# System Light-Loading Technology for mHealth: Manifold-Learning-Based Medical Data Cleansing and Clinical Trials in WE-CARE Project

Anpeng Huang, *Member, IEEE*, Wenyao Xu, *Member, IEEE*, Zhinan Li, Linzhen Xie, *Member, IEEE*,
Majid Sarrafzadeh, *Fellow, IEEE*, Xiaoming Li, *Senior Member, IEEE*, and Jason Cong, *Fellow, IEEE*

***Abstract*—Cardiovascular disease (CVD) is a major issue to public health. It contributes 41% to the Chinese death rate each year. This huge loss encouraged us to develop a Wearable Efficient teleCARdiology systEm (WE-CARE) for early warning and prevention of CVD risks in real time. WE-CARE is expected to work 24/7 online for mobile health (mHealth) applications. Unfortunately, this purpose is often disrupted in system experiments and clinical trials, even if related enabling technologies work properly. This phenomenon is rooted in the overload issue of complex Electrocardiogram (ECG) data in terms of system integration. In this study, our main objective is to get a system light-loading technology to enable mHealth with a benchmarked ECG anomaly recognition rate. To achieve this objective, we propose an approach to purify clinical features from ECG raw data based on manifold learning, called the Manifold-based ECG-feature Purification algorithm. Our clinical trials verify that our proposal can detect anomalies with a recognition rate of up to 94% which is highly valuable in daily public health-risk alert applications based on clinical criteria. Most importantly, the experiment results demonstrate that the WE-CARE system enabled by our proposal can enhance system reliability by at least two times and reduce false negative rates to 0.76%, and extend the battery life by 40.54%, in the system integration level.**

***Index Terms*—Manifold learning, Manifold-based ECG-feature Purification (MEP), mHealth (mobile health), system light-loading, Wearable Efficient teleCARdiology systEm (WE-CARE).**

## I. Introduction

CARDIOVASCULAR disease (CVD) is a major issue in public health today. Official statistics in [1], [2] show that 230 million people in China, which is 1/5 of Chinese adult people, are suffering from cardiovascular diseases. On average, one patient dies from CVD every 10 s in China, and 3 million patients are dead from this disease each year. In turn, if the death rate would be reduced by 1% in next three decades, the total social cost could be saved by up to 68% of Chinese GDP in the 2010 fiscal year (or say, 10.7 trillion US dollars). This huge loss caused by CVD attracts attention from academic and industry communities to develop an early warning system for CVD risk monitoring [3]–[7]. Recent advances in wireless mobile networking technologies have provided an opportunity to alleviate this problem; this concept is known as mobile health (mHealth) [8]–[10], which is changing health-care delivery today and is at the core of responsive health systems [8].

Based on this motivation, we developed and tested a Wearable Efficient teleCARdiology systEm (WE-CARE) [11], which combines both flexible user mobility and adequate clinical information requirements for real-time disease-risk alert. To let WE-CARE be 24/7 online for the real-time alert purposes, it must work at a light-loaded mode with necessary clinical information. Unfortunately, this objective is confronted by two major challenges given as follows.

1) Reliability of health data transmission is a serious issue in WE-CARE: In this system, electrocardiogram (ECG) data are transmitted in wireless mobile networks for real-time alert purposes. On the other hand, a wireless radio channel is unreliable due to radio channel interference, pathloss, fading, and other mobility effects. Furthermore, wireless resources are always limited. Thus, how to guarantee reliable delivery of ECG data in limited wireless bandwidth is vital in real-time health-risk monitoring applications.
2) Power consumption is a critical issue for WE-CARE terminal users: If the captured raw data are processed at the first-hand site, then the power consumption has a major impact on continuation capability of terminal devices. In WE-CARE, the 24/7 online requirement needs the terminal device running in a power-saving mode. Otherwise, the shortened battery life may affect daily monitoring applications.

In fact, these issues are caused by the overload of complex ECG data of online health-monitoring services, which causes service disruptions quite often in the WE-CARE system.

Manuscript received June 3, 2013; revised November 13, 2013 and September 12, 2013; accepted November 17, 2013. Date of publication November 22, 2013; date of current version September 2, 2014. This work was supported in part by the National Science and Technology Major Projects in Wireless Mobile Healthcare Projects under Contract 2012ZX03005013 and Contract 2013ZX03005008, in part by the National Key Technologies R&D Program under Contract 2013BAI05B07, and in part by OKAWA Foundation. This paper was presented in part at the IEEE Body Sensor Networks 2012 conference.

A. Huang, Z. Li, and L. Xie are with the Mobile Health Lab, PKU-UCLA Joint Research Institute and State Key Lab of Advanced Optical Communication Systems and Networks, Peking University, Beijing 100871, China (e-mail: hapku@pku.edu.cn; lizhinan@pku.edu.cn, tydxlz@pku.edu.cn).

W. Xu is with the Department of Computer Science and Engineering, University at Buffalo, the State University of New York, NY 12246 USA (e-mail: wenyaoxu@buffalo.edu).

M. Sarrafzadeh and J. Cong are with the PKU-UCLA Joint Research Institute and Department of Computer Science, University of California, Los Angeles, CA 90095 USA (e-mail: majid@cs.ucla.edu; cong@cs.ucla.edu).

X. Li is with the PKU-UCLA Joint Research Institute and Computer Science Department, Peking University, Beijing 100871, China (e-mail: lxm@pku.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JBHI.2013.2292576

Obviously, this challenge revealed in our experiments and clinical trials is raised from system integration. To handle the challenge, the acquired ECG raw data should be cleansed to get rid of redundancy inside, so that WE-CARE can be light-loaded. On the other hand, to avoid any potential mistakes during a diagnosing process, ECG raw data should be cleansed at a high-fidelity level. In this study, such an enabling technology solution is expected to maintain service continuation, reduce false negative rates, and improve the battery life, in the system integration level, called system light-loading technology.

To maintain the high-fidelity requirement, we propose a system light-loading approach based on manifold learning, called the Manifold-based ECG-feature Purification (MEP) algorithm. Our proposal fits with requirements in the WE-CARE by taking full advantage of favorable properties given next.

1) Redundancy in ECG raw data can be minimized while preserving clinical features. In practice, some parts of ECG data may be of no meaning to a kind of diseases (for example, arrhythmia), but it may be needed in other kinds. Thus, redundancy information is depending on its specific application objective. In other words, only prechosen information will be monitored and sent back to the data center and medical professionals. Obviously, the amount of ECG data is greatly decreased.
2) Unsupervised patterns are implemented to let MEP be self-adaptive to any variations of segmented ECG cycles dynamically. This is because the physiological nature of each person is unique. This unsupervised pattern is programmed to capture intact ECG signal features regardless of bio-signal diversities.

To the end, contributions in this paper are summarized as follows.

1) The study focuses on a new research topic—system light-loading technology. This is the first time to address such a system-level issue in an Information and Communication Technology (ICT)-based health monitoring system, which is born from system integration and exposed in our empirical study.
2) Considering system light-loading requirement, we propose a manifold-based learning method, called MEP, which can help the WE-CARE system reduce system disruption probabilities and false negative rates, and extend the battery life. In this proposal, manifold learning is applied into the cleansing process considering the manifold nature of ECG raw data.
3) Our WE-CARE is a useful tool to test the effectiveness of our proposal. The WE-CARE is a seven-lead wireless mobile ECG system [11], which is a combination of user mobility and clinical functions.
4) To evaluate our proposal, we conducted clinical trials in our WE-CARE system. Results demonstrate that our proposal is a promising light-loading solution that will enable WE-CARE for wireless mobile health-monitoring applications.

The rest of the paper is organized as follows. In Section II, we briefly introduce our WE-CARE project, which is an innovative mobile seven-lead ECG system. In Section III, the proposed MEP algorithm, a system light-loading technology, is highlighted. In Section IV, system experiments and clinical trials will be performed. Finally, this study is concluded in Section V.

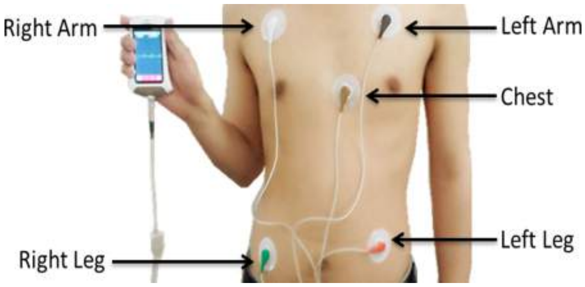


Fig. 1. Portable ECG device of our WE-CARE system.

## II. A System Light-Loading Test Tool: WE-CARE System

For better understanding our topic in this study, let us first look at the WE-CARE system. So far, the existing one-lead or three-lead wireless ECG systems are for home care users, in which the collected data are only for reference, and lack necessary clinical values [6], [7]. In hospitals, 12-lead or 18-lead systems are typically used, but their users lose mobility [6], [7]. It is desired if a system can be designed for health-risk alert purposes, which can combine user mobility and intelligent clinical function. Motivated by this trend, the WE-CARE system was developed for seven-lead ECG real-time monitoring service over mobile networks in PKU mHealth lab (a wireless network may be not mobile, but a mobile network must be wireless).

In our WE-CARE system, the size of the mobile ECG devices is at the scale of normal smart phones (10.3 cm×5.5 cm) in Fig. 1(a). Only five electrodes are needed to derive seven-lead ECG signals [12]. Fig. 1(b) shows how to use the device in clinical trials without disturbing normal daily life. The five electrodes are placed on the patient's body properly (in five attachment locations on the upper body: Left Arm, Right Arm, Left Leg, Right Leg, and Chest, respectively). Then, ECG signals of seven leads can be acquired from the body and visualized in real time on the touch screen of the device. About WE-CARE design principles, refer to [11] for more details.

When our system was tested in lab experiments and clinical trials, its online risk alert broke down quite often. To find out the real reason of system interruption, we did test related enabling technologies on purpose, e.g., ECG detection algorithms, terminal devices, transmission protocols, data center configuration, etc. We found that the upper bound capacity of General Packet Radio Service (GPRS) is 20 Kb/s (only a few Kb/s in most test scenarios), while seven-lead ECG data are more than 28 Kb/s. It means that this system is often crashed due to overload. In the beginning, we use a conventional approach, data compression (please refer to Section IV), in this system. But its average system-interruption frequency and false negative rate are still higher than clinical requirements. This challenge motivates us to study system light-loading technology, which is to reduce the system load without dropping health risk information. For this objective, we propose MEP solution, which is embedded in the

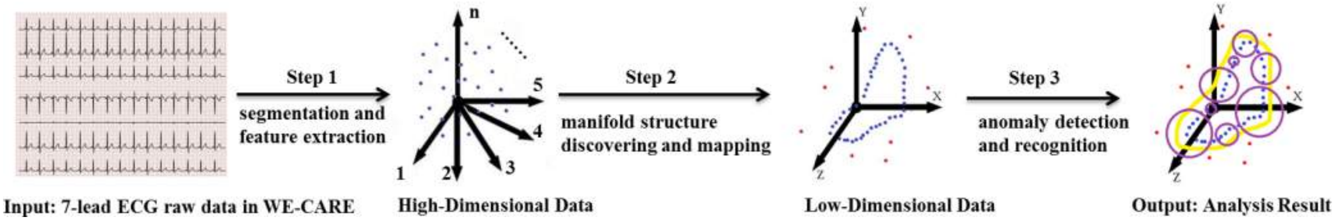


Fig. 2. Flowchart of the proposed the MEP algorithm.

TABLE I
RELATED SYMBOLS AND DEFINITIONS

| Symbols | Definitions |
|---|---|
| $X$ | A high dimensional data vector, which is a representation of an original ECG cycle (the number of dimensions in $\{X\}$ = 42, in Eqn. (1)). |
| $Y$ | A low dimensional data vector, which is a representation of a purified ECG cycle, in Eqn. (6). |
| $NN_i$ | The $i^{th}$ nearest neighbor of $X$, in Eqn. (2). |
| $K$ | The number of nearest neighbors, in Eqn. (2). |
| $N$<br>$w_i$ | The number of all ECG cycles, in Eqn. (3).<br>Reconstruction weight from the neighbor $NN_i$ to $X$, in Eqn. (2). |
| $w_{ij}$ | Reconstruction weight from $X_i$ to $X_j$, in Eqn. (3). |
| $e$ | *Reconstruction error* of $X$, in Eqn. (2). |
| $E$ | The total reconstruction error of all data units in the original coordinate system, in Eqn. (3). |
| $E$' | The total reconstruction error of all data units in the target coordinate system, in Eqn. (6). |

developed portable devices. This solution can help our system to offer 24/7 online reliable services in a power-saving mode for CVD patients. In the WE-CARE system, the raw ECG data are cleansed and transmitted to the data center stably. When a CVD risk is detected, this system can be triggered to transmit original ECG data for professional clinical analysis and activate a necessary emergency response. Next, we highlight how our proposal works.

## III. SYSTEM LIGHT-LOADING PROPOSAL: PURIFYING CLINICAL FEATURES FROM ECG RAW DATA BASED ON MANIFOLD LEARNING

Today, manifold learning theory is applied into many areas, for example, image processing [13], [14], to reduce data dimensionality. In this study, we extend manifold theory to purify clinical features from ECG raw signals in order to light-load the WE-CARE system (or other similar real-time wireless medical systems). To be specific, we intend to reduce redundancy by mapping high-dimensional ECG raw data into a low-dimensional space. In our proposal, manifold learning theory is adopted considering the manifold nature of ECG raw data, and the demand of original clinical information preservation [15]. Based on internal logics between terms above, the proposal is called MEP algorithm.

As shown in Fig. 2, our algorithm consists of three components, i.e., segmentation and feature extraction in Step 1, Manifold-structure discovering and mapping in Step 2, and anomaly detection and recognition in Step 3. Here, related symbols and definitions are listed in Table I.

Below, the proposed MEP algorithm is illuminated step by step.

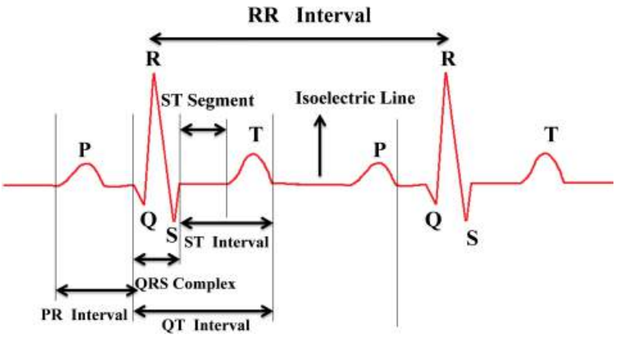


Fig. 3. Cardiac working cycle of an ECG signal.

### *A. Step-1: Segmentation and Feature Selection*

For enabling the time-sensitive risk alert function in our WE-CARE system, the captured ECG signals are processed within a time window. Thus, the ECG signal flows should be cut into a sequence of length-limited data unit. As well known, ECG signals are of a typical periodical character (please see Fig. 3). In order to preserve the original periodical structure of ECG signals, the time-window of data segmentation should be self-adaptive to avoid clinical information loss. Based on these investigations, it is better to take an inherent feature in ECG signals for ECG data segmentation, e.g., R-R interval (the interval between two adjacent R peaks as shown in Fig. 3). To segment the R-R intervals, QRS wave (see Fig. 3) detection method is useful (please refer to [16]). Inside a cycle of ECG signals, we need to extract clinical features from ECG raw data so that ECG abnormal detection is still efficient and effective after predigested process (please see Section IV. A).

Conventionally, the ECG features are selected based on the signal properties, e.g., QRS wave [16], P-Wave [17], [18]. However, these features are difficult to be quantified and cause high computing complexity. For these concerns above, we use statistical approaches to express clinical information in ECG signals within each segmentation cycle. Here, six statistical quantities are defined as the features in Table II.

These six features of each cycle are formed into a sequence, named as $X$-vectors:

$$X = \{x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, \ldots \\ \ldots, x_{n1}, x_{n2}, x_{n3}, x_{n4}, x_{n5}, x_{n6}\}, \qquad (1)$$

where $n$ is the number of ECG leads. In Function (1), each statistical feature can be regarded as a vector, and each lead is also taken as a vector. If there are $m$ statistical features, and $n$ leads, we can get an $m \times n$ dimensional coordinate system.

TABLE II
SELECTED FEATURES AND THEIR DEFINITIONS

| Features | Definition Descriptions |
|---|---|
| Arithmetic mean | The direct current component (average value) of the signal over the segmented window. |
| Standard deviation | The variability or the spread of the signal over the segmented window. |
| Derivative mean | Average value of the first order derivative sequence of the signal. |
| Derivative variance | The diversity measurement of the first order derivative sequence of the signal. |
| Correlation mean | The similarity of self-correlated sequential signal over the segmented window. |
| Correlation variance | The change or variation of the similarity of correlation over the segmented window. |

The coordinate system turns statistical features into geometrical presentation, which can help us understand our proposal more easily. For example, an $X$-vector data can be labeled as a vector point in the coordinate system. In our WE-CARE system, an $X$-vector data unit contains total 42 features within a "seven-lead" ECG-signal cycle.

## *B. Step-2: Manifold Structure Discovering and Mapping*

Based on H. Kantz and T. Schreiber's discussion on the determinism and randomness of ECG signals [19], our further investigation showed that ECG signals are of manifold nature. This means there is a possibility to scale down the dimensions of ECG raw data so that an analysis process can be more efficient.

To make full use of this unique nature, these 42-dimensional $X$-vector data can be mapped into a low-dimensional space. In our proposal, the dimensional mapping method is based on Locally Linear Embedding algorithm [20], which can preserve the locality property of ECG signals during the geometrical transformation. Thanks to this locality preservation, anomaly detection based on the data in the low-dimensional space is still valid. Here, our algorithm is implemented in an unsupervised pattern, so that no training procedures are needed. In response, the unsupervised pattern can turn our algorithm to be self-adaptive to any variations of segmented cycles dynamically. The significance of this unsupervised pattern can be further revealed by the example of the R-R interval used for the segmentation cycle in Step 1. Since anyone is unique in physiological natures, it can be translated that the length of the R-R interval is different one to one. To handle bio-signal diversities, this unsupervised pattern is very helpful to capture intact ECG signal features during our algorithm's segmentation and transformation processes.

Here, three substeps of Step 2 are highlighted below.

1) $K$-nearest neighbor searching: The first substep is to calculate the locality of ECG signals by searching $K$-nearest neighbors for each $X$-vector data unit. We choose a data unit as a basis first. Then, all the nearest neighbors to this basis are grouped together at the locality formation. Between neighboring data units, we use the Euclidean distance to evaluate their similarity degree [21]. Then all the candidate data units are ranked according to the similarity degree, in which top $K$ data units are selected as the $K$-nearest neighbors. We will further discuss the impact of the $K$ size on our algorithm in the Section IV.

2) High-dimensional data structure preservation: This substep is to obtain a reconstruction weight matrix that preserves the data geometrical locality in the original dimensional space. The geometrical locality means the data structure between two feature-vector points in the coordinate system. This is because the locality can reflect similarity between two points in the coordinate system. Certainly, the similarity is the foundation to distinguish abnormal points from normal ones. Thus, the locality (or say, data structure) is full of vital clinical information.

For an arbitrary $X$-vector data unit ($X$) with $K$ nearest neighbors $NN_i$, its reconstruction error ($e$) is defined as,

$$e = \left| X - \sum_{i=1}^{K} (w_i * NN_i) \right|, \tag{2}$$

where $w_i$ is the reconstruction weight from the neighbor $NN_i$ to X.

In this substep, we try to minimize the reconstruction error "$e$" by adjusting values of $w_i$. Based on Function (2), the *total reconstruction error* ($E$) of all data units in the coordinate system can be written as follows:

$$E = \sum_{j=1}^{N} \left( \left| X_j - \sum_{i=1}^{N} (w_{ij} * X_i) \right| \right), \tag{3}$$

s.t.,

$$w_{ij} = 0, \tag{4}$$

if $X_i$ is not in the nearest neighbor list of $X_j$.

$$\sum_{j=1}^{N} \sum_{i=1}^{N} w_{ij} = 1, \tag{5}$$

where (3) expresses the reconstruction problem in a closed least-square form, which can solve the weight $w_{ij}$ [22]. Equation (4) stands for an exclusiveness constraint when a data unit looks for its nearest neighbors, and (5) shows a normalization requirement, in which the sum of nearest neighbors' weights is normalized to 1.

With the constraints of (4) and (5) above, the reconstruction weight matrix $w_{ij}$ can be obtained, which preserves the data structure of the original high-dimensional space.

3) Low-dimensional embedding reconstruction: In the third substep, we use the aforementioned reconstruction weight matrix to reconstruct data in a low-dimensional space (which is our target space in this study). As explained in the second substep, the inherent locality between data units is characterized by $w_{ij}$. This substep is to search the low-dimensional representation $Y$ or $X$, which is obtained through minimizing the following error $E'$:

$$E' = \sum_{j=1}^{N} \left( \left| Y_j - \sum_{i=1}^{N} (w_{ij} * Y_{ij}) \right| \right), \tag{6}$$

where the weight $w_{ij}$ is from the second substep, and the object $Y_j$ is the low-dimensional manifold. Equation (6) is in a quadratic form, which can make the embedding reconstruction optimization process be solved. Furthermore, all the manifold points $Y_i$ in the low-dimensional space can be generated globally and concurrently. In other words, all vector points are generated together in both space and time domains. As a result, our reconstruction results can avoid local optimal cases.

From (3), the low dimension reconstruction only depends on the locality of the high dimensional data. This means that the manifold $Y_i$ can be transformed with an arbitrary displacement, or rotated with a random angle, without affecting results in (6). These geometric attributes of displacement and rotation can be formulated as follows:

$$\sum_{i=1}^{N} Y_i = 0. \tag{7}$$

$$\frac{1}{N}\sum_{i=1}^{N} Y_i \cdot Y_i = 1. \tag{8}$$

Therefore, this manifold reconstruction becomes an eigenvalue problem [22], in which we select the matrix rank to achieve the expected manifold dimension.

The Step 2 above is described in an algorithm below.

**Algorithm 1** : Manifold Structure Discovering and Mapping.

```
// Input X: A high dimensional representation of an ECG cycle.
// Input N: The number of X.
// Output Y: A low dimensional representation of an ECG cycle.

procedure SUB-STEP 1. TO SEARCH $K$-NEAREST NEIGHBORS FOR EACH $X$.(X)
    for $i \leftarrow 1, N$ do
        Compute the distance from $X_i$ to every other vector $X_j$.
        Find the $K$ smallest distances among results above.
        Add these $K$ nearest neighbors to the neighboring list of $X_i$.
    end for
end procedure
procedure SUB-STEP 2. TO CONFIGURE reconstruction weight matrix $W$ (WHICH IS USED FOR HIGH-DIMENSIONAL DATA STRUCTURE PRESERVATION).(X)
    Configure $W$, in which the element $w_{ij}$ is set as 0, if $X_i$ is not in the nearest neighbor list of $X_j$. Otherwise, $w_{ij}$ is an undetermined factor.
    Get the total reconstruction error ($E$) in the original space, which is the function of $W$, in Eqn. (3).
    Compute $W$ under the constraint of $E$ minimization.
end procedure
procedure SUB-STEP 3. TO OBTAIN $Y$, WHICH IS A LOW-DIMENSIONAL EMBEDDING RECONSTRUCTION.($E$)
    Get the total reconstruction error ($E'$) in the target space which is the function of $Y$ with a factor $W$ (derived in Sub-step 2), in Eqn. (6).
    Obtain the expected $Y$ by minimizing $E'$.
end procedure
```

### C. Step-3: Anomaly Detection and Recognition

In the end of Step 2, by computing $Y$, we get the low-dimensional representation of ECG cycles. Fig. 4 shows an example of $Y$ in a 3-D target space. The original ECG data are from an arrhythmia patient. Each point in this figure represents a data unit ($Y$) that is corresponding to an ECG cycle. As shown in the figure, some data units with high similarity form a cluster. In our clinical trials, it is revealed that repeated anomaly episodes are still out of the ECG normal cluster, no matter how often the anomaly occurs. This benefit is from the unique design in our proposal, which can abstract clinical information by using statistical features. Based on the basic principles of clustering-based anomaly detection [23], data units inside the cluster are marked as "normal" and data units outside are annotated with "abnormal," in which the universal $K$-means method [23] can be adopted for anomaly detection. In our study, only one "normal" cluster is formed in terms of the anomaly detection. Therefore, this $K$-means method is downgraded into a 1-means method. In this algorithm, the geometrical locality of medical data is the key feature for clustering-based anomaly detection methods, which is to extract this key feature from original data and represent it in a simple data structure. Consequently, the time computing complexity of $K$-means method is reduced to O($Nt$), where $N$ is the number of ECG cycles, and $t$ is the number of 1-means iterations [24].

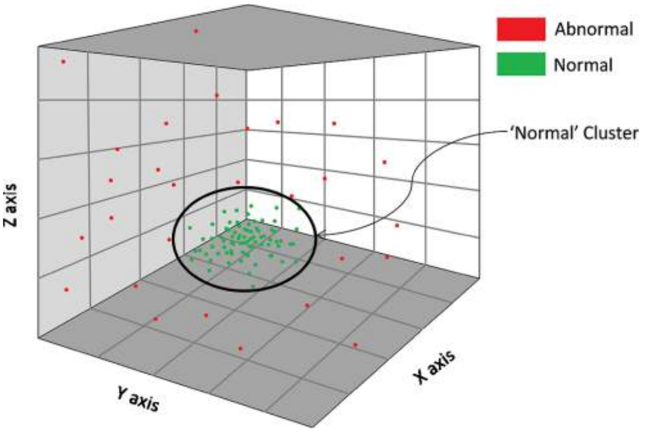


Fig. 4. Anomaly detection results of our proposed MEP algorithm.

In this section, we discussed the overall time and space complexities of the algorithm. Below, we will discuss the algorithm performance from the point of system integration perspective. Specially, we will first investigate whether the clinical information in ECG is preserved or not after the manifold process, and then evaluate how much reduction of system overload in terms of the communication bit rates, including power consumption reduction tests.

## IV. EXPERIMENTS AND EVALUATION

To evaluate our proposal, clinical trials and experiments are performed in the WE-CARE system. To show performance gain of system light-loading technology, ECG compression methods are cited as contrasted references. In reality, compression is not a light-loading technology according to the definition in Section I. This is because a compression method may reduce the system load somehow, but it could not handle issues related with system integration, e.g., battery life, false negative rates. Additionally, higher compression ratio causes higher data distortion [25], [26]. However, the compression is still an option to reduce the system load in most applications. Thus, we compare our proposal with two typical compression methods, Fast Fourier Transform (FFT)-based compression, and Discrete Cosine Transform (DCT)-based compression (for more advanced compression methods, e.g., Discrete Wavelet Transform (DWT)-based method, they are not discussed here because of their high complexities unsuitable for the WE-CARE system). Furthermore, we also show the results of our WE-CARE system without any data preprocessing [this choice is named as Non-Preprocessing Method (NPM)].

In this study, our proposal aims at system integration technology, namely, system light-loading issue without losing clinical features. Thus, our tests include two tasks: to verify the proposal itself first, and then to evaluate how much gain when the WE-CARE system is equipped with our proposal. To accomplish these tasks, 225 subjects (159 patients with mild disorder symptoms, 66 patients with disease symptoms) from 246 samples from Peking University hospital during last two years were invited for clinical trials. In these tests, each subject wears a WE-CARE device at home for risk monitoring.[1]

### A. Effectiveness Verification of the Proposal

*1) Anomaly Detection for CVD Prevention:* First of all, we investigated the impact of our proposal's two parameters on *anomaly recognition rate*—the group size $K$ discussed in Section III (Step 2) and the targeted dimension.

Based on the fundamentals of our proposal, if the group size $K$-value is too small, the manifold reconstruction in Section III (Step 3) has a lack of searching freedom; if the $K$-value is too large (more than the original input dimension), the locality of raw data described in Section III (Step 3) may lose the unique clinical information (please refer to [22] for details). Given the fact that the original input dimension is 42 in our WE-CARE system, we evaluated the algorithm with $K$ from 2 to 42. Fig. 5 shows the influence of $K$ setup on the anomaly detection when the targeted low-dimensional spaces are a 2-D space and a 3-D space, respectively. From the experimental results, we can observe that the recognition rate reaches the maximum at a $K$-value of 10. The maximum recognition rate is 90% in the 2-D case and 94% in the 3-D case. Based on test results, we observed a phenomenon: the recognition rate did not increase significantly when target space dimensionality is higher than 3-D. From now on, a 3-D space is chosen for the targeted dimension and the $K$-value is set as 10 in our experiments.

Now, let us investigate the differences between our proposal and references in term of *anomaly recognition rates*. In our trials, the anomaly recognition rate of NPM gets 95.6%, that of DCT-based compression reaches 83.4%, and that of FFT-based compression is at 87.7%. In terms of anomaly recognition rates,[2] compression methods get lower performance because there are data distortion and information loss in the ECG compression–decompression process. Although that the NPM has the highest recognition rate, it frequently leads system failures to WE-CARE due to the overload issue. In fact, the WE-CARE system cannot work properly at all, when it is equipped with NPM. Compared with compression methods, our proposal enables the system to work at a required fidelity level, and also maintains online service stability. We will further investigate this benefit below.

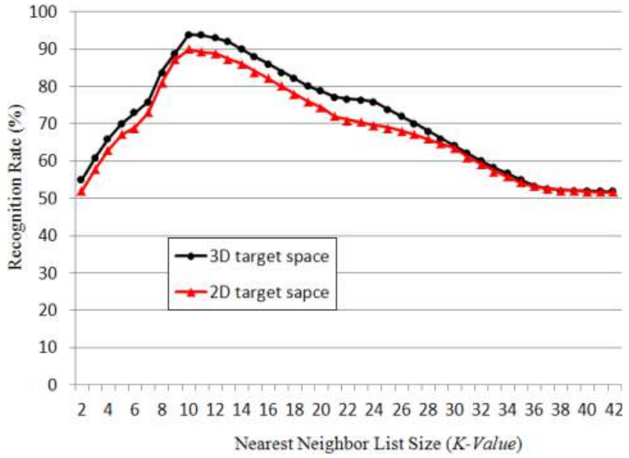


Fig. 5. Relationship between K-Value and Recognition Rate.

TABLE III
EXPERIMENT PARAMETERS

| Experiment Parameters | Values |
|---|---|
| The average ECG cycle length | 0.79 second |
| Resolution of original ECG signal sampling | 11 bits/sample |
| Resolution of Dimensionality Reduction Output | 32 bits/symbol |
| Number of ECG leads | 7 leads |
| Wireless Communication Network | GPRS Network |

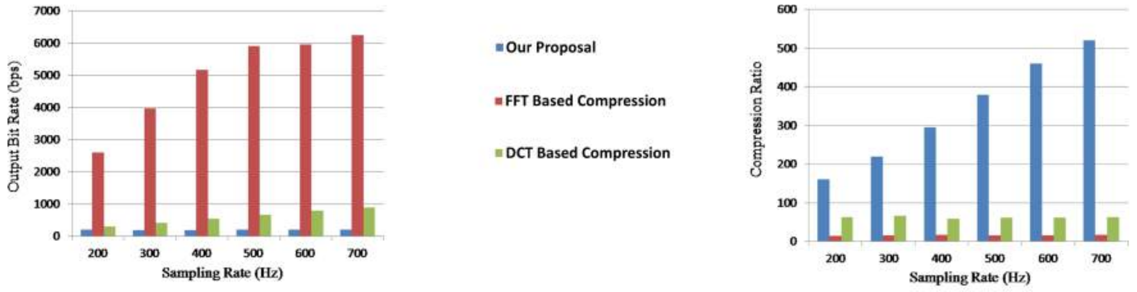
*2) Redundancy Reduction From ECG Raw Data:* In the WE-CARE project, ECG data are carried over mobile networks. In any mobile networks, limited wireless bandwidth is a common challenge to any mobile broadband applications. Thus, traffic-carrying capacity of a radio channel is always constrained. To make full use of limited carrying capacity in a radio channel, the redundancy in ECG data should be reduced as much as possible. Since our proposal is compared with ECG compression methods, we use the reduced amount of transmitted ECG data as an evaluation criterion (a compression method usually reduces traffic amount including both useful and redundant information. In our proposal, its objective is to only minimize the redundant part without loss of clinical information).

Related parameters in these tests are shown in Table III.

Fig. 6 shows comparison between compression methods and our proposal in terms of raw ECG data reduction when they are embedded into WE-CARE. In this figure, the redundancy

[1] In this project, 6 hours are chosen as a monitoring time window, because users may feel uncomfortable itch if wearing ECG electrode pastes at 24/7 (a novel type of ECG electrode pastes is expected for making daily users comfortable). Please note, among 246 patients, there are 21 patients with a potential life-threatening attack who have to be stayed in hospital for intensively treatment, and are not involved for WE-CARE tests.

[2] In this study, the regular motion noise can be treated as clinical normal features within a period of unsupervised working. For occasional motion noise and signal noise, they may lower down WE-CARE system performance, which are left for future studies.



(a)

(b)

Fig. 6. Comparison between FFT-based Compression, DCT-based Compression and the proposed MEP. (a) Comparison in terms of output bit rates. (b) Comparison in terms of compression rates.

TABLE IV
METRICS TO EVALUATE SYSTEM SERVICE RELIABILITY

| Performance Metrics | Definitions |
|---|---|
| Average reliable service duration (ARSD) | An average duration when WE-CARE provides reliable service without any disruption. |
| False negative rate (FNR) | A possibility that an alert is missed or failed to alarm the health-risk monitoring center. |

TABLE V
PERFORMANCE OF EACH METHOD

| Methods | Features | Results |
|---|---|---|
| NPM | Large data amount<br>No information loss | ARSD:0.31 hours<br>FNR:14.96% |
| FFT-based compression | Medium data amount<br>Large information loss | ARSD:6.82 hours<br>FNR:7.16% |
| DCT-based compression | Medium data amount<br>Large information loss | ARSD:8.71 hours<br>FNR:9.44% |
| Our proposal | Small data amount<br>Small information loss | ARSD:18.94 hours<br>FNR:0.76% |

amount removed by our proposal is also quantified in the compression ratio. As shown in Fig. 6(a), regardless of any ECG sampling rates, our proposal can maintain the carrying load around 103 bit/s. The carrying load of our proposal is approximately stationary because its output bit rate is not related to the input bit rate. For a given cleansed level, the output bit rate is only determined by the cycle length of ECG signals (please refer to Section III for details), which is the unique gain of our proposal. As a result, the comparative compression ratio (input bit rate over output bit rate) of our proposal became significantly larger when the input data bit rate increases, as shown in Fig. 6(b).

### *B. System-Level Performance Enhancement*

*1) Data Transmission in Wireless Mobile Networks:* Ideally, our WE-CARE system is designed to provide continuous services for users in mobile networks. Since health risk happens unexpectedly for CVD patients, CVD risk monitoring service is expected to be 24/7 online. Unfortunately, in any wireless mobile networks, radio channels are suffering from interference, pathloss, fading, and other mobility effects. In such a wireless mobile network with numerous intangibles, data transmission is interrupted frequently. The usual technique to enhance wireless transmission is to utilize low-order modulation and coding methods. Considering a low-order modulation and coding scheme has low carrying capacity [27], it is key to reduce data amount in order to stabilize WE-CARE services. As shown in the last paragraph, our proposal is an effective approach to minimize redundant information. Here, we test the proposal in the WE-CARE system in the ward indoor scenario at PKU's affiliated hospitals. In our experiments, two performance metrics are defined in Table IV for evaluating service reliability in the WE-CARE system. Experiment results are shown in Table V.

From these results in Table V, our proposal can maintain the longest reliable service time at 18.94 h and the lowest *false negative rate* at 0.76%. In terms of *false negative rate*, our proposal obtains significant gains, which is around 1/20 of NPM's outcome, and 1/12 of DCT-based compression's. This is because NPM keeps the raw data for system transmission. In turn, overloading ECG data lead to system interruption more often, owing to vulnerable wireless communications. As for the compression methods, wireless transmission reliability may be improved at some extent, but clinical information lost due to compression may introduce some false negative rates to the WE-CARE system. That is why the compression methods can get a comparative performance in contrast with NPM, but they lose superiority in contrast to our proposal. These results prove that our proposal is effective to overcome the system interruption problem due to the overload issue.

*2) Power Consumption in Terminal Devices:* Without doubt, power consumption dominates continuation capability of any terminal services. For a healthcare application, the continuation capability of terminal devices can help to avoid missing health-risk alerts. To investigate this issue, we compared our proposal with alternates in terms of battery life. In our experiments, the tested devices were fully charged (SAMSUNGEB595675LUCCHA with a capacity of 3100mAh), and the system was working at the 24/7 online mode. The tested results are shown in Fig. 7. In Fig. 7, ECG sampling rates are set as 300, 400 500, 600, and 700 Hz, respectively. Comparing with NPM, our proposal can extend the battery life more than 7.5 h, when the sampling rate is lower than 500 Hz. This is significant because the 2-h continuation capability is regarded as the benchmarked gain for consumer electronic devices [28]. This gain is benefited from the favorable features of the MEP algorithm. Since clinical information is preserved, ECG data analysis can be performed after MEP processing.

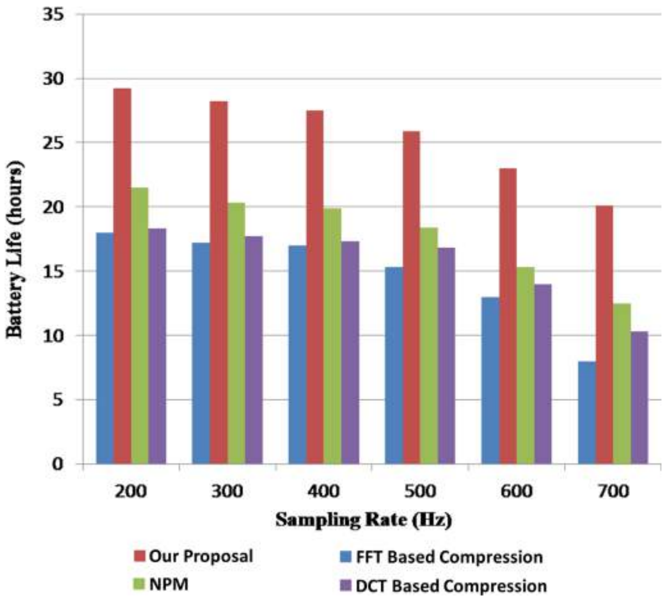


Fig. 7. Terminal battery life with different algorithms.

In contrast, compression methods do shorten the battery life when compared with NPM. This is because the compression methods need more extra power to support data processing, specifically, in which local data analysis and computing are still performed based on the raw data before compression or after decompression processes.

To summarize, let us take the typical sampling rate of 500 Hz as an example. Our proposal extends battery life by 40.54% compared with NPM and 59.50% compared with DCT-based compression. This is valuable to improve user experience because battery recharge frequency can be reduced by half, and potential failures related with power continuation can be decreased largely.

## V. Conclusion

To offer early-warning services to CVD patients, we developed a WE-CARE mHealth system to offer real-time risk monitoring. In such online risk-monitoring systems, probabilities of service discontinuation and alert missing events should be avoided or minimized. Unfortunately, system experiments and clinical trials demonstrate that the WE-CARE system operation is often crashed due to overload. If simply reducing transmitted data amount using compression methods, the alert missing events are becoming sensitive. To handle this dilemma, we studied a new research topic: system light-loading technology, which can reduce data amount without loss of clinical information. For this objective, we proposed MEP algorithm, to purify clinical features from ECG raw data based on manifold learning. Compared with conventional data compression approaches, our solution can remove redundant information from ECG raw data while required clinical information is preserved. This performance gain benefits from the ECG-feature purification principles in our algorithm design, e.g., statistical feature configuration, clinical-featured locality preservation, cluster-based detection, etc. Experiments and clinical trials demonstrate that our proposal is an effective and efficient light-loading technology to a real-time wireless mobile medical/health system for health risk alert. With the light-loading technology of our proposal, the WE-CARE system can be enabled to serve CVD patients 24/7 in real-time and avoid to miss a risk alert event.

## References

[1] *Central Government Report: Report on Cardiovascular Disease in China 2010*. Beijing, China: Ministry of Chinese Public Health, Jan. 2011

[2] *Central Government Report: Report on Cardiovascular Disease in China 2011*. Beijing, China: Ministry of Chinese Public Health, Jan. 2012

[3] B. N. Steele, M. T. Draney, J. P. Ku, and C. A. Taylor, "Internet-based system for simulation-based medical planning for cardiovascular disease," *IEEE Trans. Inf. Technol. Biomed.*, vol. 7, no. 2, pp. 123–129, Jun. 2003.

[4] H. Cao, H. Li, L. Stocco, and V. C. M. Leung, "Wireless three-pad ECG system: Challenges, design, and evaluations," *J. Commun. Netw.*, vol. 13, no. 2, pp. 113–124, Apr. 2011.

[5] C. Capua, A. Meduri, and R. Morello, "A smart ECG measurement system based on web-service-oriented architecture for telemedicine applications," *IEEE Trans. Instrum. Meas.*, vol. 59, no. 10, pp. 2530–2538, Oct. 2010.

[6] J. C. Hsieh, K. C. Yu, H. C. Chuang, and H. C. Lo, "The clinical application of an XML-based 12 lead ECG structure report system," in *Proc. 2009 Comput. Cardiol.*, Park City, UT, USA, Sep. 13–16, pp. 533–536.

[7] R. D. Chiu and S. H. Wu, "A BAN system for realtime ECG monitoring: From wired to wireless measurements," in *Proc. 2011 IEEE Wireless Commun. Network. Conf. (WCNC)*, Cancun, Mexico, Mar. 28–31, pp. 2107–2112.

[8] Global survey report. (2011). *mHealth: New horizons for health through mobile technologies*, [Online]. Available: www.who.int/goe/publications/goe_mhealth_web.pdf, World Health Organization

[9] D. Estrin and I. Sim, "Open mHealth architecture: An engine for health care innovation," *Science*, vol. 330, no. 6005, pp. 759–760, Nov. 2010.

[10] F. Collins. (2012, Jul.). "The real promise of mobile health apps: Mobile devices have the potential to become powerful medical tools," *Scientific american* [Online]. Available:http://www.scientificamerican.com/article.cfm?id=real-promise-mobile-health-apps

[11] A. Huang, C. Chen, K. Bian, X. Duan, M. Chen, H. Gao, C. Meng, Q. Zheng, Y. Zhang, B. Jiao, and L. Xie, "WE-CARE: An intelligent mobile telecardiology system to enable mHealth applications," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 2, pp. 693–702, Mar. 2014.

[12] W. Einthoven, G. Fahr, and A. Waart, "Uber die richtung und die manifeste grosse der potentialschwankungen im menschlichen herzen und uber den einfluss der herzlage auf die form des elektrokardiogramms," *Pfluegers Arch.*, vol. 150, pp. 275–315, 1913. (Translation: Hoff HE, Sekelj P. *Am. Heart J.* 1950-40: 163–194)

[13] X. He, S. Yan, Y. Hu, P. Niyogi, and H. J. Zhang, "Face recognition using Laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.

[14] J. A. Costa and A. O. Hero III, "Manifold learning using Euclidean k-nearest neighbor graphs [image processing examples," in *Proc. IEEE Int. Conf. onAcoust.*, Speech, Signal Process., Montreal, QC, Canada, May 17–21, 2004, vol. 3, pp. 988–991.

[15] Z. Li, W. Xu, A. Huang, and M. Sarrafzadeh, "Dimensionality reduction for anomaly detection in electrocardiography: A manifold approach," in *Proc. 9th Int. Conf. Wearable Implantable Body Sensor Netw.*, London, U.K., May 9–12, 2012, pp. 161–165.

[16] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Trans. Biomed. Eng.*, vol. BME-32, no. 3, pp. 230–236, Mar. 1985.

[17] J. Carlson, R. Johansson, and S. B. Olsson, "Classification of electrocardiographic P-wave morphology," *IEEE Trans. Biomed. Eng.*, vol. 48, no. 4, pp. 401–405, Apr. 2001.

[18] J. Boineau, R. Schuessler, C. Mooney, A. Wylds, C. Miller, R. Hudson, J. Borremans, and C. Brockus, "Multicentric origin of the atrial depolarization wave: The pacemaker complex. relation to dynamics of atrial conduction, p-wave changes and heart rate control," *Circulation*, vol. 58, pp. 1036–1048, Nov. 1978.

[19] H. Kantz and T. Schreiber, "Human ECG: Nonlinear deterministic versus stochastic aspects," *IEE Proc. Sci., Meas. Technol.*, pp. 279–284, Nov. 1998.

[20] S. T. Roweis and L. K. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 22, 2000.

[21] V. Yasami, S. Khorsandi, S. P. Mozaffari, and A. Jalalian, "An unsupervised network anomaly detection approach by k-Means clustering & ID3 algorithms," in Marrakech, Morocco, Jul. 6–9, 2008, pp. 398–403

[22] L. Saul and S. Roweis, "Think globally, fit locally: Unsupervised learning of low dimensional manifolds," *J. Mach. Learn. Res.*, vol. 4, pp. 119–155, Dec. 2003.

[23] C. Yang, F. Deng, and H. Yang, "An Unsupervised Anomaly Detection Approach using Subtractive Clustering and Hidden Markov Model," in *Proc. 2nd Int. Conf. Commun. Network. China, 2007*, Shanghai, China, Aug. 22–24, 2007, pp. 313–316.

[24] M. Inaba, N. Katoht, and H. Imai, "Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering," in *Proc. 10th ACM Symp. Comput. Geometry*, New York, NY, USA, Jun. 6–8, 1994, pp. 332–339.

[25] A. A. Shinde and P. Kanjalkar, "The comparison of different transform based methods for ECG data compression," in *Proc. 2011 Int. Conf. Signal Process., Commun., Comput. Network. Technol.*, Thuckalay, India, Jul. 21–22, 2011, pp. 332–335.

[26] B. H. Kwan and R. Paramesran, "Comparison between Legendre moments and DCT in ECG compression," in *Proc. IEEE Region 10 Conf.*, Chiang Mai, Thailand, Nov. 21–24, 2004, vol. 1, pp. 167–170.

[27] 3GPP TS 23.203, "Policy and charging control architecture," [Online]. Available: http://www.3gpp.org/specifications, checked in Nov. 2013

[28] F. Ferguson, B. Chen, and A. Mauskar, "Power Management in High-Level Design," in *Proc. 2nd Int. Conf. ASIC, 1996*, Shanghai, China, Oct. 21–24, 1996, pp. 357–363.

**Anpeng Huang** (M'05) received the M.S. degree from the University of Electronic Science and Technology of China, Chengdu, China, in July 2000, and the Ph.D. degree from Peking University, Beijing, China, in June 2003.

From July 2003 to April 2004, he was an Advanced Engineer at Zhongxing Telecommunication Equipment Corporation in China. From May 2004 to January 2005, he was a Visiting Scholar at the University of Waterloo, Waterloo, ON, Canada. From February 2005 to March 2008, he was a Postdoctoral Researcher in the Department of Computer Science, University of California, Davis (UC Davis), CA, USA. Since November 2007, he has been an Associate Pofessor in the State Key Lab of Advanced Optical Communication Systems and Networks, Wireless Communications Lab, and PKU-UCLA joint research institute, Peking University. He has more than 40 journal and conference papers, is the holder of 36 patents and US pending patents (with PCT application). His research interests include mobile health, mobile networks, and optical networks. His research interest focuses on mobile health.

Dr. Huang was the advisor of "Best Student Paper Award" winner at 2012 14th IEEE HEALTHCOM conference, the winner of 2012 OKAWA foundation international research grant, the winner of 42th Geneva Invention Award in 2014, and the founder of mobile health lab in PKU.

**Wenyao Xu** (M'13) received the Ph.D. degree from the Department of Electrical Engineering, University of California, Los Angeles, CA, USA, in 2013.

He is currently an Assistant Professor with the Department of Computer Science and Engineering, University at Buffalo, the State University of New York (SUNY), New York, NY, USA. His current research interests include embedded system design, computational modeling, algorithm design, human computer interaction, integrated circuit design technologies, and their applications in medical and health applications.

Dr. Xu received the Best Medical and Performance Application Paper Award of the IEEE Conference on Implantable and Wearable Body Sensor Networks in 2013 and the Best Demonstration Award of ACM Wireless Health Conference in 2011.

**Zhinan Li** received the B.S. degree in computer science from the School of Electronic Engineering, Peking University (PKU), Beijing, China, in July 2010, and the M.S. degree from the Joint Research Institute of PKU and University of California, Los Angeles, CA, USA, in July 2013.

Since July 2013, he has been working for Microsoft. His research interests include mHealth system design, data mining, and search engine technologies.

**Linzhen Xie** (M'13) received the B.S. degree from Peking University, Beijing, China, in 1963,

He has been a Professor at Peking University in China since 1978, and was a Visiting Scholar at Berkeley EECS, University of California, CA, USA, from 1980 to 1982. He is the founder of the State Key Laboratory of Advanced Optical Communication Systems and Networks at Peking University, China. One of his Ph.D. students was the winner of "100 Distinguished Ph.D. Dissertations in China" in 2000. He has published more than 140 papers in journals and at conferences in these areas, and is the holder of 21 patents. His research interests include optical network and switching, optical waveguide technology, and wireless communications.

**Majid Sarrafzadeh** (F'96) received the Ph.D. degree in 1987 from the University of Illinois at Urbana-Champaign, Urbana, IL, USA, in electrical and computer engineering.

He joined Northwestern University as an Assistant Professor in 1987. In 2000, he joined the Department of Computer Science, University of California at Los Angeles (UCLA). He is a Cofounder and Codirector of the UCLA Wireless Health Institute. His recent research interests include area of embedded computing with emphasis on healthcare. He has published approximately 450 papers, co-authored 5 books, and is a named inventor on 15 US patents. He has collaborated with many industries in the past 25 years. He co-founded two companies around 2000—they were both acquired around 2004. He has recently co-founded two companies both in the area of Technology in Healthcare.

**Xiaoming Li** (SM'03) is a professor of Peking University. His current research interests include web search and mining, data analytics, and mobile computing. He is a fellow of CCF.

**Jason Cong** (F'00) received the B.S. degree in computer science from Peking University in 1985, the M.S. and Ph.D. degrees in computer science from the University of Illinois at Urbana-Champaign in 1987 and 1990, respectively.

Currently, he is a Chancellors Professor at the Computer Science Department of University of California, Los Angeles, the director of Center for Domain-Specific Computing (CDSC), co-director of UCLA/Peking University Joint Research Institute in Science and Engineering, and co-director of the VLSI CAD Laboratory. He served as the chair the UCLA Computer Science Department from 2005 to 2008. His research interests include synthesis of VLSI circuits and systems, programmable systems, novel computer architectures, nano-systems, and highly scalable algorithms.

Dr. Cong has graduated 31 PhD students. Many of them are now faculty members in major research universities, including Cornell, Georgia Tech., Peking University, Purdue, SUNY Binghamton, UCLA, UIUC, and UT Austin. He has successfully co-founded three companies for technology transfer, including Aplus Design Technologies (acquired by Magma in 2003, now part of Synopsys), AutoESL Design Technologies (acquired by Xilinx in 2011), and Neptune Design Automation (acquired by Xilinx in 2013). He is also a distinguished visiting professor at Peking University. He has over 350 publications in these areas, including 10 best paper awards, and the 2011 ACM/IEEE A. Richard Newton Technical Impact Award in Electric Design Automation. He was ACM Fellow in 2008. He is the recipient of the 2010 IEEE Circuits and System (CAS) Society Technical Achievement Award "For seminal contributions to electronic design automation, especially in FPGA synthesis, VLSI interconnect optimization, and physical design automation."