A Novel Ensemble Deep Learning Approach for Sleep-Wake Detection Using Heart Rate Variability and Acceleration

Zhenghua Chen, Min Wu, Kaizhou Gao, Jiyan Wu, Jie Ding, Zeng Zeng, and Xiaoli Li, *Senior Member, IEEE*

***Abstract*—Sleep-wake detection is of great importance for the measurement of sleep quality. In this article, a novel ensemble deep learning framework is proposed to detect sleep-wake states based on heart rate variability (HRV) and acceleration. We firstly design a local feature based long short-term memory (LF-LSTM) network to encode temporal dependency and learn features from acceleration data with high sampling frequency. In the meantime, some handcrafted features are extracted from HRV which has a special data format. After that, we develop a unified framework to integrate these two types of features, i.e., the features extracted from HRV and the features learned by LF-LSTM from acceleration, to form a complete feature set. Finally, an efficient ensemble learning scheme is proposed to further boost the performance of sleep-wake classification. A real dataset has been collected to verify the effectiveness of the proposed approach. We also compare with some well-known benchmark approaches for sleep-wake detection. The results demonstrate that the proposed ensemble deep learning method outperforms all the benchmark approaches.**

***Index Terms*—Sleep-wake detection, Sensor data, Local features, LSTM, Ensemble deep learning.**

I. Introduction

SLEEP, as an important physiological function for human, affects the performance of various daily activities, including attention, learning, memory and productivity [1]. Inadequate sleep increases the risk of heart disease, stoke and type 2 diabetes. Sleep restrictions and disorders are also linked to the physical and mental health conditions of human [2]. To maintain good health conditions and improve daily performance, it is thus highly desirable to measure both the sleep duration and quality through sleep monitoring and sleep-wake stage detection.

Polysomnography (PSG) [3] is often considered as the gold standard for sleep stage detection. However, it is a labour-intensive and costly procedure to use PSG for sleep monitoring. Electroencephalogram (EEG) can also be used for sleep stage monitoring based on the captured data for brain activity [4]. Recently, economical wearable sensors including actigraphy and accelerometer units are widely used to measure the physical activities and further reflect the sleep patterns [5], [6]. Commercial products including FitBit, Jawbone and Basis have actigraphy or accelerometer units embedded within a watch for activity and sleep monitoring. Other wearable sensors, which collect heart rate variability (HRV) data from electrocardiogram (ECG) or respiratory data, have also been utilized for sleep-wake detection and sleep tracking [7], [8]. 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 may be able to boost the performance of sleep-wake detection.

A plethora of methods have been developed for the detection of sleep and wake stages. These methods can be divided into two categories, i.e., shallow methods and deep methods. Shallow methods usually consist of two steps: 1) feature extraction from the sensory data, and 2) sleep-wake classification by applying traditional machine learning algorithms. Such machine learning algorithms include linear discriminant (LD) classifier [9], [10], decision tree (DT) [5], support vector machine (SVM) [8], [11], artificial neural network (ANN) [5], [7], random forest (RF) [12] and conditional random fields (CRF) [13]. Meanwhile, deep methods [14], [15] directly take the raw wearable sensor data as inputs for sleep-wake classification. For example, a convolutional neural network (CNN) and a bidirectional long short-term memory (Bi-LSTM) were implemented to classify sleep-wake states in [16] and [17].

In this paper, we work on sleep-wake detection by leveraging both HRV and acceleration. However, there are two main technical challenges for this problem, shown as follows.

Firstly, as the acceleration data is a typical time series with temporal dependency, the long short-term memory (LSTM) network with strong capacity for modeling time series data can be an idea candidate. However, as the accelerometer has a high sampling frequency, each sample is thus an extremely long sequence which cannot be processed by the conventional LSTM (or Bi-LSTM [17]), given the limited computational power and memory.

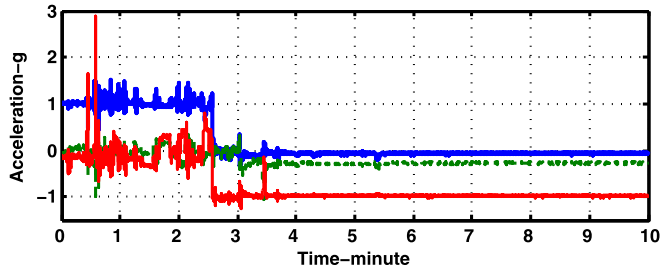
Secondly, acceleration and HRV have different properties and formats. Fig. 1 is an illustration of real streams for acceleration

Manuscript received June 6, 2019; revised January 30, 2020 and March 2, 2020; accepted May 13, 2020. *(Corresponding authors: Xiaoli Li; Min Wu.)*

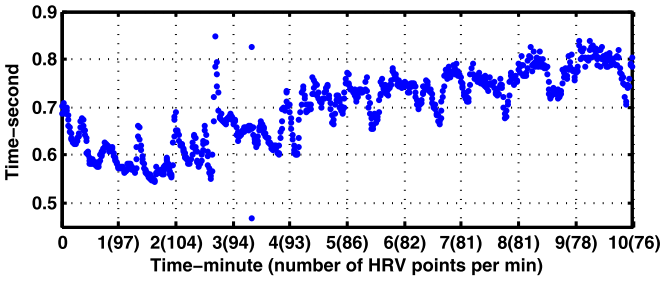
Zhenghua Chen, Min Wu, Jiyan Wu, Jie Ding, Zeng Zeng, and Xiaoli Li are with the Institute for Infocomm Research, A*STAR, Singapore 138632, Singapore (e-mail: chen0832@e.ntu.edu.sg; wumin@i2r.a-star.edu.sg; wu_jiyan@i2r.a-star.edu.sg; ding_jie@i2r.a-star.edu.sg; zengz@i2r.a-star.edu.sg; xlli@i2r.a-star.edu.sg).

Kaizhou Gao is with the Computer School, Liaocheng University, Liaocheng 252000, China, and also with the Macau Institute of Systems Engineering, Macau University of Science and Technology, Macau 999078, China (e-mail: gaokaizh@aliyun.com).

Digital Object Identifier 10.1109/TETCI.2020.2996943



(a) Acceleration



(b) HRV

Fig. 1. An illustration of real acceleration and HRV data with 10 minutes interval. For HRV data, the number of HRV data points per minute (i.e., heart rate) is also shown to indicate its special format. Note that the sampling rate of acceleration is 100 Hz.

and HRV data. We understand that the number of data points for acceleration data in a fixed time window is also fixed. However, as HRV data shows the time intervals between heart beats, the number of data points for HRV data in a fixed time window may be changing over time. For example, the numbers of HRV data points in the first 4 1-min windows (i.e., the heart rates in first 4 minutes) are 97, 104, 94 and 93 as shown in Fig. 1. Therefore, it is difficult to automatically learn features from HRV data by using deep learning techniques. It is also challenging to integrate these two types of sensory data into a unified deep learning architecture for the task of sleep-wake classification.

To address these two issues, we firstly design a local feature based LSTM (LF-LSTM) method to encode temporal dependency and learn high-level features from acceleration data. Specifically, the raw acceleration data in each sample is divided into small windows where local statistical features are extracted within each window. By doing this, we are able to shorten the acceleration time series, find a good representation of the raw acceleration and preserve temporal dependency of the data. Then, the LSTM network is leveraged to learn high-level features from these local sequential features. Meanwhile, we extract some handcrafted features (i.e., features in time domain and frequency domain) from HRV data with a special format. After that, the features learned from both acceleration and HRV are integrated to form a complete feature set. Finally, a classification layer can be employed for sleep-wake detection. To further enhance the performance of the proposed deep learning framework for sleep-wake classification, we design an efficient ensemble learning scheme which leverages on multiple classification layers with the shared feature set. Thus, the ensemble deep learning framework is able to improve the performance with minimal increase of computational complexity. The results on a real dataset demonstrate that our proposed approaches outperform all the benchmark methods for sleep-wake detection.

A preliminary version of this work has been reported in [18]. The main changes made to the original paper are summarized as follows. 1) we propose an efficient ensemble learning scheme on top of the original deep learning framework for sleep-wake detection. As such, we are able to improve the prediction performance with limited increase of computational complexity. 2) we conduct an additional leave-one-subject-out cross-validation experiment and a testing with the data from the latest nights to further show the robustness of the proposed approaches. Experimental results show that the proposed ensemble deep learning framework outperforms the original algorithm in [18] and the benchmark approaches. 3) we report the training and testing time for all the approaches. With the efficient ensemble scheme, the relative increases of training and testing time of our ensemble deep learning approach are 16.6% and 1.5% respectively, when compared with the original approach. Due to the short testing time of the proposed algorithms, they are efficient enough for real-time applications.

The main contributions of this work are summarized as follows:

- We propose a novel unified deep learning framework for sleep-wake classification with two heterogeneous sensors, i.e., acceleration and HRV, with different properties and formats. To the best of our knowledge, this is the first work to effectively combines these two sensors using a deep learning framework.
- We develop an innovative LF-LSTM network to effectively encode temporal dependencies and learn high-level features from long acceleration sequences, which cannot be directly handled by existing LSTM based methods.
- To further improve the performance of the proposed framework, we design an efficient ensemble strategy with limited increase of computational complexity.
- We perform extensive experiments to evaluate the effectiveness of the proposed approaches. The results show that the proposed approaches outperform state-of-the-arts.

The remaining of the paper is orgnized as follows: Section II performs a comprehensive review of related works for sleep-wake classification by using different algorithms. Section III firstly describes the data collection process, followed by the LF-LSTM for acceleration data, feature extraction for HRV data and the efficient ensemble learning strategy. Finally, the proposed framework for sleep-wake detection is presented in this section. Section IV shows the experimental setup, followed by the evaluation results and discussions. Section V concludes this work and shows some potential future works.

II. Related Works

Many algorithms have been reported for sleep-wake classification. Generally, we can divide them into two categories of shallow and deep algorithms.

For shallow model based sleep-wake detection, the first step is to derive representative features from data, such as statistical

features. Then, conventional machine learning algorithms can be implemented with these features to detect the states of sleep and wake. Karlen *et al.* presented an ANN based sleep-wake classification system based on ECG and respiratory [7]. They applied a fast Fourier transform to extract features from ECG and respiratory and then implemented the ANN for classification. Tilmanne *et al.* adopted the ANN and DT for sleep-wake identification with actigraphy [5]. Firstly, twenty-five handcrafted features were extracted from actigraphy data. Then, the ANN and DT algorithms were utilized for sleep-wake classification. Long *et al.* proposed a dynamic warping (DW) method for feature extraction from actigraphy and respiratory [10]. The extracted DW-based features were fed into a LD classifier for the detection of sleep and wake states. In [8], the authors firstly extracted features from the heart rate series, and then utilized the SVM to classify sleep-wake stages. Chen *et al.* developed a sleep stage detection system based on subthalamic local field potentials where features from time domain, frequency domain and entropies were extracted [19]. Then, the classification algorithms of SVM and DT were adopted to identify sleep stages.

Recently, deep models have also achieved great success for the classification of sleep-wake states. Sokolovsky *et al.* presented a deep CNN model for the classification of sleep stages based on multi-channel PSG [20]. Granovsky *et al.* proposed a multi-task CNN model for the detection of sleep-wake patterns [16]. The multi-task CNN which was built upon actigraphy data estimated both sleep/wake stages and the total sleep time. By using the multi-modal data from smartphone and wearable devices, i.e., acceleration, skin conductance and temperature, the authors in [17] presented a Bi-LSTM approach for sleep-wake classification and sleep episode on/off detection. Zhang *et al.* also developed a Bi-LSTM method for sleep-wake classification with cardiorespiratory signals from wearable devices [21].

Existing works did not consider the heterogeneous data with different properties and formats for sleep-wake classification. In this paper, we propose a unified deep learning framework to make full use of two heterogeneous data, i.e., acceleration and HRV. To further improve the performance of sleep-wake classification, we propose an ensemble learning framework with limited increase of model complexity.

III. Methodology

In this section, we firstly introduce the sleep data collected for this study. We then present our proposed ensemble deep learning framework for sleep-wake detection with acceleration and HRV data.

A. Dataset

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 FAROS device, a CamNtech MotionWatch and a Zeo sleep monitor headband, that are shown in Fig. 2. Specifically, the FAROS device is able to collect both acceleration (a sampling rate of 100 Hz) and HRV data. The CamNtech and the Zeo can report the sleep-wake states of subjects. Note that 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, we also ask the subjects to record their key events during these sleep nights.

Fig. 2. The devices for data collection.

TABLE I
Statistics for the Experimental Data

	Total segments	Training data	Testing data
Sleep	1658	1152	506
Wake	200	148	52

To analyze the sleep data, we split it into 5-min segment which is widely adopted for sleep detection [22], [23]. We derive 3 sleep-wake labels for each segment from MotionWatch, Zeo and participant's event log and only keep the segments whose 3 labels are consistent to avoid wrong labelling. Considering that the quality of labels from MotionWatch [24] and Zeo [25] is good, such a consensus process will further improve the quality of labels. Finally, we obtain 1,658 sleep segments and 200 wake segments in this study. For model evaluation, we randomly select about 30% of data for testing and the remaining for training, as shown in Table I. Note that this data is naturally imbalanced and we have many more sleep segments than wake segments.

B. LF-LSTM for Acceleration Data

The acceleration is widely adopted for sleep-wake detection due to the low-cost and easy-to-use properties [6]. To process the acceleration data which is typical time series, the first step is to perform data segmentation by using sliding windows. With a window size of d seconds and a sampling rate of r, each segment will have a size of $dr \times 3$. Since the raw acceleration segment can be noisy and not indicative for the separation of sleep and wake states. Conventional machine learning based sleep-wake detection contains a compulsory step, i.e., manually feature engineering. 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 on acceleration data. 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 for persevering or discarding has been developed in [26]. It has been shown to be powerful for the modeling of long-term dependencies of data.

In real experiments, we segment the acceleration data (100 Hz) with a window size of 5 minutes. Hence, each sample

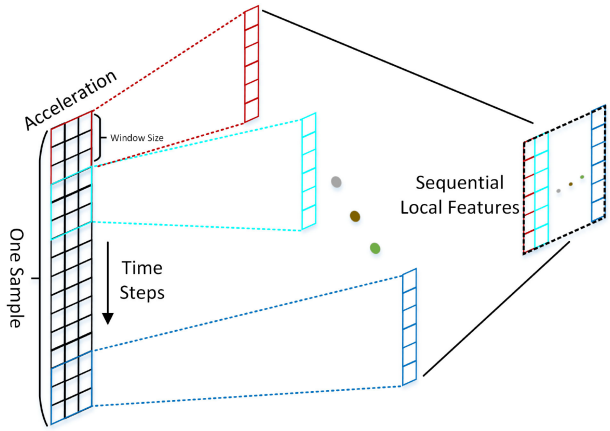


Fig. 3. The process of local feature extraction.

will have a size of $30,000 \times 3$ (three dimensional acceleration). If the normal LSTM network is applied to learn features from this extremely long sequence, it requires 30,000 LSTM cells in cascade connection for feature learning, which is not feasible due to the constrains on computational power and memory. Therefore, we can claim that the conventional LSTM is not applicable for this task. To address this issue, a LF-LSTM is proposed, which consists of two steps, local feature extraction and high-level feature learning. The details are shown in the following paragraphs.

1) Local Features: Due to the extremely long sequence of acceleration segment, we utilize sliding windows over each segment to further divide it into small windows. Then, we extract some representative features for each dimensional of acceleration and combine them to form a feature vector for each small window. Since the small windows are slide sequentially, the temporal dependency of raw data will be preserved. The main objective of this operation is to shorten the length of the sample, extract more abstract representations in each small window and preserve the temporal dependency of the data. Fig. 3. illustrates the process of local feature extraction.

In this work, the local features extracted on each small window are mean, absolute mean, maximum, minimum, range, variance, root mean square, interquartile range, and quantile at 25%, 50% and 75%. A total of eleven features are extracted on each dimension of acceleration for local features. In this work, the size of sliding window, i.e., s in Fig. 3, is chosen as 300 by using cross-validation on the training data. After local feature extraction, the dimension of the data sample is changed to 100×33. Note that, the raw data sample has a size of $30,000 \times 3$. It can be found that the length of the sample is reduced from 30,000 to $30,000/300 = 100$ and the dimension of the sample is augmented from 3 to $3 \times 11 = 33$ by performing local feature extraction.

2) LSTM Based Feature Learning: Since the LSTM network has strong capacity for modeling sequential data, it has been widely adopted to analyze time series data, such as natural language processing [27], occupancy estimation [28] and activity recognition [29]. Fig. 4 shows a typical LSTM structure, where x^t, h^t and C^{t-1} are the input, the hidden state, and the memory cell state respectively, w^f, w^i, w^C and w^o are the weights, b^f, b^i, b^C and b^o are the biases, and $tanh$ and $\sigma(\cdot)$ are the *tanh* and *sigmoid* functions, respectively.

According to Fig. 4, given the previous memory cell state C^{t-1}, the first step of LSTM intends to decide which information should be discarded based on the previous hidden state of h^{t-1} and the current input of x^t by using a forget gate. We can formulate the forget gate as follows:

$$f^t = \sigma\left(w^f[h^{t-1}, x^t] + b^f\right), \tag{1}$$

where $f^t = 0$ represents to totally remove the information from previous steps and $f^t = 1$ represents to keep all the information from previous steps. Next, based on the current input, we require to decide which new information should be included. It contains two parts. The first part is to decide what should be updated by using an input gate, shown as

$$i^t = \sigma\left(w^i[h^{t-1}, x^t] + b^i\right). \tag{2}$$

The second part attempts to produce a candidate cell state $\tilde{C}^t$ with a *tanh* function, which can be expressed as

$$\tilde{C}^t = \tanh\left(w^C[h^{t-1}, x^t] + b^C\right). \tag{3}$$

Then, we update the current cell state C^t with the following equation

$$C^t = f^t * C^{t-1} + i^t * \tilde{C}^t. \tag{4}$$

Eventually, the output gate will decide which information should be preserved from the compressed cell state $\tanh(C^t)$. The determination will be based on the value of o^t which can be calculated as

$$o^t = \sigma\left(w^o[h^{t-1}, x^t] + b^o\right). \tag{5}$$

The final hidden output of the LSTM can be expressed as

$$h^t = o^t * \tanh\left(C^t\right) \tag{6}$$

Here, we leverage on the LSTM network to learn high-level features on local sequential features extracted from raw acceleration data. The authors in [30] have shown that stacking multiple layers will enhance the modeling capacity. Therefore, in this work, we intend to use multiple layers of LSTM to learn more representative features for accurate sleep-wake classification.

C. Handcrafted Features From HRV Data

Heart Rate Variability (HRV) data shows the variation of time intervals (i.e., R-R intervals) between heart beats. Given its special format as shown in Fig. 1, we are not able to feed it into deep learning algorithms for automatic feature learning. Instead, we extract handcrafted features from HRV data. In particular, 4 types of features are computed from HRV data [31], namely, time-domain features, frequency-domain features, Poincaré plots features and DFA (detrended fluctuation analysis) features.

First, 8 time-domain features are directly derived from the R-R interval values, i.e., meanRR, meanHR, StdRR, cvRR, RMSSD, SDSD, RR50 and pRR50. Given a 5-min window, meanRR is the average of all the R-R values in this window, while meanHR

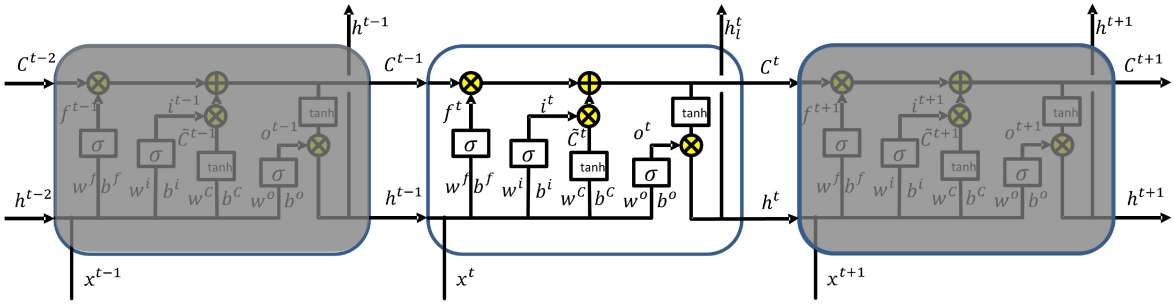


Fig. 4. The structure of LSTM.

is the average heart rate in the 5 minutes. StdRR is the standard deviation of the R-R 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 values, respectively. RR50 (pRR50) refers to the number (portion) of R-R values larger than 50 ms.

Second, we perform Fast Fourier Transform (FFT) on the R-R interval values and then extract 7 frequency-domain features from the power spectrum generated by FFT. In particular, we calculate the power for different frequency bands, e.g., VLF is the power for very low frequency (0-0.04 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 used 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 detrended fluctuation analysis (DFA). Please refer to [32] and [33] for more details about these 2 types of features. In total, we have 21 handcrafted features extracted from HRV data.

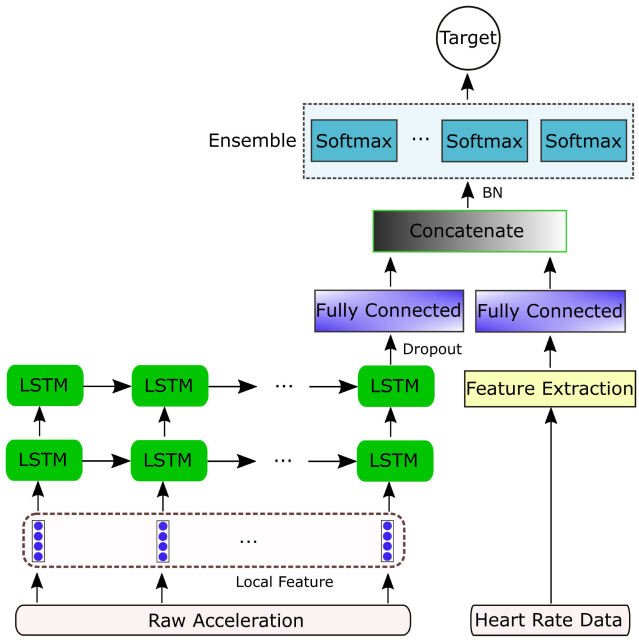


Fig. 5. Proposed EnsemDL framework for sleep-wake detection.

D. Efficient Ensemble Learning

Ensemble learning consists on integrating multiple base learners to boost the performance of learning. It has been shown to be effective for classification and regression problems. Due to the instability of decision trees and neural networks, they are naturally suitable for ensemble learning [34]. Deep neural networks (DNNs) have also been used as base learners for a regression task in [35] where the authors applied multiple DNNs. In this way, the ensemble learning will suffer from huge computational cost. Here, we consider how to further boost the performance of sleep-wake detection by developing an ensemble learning architecture upon the proposed deep learning framework with minimal increase of computational cost.

In this work, we firstly combine the features learned by the LF-LSTM from acceleration and the handcrafted features from HRV to form a complete feature set from raw data. Then, on top of this feature set, we leverage on multiple classification (softmax) layers for the classification of sleep and wake states. The parameters for feature learning and combination which are shared for all the classification layers are in the majority. Assume that the number of parameters for feature learning and combination is M and the number of parameters for a classification layer is N, our designed ensemble learning with k classification layers will lead to an increase of $(k-1)N$ parameters. In our empirical studies, $k = 50$ is good enough for our application. Thus, the relative increase of the number of parameters is $r = \frac{(k-1)N}{M+N}$. In our work, since M is much larger than N, r will be acceptable.

E. Proposed Framework

Fig. 5 shows the proposed ensemble deep learning architecture for sleep-wake classification. Specifically, we firstly develop a LF-LSTM network to learn high-level features from acceleration data with high sampling rate. In the meantime, some representative features are extracted from HRV that has a unique format. Next, we employ two fully connected layers (FCLs) to get more abstract representations for the features from the two heterogeneous sensor data, i.e., acceleration and HRV. After that, we combine the two types of features to form a complete feature set for sleep-wake classification. By using multiple classification layers with different initial parameters, we can obtain multiple

Algorithm 1: Learning of the Proposed Framework.

INITIALIZATION

Generate network weights and biases randomly.

for t $<$ Training epoch **do**

FORWARDS

1. Input raw acceleration $\mathbf{a}$ into LF-LSTM where the output is fed into a FC layer, describing as $\mathbf{feature}_1 = \text{FC}(\text{LF-LSTM}(\mathbf{a}))$;
2. Extract features $\mathbf{e}$ from HRV data and feed them into a FC layer, describing as $\mathbf{feature}_2 = \text{FC}(\mathbf{e})$;
3. Concatenate two types of features, describing as $\mathbf{feature} = \text{Concat}(\mathbf{feature}_1, \mathbf{feature}_2)$;
4. Get ensemble output, describing as $o = \text{EnsemNet}(\mathbf{feature})$.

BACKWARDS

5. Calculate cross-entropy loss based on model outputs o and true labels l, describing as $\text{loss} = \text{cross-entropy}(o, l)$;

for layer in layers **do**

a. Compute the derivatives of the loss function with respect to weights and biases;

b. Update parameters based on the optimization algorithm of *Adam* [36].

end for

end for

predictions. Finally, we perform a majority voting over all the predictions for final classification of sleep and wake states. The proposed ensemble deep learning framework in Fig. 5 is denoted as EnsemDL, while the original deep learning framework we proposed in [18] is denoted as BaseDL. The learning of the proposed frameworks consists of forward and backward propagation. The details are illustrated in Algorithm 1.

The hyperparameters of the proposed framework are determined by using cross-validation on the training data. Specifically, the hidden nodes of the two LSTM layers are set to be 50 and 100, respectively. The dropout layer after the LSTM layers has a dropout rate of 0.5. The hidden nodes of the two fully connected layers are set to be 100. Finally, we utilize 50 softmax layers for ensemble learning in this work.

IV. EVALUATION

A. 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 [5], LD [9], [10], SVM [8], ANN [7] and random forest (RF) [12], and the deep learning method of CNN [16]. Here, the traditional machine learning methods use both the HRV features and the same local features for the acceleration data. The CNN in [16] 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. Since each data sample has 30,000 steps, we cannot implement the conventional Bi-LSTM in [17] due to the general constrains on computational power and memory.

TABLE II
COMPARISON AMONG VARIOUS METHODS FOR SLEEP-WAKE DETECTION

Models	Accuracy (%)	G-mean
DT [5]	89.6	0.718
LD [9], [10]	86.9	0.802
SVM [8]	82.4	0.816
ANN [7]	91.4	0.817
RF [12]	92.3	0.854
CNN	90.0	0.850
BaseDL	95.1	**0**.884
EnsemDL	**95.8**	**0.896**

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 may overlook the minority class that is "wake" in this work. Therefore, we adopt the evaluation criterion of G-mean that is popular for evaluating the performance of a model on imbalanced datasets [37]. Given the True Positives (TP), False Positive (FP), False Negative (FN), and True Negative (TN) values, The G-mean can be defined as follows:

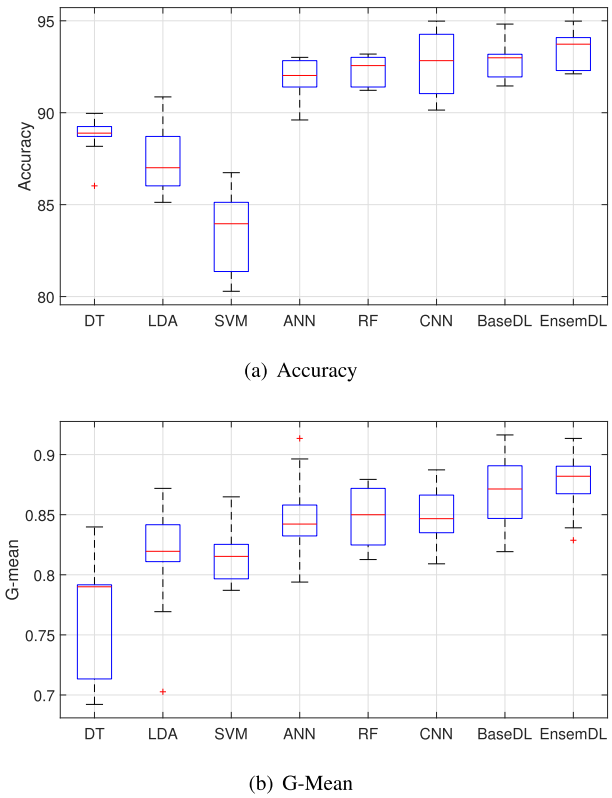
$$\begin{aligned} \text{sensitivity} &= \text{TP}/(\text{TP}+\text{FN}) \\ \text{specificity} &= \text{TN}/(\text{TN}+\text{FP}) \\ \text{G-mean} &= \sqrt{\text{sensitivity} * \text{specificity}} \end{aligned} \tag{7}$$

As mentioned above, the data for sleep-wake detection is imbalanced, we thus adopt the pre-processing technique of oversampling for data imbalance correction on the training data, such that the number of samples for the two classes, i.e., sleep and wake, is the same. In this work, we utilize the oversampling technique of SMOTE (Synthetic Minority Over-sampling Technique) [38] which has been shown to be effective in many tasks.

B. Experimental Results

Next, we present the experimental results based on the data split in Table I, i.e., we randomly select about 70% of data for model training and the remaining for testing.

1) Results: Table II 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



(a) Accuracy

(b) G-Mean

Fig. 6. The experimental results of all the approaches with ten different random selections.

TABLE III
THE ABLATION STUDY FOR THE PROPOSED METHOD

Models	SMOTE	Sensor	Accuracy (%)	G-mean
BaseDL	No	Only Acceleration	94.8	0.794
		Acceleration + HRV	95.2	0.818
	Yes	Only Acceleration	93.6	0.876
		Acceleration + HRV	95.1	0.884
EnsemDL	No	Only Acceleration	95.2	0.847
		Acceleration + HRV	**96.4**	0.868
	Yes	Only Acceleration	93.8	0.891
		Acceleration + HRV	95.8	**0.896**

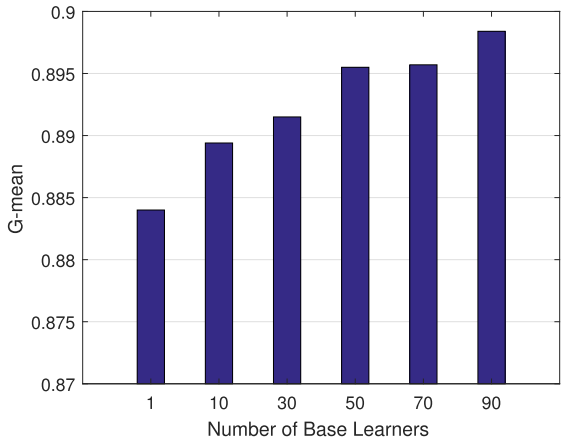


Fig. 7. The G-mean of the proposed EnsemDL approach with different number of based learners.

average results are shown. It can be found that the RF method has a superior performance than the other traditional machine learning methods of DT, LD, SVM and ANN, and the deep learning method of CNN. This indicates the effectiveness of ensemble learning for sleep-wake classification. The CNN method which cannot capture temporal dependencies of data has a limited performance. With the efficient LF-LSTM network for feature learning on acceleration data and the unified framework to combine the extracted features from HRV, the proposed BaseDL performs better than all the benchmark approaches in terms of both accuracy and G-mean. Moreover, with the proposed ensemble learning scheme, the EnsemDL can further enhance the performance for sleep-wake detection.

To further show the effectiveness of the proposed methods, we perform additional experiments with ten different random selections of the training and testing data. The results are shown in Fig. 6. It is consistent with the above analysis. The BaseDL outperforms all the benchmark approaches and the proposed ensemble version achieves the best performance in terms of both accuracy and G-mean.

2) Ablation Study for the Proposed Method: To investigate the impacts of the major components of the proposed method, we perform a comprehensive ablation study. Specifically, we consider three different types of settings for the proposed method, including with SMOTE vs without SMOTE, only acceleration vs acceleration + HRV, and original version vs ensemble version. The results are shown in Table III. We can find that the models without SMOTE achieve higher accuracy and lower G-mean than that with SMOTE. The reason why this occurs is that when not considering SMOTE for data augmentation, the outputs of classifiers 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 evaluation criterion of G-mean is more reliable for the evaluation of imbalanced data [37]. Thus, we will compare the G-mean of various settings for evaluation.

Based on the results in Table III, it is obvious that the SMOTE will dramatically improve the performance of sleep-wake detection due to the high imbalance of the data. In addition, the HRV data is able to further improve the performance of the proposed approaches. This indicates the effectiveness of SMOTE for imbalance data and the usefulness of the HRV data for sleep-wake detection. Among all the different configurations, the proposed EnsemDL further enhances the performance for sleep-wake classification, which clearly indicates the effectiveness of the proposed ensemble learning scheme.

3) Number of Base Learners for Ensemble Learning: In ensemble learning, one of the key parameters is the number of base learners, which will influence both model performance and computational complexity. Here, we investigate the performance of the proposed ensemble learning with different number of based learners. The results are shown in Fig. 7. Note that, the method with one base learner is the original proposed deep learning framework without ensemble learning, which is treated as a baseline.

TABLE IV
THE TRAINING AND TESTING TIME OF VARIOUS APPROACHES

Time	DT	LD	SVM	ANN	RF	CNN	BaseDL	EnsemDL
Training (sec)	0.048	0.013	0.039	1.11	0.040	759.87	801.89	934.99
Testing (sec)	0.00024	0.00012	0.012	0.00073	0.0011	1.31	1.35	1.37

It can be found that, with the proposed ensemble learning scheme, the performance of the model dramatically improves when compared with the baseline. With the increase of the number of base learners, the performance of the model enhances. However, more base learners will require longer training and testing time. According to Fig. 7, when the number of base learner is larger than 50, the relative improvements become marginal. Hence, considering both accuracy and efficiency, we choose the number of base learners as 50 in this work.

4) Computational Time: For ensemble learning, one big concern is the computational complexity. Here, we explore the training and testing time of the proposed approaches, and compare them with benchmark approaches. The workstation for the experiments has twelve core CPUs of Intel i7-8700 3.20 GHz and a GPU of NVIDIA GeForce GTX1080Ti. The results are shown in Table IV.

It is clear that shallow learning algorithms have much shorter training and testing time when compared with deep algorithms. The proposed approach without ensemble, i.e., BaseDL, has slightly longer training and testing, comparing with CNN. The proposed ensemble deep learning approach requires the longest training and testing time. However, due to the proposed efficient ensemble learning scheme, the relative increases of training and testing time of the EnsemDL are 16.6% and 1.5% respectively, comparing with the BaseDL. The conventional ensemble learning in [35] will increase at least k times, where k represents the number of base learners ($k = 50$ in this work). This clearly indicates the efficiency of the propose ensemble learning.

Even though, the training time of the proposed approaches is large, this tedious training process only needs to be done once in offline. The online testing time of the proposed approaches are 1.35 and 1.37 seconds for all testing samples (558 samples). This means that the testing time of the proposed approaches for each testing sample is only 2.42×10^{-3} and 2.46×10^{-3} seconds, which can be neglected in real implementations. Hence, we can claim that our proposed approaches are suitable for real-time applications.

C. Leave-One-Subject-Out Cross-Validation Results

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 testing, and the remaining for training. This cross-subject test is more challenging as the testing data is unseen by the models and thus it is a more realistic scenario to validate the generalization capability of various models.

The LOSO results are shown in Table V. Compared with Table II, 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 pattern and HRV pattern), it is reasonable that all the approaches obtain degraded performance.

TABLE V
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
BaseDL	**88.5**	0.761
EnsemDL	86.6	**0.804**

TABLE VI
EXPERIMENTAL RESULTS ON THE LATEST NIGHTS

Models	Accuracy (%)	G-mean
DT	85.0	0.668
LD	75.6	0.684
SVM	85.6	0.807
ANN	81.8	0.679
RF	88.3	0.757
CNN	89.7	0.772
BaseDL	89.1	0.804
EnsemDL	**92.1**	**0.823**

In this challenging LOSO setting, the proposed approaches significantly outperform all the benchmark approaches, and the EnsemDL achieves the best performance. The improvements of the proposed ensemble deep learning approach over the benchmark approaches range from 13.4% to 75.2%. This clearly shows the robustness of the proposed approach in the cross-subject validation.

D. Testing on the Latest Nights

To demonstrate the robustness of the proposed methods on newly available data, we test all the approaches on the data from the latest nights. In particular, we use the latest 8 nights of data for testing and the first 20 nights of data for training (total 28 nights of data).

The results are shown in Table VI. It can be found that the proposed BaseDL outperforms most of benchmark approaches. Consistently, the proposed EnsemDL performs the best over all the other methods. This further indicates the robustness of the proposed ensemble learning.

V. CONCLUSION

In this paper, we developed a sleep-wake classification system with two types of sensors, i.e., heart rate variability (HRV) and acceleration, by using a novel ensemble deep learning framework. Firstly, we presented a local feature based LSTM

(LF-LSTM) approach for feature learning from the acceleration data. Meanwhile, representative features were manually extracted from the HRV data. Next, we concatenated these two types of features to make full use of all the available information from these two types of sensors. Finally, we designed an efficient ensemble learning scheme on all the features to boost the performance of sleep-wake detection.

We have collected real experimental data for evaluation and compared the proposed approach with several benchmark approaches in the literature. To address the data imbalance issue, we employed the technique of SMOTE (Synthetic Minority Over-sampling Technique) for imbalance correction. The results indicated that the proposed method outperforms the benchmark approaches including conventional machine learning and deep learning methods. In addition, the experimental results demonstrated that the data imbalance correction (i.e., SMOTE) and the HRV data will boost the model performance. Lastly, to show the robustness of the proposed approach, we conducted a leave-one-subject-out (LOSO) cross-validation for all the approaches. The proposed approach significantly outperforms all the benchmark approaches with the improvements ranging from 13.4% to 75.2%. This clearly indicates the robustness of the proposed approach in this challenging and practical scenario.

In future works, we intend to explore methods to automatically learn features from HRV data with a special format. Besides, the contributions of the two sensors may be different for the identification of sleep-wake states. Thus, we attempt to design an attention mechanism [39] to automatically learn the importance of the two sensors and assign larger weights to more important ones. 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] K. A. Wilckens, S. G. Woo, A. R. Kirk, K. I. Erickson, and M. E. Wheeler, "Role of sleep continuity and total sleep time in executive function across the adult lifespan," *Psychol. Aging*, vol. 29, no. 3, pp. 658–665, 2014.

[2] O. M. Buxton *et al.*, "Adverse metabolic consequences in humans of prolonged sleep restriction combined with circadian disruption," *Sci. Transl. Med.*, vol. 4, no. 129, 2012, Art. no. 129ra43.

[3] C. A. Kushida *et al.*, "Practice parameters for the indications for polysomnography and related procedures: An update for 2005," *Sleep*, vol. 28, no. 4, pp. 499–523, 2005.

[4] S. Supriya, S. Siuly, H. Wang, and Y. Zhang, "EEG sleep stages analysis and classification based on weighed complex network features," *IEEE Trans. Emerg. Topics Comput. Intell.*, to be published, doi: 10.1109/TETCI.2018.2876529.

[5] 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," *J. Sleep Res.*, vol. 18, no. 1, pp. 85–98, 2009.

[6] V. T. van Hees *et al.*, "A novel, open access method to assess sleep duration using a wrist-worn accelerometer," *PloS One*, vol. 10, no. 11, 2015, Art. no. e0142533.

[7] W. Karlen, C. Mattiussi, and D. Floreano, "Sleep and wake classification with ECG and respiratory effort signals," *IEEE Trans. Biomed. Circuits Syst.*, vol. 3, no. 2, pp. 71–78, Apr. 2009.

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

[9] S. Devot, R. Dratwa, and E. Naujokat, "Sleep/wake detection based on cardiorespiratory signals and actigraphy," in *Proc. IEEE 32th Annu. Int. Conf. Eng. Med. Biol. Soc.*, 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 J. Biomed. Health Informat.*, vol. 18, no. 4, pp. 1272–1284, Jul. 2014.

[11] T. Willemen *et al.*, "An evaluation of cardiorespiratory and movement features with respect to sleep-stage classification," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 2, pp. 661–669, Mar. 2014.

[12] M. B. Pouyan, M. Nourani, and M. Pompeo, "Sleep state classification using pressure sensor mats," in *Proc. IEEE 37th Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2015, pp. 1207–1210.

[13] P. Fonseca, N. den Teuling, X. Long, and R. M. Aarts, "Cardiorespiratory sleep stage detection using conditional random fields," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 4, pp. 956–966, Jul. 2017.

[14] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," *Briefings Bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[15] Q. Chen *et al.*, "A survey on an emerging area: Deep learning for smart city data," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 3, no. 5, pp. 392–410, Oct. 2019.

[16] L. Granovsky, G. Shalev, N. Yacovzada, Y. Frank, and S. Fine, "Actigraphy-based sleep/wake pattern detection using convolutional neural networks," 2018, *arXiv:1802.07945*.

[17] W. Chen *et al.*, "Multimodal ambulatory sleep detection," in *Proc. IEEE-EMBS Int. Conf. Biomed. Health Informat.*, 2017, vol. 2017, pp. 465–468.

[18] Z. Chen *et al.*, "A deep learning approach for sleep-wake detection from HRV and accelerometer data," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat.*, 2019, pp. 1–4.

[19] Y. Chen *et al.*, "Automatic sleep stage classification based on subthalamic local field potentials," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 2, pp. 118–128, Feb. 2019.

[20] M. Sokolovsky, F. Guerrero, S. Paisarnsrisomsuk, C. Ruiz, and S. A. Alvarez, "Deep learning for automated feature discovery and classification of sleep stages," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, to be published, doi: 10.1109/TCBB.2019.2912955.

[21] Y. Zhang *et al.*, "Sleep stage classification using bidirectional LSTM in wearable multi-sensor systems," in *Proc. IEEE Conf. Comput. Commun. Workshops*, 2019, pp. 443–448.

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

[23] 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," *Comput. Methods Programs Biomedicine*, vol. 112, no. 1, pp. 47–57, 2013.

[24] 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.

[25] J. R. Shambroom, S. E. Fábregas, and J. Johnstone, "Validation of an automated wireless system to monitor sleep in healthy adults," *J. Sleep Res.*, vol. 21, no. 2, pp. 221–230, 2012.

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

[27] S. Wang and J. Jiang, "Learning natural language inference with LSTM," in *Proc. 2016 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technologies*, 2016, pp. 1442–1451.

[28] Z. Chen, R. Zhao, Q. Zhu, M. K. Masood, Y. C. Soh, and K. Mao, "Building occupancy estimation with environmental sensors via CDBLSTM," *IEEE Trans. Ind. Electron.*, vol. 64, no. 12, pp. 9549–9559, Dec. 2017.

[29] Z. Chen, L. Zhang, C. Jiang, Z. Cao, and W. Cui, "WiFi CSI based passive human activity recognition using attention based BLSTM," *IEEE Trans. Mobile Comput.*, vol. 18, no. 11, pp. 2714–2724, Nov. 2019.

[30] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[31] M. Wu, H. Cao, H.-L. Nguyen, K. Surmacz, and C. Hargrove, "Modeling perceived stress via HRV and accelerometer sensor streams," in *Proc. IEEE 37th Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2015, pp. 1625–1628.

[32] 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 J.*, vol. 6, no. 1, pp. 35–41, 2006.

[33] 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 Trans. Biomed. Eng.*, vol. 50, no. 10, pp. 1143–1151, Oct. 2003.

[34] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. London, U.K.: Chapman and Hall, 2012.

[35] X. Qiu, L. Zhang, Y. Ren, P. N. Suganthan, and G. Amaratunga, "Ensemble deep learning for regression and time series forecasting," in *Proc. IEEE Symp. Comput. Intell. Ensemble Learn.*, 2014, pp. 1–6.
[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[37] Y. Tang, Y.-Q. Zhang, N. V. Chawla, and S. Krasser, "SVMs modeling for highly imbalanced classification," *IEEE Trans. Syst., Man, Cybern., Part B*, vol. 39, no. 1, pp. 281–288, Feb. 2009.
[38] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
[39] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

Zhenghua Chen received the B.Eng. degree in mechatronics engineering from the 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 the 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 received his Ph.D. degree in Computer Science from Nanyang Technological University (NTU), Singapore, in 2011 and B.S. degree in Computer Science from the University of Science and Technology of China (USTC) in 2006. He is currently a Senior Scientist in Data Analytics Department, Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore. 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.

Kaizhou Gao received the B.Sc. and master's degrees from Liaocheng University and Yangzhou University, China, in 2005 and 2008, respectively, and the Ph.D. degree from Nanyang Technological University (NTU), Singapore, in 2016. From 2008 to 2012, he was with the School of Computer, Liaocheng University, China. From 2012 to 2013, he was a Research Associate with the School of Electronic and Electrical Engineering, NTU, where he has been a Research Fellow from 2015 to 2018. He is currently an Assistant Professor with the Macau Institute of Systems Engineering, Macau University of Science and Technology. His research interests include intelligent computation, optimization, scheduling, and intelligent transportation. He has published over 70 refereed papers. He is an Associate Editor of international journal Swarm and Evolutionary Computation.

Jiyan Wu received the Ph.D. degree in computer science and technology from the Beijing University of Posts and Telecommunications, Beijing, China, in June 2014. He was a Post-Doctoral Research Fellow with the Singapore University of Technology and Design, Singapore, from 2014 to 2016, and a Senior Software Engineer with OmniVision Technologies Singapore from 2016 to 2017. Currently, he is working on applying machine/deep learning algorithms on industrial sensory data analysis. His research interests include data analysis, video communication, IoT applications, and heterogeneous wireless networks.

Jie Ding received her B.Eng. degree in automation from Harbin Engineering University, China in 2012 and Ph.D. degree in electrical and electronic engineering from Nanyang Technological University, Singapore in 2018. She was a Scientist in Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR), Singapore, until August 2019. Since September 2019, she has been with the Electronic Engineering Department, Fudan University, P.R. China, where she is currently a pretenure associate professor. Her research interests include machine learning, pattern recognition, control & optimization and complex networks.

Zeng Zeng received the Ph.D. degree in electrical and computer engineering from the National University of Singapore, Singapore. Currently, he works as Senior Scientist, Program Head, I2R, A*Star, Singapore. From 2011 to 2014, he worked as a Senior Research Fellow with the National University of Singapore. From 2005 to 2011, he worked as Professor in Computer and Communication School, Hunan University, China. His research interests include distributed/parallel computing systems, data stream analysis, deep learning, multimedia storage systems, wireless sensor networks.

Xiaoli Li (Senior Member, IEEE) 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.