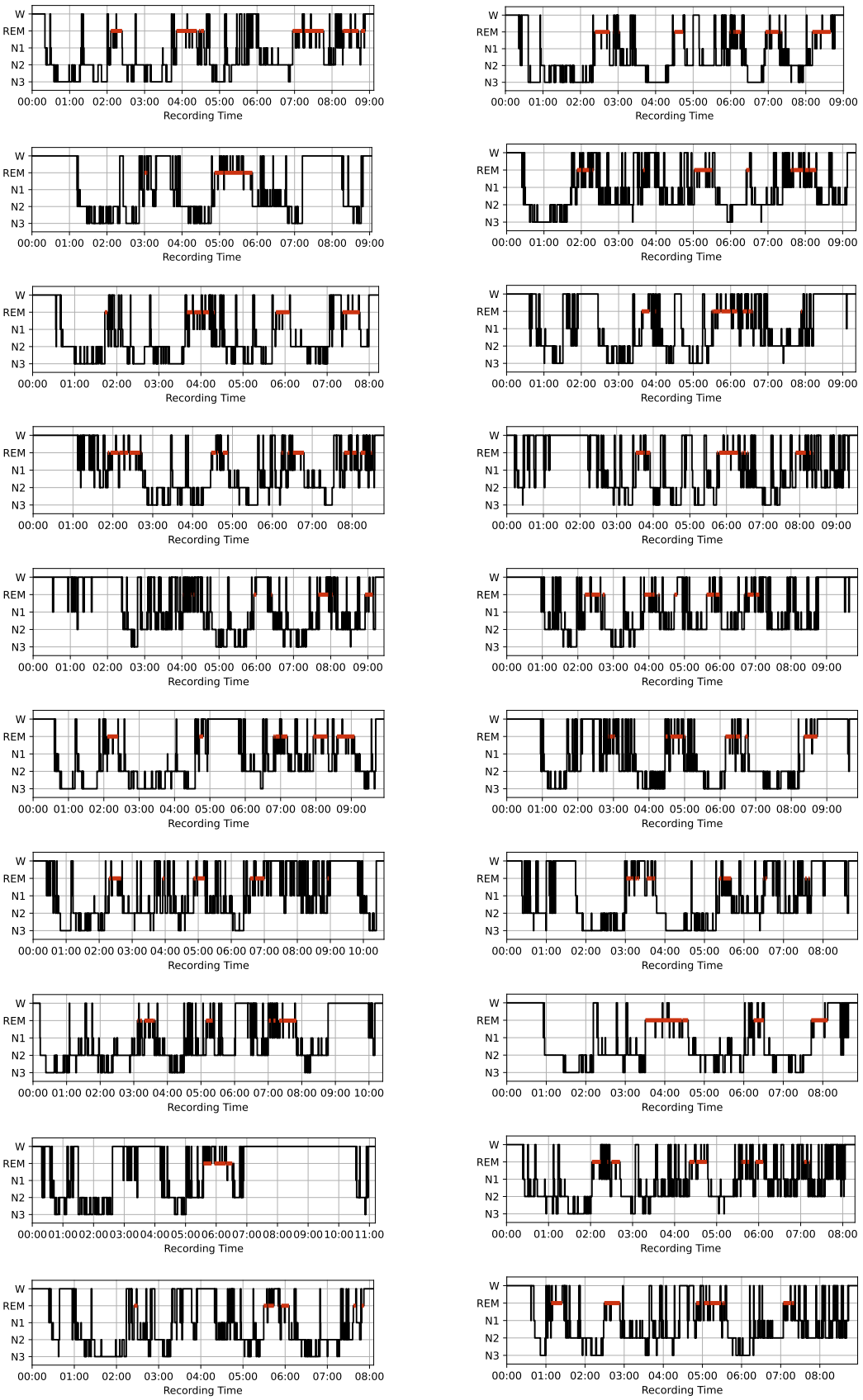
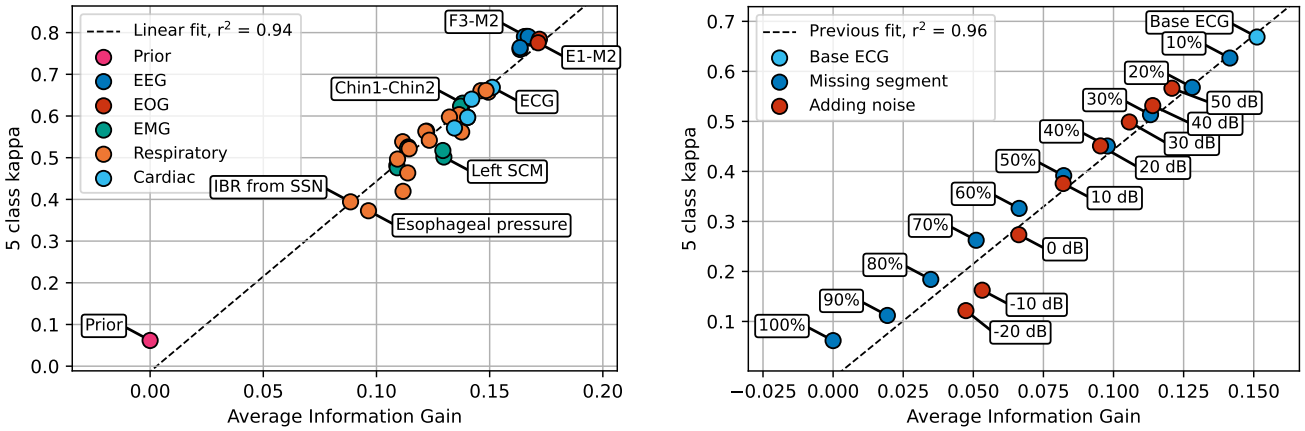


Fig. S.8. We here show samples taken from the global prior score network (without the use of any measurement data).

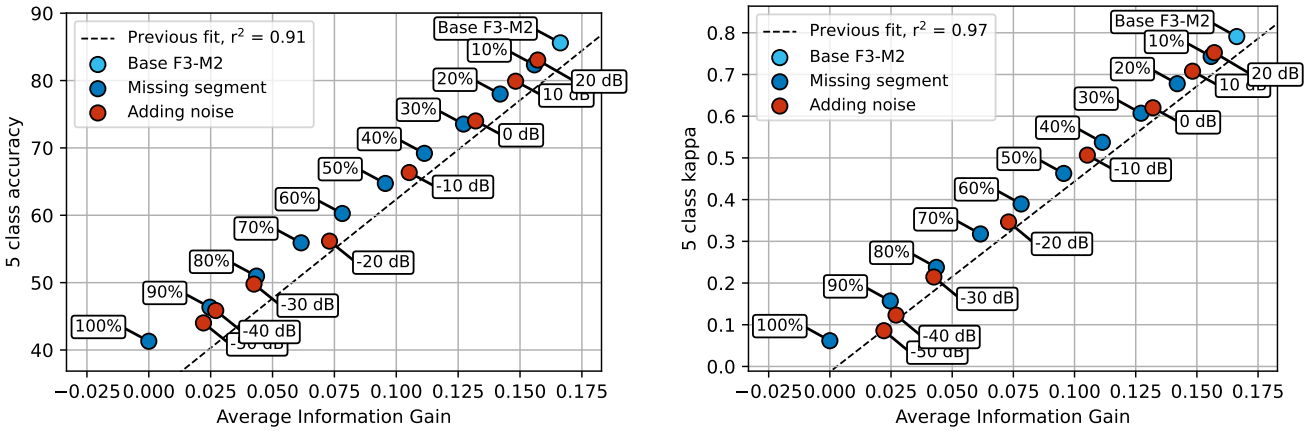
D. Information Gain vs Performance

We here show the same experiment as shown in Fig.7 from the manuscript, but now for Cohen's kappa instead of accuracy. Additionally, we show the results for this experiment for the F3-M2 sensor.



(A) **(B)**

Fig. S.9. Quantitative results for information gain. **(A)** The average information gain per sensor over all test recordings shows a clear linear correlation with respect to Cohen's kappa. **(B)** Reducing the usefulness of the ECG signal by removing segments or adding noise reduces down-stream accuracy and information gain. The linear relationship as fitted on the data from (A) still provides a good fit here.



(A) **(B)**

Fig. S.10. Quantitative results for information gain for the F3-M2 sensor. **(A)** Results of noise and sensor disconnection for the 5-class accuracy and information gain of the F3-M2 sensor. **(B)** The same experiment as (A), but now expressed for Cohen's kappa.

E. DISCONNECTION OF ANOTHER SENSOR

We here show the same experiment as shown in Fig.8 from the manuscript, but now for a nasal cannula that is disconnected halfway through the night.

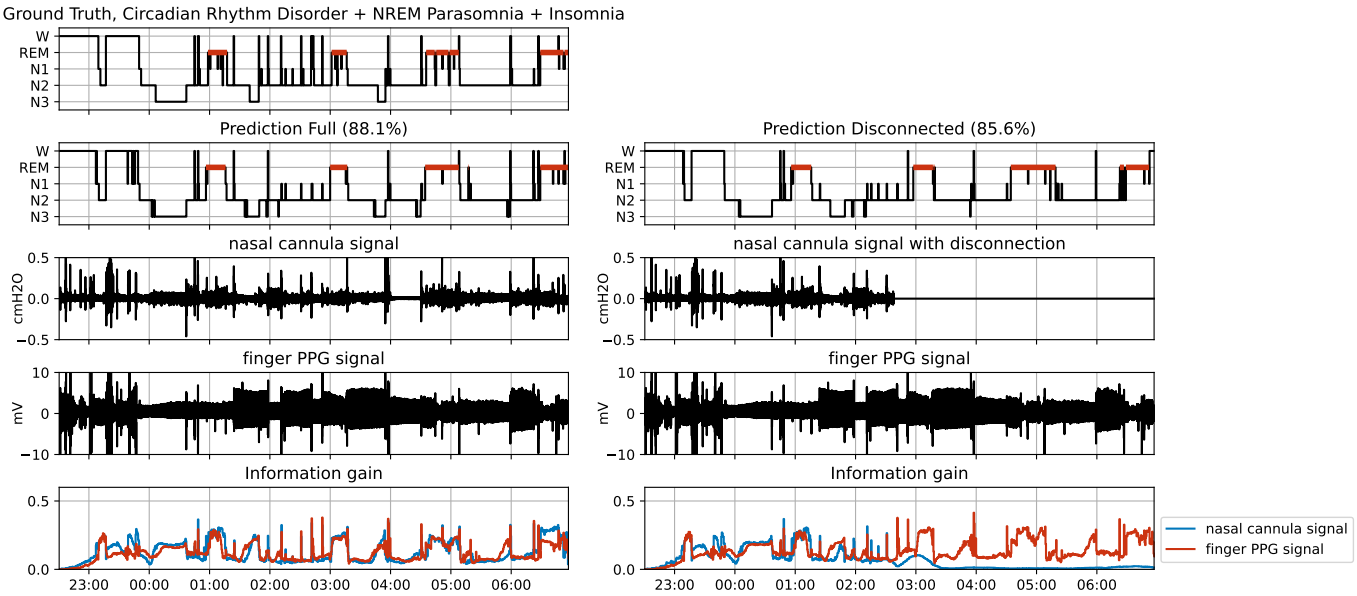


Fig. S.11. Example of disconnecting the nasal cannula signal halfway through the night. Left, output of using the nasal cannula and finger PPG signals over the entire night. Right, artificially created example of what would happen if the user took of the nasal cannula halfway through the night at 3:15. Note that in the original signal there are already two periods of disconnection for the nasal cannula (2:00-2:30 and 4:00-4:45). Additionally, this subject was diagnosed with circadian rhythm disorder, NREM parasomnia, and insomnia.

F. Individual diagnoses and their clusters

TABLE S.IV: Specific diagnoses and how they were clustered. Subjects could have multiple sleep disorders within the same cluster, e.g., a pediatric obstructive sleep apnea diagnosis and an adult obstructive sleep apnea diagnsosis.

Diagnosis	**#**	**Cluster**	**#**
Chronic insomnia disorder	217	Insomnia disorders	613
Psychophysiological insomnia	206		
Idiopathic insomnia	11		
Paradoxical insomnia	23		
Inadequate sleep hygiene	55		
Behavioral insomnia of childhood	0		
Insomnia due to (another) mental disorder	69		
Insomnia due to (a) medical condition	86		
Other insomnia disorder	31		
Obstructive sleep apnea, adult	1037	Obstructive Sleep Apnea	1037
Obstructive sleep apnea, pediatric	4		
Central sleep apnea with Cheyne-Stokes breathing	19	Central Sleep Apnea	42
Central sleep apnea due to a medical disorder without Cheyne-Stokes breathing	9		
Central sleep apnea due to a medication or substance	3		
Primary central sleep apnea	11		
Treatment emergent central sleep apnea	6	Treatment emergent central sleep apnea	6
Obesity hypoventilation syndrome	1	Hypoventilation	8
Idiopathic central alveolar hypoventilation	1		
Sleep related hypoventilation due to a medication or substance	1		
Sleep related hypoventilation due to a medical disorder	0		
Sleep related hypoxemia disorder	5		
Narcolepsy type 1	25	Narcolepsy	31
Narcolepsy type 1 due to a medical condition	1		
Narcolepsy type 2	5		
Narcolepsy type 2 due to a medical condition	1		
Idiopathic hypersomnia	8	Other Hypersomnolence Disorders	54
Idiopathic hypersomnia with normal sleep time	14		
Idiopathic hypersomnia with long sleep time	8		
Kleine-Levin syndrome	1		
Hypersomnia due to a medical disorder	11		
Hypersomnia secondary to Parkinson disease	4		
Residual hypersomnia in OSA patients with adequately treated OSA	2		
Hypersomnia associated with a psychiatric disorder	7		
Hypersomnia associated with mood disorder	1		
Hypersomnia associated with a conversion disorder or somatic symptom disorder	1		
Insufficient sleep syndrome	66	Insufficient sleep syndrome	66

TABLE S.IV: Specific diagnoses and how they were clustered. Subjects could have multiple sleep disorders within the same cluster, e.g., a pediatric obstructive sleep apnea diagnosis and an adult obstructive sleep apnea diagnsosis.

Diagnosis	**#**	**Cluster**	**#**
Delayed sleep-wake phase disorder	33	Circadian rhythm disorders	46
Advanced sleep-wake phase disorder	2		
Irregular sleep-wake rhythm disorder	1		
Shift work disorder	9		
Circadian sleep-wake disorder NOS	1		
Confusional arousals	74	NREM parasomnias	115
Sleepwalking	48		
Sleep terrors	24		
Sleep related abnormal sexual behaviors	2		
Sleep related eating disorder	1		
REM sleep behavior disorder	122	RBD	122
Recurrent isolated sleep paralysis	11	REM parasomnias other than RBD	55
Nightmare disorder	39		
Sleep related hallucinations	12		
Parasomnia overlap disorder	7	Other parasomnias	45
Exploding head syndrome	1		
Sleep enuresis	2		
Parasomnia due to a medical disorder	4		
Parasomnia, unspecified	31		
Restless legs syndrome	185	RLS/PLMD	268
Periodic limb movement disorder	114		
Sleep related leg cramps	2	Other movement disorders	58
Sleep related bruxism	27		
Sleep related rhythmic movement disorder	5		
Propriospinal myoclonus at sleep onset	2		
Sleep related movement disorder, unspecified	18		
Sleep starts (hypnic jerks)	5		
Other sleep disorder	5	Other	16
Sleep related epilepsy	2		
Sleep related headache	1		
Sleep related laryngospasm	4		
Sleep related gastroesophageal reflux	3		
Sleep disorder due to sedative, hypnotic or anxiolytic	1		
No primary sleep diagnosis	45	No primary sleep diagnosis and/or normal variants	99
Short sleeper	2		
Snoring	31		
Catathrenia	7		
Long sleeper	14		

TABLE S.V
PERFORMANCE OF THE 'RECOMMENDED PSG' SETUP ON THE DIFFERENT DIAGNOSTIC GROUPS. DIFFERENCE WERE TESTED WITH A MANN-WHITNEY U RANK TEST AT SIGNIFICANCE LEVEL P = 0.05, COMPARING EACH GROUP WITH RESPECT TO ALL OTHER RECORDINGS.

Diagnostic Group	#	5-class Accuracy	5-class Kappa	F1 Wake	F1 N1	F1 N2	F1 N3	F1 REM
Insomnia disorders	166	87.4	0.818	0.910	0.606	0.892	0.847	0.881
Obstructive sleep apnea	278	85.4	† 0.786	0.887	0.596	0.877	0.830	0.877
Central sleep apnea	14	85.5	0.781	0.911	0.565	0.874	0.832	0.869
Hypoventilation	2	79.5	0.716	0.903	0.435	0.792	0.759	0.862
Narcolepsy	10	85.5	0.789	0.823	0.648	0.891	0.836	0.875
Other hypersomnolence disorders	11	88.4	0.831	0.881	0.636	0.922	0.904	0.859
Insufficient sleep syndrome	11	88.2	0.833	0.903	0.655	0.898	0.907	0.896
Circadian rythm disorder	8	86.3	0.801	0.840	0.621	0.898	0.824	0.828
NREM Parasomnias	24	88.1	0.830	0.896	0.651	0.898	0.888	0.900
RBD	30	81.7	† 0.741	0.865	0.513	0.847	0.799	0.831
REM Parasomnias other than RBD	6	85.6	0.799	0.906	0.660	0.877	0.884	0.880
Other Parasomnia	11	84.8	0.702	0.827	0.602	0.880	0.848	0.890
RLS/PLMD	59	87.1	0.798	0.912	0.591	0.888	0.855	0.885
Other movement disorders	18	86.8	0.811	0.873	0.590	0.898	0.835	0.904
Other sleep disorders	2	89.1	0.424	0.931	0.615	0.889	0.939	0.916
No primary sleep diagnosis and/or normal variants	18	87.6	0.821	0.891	0.602	0.902	0.870	0.876
Healthy	27	88.8	0.838	0.866	0.659	0.901	0.898	0.929

† Performance is significantly different from the rest.

TABLE S.VI
PERFORMANCE OF AN HSAT SETUP (CANNULA FLOW, THORACIC BELT, FINGER PPG) ON THE DIFFERENT DIAGNOSTIC GROUPS. DIFFERENCE WERE TESTED WITH A MANN-WHITNEY U RANK TEST AT SIGNIFICANCE LEVEL P = 0.05, COMPARING EACH GROUP WITH RESPECT TO ALL OTHER RECORDINGS.

diagnostic group	#	5-class Accuracy	5-class Kappa	F1 Wake	F1 N1	F1 N2	F1 N3	F1 REM
Insomnia disorders	149	80.2	0.715	0.864	0.424	0.816	0.734	0.850
Obstructive sleep apnea	219	78.0	† 0.683	0.834	0.444	0.801	0.712	0.838
Central sleep apnea	7	71.6	0.575	0.742	0.433	0.770	0.568	0.835
Hypoventilation	2	65.7	0.521	0.815	0.172	0.594	0.544	0.845
Narcolepsy	10	77.3	0.672	0.730	0.465	0.826	0.669	0.847
Other hypersomnolence disorders	10	83.0	0.750	0.821	0.513	0.869	0.832	0.843
Insufficient sleep syndrome	11	80.6	0.728	0.823	0.493	0.825	0.803	0.861
Circadian rythm disorder	8	78.3	0.679	0.764	0.352	0.831	0.720	0.843
NREM Parasomnias	24	79.9	0.711	0.833	0.502	0.825	0.727	0.876
RBD	30	72.6	† 0.609	0.807	0.362	0.755	0.624	0.747
REM Parasomnias other than RBD	6	79.0	0.703	0.849	0.508	0.820	0.787	0.799
Other Parasomnia	11	79.2	0.696	0.804	0.444	0.802	0.693	0.818
RLS/PLMD	56	80.2	0.712	0.857	0.409	0.818	0.706	0.843
Other movement disorders	18	80.3	0.714	0.821	0.391	0.828	0.733	0.864
Other sleep disorders	3	89.3	0.850	0.927	0.548	0.859	0.812	0.872
No primary sleep diagnosis and/or normal variants	17	81.5	0.733	0.849	0.440	0.841	0.801	0.853
Healthy	27	81.3	0.730	0.795	0.472	0.835	0.782	0.902

† Performance is significantly different from the rest.

G. PERFORMANCE PER DIAGNOSIS

To test whether the underlying sleep disorder impacts sleep staging results, we calculated the performance of two signal combinations for each group separately. We used the groups as indicated in Table S.IV in Supplement F. The results for the recommended PSG setup are shown in Table S.V, while the results for an HSAT (cannula flow, thoracic belt, finger PPG) are shown in Table S.VI.

To test whether differences in performance were statistically significant, we used a Mann-Whitney U rank test. We compared the 5-class Kappa of subjects with a certain sleep disorder to those without that sleep disorder. We used a significance level of p = 0.05 and applied a Bonferroni correction for repeated testing. Using this setup, two significant differences were found in both sensor setups. The performance of the OSA and RBD groups was found to be significantly different from the rest. All other diagnostic groups did not show significant differences.

The lower performance for the OSA group is probably due to the increased sleep fragmentation found in this group. The differences in performance for the RBD group are probably due to the large differences in sleep structure of these subjects. By definition, these subjects display REM sleep without atonia (RSWA), which has completely different characteristics from the REM sleep found in all other diagnostic groups. Additionally, this group tends to contain older subjects, whose sleep, especially stage N3, is also more difficult to score [47].

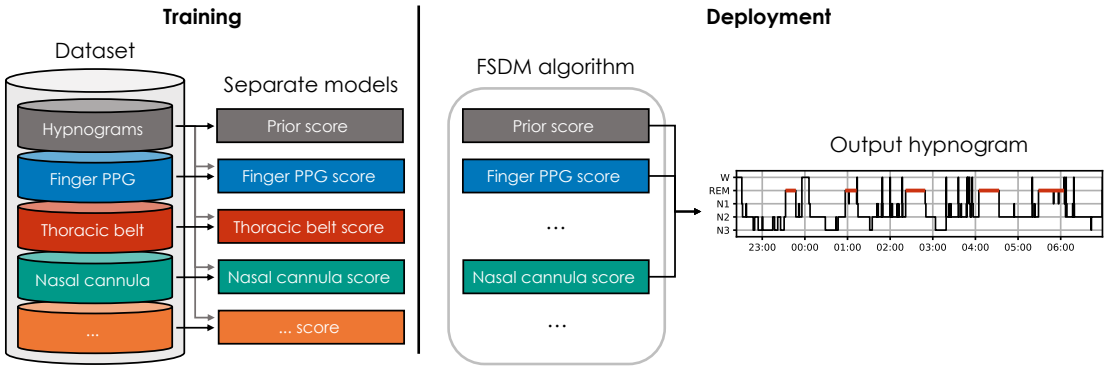


Fig. S.12. Visual overview of the proposed FSDM pipeline. During training, each signal-specific score-network is trained on the subset of data where its sensor was used, including the ground-truth hypnogram. During deployment, any combination of signal modalities can occur. Each signal that was present in the measurement is used with its specific score-network. The proposed FSDM algorithm then fuses the results with a prior score to obtain a posterior using equation (1), from which the hypnogram is sampled.

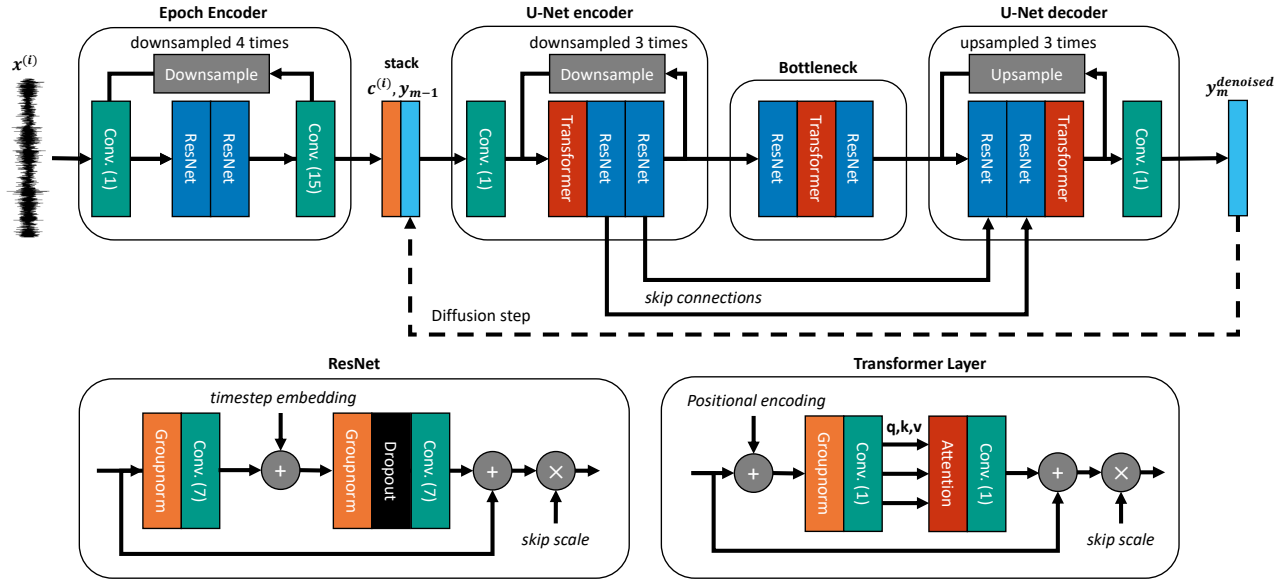


Fig. S.13. Overview of a the neural network used for each denoiser $\boldsymbol{D}_{\boldsymbol{\theta}^{(i)}}\left(\boldsymbol{y_m}, \boldsymbol{x}^{(i)}, \boldsymbol{\sigma}(\boldsymbol{t_m})\right)$. The signal, $\boldsymbol{x}^{(i)}$, is used as the inital input to the network and encoded into a context vector. This is then stacked with the sample at the current timestep $\boldsymbol{y_m}$ and fed though a U-Net structure. At the end, a hypnodensity is given as output through the use of a softmax activation function. The current noise variance, $\boldsymbol{\sigma}(\boldsymbol{t_m})$, is additionally embedded into a timestep embedding, which is added inside the ResNet layers of the U-Net encoder and U-Net decoder. To avoid needing to run the Epoch Encoder $\boldsymbol{M}$ times, the timestep embedding is not added to its ResNet layers.

H. Details regarding the network architecture

We leveraged the DDPM++ model as implemented by Karras *et al.* [13], and modified to work on 1D timeseries in our previous work on EOG-driven sleep staging [27]. See Fig. S.13 for an overview. The neural network architecture used in this work differs from [27] in two aspects. Firstly, there is an additional input between the epoch encoder and U-Net encoder in order to add the $\boldsymbol{y}_{m-1}$ of the previous output (in order to solve the ODE). Secondly, there is an additional embedding in ResNets to add the current diffusion timestep, a common practice in score-based diffusion models [13]. We will now discuss each neural network component.

A. Epoch encoder

Because the signals and the hypnograms had different sampling frequencies (128 Hz vs. 1/30 Hz), we first needed to downsample the input signal before we could use the U-Net structure of our model. To that end, we employed a context encoder, which downsampled the signals from $\mathbb{R}^{1792\cdot30\cdot128\times1}$ to $\mathbb{R}^{1792\times16}$, i.e. a context encoding of length number of epochs with 16 channels.

The context encoder worked as follows. First, a convolution of kernel size 1 expanded the number of channels from 1 to 16. Then, a series of two ResNets was employed to extract meaningful features from the input signal (see the ResNet section for further details). This pattern was repeated 5 times with 4 downsampling operations between the 5 blocks. Each downsampling operation used a kernel of [1,1,1,1] and a stride of 4, to effectively downsample the input by a factor of 4. At the end of the epoch encoder, another convolution of kernel and stride 15 was used, thus compressing the signals to a feature map of size $\mathbb{R}^{1792\times16}$ which was used as input to the U-Net encoder.

B. Stacking

After the epoch encoder, the feature map is concatenated channel wise with the previous estimate of the diffusion step. Following [13], we apply input scaling to the previous estimate of $\boldsymbol{y}_m$ as:

$$\tilde{\boldsymbol{y}}_m = \frac{1}{\sqrt{\sigma_{data}^2 + \sigma_t^2}} \cdot \boldsymbol{y}_m, \tag{23}$$

Where σ_{data} was estimated from data as $\sigma_{data} = 0.3160$ and σ_t is the current variance of the diffusion ODE. The stacking then results in an input of $\mathbb{R}^{1792\times(16+5)} = \mathbb{R}^{1792\times21}$ for the U-Net encoder.

C. Note on prior networks

When we are using the network as a prior network, no input conditioning data is used. Thus, we skip the epoch encoder and the stacking operation, and we only input the previous diffused hypnogram $\tilde{\boldsymbol{y}}_m$ into the rest of the network.

D. U-Net encoder

The U-Net encoder first employed a convolution of kernel size 1 to increase the channel size from 21 to 32. Then, a Transformer layer together with two ResNet blocks was employed (see the Transformer layer section for further details). After each ResNet block, a skip connection was added to the U-Net decoder at the same resolution. This pattern of a transformer with two ResNets was repeated 4 times with 3 downsampling operations in between. Again, a kernel of [1,1,1,1] and a stride of 4 was used in the downsampling operations. The number of channels was left the same throughout the network, at a fixed 32 channels. Note that in the original DDPM++ implementation [12], an attention layer was added after each ResNet in the encoder. However, to bring down the computational complexity of our method and to make the encoder symmetric with the decoder, we employed only a single transformer layer at the start of each resolution level in the U-Net encoder.

E. Bottleneck

In the bottleneck, the feature map was of its smallest size, namely $\mathbb{R}^{28\times32}$. Here, one transformer layer sandwiched between two ResNet blocks was used to learn the highest-level features of the hypnogram.

F. U-Net decoder

The decoder followed a mirrored structure to the encoder. The skip connections from the corresponding resolution levels were concatenated to the inputs of each ResNet block. These connections allowed the feature maps to skip the downward path of the 'U' and enabled the model to learn both high-and-low level features of the hypnogram. The upsampling operation of the decoder was implemented using a transposed convolution with the same filter of [1,1,1,1].

As a final step toward creating the $\boldsymbol{y}_m^{denoised}$, the U-Net decoder employed a convolution of kernel size 1 to map the input to 5 channels, where each channel corresponded to one of the five sleeps stages. A softmax activation function was then used to map each channel to a class probability. This creates a 'hypnodensity', a soft version of the hypnogram where each epoch is partially associated with each sleep stage according to some probability [48].

G. ResNet

The ResNet, or Residual Network, was repeated throughout the architecture. It consists of two group normalization layers and two convolutions in an alternating pattern. Group normalization, as described by [49], applies a learned normalization across groups of channels, enabling faster training. In our case, each group consisted of 4 channels. The 1D convolutions of the ResNet each used a kernel of size 7 and zero-padding set to 'same'. Each convolution was followed by SiLU (Sigmoid Linear Unit) activation [50]. Additionally, a spatial dropout layer was added before the second convolution, which drops out entire channels during training with a probability of 10%. Spatial dropout is a better regularizer for convolutional neural networks, since neighbouring samples are often highly correlated [51]. Finally, a residual connection was added to help combat vanishing gradient problems. To limit the magnitude of the signals, scaling with a factor of $skip\ scale = \sqrt{0.5}$ was applied.

In the case that the ResNet was part of the epoch encoder, the additive timestep embedding was equal to zero, and we effectively do not add it. Otherwise. an additive timestep embedding is generated from the current noise level of the diffusion process to tell it about what kind of noise level to expect in $\boldsymbol{y}_{m-1}$, which has been found to be helpful in the score-based diffusion literature. This additive timestep embedding is explained in the sequel.

H. Timestep embedding

Following [13], the current noise level of the diffusion process, σ_t, is given as an additional input to the network in a scaled and embedded form. To that end, it is first scaled as follows:

$$\tilde{\sigma}_t = 0.25 \log(\sigma_t). \tag{24}$$

After this scaling, a sine-cosine embedding scheme was used that embedded the noise level as follows:

$$\boldsymbol{c} = [0, 1, \ldots, C/2 - 1]^T, \tag{25}$$

$$\boldsymbol{f} = 1000\hat{}(-\boldsymbol{c}/(C/2 - 1)), \tag{26}$$

$$\boldsymbol{z} = [\cos(\tilde{\sigma}_t \cdot \boldsymbol{f}),\ \sin(\tilde{\sigma}_t \cdot \boldsymbol{f})]^T, \tag{27}$$

where C is the number of channels in the ResNet, which is equal to 32 throughout the U-Net. Subsequently, a multilayer perceptron (MLP) was applied of 2 layers, with 8 hidden nodes each and SiLU activation. Then, for each ResNet separately, a local linear layer was applied to increase the noise level embedding to 32 again. As a final step it is broadcasted into the input length of the feature map at the ResNet and added to it element-wise.

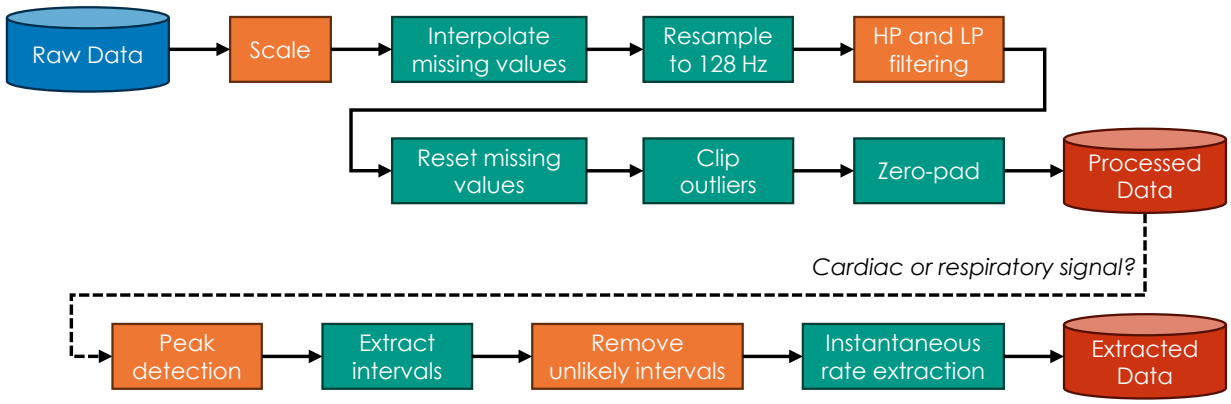


Fig. S.14. The preprocessing pipeline is the same for all the signals. The orange blocks are set by signal-type specific parameters, e.g. the cutoff frequencies of the filters can be different for different signal types, see Table III. If a signal is a cardiac or respiratory signal we also extract the instantaneous heart-rate or breathing-rate.

I. Transformer

The original transformer architecture is a sequence-to-sequence model composed of both an encoder and a decoder [52]. Where each element consists of a scaled dot-product attention layer and an element-wise feed-forward network. Additionally, positional encoding is added at the start of the encoding and decoding stacks. We adapt the transformer architecture to be suited for our network. Firstly, we did not use the decoder, since it is used to generate new sequence in an auto-regressive manner. Secondly, since we embedded the layers within a larger convolutional neural network, there was no need for separate element-wise feed-forward networks. lastly, because the attention layers operated at different time scales, we added positional encoding to each of them.

The positional encoding was also implemented using sine-cosine embedding. The encoding scheme used is similar to the timestep embedding, with some differences. Namely, we here create a full matrix embedding instead of only a vector embedding, no MLP is applied, and the sine and cosine terms are interleaved.

In the transformer layer positional encoding scheme, a positional encoding matrix is added element-wise to the input sequence of the transformer. To that end, the input sequence $\mathbf{S}$ and positional encoding matrix $\mathbf{P}$ should be of the same size: $\mathbf{S}, \mathbf{P} \in \mathbb{R}^{L \times C}$, where L is the length of the input sequence and C is the number of channels (32 in our case). The positional encoding matrix for the transformer layers is given by:

$$\begin{aligned} \mathbf{P}_{(l,2c)} &= \sin\left(l \cdot 1000^{-2c/C}\right) \\ \mathbf{P}_{(l,2c+1)} &= \cos\left(l \cdot 1000^{-2c/C}\right), \end{aligned} \tag{28}$$

with $l \in [0, 1, \ldots, L-1]$ and $c \in [0, 1, \ldots, C/2-1]$. This type of encoding enables the transformer to exploit information about both the absolute and relative positions of samples along the night.

Each of the transformer layers used scaled dot-product self-attention. While the attention mechanism can be implemented using multiple attention-heads for added complexity, we here only made use of a single head. In scaled dot-product self-attention, three linear projections are applied to transform the sequence to a query, key, and value matrix:

$$\mathbf{Q} = \mathbf{S}\mathbf{W}_Q, \ \mathbf{K} = \mathbf{S}\mathbf{W}_K, \ \mathbf{V} = \mathbf{S}\mathbf{W}_V, \tag{29}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{C \times C}$ are learned linear projection weights and $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{L \times C}$ are the query, key, and value matrices, respectively. These linear projections can be implemented efficiently by a single convolutional layer of kernel size 1 and output channel size of $3C$, as its output can be split along the channel dimension into the three separate components.

Following a database analogy, the queries are going to look for matching keys and propagate the associated values to the output, where each individual query, key, and value are found along the rows of their respective matrices. This process is defined by the scaled dot-product self-attention mechanism:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{C}}\right)\mathbf{V}, \tag{30}$$

where $\mathbf{K}^T$ denotes the transpose of the key matrix. Moreover, $\mathbf{Q}\mathbf{K}^T \in \mathbb{R}^{L \times L}$ denotes the attention map. To ensure that the magnitudes in the attention map do not grow too large, it is scaled down by a factor of $1/\sqrt{C}$. Additionally, a softmax activation is applied along the rows of the attention map in order to ensure that the attention sums to 1.

After the scaled dot-product attention layer, another linear projection using a 1D convolution was applied. Similar to the ResNet, a residual connection was applied with a scaling of $skip\ scale = \sqrt{0.5}$.

I. Preprocessing

We applied a common preprocessing pipeline to all signals as shown in Fig. S.14 and Table S.VII. We will briefly describe each preprocessing operation. First, each of the signals is scaled by a constant value in order to bring its approximate magnitude around 1. The scaling factor is chosen specific to each signal type and is shown in Table III. We for example scale all the EEG channels by a factor 10^4, making it so that an amplitude of 100μV corresponds to the value 1 and the slow-wave amplitude threshold of 75μV corresponds to a value of 0.75 [1]. This scaling enables faster training of the neural networks.

TABLE S.VII
OVERVIEW OF THE SIGNALS EXTRACTED FROM THE DATASETS. WE CLUSTERED THE SIGNALS INTO GROUPS SUCH AS EEG AND RIP BELTS. 'HP' AND 'LP' DENOTE THE HIGH-PASS AND LOW-PASS FILTER RESPECTIVELY WHERE WE SHOW THE CUT-OFF FREQUENCY IN HZ. 'FDS' AND 'SCM' REFER TO THE FLEXOR DIGITORUM SUPERFICIALIS AND STERNOCLEIDOMASTOID MUSCLES, RESPECTIVELY.

Signal group	Signals	Unit	Scale	HP	LP	Total	Train	Val	Test
EEG	F3-M2, F4-M1, C3-M3, C4-M1, O1-M2, O2-M1	V	10^4	0.3	49	11681	8081	600	3000
EOG	E1-M2, E2-M2	V	10^4	0.3	49	3886	2689	200	997
EMG chin	Chin1-Chin2, Chin1-Chin3, Chin2-Chin3	V	10^4	10	49	5838	4038	300	1500
ECG	ECG	V	10^3	0.3	49	1947	1347	100	500
RIP belts	Abdomen, Thorax	V	10^{-2}	0.1	15	3892	2692	200	1000
Thermistor	Thermistor	V	10^4	0.1	15	1706	1184	88	434
Nasal cannula	Nasal cannula	cmH2O	1	0.03	49	1706	1184	88	434
PAP flow	PAP flow	cmH2O	10	0.03	49	241	163	12	66
Suprasternal notch	Suprasternal notch	V	10	0.03	49	289	199	18	72
Esophageal pressure	Esophageal pressure	mmHg	10^{-1}	0.03	49	97	65	8	24
Snore microphone	snore microphone	V	10^3	10	49	1947	1347	100	500
Finger PPG	Finger PPG	V	10^{-2}	0.3	49	1976	1364	101	511
SpO2	SpO2	%	10^{-2}	-	-	1944	1344	100	500
EMG FDS	FDS L, FDS R	V	10^4	10	49	508	368	20	120
EMG legs	Leg L, Leg R	V	10^4	10	49	3894	2694	200	1000
EMG SCM	SCM L, SCM R	V	10^4	10	49	296	206	24	66
Instantaneous heart rate	ECG, PPG	Bpm	1/60	-	-	3923	2711	201	1011
Instantaneous breath rate	RIP Belts, Thermistor, Nasal cannula, PAP flow, Suprasternal notch, Esophageal pressure	Brpm	1/60	-	-	8163	5642	426	2095

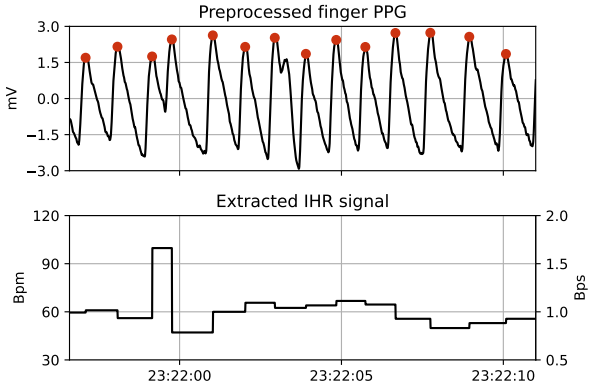


Fig. S.15. Example of IHR extraction from a finger PPG signal. The red dots denote the peaks found by the automatic multi-scale peak detection algorithm.

Second, we identify missing values in the signals as those locations where they are exactly equal to 0. We linearly interpolate these values by using neighbouring sample values. This interpolation is performed in order to reduce filtering artifacts that can occur by the large jumps in magnitude that these missing values cause. After all the filtering operations, the samples with the missing values are reset to exact zeroes, this enables the neural networks to understand which values are missing in the signal.

Third, we resample all signals to 128 Hz using a polyphase filter, except for the SpO2 signal which was up-sampled from 32 Hz to 128 Hz using a sample and hold scheme. After resampling, we apply a low-pass and a high-pass filter, both implemented using fifth order butterworth filter. The high-pass and low-pass filter settings are signal type specific and reflect the recommendations of the AASM manual [1] with some minor adjustments to the low-pass filter settings.

After filtering we reset the samples values of the missing value indices back to zero. We then clip the signals between -5 and 5. For example, the EEG signals are clipped between -500μV and +500μV. Lastly, we zero-pad all the signals to a common length of 1792 sleep epochs, or 1792·30·128 = 6,881,280 samples. This zero-padding was solely done for implementation purposes, as it allows to stack the signals of different nights, and the zero-padded segment was not used to calculate any of the overnight statistics or performance metrics such as accuracy and Cohen's kappa.

In the case that the extracted signal was a cardiac or respiratory signal, such as ECG, finger PPG, or RIP Belt, we also extracted the instantaneous heart-rate (IHR) or the instantaneous breath-rate (IBR). This extraction was performed by extracting the peaks of the cardiac pulses and the breaths using the automatic multi-scale peak detection algorithm [46]. We used the following settings: window length = 600 seconds, window overlap = 120 seconds, and maximum scale = 5 seconds for cardiac signals, while maximum scale = 60 seconds for respiratory signals. After peak detection, we convert the peaks to the inter-beat and inter-breath interval length using a sample and hold scheme. To reduce the impact of artifacts, biologically implausible intervals are removed from the sequence. We remove all cardiac inter-beat intervals outside the range of [0.3s - 2s], and we remove all respiratory inter-breath intervals outside the range of [1s - 30s]. As a final step, the intervals are converted into the IHR or IBR. While typically expressed in beats per minute (Bpm) or breaths per minute (Brpm), we scale the signals by 1/60 to get the beats/breaths per second, resulting in a better magnitude for use by the neural networks. Fig. S.15 shows an example of how we extract the peaks of a finger PPG signal, which we then convert to inter-beat intervals, to subsequently convert to the IHR.

J. COMPUTATIONAL COST

To analyze the computational cost of sampling from the system we timed the procedure 100 times on an NVIDIA GeForce RTX 3080 TI, to provide some indication of the computational

TABLE S.VIII
SLEEP STAGING RESULTS FOR SOME ADDITIONAL COMBINATIONS OF SENSORS. WE SHOW BOTH THE 5-CLASS (W/N1/N2/N3/REM) AND THE 4-CLASS (W/N1-N2/N3/REM) PERFORMANCE.

Signals	Accuracy 5	Accuracy 4	Kappa 5	Kappa 4
Odd EEG	85.5	89.5	0.789	0.822
ECG	76.9	82.2	0.669	0.711
Chin1-Chin2	74.9	80.9	0.630	0.677
Chin1-Chin3	74.5	80.6	0.624	0.672
Chin2-Chin3	74.9	80.9	0.631	0.678
Odd EEG + ECG	85.5	89.5	0.788	0.822
ECG + Chin1-Chin2	78.8	83.9	0.691	0.732
ECG + Chin1-Chin3	78.6	84.0	0.688	0.735
ECG + Chin2-Chin3	78.7	83.9	0.689	0.732
Odd EEG + ECG + Chin1-Chin2	85.2	89.1	0.783	0.816
Odd EEG + ECG + Chin1-Chin3	85.1	89.1	0.782	0.815
Odd EEG + ECG + Chin2-Chin3	85.1	89.1	0.782	0.816

burden of the system. Both theoretically and empirically, the computational complexity of sampling from our model can be described as:

$$T = (2n + 1) \cdot c, \tag{31}$$

with T the time it takes to sample from the model for a single recording, and c a hardware dependent constant. In the case of our machine with an NVIDIA GeForce RTX 3080 TI, c was found to be equal to 1.6 seconds. This means that for the recommended PSG ($n = 6$), our method takes about 20.8 seconds to come to an output.

K. ADDITIONAL COMBINATIONS OF SENSORS

Table S.VIII presents the sleep staging results for various combinations of sensors, specifically EEG, ECG, and EMG. Since EEG alone already reaches the upper limit of performance, equivalent to human inter-rater agreement, adding additional sensor modalities does not enhance its sleep staging performance. For instance, the combination of EEG and ECG achieves the same performance as EEG alone. Moreover, if an additional sensor introduces confusion, such as chin EMG, the combination with EEG may result in slightly lower performance. However, combining two sensors that have not yet reached the performance limit can be very beneficial. As shown in Table S.VIII, the combination of ECG and chin EMG achieves better sleep staging performance than either sensor alone. This effect was also observed in the combinations tested in the main manuscript, where any combination involving EEG reached the performance limit, while HSAT combinations outperformed their individual constituent signals.

L. RESULTS ON SLEEP-EDF EXPANDED

To test the proposed FSDM model on a different dataset and to compare it to other sleep staging networks proposed in the literature, we applied it to the EDF expanded dataset from 2018 [59], [60]. We used the Sleep Cassette set, which consists of 78 healthy sleepers who underwent 2 consecutive nights of recording. Each recording lasts up to 20 hours, and subjects wore a modified Walkman-like cassette-tape recorder that measured the following signals: EEG Fpz-Cz, EEG Pz-Oz, Horizontal EOG, chin EMG, oro-nasal respiration, and rectal temperature. The EEG and EOG channels were both sampled at 100 Hz. However, the other signals were only sampled at 1 Hz, making them unsuitable for our approach. Scoring was performed by experts following the Rechtschaffen & Kales rules [61]. To make this scoring more similar to the AASM standard, we merged S3 and S4 into N3 and did not consider epochs scored as 'Movement Time' when calculating metrics such as accuracy and Cohen's kappa. Additionally, because the recordings were 20 hours long, they mostly contained Wake. Following the literature, we cropped the recordings to ± 30 minutes from time in bed to focus solely on the sleep part of the recording [32].

We tested two approaches to see how our model would generalize to this unseen dataset. First, we tested direct application, where we directly used the model weights trained on SOMNIA and HealthBed on Sleep-EDF Expanded. Such direct transfer is made more difficult because the channels recorded in Sleep-EDF expanded do not match any of the channels in SOMNIA. For example, both EEG and EOG in SOMNIA are referenced to the mastoid, which is not the case in the Sleep-EDF set. Moreover, the electrode positions of Fpz, Cz, Pz, and Oz are not present in SOMNIA. Additionally, the Sleep-EDF set was scored using the R&K rules. While steps were taken to make the scoring more similar to that of the AASM standard, subtle differences in scoring style probably remain. Nonetheless, we directly applied our learned models and received descent results, as shown in Table S.IX.

Second, we also experimented with training our model from scratch on this new dataset. Following the literature, we applied 10-fold cross-validation, constantly leaving 7-8 subjects out for hold-out testing. We also left one other fold out as a validation set and thus trained on the remaining 8 folds. We trained models on each of the three signals: EEG Fpz-Cz, EEG Pz-Oz, and Horizontal EOG. Results for training from scratch are also shown in Table S.IX, together with results from the literature that followed the same 10-fold setup.

From Table S.IX, it can be observed that our model achieves highly comparable performance to that of other models proposed in the literature. Only XSleepNet2 [54] achieved higher accuracy and kappa compared to our approach. Additionally, training from scratch on Sleep-EDF resulted in better metrics than the direct application approach, probably due to the very different signal acquisition, electrode placement, scoring rules, and recording equipment used in the Sleep-EDF dataset. Interestingly, the horizontal EOG from the Sleep-EDF dataset does not reach the same sleep staging performance as the E1-M2 and E2-M2 electrodes from the SOMNIA set, even when training from scratch. This indicates that the horizontal EOG acquired in the Sleep-EDF dataset has very different characteristics and is not as informative for sleep staging as the new AASM EOG placement [1]. We leave the investigation of the exact reasons for this change to future work.

TABLE S.IX

RESULTS ON THE 'EDF EXPANDED' DATASET. WE SHOW DIRECT APPLICATION OF FSDM MODELS TRAINED ON SOMNIA, AS WELL AS TRAINING FROM SCRATCH USING 10-FOLD CROSS VALIDATION WITH ± 30 MINUTES FROM TIME IN BED. THE RESULTS FROM LITERATURE ARE TAKEN FROM THEIR RESPECTIVE PAPERS AND FOLLOW THE SAME SET-UP AS SCRATCH TRAINING. NOTE THAT THERE IS A MISMATCH BETWEEN SIGNAL DERIVATIONS PRESENT IN SOMNIA AND THOSE IN EDF EXPANDED, SEE †, ‡.

Method	**Signals**	**5-class Accuray [%]**	**5-class Kappa [-]**
Direct Application FSDM †‡	Fpz-Cz + Pz-Oz + Horizontal EOG	81.9	0.746
Direct Application FSDM †	Fpz-Cz + Pz-Oz	81.7	0.741
Direct Application FSDM †	Fpz-Cz	76.1	0.662
Direct Application FSDM †	Pz-Oz	80.7	0.728
Direct Application FSDM ‡	Horizontal EOG	76.4	0.675
Scratch Training FSDM	Fpz-Cz + Pz-Oz + Horizontal EOG	82.9	0.758
Scratch Training FSDM	Fpz-Cz + Pz-Oz	82.7	0.755
Scratch Training FSDM	Fpz-Cz	82.5	0.753
Scratch Training FSDM	Pz-Oz	79.9	0.715
Scratch Training FSDM	Horizontal EOG	79.2	0.704
Catboost [53] (2023)	Fpz-Cz + Pz-Oz + Horizontal EOG	83.0	0.763
XSleepNet2 [54] (2022)	Fpz-Cz + Horizontal EOG	84.0	0.778
XSleepNet2 [54] (2022)	Fpz-Cz	84.0	0.778
LGSleepNet [55] (2023)	Fpz-Cz	82.3	0.75
SleepTransformer [32] (2022)	Fpz-Cz	81.4	0.743
TinySleepNet [56] (2021)	Fpz-Cz	83.1	0.77
DeepSleepNet-Lite [57] (2021)	Fpz-Cz	80.3	0.73
SleepEEGNet [58] (2019)	Fpz-Cz	80.0	0.73
SleepEEGNet [58] (2019)	Pz-Oz	77.6	0.689

† Using the model trained on derivations F3-M2, F4-M1, C3-M2, C4-M1, O1-M2, O2-M1.
‡ Using the model trained on derivations E1-M2, E2-M2.

M. Additional Results For Subjects From Fig. 3

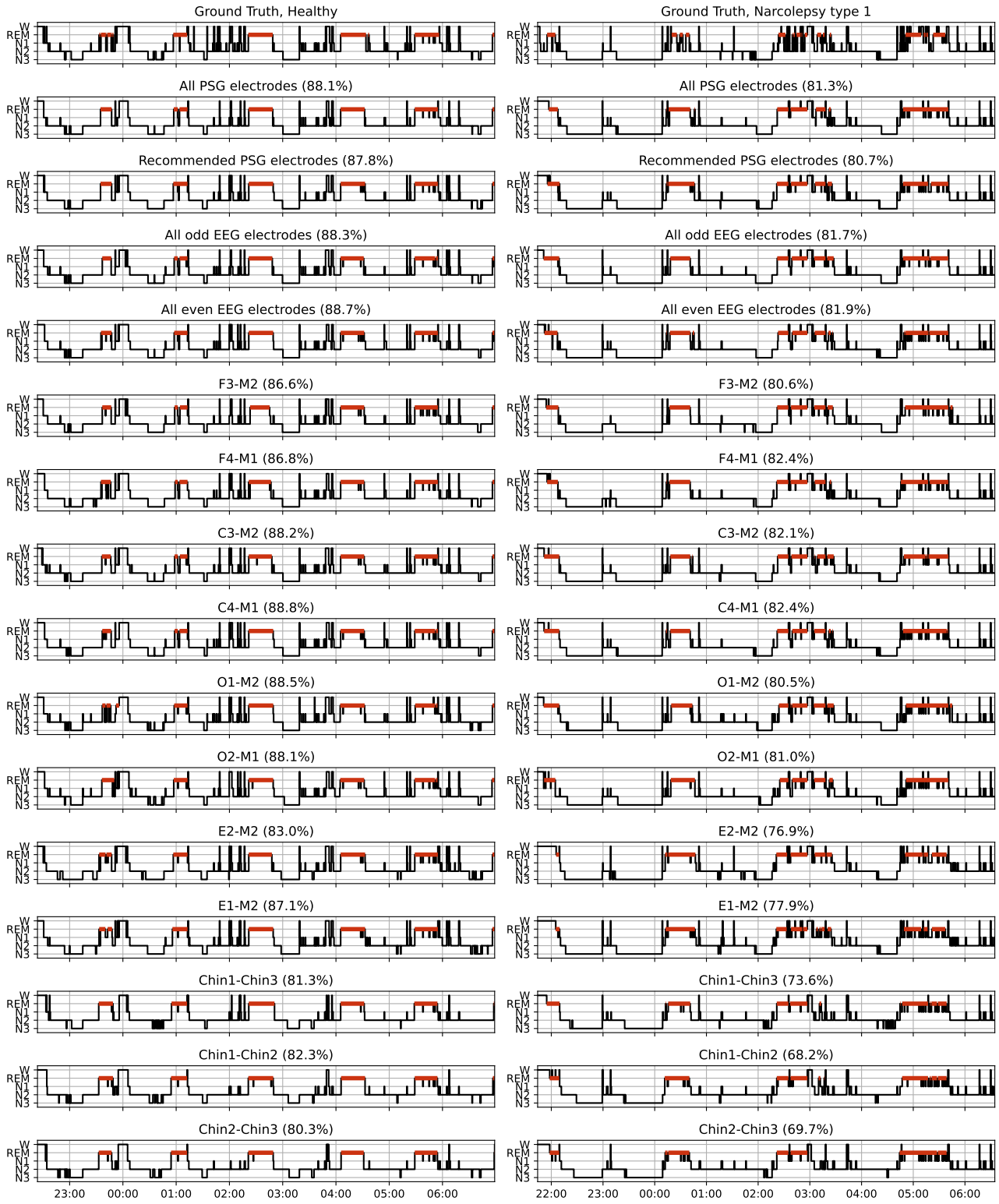


Fig. S.16. Additional Results for the two subjects from Fig. 3. Between brackets the 5 class accuracy for that specific recording is listed. Part 1 of 3.

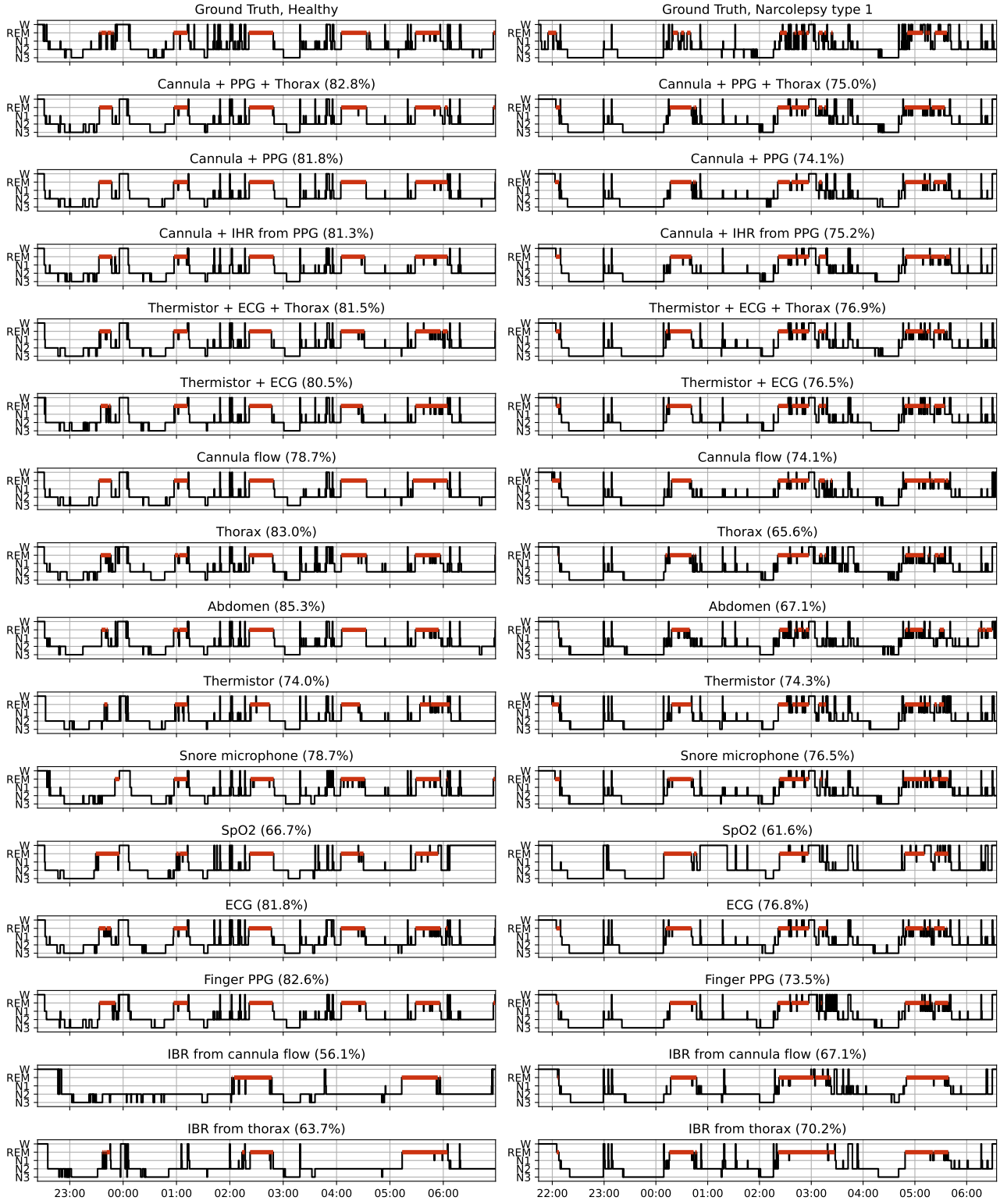


Fig. S.17. Additional Results for the two subjects from Fig. 3. Between brackets the 5 class accuracy for that specific recording is listed. Part 2 of 3.

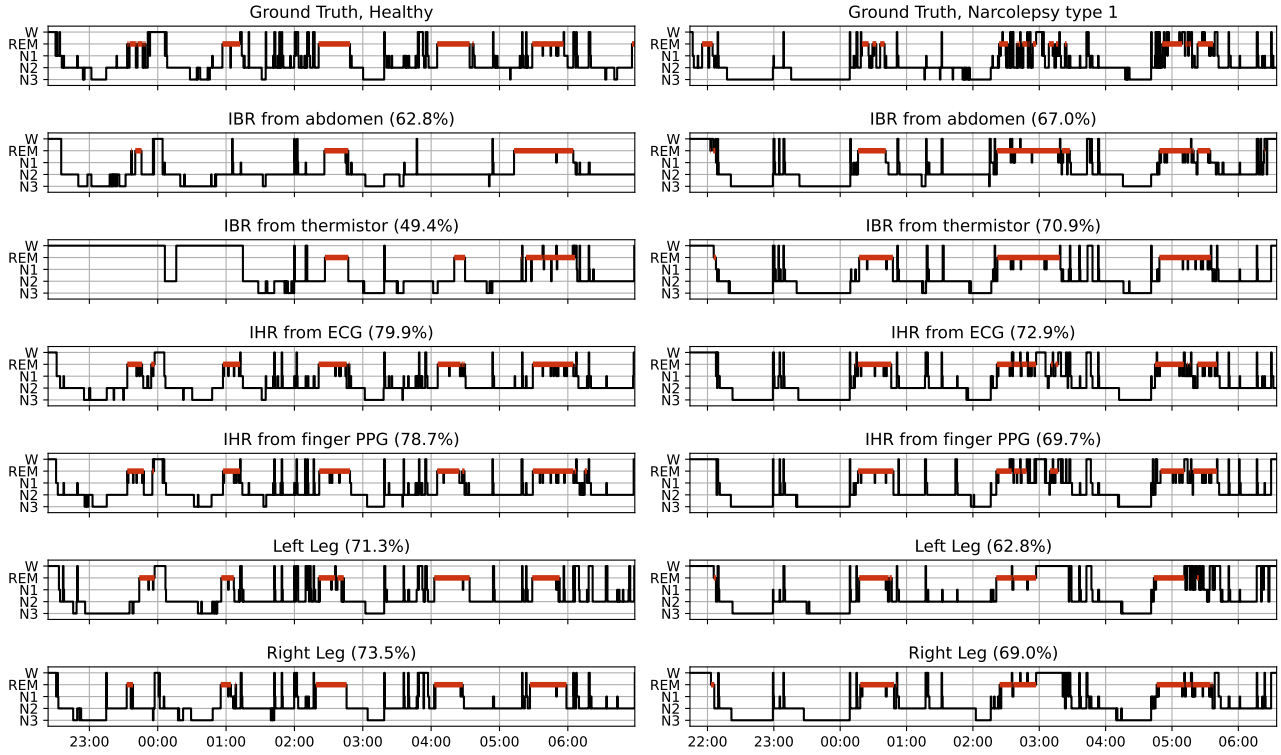


Fig. S.18. Additional Results for the two subjects from Fig. 3. Between brackets the 5 class accuracy for that specific recording is listed. Part 3 of 3.

REFERENCES

[1] M. M. Troester, S. F. Quan, and R. B. Berry, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*, Version 3. Darien, IL, USA: American Academy of Sleep Medicine, 2023.

[2] C. van der Woerd *et al.*, "Studying sleep: towards the identification of hypnogram features that drive expert interpretation," *Sleep*, vol. 47, no. 3, zsad306, Dec. 2023, ISSN: 0161-8105.

[3] H. Phan and K. Mikkelsen, "Automatic sleep staging of EEG signals: Recent development, challenges, and future directions," *Physiological Measurement*, vol. 43, no. 4, 04TR01, Apr. 2022.

[4] P. Fonseca *et al.*, "A computationally efficient algorithm for wearable sleep staging in clinical populations," *Scientific Reports*, vol. 13, no. 1, p. 9182, Jun. 2023, ISSN: 2045-2322.

[5] J. P. Bakker *et al.*, "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*, vol. 17, no. 7, pp. 1343–1354, Jul. 2021.

[6] H. Zhai, Y. Yan, S. He, P. Zhao, and B. Zhang, "Evaluation of the accuracy of contactless consumer sleep-tracking devices application in human experiment: A systematic review and meta-analysis," *Sensors*, vol. 23, no. 10, May 2023, ISSN: 1424-8220.

[7] N. Sridhar *et al.*, "Deep learning for automated sleep staging using instantaneous heart rate," *npj Digital Medicine*, vol. 3, no. 1, p. 106, Aug. 2020, ISSN: 2398-6352.

[8] L. Cerina *et al.*, "A sleep stage estimation algorithm based on cardiorespiratory signals derived from a suprasternal pressure sensor," *Journal of Sleep Research*, e14015, 2023.

[9] M. Perslev, S. Darkner, L. Kempfner, M. Nikolic, P. J. Jennum, and C. Igel, "U-Sleep: Resilient high-frequency sleep staging," *npj Digital Medicine*, vol. 4, no. 1, p. 72, Apr. 2021, ISSN: 2398-6352.

[10] L. Fiorillo *et al.*, "U-Sleep's resilience to AASM guidelines," *NPJ digital medicine*, vol. 6, no. 1, p. 33, 2023.

[11] R. Thapa *et al.*, "SleepFM: Multi-modal representation learning for sleep across brain activity, ECG and respiratory signals," in *Forty-first International Conference on Machine Learning*, 2024.

[12] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations (ICLR)*, 2020.

[13] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *Advances in Neural Information Processing Systems*, vol. 35, Dec. 2022, pp. 26 565–26 577.

[14] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[15] B. Efron, "Tweedie's formula and selection bias," *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1602–1614, 2011.

[16] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 8780–8794.

[17] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[18] J. Song, A. Vahdat, M. Mardani, and J. Kautz, "Pseudoinverse-guided diffusion models for inverse problems," in *International Conference on Learning Representations*, 2023.

[19] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye, "Diffusion posterior sampling for general noisy inverse problems," *arXiv preprint arXiv:2209.14687*, 2022.

[20] T. S. W. Stevens, F. C. Meral, J. Yu, I. Z. Apostolakis, J. L. Robert, and R. J. G. Van Sloun, "Dehazing ultrasound using diffusion models," *IEEE Transactions on Medical Imaging*, 2024.

[21] H. van Gorp, M. M. van Gilst, P. Fonseca, S. Overeem, and R. J. G. van Sloun, "Modeling the impact of inter-rater disagreement on sleep statistics using deep generative learning," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–12, Aug. 2023.

[22] A. Caticha, "Entropic inference," in *AIP Conference Proceedings*, American Institute of Physics, vol. 1305, 2011, pp. 20–29.

[23] L. Fiorillo, D. Pedroncelli, V. Agostini, P. Favaro, and F. D. Faraci, "Multi-scored sleep databases: How to exploit the multiple-labels in automated sleep scoring," *Sleep*, vol. 46, no. 5, zsad028, 2023.

[24] P. Anderer, M. Ross, A. Cerny, R. Vasko, E. Shaw, and P. Fonseca, "Overview of the hypnodensity approach to scoring sleep for polysomnography and home sleep testing," *Frontiers in Sleep*, vol. 2, Apr. 2023.

[25] M. M. van Gilst *et al.*, "Protocol of the SOMNIA project: An observational study to create a neurophysiological database for advanced clinical sleep monitoring," *BMJ open*, vol. 9, no. 11, Nov. 2019.

[26] F. B. van Meulen *et al.*, "Contactless camera-based sleep staging: The healthbed study," *Bioengineering*, vol. 10, no. 1, 2023, ISSN: 2306-5354.

[27] H. van Gorp, M. M. van Gilst, P. Fonseca, S. Overeem, and R. J. G. van Sloun, "Single-channel EOG sleep staging on a heterogeneous cohort of subjects with sleep disorders," *Physiological measurement*, pp. 1–14, 2024.

[28] American Academy of Sleep Medicine, *International classification of sleep disorders*, 3rd ed, text revision. Darien, IL, USA: American Academy of Sleep Medicine, 2023.

[29] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[30] F. Andreotti, H. Phan, and M. De Vos, "Visualising convolutional neural network decisions in automatic sleep scoring," in *CEUR Workshop Proceedings*, CEUR Workshop Proceedings, 2018, pp. 70–81.

[31] M. Dutt, S. Redhu, M. Goodwin, and C. W. Omlin, "SleepXAI: An explainable deep learning approach for multi-class sleep stage identification," *Applied Intelligence*, vol. 53, no. 13, pp. 16 830–16 843, 2023.

[32] H. Phan, K. B. Mikkelsen, O. Chen, P. Koch, A. Mertins, and M. de Vos, "Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 8, pp. 2456–2467, 2022.

[33] I. A. M. Huijben, S. Overeem, M. M. van Gilst, and R. J. G. van Sloun, "Attention on sleep stage specific characteristics," in *46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2024.

[34] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[35] A. Brink-Kjaer, K. M. Gunter, E. Mignot, E. During, P. Jennum, and H. B. D. Sorensen, "End-to-end deep learning of polysomnograms for classification of REM sleep behavior disorder," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2022, pp. 2941–2944.

[36] Z. Chen, Y. Deng, Y. Wu, Q. Gu, and Y. Li, "Towards understanding the mixture-of-experts layer in deep learning," *Advances in neural information processing systems*, vol. 35, pp. 23 049–23 062, 2022.

[37] W. Cai, J. Jiang, F. Wang, J. Tang, S. Kim, and J. Huang, "A survey on mixture of experts," *Authorea Preprints*, 2024.

[38] S. E. Yuksel, J. N. Wilson, and P. D. Gader, "Twenty years of mixture of experts," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 8, pp. 1177–1193, 2012.

[39] H. van Gorp, I. A. M. Huijben, P. Fonseca, R. J. G. van Sloun, S. Overeem, and M. M. van Gilst, "Certainty about uncertainty in sleep staging: a theoretical framework," *Sleep*, vol. 45, no. 8, Jun. 2022.

[40] R. S. Rosenberg and S. Van Hout, "The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring," *Journal of Clinical Sleep Medicine*, vol. 9, no. 1, pp. 81–87, Jan. 2013.

[41] H. Zhu, C. Fu, F. Shu, H. Yu, C. Chen, and W. Chen, "The effect of coupled electroencephalography signals in electrooculography signals on sleep staging based on deep learning methods," *Bioengineering*, vol. 10, no. 5, p. 573, 2023.

[42] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202, PMLR, Jul. 2023, pp. 32 211–32 252.

[43] J. Kohler *et al.*, "Imagine flash: Accelerating emu diffusion models with backward distillation," *Facebook preprint*, Apr. 2024.

[44] J. Xie, P. Fonseca, J. P. van Dijk, S. Overeem, and X. Long, "Assessment of obstructive sleep apnea severity using audio-based snoring features," *Biomedical Signal Processing and Control*, vol. 86, p. 104 942, 2023, ISSN: 1746-8094.

[45] J. F. van der Aar, D. A. van den Ende, P. Fonseca, F. B. van Meulen, and M. M. van Gilst, "Deep transfer learning for automated single-lead EEG sleep staging with channel and population mismatches," *Frontiers in Physiology*, vol. 14, p. 1 287 342, 2024.

[46] F. Scholkmann, J. Boss, and M. Wolf, "An efficient algorithm for automatic peak detection in noisy periodic and quasi-periodic signals," *Algorithms*, vol. 5, no. 4, pp. 588–603, 2012.

[47] M. Baumert, S. Hartmann, and H. Phan, "Automatic sleep staging for the young and the old – evaluating age bias in deep learning," *Sleep Medicine*, vol. 107, pp. 18–25, 2023, ISSN: 1389-9457.

[48] J. B. Stephansen *et al.*, "Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy," *Nature Communications*, vol. 9, no. 1, p. 5229, Dec. 2018, ISSN: 2041-1723.

[49] Y. Wu and K. He, "Group normalization," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[50] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[51] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 648–656.

[52] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[53] J. Van Der Donckt, J. Van Der Donckt, E. Deprost, *et al.*, "Do not sleep on traditional machine learning: Simple and interpretable techniques are competitive to deep learning for sleep scoring," *Biomedical Signal Processing and Control*, vol. 81, p. 104 429, 2023.

[54] H. Phan, O. Y. Chén, M. C. Tran, P. Koch, A. Mertins, and M. De Vos, "XSleepNet: Multi-view sequential model for automatic sleep staging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5903–5915, 2021.

[55] Q. Shen, J. Xin, X. Liu, *et al.*, "LGSleepNet: An automatic sleep staging model based on local and global representation learning," *IEEE Transactions on Instrumentation and Measurement*, 2023.

[56] A. Supratak and Y. Guo, "TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2020, pp. 641–644.

[57] L. Fiorillo, P. Favaro, and F. D. Faraci, "Deepsleepnet-lite: A simplified automatic sleep stage scoring model with uncertainty estimates," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 29, pp. 2076–2085, 2021.

[58] S. Mousavi, F. Afghah, and U. R. Acharya, "SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach," *PloS one*, vol. 14, no. 5, e0216456, 2019.

[59] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. Kamphuisen, and J. J. Oberye, "Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 9, pp. 1185–1194, 2000.

[60] A. L. Goldberger, L. A. Amaral, L. Glass, *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, e215–e220, 2000.

[61] A. Rechtschaffen and A. Kales, "A manual for standardized terminology, techniques and scoring system for sleep stages in human subjects," *Brain information service*, 1968.