APPENDIX
SUPPLEMENTARY FIGURES AND TABLES

TABLE A.1
SLEEP DETECTION ACCURACY WITH BEST RESULTS HIGHLIGHTED IN BOLD.

Feature combinations	Bi-LSTM RNN		ANN		Logistic regression		SVM	
	Without time	With time	Without time	With time	Without time	With time	Without time	With time
Wrist sensor	95.9	96.5	91.2	93.5	87.2	88.6	88.3	90.0
Phone	76.7	89.1	71.4	87.2	71.4	74.7	71.4	81.3
Wrist + Phone	96.0	96.3	91.4	93.7	87.6	89.0	88.7	90.3
ACC + EDA	95.6	96.3	90.0	92.9	85.1	87.9	87.3	89.4
ACC + ST	**96.2**	**96.5**	91.0	93.3	87.1	88.5	88.1	90.0
EDA + ST	92.5	94.5	84.8	91.3	82.2	84.5	82.3	84.7
ACC	95.5	96.3	89.6	92.6	87.9	88.1	87.2	89.2
EDA	90.8	93.7	78.8	89.8	72.7	78.0	72.8	80.3
ST	86.7	90.7	81.2	89.4	81.1	82.8	81.2	83.1
Actigraphy	94.1							

TABLE A.2
THE STATISTICS OF SLEEP DETECTION ACCURACY ACROSS $N = 149$ PARTICIPANTS. RIGHT-TAILED PAIRED T-TESTS WERE CONDUCTED BETWEEN EACH FEATURE COMBINATION AND THE ACTIGRAPHY-ONLY METHOD WITH T-VALUES AND P-VALUES ALSO SHOWN IN THE TABLE. A STAR AFTER A P-VALUE INDICATES THAT THE RESULT IS SIGNIFICANTLY BETTER THAN THAT OF THE ACTIGRAPHY-ONLY METHOD AT THE 5% SIGNIFICANCE LEVEL.

Feature combinations	Without time				With time			
	Mean	SD	t(149)	p	Mean	SD	t(149)	p
Wrist sensor	95.8	3.3	4.12	$3.12\text{x}10^{-5}$ *	96.3	2.9	-21.40	$1.42\text{x}10^{-7}$ *
Phone	76.9	8.6	76.9	1.00	89.0	5.4	-8.99	1.00
Wrist + Phone	95.9	3.1	4.55	$5.56\text{x}10^{-6}$ *	96.3	2.8	5.18	$3.61\text{x}10^{-7}$ *
ACC + EDA	95.5	3.7	4.16	$2.66\text{x}10^{-5}$ *	96.2	3.0	5.14	$4.29\text{x}10^{-7}$ *
ACC + ST	96.1	3.1	4.93	$1.07\text{x}10^{-6}$ *	96.4	3.0	5.41	$1.26\text{x}10^{-7}$ *
EDA + ST	92.5	5.6	-2.39	0.99	94.5	3.7	1.00	0.16
ACC	95.3	4.1	3.72	$1.41\text{x}10^{-4}$ *	95.4	4.2	3.93	$6.50\text{x}10^{-5}$ *
EDA	90.8	7.2	-4.61	1.00	72.7	93.7	-0.56	0.71
ST	86.5	5.7	-12.60	1.00	90.7	4.7	-6.10	1.00
Actigraphy	94.0	5.5						

TABLE A.3
SLEEP DETECTION ACCURACY COMPARISON (BIDIRECTIONAL LSTM MODELS WITH DATA SPLIT BY DAYS AND PARTICIPANTS AND REAL-TIME LSTM MODEL WITH DATA SPLIT BY DAYS AND PARTICIPANTS.

Feature Combination	Bidirectional LSTM model				LSTM with previous 30 min data	
	With data split by days		With data split by participants		With data split by days	
	Without time	With time	Without time	With time	Without time	With time
Wrist sensor	95.9	96.5	96.0	96.3	95.5	96.0
Phone	76.7	89.1	77.3	88.6	74.3	88.1
Wrist + Phone	96.0	96.3	96.0	96.2	95.6	96.0
ACC + EDA	95.6	96.3	95.8	96.0	95.1	95.8
ACC + ST	96.2	96.5	96.0	96.4	95.6	96.1
EDA + ST	92.5	94.5	92.8	94.7	91.7	93.8
ACC	95.5	96.3	95.8	96.2	94.9	95.7
EDA	90.8	93.7	91.9	94.1	89.3	93.2
ST	86.7	90.7	87.5	91.0	85.4	90.5

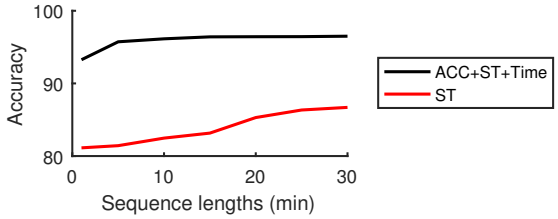

Fig. A.1. Sleep detection accuracy versus the past- and future-looking sequence lengths using two different feature combinations. The performance of sleep detection using the more informative feature combination (ACC+ST+Time) saturates when the sequence length is longer than 15 min, while the performance of using only the ST feature keeps increasing until a sequence length of 30 min. Considering the extra resources cost by longer sequences, the 30-min length we used in all the main experiments proves to be an appropriate choice.

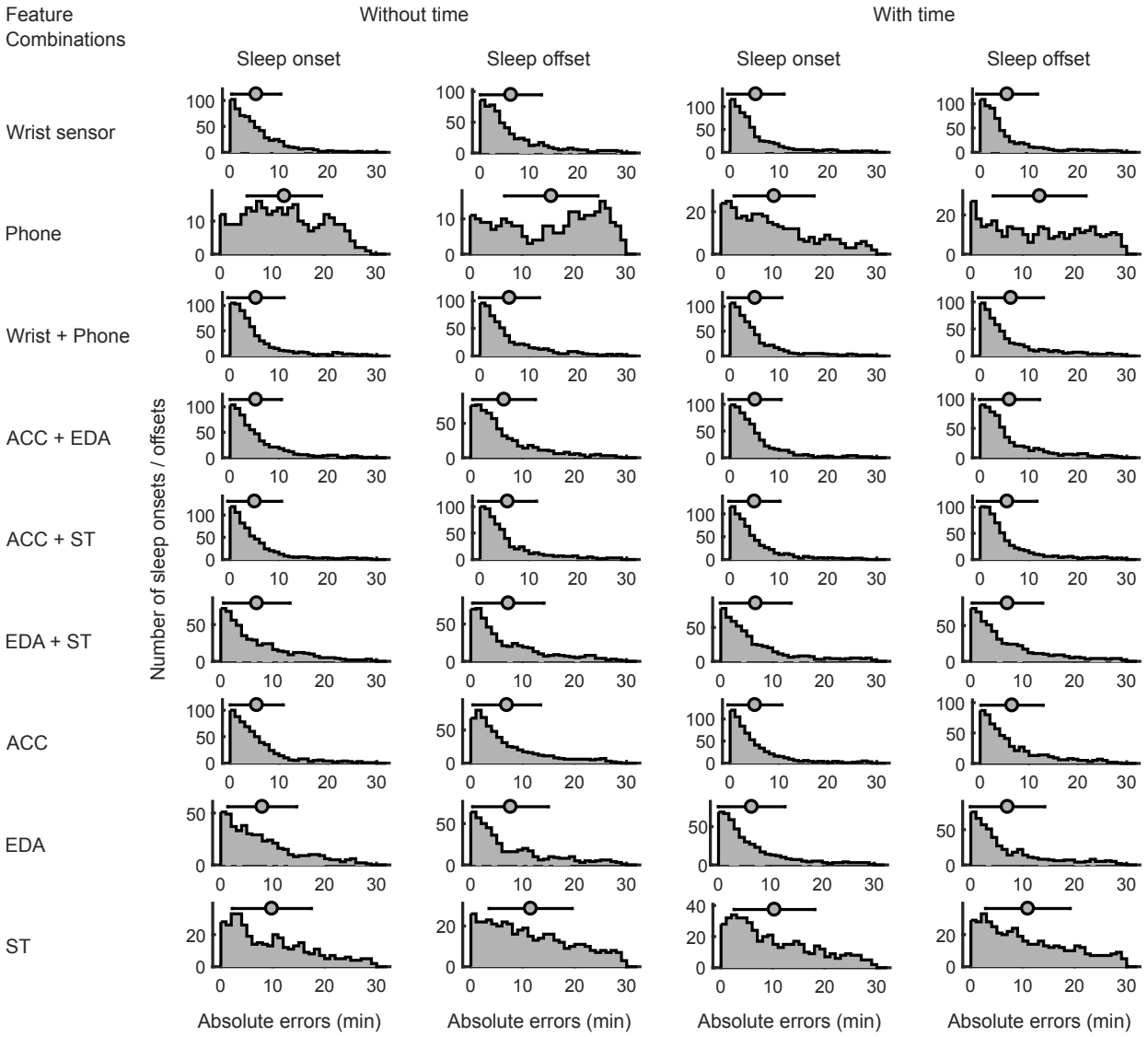

Fig. A.2. Error distributions for sleep episode onset/offset detection using differential Bi-LSTM RNN. The means and standard deviations of the absolute errors are also denoted in each plot.

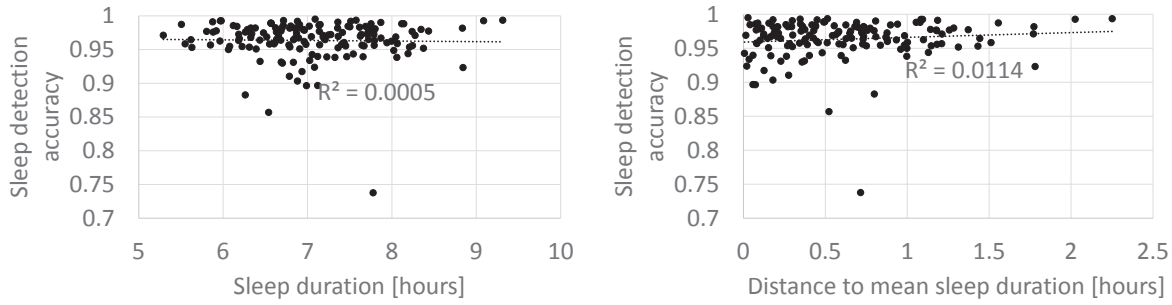

Fig. A.3. (left) Relationship between average sleep duration vs sleep detection accuracy (ACC+ST+time).(right)Relationship between distance to average sleep duration vs sleep detection accuracy (ACC+ST+time).