SUPPLEMENTARY TABLES

Table S1: Temporal context impact on model performance in validation partition ($n = 426$).

	Overall accuracy				Cohen's kappa			
	Mean	**SD**	**Median**	**95% CI, mean**	**Mean**	**SD**	**Median**	**95% CI, mean**
Hidden units								
0	0.779	0.083	0.794	[0.771-0.787]	0.645	0.126	0.660	[0.633-0.657]
64	0.818	0.079	0.837	[0.810-0.825]	0.720	0.120	0.745	[0.709-0.731]
128	0.821	0.080	0.841	[0.813-0.829]	0.724	0.121	0.745	[0.713-0.736]
256	0.820	0.082	0.843	[0.812-0.828]	0.725	0.124	0.751	[0.713-0.736]
512	0.822	0.079	0.841	[0.815-0.830]	0.727	0.119	0.752	[0.716-0.739]
1024	0.828	0.072	0.845	[0.821-0.835]	0.734	0.111	0.758	[0.723-0.744]
2048	0.823	0.080	0.843	[0.816-0.831]	0.729	0.122	0.757	[0.717-0.740]
Sequence length								
2 min	0.821	0.075	0.840	[0.814-0.828]	0.726	0.114	0.754	[0.715-0.737]
3 min	0.826	0.080	0.845	[0.818-0.833]	0.733	0.123	0.762	[0.721-0.744]
4 min	0.828	0.079	0.849	[0.820-0.835]	0.734	0.122	0.762	[0.722-0.745]
5 min	0.828	0.072	0.845	[0.821-0.835]	0.734	0.111	0.758	[0.723-0.744]
10 min	0.829	0.075	0.848	[0.822-0.836]	0.734	0.113	0.759	[0.723-0.745]
Window length								
1 s	0.824	0.074	0.843	[0.817-0.831]	0.728	0.113	0.752	[0.717-0.738]
3 s	0.824	0.074	0.845	[0.817-0.832]	0.728	0.113	0.752	[0.717-0.739]
5 s	0.825	0.074	0.843	[0.818-0.832]	0.728	0.113	0.752	[0.717-0.739]
10 s	0.825	0.074	0.844	[0.818-0.832]	0.729	0.113	0.753	[0.718-0.739]
15 s	0.826	0.074	0.845	[0.818-0.833]	0.729	0.113	0.755	[0.719-0.740]
30 s	0.829	0.075	0.848	[0.822-0.836]	0.734	0.113	0.759	[0.723-0.745]

The **Hidden units** variable corresponds to varying the complexity in the recurrent module by increasing the number of hidden units. **Sequence length** indicate the length of the sequence of 30 epochs, while **Window length** correspond to varying the evaluation frequency.

Table S2: Performance characteristics for LOCI and LOCO training configurations.

	N PSGs	Overall accuracy				Cohen's kappa			
		Mean	**SD**	**Median**	**95% CI, mean**	**Mean**	**SD**	**Median**	**95% CI, mean**
LOCI-wd									
ISRUC	1584	0.679	0.123	0.701	[0.673-0.685]	0.542	0.169	0.574	[0.533-0.550]
MrOS	1584	0.821	0.077	0.835	[0.817-0.825]	0.727	0.114	0.745	[0.721-0.733]
SHHS	1584	0.834	0.088	0.858	[0.830-0.839]	0.750	0.132	0.786	[0.744-0.757]
SSC	1584	0.762	0.094	0.774	[0.757-0.767]	0.639	0.129	0.654	[0.633-0.646]
WSC	1584	0.758	0.105	0.773	[0.753-0.764]	0.633	0.145	0.653	[0.626-0.640]
LOCI									
ISRUC	1584	0.676	0.124	0.700	[0.670-0.682]	0.539	0.170	0.574	[0.531-0.547]
MrOS	1584	0.826	0.074	0.839	[0.822-0.829]	0.732	0.111	0.748	[0.726-0.737]
SHHS‡	1584	0.837	0.084	0.858	[0.833-0.841]	0.754	0.127	0.786	[0.748-0.761]
SSC	1584	0.773	0.088	0.785	[0.769-0.777]	0.657	0.125	0.671	[0.651-0.663]
WSC	1584	0.763	0.101	0.776	[0.758-0.768]	0.641	0.140	0.659	[0.635-0.648]
LOCO									
ISRUC†	52	0.749	0.081	0.764	[0.727-0.771]	0.648	0.119	0.682	[0.616-0.680]
	126	*0.757*	*0.071*	*0.766*	*[0.744-0.769]*	*0.661*	*0.101*	*0.682*	*[0.643-0.678]*
MrOS†	371	0.843	0.066	0.851	[0.836-0.849]	0.757	0.104	0.776	[0.746-0.767]
	3932	*0.841*	*0.069*	*0.854*	*[0.838-0.843]*	*0.752*	*0.107*	*0.775*	*[0.749-0.755]*
SHHS	846	0.805	0.076	0.815	[0.800-0.810]	0.705	0.109	0.722	[0.698-0.712]
	8444	*0.800*	*0.081*	*0.811*	*[0.798-0.801]*	*0.697*	*0.115*	*0.713*	*[0.694-0.699]*
SSC	76	0.793	0.086	0.809	[0.744-0.812]	0.680	0.120	0.700	[0.653-0.707]
	766	*0.798*	*0.086*	*0.815*	*[0.792-0.805]*	*0.690*	*0.123*	*0.711*	*[0.681-0.699]*
WSC†	239	0.826	0.065	0.835	[0.818-0.834]	0.720	0.096	0.736	[0.708-0.732]
	2411	*0.824*	*0.068*	*0.837*	*[0.821-0.827]*	*0.718*	*0.100*	*0.736*	*[0.714-0.722]*

Metrics are aggregated across all subjects for each cohort in test partition ($N = 1{,}584$ PSGs). Statistics in italics correspond to evaluating performance on entire cohort. PSG: polysomnography; LOCI-wd: leave-one-cohort-in with weight decay; LOCO: leave-one-cohort-out; ISRUC: Institute of Systems and Robotics, University of Coimbra Sleep Cohort; MrOS: The Osteoporotic Fractures in Men Sleep Study; SHHS: Sleep Heart Health Study; SSC: Stanford Sleep Cohort; WSC: Wisconsin Sleep Cohort; †: significantly better than corresponding LOCI; ‡: significantly better than corresponding LOCO.

Table S3: Model performance of test partition with varying fractions of training data.

	Overall accuracy				**Cohen's kappa**			
	Mean	**SD**	**Median**	**95% CI, mean**	**Mean**	**SD**	**Median**	**95% CI, mean**
Fraction (%)								
0.25	0.782	0.097	0.801	[0.777-0.787]	0.671	0.141	0.696	[0.664-0.678]
0.50	0.804	0.086	0.824	[0.800-0.808]	0.696	0.131	0.724	[0.689-0.702]
1	0.824	0.079	0.840	[0.820-0.828]	0.730	0.118	0.753	[0.724-0.736]
5	0.841	0.074	0.856	[0.837-0.844]	0.757	0.113	0.780	[0.751-0.763]
10	0.850	0.069	0.864	[0.847-0.853]	0.770	0.108	0.791	[0.765-0.775]
25	0.858	0.066	0.873	[0.854-0.861]	0.782	0.102	0.804	[0.777-0.787]
50	0.860	0.063	0.874	[0.856-0.863]	0.787	0.097	0.809	[0.782-0.792]
75	0.867	0.062	0.882	[0.864-0.870]	0.797	0.096	0.818	[0.792-0.802]
100	0.869	0.064	0.883	[0.865-0.872]	0.799	0.098	0.820	[0.794-0.804]

Increasing the available training data increased performance on the test partition ($N = 1{,}584$) shown here as aggregated metrics across all subjects. No statistical difference was found by comparing confidence intervals (CI) between models trained with 75% and 100% of available training data, which indicates a saturation in training.

Table S4: Model performance on test partition ($N = 1{,}584$) with varying number of cohorts in training partition.

Training cohorts	Overall accuracy				Kappa			
	Mean	SD	Median	95% CI, mean	Mean	SD	Median	95% CI, mean
2								
Overall	0.788	0.102	0.811	[0.787-0.790]	0.683	0.143	0.710	[0.681-0.685]
ISRUC-MrOS	0.781	0.102	0.804	[0.776-0.786]	0.675	0.143	0.703	[0.668-0.682]
ISRUC-SHHS	0.808	0.097	0.835	[0.804-0.813]	0.717	0.142	0.756	[0.710-0.724]
ISRUC-SSC	0.735	0.103	0.753	[0.729-0.740]	0.613	0.140	0.638	[0.606-0.620]
ISRUC-WSC	0.745	0.107	0.758	[0.740-0.750]	0.628	0.140	0.642	[0.621-0.635]
MrOS-SHHS	0.829	0.081	0.849	[0.825-0.833]	0.740	0.124	0.769	[0.734-0.746]
MrOS-SSC	0.796	0.090	0.816	[0.791-0.800]	0.683	0.133	0.708	[0.677-0.690]
MrOS-WSC	0.805	0.087	0.822	[0.801-0.809]	0.699	0.126	0.722	[0.693-0.705]
SHHS-SSC	0.816	0.090	0.839	[0.812-0.821]	0.722	0.129	0.755	[0.716-0.729]
SHHS-WSC	0.824	0.089	0.846	[0.820-0.828]	0.733	0.128	0.762	[0.727-0.739]
SSC-WSC	0.742	0.110	0.755	[0.737-0.748]	0.620	0.145	0.634	[0.613-0.627]
3								
Overall	0.808	0.092	0.830	[0.807-0.810]	0.711	0.131	0.739	[0.709-0.713]
ISRUC-MrOS-SHHS	0.820	0.092	0.844	[0.815-0.825]	0.732	0.134	0.766	[0.725-0.738]
ISRUC-MrOS-SSC	0.798	0.088	0.816	[0.794-0.802]	0.694	0.129	0.720	[0.688-0.700]
ISRUC-MrOS-WSC	0.811	0.083	0.828	[0.807-0.815]	0.711	0.119	0.735	[0.705-0.717]
ISRUC-SHHS-SSC	0.807	0.090	0.828	[0.803-0.812]	0.714	0.126	0.739	[0.708-0.721]
ISRUC-SHHS-WSC	0.817	0.091	0.842	[0.813-0.822]	0.728	0.128	0.759	[0.722-0.735]
ISRUC-SSC-WSC	0.755	0.109	0.775	[0.750-0.760]	0.639	0.150	0.670	[0.631-0.646]
MrOS-SHHS-SSC	0.833	0.071	0.848	[0.829-0.837]	0.744	0.109	0.766	[0.739-0.750]
MrOS-SHHS-WSC	0.840	0.073	0.854	[0.836-0.843]	0.753	0.109	0.774	[0.748-0.759]
MrOS-SSC-WSC	0.795	0.088	0.811	[0.791-0.800]	0.687	0.123	0.706	[0.681-0.693]
SHHS-SSC-WSC	0.807	0.101	0.833	[0.802-0.812]	0.710	0.142	0.744	[0.703-0.717]
4								
Overall	0.821	0.085	0.840	[0.819-0.823]	0.728	0.124	0.755	[0.726-0.731]
ISRUC-MrOS-SHHS-SSC	0.827	0.078	0.843	[0.823-0.831]	0.739	0.115	0.764	[0.733-0.744]
ISRUC-MrOS-SHHS-WSC	0.835	0.075	0.850	[0.831-0.838]	0.747	0.112	0.768	[0.742-0.753]
ISRUC-MrOS-SSC-WSC	0.794	0.097	0.817	[0.789-0.799]	0.687	0.139	0.716	[0.680-0.694]
ISRUC-SHHS-SSC-WSC	0.819	0.091	0.843	[0.814-0.823]	0.728	0.131	0.759	[0.721-0.734]
MrOS-SHHS-SSC-WSC	0.830	0.076	0.846	[0.826-0.834]	0.741	0.112	0.763	[0.736-0.747]

The total number of training records were fixed at $N = 500$ for all configurations. ISRUC: Institute of Systems and Robotics, University of Coimbra Sleep Cohort; MrOS: The Osteoporotic Fractures in Men Sleep Study; SHHS: Sleep Heart Health Study; SSC: Stanford Sleep Cohort; WSC: Wisconsin Sleep Cohort.