An Attention Based CNN-LSTM Approach for Sleep-Wake Detection with Heterogeneous Sensors

Zhenghua Chen, Min Wu, Wei Cui, Chengyu Liu, *Senior Member, IEEE* and Xiaoli Li, *Senior Member, IEEE*

***Abstract*—In this paper, we propose an attention based convolutional neural network long short-term memory (CNN-LSTM) approach for sleep-wake detection with *heterogeneous* sensor data, i.e., acceleration and heart rate variability (HRV). Since the three-dimensional acceleration data was sampled with a high frequency, we firstly design a CNN-LSTM structure to effectively learn latent features from the acceleration. Meanwhile, considering the unique format of the HRV data, some effective features are extracted based on domain knowledge. Next, we design a unified architecture to efficiently merge the features learned by CNN-LSTM approach from the acceleration and the extracted features from the HRV, which enables us to make full use of all the available information from these two heterogeneous sources. Taking into consideration that these two heterogeneous sources may have *distinct contributions* for the sleep and wake states, we propose an *attention* network to dynamically adjust the importance of features from the two sources. Real-world experiments have been conducted to verify the effectiveness of the proposed approach for sleep-wake detection. The results demonstrate that the proposed method outperforms all existing approaches for sleep-wake classification. In the evaluation of leave-one-subject-out (LOSO) cross-validation which is more challenging and practical, the proposed method achieves remarkable improvements ranging from 5% to 46% over the benchmark approaches.**

***Index Terms*—Sleep-wake detection, CNN-LSTM, attention, acceleration, HRV.**

I. Introduction

SLEEP is a critical physiological function for human as it affects both physical and mental health. Inadequate sleep increases the risk of heart disease, stoke and type 2 diabetes. Mental health issues, such as depression, are also strongly linked to poor sleep quality. Therefore, it is highly desirable to identify the sleep quality and duration through sleep monitoring and sleep-wake detection.

Polysomnography (PSG) is the gold standard for sleep stage detection and sleep quality measurement [1]. Based on its electroencephalography (EEG) data, sleep specialists or even computational approaches can distinguish different sleep stages [2]. Recently, researchers demonstrate that they are able to accurately recognition multiple sleep stages, such as wake, Rapid Eye Movement (REM), 3 non-REM stages N1, N2 and N3, with EEG data [3], [4]. However, PSG (or even EEG only) is considered to be *costly*, *labour-intensive* and *invasive* for sleep monitoring and thus is not feasible to be widely used in daily life and long-term monitoring applications.

Wearable sensors [5] which are cost effective and easy-to-use have become popular for long-term sleep tracking [6], [7], as they can easily collect different types of data from human body, e.g., acceleration, respiratory, electrocardiogram (ECG), heart rate variability (HRV), etc. Since these sensors can only obtain movement or heart rate related information, they are hard to detect multiple sleep stages. Instead, they can be good indicators for detecting sleep/wake states. And long-term monitoring of sleep/wake states is also crucial [7], [8]. How to accurately detect sleep/wake states by only using some low-cost and easy-to-use wearable sensors is attracting great attention recently.

With the data collected by wearable sensors, various traditional machine learning methods can be employed for the classification of sleep and wake states, e.g., linear discriminant (LD) classifier [9], [10], support vector machines (SVM) [11], [12], decision tree (DT) [13], random forest (RF) [14] and artificial neural network (ANN) [13], [15]. However, it is compulsory to extract representative features from complex sensor measurements before applying these machine learning methods for sleep-wake classification, while this feature extraction requires strong domain knowledge [16].

Recently, deep learning, which is capable of learning features, has achieved great successes in many challenging applications, such as image classification [17], natural language understanding [18], and time series prediction [19]. It also has been widely used for biomedical applications [20], [21], [22]. Several deep learning algorithms have also been proposed for sleep-wake classification based on wearable sensor data. In [23], the authors adopted the convolutional neural network (CNN) to identify sleep and wake states on two public datasets. Chen et al. presented a bidirectional long short-term memory (Bi-LSTM) approach for sleep-wake detection with multimodal data, such as skin temperature, skin conductance and acceleration.

Here, we adopt the wearable sensor data of acceleration and HRV for accurate sleep-wake classification. Both acceleration and HRV data have been shown to be effective for sleep-wake classification. Meanwhile, they can be treated as two different indicators, i.e., physical and physiological, for the detection of sleep and wake states. Therefore, the combination of these two types of data is expected to boost the performance of sleep-wake detection. There are some key technical challenges in this problem. Firstly, acceleration and HRV are considered as heterogeneous sensor streams, as they have different nature and format. In particular, the acceleration has a high sampling

Zhenghua Chen, Min Wu, Wei Cui and Xiaoli Li are with the Institute for Infocomm Research, A*STAR, 1 Fusionopolis Way #21-01 Connexis, Singapore 138632. Min Wu is the corresponding author. (e-mail: chen0832@e.ntu.edu.sg; wumin@i2r.a-star.edu.sg; cuiw@i2r.a-star.edu.sg; xlli@i2r.a-star.edu.sg).

Chengyu Liu is with School of Instrument Science and Engineering, Southeast University, Nanjing, 210096, China (e-mail: chengyu@seu.edu.cn).

rate, meaning that the sequence is uniform, but very long. The HRV data contains values of R-R intervals, which is typical non-uniform data. It is challenging to effectively integrate the two heterogeneous sources for the detection of sleep and wake states.

Secondly, sensor streams are typical time series and thus LSTM based methods with strong sequential modeling capability are naturally suitable. However, it is infeasible to train LSTM on the raw acceleration data with high sampling frequency, with the general constraints on the memory and computational power.

Finally, the acceleration and HRV may have distinct contributions for sleep-wake classification. It is thus very important to design a method which is able to dynamically adjust the significance of features from these two heterogeneous sources within the deep learning framework.

To effectively address the above challenges, we propose a novel attention based CNN-LSTM approach for sleep-wake classification with acceleration and HRV data. First, we design a CNN to learn sequential local features on the acceleration streams. Then, we adopt LSTM to encode the temporal dependencies of the learned local features and further learn high-level representations. Such a CNN-LSTM integration framework can automatically learn features from the raw acceleration data (high sampling frequency). Second, we extract features (e.g., features in time domain and frequency domain) for HRV data based on domain knowledge. We then design a unified architecture to integrate the features from both the acceleration and HRV. Lastly, to boost the performance of sleep-wake detection, we develop an attention network to dynamically adjust the importance of the features from the two different sources. Real experiments have been conducted to verify the effectiveness of the proposed method for sleep-wake detection.

The main contributions of this work are summarized as follows:

- We propose a novel unified deep learning framework for sleep-wake detection by combining two heterogeneous sensors, i.e., acceleration and HRV, with different properties and formats.
- We develop an innovative CNN-LSTM structure to effectively learn latent features from long acceleration sequences, which cannot be directly handled by existing LSTM based methods.
- Considering that the two heterogeneous sensors may have distinct contributions for classifying the sleep and wake states, we design an attention network to dynamically adjust the importance of features from the two heterogeneous sensors to boost the performance for sleep-wake detection.
- We perform extensive experiments to evaluate the effectiveness of the proposed approaches. The results show that the proposed approaches outperform all benchmark approaches.

This paper is a little bit similar to our previous work in [24]. However, the differences are quite obvious. We summarize the main differences as follows: 1) The work in [24] requires to extract local features from acceleration and combines local features with LSTM network, which is a labor-intensive and tedious process. In this paper, we design a CNN-LSTM network which is an end-to-end architecture for automatic feature learning from acceleration without human intervention. 2) The work in [24] does not adjust the weights for two types of features, i.e., features from acceleration and HRV, which may have different contributions for sleep-wake detection. In this paper, we develop an innovative attention network to dynamically adjust the weights of the two types of features, which has been shown to be effective.

II. Related Works

In this section, we review some existing algorithms for sleep-wake classification. The algorithms can be divided into shallow models and deep models.

A. Shallow Models

For shallow models, they generally consist of two steps: 1) feature extraction from sensory data, and 2) sleep-wake classification by applying traditional machine learning algorithms.

For instance, in [9], various features were extracted from actigraphy, ECG and respiratory data, and then a linear discriminant (LD) classifier was employed for sleep-wake classification. Similarly, the LD classifier was also used for the same purpose on dynamic frequency warping (DFW) features which were extracted from actigraphy and respiratory data [10]. Power spectral density scores were extracted from ECG and respiratory signals by using the Fast Fourier Transform (FFT), followed by an ANN model for classifying sleep and wake states [15]. In [11], firstly, HRV data was extracted from the raw ECG, and then various features from time domain, frequency domain and detrended fluctuation analysis (DFA) were extracted from the HRV data. After that, the classifier of SVM was then employed for sleep-wake classification and sleep efficiency estimation. Note that feature extraction from the raw sensory data is an essential step in shallow models, which usually requires expert knowledge and may inevitably miss some implicit useful features, leading to an unsatisfactory performance.

B. Deep Models

Currently, deep learning methods have achieved great successes for healthcare and biomedical applications [20], [21]. For sleep analysis, many algorithms have been proposed for automatic sleep staging based on the EEG data. For example, the CNN was presented to work on single-channel EEG [25]. SLEEPNET [26] and DeepSleepNet [4] leveraged on the LSTM model for sleep stage classification based on the EEG data. Meanwhile, several deep learning algorithms have also been proposed for sleep-wake classification based on wearable sensor data. Phan et al. applied the CNN to identify sleep and wake states with the actigraphy data [23]. In [27], a Bi-LSTM model performed very well for sleep-wake classification with multimodal data, i.e., skin temperature, skin conductance, and acceleration. Note that the sensor sampling rate is relatively low in [27] and it will be infeasible to learn the Bi-LSTM

model if the sensors are with very high sampling frequency. Our previous work in [24] proposed a local feature based LSTM (LF-LSTM) method to extract features from acceleration and developed a fusion framwork to combine features from acceleration and HRV for sleep-wake classification.

Existing studies that either adopt one sensor or simply combine several sensors (equal importance) using conventional machine learning or deep learning methods have limited performance for sleep-wake classification. In this study, we design an innovative deep learning framework for sleep-wake classification with wearable sensor streams, including acceleration and HRV data. The proposed method can effective learn features from these two heterogeneous sensors and dynamically adjust the importance of features from the two sensors to boost the performance of sleep-wake classification.

III. Methodology

A. Feature Learning on Acceleration via CNN-LSTM

The three-dimensional acceleration data was collected by using a FAROS sensor with a sampling rate of 100 Hz in this work. Due to the varying orientations of the sensor in use, we also include the magnitude of the acceleration as the fourth dimension to overcome the issue of orientation changing. To segment the data for model learning, we use a sliding window with a window size of 5 minutes which is widely used [28], [29]. Thus, the sample size of a 5-min segment/window is $30,000 \times 4$. As aforementioned, it is compulsory for conventional shallow models to extract informative features from the acceleration in each window based on strong domain knowledge which may not be available all the time. Besides, the feature extraction will inevitably miss some useful and implicit features and thus limit the sleep-wake classification performance.

Recently, deep learning has achieved great successes in many challenging areas and the biggest merit of deep learning is the ability of automatic feature learning from data. Therefore, it can be adopted for feature learning upon the acceleration. Owing to the sequential property of acceleration, recurrent neural network (RNN) is naturally suitable for this task. However, the traditional RNN may suffer from the issue of gradient vanishing or exploding, resulting a limited performance for long-term dependencies. To solve this issue, the LSTM network which intends to use some gates to control the information has been developed in [30]. It has been successfully used in many applications with time series sensor data. For example, it has been explored for human activity recognition with inertial sensors [31], [32], [33]. In addition, it has also been used for acoustic novelty detection with acoustic sensors [34] and occupancy detection with environmental sensors [35].

As mentioned above, each sample has a size of $30,000 \times 4$ in our experiments. If the LSTM network is directly used to learn features on this acceleration sequence which is extremely long, we need to use $30,000$ LSTM cells to learn features. It is computationally infeasible with the general constrains on memory and computational capability. To address this issue, we firstly design a 1D-CNN to learn local features on each sensor dimension. Assume a sample $\mathbf{X} = \{\mathbf{x}^i\}, i \in \{1,2,3,4\}^\top, \mathbf{x}^i \in \mathbb{R}^{L\times 1}$ where $\top$ is the transpose operation and $L = 30,000$ in this work, the 1D convolutional operation on each sensor dimension for the r-th filter can be expressed as

$$\mathbf{c}_l^r = f^r(\mathbf{x}_{l:l+s-1}^i * \mathbf{w}^r) + b^r,\ l \in \{1,2,...,L-s+1\} \quad (1)$$

where s is the filter size, $f(\cdot)$ is the activation function, $\mathbf{w}$ is the weight vector, and b is the bias. Then, the output of the 1D convolutional operation can be expressed as $\mathbf{C} \in \mathbb{R}^{(L-s+1)\times 4}$. To get more compact representations, a pooling operation can be adopted. Here, we apply 1D max-pooling on the outputs of 1D convolutional operation. The output of the 1D max-pooling operation is to take the maximal value over consecutive features from one sensor dimension, which can be expressed as

$$\mathbf{h}_k^r = max(\mathbf{c}_{kd+1}^r, ..., \mathbf{c}_{(k+1)d}^r), \quad (2)$$

where d is the pooling size, $k \in \{1,2,...,\lfloor (L-s+1)/d \rfloor\}$, and $\lfloor \cdot \rfloor$ is the rounding down operation.

By performing one 1D convolutional operation and one 1D max-pooling operation, the raw data sample at one sensor dimension $\mathbf{x} \in \mathbb{R}^{L\times 1}$ has been transferred to the feature matrix $\mathbf{h} \in \mathbb{R}^{\lfloor (L-s+1)/d \rfloor \times R}$ where R is the number of filters in 1D convolutional operation. Since the convolution window moves step by step from the beginning to the end of the raw signal at one sensor dimension, the first dimension of the feature matrix will preserve the temporal dependency. At the same time, the second dimension of the feature matrix indicates the high-level representations learned by the CNN at each sequential step of the feature matrix. In this way, we can learn features from the raw sequential acceleration data and preserve the temporal dependency of the raw data. Stacking multiple operation layers, i.e., 1D convolutional layers and 1D max-pooling layers, has been shown to be powerful for representation learning [36]. Here, we stack multiple operation layers for feature learning on raw acceleration time series.

The outputs of the CNN are high-level features of raw acceleration with temporal dependency, which can also be treated as local features with temporal dependency. Then, these local features will be fed into the LSTM to encode temporal dependency for feature learning. The final outputs of the LSTM are the efficient features learned from the big and complex acceleration data. In summary, instead of feature engineering, we design an efficient CNN-LSTM network to automatically learn representative features from the acceleration that is collected under high sampling frequency.

B. Feature Extraction from the HRV

The collected HRV shows the variation of time intervals (i.e., R-R intervals) between heart beats. Given its special format, we are not able to feed it into deep learning algorithms for automatic feature learning. Instead, we extract features from HRV data based on domain knowledge. In particular, 4 types of features are computed from HRV data [37], namely, time-domain features, frequency-domain features, Poincaré plots features and DFA features.

Firstly, we directly drive 8 time-domain features from the R-R interval values, i.e., meanHR, meanRR, StdRR, cvRR, RMSSD, SDSD, pRR50 and RR50. Given a 5-min window, meanRR is the average of all the R-R interval values in this window, while meanHR is the average heart rate in the 5 minutes. StdRR is the standard deviation of the R-R interval values and cvRR is the coefficient of variance (i.e., the ratio between meanRR and StdRR). RMSSD and SDSD are root mean square and standard deviation of the successive differences of R-R interval values, respectively. RR50 (pRR50) refers to the number (portion) of R-R interval values larger than 50 ms.

Second, we perform Fast Fourier Transform (FFT) on the R-R interval values. And then, 7 features from frequency domains are extracted with the power spectrum generated by FFT. In particular, we calculate the power for different frequency bands. For instance, VLF is the power for very low frequency (0.003-0.04 Hz), LF for low frequency (0.04-0.15 Hz), HF for high frequency (0.15-0.4 Hz) and TP for the total power. In addition, the ratios LF/(LF+HF), HF/(LF+HF) and LF/HF are also adopted as frequency-domain features.

Third, we obtain 3 features from the Poincaré plot (i.e., SD1, SD2 and SD1/SD2) and 3 slope coefficients based on DFA. Please refer to [38] and [39] for more details about these 2 types of features. In total, we extract 21 features from HRV data.

C. Attention Network

To make full use of all the available information from two heterogeneous sensors, we apply a concatenate layer to combine the features from the acceleration and HRV. The extracted representations (features) from these two modalities (i.e., HRV and acceleration) may have different contributions for the detection sleep and wake states. To achieve that, we design an attention network to dynamically adjust the importance of features from the two modalities.

The attention mechanism was firstly designed for image processing [40]. It is inspired by human vision systems, claiming that human always pay attention to a certain region of an image during recognition and adjust the focus over time. Here, an attention network is designed to adjust the weights for the features from the two different modalities. Note that, no prior information is available for the attention network to assign the weights. Hence, we design a self-attention scheme where the inputs to the attention network are all the features from the two modalities. Assume that the feature vector is $\mathbf{z} = \mathbf{z}_{ACC} \oplus \mathbf{z}_{HRV}$, where $\mathbf{z}_{ACC}$ and $\mathbf{z}_{HRV}$ are the features from the acceleration and HRV respectively, and $\oplus$ is the concatenation operation, the self attention network can be expressed as

$$softmax\left(\omega^{\top} * \mathbf{z} + \beta\right), \tag{3}$$

where $softmax(\cdot)$ is the *softmax* activation function for the attention network, ω and β are the weights and bias respectively, and $\top$ is the transpose operation. Let $\eta = \omega^{\top} * \mathbf{z} + \beta$ be a vector with N elements (equivalent to the number of features), the i-th attention output (weight) can be expressed as

$$\varpi_i = softmax(\eta_i) = \frac{\exp(\eta_i)}{\sum_{i=1}^{N} \exp(\eta_i)}. \tag{4}$$

Given that $\varpi = [\varpi_1,\ \varpi_2,\ ...,\ \varpi_N]$ are the final attention outputs, i.e., attention weights, of the attention network. Finally, we assign these attention weights to the features by using a element-wise multiplication, which can be expressed as

$$\tilde{\mathbf{z}} = \mathbf{z} \odot \varpi, \tag{5}$$

where $\tilde{\mathbf{z}}$ are the final features for sleep-wake classification, and $\odot$ is the element-wise multiplication operation. Specifically, given vectors $\mathbf{a} = \begin{bmatrix} a_1 & a_2 & ... & a_n \end{bmatrix}^{\top}$ and $\mathbf{b} = \begin{bmatrix} b_1 & b_2 & ... & b_n \end{bmatrix}^{\top}$, $\mathbf{a} \odot \mathbf{b} = \begin{bmatrix} a_1 b_1 & a_2 b_2 & ... & a_n b_n \end{bmatrix}^{\top}$.

D. Proposed Framework for Sleep-wake Detection

Fig. 1. shows the proposed attention based CNN-LSTM framework for sleep-wake classification with two heterogeneous sensor data. Firstly, a sliding window of the four-dimensional acceleration (including the magnitude) is fed into a 1D-CNN to learn local features with temporal dependency. Then, these sequential local features is passed into a LSTM network to learn latent feature representations. The outputs of the LSTM are normalized in batch by using a batch normalization (BN) layer, followed by a dropout layer to prevent over-fitting. Then, a fully connected layer (FCL) is applied to get more abstract features. In the meantime, we extracted some features from the HRV data due to its unique structure. Similarly, a FCL is used on these extracted features to get more abstract representations. Next, to make full use of the available information from these two heterogeneous sensors, we concatenate the features from these two modalities into a feature vector. Then, an attention network is leveraged to dynamically adjust the significance of features from the two modalities. The final features, i.e., the outputs of the attention network, are fed into a BN layer for normalization and a dropout layer to prevent over-fitting. Eventually, we use a *softmax* layer for binary classification between sleep and wake states.

Since the problem method in Fig. 1 is an end-to-end trainable architecture, all the model parameters including the weights of CNN-LSTM, FCLs, attention layer and *softmax* layer can be jointly trained. Specifically, given the predicted sleep/wake states and the true ones, the cross-entropy losses over training data can be calculated and back-propagated to generate the error gradients for each layer (including the attention layer). Then, the optimization method of *Adam* is adopted to optimize model parameters at each layer based on the error gradients.

The hyperparamters of the proposed method which are specified using cross-validation on the training data are shown as follows. In particular, we use 4 1D-convolutional-pooling layers with a kernel size of 10, a step size of 1 and a pooling size of 5, where the number of filters are 16, 32, 64 and 128 respectively. The LSTM has one layer with 100 hidden nodes. The both FCLs have 100 hidden nodes and the both dropout

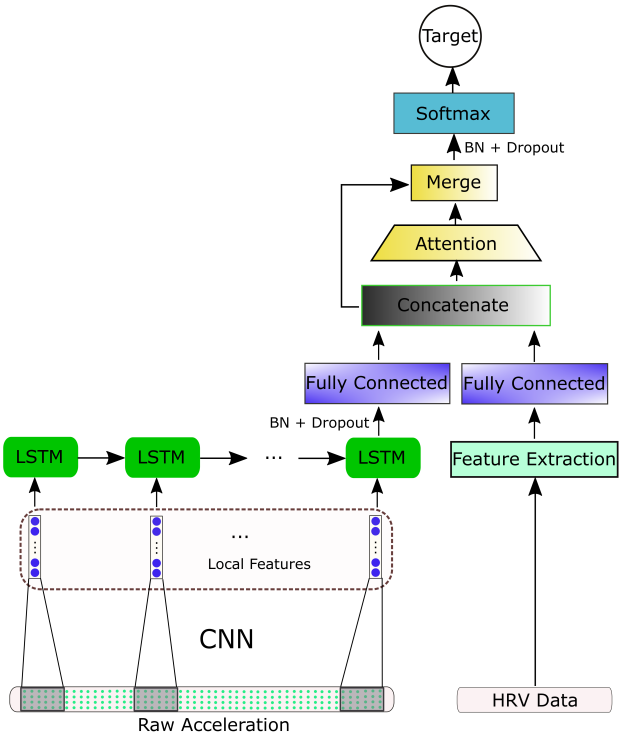


Fig. 1. The proposed attention based CNN-LSTM framework.

layers have a dropout rate of 0.5. The activation functions for the convolutional layers, FCLs and the LSTM are *ReLU*, *ReLU* and *tanh*, respectively.

IV. Experiments

A. Data Acquisition

A dataset was collected from 11 subjects for 28 sleep nights (NUS-IRB Ref Code: B-15-276). Each subject wore three types of sensors, i.e., a Zeo sleep monitor headband, a CamNtech MotionWatch and a FAROS device, that are shown in Fig. 2. Specifically, we use the FAROS device to collect both the acceleration (a sampling rate of 100 Hz) and HRV data (shown as R-R intervals). The CamNtech and the Zeo can report the sleep-wake states of subjects. In particular, CamNtech Watch provides a sleep/wake label per 1 minute. Zeo provides a sleep stage label (i.e., wake, REM, light sleep and deep sleep) per 5 minutes. In this work, we consider REM, light sleep and deep sleep from Zeo as sleep. Here, we used the time shown in FAROS sensor as a reference to synchronize both MotionWatch and Zeo, so that the data from these 3 types of sensors are matched. Note that, the subjects are also requested to record some major events in the night.

In order to perform sleep-wake detection, we split the time series sensor data into 5-min segment which is widely adopted for sleep detection [28], [29]. Three sleep-wake labels are derived for each segment from MotionWatch, Zeo and subjects' event logs. We only keep the segments that three labels are consistent to avoid wrong labelling. Considering that the quality of labels from MotionWatch [41] and Zeo [42] is good, such a consensus process will further improve the quality of labels. Finally, we obtain 1,658 sleep samples and 200 wake samples in this work. For model evaluation, we randomly select about 30% of data for testing and the remaining for training. Note that this dataset is naturally imbalanced and the number of sleep segments is larger than that of wake segments.

Fig. 2. The devices for data acquisition.

B. Experimental Setup

To verify the performance of the proposed method, a comparison has been made with some benchmark approaches which include some traditional machine learning methods, such as DT [13], LD [9], [10], SVM [11], ANN [15] and RF [14], and the deep learning method of CNN [23], [43] and LF-LSTM [24]. Here, the traditional machine learning methods use both the HRV features and the same local features extracted from the acceleration. The CNN in [43] can only use the acceleration data as input (the HRV data cannot be employed due to its special format). The empirical study shows that the CNN with acceleration has very limited performance. We have included the features from HRV into the CNN by using the same feature fusion architecture that we developed in this work, such that the comparison with the CNN can be fair enough. Because of the extremely sequence of acceleration (30,000 time steps), we cannot implement the conventional Bi-LSTM in [27] due to the general constrains on computational power and memory.

The hyperparamters of the benchmark approaches, i.e., ANN, SVM, RF and CNN, are determined by using cross-validation on the training data. Specifically, the number of hidden neurons is set to be 100 and the activation function is chosen to be Rectified Linear Unit (ReLU) for the ANN. The Radial Basis Function (RBF) kernel is adopted for the SVM. The RF algorithm contains 10 decision trees. The CNN consists of four 1D convolutional operations with kernel size of 10 and step size of 2, and four 1D pooling layers with pooling size of 3. The activation function of *ReLU* is applied for all convolutional layers.

Since sleep-wake detection is a highly imbalanced classification problem, the detection accuracy will overlook the minority class that is "wake". Therefore, we adopt the evaluation criterion of G-mean that is popular for evaluating the performance of a model on imbalanced datasets [44]. Given the True Positives (TP), False Positive (FP), False Negative (FN), and True Negative (TN) values, the G-mean is defined as follows:

$$\begin{aligned} \text{precision} &= \text{TP}/(\text{TP}+\text{FN}) \\ \text{recall} &= \text{TN}/(\text{TN}+\text{FP}) \\ \text{G-mean} &= \sqrt{\text{precision} * \text{recall}} \end{aligned} \tag{6}$$

In experiments, we randomly choose 30% of data for the test, and the rest for model training. To give a more comprehensive evaluation, we also perform a leave-one-subject-out (LOSO) cross-validation. Specifically, we use the data from one subject for test, and the remaining for training. This cross-subject test is more challenging as the test data is unseen by the models and thus it is a more realistic scenario to validate the generalization capability of the models.

In this work, in order to handle the imbalanced issue of the data, the widely used technique of SMOTE (Synthetic Minority Over-sampling Technique) [45] is adopted for data augmentation on the training data, such that the number of samples for the two classes, i.e., sleep and wake, is the same.

C. Results and Discussions

1) Comparison with State-of-the-arts: Table I shows the evaluation results of all the methods. Note that, due to the randomness of the neural network based algorithms, we run ten times of the algorithms and the average results are shown. It can be found that the RF method has a superior performance over the other traditional methods of DT, LD, SVM and ANN, and the deep learning method of CNN. The CNN has limited performance. Because it cannot model long-term dependencies in the long sequence of the acceleration.

The proposed approach and the LF-LSTM approach outperform all the other methods under the two criteria, i.e., accuracy and G-mean. These two approaches achieve comparable performance. However, the LF-LSTM method requires a tedious feature engineering process for acceleration data. The proposed method does not contain this tedious process. It is able to automatically learn features from acceleration without human intervention. More importantly, the proposed method outperforms the LF-LSTM in the more challenging and practical evaluation of Leave-One-Subject-Out Cross-Validation, which will be shown later.

TABLE I
EVALUATION RESULTS

Methods	Accuracy (%)	G-mean
DT [13]	89.6	0.718
LD [10]	86.9	0.802
SVM [11]	82.4	0.816
ANN [15]	91.4	0.817
RF [14]	92.3	0.854
CNN [23]	90.0	0.850
LF-LSTM [24]	95.1	0.884
Proposed	94.5	**0.887**

We also show the confusion matrices of all the approach in Table II. Obviously, the proposed approach performs well on the detection of both sleep and wake states. The conclusion is consistent with our previous analysis.

TABLE II
CONFUSION MATRICES OF ALL THE APPROACHES.

(a) DT [13]

	Predicted Wake	Predicted Sleep
True Wake	34	28
True Sleep	30	466

(b) LD [10]

	Predicted Wake	Predicted Sleep
True Wake	45	17
True Sleep	56	440

(c) SVM [11]

	Predicted Wake	Predicted Sleep
True Wake	50	12
True Sleep	86	410

(d) ANN [15]

	Predicted Wake	Predicted Sleep
True Wake	44	18
True Sleep	30	466

(e) RF [14]

	Predicted Wake	Predicted Sleep
True Wake	48	14
True Sleep	29	467

(f) CNN [23]

	Predicted Wake	Predicted Sleep
True Wake	40	12
True Sleep	72	434

(g) LF-LSTM [24]

	Predicted Wake	Predicted Sleep
True Wake	51	11
True Sleep	21	475

(h) Proposed

	Predicted Wake	Predicted Sleep
True Wake	50	12
True Sleep	12	484

2) Ablation Study: We perform an ablation study to show the effectiveness of each part in our method, i.e., the SMOTE, the attention and the HRV data. Table III presents the results of the ablation study. We can find that the model without SMOTE has a higher accuracy and lower G-mean than that with SMOTE. This is because that if without SMOTE for imbalance data augmentation, the classifier outputs will tend to the majority class to enhance classification accuracy, which will influence the detection of the minority class negatively, resulting a lower G-mean. Since the detection of both majority and minority classes is important, the G-mean can be more reliable for the evaluation of imbalanced data [44]. Thus, we will compare the G-mean of various settings for evaluation. It is clear that the model with SMOTE performs much better, which indicates the usefulness of data augmentation for sleep-wake classification which is a typical imbalanced data problem.

According to Table III, the including of the HRV data will improve model performance. This indicates that the HRV data is useful for the task of sleep-wake classification. By using the attention network to dynamically adjust the importance of features from the acceleration and HRV, the proposed approach is further enhanced. We also show the attention weights for the two classes by averaging over all the testing

samples in Fig. 3. Since the both FCLs of the proposed unified framework (See Fig. 1) have 100 hidden nodes. The number of concatenated features is 200. It can be found that the corresponding 200 attention weights for the two states have distinct patterns. This means that the significance of different features is varying during the detection of these two states. Thus, the attention network will be useful for this case with two different modalities.

TABLE III
THE RESULTS OF ABLATION STUDY. HERE, "ACC" STANDS FOR THE ACCELERATION DATA.

SMOTE	Sensors	Attention	Accuracy (%)	G-mean
No	ACC	-	94.4	0.811
No	ACC + HRV	No	95.7	0.830
No	ACC + HRV	Yes	**96.0**	0.846
Yes	ACC	-	90.0	0.870
Yes	ACC + HRV	No	94.5	0.881
Yes	ACC + HRV	Yes	94.5	**0.887**

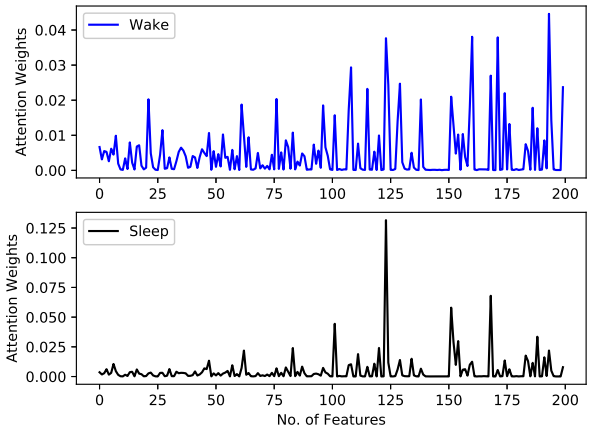


Fig. 3. Attention weights for the two classes.

3) Leave-One-Subject-Out Cross-Validation Results: To test the robustness of the proposed approach, we perform a LOSO cross-validation. The results are shown in Table IV. Compared with Table I, the performances of all the approaches degrade. In LOSO evaluation, all the approaches are tested on the data from an unseen subject. Considering that different subjects may have different behaviors (e.g., movement patterns and HRV patterns), it is reasonable that all the approaches obtain degraded performance.

In this challenging LOSO setting, thanks to the efficient the CNN-LSTM network and the attention mechanism, the proposed method achieves significant improvements over the benchmark approaches, including our previous method of LF-LSTM. This clearly shows the robustness of the proposed approach in the more challenging and practical cross-subject evaluation.The improvements of the proposed approach over the benchmark approaches range from 5% to 46%.

4) Discussion on Multi-modality in Sleep-Wake Detection: How to handle multi-modality is a common issue for detecting sleep/wake states with multiple heterogeneous sources. Sano et al. adopted the multi-modal data collected from wearable

TABLE IV
EXPERIMENTAL RESULTS FOR LOSO CROSS-VALIDATION.

Models	Accuracy (%)	G-mean
DT	82.8	0.542
LD	77.8	0.679
SVM	79.4	0.687
ANN	83.4	0.709
RF	87.5	0.717
CNN	86.6	0.685
LF-LSTM	89.1	0.804
Proposed	**91.8**	**0.845**

sensors and a smartphone for sleep detection [6]. By analyzing each modality, specific features are extracted based on domain knowledge. Finally, they leveraged a bi-directional LSTM model with the extracted features to detect sleep/wake states. Chambon et al. proposed a sleep stage classification system with multi-modal data of EEG and EMG (electromyogram) [1]. They utilized two parallel CNN models to deal with the two modalities separately, and then combined the features learned from these two modalities for classification. Similar idea can be found in [46] where the authors performed sleep detection based on multi-modalities including EEG, EOG (electrooculogram), EMG, Airflow and SaO2 signals. They applied a CNN model for each modality to learn features and combined all the learned features using fully connected layers for sleep detection.

Our proposed method is different with existing approaches in two main aspects: 1) We consider both physical and physiological sensors, i.e., acceleration and HRV respectively, for sleep-wake detection. By analyzing the properties of these two modalities, we designed a CNN-LSTM network to learn features from acceleration with high sampling rate and extracted features from HRV with unique format. 2) Instead of simply combining the features learned from these two modalities, we proposed an attention network to automatically learn the importance of features and assign larger weights to more important ones. With the designed deep learning architecture, we are able to effectively combine these two modalities and achieve the best performance for sleep-wake classification in real experiments.

V. CONCLUSION

In this paper, we proposed an attention based convolutional neural network long short-term memory (CNN-LSTM) approach with two heterogeneous sensors, that are heart rate variability (HRV) and acceleration, for sleep-wake detection. Firstly, a CNN-LSTM network was designed to learn representative features from the big and complex acceleration data. The learned features are combined with the features extracted from the HRV data to make full use of all the information from the two heterogeneous sources (modalities). To dynamically adjust the significance of features from the two modalities, we developed an attention network for efficient sleep-wake classification.

The performance of the proposed method was verified by using real experimental data. The results showed that the proposed method achieved the best performance over

all existing approaches including shallow and deep learning algorithms. In addition, the experimental results demonstrated that the data imbalance correction (i.e., SMOTE), the attention network and the HRV data will boost the model performance. Lastly, to show the robustness of the proposed approach, we conduct a leave-one-subject-out (LOSO) cross-validation for all the approaches. This clearly indicates the robustness of the proposed approach in this challenging and practical scenario. The proposed approach significantly outperforms all the benchmark approaches with the improvements ranging from 5% to 46%. In future works, we intend to work on cost-sensitive learning [47] for sleep-wake detection which is typical imbalance classification problem. Another future work is to collect more data from subjects with more diversities, such as age, race, health states, etc., to further evaluate the generalization performance of models.

REFERENCES

[1] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 4, pp. 758–769, 2018.

[2] P. Fonseca, N. den Teuling, X. Long, and R. M. Aarts, "Cardiorespiratory sleep stage detection using conditional random fields," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 4, pp. 956–966, 2016.

[3] T. Willemen, D. Van Deun, V. Verhaert, M. Vandekerckhove, V. Exadaktylos, J. Verbraecken, S. Van Huffel, B. Haex, and J. Vander Sloten, "An evaluation of cardiorespiratory and movement features with respect to sleep-stage classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 2, pp. 661–669, 2013.

[4] A. Supratak, H. Dong, C. Wu, and Y. Guo, "Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 1998–2008, 2017.

[5] L.-l. Chen, Y. Zhao, P.-f. Ye, J. Zhang, and J.-z. Zou, "Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers," *Expert Systems with Applications*, vol. 85, pp. 279–291, 2017.

[6] A. Sano, W. Chen, D. Lopez-Martinez, S. Taylor, and R. W. Picard, "Multimodal ambulatory sleep detection using lstm recurrent neural networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 4, pp. 1607–1617, 2018.

[7] C. Kuo, Y. Liu, D. Chang, C. Young, F. Shaw, and S. Liang, "Development and evaluation of a wearable device for sleep quality assessment," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1547–1557, 2017.

[8] X. Long, P. Fonseca, J. Foussier, R. Haakma, and R. M. Aarts, "Sleep and wake classification with actigraphy and respiratory effort using dynamic warping," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1272–1284, 2013.

[9] S. Devot, R. Dratwa, and E. Naujokat, "Sleep/wake detection based on cardiorespiratory signals and actigraphy," in *32th Annual International Conference of IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2010, pp. 5089–5092.

[10] X. Long, P. Fonseca, J. Foussier, R. Haakma, and R. M. Aarts, "Sleep and wake classification with actigraphy and respiratory effort using dynamic warping," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1272–1284, 2014.

[11] M. Adnane, Z. Jiang, and Z. Yan, "Sleep–wake stages classification and sleep efficiency estimation using single-lead electrocardiogram," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1401–1413, 2012.

[12] E. Alickovic and A. Subasi, "Ensemble svm method for automatic sleep stage classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 67, no. 6, pp. 1258–1265, 2018.

[13] J. Tilmanne, J. Urbain, M. V. Kothare, A. V. Wouwer, and S. V. Kothare, "Algorithms for sleep–wake identification using actigraphy: a comparative study and new results," *Journal of Sleep Research*, vol. 18, no. 1, pp. 85–98, 2009.

[14] M. B. Pouyan, M. Nourani, and M. Pompeo, "Sleep state classification using pressure sensor mats," in *37th Annual International Conference of Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 1207–1210.

[15] W. Karlen, C. Mattiussi, and D. Floreano, "Sleep and wake classification with ecg and respiratory effort signals," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 3, no. 2, pp. 71–78, 2009.

[16] V. Metsis, D. Kosmopoulos, V. Athitsos, and F. Makedon, "Non-invasive analysis of sleep patterns via multimodal sensor input," *Personal and Ubiquitous Computing*, vol. 18, no. 1, pp. 19–26, 2014.

[17] X. Yang, Z. Zeng, S. Teo, L. Wang, V. Chandrasekhar, and S. Hoi, "Deep learning for practical image recognition: Case study on kaggle competitions," in *KDD*, 2018, pp. 923–931.

[18] R. Sarikaya, G. E. Hinton, and A. Deoras, "Application of deep belief networks for natural language understanding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 778–784, 2014.

[19] C. Chen, K. Li, S. Teo, G. Chen, X. Zou, X. Yang, R. Vijay, J. Feng, and Z. Zeng, "Exploiting spatio-temporal correlations with multiple 3d convolutional neural networks for citywide vehicle flow prediction," in *ICDM*, 2018, pp. 893–898.

[20] A. Gogna, A. Majumdar, and R. Ward, "Semi-supervised stacked label consistent autoencoder for reconstruction and analysis of biomedical signals," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 9, pp. 2196–2205, 2017.

[21] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in Bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2017.

[22] A. Balamurugan, S. Teo, J. Yang, Z. Peng, X. Yang, and Z. Zeng, "Reshnet: Spectrograms based efficient heart sounds classification using stacked residual networks," in *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, 2019, pp. 1–4.

[23] H. Phan, F. Andreotti, N. Cooray, O. Y. Chn, and M. De Vos, "Joint classification and prediction cnn framework for automatic sleep stage classification," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 5, pp. 1285–1296, 2019.

[24] Z. Chen, M. Wu, J. Wu, J. Ding, Z. Zeng, K. Surmacz, and X. Li, "A deep learning approach for sleep-wake detection from hrv and accelerometer data," in *2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, 2019, pp. 1–4.

[25] A. Sors, S. Bonnet, S. Mirek, L. Vercueil, and J.-F. Payen, "A convolutional neural network for sleep stage scoring from raw single-channel eeg," *Biomedical Signal Processing and Control*, vol. 42, pp. 107–114, 2018.

[26] S. Biswal, J. Kulas, H. Sun, B. Goparaju, M. B. Westover, M. T. Bianchi, and J. Sun, "Sleepnet: automated sleep staging system via deep learning," *arXiv preprint arXiv:1707.08262*, 2017.

[27] W. Chen, A. Sano, D. L. Martinez, S. Taylor, A. W. McHill, A. J. Phillips, L. Barger, E. B. Klerman, and R. W. Picard, "Multimodal ambulatory sleep detection," in *IEEE-EMBS International Conference on Biomedical and Health Informatics*, vol. 2017. NIH Public Access, 2017, p. 465.

[28] P. K. Stein and Y. Pu, "Heart rate variability, sleep and sleep disorders," *Sleep Medicine Reviews*, vol. 16, no. 1, pp. 47–66, 2012.

[29] F. Ebrahimi, S.-K. Setarehdan, J. Ayala-Moyeda, and H. Nazeran, "Automatic sleep staging using empirical mode decomposition, discrete wavelet transform, time-domain, and nonlinear dynamics features of heart rate variability signals," *Computer Methods and Programs in Biomedicine*, vol. 112, no. 1, pp. 47–57, 2013.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[31] N. Y. Hammerla, S. Halloran, and T. Plötz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*. AAAI Press, 2016, pp. 1533–1540.

[32] Z. Chen, L. Zhang, Z. Cao, and J. Guo, "Distilling the knowledge from handcrafted features for human activity recognition," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 10, pp. 4334–4342, 2018.

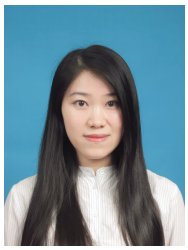
[33] X. Zhang, L. Yao, C. Huang, S. Wang, M. Tan, G. Long, and C. Wang, "Multi-modality sensor data classification with selective attention," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*. AAAI Press, 2018, pp. 3111–3117.

[34] E. Marchi, F. Vesperini, F. Eyben, S. Squartini, and B. Schuller, "A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional lstm neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 1996–2000.

[35] Z. Chen, R. Zhao, Q. Zhu, M. K. Masood, Y. C. Soh, and K. Mao, "Building occupancy estimation with environmental sensors via cd-blstm," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 12, pp. 9549–9559, 2017.
[36] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[37] M. Wu, H. Cao, H.-L. Nguyen, K. Surmacz, and C. Hargrove, "Modeling perceived stress via hrv and accelerometer sensor streams," in *37th Annual International Conference of IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 1625–1628.
[38] A. S. Khaled, M. I. Owis, and A. S. Mohamed, "Employing time-domain methods and poincaré plot of heart rate variability signals to detect congestive heart failure," *BIME Journal*, vol. 6, no. 1, pp. 35–41, 2006.
[39] T. Penzel, J. W. Kantelhardt, L. Grote, J.-H. Peter, and A. Bunde, "Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 10, pp. 1143–1151, 2003.
[40] M. Denil, L. Bazzani, H. Larochelle, and N. de Freitas, "Learning where to attend with deep architectures for image tracking," *Neural Computation*, vol. 24, no. 8, pp. 2151–2184, 2012.
[41] M. S. Ameen, L. M. Cheung, T. Hauser, M. A. Hahn, and M. Schabus, "About the accuracy and problems of consumer devices in the assessment of sleep," *Sensors*, vol. 19, no. 19, p. 4160, 2019.
[42] J. R. Shambroom, S. E. Fábregas, and J. Johnstone, "Validation of an automated wireless system to monitor sleep in healthy adults," *Journal of Sleep Research*, vol. 21, no. 2, pp. 221–230, 2012.
[43] L. Granovsky, G. Shalev, N. Yacovzada, Y. Frank, and S. Fine, "Actigraphy-based sleep/wake pattern detection using convolutional neural networks," *arXiv preprint arXiv:1802.07945*, 2018.
[44] Y. Tang, Y.-Q. Zhang, N. V. Chawla, and S. Krasser, "Svms modeling for highly imbalanced classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 1, pp. 281–288, 2009.
[45] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
[46] A. Patane, S. Ghiasi, E. P. Scilingo, and M. Kwiatkowska, "Automated recognition of sleep arousal using multimodal and personalized deep ensembles of neural networks," in *2018 Computing in Cardiology Conference (CinC)*, vol. 45. IEEE, 2018, pp. 1–4.
[47] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 8, pp. 3573–3587, 2017.

Zhenghua Chen received the B.Eng. degree in mechatronics engineering from University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2011, and Ph.D. degree in electrical and electronic engineering from Nanyang Technological University (NTU), Singapore, in 2017. He has been working at NTU as a research fellow. Currently, he is a scientist at Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore. His research interests include sensory data analytics, machine learning, deep learning, transfer learning and related applications.

Min Wu is currently a senior scientist in Data Analytics Department, Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore. He received his Ph.D. degree in Computer Science from Nanyang Technological University (NTU), Singapore, in 2011 and B.S. degree in Computer Science from University of Science and Technology of China (USTC) in 2006. He received the best paper awards in InCoB 2016 and DASFAA 2015. He also won the IJCAI competition on repeated buyers prediction in 2015. His current research interests include machine learning, data mining and bioinformatics.



Wei Cui received the M.E. and Ph.D. degrees in pattern recognition and intelligent system from Northeastern University, Shenyang, China, in 2013 and 2017, respectively. Currently, she is a scientist at Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore. Her current research interests include wireless sensor networks, localization and navigation, and machine learning.

Chengyu Liu received his B.S. and Ph.D. degrees in Biomedical Engineering from Shandong University, China, in 2005 and 2010 respectively. Dr. Liu has completed the Postdoctoral trainings at Shandong University in China (2010-2013), Newcastle University in UK (2013-2014) and Emory University in USA (2015-2017). He is currently a Professor at School of Instrument Science and Engineering, Southeast University, China, and acts as a Federation Journal Committee Member of IFMBE. His research topics include: mHealth and intelligent monitoring, signal processing and device development for CADs.

Xiaoli Li is currently a principal scientist at the Institute for Infocomm Research, A*STAR, Singapore. He also holds adjunct professor positions at Nanyang Technological University. His research interests include data mining, machine learning, AI, and bioinformatics. He has been serving as a (senior) PC member/workshop chair/session chair in leading data mining and AI related conferences (including KDD, ICDM, SDM, PKDD/ECML, WWW, IJCAI, AAAI, ACL and CIKM). Xiaoli has published more than 200 high quality papers and won numerous best paper/benchmark competition awards.