Multimodal Ambulatory Sleep Detection Using LSTM Recurrent Neural Networks

Akane Sano[1,*], Weixuan Chen[2,*], Daniel Lopez-Martinez[2,3], Sara Taylor[2], and Rosalind W. Picard[2]

***Abstract*—Unobtrusive and accurate ambulatory methods are needed to monitor long-term sleep patterns for improving health. Previously developed ambulatory sleep detection methods rely either in whole or in part on self-reported diary data as ground truth, which is a problem since people often do not fill them out accurately. This paper presents an algorithm that uses multimodal data from smartphones and wearable technologies to detect sleep/wake state and sleep onset/offset using a type of recurrent neural network with long-short-term memory (LSTM) cells for synthesizing temporal information. We collected 5580 days of multimodal data from 186 participants and compared the new method for sleep/wake classification and sleep onset/offset detection to (1) non-temporal machine learning methods and (2) a state-of-the-art actigraphy software. The new LSTM method achieved a sleep/wake classification accuracy of 96.5%, and sleep onset/offset detection F_1 scores of 0.86 and 0.84 respectively, with mean absolute errors of 5.0 and 5.5 min, respectively, when compared with sleep/wake state and sleep onset/offset assessed using actigraphy and sleep diaries. The LSTM results were statistically superior to those from non-temporal machine learning algorithms and the actigraphy software. We show good generalization of the new algorithm by comparing participant-dependent and participant-independent models, and we show how to make the model nearly realtime with slightly reduced performance.**

***Index Terms*—Sleep monitoring, sleep detection, recurrent neural networks, long-short-term memory, LSTM, wearable sensor, mobile phone, smartphone, mobile health.**

I. Introduction

HEALTHY sleep requires adequate duration, good quality, appropriate timing, and regularity [1]. Inadequate sleep increases appetite and food intake [2], decreases insulin sensitivity and glucose tolerance [3], impairs immune function [4], disturbs mood [5], leads to slowed reaction time [6] and attentional failures [7], [8], and compromises memory and learning [9]. Unfortunately, an estimated ∼30% of adults have a sleep disorder [10], and most people who have sleep disorders remain undiagnosed [11].

There are two tools that are commonly used in diagnosing and treating sleep-related problems: bedtime sleep monitoring, and ambulatory sleep detection. The former, which includes polysomnography (PSG), is usually applied only during sleep at night, and usually involves coarse or fine detection of sleep stages. Its equipment is either too obtrusive to wear during the daytime or restricted in bed so is not used in studies that continuously monitor participants for 24 hours/day for multiple days or weeks. The latter, ambulatory sleep detection, is typically used 24-hours/day to estimate sleep episode onset/offset times and sleep duration from continuous behavioral and physiological data.

PSG is the gold standard for bedtime sleep monitoring and diagnosis of sleep disorders [12], recording electroencephalogram (EEG), heart rhythm, respiratory effort, eye and leg movements and oxygen saturation over multiple nights in a sleep laboratory to produce a detailed picture of a patient's sleep patterns [12]. Home unattended polysomnography (H-PSG) is a lower cost option than in-clinic PSG. Collecting data at home is likely to have more validity for understanding a person's typical sleep patterns since a person's sleep is impacted by their environment [13]. However, both PSG and H-PSG tend to involve bulky equipment that requires significant time to put on properly and to interpret; the systems tend to be cumbersome, and might themselves interact with the sleep behavior, which make them impractical for long-term sleep/wake detection. More recent methods involve less cumbersome equipment such as using a smartphone placed on the bed to track movement via accelerometers [14], using contact [15] and non-contact microphones [16], short-range doppler radar [17] and WiFi [18].

Clinical sleep studies use two standard instruments for performing ambulatory sleep detection: actigraphy and diaries. Actigraphy [19], [20] is based on the observation that there is less movement during sleep and more movement during wake [21], [22]. Since actigraphs are typically small and comfortable to wear, actigraphy can conveniently be recorded continuously for 24-hours/day for weeks or longer. However, it has been shown to fail in special populations and to be unreliable for detecting wakefulness during motionless periods. Sleep diaries and questionnaires [23] also have several drawbacks, namely users' adherence and reporting bias [23], [24]. Significant effort is required by users to maintain accurate diaries, and by researchers to check their entries for anomalies. Today's best ambulatory studies combine diaries and actigraphs in a labor-intensive human-validation process to merge their combined measures [25]. There is thus a need for better automated tools to enable accurate long-term evaluation of sleep timing and duration in daily life. In this paper, we focus on ambulatory sleep detection.

Smartphones and wearables that measure acceleration, light, heart rate, skin conductance, skin temperature, phone usage, and other behaviors and physiology offer a low-cost, easy-to-

*Both authors contributed equally to this work
[1]Department of Electrical and Computer Engineering, Rice University Akane.Sano@rice.edu
[2]Affective Computing Group, Media Lab, Massachusetts Institute of Technology
[3]Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology.
Manuscript received April 19, 2005; revised August 26, 2015.

use possibility for ambulatory long-term sleep detection; for example accelerometer [26], smartphone [27], [28], or biosensor [29] data alone have been used to achieve ambulatory sleep/wake detection. However, none of these have compared their results with the current best ambulatory detection that merges actigraphy and sleep diaries [25].

In this paper, we present a novel automated method to detect sleep/wake state and sleep onset/offset times using recurrent neural networks with long short-term memory (LSTM) cells [30] applied to multimodal data from a smartphone and a wrist-worn sensor, and labeled by both actigraphy and sleep diaries. The proposed method combines multimodal ambulatory physiological and behavioral data and improves upon our team's earlier work showing that refining sleep onset/offset times can help improve sleep characterization [31].

This paper makes several novel contributions:

- We develop a fully automated machine learning algorithm for (i) sleep/wake classification and (ii) sleep onset/offset detection, using physiological and behavioral data from a mobile phone and a wearable sensor. This automated algorithm requires much less human effort than the actigraphy + sleep diary method.
- The algorithm achieves higher performance in real-life ambulatory settings than actigraphy + an existing fully automated sleep detection algorithm (actigraphy software: Action4).
- We compare the effectiveness of different physiological and behavioral modalities and determine the best combination for sleep/wake detection.
- The new bidirectional LSTM model is shown to outperform three other machine learning models.
- We show the new model generalizes well to people not included in its training data.
- We show that the new algorithm, with minor adjustments, can give near real-time sleep/wake estimates.

II. Related Work

A. Ambulatory Sleep detection systems

The most common ambulatory method for sleep evaluation is actigraphy, in which movement detectors (typically accelerometers) are placed on the wrist, ankle or trunk, and used to sample movement several times per second to derive sleep/wake parameters such as total sleep time, percent of time spent asleep, total wake time, percent of time spent awake, number of awakenings and sleep efficiency [32]. Watch-sized, consumer-oriented, wearable sensors such as the wrist-worn FitBit are able to record actigraphy and perform sleep detection [33], [34]. While actigraphy has reasonable validity and reliability in assessing sleep-wake patterns in normal individuals with average or good sleep quality [35] and has been shown to be more reliable than subjective or self-reported sleep diaries and behavior logs [36], it may fail in special populations (e.g., elderly people, individuals with other major health problems or individuals with poor sleep quality, such as patients with movement disorders [37] and shift workers [38] [39]). The main problem associated with actigraphic sleep-wake detection is the false labeling of sleep when people are awake but relatively motionless [39], [40].

Smartphone usage patterns (e.g., the time and length of smartphone usage or recharge events) and environmental observations (e.g., prolonged silence and darkness) have also been used to infer periods of likely sleep [27], [28], [41]. Chen et al. automatically inferred sleep duration with $\pm$42 minute error using smartphone data (light sensor, phone lock, off and charge logs, activity and audio features) and regression models [27].

Min et al. used sound amplitude, acceleration, light intensity, screen proximity, app usage, battery and screen status and detected sleep episodes with 93% accuracy and showed $\pm$44 minute, $\pm$42 minute, and $\pm$64 minute errors for bedtime, waketime, and duration compared with sleep diaries, using a Bayesian Network [28]. Saeb et al. used random forest classifiers to develop both personalized and global self-reported sleep detection models state from the phone sensor data (location, motion, light, sound, and in-phone activity data from Android phone). They obtained 88.8% accuracy to detect sleep segments and 91.8% after correcting missing sensor data and sleep diaries with an average median absolute deviation (MAD) of 38 min for sleep episode onset detection and 36 min for sleep episode offset detection [41]. These previous studies used self-reported sleep diaries as ground truth and did not use temporal machine learning models, which we show can improve the accuracy of sleep/wake detection and sleep onset/offset timing.

B. Recurrent neural networks for sleep data

A recurrent neural network (RNN) is a type of artificial neural network where connections between units form a directed cycle. This allows it to exhibit dynamic temporal behavior and process arbitrary sequences of inputs. A significant limitation of vanilla RNN models, which strictly integrate state information over time, is known as the "vanishing gradients" effect, where the gradient signal gets so small that learning either becomes very slow or stops working altogether. Long-short-term memory networks (LSTMs) are a special kind of RNN, capable of learning long-term dependencies [42]. They contain gate functions that determine when the input is significant enough to remember and when it should continue to remember or forget the value, and when it should output the value. LSTMs have recently shown great success in temporal sequence tasks such as speech recognition [43] and machine translation [44], [45].

RNNs, especially LSTMs, have been successfully used in sleep studies. They have been applied to actigraphy during awake time to predict sleep quality (good/poor binary prediction), performing with 79.6 % accuracy and 0.85 F_1 score [46]. In combination with deep belief networks (DBN), they have also been applied to a single channel of EEG and EOG data for classifying 5 sleep stages (Wake, non-REM1/2/3 and REM) with the best performance of overall accuracy 85.92 % and macro F_1 score 80.50 [47]. In another study, DBN and LSTM have been applied to EEG, EOG, and EMG data for classification of the 5 sleep stage classification (overall accuracy 98.8 % and F_1 score 0.99) [48].

In this paper, we aim to develop an LSTM-based detector of sleep/wake and sleep episode onset/offset timing using multi-modal data from a wearable sensor and mobile phone.

III. Methods

A. Data acquisition

186 undergraduate students in 5 cohorts participated in an ∼30-day study (120 males, 66 females, age: 18-25) that produced 5580 days of data. The study protocol was approved by the Massachusetts Institute of Technology (# 1209005240). Participants were recruited through email. During the ∼30-day experiment, participants (i) wore a wrist sensor on their dominant hand (Q-sensor, Affectiva, USA) to measure 3-axis acceleration (ACC), skin conductance (SC), and skin temperature (ST) at 8 Hz; (ii) installed an Android phone application using the *funf* open source framework [49] to measure timing of calls, timing of short message service (SMS), location, and timing of screen-on; (iii) wore a wrist actigraphy monitor on their non-dominant hand (Motion Logger, AMI, USA) to measure activity and light exposure levels every 1 minute; and (iv) completed a sleep diary every morning to record bed time, sleep latency, wake time, and the number and timing of awakenings. The sleep diary was inspected by an experimenter every day to check completion and to obtain corrections or clarifications from the student if there were any clear errors or missing data.

To obtain the ground truth of sleep/wake epochs and sleep onset and offset, we used a method previously established by Harvard Medical School sleep experts to score sleep from diaries and actigraphy data [25]. An experienced investigator first reviewed the data and selected analysis windows for potential sleep episodes based on the combined diary and activity data. Actigraphy software (Action4, AMI, USA) set the sleep episode onset/offset times and classified each epoch as sleep or wake. Based on the sleep episode time and duration, the investigator labeled each sleep episode as either a main sleep or a nap based on sleep diaries. From this procedure, we obtained (1) a classification of sleep or wake for every 1-min epoch, (2) sleep episode onset/offset times, (3) whether a sleep episode was from main sleep or a nap. These labels were used as "ground truth". Fig. 1 shows an exemplary day of raw data collected in our study in which the first stage labels are superimposed. These assessments were used to train and test results from the Q-sensor and phone data. The diaries and actigraphy data were not used as inputs for sleep/wake and sleep episode onset/offset detection because they were used to generate the ground truth labels.

B. Feature preparation

Table I shows the features we computed for each time window. A window length of 20 or 30 seconds is the convention for PSG sleep scoring [50], while other studies using ambulatory data adopt 10-min [28] or 5-min [27] windows. In this study, we used a window length of 1 minute without overlap to match the scale of our ground truth labels.

There were several reasons why we chose these feature variables. First, it has been shown that SC is more likely to have

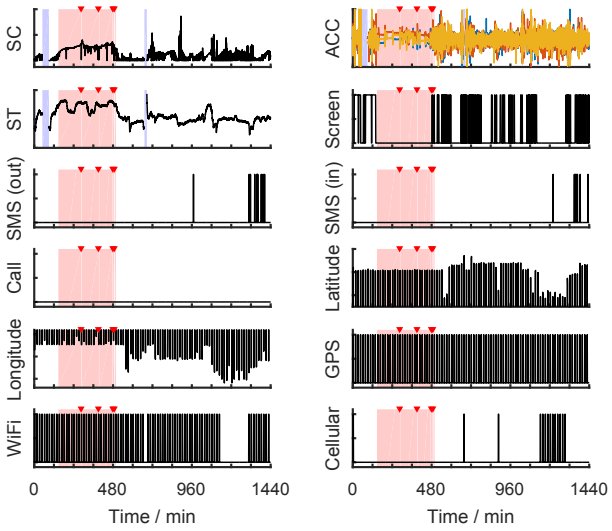

Fig. 1. Raw data streams from an exemplary day. The pink bars mark sleep epochs and the red triangles indicate waking up during the night as determined from actigraphy and sleep diary. The blue bars denote missing data. (SC = skin conductance, ACC = accelerations of three axes, ST = skin temperature)

TABLE I
Feature sets for sleep detection

Source	Modality	Feature variables
Wrist sensor	Skin conductance (SC)	Mean, SD, power within 0-0.1, 0.1-0.2, 0.2-0.3, 0.3-0.4, and 0.4-0.5Hz bands, the number of SC responses, storm flag, elapsed time since a storm started
	Acceleration (ACC)	Mean, SD
	Skin temperature (ST)	Mean, SD
Phone	Screen	Screen was on, the time the screen was turned on
	SMS	Sent a message
	Call	On a call, missed a call
	Location	Movement index, connected to WiFi, connected to cellular nets
Time	Time	Elapsed minutes since 12:00 AM

periods of high frequency activity called "storms" during deep sleep [51]; we used algorithms developed to automatically detect storms in SC data [52]. Therefore, for the SC modality we computed mean, standard deviation, and frequency-domain features, including powers for five frequency bands (0–0.5 Hz, in 0.1 Hz intervals), and three storm features according to [53], including the number of SC responses, storm flag (whether we observe a storm in that minute), and the elapsed time since a storm started. Also, ST rises during sleep in individuals in living environments similar to those in our experiment [54]. Second, our phone app recorded time stamps both when phone users sent a SMS and also when they received a SMS. Since receiving a SMS is a passive behavior that could happen during sleep, we only kept SMS-sending events as a feature variable. Third, the raw location data acquired in our experiments were the latitude and longitude of phones, whose absolute numbers are nearly meaningless for sleep estimation across subjects. Hence, we developed a movement index for each minute, formulated as the arithmetic mean of the variances of the

latitude and the longitude, to indicate whether a user was actively changing location in that minute.

We had missing data lasting from a few minutes to several hours due to phone and sensor charging, and activities such as removal for a shower. We used a two-step strategy to solve this problem. First, a 25% missing tolerance threshold was applied to the wrist-worn sensor data: if any modalities had a missing rate higher than 25% within a day, the whole day's data were dropped. This rule was not adopted on the phone data for sporadic events such as sending a SMS, because we cannot discriminate if such events did not happen or were missed. Second, for the remaining days, we filled minutes with missing data using the average of the same feature variable over the remaining part of the same day. (We also tried linear interpolation for filling in missing data, but obtained consistently slightly worse performance.) After dealing with the missing data, we had 3439 days of data (average sleep onset time: 2:45AM (SD 1:49), sleep offset time: 9:50AM (SD 1:50), sleep duration: 7.3 hours (SD 1.8), sleep efficiency [32]: 95.7 (SD 5.1), sleep data:wake data=1:2.5). Note that the class ratio between sleep and wake is not extreme and we have abundant samples belonging to the minority class (wake), so we did not balance the two classes using oversampling or undersampling. To equalize features and help with gradient descent optimization, every feature variable was also normalized to the $[0, 1]$ range within each day.

Fig. 2 shows two ways to split the data for training and testing — by days or by participants. The latter ensures that a person who is in the test data is not in the training data. An LSTM trained this way is considered to be participant-independent. While such a model is less likely to perform as well as a participant-dependent model, its performance is more likely to reflect realistic future performance on new people whose data have not helped train the model.

The simplest way to make a participant-dependent model is shown in Fig. 2 (a) where the data are split by days. However, because sleep can overlap multiple days we want to be careful how this is done and not simply assign the days in randomized order. For example, if we assign two consecutive days to the training and test sets respectively, the first period of the second day will lose its past information. Therefore, we connected consecutive days of each subject into chunks, and then randomly cut the first or last 20% of each chunk as the test set. Our process resulted in 2772 days assigned to the training set, and 667 days to the test set (a roughly 80-20 split of the 3439 days). For the participant-independent model (splitting the data by participants), we randomly assigned 80% of the participants as the training set and the remaining 20% as the test set.

C. Sleep/wake detection

Our goal is to automatically classify every minute of data as sleep or wake, which is a binary sequential classification problem. We wish to use a model that exploits how current feature variables can depend on both past and future ones. For example, if a participant turned on her phone screen at 11:01pm, it would be highly likely that she was still awake at 11:00pm.

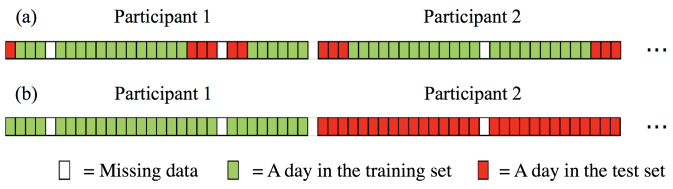

Fig. 2. Two ways to split the features and labels into a training set and a test set: (a) splitting by days, (b) splitting by participants.

Fig. 3 shows the structure of the bidirectional LSTM we used for sleep detection. The vector x_t contains all the features at time t, and $y_t \in [0, 1]$ is the estimated sleep probability for each minute. The activation function used in the fully-connected layer is rectified linear units. The bidirectional neural network was trained using RMSprop [55] with binary cross-entropy loss. The optimizer parameters were adopted from [56]. We set the past- and future-looking sequence lengths to 30 min each based on the work of Min et al. [28]. We also ran experiments to verify that 30 min is an appropriate choice, which were introduced in the supplementary Fig. A.1. The other hyper-parameters including the number of hidden units in LSTM, the number of LSTM layers, and the drop-out rate were tuned and selected on the training set using 5-fold cross-validation. The whole algorithm was implemented using deep learning frameworks Theano 0.8.2 and Keras 1.0.5.

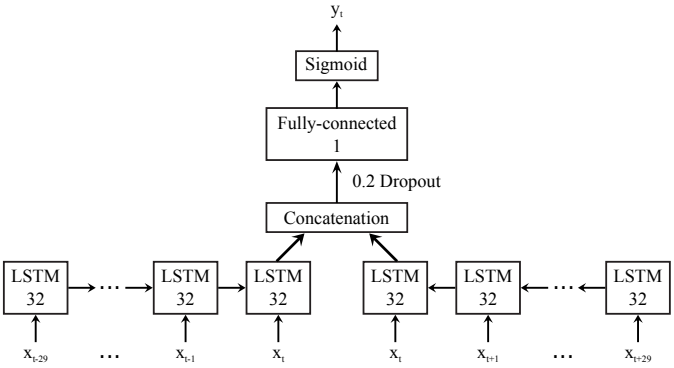

Fig. 3. Bidirectional LSTM model for sleep detection. The input x_t is the feature matrix at time t, and the output $y_t \in [0, 1]$ is the estimated sleep probability. Output dimensions are denoted in each box.

D. Sleep episode onset/offset estimation

After sleep/wake detection, we estimated sleep episode onset/offset points. In the proposed sleep detection model (Fig. 3), information from both the past and the future contribute to the estimation of sleep probabilities. A high sleep probability y_t can be achieved, only when the past feature matrices $\{x_{t-29}, \cdots, x_{t-1}, x_t\}$ and the future feature matrices $\{x_t, \cdots, x_{t+1}, x_{t+29}\}$ both show sleep patterns. Note that sleep patterns and awake patterns in feature matrices generally make opposite contributions to the final output of the model. Thus if we want to output sleep offset probability instead of sleep probability, an inverter can be inserted between the backward network and the concatenation layer, as shown in Fig. 4(a). In this way, a high y_t will be triggered when the

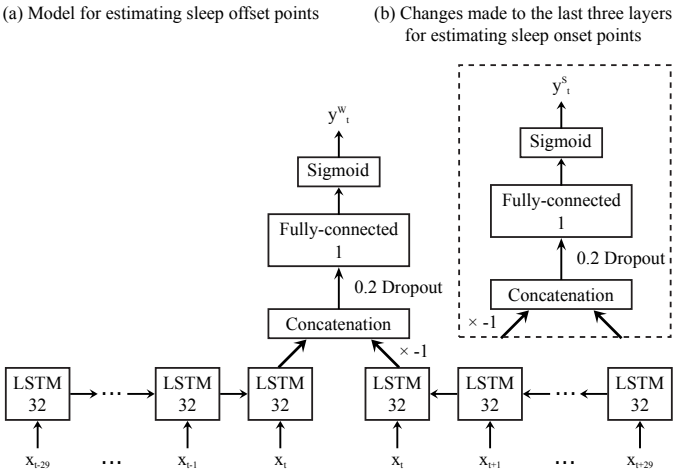

Fig. 4. Differential bidirectional LSTM model for sleep episode onset/offset estimation. The input x_t is the feature vector at time t, and $y_t^S, y_t^W \in [0, 1]$ are the estimated sleep episode onset/offset probabilities. Output dimensions are denoted in each box.

past features show sleep patterns and the future features show awake patterns. We call this new y_t, corresponding to waking up, y_t^W. Likewise, if an inverter is inserted at the output of the forward network (Fig. 4(b)), the final output will be sleep episode onset probability y_t^S indicating falling asleep.

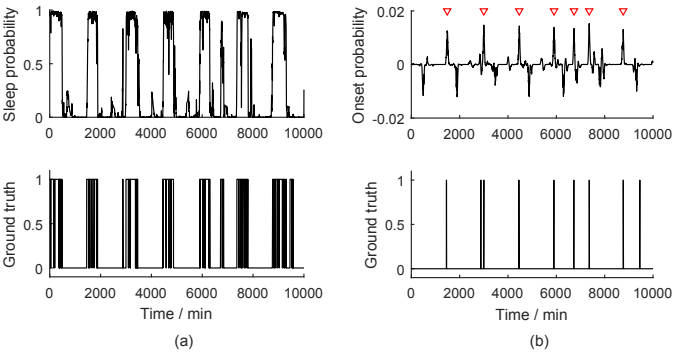

Fig. 5. (a) Exemplary sleep detection results and the corresponding ground truth. (b) Exemplary sleep episode onset detection results and the corresponding ground truth. The red triangles indicate the detected sleep onset points.

Fig. 5 (b) displays an exemplary output of the sleep onset estimation model (y_t^S), in which the peaks are sleep episode onset point candidates. To localize them, we applied the *findpeaks* function in MATLAB R2015b to the signal to find local maxima satisfying two conditions: First, to eliminate potential false positives, the height of a detected peak needs to be higher than a threshold, which was optimized on the training set towards a higher F_1 score (introduced below) and applied to the test data. Second, the distance between two peaks needs to be longer than 30 min. This rule was set to avoid false positives, since the shortest time interval between two neighboring sleep episode onset/offset points was 45 min in our data.

The detected peaks in y_t^S and y_t^W of the test set were then compared to the ground truth quantitatively. If the distance between a peak and its closest sleep episode onset/offset point in the ground truth was less than 30 min, then we defined the peak as a true positive. Based on this, we computed the precision, recall, and F_1 score (the harmonic mean of precision and recall) of our estimation. For all the true positive points, we also reported the average of their distances to the ground truth as the estimation errors.

IV. Experiments

In this paper, we conducted the following experiments to evaluate performance of the proposed algorithm.

A. Sleep detection and sleep episode onset/offset detection using the participant-dependent bidirectional LSTM model: Determine the best combinations of features

We compared performance of sleep detection and sleep episode onset/offset detection using the proposed bidirectional LSTM model and different combinations of features from a wearable sensor and a mobile phone. We compared (i) wrist sensor (ACC, EDA, ST), (ii) phone (screen, SMS, Call, location), (iii) wrist sensor + phone and (iv-ix) all combinations of wrist sensor features. We also compared performance using time feature.

B. Sleep detection: Compare the LSTM model to other machine learning models using the same participant-dependent setup

We compared the proposed temporal LSTM model and three other machine learning models: (i) a vanilla artificial neural network (ANN, a feed forward neural network with one hidden layer), (ii) a logistic regression model with ridge regularization, and (iii) a regularized linear support vector machine model (SVM).

As described in the Methods/feature preparation subsection, we split our data into training set (80%) and test data (20%). For ANN, we used the training data for optimizing the number of neurons and training the model (using 70/15/15 % of the training data for training, validation and testing respectively) and used the test data for testing the model. For Logistic Regression and SVM, we applied 10-fold cross validation to the training dataset for tuning hyper-parameters and training the models, and tested the models on the test data.

C. Sleep detection: Participant-dependent LSTM model vs existing fully automated sleep detection algorithm (Actigraphy software (Action4))

We compared our proposed LSTM model with an existing fully automated sleep/wake detection algorithms: Actigraphy software (Action4). The actigraphy device is bundled with their own software (Action4, AMI, USA) for sleep scoring. We computed the sleep/wake classification performance of the Action4 software by comparing sleep/wake output from the Action4 Software using actigraphy data (using ZCM and UCSD zero-crossing algorithm) (without diaries or human assessment) and the ground truth.

D. Sleep detection: participant-dependent LSTM model vs participant-independent LSTM model

We compared the performance of participant-dependent LSTM models and participant-independent LSTM models for sleep detection.

E. Bidirectional LSTM model vs real-time LSTM model (participant-dependent)

We compared the performance of bidirectional LSTM models that use both past and future 30-min sequences of data with LSTM models that use only the past 30 mins of data (real-time implementation, Fig. 6). We hypothesized that a model using both past and future data would have higher performance; however, a model that only requires the past data would be useful for real-time sleep detection. Thus, we compare if the performance from the two models differs significantly.

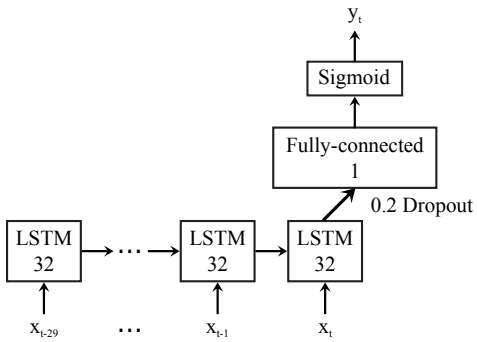

Fig. 6. Real-time implementation of the proposed sleep detection algorithm. The LSTM model only reads historical information. Output dimensions are denoted in each box.

F. Sleep episode onset/offset detection: template-matching-based method vs differential bidirectional LSTM model (participant-dependent)

In our previous paper [31], we proposed a sleep episode onset/offset detector using cross-correlation-based template matching. Specifically, we first computed the sleep probability patterns (for both sleep onset and offset) from the sleep detection results within the training set. We denoted these training set patterns as templates. The cross-correlation between the templates and sleep probabilities in the test set were then computed. The time points with the highest similarities to the templates were then labeled as sleep onsets or sleep offsets.

In this paper, we compared the performance of differential bidirectional LSTM based sleep episode onset/offset detection (Fig. 4) with the template-matching based algorithm.

V. Results

A. Sleep/wake detection using the participant-dependent bidirectional LSTM model: modality comparison

We summarize sleep detection performance of the proposed bidirectional LSTM model (participant-dependent) in Fig. 7 (Supplementary Material Table A.1), comparing two cases: whether the algorithm used the clock time as an input feature ("with time") or not ("without time"). This quantifies how much the algorithm is biased by sleep being more likely to occur at night. The detection accuracy of using only the time feature, 86.7 % is also shown in the figure as a baseline. Among all feature combinations, ACC + ST showed the best performance, 96.2% and 96.5% accuracy without and with time, respectively, and phone features showed the worst performance.

To show the relative importance of different features, we visualized the mean absolute weights of each feature across all connected nodes in the input layer of LSTM, as shown in Fig. 8. In the visualization, a higher weight indicates that the feature has a stronger influence sleep detection. In descending order, the most important feature are the standard deviation of acceleration, whether a phone call was missed, and the movement index calculated from the GPS data. Note that the estimated importance is affected by the collinearity of features. For example, we observed that the five frequency bands of EDA were sometimes correlated with each other, which dispersed their importance.

We also computed the percentages of main sleep and naps that were successfully detected. Our results showed that with ACC + ST and time features, 93% of epochs within main sleep were correctly classified (3% sleep was misrecognized as wake, 4% wake was misrecognized as sleep) and 65% of epochs within naps (33% sleep was misrecognized as wake, 2% wake was misrecognized as sleep) were successfully detected. As expected, given the irregular timing of naps, the performance in detecting naps was lower when using time features.

B. Sleep episode onset/offset detection using the participant-dependent differential bidirectional LSTM model: modality comparison

Table II shows a summary of sleep episode onset/offset detection using the differential bidirectional LSTM RNN. We obtained the best F_1 score and smallest mean absolute errors with ACC + ST features both for sleep episode onset/offset without and with time features. F_1 scores are 0.84–0.86 and mean absolute errors are 5.0–5.6 minutes. Sleep episode offset detection performed slightly worse than sleep episode onset detection. The details about the error distributions for each feature combination are shown in Supplementary Material Fig. A.2.

C. Sleep detection performance: the participant-dependent bidirectional LSTM models vs other machine learning models

Fig. 9 shows the results comparing participant-dependent LSTM to other models that did not include temporal information (neural networks, logistic regression, SVM). The LSTM models showed higher accuracy in sleep detection (see detailed numbers in Supplementary Material Table A.1). We applied the McNemar test and found that LSTM models with any combinations of features (both with time and without time) showed higher accuracy than neural networks, logistic regression and SVM models ($p < 0.05$).

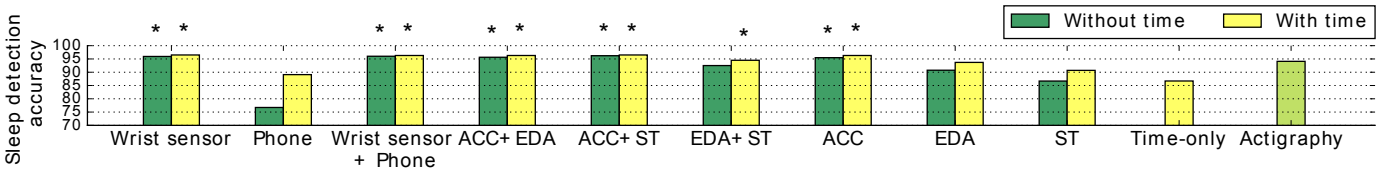

Fig. 7. Sleep detection accuracies using the proposed bidirectional LSTM model compared among feature combinations (participant-dependent models). Star marks indicate the performance of the LSTM model was statistically significantly better than that of actigraphy (the Action4 algorithm) (McNemar test).

TABLE II
SLEEP EPISODE ONSET/OFFSET DETECTION PERFORMANCE WITH BEST RESULTS HIGHLIGHTED IN BOLD (MAE: MEAN ABSOLUTE ERROR (MIN)).

Feature combinations	Differential Bi-LSTM RNN								Cross-correlation-based template matching							
	Without time				With time				Without time				With time			
	Sleep onset		Sleep offset		Sleep onset		Sleep offset		Sleep onset		Sleep offset		Sleep onset		Sleep offset	
	F_1	MAE	F_1	MAE	F_1	MAE	F_1	MAE	F_1	MAE	F_1	MAE	F_1	MAE	F_1	MAE
Wrist sensor	0.85	5.3	0.82	6.3	0.84	5.3	0.84	5.5	0.83	5.0	0.80	5.8	0.83	5.3	0.81	5.7
Phone	0.27	12.4	0.21	15.6	0.43	10.3	0.36	13.2	0.24	11.5	0.16	16.2	0.41	9.6	0.35	11.7
Wrist + Phone	0.84	5.3	0.82	6.0	0.84	5.1	0.82	6.3	0.83	**4.7**	0.80	5.8	0.83	5.7	0.80	6.0
ACC + EDA	0.84	5.3	0.83	6.4	0.84	5.1	0.83	6.0	0.83	5.1	0.80	6.3	0.84	5.4	0.81	5.9
ACC + ST	**0.86**	**5.0**	**0.84**	**5.6**	**0.86**	**5.0**	**0.84**	**5.5**	**0.84**	4.9	**0.81**	**5.1**	**0.85**	5.3	**0.82**	**5.5**
EDA + ST	0.74	6.9	0.73	7.2	0.74	6.7	0.74	7.1	0.70	7.3	0.69	7.3	0.73	6.9	0.71	6.3
ACC	0.85	5.4	0.81	6.9	0.84	5.1	0.81	6.5	0.82	5.2	0.79	7.2	0.82	**5.1**	0.79	6.3
EDA	0.71	8.0	0.68	7.7	0.70	6.4	0.70	7.1	0.67	8.1	0.64	8.1	0.70	6.8	0.68	7.3
ST	0.59	9.8	0.56	11.5	0.52	10.3	0.60	11.0	0.49	10.7	0.48	11.6	0.51	9.6	0.56	9.1

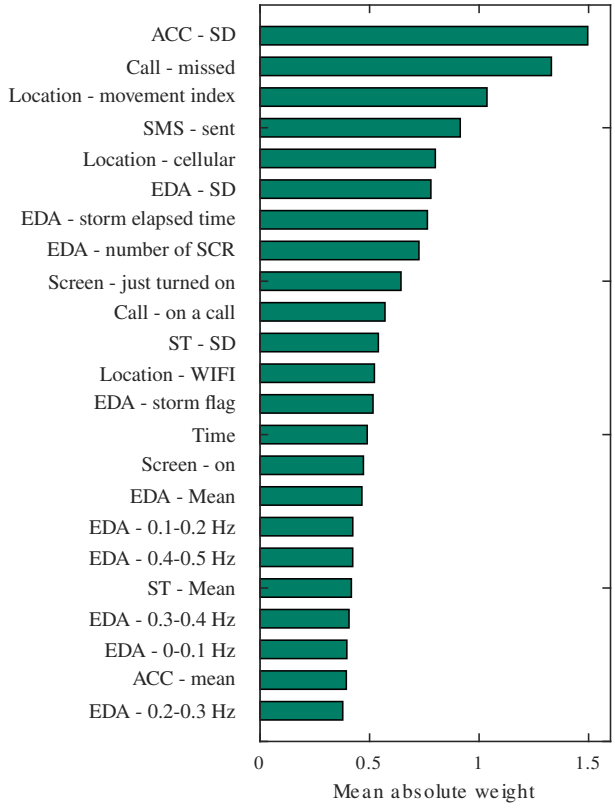

Fig. 8. Mean absolute weights of each feature across all connected nodes in the input layer of LSTM.

D. Sleep detection performance: the participant-dependent bidirectional LSTM vs the actigraphy (Action4) algorithm

We compared the performance between the LSTM models (participant-dependent) and the actigraphy (Action4) algorithm. We obtained accuracy 93.5% (SD: 5.1%) and F_1 score 0.94 (SD: 0.1) for the actigraphy algorithm. We tested if the proposed LSTM model performs better than the actigraphy algorithm using the McNemar test. Feature combinations such as ACC and ACC + ST showed statistically higher accuracy than the actigraphy algorithm (ACC: 95.5 % without time, 96.3 % with time, ACC + ST: 96.2 % without time, 96.5 % with time). The 3.2 % accuracy difference is equivalent to 46 mins per day (24 hours). Stars on the bars indicates that the result is significantly better than that of the actigraphy algorithm at the 5% significance level even after Bonferroni correction (Fig. 7). We also compared false positive rates for actigraphy and our algorithm. The false positive rate for actigraphy was 7.8 % and the one for LSTM with ACC data was 3.7 %; therefore our proposed method reduced false positives which are the main weakness of actigraphy.

E. Sleep detection performance: participant-dependent bidirectional LSTM models vs participant-independent bidirectional LSTM models

The models with data split by days and the models with data split by participants were very similar in terms of sleep detection accuracy, with very small differences (-0.4–0.5%: -6–7 mins per day(24 hours)). The detailed accuracy is shown in Supplementary Material Table A.3).

F. Sleep detection performance: the bidirectional LSTM models vs the realtime LSTM models

Fig. 10 shows the accuracy comparison between bidirectional LSTM models and realtime LSTM models (participant-dependent models), both also using time as a feature. The bidirectional models showed 0.2–1% higher accuracy then the realtime LSTM models (3–14 mins per day (24 hours), See

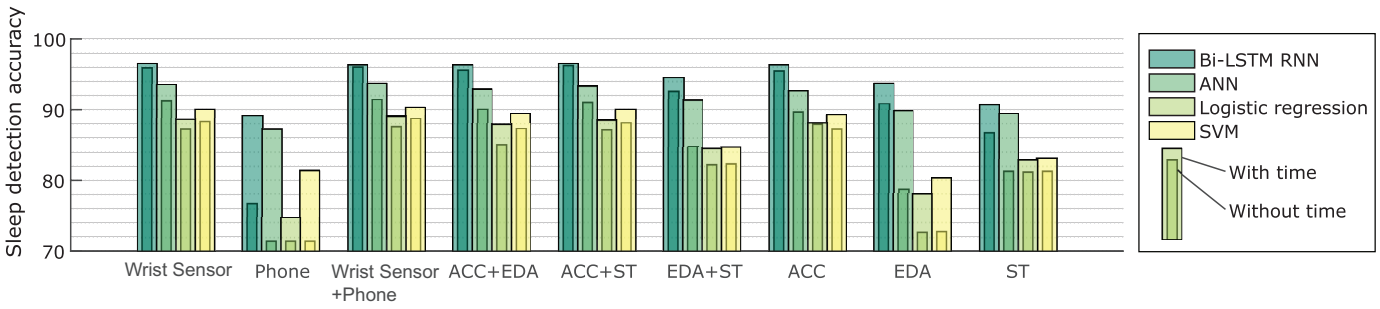

Fig. 9. Sleep detection accuracies compared among feature combinations and algorithms.

Supplementary Material Table A.3 for the detailed results). The McNemar test was conducted to examine if the bidirectional model was statistically superior to the realtime model. In all combinations of features (both with time and without time), the bidirectional models showed statistically higher performance than the realtime models ($p < 6.3 \times 10^{-23}$), confirming our hypothesis that it would be better.

G. Sleep episode onset/offset detection performance: Differential bidirectional LSTM model vs cross-correlation-based template matching method

Fig. 11 shows the results comparing sleep episode onset/offset detection using the LSTM model vs using cross-correlation-based template matching. Here, both methods used the ACC+ST+time feature combination, which allows us to compare the best performances for both. Red lines from the proposed LSTM models are above the black ones over almost the whole recall range, which indicates that the proposed LSTM-based sleep episode onset/offset detection performs better than the previous cross-correlation-based template matching method. Due to the way we defined a true positive in onset/offset detection, the precision-recall curves were undefined when recall was very close to 1. As a result, we were unable to calculate the area under curve (AUC) values for these curves.

VI. Discussion

Results show that statistically significantly better performance in sleep/wake detection can be achieved using LSTM models applied to wearable sensor and smartphone data than when using the Action4 software or non-temporal machine learning models. The LSTM models further showed superior performance in sleep episode onset/offset detection than our previously proposed cross-correlation based template matching method. When comparing features measured from a wrist sensor and a phone, the combination of ACC + ST + time features performed the best. The ST boosts the sleep detection performance along with the ACC. Previous studies have shown that ST increases during sleep (e.g. [57]). Our finding also replicates previous sleep detection work that showed that ST on the wrist coupled with motion is a strong discriminant for sleep/wake state [53]. Our algorithm showed the lowest performance with phone features among the comparisons of different modalities; however our results (sleep onset/offset detection MAE: 10.3 min and 13.2 min) still excelled previous work (errors in sleep onset: 44 min, offset: 42 min [28], sleep onset: 38 min and offset: 36 min [41]).

In order to confirm the generalizability of our algorithm, we also examined the relationship between participant-wise average sleep duration and sleep detection algorithm performance. Our participants had a range of average sleep duration: 5 hours 18 mins to 9 hours 19 mins). Participant-wise sleep detection accuracy was constant and no trend was found either in the relationship between average sleep duration vs sleep detection accuracy and between the distance to average sleep duration across participants and sleep detection accuracy (ACC+ST+time) (Figs. A.3).

To better understand the performance benefits of using temporal information, we analyzed errors made by the models that did not use the temporal information. We specifically consider the models that performed best, which are those that used the ACC+ST+time features. Fig. 12 shows the past and future 30 minutes of median values of the feature vectors (ACC mean, ST mean, and time) for the sleep and wake detection test data. We can see that the non-temporal neural network model made errors and only the LSTM model detected sleep/wake properly when the neighboring minutes showed different patterns from the detection point (ACC and ST mean, x=0).

Our results showed that the LSTM models trained and tested with data split by days and by participants showed very little difference in performance (-0.4–0.5%). These results show that the proposed LSTM models can be robust in real world person-independent applications. We also showed that the bidirectional LSTM models performed statistically significantly better than the real-time LSTM model in all feature combinations. However, the performance differences were small (0.2–1%). Thus, a real-time LSTM would probably give about the same performance as noticed by a typical wearer, but if it were important to maximize the accuracy for scientific or medical purposes, then a slightly slower but bidirectional LSTM works better.

In addition to using a McNemar test to compare the LSTM models and the actigraphy algorithm (Action4), we also examined participant-wise performance using paired t-test (Supplementary Material Table A.2). The results showed that feature combinations such as ACC and ACC + ST showed statistically higher accuracy than the actigraphy algorithm, as similarly observed in the McNemar test.

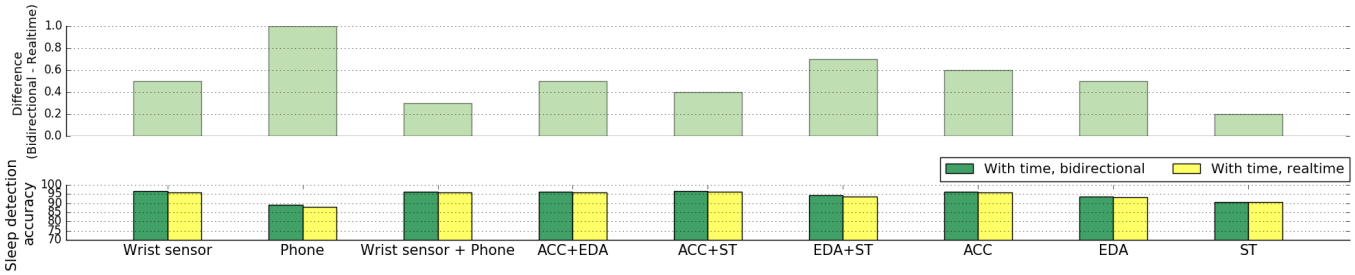

Fig. 10. Comparison of sleep detection performance between bidirectional LSTM models and realtime LSTM models (with time).

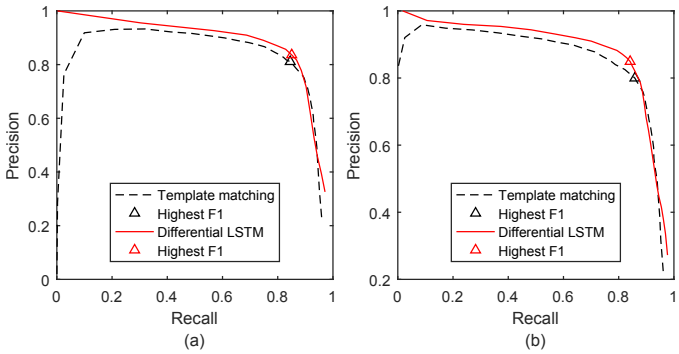

Fig. 11. Precision-recall curves for sleep episode onset detection (a) and offset detection (b) using the ACC + ST + time feature combination. The curves generated by cross-correlation-based template matching [31] and the new differential LSTM method are both shown for comparison.

Our work has several limitations. One is that the ground truth we used was provided by an experienced investigator using Harvard Medical School's standard operating procedures for reconciling diary and actigraphy data; this process is time consuming and does not always produce the same results as PSG when PSG and actigraphy are collected simultaneously. However, we believe this is more reliable than previously published ambulatory ground truth sources such as self reports alone or actigraphy alone. Another limitation arises from keeping participants comfortable: because two sensors were needed, they were worn on left and right wrists, so that the LSTM trained from data on one wrist was ultimately compared to ground truth data derived from the other wrist. We would expect that if it were possible to have worn both sensors on the same side, then the results of the LSTM would be even more accurate than they are here. Thus, the actigraphy algorithm was probably given a slight advantage in our study because its sensor was also used to produce the ground truth data. Furthermore, our phone data did not distinguish if we had missing data (e.g., phone battery ran out and phone was off), or if participants did not use their phone or were not carrying their phone. Although we asked our participants to charge their phone every day, our data might include some time when a phone was not operating. Lastly, our LSTM models could take more time to train compared to simple logistic regression or SVM models, though the gap can be minimized by using a modern GPU. We have shown that our algorithm can be generalized to participants with different sleep habits, so retraining the LSTM models could be avoided to the greatest extent possible.

For future work, we need to test the proposed algorithm in other populations, including people who do not intensively use their mobile phones, people with medical conditions and/or on medications that may affect the Q-sensor data, and people whose primary wake episode is not during the day (e.g., night or shift workers). While here we describe a general model for all users, we could also build personalized models with individual data or keep updating a model while capturing daily sleep data. In this way, the model performance could improve especially for irregular sleepers, shift workers or frequent travelers. Furthermore, sleep/wake detection performance when using only the phone data might further improve if our phone application was modified to collect acceleration, audio and ambient light. Finally, compared with implicit knowledge learned in LSTM, RNNs with explicit memory [58], [59] can memorize facts and common sense, which are potentially useful for the task of sleep detection. We plan to adapt these algorithms to our sleep data in the future.

VII. Conclusion

We presented the design and evaluation of a novel algorithm for long-term ambulatory sleep/wake detection that utilizes data from smartphones and a wrist-worn sensor. The algorithm uses an LSTM RNN model to assign a sleep probability to each 1-minute epoch to detect sleep/wake state and sleep onset/offset episodes. The novel method achieved a sleep/wake classification accuracy of 96.5%, and F_1 scores of 0.86 and 0.84 for sleep episode onset and offset detection, with mean absolute errors of 5.0 and 5.5 mins respectively, using the acceleration, skin temperature and time data.

We evaluated the LSTM algorithm also by comparing it with other machine learning models (neural networks, logistic regression, SVM) and with a commercial actigraphy software algorithm and showed that the LSTM algorithm was statistically significantly superior to these others. We also assessed the generalizability of the method by training and testing the model with data from two randomly separated groups of participants, and found that the accuracy achieved was person-independent. We also confirmed that our sleep detection algorithm performed stably across participants with different sleep duration. Future studies should continue to collect a wider variety of data with more types of sleepers, and see if the performance shown here from using the temporal LSTM method continues to hold.

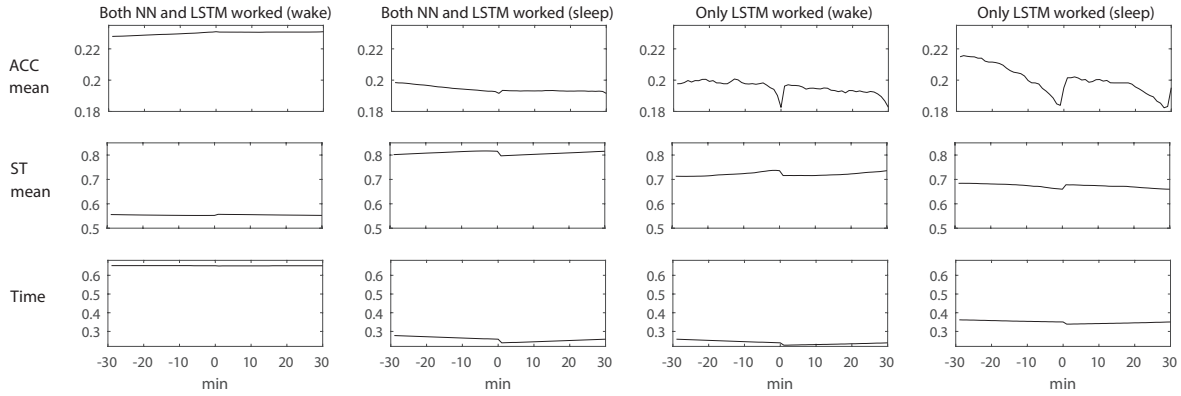

Fig. 12. An example of features where both neural network (NN) model and LSTM model worked and only LSTM model worked and neural network model made errors. Only LSTM models detected sleep and wake correctly when the neighboring minutes showed different patterns from the detection point (x=0). ACC = acceleration, ST = skin temperature.

ACKNOWLEDGMENT

This work was supported by NIH (R01GM105018), Samsung Electronics, NEC, and MIT Media Lab Consortium. The authors also would like to thank SNAPSHOT study project members and participants (https://snapshot.media.mit.edu/).

REFERENCES

[1] C. A. Czeisler, "Duration, timing and quality of sleep are each vital for health, performance and safety," *Sleep Health*, vol. 1, no. 1, pp. 5–8, 3 2015.

[2] R. R. Markwald *et al.*, "Impact of insufficient sleep on total daily energy expenditure, food intake, and weight gain," *Proceedings of the National Academy of Sciences*, vol. 110, no. 14, pp. 5695–5700, 2013.

[3] E. Van Cauter *et al.*, "Metabolic consequences of sleep and sleep loss," *Sleep Medicine*, vol. 9, no. 10, pp. S23–S28, 9 2008.

[4] S. Cohen *et al.*, "Sleep Habits and Susceptibility to the Common Cold," *Archives of Internal Medicine*, vol. 169, no. 1, p. 62, 2009.

[5] D. F. Dinges *et al.*, "Cumulative Sleepiness, Mood Disturbance, and Psychomotor Vigilance Performance Decrements During a Week of Sleep Restricted to 4-5 Hours per Night," *Sleep*, vol. 20, no. 4, pp. 267–77, 4 1997.

[6] H. O. Lisper and A. Kjellberg, "Effects of 24-hour sleep deprivation on rate of decrement in a 10-minute auditory reaction time task." *Journal of Experimental Psychology*, vol. 96, no. 2, pp. 287–290, 1972.

[7] S. M. Doran *et al.*, "Sustained attention performance during sleep deprivation: Evidence of state instability," *Archives Italiennes de Biologie*, vol. 139, no. 3, pp. 253–267, 2001.

[8] J. Lim and D. F. Dinges, "Sleep Deprivation and Vigilant Attention," *Annals of the New York Academy of Sciences*, vol. 1129, no. 1, pp. 305–322, 5 2008.

[9] R. Stickgold *et al.*, "Visual discrimination learning requires sleep after training," *Nature Neuroscience*, vol. 3, no. 12, pp. 1237–1238, 12 2000.

[10] M. M. Ohayon and C. Guilleminault, "Epidemiology of Sleep Disorders," in *Sleep: A Comprehensive Handbook*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 11 2005, pp. 73–82.

[11] M. A. Grandner and A. Malhotra, "Sleep as a vital sign: why medical practitioners need to routinely ask their patients about sleep," *Sleep Health*, vol. 1, no. 1, pp. 11–12, 3 2015.

[12] R. B. Berry *et al.*, "The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology, and Technical Specifications, Version 2.2," in *American Academy of Sleep*, 2016, vol. 28, no. 3.

[13] F. Gagnadoux *et al.*, "Home Unattended vs Hospital Telemonitored Polysomnography in Suspected Obstructive Sleep Apnea Syndrome," *Chest*, vol. 121, no. 3, pp. 753–758, 3 2002.

[14] V. Natale *et al.*, "Monitoring sleep with a smartphone accelerometer," *Sleep and Biological Rhythms*, vol. 10, no. 4, pp. 287–292, 2012.

[15] R. Soltanzadeh and Z. Moussavi, "Sleep Stage Detection Using Tracheal Breathing Sounds: A Pilot Study," *Annals of Biomedical Engineering*, vol. 43, no. 10, pp. 2530–2537, 10 2015.

[16] E. Dafna *et al.*, "Automatic Detection of Whole Night Snoring Events Using Non-Contact Microphone," *PLoS ONE*, vol. 8, no. 12, p. e84139, 12 2013.

[17] T. Rahman *et al.*, "DoppleSleep: A Contactless Unobtrusive Sleep Sensing System Using Short-Range Doppler Radar," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '15*. New York, New York, USA: ACM Press, 2015, pp. 39–50.

[18] J. Liu *et al.*, "Tracking Vital Signs During Sleep Leveraging Off-the-shelf WiFi," in *Proceedings of the 16th ACM International Symposium on Mobile Ad Hoc Networking and Computing - MobiHoc '15*. New York, New York, USA: ACM Press, 2015, pp. 267–276.

[19] S. Ancoli-Israel *et al.*, "The Role of Actigraphy in the Study of Sleep and Circadian Rhythms," *Sleep*, vol. 26, no. 3, pp. 342–392, 5 2003.

[20] A. Sadeh and C. Acebo, "The role of actigraphy in sleep medicine," *Sleep Medicine Reviews*, vol. 6, no. 2, pp. 113–124, 5 2002.

[21] J. F. Sassin and L. C. Johnson, "Body motility during sleep and its relation to the K-complex," *Experimental Neurology*, vol. 22, no. 1, pp. 133–144, 9 1968.

[22] S. Gori *et al.*, "Body movements during night sleep in healthy elderly subjects and their relationships with sleep stages," *Brain Research Bulletin*, vol. 63, no. 5, pp. 393–397, 6 2004.

[23] C. E. Carney *et al.*, "The Consensus Sleep Diary: Standardizing Prospective Sleep Self-Monitoring," *Sleep*, vol. 35, no. 2, pp. 287–302, 2012.

[24] D. S. Lauderdale *et al.*, "Objectively Measured Sleep Characteristics among Early-Middle-Aged Adults: The CARDIA Study," *American Journal of Epidemiology*, vol. 164, no. 1, pp. 5–16, 6 2006.

[25] L. K. Barger *et al.*, "Prevalence of sleep deficiency and use of hypnotic drugs in astronauts before, during, and after spaceflight: an observational study," *The Lancet Neurology*, vol. 13, no. 9, pp. 904–912, sep 2014.

[26] R. J. Cole *et al.*, "Automatic sleep/wake identification from wrist activity." *Sleep*, vol. 15, no. 5, pp. 461–469, 1992.

[27] Z. Chen *et al.*, "Unobtrusive Sleep Monitoring using Smartphones," in *Proceedings of the ICTs for improving Patients Rehabilitation Research Techniques*. IEEE, 2013.

[28] J.-K. Min *et al.*, "Toss 'n' turn," in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*. New York, New York, USA: ACM Press, 2014, pp. 477–486.

[29] W. Karlen, "Adaptive wake and sleep detection for wearable systems," Ph.D. dissertation, EPFL, 2009.

[30] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural computation*, vol. 9, no. 8, pp. 1735–80, 1997.

[31] W. Chen *et al.*, "Multimodal ambulatory sleep detection," in *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2017, pp. 465–468.

[32] D. L. Reed and W. P. Sacco, "Measuring Sleep Efficiency: What Should the Denominator Be?" *Journal of Clinical Sleep Medicine*, vol. 12, no. 02, pp. 263–266, 2 2016.

[33] J. Lee and J. Finkelstein, "Consumer Sleep Tracking Devices: A Critical Review," *Studies in Health Technology and Informatics*, vol. 210, pp. 458–460, 2015.

[34] L. J. Meltzer *et al.*, "Comparison of a Commercial Accelerometer with Polysomnography and Actigraphy in Children and Adolescents," *Sleep*, 2015.
[35] C. A. Kushida *et al.*, "Comparison of actigraphic, polysomnographic, and subjective assessment of sleep paramenters in sleep disordered patients," *Sleep Medicine*, vol. 2, pp. 389–396, 2001.
[36] T. Arora *et al.*, "An Investigation into the Strength of the Association and Agreement Levels between Subjective and Objective Sleep Duration in Adolescents," *PLoS ONE*, vol. 8, no. 8, p. e72406, 8 2013.
[37] P. J. Hauri and J. Wisbey, "Wrist Actigraphy in Insomnia," *Sleep*, vol. 15, no. 4, pp. 293–301, 7 1992.
[38] K. Reid and D. Dawson, "Correlation Between Wrist Activity Monitor and Electrophysiological Measures of Sleep in a Simulated Shiftwork Environment for Younger and Older Subjects," *Sleep*, vol. 22, no. 3, pp. 378–385, 5 1999.
[39] A. Sadeh, "The role and validity of actigraphy in sleep medicine: An update," *Sleep Medicine Reviews*, vol. 15, no. 4, pp. 259–267, 8 2011.
[40] A. Sadeh *et al.*, "Actigraphically-based automatic bedtime sleep-wake scoring: Validity and clinical applications," *Journal of Ambulatory Monitoring*, vol. 2, no. 3, pp. 209–216, 1989.
[41] S. Saeb *et al.*, "Scalable Passive Sleep Monitoring Using Mobile Phones: Opportunities and Obstacles," *Journal of Medical Internet Research*, vol. 19, no. 4, p. e118, apr 2017.
[42] S. H. Schmidhuber and J., "Long short-term memory," *Neural Computation*, 1997.
[43] A. Graves and N. Jaitly, "Towards End-To-End Speech Recognition with Recurrent Neural Networks," in *JMLR Workshop and Conference Proceedings*, vol. 32, no. 1, 2014, pp. 1764–1772.
[44] I. Sutskever *et al.*, "Sequence to Sequence Learning with Neural Networks," in *NIPS*, 2014, pp. 3104–3112.
[45] K. Cho *et al.*, "On the Properties of Neural Machine Translation : Encoder Decoder Approaches," *Ssst-2014*, pp. 103–111, 2014.
[46] A. Sathyanarayana *et al.*, "Sleep Quality Prediction From Wearable Data Using Deep Learning," *JMIR mHealth and uHealth*, vol. 4, no. 4, p. e125, 11 2016.
[47] H. Dong *et al.*, "Mixed Neural Network Approach for Temporal Sleep Stage Classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, pp. 1–1, 2017.
[48] E. P. Giri *et al.*, "Combining Generative and Discriminative Neural Networks for Sleep Stages Classification," pp. 1–13, 10 2016.
[49] N. Aharony *et al.*, "Social fMRI: Investigating and shaping social mechanisms in the real world," in *Pervasive and Mobile Computing*, vol. 7, no. 6, 2011, pp. 643–659.
[50] A. A. o. S. Medicine, "The AASM Manual for the scoring of sleep and associated events," *Am. Acad. Sleep Med.*
[51] K. Asahina and K. Omura, "Phenomenological Study of Paradocial Phase and Reverse Paradoxical Phase of Sleep," *The Japanese Journal of Physiology*, vol. 14, no. 4, pp. 365–372, 1964.
[52] A. Sano *et al.*, "Quantitative analysis of wrist electrodermal activity during sleep," *International Journal of Psychophysiology*, vol. 94, no. 3, pp. 382–389, dec 2014.
[53] A. Sano and R. W. Picard, "Comparison of sleep-wake classification using electroencephalogram and wrist-worn multi-modal sensor data," *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, vol. 2014, pp. 930–933, 2014.
[54] J. A. Sarabia *et al.*, "Circadian rhythm of wrist temperature in normal-living subjects. A candidate of new index of the circadian system," *Physiology and Behavior*, vol. 95, no. 4, pp. 570–580, 2008.
[55] T. Tieleman and G. Hinton, "Lecture 6.5 - RMSProp," *COURSERA: Neural Networks for Machine Learning. Technical report*, 2012.
[56] G. Hinton *et al.*, "Rmsprop: Divide the gradient by a running average of its recent magnitude," *Neural networks for machine learning, Coursera lecture 6e*, 2012.
[57] A. Martinez-Nicolas *et al.*, "Uncovering different masking factors on wrist skin temperature rhythm in free-living subjects." *PloS one*, vol. 8, no. 4, p. e61142, jan 2013.
[58] S. Sukhbaatar *et al.*, "End-to-end memory networks," in *Advances in neural information processing systems*, 2015, pp. 2440–2448.
[59] A. Graves *et al.*, "Neural turing machines," *arXiv preprint arXiv:1410.5401*, 2014.

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

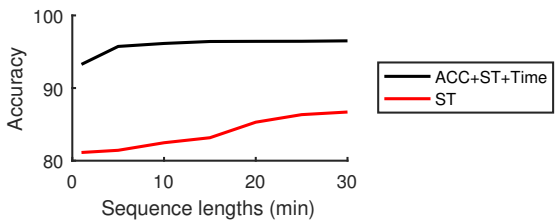

Fig. A.1. Sleep detection accuracy versus the past- and future-looking sequence lengths using two different feature combinations. The performance of sleep detection using the more informative feature combination (ACC+ST+Time) saturates when the sequence length is longer than 15 min, while the performance 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.

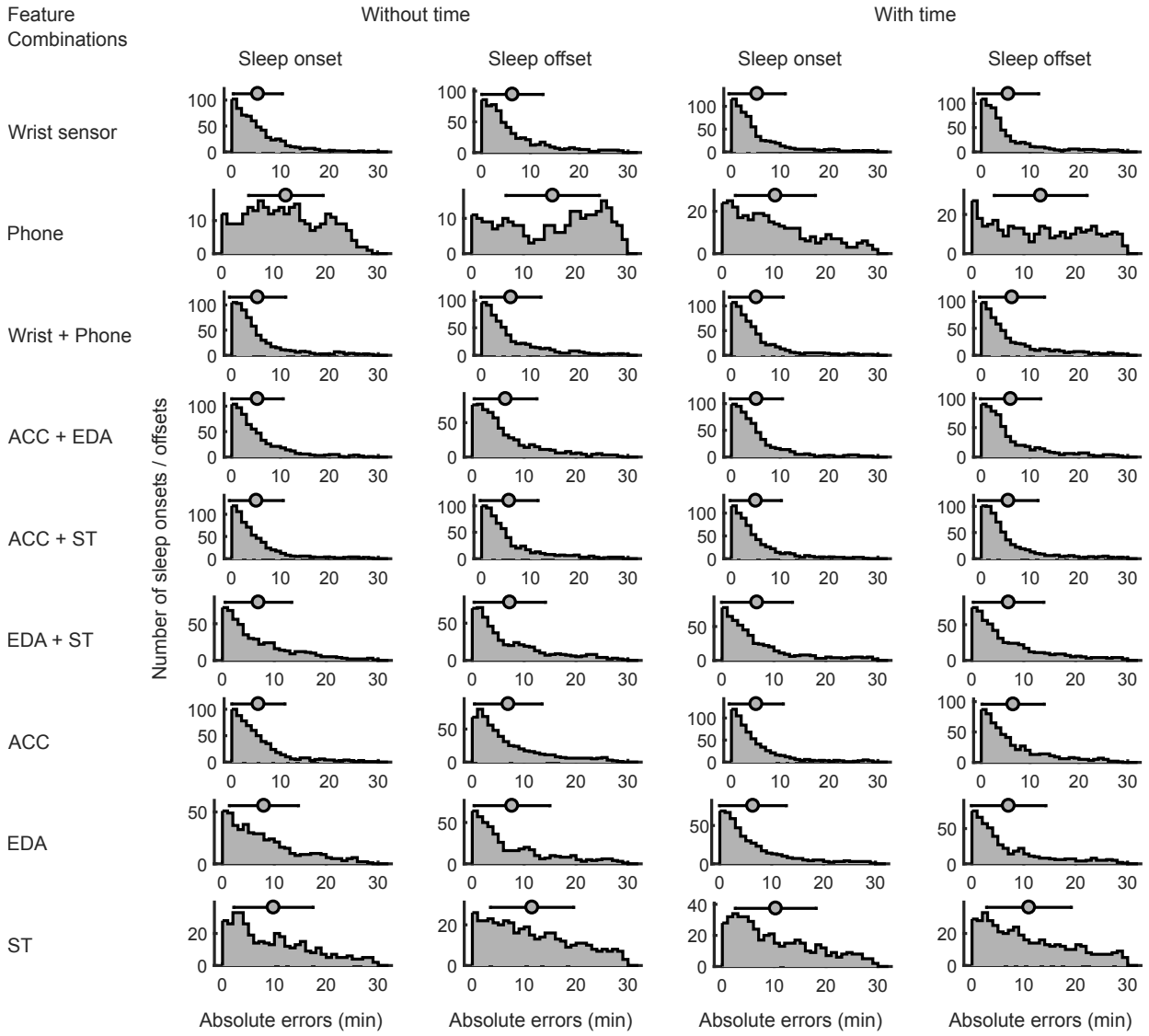

Fig. A.2. Error distributions for sleep episode onset/offset detection using differential Bi-LSTM RNN. The means and standard deviations of the absolute errors are also denoted in each plot.

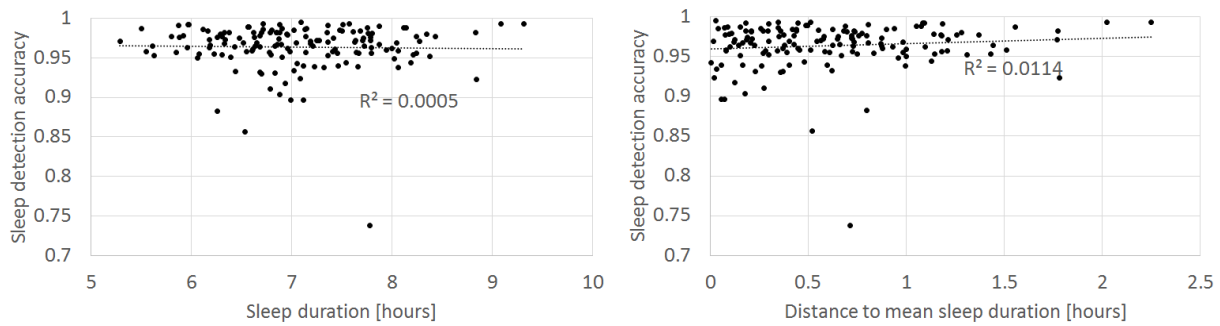

Fig. A.3. (left) Relationship between average sleep duration vs sleep detection accuracy (ACC+ST+time).(right)Relationship between distance to average sleep duration vs sleep detection accuracy (ACC+ST+time).