IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON DATA-ENABLED INTELLIGENCE FOR DIGITAL HEALTH

Received February 5, 2019, accepted February 23, 2019, date of publication March 1, 2019, date of current version April 3, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2902467

IFFLC: An Integrated Framework of Feature Learning and Classification for Multiple Diagnosis Codes Assignment

YUWEN LI[1], WEITONG CHEN[2], DEYIN LIU[2,3], ZHIMIN ZHANG[2,4], SHUNXIANG WU[1], AND CHENGYU LIU[5], (Member, IEEE)

[1]Department of Automation, Xiamen University, Xiamen 361005, China
[2]School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, QLD 4072, Australia
[3]School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China
[4]School of Control Science and Engineering, Shandong University, Jinan 250100, China
[5]School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China

Corresponding authors: Shunxiang Wu (sxwu@xmu.edu.cn) and Chengyu Liu (chengyu@seu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 81871444, and in part by the China Scholarship Council under Grant 201706310048.

ABSTRACT The International Classification of Diseases, Version 9 (ICD-9) is often used to identify patients with specific diagnoses. However, certain conditions may not be accurate reflected by the ICD-9 codes, and diagnoses code assignments are complex time-consuming processes. Although there are existing methods for automotive disease diagnostic assignment techniques, they have limitations on the descriptiveness and interpretability of diseases based on features. More importantly, they ignored the importance of different features with respect to different diseases. To address the above-mentioned challenges, we propose a novel framework, namely IFFLC, which can select the most relevant features, learn disease-specific features for each disease, and perform multiple diagnosis codes' assignment. Specifically, we first develop feature selection based on disease information entropy to remove redundant and irrelevant features in both medical chart data and medical laboratory data. Then, we build a novel multiple diagnosis codes' classifier by learning the disease-specific features and exploring the intra-correlations between diseases. We employ an alternating direction method of multipliers to iteratively solve the related optimization problem. The extensive experiments on a real-world ICU database verify the superiority of the proposed method over state-of-the-art approaches.

INDEX TERMS Disease correlation, disease-specific feature learning, ICD code labeling, multi-label classification.

I. INTRODUCTION

ICD code is designed as a health care classification system [1], [2], providing a system of diagnostic codes for classifying diseases, and used in assigning diagnostic and procedures [3]–[7]. It defines the universe of diseases, disorders, injuries and other related health conditions, listed in a hierarchical structure. ICD is the international standard for reporting diseases and health conditions and the diagnostic classification standard for all clinical and research purposes that is proposed and periodically revised by the World Health Organization (WHO). In medical records learning, the ninth version ICD-9 has been widely utilized. A complete, timely and accurate diagnosis codes assignment is very important, especially in Intensive Care Unit (ICU) [8]–[11] since it can be the best practice guideline to provide better treatments for patients. However, the assignment of ICD codes to patients in ICU is traditionally done by medical professionals. The task of ICD coding is by nature complex to be completed manually, as it consists of a multi-label classification over a tree structure. To free medical professionals from time-consuming and tedious medical record reviews, a system that can automatically annotate ICD-9 codes for patients is desirable. Although there are exiting automotive disease diagnostic assignment techniques [12]–[14], code assignment is still remain as a challenge task with following reasons. Firstly, because of the high-granularity of ICD-9 code, code assignment may be inconsistent or inexact. Secondly, the correlations between corresponding diagnosis

The associate editor coordinating the review of this manuscript and approving it for publication was Yonghong Peng.

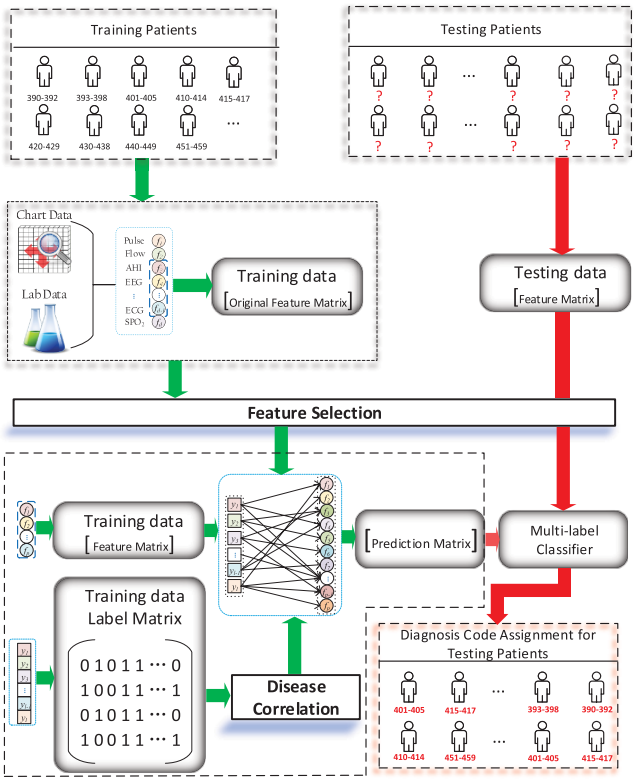


FIGURE 1. **Workflow demonstration of the proposed framework. The green arrowheads on the left represent the training process. The red arrowheads on the right represent the testing process. Two types of patient data, i.e. medical chart and lab data, are in use. Feature selection based on disease information entropy is developed to remove redundant and irrelevant features. The disease-specific features are learned and the intra-correlations between all diseases are explored. Finally, a multi-label classifier is proposed to assign diagnosis codes for testing patients.**

codes are easily ignored, resulting in each disease viewed independently and the loss of useful information. Finally, lack of selecting most relevant features for each disease leads to the uninterpretable of coding assignment.

To address above challenge, we build an integrated framework, namely IFFLC, which include three tasks: remove redundant and irrelevant features, learn disease-specific features for each disease and perform multiple diagnosis codes assignment. Firstly, we utilize fuzzy mutual information as assessment criterion to develop our feature selection algorithm. Then, we build a novel multiple diagnosis codes classifier by learning a map from the selected features to diagnosis codes. Considering each disease highly depends on only a few specific features (i.e., disease-specific features), the map matrix has sparsity involving discriminative information. Moreover, the diseases are often not independent, so the disease correlations are embedded into the map. Therefore, this classifier integrates the map sparsity and disease correlations simultaneously, and can be used to assign diagnosis codes for a new patient. Extensive experiments are carried out to make clear the effectiveness of the proposed algorithm. The demonstration of the proposed framework is shown in Figure 1. The contributions of IFFLC are summarized as follows.

- **Effective integrated framework**. We advance an integrated framework, which can select most relevant features, learn disease-specific features for each disease and perform multiple diagnosis codes assignment.
- **Obtaining disease-specific features**. We select and analyze some specific features containing discriminative information for each diseases and possess advantage with respect to interpretability.
- **Embedding diseases correlations**. We consider diseases are related to each other, and incorporate this information into our model.
- **Superior experiments results**. We conduct extensive experiments on a real-world ICU patient database, and compare the proposed method with six comparative approaches to demonstrate the effectiveness of IFFLC.

II. RELATED WORK

Automated ICD coding approaches to classification of patient records against multiple diagnosis codes fall into multi-label classification task. Multi-label classification deals with one sample having more than one labels simultaneously. In recent years, many well-established multi-label learning algorithms [15], [16] have been proposed. These algorithms can be grouped into two categories: problem transformation methods (fitting data to algorithm) and algorithm adaption methods (fitting algorithm to data). Problem transformation methods transform one multi-label learning task to several binary single-label learning tasks, each for one label, such as Binary Relevance (BR) [17] and Classifier Chain (CC) [27]. CC transforms the multi-label learning problem into a chain of binary classification problems. BR is a well-known framework for multi-label classification. It transforms one multi-label learning task to several binary single-label learning tasks. The BR approach is a simple and straight-forward solution to multi-label learning. However, it ignores label correlations which may provide helpful extra information. Algorithm adaption methods adapt or extend existing single-label algorithms to multi-label learning, which can handle multi-label data directly, such as ML-KNN [30], Multi-Label Naive Bayes classification (MLNB) [29] and RankSVM [28]. ML-KNN [30] adapts KNN to multi-label classification. MLNB [29] extends the traditional naive Bayes classifiers to deal with multi-label data. RankSVM [28] improves the maximum margin strategy of single-label classifier SVM to construct a multi-label classifier.

At present, multi-label classification model has been widely used to construct diagnosis codes assignment algorithms [10], [12]–[14], [18]–[21]. Yan *et al.* [13] introduced a multi-label large-margin classifier that automatically learnt the underlying inter-code structure and predicted diagnosis codes for patients. Ferrao *et al.* [18] proposed a methodology entailing an adaptive data processing method to support ICD coding based on structured electronic health record data and SVM. Perotte *et al.* [10] automated ICD code assignment using flat classifier and hierarchy-based classifier

based on ICD9 diagnosis codes and discharge summaries. Zufferey *et al.* [14] provided a performance comparison of state-of-the-art multi-label classification algorithms for the assignment of chronic diseases to patients' records. Baumel *et al.* [12] constructed multi-label classification to assign multiple ICD codes and high-light the elements in the clinical documents that explain and support the predicted results. These methods automatically assign diagnosis codes to patients according to their clinical records. However, they do not take into account correlations between diseases, resulting in each disease viewed independently and the loss of useful information. In practice, diseases are often related to each other. Doctors can recognize disease correlations using their professional knowledge to achieve more accurate diagnosis results. Therefore, disease correlations should be incorporated into diagnoses code assignments based on multi-label classification as doctors do.

When diseases correlations are considered, Wang *et al.* [19] proposed a multi-label classifier based on both global information and local diseases correlations to assign diagnosis code using sparsity-based disease correlation embedding. Wang *et al.* [20] used an encoded vector to locally exploit disease correlation, and built multi-label classifier with local disease correlation mining to learn multiple diagnosis codes for ICU patients. However, the above multi-label classifiers have limitations on the descriptiveness and interpretability of diseases based on features, because they automatically assign diagnostic codes without taking into account each disease highly depends on only a few specific features (i.e., disease-specific features). Disease-specific features contain the most discriminating information about the corresponding disease, and they can be used to explain and describe this disease.

According to the above analysis, in order to solve the existing problems of the existing models simultaneously, we construct a multi-label classifier with disease correlations, which can learn disease-specific features for each disease and perform multiple diagnosis codes assignment.

III. METHODOLOGY

A. DATABASE AND DATA DENOISING

MIMIC-III (Medical Information Mart for Intensive Care III) is a real-world and freely-available clinical database associated with over forty thousand patients who stayed in ICU of the Beth Israel Deaconess Medical Center between 2001 and 2012. The database includes information such as demographics, vital sign measurements made at the bedside, chart test results, laboratory test results, procedures, medications, nurse and physician notes, imaging reports, and out-of-hospital mortality. MIMIC-III supports a diverse range of analytic studies spanning epidemiology, clinical decision-rule improvement, and electronic tool development [22]. Every patient in MIMIC-III is diagnosed with one or several diseases. According to the coding scheme of ICD-9, there are 19 categories, and *diseases of the circulatory system* (390-459) have the highest morbidity. Thus, we conduct our research on this kind of diseases in this paper. For a more specific classification, *diseases of the circulatory system* (390-459) can be divided into 9 subclasses as shown in Table 1. We use these 9 subclasses as labels in multi-label learning. In this paper, we extract adult patients (16 years old) with diseases of the circulatory system in the first step. Then, we only choose 5,063 patients who stay in the ICU for 12-72 hours to obtain stable and reliable medical data values.

TABLE 1. ICD-9 codes considered for building the 9 subclasses of circulatory system diseases.

ICD-9 codes	Label diseases
390-392	Acute Rheumatic Fever
393-398	Chronic Rheumatic Heart Disease
401-405	Hypertensive Disease
410-414	Ischemic Heart Disease
415-417	Diseases of Pulmonary Circulation
420-429	Other Forms of Heart Disease
430-438	Cerebrovascular Disease
440-449	Diseases of Arteries, Arterioles, and Capillaries
451-459	Diseases of Veins and Lymphatics, and Other Diseases of Circulatory System

The medical data in MIMIC-III include two major data sources: medical chart event data and medical laboratory event data [14]. For chart event data, it contains all the charted data available for a patient and displays patients' routine vital signs and any additional information relevant to their care, such as fluid assessment, or physiological measure. For laboratory event data, it contains all laboratory measurements. Since most of the textual items in chart and laboratory event data are full of noise [19], [20], it is difficult to reflect personal health conditions. Therefore, we exclude textual items and extract 2,240 numerical items as input features of patients. Meanwhile, we exclude the patients whose chart or laboratory event data are empty or corrupted, in order to guarantee the completion of both chart and laboratory event data for each patient. As a result, we obtain 5,058 adult patients records out of 5,063.

B. PROBLEM STATEMENT

Suppose patients as sample set $\mathcal{X} = \{x_1, x_2, \cdots, x_n\}$ has a finite set of l possible diagnosis labels $\mathcal{Y} = \{y_1, y_2, \cdots, y_l\}$ with d features $\mathcal{F} = \{f_1, f_2, \cdots, f_d\}$. The features are made up of chart event data and laboratory event data. $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)|1 \leq i \leq n\}$ is a training dataset in a given multi-label dataset, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is a sample, $\boldsymbol{y}_i \in \{0, 1\}^l$ is the corresponding diagnosis label set. If x_i has the label y_j then $y_{ij} = 1$, otherwise $y_{ij} = 0$. $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]^T \in \mathbb{R}^{n\times d}$ is the sample matrix, and $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n]^T \in \{0, 1\}^{n\times l}$ is the ground truth label matrix.

C. FEATURE SELECTION FOR MULTIPLE DIAGNOSIS CODES

As mentioned above, 2,240 items constitute feature space of multiple diagnosis codes learning. However, many of the features may be redundant and/or irrelevant. The reason for the redundancy is that the chart data and laboratory data are partially duplicated. Even though laboratory values are cap-

tured elsewhere laboratory event, they are frequently repeated within chart event, according to the related description about MIMIC-III [22]. Furthermore, the reason for the existence of irrelevant features is that 2,240 items currently obtained are associated with all diseases in ICD-9 codes system, but for *diseases of the circulatory system* (390-459), some features are irrelevant obviously. The existence of redundant and/or irrelevant features degrades the performance and increases the time and space complexity in learning.

Therefore, our goal is to learn a small feature subset $\mathcal{S} = \{f'_1, f'_2, \cdots, f'_p\}(p \ll d)$ containing the most discriminative and representative information. To guarantee that $\mathcal{S}$ can achieve optimal performance, we expect it to possess two properties. 1) Max-relevance: $\mathcal{S}$ is completely relevant to the 9 subclasses of the circulatory system. 2) Min-redundancy: Features in $\mathcal{S}$ are not redundant with each other.

To achieve these goals, we develop feature selection based on disease information entropy. Motivated by Lin *et al.* [23], Fuzzy Mutual Information (FMI) is utilized as an evaluation of multi-label feature selection according to the following equation

$$\mathrm{FMI}(F_1; F_2) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_{F_1}| \cdot |[x_i]_{F_2}|}{n \cdot |[x_i]_{F_1} \bigcap [x_i]_{F_2}|}, \tag{1}$$

where $[x_i]_{F_1}([x_i]_{F_2})$ is fuzzy equivalence class associated with x_i and fuzzy set $F_1(F_2)$. The fuzzy cardinality of $[x_i]_F$ is calculated by Equation (2).

$$|[x_i]_F| = \sum_{i=1}^{n} r_{ij}, |[x_i]_{\mathcal{Y}}| = \sum_{i=1}^{n} c_{ij}, \tag{2}$$

where r_{ij} is the degree of x_i equivalent to x_j. For the diagnosis label set $\mathcal{Y}$, c_{ij} is computed by cosine similarity to map categorical label data to Euclidean space. Furthermore, based on max-relevance and min-redundancy strategy, a candidate feature f_i is selected if it has the maximal relevance with $\mathcal{Y}$, and the minimal redundancy with the selected features in $\mathcal{S}_{k-1}$. The objective function is calculated by

$$\max_{f_i \in F - S_{k-1}} [\mathrm{FMI}(f_i; \mathcal{Y}) - \frac{1}{k-1}\sum_{f_j \in S_{k-1}} (\mathrm{FMI}(f_i; f_j))]. \tag{3}$$

Based on Equation (3), we summarize the pseudo-code of feature selection for multiple diagnosis codes in Algorithm 1.

Algorithm 1 Feature Selection for Multiple Diagnosis Codes

Input:Feature set $\mathcal{F}$, sample set $\mathcal{X}$, label set $\mathcal{Y}$, candidate feature f.
Output: $\mathcal{S}$.
Initialization: $\mathcal{S} \leftarrow []$, $k \leftarrow 1$.

```
while |F| ≠ ∅ do
    find f ∈ F by maximizing Eq.(3);
    S_k ← f;
    F ← F − {f};
    k ← k + 1;
end while
return S.
```

D. IFFLC FOR MULTIPLE DIAGNOSIS CODES ASSIGNMENT

Based on Algorithm 1, we can obtain a dimensionality reduced feature subset $\mathcal{S} = \{f'_1, f'_2, \cdots, f'_p\}(p < d)$ to reconstruct $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]^T \in \mathbb{R}^{n \times p}$.

In this section, in order to capture intrinsic relationships between features and diseases for multi-label classification, a map matrix $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_l] \in \mathbb{R}^{p \times l}$ will be learned. Moreover, inspired by Zhang and Wu [16], and Huang *et al.* [24], [25], we aim to learn $\boldsymbol{W}$ to indicate the corresponding discriminative features of each disease, i.e., the disease-specific features can be captured by $\boldsymbol{W}$. Consequently, in multiple diagnosis codes assignment, we expect $\boldsymbol{W}$ has three properties: 1) $\boldsymbol{W}$ can map the selected features to diagnosis codes. 2) Disease-specific features are generated by the nonzero entries of $\boldsymbol{W}$. 3) $\boldsymbol{W}$ should contain the correlations between diseases, because similar diseases share more features.

Given the comprehensive consideration of the three properties, we obtain the following optimization problem for multiple diagnosis codes assignment:

$$\min_{\boldsymbol{W}} \mathrm{loss}(\boldsymbol{W}) + \alpha\Theta(\boldsymbol{W}) + \beta\Omega(\boldsymbol{W}), \tag{4}$$

where $\mathrm{loss}(\cdot)$ is loss function, $\Theta(\cdot)$ is built to model the disease correlations, and $\Omega(\cdot)$ is the sparsity regularisation term. $\alpha > 0$ and $\beta > 0$ are trade-off parameters.

1) DISCRIMINANT AND SPARSITY OF DISEASE-SPECIFIC FEATURES

In Problem (4), we leverage the least squared loss as the loss function $\mathrm{loss}(\cdot)$, because of its simplicity and efficiency. Thus, $\mathrm{loss}(\boldsymbol{W})$ can be formulated as

$$\mathrm{loss}(\boldsymbol{W}) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{W} - \boldsymbol{Y}\|_F^2. \tag{5}$$

In multiple diagnosis codes assignment, different diseases have distinct characteristics of their own. Each disease highly depends on only a few specific features. Compared with the feature set $\mathcal{S}$, disease-specific features are sparse. To model the sparsity of disease-specific features, l_1-norm is employed on $\boldsymbol{W}$

$$\Omega(\boldsymbol{W}) = \|\boldsymbol{W}\|_1. \tag{6}$$

If $w_{ij} = 0$, it indicates that the i-th feature has no use for the discrimination of the j-th disease. On the contrary, if $w_{ij} \neq 0$, it reveals that the corresponding feature is discriminative to the j-th disease.

2) EMBEDDING DISEASE CORRELATIONS

Diseases of circulatory system are not independent because we notice that certain diseases would always appear simultaneously. For example, *Hypertensive Disease* (401-405) is highly correlated with *Other Forms of Heart Disease* (420-429). Therefore, two strongly correlated diseases would

share more features than two weakly correlated or uncorrelated diseases. In light of this, to embed disease correlations in multi-label classification, $\Theta(\boldsymbol{W})$ can be formulated as

$$\Theta(\boldsymbol{W}) = \frac{1}{2}\sum_{i=l}^{n}\sum_{i=l}^{n} r_{ij}\boldsymbol{w}_i^T\boldsymbol{w}_j = \frac{1}{2}\mathrm{Tr}(\boldsymbol{R}\boldsymbol{W}^T\boldsymbol{W}), \tag{7}$$

where $\boldsymbol{R} = [r_{ij}]_{l\times l}$ stores the correlation information between diseases y_i and y_j. In [19], [24], and [25], $r_{ij} = 1 - c_{ij}$, and c_{ij} is calculated by cosine similarity. However, in MIMIC-III, the number of patients is far less than that of normal people with respect to each disease. Cosine similarity ignores this situation. Therefore, we redefine a new metric to measure the similarity between diseases y_i and y_j as follow

$$r_{ij} = 1 - \frac{d(\boldsymbol{y}_i, \boldsymbol{y}_j)}{\max(d(\boldsymbol{y}_s, \boldsymbol{y}_t)) - \min(d(\boldsymbol{y}_s, \boldsymbol{y}_t))}, \tag{8}$$

where

$$d(\boldsymbol{y}_i, \boldsymbol{y}_j) = (\sum_{r=1}^{l}(c_{ir} - c_{jr})^2)^{\frac{1}{2}}, \tag{9}$$

c_{ij} is cosine similarity and $s, t = 1, 2, \cdots, l$.

Combining Equations (5), (6) and (7), Problem (4) can be rewritten as:

$$\min_{\boldsymbol{W}} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{W} - \boldsymbol{Y}\|_F^2 + \frac{\alpha}{2}\mathrm{Tr}(\boldsymbol{R}\boldsymbol{W}^T\boldsymbol{W}) + \beta\|\boldsymbol{W}\|_1 \tag{10}$$

3) OPTIMIZATION WITH ALTERNATING DIRECTION METHOD OF MULTIPLIERS (ADMM)

We apply ADMM [26] to solve the optimization problem (10) in this paper.

Firstly, we introduce two auxiliary variables $\boldsymbol{U}$ and $\boldsymbol{V}$ to make the objective function separable. Problem (10) can be rewritten as:

$$\begin{aligned}&\min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{W} - \boldsymbol{Y}\|_F^2 + \frac{\alpha}{2}\mathrm{Tr}(\boldsymbol{R}\boldsymbol{U}^T\boldsymbol{U}) + \beta\|\boldsymbol{V}\|_1,\\ &\text{s.t. } \boldsymbol{U} = \boldsymbol{W}, \boldsymbol{V} = \boldsymbol{W}.\end{aligned} \tag{11}$$

Problem (11) can be transformed into its augmented Lagrangian function form:

$$\begin{aligned}&\min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{W} - \boldsymbol{Y}\|_F^2 + \frac{\alpha}{2}\mathrm{Tr}(\boldsymbol{R}\boldsymbol{U}^T\boldsymbol{U}) + \beta\|\boldsymbol{V}\|_1\\ &\quad + \langle \boldsymbol{L}_1, \boldsymbol{U} - \boldsymbol{W}\rangle + \langle \boldsymbol{L}_2, \boldsymbol{V} - \boldsymbol{W}\rangle\\ &\quad + \frac{\mu}{2}(\|\boldsymbol{U} - \boldsymbol{W}\|_F^2 + \|\boldsymbol{V} - \boldsymbol{W}\|_F^2).\end{aligned} \tag{12}$$

Problem (12) will be solved iteratively. In each iteration, $\boldsymbol{U}$, $\boldsymbol{V}$ and $\boldsymbol{W}$ will be optimized alternately (update one with the others fixed).

[Update $\boldsymbol{W}$]: In the $k+1$ iteration, the subproblem for optimizing $\boldsymbol{W}$ is described as follows:

$$\begin{aligned}&\min_{\boldsymbol{W}} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{W} - \boldsymbol{Y}\|_F^2 + \left\langle \boldsymbol{L}_1^k, \boldsymbol{U}^k - \boldsymbol{W}\right\rangle + \left\langle \boldsymbol{L}_2^k, \boldsymbol{V}^k - \boldsymbol{W}\right\rangle\\ &\quad + \frac{\mu^k}{2}(\|\boldsymbol{U}^k - \boldsymbol{W}\|_F^2 + \|\boldsymbol{V}^k - \boldsymbol{W}\|_F^2).\end{aligned} \tag{13}$$

Problem (13) can be solved by taking the gradient of its objective function w.r.t. $\boldsymbol{W}$ and setting it to zero. Then $\boldsymbol{W}$ can be obtained by

$$\boldsymbol{W}^{k+1} = (\boldsymbol{X}^T\boldsymbol{X} + 2\mu^k\boldsymbol{I})^{-1}(\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{L}_1^k + \boldsymbol{L}_2^k + \mu^k(\boldsymbol{U}^k + \boldsymbol{V}^k)) \tag{14}$$

[Update $\boldsymbol{U}$]: The subproblem for optimizing $\boldsymbol{U}$ is described as follows:

$$\min_{\boldsymbol{U}} \frac{\alpha}{2}\mathrm{Tr}(\boldsymbol{R}\boldsymbol{U}^T\boldsymbol{U}) + \left\langle \boldsymbol{L}_1^k, \boldsymbol{U} - \boldsymbol{W}^{k+1}\right\rangle + \frac{\mu^k}{2}\|\boldsymbol{U} - \boldsymbol{W}^{k+1}\|_F^2 \tag{15}$$

Taking the gradient of the objective function w.r.t. $\boldsymbol{U}$ and setting it to zero. Then $\boldsymbol{U}$ can be calculated by

$$\boldsymbol{U}^{k+1} = (\alpha\boldsymbol{R} + \mu^k\boldsymbol{I})^{-1}(\mu^k\boldsymbol{W}^{k+1} - \boldsymbol{L}_1^k) \tag{16}$$

[Update $\boldsymbol{V}$]: The subproblem for optimizing $\boldsymbol{V}$ is described as follows:

$$\min_{\boldsymbol{V}} \beta\|\boldsymbol{V}\|_1 + \left\langle \boldsymbol{L}_2^k, \boldsymbol{V} - \boldsymbol{W}^{k+1}\right\rangle + \frac{\mu^k}{2}\|\boldsymbol{V} - \boldsymbol{W}^{k+1}\|_F^2 \tag{17}$$

Then $\boldsymbol{V}$ can be calculated by

$$\boldsymbol{V}^{k+1} = S_{\frac{\beta}{\mu^k}}[\boldsymbol{W}^{k+1} - \frac{\boldsymbol{L}_2^k}{\mu^k}] \tag{18}$$

where S is the soft-thresholding operator.

The overall procedures of the optimization via ADMM are summarized in Algorithm 2. Then, the Lagrangian multipliers will also be updated in each iteration, which is shown in the procedure 4 in Algorithm 2. For faster convergence, can be adjusted using the updating strategy as shown in the procedure 5 in Algorithm 2. The termination condition is either when the differences between the objective variables of two adjacent iterations are all below the pre-set threshold or the maximum number of iteration is reached.

Algorithm 2 Algorithm to Solve the Problem (10)

Input:Data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, Label matrix $\boldsymbol{Y} \in \{0, 1\}^{n\times l}$, Disease correlation matrix $\boldsymbol{R} \in \mathbb{R}^{l\times l}$, Parameters α, β, ρ.
Output: Map matrix $\boldsymbol{W} \in \mathbb{R}^{p\times l}$.
Initialization: $\boldsymbol{W}^0 = \boldsymbol{U}^0 = \boldsymbol{V}^0 = \boldsymbol{L}_1^0 = \boldsymbol{L}_2^0 = \boldsymbol{0}$, $\mu^0 = 10^{-6}$, $\mu^{\max} = 10^6$.

1: **repeat**
2: update $\boldsymbol{W}$ by solving (14);
3: update $\boldsymbol{U}$ by solving (16);
4: update $\boldsymbol{V}$ by solving (18);
5: update Lagrangian multipliers $\boldsymbol{L}_1, \boldsymbol{L}_2$.
$\boldsymbol{L}_1^{k+1} = \boldsymbol{L}_1^k + \mu^k(\boldsymbol{U}^{k+1} - \boldsymbol{W}^{k+1})$
$\boldsymbol{L}_2^{k+1} = \boldsymbol{L}_2^k + \mu^k(\boldsymbol{V}^{k+1} - \boldsymbol{W}^{k+1})$
6: update μ
$\mu^{k+1} = \min(\mu^{\max}, \rho\mu^k)$, $\rho > 1$.
7: **until** termination condition:
$\max\{\|\boldsymbol{W}^{k+1} - \boldsymbol{W}^k\|, \|\boldsymbol{U}^{k+1} - \boldsymbol{U}^k\|, \|\boldsymbol{V}^{k+1} - \boldsymbol{V}^k\|\} < 10^{-6}$.

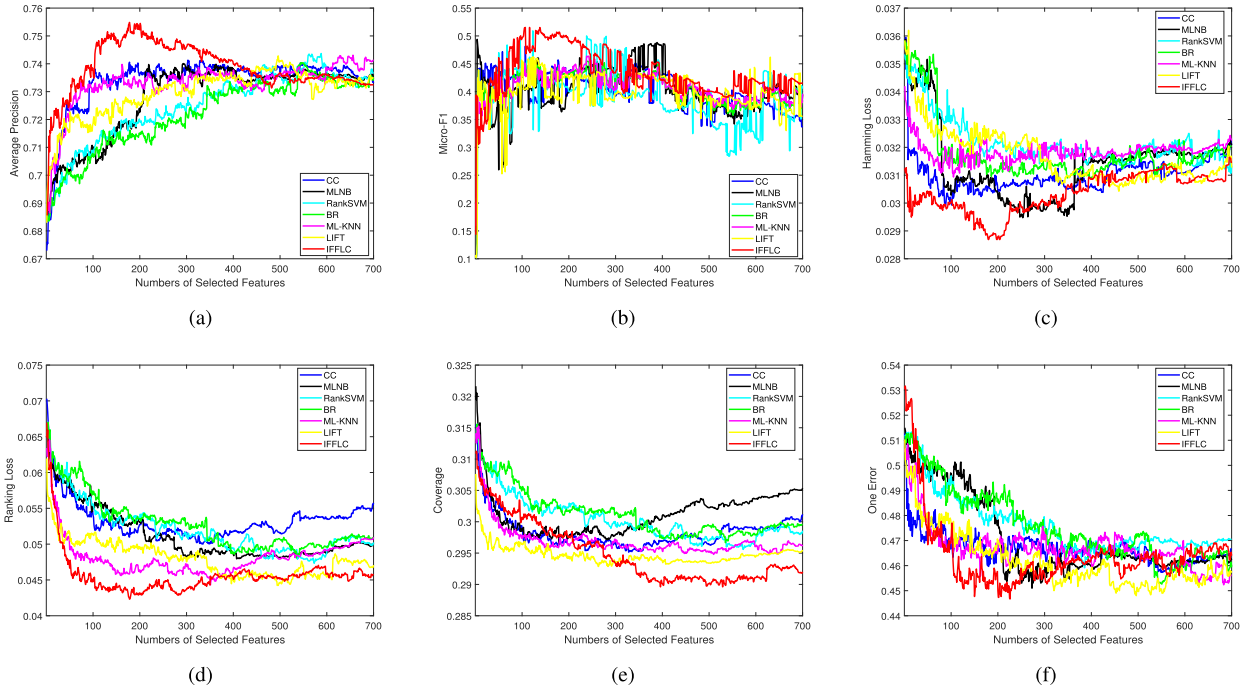


FIGURE 2. Performance variations for the classification algorithms. (a) Average precision. (b) Micro-F1. (c) Hamming loss. (d) Ranking loss. (e) Coverage. (f) One-error.

TABLE 2. Performance comparison between IFFLC and all comparative methods.

Method	Average Precision	Hamming Loss	Ranking Loss	One Error	Coverage	Micro-F1
BR	0.7146 ± 0.0156	0.0312 ± 0.0004	0.0535 ± 0.0036	0.4867 ± 0.0018	0.3018 ± 0.0025	0.4305 ± 0.0050
CC	0.7363 ± 0.0089	0.0304 ± 0.0013	0.0528 ± 0.0020	0.4674 ± 0.0045	0.2977 ± 0.0023	0.4380 ± 0.0068
RankSVM	0.7192 ± 0.0075	0.0321 ± 0.0018	0.0538 ± 0.0030	0.4798 ± 0.0038	0.3024 ± 0.0031	0.4003 ± 0.0060
MLNB	0.7206 ± 0.0092	0.0303 ± 0.0017	0.0531 ± 0.0018	0.4783 ± 0.0072	0.2976 ± 0.0020	0.3980 ± 0.0057
ML-KNN	0.7339 ± 0.0101	0.0318 ± 0.0020	0.0464 ± 0.0019	0.4667 ± 0.0044	0.2965 ± 0.0027	0.4338 ± 0.0061
LIFT	0.7252 ± 0.0084	0.0324 ± 0.0014	0.0496 ± 0.0022	0.4709± 0.0050	**0.2945 ± 0.0023**	0.4225 ± 0.0055
IFFLC	**0.7521± 0.0090**	**0.0288 ± 0.0018**	**0.0432 ± 0.0019**	**0.4651± 0.0024**	0.2975±0.0011	**0.4903± 0.0046**

IV. EXPERIMENTS

In this section, we conduct experiments on MIMIC-III database, so as to evaluate the effectiveness of our proposed algorithm IFFLC.

A. EXPERIMENT SETTINGS

To illustrate the performance of IFFLC, we compare it with the following state-of-the-art multi-label classification algorithms, including Binary Relevance (BR) [17], Classifier Chain (CC) [27], RankSVM [28], Multi-Label Naive Bayes classification (MLNB) [29], Multi-Label KNN (MLKNN) [30], and Label specific FeaTures (LIFT) [16].

- **BR** [17]: BR is a transformation approach, which divides the multi-label classification problem into many binary classification problems. In this experiment, SVM is used as base classifier.
- **CC** [27]: CC is composed of a chain of binary classifiers, where the prediction results of preceding binary classifiers act as additional features for constructing latter ones. In this experiment, SVM is used as base classifier.
- **RankSVM** [28]: RankSVM extends maximum margin strategy to deal with multi-label data, where a set of linear classifiers are optimized to minimize the empirical ranking loss and enabled to handle nonlinear cases with kernel tricks.
- **MLNB** [29]: MLNB improves the traditional naive Bayes classifiers to deal with multi-label data.
- **ML-KNN** [30]: ML-KNN adapts k-nearest neighbor techniques to deal with multi-label data, where maximum a posteriori (MAP) rule is utilized to make predictions by reasoning with the labeling information embodied in the neighbors.
- **LIFT** [16]: LIFT first performs clustering on features with respect to each class first. Afterwards, training and testing are conducted by querying the clustering results. In this method, label-specific features with respect to a certain class are exploited.

For the comparative methods, the parameter values of each algorithm are used as the default settings according to the corresponding literatures. For IFFLC, we tune parameters α, β in $\{2^{-10}, 2^{-9}, \cdots, 2^{10}\}$ and $\rho > 1$ as the default set-

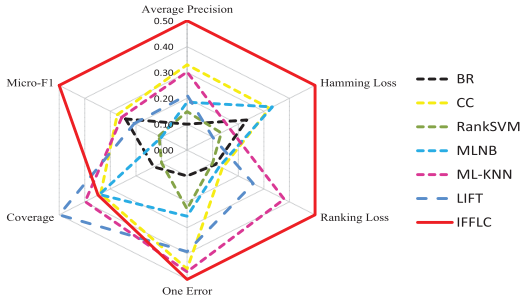


FIGURE 3. Spider Web diagram showing the stability index values with different evaluation metrics.

tings, and report the best results. Besides, six commonly used multi-label criteria [15] are utilized to make the comparison from different evaluation aspects, including Average Precision, Hamming Loss, Ranking Loss, One Error, Coverage and Micro-F1. Generally, few algorithms can outperform other algorithms on all these criteria. For Average Precision and Micro-F1, the larger the value is, the better the performance will be. For Hamming Loss, Ranking Loss, One Error and Coverage, the smaller the value is, the better the performance will be.

B. MULTI-LABEL CLASSIFICATION RESULTS

According to Algorithm 1, we can obtain the rank list of features directly. For the comparability of performances among all classification algorithms, the rank list that contains 2,240 features is fed to CC, MLNB, RankSVM, BR, MLKNN, LIFT and the proposed algorithm in this study as input. Figure 2 illustrates the change tendency of classification performance as the number of selected features increases. Since 700 selected features have been able to fully represent the trend of performance, we intercept 700 features from the rank list of features. In these subfigures, the number of the selected features is regarded as the horizontal axes, and the classification performance is regarded as the vertical axes. Moreover, seven different colored lines represent CC, MLNB, RankSVM, BR, MLKNN, LIFT and IFFLC, respectively. As shown in Figure 2, IFFLC can obtain superior classification performance with the growing number of features selected no matter how the variation tendency changes.

As shown in Figure 2, we find 200 is a trade-off number between effectiveness and efficiency for the rank list of features. A 10-fold cross-validation is used to evaluate the performance systematically. Table 2 reports the classification performance based on Average Precision, Hamming Loss, Ranking Loss, One Error, Coverage, and Micro-F1. Best results are highlighted in bold. The following observations can be easily drawn from Table 2: (1) For all evaluation indices except Coverage, IFFLC obtains superior performance against the comparative algorithms. (2) Note that the performance of IFFLC is extremely close to the best value with respect to best Coverage results obtained by LIFT method.

In addition, we draw Figure 3 to examine the stability of different multi-label classification algorithms on the six criteria. The rounder the spider web diagram is, the more stability each classification algorithm achieves. Specifically, the red line denotes the stability value of IFFLC. The result in Figure 3 demonstrates that IFFLC outperforms the other approaches and achieves stable performances on six evaluation criteria.

In order to systematically analyze the significant difference in 10-fold cross-validation, pairwise t-test is performed between IFFLC and each comparative method. The results are given in Table 3, in which the significance level (or p value) in most cells are 0.000 ($p < 0.05$), implying significant difference between the corresponding pair for all evaluation indices that except Coverage. Meanwhile, it is worth noting that the p-value in bold indicates that IFFLC and LIFT have no significant difference with respect to Coverage.

The success of IFFLC is due to feature selection based on FMI, modeling the sparsity of disease-specific features, and embedding with disease correlations. By feature selection, we remove redundant and irrelevant features to alleviate the influence of high dimensionality for MIMIC III data. It is easier to capture intrinsic relationships between features and diseases for multiple diagnosis codes assignment. By modeling the sparsity of disease-specific features, we can obtain the strong discriminability to the corresponding diseases in a lower dimensionality. Besides, the correlations between diseases are discovered to learn extra latent medical information and model a practical and stable relationship. The other comparative approaches only exploit some of the above aspects. For example, BR transforms this multi-label learning task to several independent binary single-label learning tasks. Obviously, BR does not consider disease correlations, and its performance is always the worst or second-worst.

C. THE SUPERIORITY OF FEATURE SELECTION

Most of the existing works [14], [19], [20] rank the frequencies of features occurrences and select the top p most frequently recorded features to form the feature space. In contrast, we utilize feature selection to form the input space in this paper. To extrude the superiority of feature selection, 200 features by feature selection and 200 features by frequency ranking are compared as the input feature space of our multi-label classification algorithm, respectively. These two methods are denoted as Frequency Ranking (FR) and Feature Selection (FS). As shown in Table 4, we can see that feature selection achieves decent performance and is significantly superior to the frequency ranking method on all evaluation criteria. This is because the features obtained by frequency ranking are for all disease codes. In contrast, the features obtained through feature selection are associated with specific diseases, and these features are more relevant. In other words, there may be some features that are high in frequency but not useful for multiple diagnosis codes assignment.

TABLE 3. Significance between IFFLC and each comparative method.

Method	BR	CC	RankSVM	MLNB	ML-KNN	LIFT
Average Precision	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**
Hamming Loss	0.000**	0.000**	0.000**	0.000**	0.000**	0.003**
Ranking Loss	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**
One Error	0.000**	0.000**	0.000**	0.000**	0.001**	0.000**
Coverage	0.000**	0.036*	0.000**	0.000**	0.019*	0.174
Micro-F1	0.000**	0.000**	0.000**	0.000**	0.000**	0.005**

*Significant at the 0.05 level (two-tailed).
**Highly Significant at the 0.01 level (two-tailed).

TABLE 4. Performance comparison of feature selection scheme on IFFLC.

Method	FS Scheme	FR Scheme
Average Precision	**0.7521±0.0090**	0.7282 0.0157
Hamming Loss	**0.0288 0.0018**	0.0306±0.0022
Ranking Loss	**0.0432±0.0019**	0.0521 ±0.0024
One Error	**0.4651 ±0.0024**	0.4767 ±0.0030
Coverage	**0.2975 ±0.0011**	0.3092 ±0.0023
Micro-F1	**0.4903 ±0.0046**	0.3887 ±0.0059

V. CONCLUSION

In this paper, we proposed an integrated framework that performed feature selection, disease-specific features learning for each disease and multiple diagnosis codes assignment. This integrated framework intends to intelligently imitate the diagnosis process of doctors to assign multiple diagnosis codes to patient records automatically and effectively. In IFFLC, the most discriminative features are learned and the influence of disease correlations is considered. Extensive experiments validated that IFFLC was superior over the other state-of-the-art algorithms. In the future, we will consider modeling the more complex levels of disease correlations. In this paper, we exploited pairwise relations between diseases, i.e., second-order disease correlations. The proposed approach can achieve good generalization performance. However, disease correlations go beyond the second-order assumption in the real-world medical diagnosis process. Therefore, in our future work, we can construct a multi-label classifier for diagnosis codes assignment by considering high-order relations among labels such as imposing all other diseases' influences on each disease, or addressing connections among random subsets of diseases, etc. Theoretically, high-order relations have stronger correlation-modeling capabilities than second-order relations.

ACKNOWLEDGMENT

The authors will thank Professor Xue Li from University of Queensland and Professor Minling Zhang from Southeast University for their great help and guidance.

REFERENCES

[1] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 5, pp. 1589–1604, Sep. 2018.

[2] C. Liu, F. Wang, J. Hu, and H. Xiong, "Temporal phenotyping from longitudinal electronic health records: A graph based framework," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 705–714.

[3] K. Bhatia, H. Jain, P. Kar, M. Varma, and P. Jain, "Sparse local embeddings for extreme multi-label classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 730–738.

[4] M. F. Ghalwash, V. Radosavljevic, and Z. Obradovic, "Extraction of interpretable multivariate patterns for early diagnostics," in *Proc. 13th Int. Conf. Data Mining*, Dec. 2013, pp. 201–210.

[5] H. Harutyunyan, H. Khachatrian, D. Kale, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 1–11. [Online]. Available: https://arxiv.org/abs/1703.07771

[6] M. S. Mohktar *et al.*, "Predicting the risk of exacerbation in patients with chronic obstructive pulmonary disease using home telehealth measurement data," *Artif. Intell. Med.*, vol. 63, no. 1, pp. 51–59, 2015.

[7] M. Peng *et al.*, "Exploration of association rule mining for coding consistency and completeness assessment in inpatient administrative health data," *J. Biomed. Informat.*, vol. 79, pp. 41–47, Mar. 2018.

[8] L. Chen *et al.*, "Mining health examination records—A graph-based approach," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 9, pp. 2423–2437, Sep. 2016.

[9] S. Fodeh and Q. Zeng, "Mining big data in biomedicine and health care," *J. Biomed. Informat.*, vol. 63, pp. 400–403, Oct. 2016.

[10] A. Perotte, R. Pivovarov, K. Natarajan, N. Weiskopf, F. Wood, and N. Elhadad, "Diagnosis code assignment: Models and evaluation metrics," *J. Amer. Med. Informat. Assoc.*, vol. 21, no. 2, pp. 231–237, 2014.

[11] H. Xiong, J. Zhang, Y. Huang, K. Leach, and L. Barnes, "Daehr: A discriminant analysis framework for electronic health record data and an application to early detection of mental health disorders," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 3, p. 47, 2017.

[12] T. Baumel, J. Nassour-Kassis, R. Cohen, M. Elhadad, and N. Elhadad, "Multi-label classication of patient notes: Case study on ICD code assignment," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1–8. [Online]. Available: https://arxiv.org/abs/1709.09587

[13] Y. Yan, G. Fung, J. Dy, and R. Rosales, "Medical coding classification by leveraging inter-code relationships," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 193–202.

[14] D. Zufferey, T. Hofer, J. Hennebert, M. Schumacher, R. Ingold, and S. Bromuri, "Performance comparison of multi-label learning algorithms on clinical data for chronic diseases," *Comput. Biol. Med.*, vol. 65, no. 10, pp. 34–43, 2015.

[15] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.

[16] M.-L. Zhang and L. Wu, "LIFT: Multi-label learning with label-specific features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 1, pp. 107–120, Jan. 2015.

[17] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognit.*, vol. 37, no. 9, pp. 1757–1771, Sep. 2004.

[18] J. C. Ferrão, F. Janela, M. D. Oliveira, and H. M. G. Martins, "Using structured EHR data and SVM to support ICD-9-CM coding," in *Proc. IEEE Int. Conf. Healthcare Informat.*, Sep. 2013, pp. 511–516.

[19] S. Wang, X. Chang, X. Li, G. Long, L. Yao, and Q. Z. Sheng, "Diagnosis code assignment using sparsity-based disease correlation embedding," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 12, pp. 3191–3202, Dec. 2016.

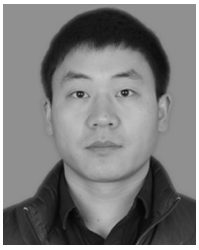
[20] S. Wang, X. Li, L. Yao, Q. Z. Sheng, and G. Long, "Learning multiple diagnosis codes for ICU patients with local disease correlation mining," *ACM Trans. Knowl. Discovery Data*, vol. 11, no. 3, pp. 1–31, 2017.

[21] Y. Zhu, J. T. Kwok, and Z.-H. Zhou, "Multi-label learning with global and local label correlation," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 6, pp. 1081–1094, Jun. 2018.
[22] A. E. W. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, May 2016, Art. no. 160035.
[23] Y. Lin, Q. Hu, J. Liu, J. Li, and X. Wu, "Streaming feature selection for multilabel learning based on fuzzy mutual information," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 6, pp. 1491–1507, Dec. 2017.
[24] J. Huang, G. Li, Q. Huang, and X. Wu, "Learning label specific features for multi-label classification," in *Proc. IEEE Int. Conf. Data Mining*, Nov. 2015, pp. 181–190.
[25] J. Huang, G. Li, Q. Huang, and X. Wu, "Joint feature selection and classification for multilabel learning," *IEEE Trans. Cybern.*, vol. 48, no. 3, pp. 876–889, Mar. 2018.
[26] C. Lu, J. Feng, S. Yan, and Z. Lin, "A unified alternating direction method of multipliers by majorization minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 527–541, Mar. 2018.
[27] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *J. Mach. Learn.*, vol. 85, no. 3, pp. 333–359, Dec. 2011.
[28] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 681–687.
[29] M.-L. Zhang, J. M. Peña, and V. Robles, "Feature selection for multi-label naive Bayes classification," *Inf. Sci.*, vol. 179, no. 19, pp. 3218–3229, 2009.
[30] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, Jul. 2007.

YUWEN LI is currently pursuing the Ph.D. degree with the Department of Automation, Xiamen University. She is also a joint training student with the School of Information Technology and Electrical Engineering, The University of Queensland. Her research interests include data mining and granular computing.

WEITONG CHEN received the B.S. degree from Griffith University, Australia, in 2011, and the M.S. degree from The University of Queensland, Australia, in 2013. He is currently pursuing the Ph.D. degree with The University of Queensland, Australia. He has published nearly 20 peer-reviewed papers in prestigious journals and top international conferences, including the IEEE ICDE, AAAI, IJCAI, WWW, WWWJ, ECML, CIKM, and SIAM SDM. His current main research interests include medical data analytic, deep learning, data mining, pattern recognition, and social computing. He has been actively engaged in professional services by serving as a Conference Organizer, a Conference PC Member, and a Reviewer for journals, such as ADMA, WWW, MobiQuitous, JCST, KAIS, and MobiSPC.

DEYIN LIU received the bachelor's degree from Zhengzhou University, Zhengzhou, Henan, China, where he is currently pursuing the Ph.D. degree. He is also pursuing the Joint Ph.D. degree with The University of Queensland. His major research interests include computer vision, pattern recognition, and machine learning.



ZHIMIN ZHANG received the B.E. degree in biomedical engineering from Shandong University, Jinan, China, in 2014. He is currently pursuing the Ph.D. degree with Shandong University. He is also a Research Student with the School of Information Technology and Electrical Engineering, The University of Queensland, where he is mainly working on ECG and EEG signals for cardiovascular disease and sleep stages' classification using compressed sensing, machine learning, and entropy analysis techniques.

SHUNXIANG WU was born in Shaoyang, Hunan, China, in 1967. He received the M.S. degree from the Department of Computer Science and Engineering, Xi'an Jiaotong University, in 1991, and the Ph.D. degree from the School of Economics and Management, Nanjing University of Aeronautics and Astronautics, in 2007. He is currently a Professor with the Department of Automation, School of Aerospace Engineering, Xiamen University. His research interests include intelligent computing, data mining and knowledge discovery, and systems' engineering theory and application.

CHENGYU LIU (M'14) received the B.S. and Ph.D. degrees in biomedical engineering from Shandong University, China, in 2005 and 2010, respectively. He completed the Postdoctoral training at Shandong University, China, from 2010 to 2013, Newcastle University, U.K., from 2013 to 2014, and Emory University, USA, from 2015 to 2017.

He is currently the Director and a Professor with the Southeast-Lenovo Wearable Heart-Sleep-Emotion Intelligent Monitoring Laboratory, School of Instrument Science and Engineering, Southeast University, Nanjing, China. He was a PI on over ten awarded grants attracting a total of over 1 million. He has published more than 130 journal/conference papers and eight chapters in books. He holds 15 invention patents. His research interests include mHealth and intelligent monitoring, machine learning and big data processing for physiological signals, early detection of CADs, device development for CADs, and sleep and emotion monitoring. He is currently a Federation Journal Committee Member of the International Federation for Medical and Biological Engineering.