The following publication G. Wang, Z. Deng and K. -S. Choi, "Tackling Missing Data in Community Health Studies Using Additive LS-SVM Classifier," in IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 2, pp. 579-587, March 2018 is available at https://doi.org/10.1109/JBHI.2016.2634587.

Tackling missing data in community health studies using additive LS-SVM classifier

Guanjin Wang, Zhaohong Deng, *Senior Member, IEEE*, and Kup-Sze Choi, *Member, IEEE*

***Abstract*—Missing data is a common issue in community health and epidemiological studies. Direct removal of samples with missing data can lead to reduced sample size and information bias, which deteriorates the significance of the results. While data imputation methods are available to deal with missing data, they are limited in performance and could introduce noises into the dataset. Instead of data imputation, a novel method based on additive least square support vector machine (LS-SVM) is proposed in this paper for predictive modeling when the input features of the model contain missing data. The method also determines simultaneously the influence of the features with missing values on the classification accuracy using the fast leave-one-out cross-validation strategy. The performance of the method is evaluated by applying it to predict the quality of life (QOL) of elderly people using health data collected in the community. The dataset involves demographics, socioeconomic status, health history and the outcomes of health assessments of 444 community-dwelling elderly people, with 5% to 60% of data missing in some of the input features. The QOL is measured using a standard questionnaire of the World Health Organization. Results show that the proposed method outperforms four conventional methods for handling missing data – case deletion, feature deletion, mean imputation and K-nearest neighbor imputation, with the average QOL prediction accuracy reaching 0.7418. It is potentially a promising technique for tackling missing data in community health research and other applications.**

***Index Terms*—Missing data, community health, predictive models, support vector machine, quality of life**

I. Introduction

Missing data is a common problem in community health and epidemiological studies. It does not only reduce the amount of samples available for analysis, but also introduces bias into the studies. Missing data can be caused by various reasons, e.g. respondents refuse to provide information due to privacy, withdraw or relocate in the middle of the study, or interviewers miss out some questions due to carelessness. As the recovery of missing data is usually difficult, especially after the completion of the study, it would be helpful if the incomplete dataset can still be used to produce useful results.

In studies concerning outcome prediction by pattern classification, the process involves the handling of missing values in the dataset and the construction of the pattern classification models. In general, there are four categories of approaches developed for handing missing data [1]. The first category simply discards the samples that contain missing data [2] and construct the classification models using only the samples with complete data. This method results in a loss of information and the introduction of bias into the analysis, especially when the missing data are not randomly distributed [2, 3].

The second category “recovers” the dataset by imputing the missing data with surrogate values and then constructs the classification models. Here, imputation can be achieved statistically by using mean imputation [4] or regression-based imputation [2]. For a sample where the value of a feature is missing, mean imputation fills in the missing value by the average of the values of that feature in the other samples. However, this method does not consider the correlations with the other features of the samples in the dataset [4]. In regression-based imputation, the feature with missing values is estimated by a regression model constructed using the remaining samples where the data are complete. Multivariate regression model is applied when there are two or more features containing missing values, while it is highly dependent on the nature of the data [2]. The choice between linear or non-linear regression is determined by the feature dependency relationship. This method takes the correlations between the features into consideration but the imputation is only based on a regression curve and limited by the inherent variation of the data [2, 4].

Besides, imputation can also be achieved using machine learning techniques to construct predictive models for estimating the missing values, e.g. K-nearest neighbor (KNN), self-organizing map (SOM) [5] and neural networks [6]. Among them, KNN is a common method where K nearest neighbors are selected from the group of samples with complete data to estimate the missing feature values. It outperforms other machine learning methods like decision trees (e.g. C4.5 and CN2) and mean/mode imputation methods, even when the amount of missing data is large [3]. In a DNA research, KNN is also found to be superior to mean imputation and imputation based on singular value decomposition [7]. However, the main drawback of KNN is

G. Wang is with the Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hung Hom, Hong Kong (e-mail: guanjin.br.wang@connect.polyu.hk).

Z. Deng is with the School of Digital Media, Jiangnan University, Wuxi, Jiangsu, China (e-mail: dzh666828@aliyun.com).

K. S. Choi is with the Centre for Smart Health, School of Nursing, and the Interdisciplinary Division of Biomedical Engineering, The Hong Kong Polytechnic University, Hong Kong (e-mail: hskschoi@polyu.edu.hk).

that its performance is dependent on the setting of a number of parameters, including the value of K, the distance function and the weighting function, which cannot be readily determined using theoretical approaches. In addition, the search for the nearest neighbors, i.e. the most similar cases, within the portion of samples with complete data is computationally expensive.

The third category is model-based procedures which estimate the input data distribution and use it for pattern classification. A popular approach is to develop mixture models with the expectation maximization (EM) algorithm [8] to estimate the data distribution. Once the model is constructed, the Bayes decision theory is used for classification [9]. Despite the effectiveness, computational complexity presents a difficulty to this category of methods. For example, the calculation of the standard errors of the estimates [10], and the Monte Carlo implementation of the EM algorithm (MCEM) for modeling the joint distribution of the covariates [11] are complicated that limits the applicability of the method.

The fourth category deals with missing data and constructs the pattern classification model at the same time, thus avoiding the need of imputation. For example, neural network ensembles have been designed for classification of incomplete data [12-14]. Several complete sub-datasets can be generated from the dataset with missing data to serve as the training datasets for the neural networks [15]. This method maximizes the use of information in the dataset while preserving the characteristics of the original data as much as possible without making any assumption on the data distribution. Decision trees (e.g. ID3, C4.5 and CN2) are also commonly used to deal with missing data in the training and testing datasets simultaneously. Besides, fuzzy approaches have also been proposed where fuzzy rule based classifiers are used for handing missing data. This can be achieved using on fuzzy C-means algorithms [16] or one-dimensional membership functions [17]. In the latter, a set of rules are defined for each class and each of the rules are further reduced to one-dimensional membership functions based on the fuzzy sets. These one-dimensional membership functions facilitate the generation of rules for the inputs with missing data for classification.

Recently, support vector machines (SVMs) are also extended for handling missing data [18-22]. For example, SVM and Gaussian processes are combined such that the estimation of missing values is equivalent to the identification of efficient optimization methods (e.g. EM algorithm) [19]. A max-margin learning framework is proposed using geometrically-inspired objective function to directly classify incomplete data at reduced computational cost [20]. Standard SVM classifier is also extended for classifying missing data by using probabilistic classification constraints [22].

An increasing amount of studies have been conducted to improve the classification performance using the fourth category of approaches. Many of them have demonstrated effectiveness in categorizing datasets with missing data using machine learning approaches [1]. In addition, the influence of the features with missing values on the classification performance can be readily evaluated at the same time during pattern classification. The information gives the relative importance of those features and provides insights into the data collection process. The approach proposed in this paper belongs to this category.

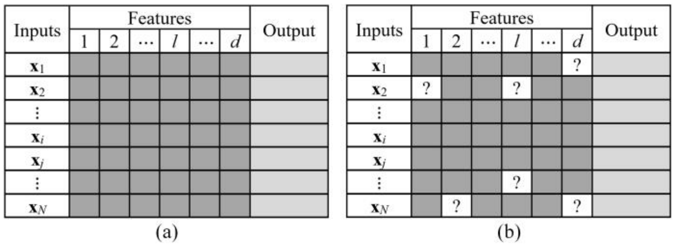
This study attempts to address the problem of missing data by developing a classifier based on additive Least Square Support Vector Machine (LS-SVM) [23, 24]. The additive LS-SVM classifier integrates LS-SVM with additive Gaussian kernels, where the classification of data with missing values is performed simultaneously with the evaluation of the influence of the features containing missing entries on the classification performance. The performance of the proposed method is demonstrated by the prediction of the quality of life (QOL) of elderly people using health data, with missing entries, collected in the community. The data attributes are of mixed types and with complex nonlinear interactions. In this study, the proposed additive LS-SVM classifier also handles missing data by the application of a fast leave-one-out cross-validation strategy on the training dataset. The proposed classifier inherits a desirable feature of the LS-SVM that analytical solution of the corresponding convex optimization problem is available by minimizing the LS-SVM based objective function [23]. The analytical solution is indeed essential for the development of the fast leave-one-out cross-validation strategy in the study to determine the influence of the features with missing values on the classification error. The performance of the proposed classifier is compared with that of the existing data imputation approaches, namely, case deletion, feature deletion, mean imputation and KNN imputation.

II. METHODS

Strong generalization capability is always desirable to pattern classification models. This is also of particular significance for models developed using data with missing values. To achieve this goal, cross validation is a strategy that is often used to determine the appropriate parameters such that an unbiased and generalized classification model can be obtained. In this paper, a novel additive LS-SVM based classifier is proposed using additive kernel functions, where a fast leave-one-out cross-validation strategy is also employed to estimate the upper bounds of the influences of the features with missing values on the classification error.

A. *Additive LS-SVM based classifier*

Without loss of generality, a binary classification task is first considered. Given a training dataset $\mathbf{T}$ with N samples, an input dataset X and the corresponding output dataset Y, where $\mathbf{T}=\{(\mathbf{x_1},y_1),...,(\mathbf{x_N},y_N)\}\in(X\times Y)$, $\mathbf{x}_i=(x_1^i,x_2^i,x_l^i,...,x_d^i)\in X\subset\mathbf{R}^d$ and $y_i\in Y=\{+1,-1\}$. Here, the input dataset X is associated with two separate classes with the class labels +1 and -1 stored in the output dataset Y, and each input sample $\mathbf{x}_i$ contains d features. Fig. 1(a) shows the scenario of a general pattern classification problem, where all the features contain a value. On the contrary, the scenario shown in Fig. 1(b) illustrates pattern classification on an incomplete dataset, where the values for some of the features in some of the samples are missing, as denoted by the symbol ‘?’.



Inputs	Features						Output
	1	2	⋯	l	⋯	d	
$\mathbf{x}_1$							
$\mathbf{x}_2$							
⋮							
$\mathbf{x}_i$							
$\mathbf{x}_j$							
⋮							
$\mathbf{x}_N$							

(a)

Inputs	Features						Output
	1	2	⋯	l	⋯	d	
$\mathbf{x}_1$						?	
$\mathbf{x}_2$	?			?			
⋮							
$\mathbf{x}_i$							
$\mathbf{x}_j$							
⋮				?			
$\mathbf{x}_N$		?				?	

(b)

Fig. 1. Pattern classification on (a) complete and (b) incomplete dataset.

In the proposed additive LS-SVM classifier, the upper bound of the classification error due to the missing value of the lth feature is denoted as c_l, with $l = 1,2,\dots,d$. The upper bound c_l is determined simultaneously during the process of classification using the fast leave-one-out cross-validation method (to be discussed in Section II-B). For the ith sample, the upper bound of the *total classification error* caused by all those features with missing values is given by $\sum_{l=1}^{d} c_l I_l^i$, where I_l^i is an indicator defined as

$$I_l^i = \begin{cases} 1 & \text{if the value of the } l^{th} \text{ feature in } \mathbf{x}_i \text{ (i.e. } x_l^i \text{) is mising} \\ \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Refer to the standard formulation of the LS-SVM [25],

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{N}\xi_i^2 \tag{2}$$

$$s.t.\ y_i = \mathbf{w}^T\phi(\mathbf{x}_i) + b + \xi_i,\ i = 1,2,\dots,N,$$

where C is the regularization parameter and ξ_i is the slack variable for $\mathbf{x}_i$, $\mathbf{w} = (w_1, w_2, \dots, w_d)$ is the d-dimensional weight vector, $\phi(\cdot)$ is the mapping function that maps $\mathbf{x}_i$ into a high dimensional feature space and b is the bias term.

When data are missing, by introducing the upper bound c_l as defined above, the formulation of the LS-SVM becomes

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{N}\xi_i^2 \tag{3}$$

$$s.t.\ y_i = \mathbf{w}^T\phi(\mathbf{x}_i) + b + \sum_{l=1}^{d} c_l I_l^i + \xi_i,\ i = 1,2,\dots,N$$

which is mathematically equivalent to

$$\min \frac{1}{2}\mathbf{w}^2 + \frac{C}{2}\sum_{i=1}^{N}\left(\xi_i - \sum_{l=1}^{d} c_l I_l^i\right)^2 \tag{4}$$

$$s.t.\ y_i = \mathbf{w}^T\phi(\mathbf{x}_i) + b + \xi_i, i = 1,2,\dots,N$$

where $\varphi(\mathbf{x}_i) = \left(\tilde{\varphi}(x_1^i), \tilde{\varphi}(x_2^i), \dots, \tilde{\varphi}(x_d^i)\right)$ and $\tilde{\varphi}(x_l^i)$ is a feature mapping such that the kernel K below can be adopted in (4). That is,

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j) = \sum_{l=1}^{d} k(x_j^i, x_l^j), \tag{5}$$

where

$$k(x_l^i, x_l^j) = \begin{cases} \tilde{k}(x_l^i, x_l^j) & \text{both } x_l^i \text{ and } x_l^j \text{ are not missing} \\ \\ 0 & \text{otherwise} \end{cases}$$

and $\tilde{k}(x_l^i, x_l^j)$ is a kernel function. In this paper, Gaussian function is employed as the kernel, i.e., $\tilde{k}(x_l^i, x_l^j) = e^{-(x_l^i - x_l^j)^2 / \sigma^2}$, where σ is the kernel width. Obviously, $K(\mathbf{x}_i, \mathbf{x}_j)$ in (5) is an additive Gaussian kernel [4, 13] which can be readily used to calculate the corresponding values of the kernel depending on whether the features contain missing values or not, and consequently introducing the upper bounds of the influences of features into Eq.(3).

In fact, it can be seen that after subtracting the total classification error caused by the features that contain missing entries in the training dataset, the aim of the primal problem in (4) is essentially to minimize the total classification error caused by all the features *without* missing values. Equation (4) is reduced to the standard LS-SVM when there are no missing values in the all samples of the training dataset, i.e., all I_l^i is zero. The advantage of the LS-SVM framework in (4) is that its analytical solution can be used to develop the fast leave-one-out cross-validation method and determine the upper bound of the classification error c_l. The Lagrangian J of (4) is given by

$$J = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{N}\left(\xi_i - \sum_{l=1}^{d} c_l I_l\right)^2 + \sum_{i=1}^{N}\alpha_i\left(y_i - \mathbf{w}^T\phi(\mathbf{x}_i) - b - \xi_i\right) \tag{6}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_N)$ is a vector of all the Lagrangian multipliers. The optimality conditions with respect to $\mathbf{w}$, ξ and α_i are

$$\frac{\partial J}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{N}\alpha_i\phi(\mathbf{x}_i), \tag{7}$$

$$\frac{\partial J}{\partial \xi_i} = 0 \Rightarrow \xi_i = \frac{1}{C}\alpha_i + \sum_{l=1}^{d} c_l I_l^i, \text{ and} \tag{8}$$

$$\frac{\partial J}{\partial \alpha_i} = 0 \Rightarrow y_i = \mathbf{w}^T\varphi(\mathbf{x}_i) + b + \xi_i. \tag{9}$$

By combining (7), (8) and (9), the system of linear equations below can be obtained.

$$\sum_{i=1}^{N}\alpha_i\varphi(\mathbf{x}_i)^T\varphi(\mathbf{x}_j) + b + \frac{1}{C}\alpha_i = y_i - \sum_{l=1}^{d} c_l I_l^i \tag{10}$$

By using (5), (10) can be further written in the compact matrix form as

$$\begin{bmatrix} \mathbf{K}+\frac{1}{C}\mathbf{\Lambda} & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix}=\begin{bmatrix} \mathbf{y}-\sum_{l=1}^{d} c_l \mathbf{I}_l \\ 0 \end{bmatrix}, \tag{11}$$

where $\mathbf{y}$ is the actual label vector of all the samples in the training dataset, i.e. $\mathbf{y}=(y_1, y_2,..., y_N)^T$, $\mathbf{\Lambda}$ is a diagonal matrix with unity diagonal entries, and $\mathbf{I}_l=(I_l^1, I_l^2,..., I_l^N)^T$. Finally, let $\mathbf{H}$ be the first matrix on the left hand side of (11), the model parameters can be calculated simply by matrix inversion, i.e.,

$$\begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix}=\mathbf{Q}\begin{bmatrix} \mathbf{y}-\sum_{l=1}^{d} c_l \mathbf{I}_l \\ 0 \end{bmatrix}, \tag{12}$$

where $\mathbf{Q}=\mathbf{H}^{-1}$. Once c_l is determined for all the features, $\boldsymbol{\alpha}$ and b can be readily obtained from (12), and hence $\mathbf{w}$ and b from (7) and (10) respectively. Therefore, for a new input sample $\mathbf{x}_t$, the decision rule for its labeling prediction becomes

$$y_t = \begin{cases} \mathbf{w}^T\phi(\mathbf{x}_t)+b & \text{if } \mathbf{x}_t \text{ is not missing} \\ \mathbf{w}^T\phi(\mathbf{x}_t)+b+\sum_{l=1}^{d} c_l I_l^t & \text{otherwise} \end{cases} \tag{13}$$

To extend the proposed LS-SVM classifier presented above for multi-class classification tasks, the one-against-all LS-SVM [26] is utilized to determine the multiple decision functions that separate one class from the other. In the end, the new input data sample $\mathbf{x}_t$ is classified into the class with

$$\max_{k=1,...,M} y_k(\mathbf{x}_t), \tag{14}$$

where *M* denotes the number of the classes.

B. Fast Leave-One-Out Cross Validation

From the discussion in the previous section, it can be seen that the classification performance of the proposed LS-SVM based classifier depends on the choice of the parameter c_l. While the conventional cross-validation method is proven an unbiased estimator and has been extensively applied to determine the parameters for various algorithms, the computation is very time-consuming. A fast version of the leave-one-out cross-validation method is thus developed in this paper, which is used to determine the optimal value for c_l by using (12). With the standard LS-SVM, it is possible to formulate the leave-one-out cross validation in a closed form and the additional computational cost is negligible [27]. The proposed LS-SVM in (4) inherits this desirable property which will be discussed as follows.

By decomposing $\mathbf{H}$ into block representation with the isolation of the first row and the first column, i.e.,

$$\mathbf{H}=\begin{bmatrix} \mathbf{K}+\frac{1}{C}\mathbf{\Lambda} & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix}=\begin{bmatrix} h_{11} & \mathbf{h}_1^T \\ \mathbf{h}_1 & \mathbf{H}_{(-1)} \end{bmatrix}, \tag{15}$$

let $\boldsymbol{\alpha}_{(-i)}$ and $b_{(-i)}$ be the model parameters during the i^{th} iteration of the leave-one-out cross validation procedure. In the first iteration where the first training sample is excluded,

$$\begin{bmatrix} \alpha_{(-i)} \\ b_{(-i)} \end{bmatrix}=\mathbf{Q}_{(-1)}\left(\mathbf{y}_{(-1)}-\sum_{l=1}^{d} c_l \mathbf{I}_{l(-1)}\right), \tag{16}$$

where $\mathbf{Q}_{(-1)}=\mathbf{H}_{(-1)}^{-1}$ and $\mathbf{y}_{(-1)}=(y_2, y_3,..., y_N, 0)^T$. Let $\tilde{y}_i$ be the prediction on the i^{th} sample when this sample is removed from the training dataset, the leave-one-out prediction for the first sample is then given by

$$\tilde{y}_1=\mathbf{h}_1^T\begin{bmatrix} \alpha_{(-i)} \\ b_{(-i)} \end{bmatrix}+\sum_{l=1}^{d} c_l I_l^1 \tag{17}$$
$$=\mathbf{h}_1^T\mathbf{Q}_{(-1)}\left(\mathbf{y}_{(-1)}-\sum_{l=1}^{d} c_l \mathbf{I}_{l(-1)}\right)+\sum_{l=1}^{d} c_l I_l^i \cdot$$

Considering the last N equations in the system of equations in (11), it is clear that $\left[\mathbf{h}_1 \;\; \mathbf{H}_{(-1)}\right]\left[\boldsymbol{\alpha}^T \;\; b\right]^T=\left(\mathbf{y}_{(-1)}-\sum_{l=1}^{d} c_l \mathbf{I}_{l(-1)}\right)$ and

$$\tilde{y}_1=\mathbf{h}_1^T\mathbf{Q}_{(-1)}[\mathbf{h}_1\mathbf{H}_{(-1)}][\alpha_1,...,\alpha_N,b]^T+\sum_{l=1}^{d} c_l I_l^1 \tag{18}$$
$$=\mathbf{h}_1^T\mathbf{Q}_{(-1)}\mathbf{h}_1\alpha_1+\mathbf{h}_1^T[\alpha_2,...,\alpha_N,b]^T+\sum_{l=1}^{d} c_l I_l^1$$

From (11), the first equation of the system is $y_1-\sum_{l=1}^{d} c_l \mathbf{I}_l^1=h_{11}\alpha_1+\mathbf{h}_1^T[\alpha_2,\alpha_3,...,\alpha_N,b]^T$, and hence $\tilde{y}_1=y_1-\alpha_1(h_{11}-\mathbf{h}_1^T\mathbf{Q}_{(-1)}\mathbf{h}_1)$. Finally, with $\mathbf{Q}=\mathbf{H}^{-1}$, the block matrix inversion is given by

$$\mathbf{Q}=\begin{bmatrix} v^{-1} & -v^{-1}\mathbf{h}_1\mathbf{Q}_{(-1)} \\ \mathbf{Q}_{(-1)}+v^{-1}\mathbf{Q}_{(-1)}\mathbf{h}_1^T\mathbf{h}_1\mathbf{Q}_{(-1)} & -v^{-1}\mathbf{Q}_{(-1)}\mathbf{h}_1^T \end{bmatrix}, \tag{19}$$

where $v=h_{11}-\mathbf{h}_1^T\mathbf{Q}_{(-1)}\mathbf{h}_1$. Since the system of linear equations in (11) is insensitive to the permutations of the order of the equations,

$$\tilde{y}_i=y_i-\alpha_i/\mathbf{Q}_{ii}. \tag{20}$$

Define $\left[\mathbf{a}'^T, b'\right]^T=\mathbf{Q}\left[\mathbf{y}^T, 0\right]$, $\left[\mathbf{a}''^T, b''\right]^T=\mathbf{Q}\left[\mathbf{I}_l^T, 0\right]$ and $\boldsymbol{\alpha}=\boldsymbol{\alpha}'-\sum_{l=1}^{d} c_l \boldsymbol{\alpha}_l''$, then (20) is given by

$$\tilde{y}_i=y_i-\frac{\alpha_i'}{\mathbf{Q}_{ii}}+\frac{\sum_{l=1}^{d} c_l \alpha_{li}''}{\mathbf{Q}_{ii}}. \tag{21}$$

It can be seen from (21) that the leave-one-out prediction $\tilde{y}_i$ can be expressed in an analytical form and that $\boldsymbol{\alpha}$ depends linearly on $\mathbf{c}=(c_1, c_2, ..., c_d)$, which makes it straightforward to obtain the learning model once all the parameters c_l are selected. The best values for c_l are those producing positive values of $\tilde{y}_i y_i$ for all sample i in the training dataset. However, this would yield a non-convex solution with many local minima if the sign of $(\tilde{y}_i y_i)$ is only considered.

Instead of the conventional hinge loss function, the following loss function is used,

$$l(\tilde{y}_i, y_i) = |1 - \tilde{y}_i y_i|_+ = \left| y_i \frac{\alpha_i^{'} - \sum_{l=1}^{d} c_l \alpha_{li}^{''}}{\mathbf{Q}_{ii}} \right|_+ , \qquad (22)$$

where $|x|_+ = \max\{0, x\}$. The loss function gives the convex upper bound to the leave-one-out misclassification loss. The solutions can be obtained when $\tilde{y}_i$ has an absolute value greater than or equal to one, and has the same sign of y_i. Finally, the objective function is given by

$$\sum_{i=1}^{N} l(\tilde{y}_i, y_i)$$

(23)

s.t. $\|\mathbf{c}\|_2 \le D$,

where D is a constant and the L_2 constraint is imposed on the vector $\mathbf{c}$ such that a solution to $\mathbf{c}$ exists. The optimization process can be implemented by using the projected sub-gradient descent algorithm. The pseudo-code is given in Algorithm 1.

Algorithm 1: Projected Sub-gradient Descent Algorithm

```
Input: set $\boldsymbol{\alpha}^{'}$, $\boldsymbol{\alpha}_l^{''}$ and $\mathbf{I}_l$, $l = 1,2,\ldots, d$
Initialize: $\mathbf{c} \leftarrow \mathbf{0}$ and $t \leftarrow 1$
Repeat
    $\tilde{y}_i = y_i - \frac{\alpha_i^{'}}{\mathbf{Q}_{ii}} + \frac{\sum_{l=1}^{d} c_l \alpha_{li}^{''}}{\mathbf{Q}_{ii}}, i = 1,2,\ldots, N$
    $d_i \leftarrow \mathbf{1}\{\tilde{y}_i y_i > 0\}$, $i = 1,2,\ldots, N$
    $c_l \leftarrow c_l - \frac{1}{\sqrt{t}} \sum_{i=1}^{N} d_i y_i \frac{\alpha_{li}}{Q_{ii}}$, $l = 1,2,\ldots, d$
    If $\|\mathbf{c}\|_2 > D$ then $\mathbf{c} \leftarrow \frac{\mathbf{c}}{\|\mathbf{c}\|_2} D$
    End if
    $c_l \leftarrow \max(c_l, 0)$, $l = 1,2,\ldots, d$
    $t \leftarrow t + 1$
Until convergence
Output: $\mathbf{c}$
```

C. *Influence of the Features with Missing Values*

The influence of the individual features that contain missing values on the classification performance provides information about the relative importance of these features in the classification model, which can in turn, provides guidance to the data collection process. Refer to (12), for an M-class classification task, after obtaining the value of c_k for each of the M classes $(k = 1,2,\ldots,M)$, the fast leave-one-out cross-validation method is used to evaluate the influence of the features l that contain missing values. The approach is developed by considering the two cases below.

Case 1: If all the c_l^k of the feature l equals 0 or $\max_{k=1,2,\ldots,M} |c_l^k|$ is less than a given small positive threshold, the upper bound of the influence of the feature l can be regarded as negligible.

Case 2: If the value of $\min_{k=1,2,\ldots,M} |c_l^k|$, denoted as *Inf*, is greater than 0 or a given small positive threshold, the feature l has some influence on the classification performance. Here, *Inf* is a measure of the degree of influence. The greater the value of *Inf*, the more significant the influence of a feature with missing values on the classification performance.

III. Experiments

The proposed method was evaluated with a real-world practical application in the fields of community healthcare, where the data collected from health screening services were used to predict the quality of life (QOL) of community-dwelling elderly people. The performance was compared with that of four conventional methods of handling missing data, i.e., case deletion, feature deletion, mean imputation and K-nearest neighbor imputation.

A. *Community Health Data*

The data used in the experiment were collected from the PolyU-Henry G. Leong Mobile Integrative Health Centre (MIHC) which is a nurse-led mobile clinic in Hong Kong providing free health screening services for elderly people at the age of 60 years or above [28]. The data were collected in August 2013 from two communities, which include demographics, socioeconomic status, health history and the outcomes of several health assessments of 444 clients. The clients were also asked to complete a questionnaire about their QOL. In the dataset, each sample contained 33 features which were used to construct a predictive model of QOL by supervised learning.

Some data were missing from the dataset, which was mainly caused by (i) language barriers due to incomprehension of dialectical differences, (ii) physical frailty, hearing or cognitive impairment, (iii) clients lacking patience to finish the health assessments, (iv) time conflicts, and (v) reluctance to disclose personal information due to privacy concerns. Among the 33 features, 14 of them did not have any missing entries; the entries of 14 features were found missing in 5% of the samples, and the entries of 4 features in 5% to 10% of the samples were not available. The number of comorbidities in more than half (60.1%) of the samples were missing. That is, 19 out of the 33 features in the dataset contained missing values. Table I shows the extent of data missing for some of the features.

In the dataset, demographic data including gender, age and marital status; socioeconomic data including the type of residency, relationships with roommates and social participation; and health history data including smoking and drinking habits and chronic health conditions were available. Data obtained from a series of health assessments were also available and described as follows.

1) *Bio-measurements*

Major vital signs of the clients, e.g. body temperature, pulse rate, oxygen saturation (SpO_2), blood pressure and waist-hip ratio (WHR), were measured.

2) *Berg Balance Scale (BBS)*

The balance ability of the clients was evaluated using the BBS [29]. By performing 14 tasks that reflect the ability to balance, e.g. standing up from a sitting position and standing on one foot, a score between 0 and 4 (best) is given to each task and the total score, ranging from 0 to 56, is a measure of the overall balance ability.

3) *Timed up and go test (TUG)*

The TUG was conducted to measure basic functional mobility [30]. It is a common test used and has good reliability [31]. In the test, clients were required to rise from a chair, walk forward for 3 meters, turn around, walk back to the same chair, and finally sit on the chair. The time elapsed in two trials was recorded and the one with the shortest time was adopted.

4) *Visual analogue scale (VAS) for pain*

The VAS for pain was used to measure the extent of pain to which a client felt at the most painful location of the body. It is a vertical line of exactly 10 cm joining the statement “no pan” at the lower end and “unbearable pain” at the upper end. Clients put a mark on the VAS to express the extent of their pain [32].

5) *The 30-second chair stand test (30-s CST)*

The test was used to estimate the lower body strength and endurance that are related to demanding tasks in activities of daily living, e.g. climbing stairs, getting out of a chair or bath tub [33-35]. In this test, clients were required to repeatedly rise from a chair to a full standing position and then sit down again. The number of repetitions performed within 30 seconds was recorded.

6) *Body composition analysis*

Body mass index (BMI), skeletal muscle mass, body fat mass and body fat percentage (BFP) of the clients were recorded to assess the degrees of obesity and fitness.

7) *Handgrip strength*

The strength of handgrip was measured using a dynamometer. Clients were required to stand upright and hold the dynamometer in a hand, and squeeze it with maximum effort. Three trials were conducted respectively with the dominant and non-dominant hand, and the average strength of each hand was obtained.

8) *Quality of life*

The QOL of the clients was measured using the World Health Organization Questionnaire on Quality of Life: Short Form – Hong Kong version (WHOQOL-BREF(HK)) [36, 37]. It has good validity and reliability, and has been adopted in Hong Kong to evaluate the QOL. The questionnaire employs a 5-point Likert scale (with “1” indicating most negative response and “5” most positive) and covers 4 specific domains of QOL, i.e. physical health (7 items), psychological (8 items), social relationships (3 items) and environment (8 items). Besides, there are 2 items on the overall perception of the QOL and general health respectively. In this study, the item “overall QOL” is the output of the prediction model.

B. Data pre-processing

The data collected from the WHOQOL-BREF (HK) indicated a heterogeneous distribution in the response of the item “overall QOL”, with relatively few clients choosing “1” and “5”, which is undesirable for model training. The scores given were thus re-categorized into three classes, namely, (i) “1” indicating poor QOL, (ii) “2” for neutral QOL and (iii) “3” for good QOL, by grouping those choosing “1” and “2” in the original 5-point Likert scale into the first class and those choosing “4” and “5” into the third class, as shown in Table II.

TABLE I
EXTENT OF DATA MISSING IN SOME OF THE FEATURES (N=444)

Features	No. (%) of missing values	Mean±SD	No. (%) of patients
Age	0	75.30±7.87	
Gender	0		
Male			136(30.6)
Female			308(69.4)
Mobility	0		
Wheel chair			8(1.8)
Walking stick			53(11.9)
Independent			382(86.0)
Walking frame			1(.2)
Social participation	0		
Unengaged			99(22.3)
Partial unengaged			119(26.8)
Engaged			226(50.9)
Marital status	7 (1.6)		
Single			25(5.6)
Married			250(56.3)
Widowed			138(31.1)
Separated/divorced			24(5.4)
Residence	7 (1.6)		
Private housing			95(21.4)
Public housing			197(44.4)
Elderly home			2(0.5)
Nursing home			7(1.6)
Others			136(30.6)
Smoking habit	7 (1.6)		
Smoker			24(5.4)
Non-smoker			413(93.0)
Drinking habit	7 (1.6)		
Drinker			52(11.7)
Non-drinker			385(86.7)
Hypertension	0		
Without hypertension			185(41.7)
With hypertension			259(58.3)
Number of comorbidities	267(60.1)	2.16±1.62	

C. Classification Performance

In the experiment, five methods were used to deal with the missing data in the original dataset. They were denoted respectively as methods A to E, referring to (A) the proposed additive LS-SVM classifier, (B) case deletion, (C) feature deletion, (D) mean imputation and (E) KNN imputation respectively. In method B, all the *samples* with missing values were removed, whereas in method C, all the features with missing values were removed. In method D, for a missing feature value in a sample, it was filled with the average of the values of that feature in the other samples. In method E, a missing value was filled by the value of the corresponding feature in a sample that was nearest to the sample with missing entry [38].

For the proposed approach, i.e., method A, the missing data were handled simultaneously with the construction of pattern classification model. For methods B to E, the missing feature values were first handled, followed by the use of standard SVM to classify the data. The classification performance of the five methods was then compared.

For model construction, the original dataset was divided respectively into three subsets for training, validation and testing at the ratio of 70:15:15, such that the distribution of the three QOL classes in each subset was similar to that in the original dataset. The validation subset was used to evaluate the performance of the trained model and to estimate the model properties, such as the regularization parameter, the parameters of the additive Gaussian kernel in the proposed classifier and the radial basis function of the methods to be compared. The training and testing subsets required for evaluating the classification performance were generated using the 10-fold cross-validation in order to reduce the biasing effect that might lead to over-fit or under-fit models. The mean and standard deviation (SD) of the accuracy in the 10 runs were the calculated [39]. The experimental results are shown in Table III. The classification accuracy of method A was also compared with that of the other methods using t-test. It is found that method A exhibited better average accuracy, with p-values smaller than 0.05.

TABLE II
RE-CATEGORIZATION OF THE RESPONSES TO OVERALL QOL

Original score	Re-categorized score	QOL
1, 2	1	poor
3	2	neutral
4, 5	3	good

TABLE III
CLASSIFICATION ACCURACY OF THE FIVE METHODS

		Method				
		A	B	C	D	E
Accuracy	Mean	**0.7438**	0.7149	0.7187	0.6896	0.7052
	SD	0.0215	0.0441	0.0407	0.0163	0.0297
	Max.	0.7612	0.7447	0.7836	0.7164	0.7487
	Min.	0.7164	0.6783	0.6716	0.6643	0.6816
p-value		-	0.002	0.003	0.000	0.008

TABLE IV
INFLUENCE OF THE FEATURES WITH MISSING VALUES

Features	*Inf*
Number of co-morbidity	**0.5837**
Duration of doing exercise (each time)	0.0859
Skeletal muscle mass	0.0585
Body fat mass	0.0585
BFP	0.0585
BMI	0.0311
Roommate	0.0173
Marital status	0.0173
Residence	0.0173
Smoking habit	0.0173
Drinking habit	0.0173
VAS for pain	0.0132
Day of doing exercise (per week)	0.0125
Blood glucose	0.0124
Body temperature	0.0124
Relation with roommate(s)	0.0111
Abbreviated mental test (AMT) score	0.0072
WHR	0.0000
30-s CST test	0.0000

D. The Inf Values

As discussed in Section II-C, the influence of the features with missing values on QOL prediction can be determined using the values of Inf, which were obtained during the classification process. Table IV lists the Inf values of the features in descending order.

IV. DISCUSSION

From Table III, the proposed approach, method A showed the best classification performance. The mean accuracy was 0.7438 and the maximum accuracy was 0.7612. The prediction models developed using methods B, C, D and E had lower performance. Their mean classification accuracies were 0.7149, 0.7187, 0.6896 and 0.7052 respectively. The performance of method A was statistically better than that of the others as evidence from the results of the t-tests. The experimental results show that in this practical application, the proposed additive LS-SVM classifier outperformed the approaches that employed conventional methods for handling missing data and the standard SVM classifier for pattern classification.

The results also verified that speed of the leave-one-outcross-validation strategy adopted in the proposed method was relatively fast. Based on the mathematical treatment and settings discussed in section II-B, the running time of the proposed additive LS-SVM classifier was about 60 seconds (on a computer equipped with a 3.4 GHz Intel Core i7-4930K processor and 16 GB RAM). Such a timing performance cannot be achieved if the missing data were handed using the standard leave-one-out cross-validation strategies.

In the present study, data were missing in some of the d features in the training dataset of size *N*. Suppose it takes *t* seconds to randomly assign a certain value to c_l for all the *d* features (*l*=1,2,…,*d*) and perform classification using the standard LS-SVM classifier defined in (4). In the experiment, since the value of c_l of the l^{th} feature is selected within the range from 0.0 to 1.0 at a step size of 0.001, according to the standard leave-one-out cross-validation strategy, the running time is given by $(1/0.001)^d \times N \times t = (1000)^d Nt$. This means that it is impractical to use the standard LS-SVM classifier when d is large. For the proposed method, it only took 60 seconds to complete pattern classification although there were totally 19 features in the training dataset that contained missing entries.

The proposed additive LS-SVM classifier provides information about the influence of the features with missing entries on the classification performance with the Inf value. If the Inf value of a feature with missing entries is equal to zero or negligibly small, it can be inferred that this feature, even though there are missing values in the dataset, has little effect in the pattern classification process. As shown in Table IV, this is the case for the features “30-s CST test” and “WHR”, both with Inf equal to zero. While data were missing for these two features, the proposed method suggests that they were not significant for the prediction of QOL and the effect of missing data was minimal.

On the other hand, the Inf value of the feature 'Number of co-morbidity' was highest (0.5837) among all the features with missing entries. It can thus be inferred that this feature was important for the prediction of QOL and the effect of missing date was significant. Besides, it is found that the Inf of the features related to body composition analysis (e.g. skeletal muscle mass, body fat mass, BFP and BMI) were higher than that related to socio-demographic characteristics (e.g. marital status, residence, roommate), which were in turn higher than that related to health history (e.g. drinking habit and smoking habit) and health assessments (e.g. pain and AMT score). The results suggest that more attention should be paid to those features with high values of Inf to ensure complete collection.

V. Conclusion

In this study, a novel additive LS-SVM classifier was developed to tackle the issues of missing data that are common in community health research. In the proposed approach, the handling of missing data and the construction of pattern classification model were carried out at the same time. A fast leave-one-out cross-validation strategy was also adopted. Furthermore, the influence of the features that contain missing data on pattern classification could be evaluated with the Inf value.

To evaluate the performance of the proposed approach, experiments based on the health and socio-demographic data of elderly people were conducted to predict their QOL. The data were collected from a nurse-led mobile health center that provides primary and preventive healthcare services in the community. Out of the 33 features in the dataset, 19 of them contained missing data. Results show that the proposed approach had a higher accuracy than the methods that handled missing data with conventional techniques and applied the standard SVM as classifier. Besides, the influence of the 19 features that contained missing data reflected the relative importance of these features in the prediction of QOL, highlighting the ones with high Inf values for more attention to ensure complete data collection.

However, the study has some limitations. First, the data used to build the predictive model were collected from two communities. The model thus developed using the data may not be representative of a broader population of elderly people. Second, in the study, the item "overall QOL" of the questionnaire WHOQOL-BREF (HK) is used as the output of the classification model. While the underlying framework of the questionnaire is based on four domains of QOL (i.e. physical health, psychological, social relationships and environment), each of which further encompasses specific attributes under, the data collected from the mobile clinic indeed do not correspond exactly to the attributes considered by the framework of the questionnaire. Nevertheless, to demonstrate the performance of the proposed approach in handling missing data, the dataset had satisfied this purpose and showed that the additive LS-SVM classifier is a promising technique with a practical application in community health studies. In this regard, the study of the proposed additive LS-SVM classifier will be further extended to investigate separately the associations between each of the four specific domains and the overall QOL in attempt to enhance the prediction performance. Besides, the prediction of QOL for elderly people with certain diseases, e.g. hypertension or diabetes mellitus, can be conducted with the proposed approach. This will allow for a more specific understanding of the QOL status of the elderly people with respect to the diseases, thereby enabling early administration of appropriate preventive or remedial actions.

Acknowledgment

We thank the support of the PolyU-Henry G. Leong Mobile Integrative Health Centre which is donated the Tai Hung Fai Charitable Foundation. G. Wang is supported by the Y.C. Yu Scholarship for Centre for Smart Health.

References

[1] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: a review," *Neural Computing and Applications,* vol. 19, pp. 263-282, 2010// 2010.

[2] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. New York: John Wiley, 2014.

[3] G. E. Batista and M. C. Monard, "An analysis of four missing data treatment methods for supervised learning," *Applied Artificial Intelligence,* vol. 17, pp. 519-533, 2003/05/01 2003.

[4] J. L. Schafer, *Analysis of Incomplete Multivariate Data*. London, England: Chapman & Hall, 1997.

[5] F. Fessant and S. Midenet, "Self-organising map for data imputation and correction in surveys," *Neural Computing & Applications,* vol. 10, pp. 300-310, 2002.

[6] P. J. García-Laencina, J. Serrano, A. R. Figueiras-Vidal, and J.-L. Sancho-Gómez, "Multi-task neural networks for dealing with missing inputs," in *International Work-Conference on the Interplay Between Natural and Artificial Computation*, 2007, pp. 282-291.

[7] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani*, et al.*, "Missing value estimation methods for DNA microarrays," *Bioinformatics,* vol. 17, pp. 520-525, 2001.

[8] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Magazine,* vol. 13, pp. 47-60, 1996.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (methodological),* vol. 39, pp. 1-38, 1977.

[10] N. J. Horton and K. P. Kleinman, "Much Ado About Nothing," *The American Statistician,* vol. 61, pp. 79-90, 2007/02/01 2007.

[11] J. G. Ibrahim, M.-H. Chen, and S. R. Lipsitz, "Monte Carlo EM for Missing Covariates in Parametric Regression Models," *Biometrics,* vol. 55, pp. 591-596, 1999.

[12] P. Juszczak and R. P. Duin, "Combining one-class classifiers to classify missing data," in *International Workshop on Multiple Classifier Systems*, 2004, pp. 92-101.

[13] J. Kai, C. Haixia, and Y. Senmiao, "Classification for Incomplete Data Using Classifier Ensembles," in *2005 International Conference on Neural Networks and Brain*, 2005, pp. 559-563.

[14] S. Krause and R. Polikar, "An ensemble of classifiers approach for the missing feature problem," in *Proceedings of the International Joint Conference on Neural Networks*, 2003, pp. 553-558 vol.1.

[15] K. Jiang, H. Chen, and S. Yuan, "Classification for incomplete data using classifier ensembles," in *2005 International Conference on Neural Networks and Brain*, 2005, pp. 559-563.

[16] H. Ichihashi and K. Honda, "Fuzzy c-means classifier for incomplete data sets with outliers and missing values," in *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)*, 2005, pp. 457-464.

[17] D. Nauck and R. Kruse, "Learning in neuro-fuzzy systems with symbolic attributes and missing values," in *Proceedings of 6th*

International Conference on Neural Information Processing, Perth, WA, Australia, 1999, pp. 142-147.

[18] K. Pelckmans, J. De Brabanter, J. A. K. Suykens, and B. De Moor, "Handling missing values in support vector machine classifiers," *Neural Networks*, vol. 18, pp. 684-692, 2005.

[19] A. J. Smola, S. Vishwanathan, and T. Hofmann, "Kernel methods for missing variables," presented at the International Workshop on Artificial Intelligence and Statistics, Barbados, 2005.

[20] G. Chechik, G. Heitz, G. Elidan, P. Abbeel, and D. Koller, "Max-margin classification of incomplete data," in *Advances in Neural Information Processing Systems 19*, 2006, pp. 233-240.

[21] J. Bi and T. Zhang, "Support Vector Classification with Input Data Uncertainty," in *Advances in Neural Information Processing Systems 17*, Cambridge, MA, 2004, pp. 161-168.

[22] P. K. Shivaswamy, C. Bhattacharyya, and A. J. Smola, "Second Order Cone Programming Approaches for Handling Missing and Uncertain Data," *Journal of Machine Learning Research*, vol. 7, pp. 1283-1314, 2006.

[23] J. A. K. Suykens, T. V. Gestel, J. D. Brabanter, B. D. Moor, and J. Vandewalle, *Least Squares Support Vector Machines*. River Edge, New Jersey: World Scientific, 2002.

[24] K. Pelckmans, I. Goethals, J. D. Brabanter, J. A. K. Suykens, and B. D. Moor, "Componentwise Least Squares Support Vector Machines," in *Support Vector Machines: Theory and Applications*, L. Wang, Ed., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 77-98.

[25] J. A. K. Suykens and J. Vandewalle, "Least Squares Support Vector Machine Classifiers," *Neural Processing Letters*, vol. 9, pp. 293-300, 1999.

[26] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *Trans. Neur. Netw.*, vol. 13, pp. 415-425, 2002.

[27] G. C. Cawley, "Leave-One-Out Cross-Validation Based Model Selection Criteria for Weighted LS-SVMs," in *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, 2006, pp. 1661-1668.

[28] K.-S. Choi, R. K. P. Wai, and E. Y. T. Kwok, "Healthcare Information System: A Facilitator of Primary Care for Underprivileged Elderly via Mobile Clinic," in *Smart Health: International Conference, ICSH 2013, Beijing, China, August 3-4, 2013. Proceedings*, D. Zeng, C. C. Yang, V. S. Tseng, C. Xing, H. Chen, F.-Y. Wang, *et al.*, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 107-112.

[29] L. Blum and N. Korner-Bitensky, "Usefulness of the Berg Balance Scale in Stroke Rehabilitation: A Systematic Review," *Physical Therapy*, vol. 88, pp. 559-566, 2008.

[30] D. Podsiadlo and S. Richardson, "The Timed “Up & Go”: A Test of Basic Functional Mobility for Frail Elderly Persons," *Journal of the American Geriatrics Society*, vol. 39, pp. 142-148, 1991.

[31] A. Shumway-Cook, S. Brauer, and M. Woollacott, "Predicting the Probability for Falls in Community-Dwelling Older Adults Using the Timed Up & Go Test," *Physical Therapy*, vol. 80, pp. 896-903, 2000.

[32] B. Wolf, H. Feys, W. De Weerdt, J. van der Meer, M. Noom, and G. Aufdemkampe, "Effect of a physical therapeutic intervention for balance problems in the elderly: A single-blind, randomized, controlled multicentre trial," *Clinical Rehabilitation*, vol. 15, pp. 624-636, June 1, 2001 2001.

[33] C. J. Jones, R. E. Rikli, and W. C. Beam, "A 30-s Chair-Stand Test as a Measure of Lower Body Strength in Community-Residing Older Adults," *Research Quarterly for Exercise and Sport*, vol. 70, pp. 113-119, 1999/06/01 1999.

[34] T. Nakatani, M. Nadamoto, and M. Itoh, "Validation of a 30-sec chair-stand test for evaluating lower extremity muscle strength in Japanese elderly adults," *Japanese Society of Physical Education*, vol. 47, pp. 451-461, 2002.

[35] D. J. Macfarlane, K. L. Chou, Y. H. Cheng, and I. Chi, "Validity and normative data for thirty-second chair stand test in elderly community-dwelling Hong Kong Chinese," *American Journal of Human Biology*, vol. 18, pp. 418-421, 2006.

[36] K. F. Leung, W. W. Wong, M. S. M. Tay, M. M. L. Chu, and S. S. W. Ng, "Development and validation of the interview version of the Hong Kong Chinese WHOQOL-BREF," *Quality of Life Research*, vol. 14, pp. 1413-1419, 2005// 2005.

[37] K. Leung, M. Tay, S. Cheng, and F. Lin, "Hong Kong Chinese Version World Health Organization Quality of Life Measure Abbreviated Version," ed: Hong Kong Hospital Authority, 1997.

[38] E. Acuña and C. Rodriguez, "The Treatment of Missing Values and its Effect on Classifier Accuracy," in *Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, 15–18 July 2004*, D. Banks, F. R. McMorris, P. Arabie, and W. Gaul, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 639-647.

[39] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," presented at the Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2, Montreal, Quebec, Canada, 1995.

Guanjin Wang received both bachelor and master degrees in Information Technology from Monash University, Australia in 2012 and 2014 respectively. She is currently a joint PhD student in the Faculty of Engineering and Information Technology at the University of Technology, Sydney and Centre of Smart Health, school of Nursing at the Hong Kong Polytechnic University. Her research interest includes, machine learning, computational intelligence and health informatics.

Zhaohong Deng (M’2012, SM’2014) received the B.S. degree in physics from Fuyang Normal College, Fuyang, China, in 2002, and the Ph.D. degree in light industry information technology and engineering from Jiangnan University, Wuxi, China, in 2008. He is currently a professor in the School of Digital Media, Jiangnan University and a visiting associate researcher in the University of California, Davis, CA, USA. His current research interests include computational intelligence and pattern recognition. He serves as the associate editor of four international editors including Neurocomputing, PLOS one. He has published about 80 papers in international/national journals.

Kup-Sze Choi (M’97) received his Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong. He is currently an associate professor at the School of Nursing, the Hong Kong Polytechnic University, and the director of the Centre for Smart Health and the PolyU-Henry G. Leong Mobile Integrative Health Centre. His research interests include computational intelligence and virtual reality, and their applications in medicine and health care.