A Clinical Decision Support Framework for Heterogeneous Data Sources

Mengxing Huang, *Member, IEEE*, Huirui Han, Hao Wang, *Member, IEEE*, Lefei Li, *Member, IEEE*,
Yu Zhang, *Member, IEEE*, Uzair Aslam Bhatti

***Abstract*—To keep pace with the developments in medical informatics, health medical data is being collected continually. But, owing to the diversity of its categories and sources, medical data has become highly complicated in many hospitals that it now needs *Clinical Decision Support* (CDS) system for its management. To effectively utilize the accumulating health data, we propose a CDS framework that can integrate heterogeneous health data from different sources, such as laboratory test results, basic information of patients, and health records into a consolidated representation of features of all patients. Using the electronic health medical data so created, multi-label classification was employed to recommend a list of diseases and thus assist physicians in diagnosing or treating their patients' health issues more efficiently. Once the physician diagnoses the disease of a patient, the next step is to consider the likely complications of that disease, which can lead to more diseases. Previous studies reveal that correlations do exist among some diseases. Considering these correlations, a k-nearest neighbors algorithm is improved for multi-label learning by using correlations among labels (CML-*k*NN). The CML-*k*NN algorithm first exploits the dependency between every two labels to update the origin label matrix and then performs multi-label learning to estimate the probabilities of labels by using the integrated features. Finally, it recommends the top N diseases to the physicians. Experimental results on real health medical data establish the effectiveness and practicability of the proposed CDS framework.**

***Index Terms*—Diagnosis recommender systems, clinical decision support system, heterogeneous data sources, multi-label classification**

I. Introduction

IN the era of fourth revolution of industry (Industry 4.0), the associated services (smart services) develop rapidly [1]. In this context, Health 4.0 has been growing as a vital strategic concept for health domain, which is aimed at providing real-time and personalized health services (called smart healthcare services) for patients and professionals [2]. Health data from numerous medical sources (Cyber-Physical Systems, the Internet of Things) is collected continuously, facilitating the growth of healthcare industry [3].

It is widely accepted that health information tools and machine learning techniques can be exploited successfully to help doctors in diagnosing and treating their patients more efficiently [4]. Using their experience and knowledge, the physicians classify patients and diagnose their diseases, but in doing so, it is probable that they commit some mistakes, particularly when they lack adequate experience or when their faculty of judgment is poor. In such situations, *Clinical Decision Support* (CDS) systems, including systems that provide diagnosis, personalized medical measurement, treatment and relevant knowledge, would be helpful to the physicians by way of providing them with specific knowledge, patients' information and intelligent applications, which can improve the efficiency of their decision-making processes [5]. CDS systems focus on extracting characteristics of patients, based on which they classify patients and provide corresponding clinical suggestions to the physicians. Patients' medical information is extracted from their personal medical data, such as the physiological data, electronic health records (EHRs), 3D images, radiology images, genomic sequencing, and clinical and billing data. Through CDS applications, the physicians can avoid the mistakes that are likely to arise from medical negligence and thus improve the quality of their medical service. In the medical field, the demand for high-quality clinical support systems has been steadily on the increase [6]. In medical scenes, the specifies of scoring standards and the context complexity of medical field are the challenges of clinical decision support system.

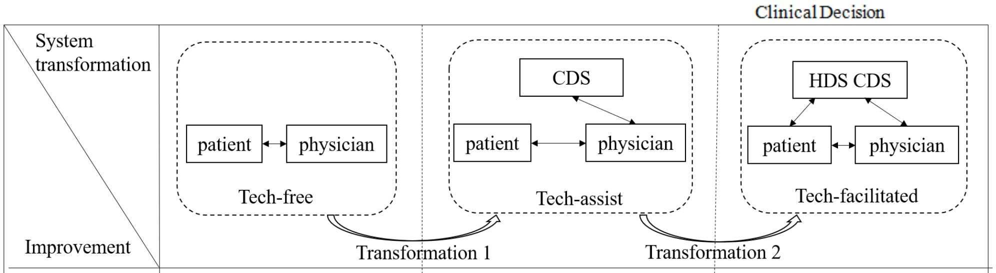
Medical institutions keep accumulating health medical data, which is highly complex in most of the recognized research labs and hospitals. Government agencies have been working hard to utilize such complex and diverse types of medical data to diagnose patients' diseases correctly and offer them the right treatment [7]. To make this happen, the physicians have to consider multiple types of health information of patients, like laboratory test results, basic attributes, health records and monitoring data. Medical data comes from different sources, and most of it is unstructured. Integrating complex medical

This work was supported by the Chinese National Natural Science Foundation of China under Grant 61462022, the Chinese National Key Technology Support Program under Grant 2015BAH55F04 and Grant 2015BAH55F01, Major Science and Technology Project of Hainan province under Grant ZDKJ2016015. (Corresponding authors: Hao Wang; Lefei Li; Yu Zhang.

Mengxing Huang, Huirui Han, Yu Zhang and Uzair Aslam Bhatti are with State Key Laboratory of Marine Resource Utilization in South China Sea, College of Information Science & Technology, Hainan University, Haikou, China (email: huangmx09@163.com; hanhr26@163.com; yuzhang_nwpu@163.com; uzairaslambhatti@hotmail.com)

Hao Wang is with Big Data Lab, Dept. of ICT and Natural Sciences, Norwegian University of Science and Technology, Aalesund, 6009, Norway (email: hawa@ntnu.no)

Lefei Li is with the Department of Industrial Engineering, Tsinghua University, Beijing, 100084, China (email: lilefei@tsinghua.edu.cn)



Improvement	Tech-free	Tech-assist	Tech-facilitated
Efficiency	• Repeated inquiry process • Long waiting time	• More convenient for physician to obtain clinical information assisted by CDS	• More convenient for patient and physician to obtain clinical information assisted by HDS CDS
Effectiveness	• difficult to diagnose complicated illnesses for inexperienced physician	• Integrate more historical medical records • narrow range of diagnosis • reduce faults caused by work fatigue	• Provide more comprehensive analysis • Improve diagnosis precision based on multiple information and multi-label datasets
Wholeness	• No decision support technology	• No connection between CDS system and patient	• Enhance interactivity and circulation of information

Fig. 1. The Improvement of Clinical Decision Support System

data and transforming it into appropriate format are challenging tasks for the CDS systems.

At present, although many CDS systems have been developed to assist the doctors, they are still unsatisfactory, because they consider only a single medical condition [8-9]. But, it is not uncommon that a patient has more than one medical condition, at the same time, because of the complications of the first disease. For instance, a patient with diabetes mellitus type 2 may develop some cardiovascular diseases [10] and a patient with hypertensive heart disease may get coronary disease or/and angina pectoris [11]. A clinical decision support system might be more complicated to discover more possible diseases as references to clinics. After analyzing real-world diagnostic information, it is found that the majority of the patients have more than one disease. Therefore, the clinical decision support system recommends a list of possible diseases rather than only a single disease. Consequently, the task of recommendation transforms itself into a multi-label classification problem [12]. In dealing with this problem, ML-*k*NN algorithm [13] has gained wide acceptance because of its simple procedure and effectiveness. But, this method estimates each label independently and ignores the correlation among labels. Considering the correlations among diseases, we applied a novel multi-label classification approach in the clinical decision support framework.

What one gets to see today is the emergence of Health 4.0. It unfolds the coming-together of all these technologies, coupled with real-time data collection, increased use of Artificial Intelligence (AI) and an overlay of invisible user interfaces. By focusing on collaboration, coherence, and convergence the healthcare can be made more accurately predictive and personalized. The main focus of this work is to build a clinical decision support framework for heterogeneous data sources (HDS CDS) for assisting doctors in diagnosing the diseases of their patients and treating them more efficiently. Fig. 1 illustrates the proposed improvement in HDS CDS system. In a tech-free system, the patients need to wait for a long time before they get to consult the doctor; besides, their repeated inquiries reduce diagnosis efficiency of their doctors. Also, inexperienced doctors may find it difficult to diagnose complicated illnesses. Traditional CDS systems (clinical decision support systems) improve efficiency and effectiveness of diagnosis by giving decision support to the physicians. They integrate historical medical records, which would be helpful in identifying potential diseases and thus in reducing fault risks. HDS CDS systems further improve the efficiency and effectiveness of CDS systems, especially by enhancing the wholeness of the system. HDS CDS systems collect and analyze healthcare data from diverse sources, rather than a single source. They suggest several correlative diseases by formulating a multi-label estimation model. Thus, HDS CDS system improves its performance in terms of comprehensiveness and accurate diagnosis. In the ongoing HDS CDS system, the patients initially offer information and, in return, they get information from both the system and the physician. As a result, the interaction among patients, physicians and system will be enhanced and information dissemination will improve. The proposed clinical decision support framework for heterogeneous data sources is depicted in Fig. 2.

The main contributions of this paper are as follow:

1) A novel framework is proposed for retrieving the most relevant information of patients from multiple data sources, such as laboratory test data, basic information of patients, symptoms of patients and electrocardiogram data, and for combining them to generate integrated features.
2) Considering the likely complications due to multiple medical diseases (conditions), k-nearest neighbors algorithm is proposed for multi-label learning, by using correlations among labels (CML-*k*NN) and for anticipating more potential diseases of a patient, so that a list of diseases can be recommended to the physician simultaneously.
3) Using the laboratory test data and basic information of patients, a set of experiments of different multi-label learning methods were performed to confirm the effectiveness and practicality of the proposed framework.

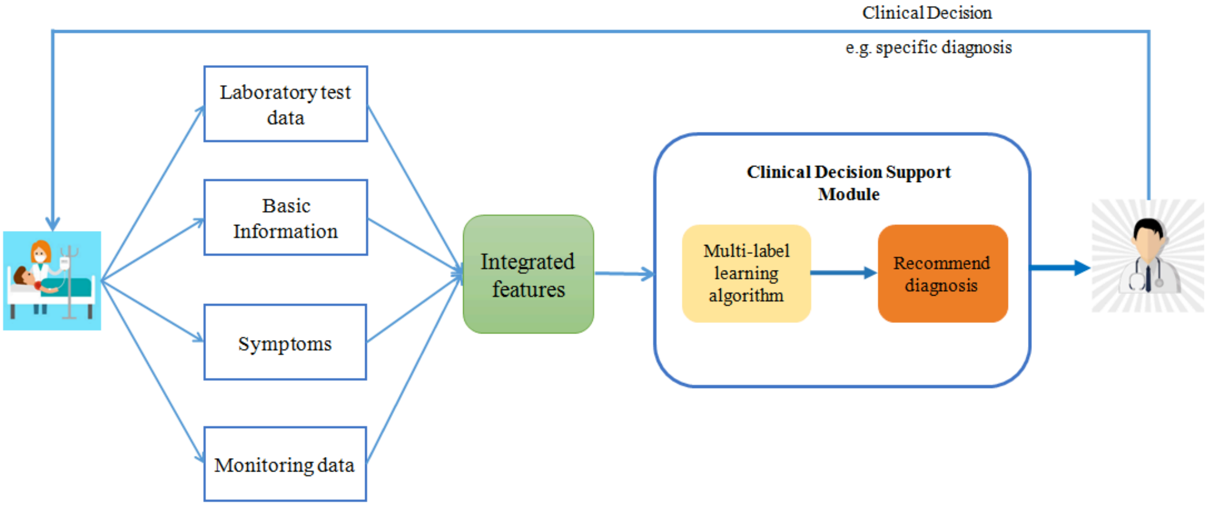


Fig. 2. The Description of Clinical Decision Support Framework for Heterogeneous Data Sources

The remainder of the paper is organized as follows. Section II presents some related works about clinical decision support systems and multi-label learning algorithms. We describe methods in section III, and the analysis of the correlations in the health data is given in section IV. In Section V, the paper introduces the model validation. Finally, section VI concludes the paper and gives an outlook on future works.

II. Related Works

Big data has five attributes: Volume, Velocity, Variety, Value and Veracity [14]. The number of visiting patients in a general hospital is generally large. For example, Haikou peoples' hospital had received 46,000 inpatients and 95,000 outpatients approximately last year. The health medical data of such a large number of patients would obviously be of peta or zeta bytes, and this refers to the Volume of Big data. The patient's information will be updated as soon as he/she visits the hospital again, and this represents the Velocity of health medical data. The health medical data consists of structured, semi-structured and unstructured data. Moreover, the health medical data is of different categories: electronic health medical records written by physicians; data from real-time monitors; images collected by computed tomography (CT); nuclear magnetic resonance images (MRI); cardiac angiographs etc. Each patient's medical records from professional physicians and medical instruments reflect his/her real physical condition, which represent the Veracity of health medical data. Furthermore, the amenability of the collected health data for transformation into useful and meaningful knowledge, which represents the Value of the data. Therefore, health medical data is a kind of "Big data" to some extent.

Nowadays, the data-intensive applications require a large number of efficient models. Numerous stochastic methods [15] were exploited by different researchers for healthcare parameter analysis. Furthermore, physicians consider the similarity between the health parameters of a patient for accurate diagnosis decision [16]. Analysis of big data is applied in healthcare to identify clusters of patients and groups of diseases, which are used to estimate future health condition with the help of different machine learning techniques [17]. Utilization and analysis of health medical data play a vital role in healthcare system.

The clinical decision support (CDS) system is an information system that offers knowledge and personalized information to users in enhancing health and healthcare outcomes [18]. The system is aimed at aiding physicians in diagnosis and treatment planning. CDS system can be used for routine requirements or applied to a specific situation, like implant placement, and the output can be submitted to users before, during or after clinical decision [19]. For designing and implementing CDS system, the following five concepts must be followed:

1) Correct information (treatment planning and drug interaction)
2) Appropriate user (clinicians, patients)
3) Through the applicable channels (mobile devices, working stations)
4) Applying the right intervention format (alarm, graphics, buttons)
5) At the right time within clinical working process (before working on the drug prescription, at the point of nursing)

The CDS system can utilize appropriate computing technology to improve the efficiency of decision-making [20]. The big data of health has been recognized as a great opportunity for improving CDS systems [21]. Many machine learning techniques, such as ensemble learning [22], SVMs [23], deep neural networks [24] and rule-based algorithm [25] were employed in developing clinical decision support systems for some specific diseases. Although these systems achieved high accuracy, the effectiveness of improvement is unsatisfactory, because they considered only a small number of selected features [26]. These methods may prove ineffective if they are to deal with a large number of different kinds of features. The existing clinical decision support systems are poor in processing large volumes of multi-structured healthcare data and in providing accurate health recommendation, in practice [27-28].

Traditional classification methods belong mainly to binary-class or multi-class, in which each sample belongs to a single class. For analyzing some samples with multiple labels, those methods are transformed to multi-label learning methods, whose task is to estimate possible labels of target samples. It is possible for a sample from multi-label data to own multiple labels, while the sample from single-label data owns only one distinct label. Thus, multi-label data is more complex and varied. Multi-label learning methods can be divided into two groups: problem transformation method (PTM) and algorithm adaption method (AAM) [29].

Problem transformation method first transforms one multi-label dataset into multiple single-label datasets, and then exploits existing single-label learning algorithm to process each single-label dataset. Problem transformation method makes multi-label data to adapt algorithms. Traditional problem transformation approaches are BR (binary relevance) method [30], LP (label powerset) method etc. However, algorithm adaption method relies on a certain machine learning algorithm, such as decision tree [31], support vector machine [32], BP neural network [33] etc., and enables it to tackle multi-label data directly.

Compared to other algorithms, *k*NN algorithm, which does not require training model or optimizing parameters in advance, has the advantages of low complexity and better performance in classification. BR*k*NN, proposed by Spyromitros et al. [34], improves the running efficiency of classification algorithm, but it retains the disadvantage of BR, namely ignoring the correlations among labels. Zhang et al. [13] proposed a multi-label lazy learning algorithm by applying *k*NN algorithm and Bayesian theory in multi-label learning. ML-*k*NN received wide attention immediately, because it is simple and effective. However, ML-*k*NN algorithm estimates each label independently, but it ignores the correlations among labels. Based on ML-*k*NN algorithm, the authors propose here a k-nearest neighbors algorithm for multi-label learning by using correlations among labels (CML-*k*NN) to recommend diagnosis to physicians.

A Shared Decision-Making System for Diabetes Medication Choice is adopted as a decision aid for clinical purpose by using Random Forest to predict a list of appropriate medications simultaneously [12]. Xu et al. developed an effective Chinese Medicine diagnostic model for coronary heart disease by using a multi-label learning algorithm, which is based on mutual information feature selection [35]. However, the correlations among diseases are not considered by these systems. Hence, the authors exploit the proposed CML-*k*NN in the HDS CDS framework to recommend a list of diseases of a patient to his/her physician simultaneously.

III. Methods

A. Overview of the Framework

The authors propose a novel framework–A Clinical Decision Support Framework for Heterogeneous Data Sources (HDS CDS). The HDS CDS framework is designed to process and analyze huge volumes of various types of healthcare data in a medical context. Patients' features, extracted from different sources, such as laboratory test data, health medical records and monitoring data, are exploited by a novel multi-label learning method for recommending possible diseases to physicians.

The HDS CDS framework is divided into two stages. The first stage is to integrate different categories of health medical data from different sources. The hospital stores medical data of each patient daily, in different databases, in terms of the datatype. For example, the laboratory test data, in terms of some blood parameters and symptoms of patients, has been found to be relevant in diagnosing some diseases [36]. Thus, different features may be relevant to different diagnoses. The proposed framework considers four categories of patient's information, including laboratory test data, basic information of patients, symptoms of patients and electrocardiogram data, to generate the integrated features. The first stage consists of the following modules.

1) The module of analyzing laboratory test data: Comparing the results of testing items with the reference ranges of those items to identify the abnormal items and quantify the levels of their abnormality.
2) The module of analyzing basic information of patients: Analyzing the texts to extract information, in terms of gender, age, temperature, height and weight of patients, from the health medical records to build the basic attributes of patients.
3) The module of analyzing symptoms: Extracting and building a set of specific symptoms and quantifying their degrees by natural language processing techniques.
4) The module of analyzing monitoring data: Identifying the exceptions from monitoring data and classifying them.

The second stage is to employ the proposed multi-label learning method to generate the list of diseases to be recommended.

5) The module of reconstructing label matrix: Using cosine similarity in estimating the relevance between every two labels to construct label-to-label similarity matrix, and then to reconstruct the label matrix by label-to-label similarity matrix.
6) The module of diagnoses to be recommended: Exploiting ML-*k*NN [13] to identify possible diseases of target patients and recommend the same to physicians.

B. Available Features of Patients

1) Features from Laboratory Test Data: The results of testing items are shown in each patient's laboratory test report and stored in the laboratory table of the database. Once the results come out, each item's result is compared with the corresponding biological reference interval. If the lab result is out of the biological reference range, the testing item is regarded as abnormal; otherwise, it is regarded as normal. All testing items are considered as features of laboratory test data and the features are denoted by a vector. Therefore, the lab test vector of a patient is defined as $L=\{l_1,l_2,\ldots,l_n\}$, where l indicates the item of lab test reports, and n indicates the number of items in the lab test reports.

2) Features from Basic Information of Patients: The basic information of patients (gender, age etc.), listed in the textual medical health record, is extracted and categorized as basic attributes of patients by text processing methods. Hence, the vector of these attributes is defined as $A = \{a_1, a_2, ..., a_n\}$, where n is the number of attributes of the patient.

3) Features from Symptoms of Patients: The symptoms are recorded, while describing the illness, in the form of a few sentences in medical health records. The Natural Language processing can carry out the task of extracting features of each patient from medical health record [37]. At first, each sentence of illness description is divided into several words by Chinese word segmentation technique. After extracting useful words, LDA (Latent Dirichlet Allocation) model [38] is performed to obtain patient-based topic distribution $\theta = \{\theta^1, \theta^2, \mathrm{K}\ , \theta^K\}$, where K is the number of topics in patients.

4) Features from ECG of Patients: The hospital has many devices for monitoring some physical signals of patients, such as cardiac monitoring system and glucose monitors. Electrocardiogram, one form of monitoring data, can be easily collected by the electrocardiograph (ECG). Signal processing, high performance computing and data mining techniques, used in the analysis of electrocardiogram, are helpful to doctors in improving the quality of their diagnosis. Some features of electrocardiogram can be identified only by referring techniques and not by naked eye. Many researchers have worked to identify different heartbeat classes from electrocardiograms [39-41]. This study is focused on five most common heartbeat cases in MIT-BIH arrhythmia database [42], including (i) normal heartbeat (NORM); (ii) left bundle branch block (LBBB); (iii) right bundle branch block (RBBB); (iv) ventricular premature contractions (VPC); (v) atrial premature complexes (APC). Based on the heartbeat class of the patient, the ECG vector of each patient is created as $E = \{e_1, e_2, \mathrm{K}\ , e_n\}$, where e is a heartbeat class and n is the number of heartbeat classes.

5) Integrating Features of Patients: Features from different health data sources are integrated to generate the final features of patients. With increasing number of new patients, the number of final features may increase rapidly. To maintain the efficiency of clinical decision support system, some dimension-reduction approaches, like PCA (Principal Component Analysis) [43] and LDA (Linear Discriminant Analysis) [44], are used for reducing the number of features.

C. Model of Disease Recommendation

1) Problem Transformation: For a patient having one or more diseases at the same time, the disease recommendation will be made by using multi-label learning method. The input of the model will be the integrated features of the target patient and the output will be one or more possible diseases of the target patient. In this study, “labels” are denoted as the diseases to be recommended, which are the results recommended by the model. Furthermore, the patients are denoted as “samples” in the recommendation model.

$F = \{f_1, f_2, \mathrm{K}\ , f_b\}$ is the space of feature with b dimensions, and $L = \{l_1, l_2, \mathrm{K}\ , l_q\}$ is the space of label with q dimensions. Given a multi-label data $T = \{(X_1, Y_1), (X_2, Y_2), \mathrm{K}\ , (X_n, Y_n)\}$, where $X_i = (x_i^1, x_i^2, \mathrm{K}\ , x_i^b)$ denotes a feature vector of the sample X_i and x_i^j is the value of X_i in feature f_j , and $Y_i = (y_i^1, y_i^2, ..., y_i^b)$ denotes the label vector of the sample X_i , $y_i^j = 1$ when X_i has label l_j ; otherwise, $y_i^j = 0$. The task of multi-label learning is to learn a classification model to estimate the possible label vector for testing sample X, which has no known label.

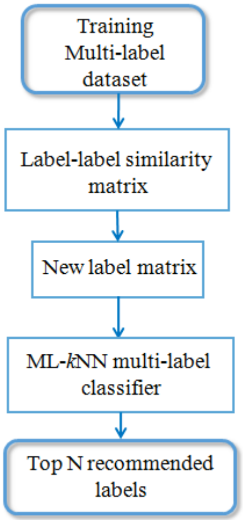

Training Multi-label dataset
Label-label similarity matrix
New label matrix
ML-kNN multi-label classifier
Top N recommended labels

Fig. 3. The Flowchart of CML-*k*NN

2) Multi-label Disease Recommendation Model: Owing to the correlation and dependency among some common labels, the relevance between every two labels should be considered in multi-label learning method. For example, if a patient is ill, it is very likely that he or she may, sooner or later, develop complications of this illness, because the current disease(s) may give rise to its complications. Therefore, the co-currency and dependency between every two labels are exploited to update the origin label matrix and obtain a potential and abundant label matrix. Motivated by the idea of dependency propagation of labels [45], a k-nearest neighbors algorithm for multi-label learning is applied to the system by using correlations among labels (CML-*k*NN). Fig. 3 describes the flowchart of the CML-*k*NN.

First, the label-to-label similarity matrix is generated. Every two labels’ frequencies of co-occurrence in the same patient can be used to evaluate the similarity between two such labels. Cosine similarity is employed as follows to evaluate the similarity between every two labels:

$$sim(I_i, I_j) = \frac{\sum_{k \in P_{ij}} r_{ik} r_{jk}}{\sqrt{\sum_{k \in P_i} r_{ik} \sum_{k \in P_j} r_{jk}}} \qquad (1)$$

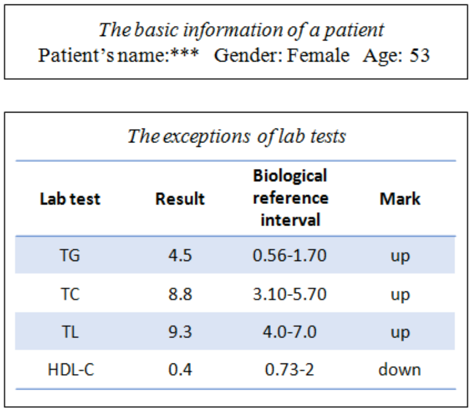
where P_{ij} is the set of samples with label I_i and I_j . r_{ik} is 1, when sample k got label i; otherwise, r_{ik} is 0. $sim(I_i, I_j)$ falls into [0,1]; when it is closer to 1, it implies that label I_i and label I_j are more related.

Based on the similarity between the labels, label-to-label similarity matrix $S \in °\ ^{q \times q}$ is generated as follows:

```
gorithm 1 A k-nearest neighbors algorithm for multi-label
rning by using correlations among labels
ut: Testing sample X, the number of neighbors k,
aining set $T=\{(X_1,Y_1),(X_2,Y_2),K\ ,(X_n,Y_n)\}$, the space of
bel L, the number of recommended label N
tput: The suggested label vector of X
p 1 : Generate label-label similarity matrix
R i= 1 to q:
    FOR j = 1 to q:
is computed by formula (1)
p 2 : Update each label vector by formula (3)
p 3 : Obtain N(X),which is the set of k neighbors of X
R $l_i \in L$
 Step 4 : Based on traditional ML-kNN [34],
 collect $C_x(l_i)$, which is the number of sample with label
 $i$ in the N(X), and achieve the probability that X has
 label $l_i$:
 $p_i=\frac{p(H_1^i)p(E^i_{C_x(l_i)}|H_1^i)}{\sum_{a=0}^{1}p(H_a^j)p(E^i_{C_x(l_i)}|H_a^i)}$
D FOR
p 5 : Combine probabilities of all labels:
$P=(p_1,p_2,K\ ,p_q)$
p 6 : Recommend N labels with probabilities ranked top
n $P$ to physicians.
```

$$S=\begin{bmatrix} sim(I_1,I_1) & sim(I_1,I_2) & L & sim(I_1,I_k) \\ sim(I_2,I_1) & sim(I_2,I_2) & L & sim(I_2,I_k) \\ L & L & L & L \\ sim(I_k,I_1) & sim(I_k,I_2) & L & sim(I_k,I_k) \end{bmatrix} \quad (2)$$

where each entry of similarity matrix $S_{ij}=sim(I_i,I_j)$ represen the similarity between I_i and I_j.



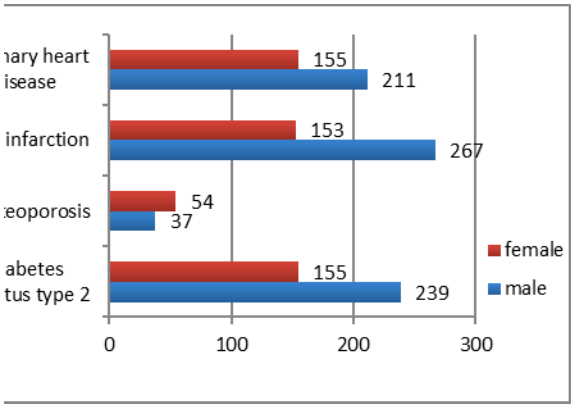
The basic information of a patient
Patient's name:*** Gender: Female Age: 53

The exceptions of lab tests

Lab test	Result	Biological reference interval	Mark
TG	4.5	0.56-1.70	up
TC	8.8	3.10-5.70	up
TL	9.3	4.0-7.0	up
HDL-C	0.4	0.73-2	down

Fig. 4. The Basic Information and Lab Tests of a Patient

n an unstructured
; technology was
each patient from
r her disease [46].
ords, 9 common



g. 5. The Statistics of Different Diseases, in terms of gender

Then, the label matrix is recreated by label-to-label similarity matrix, as follows:

$$\mathbf{Y}=g(Y\times S) \quad (3)$$

where $g(g)$ is a function for transforming each entry of matrix to 0 or 1. $g(g)$ is defined as follows:

$$g(X)=\begin{cases} 1, & X_{ij}\geq 0.5 \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

After updating each label vector, traditional multi-label learning algorithm ML-*k*NN [34] is carried out to estimate possible labels for testing samples, as detailed under Algorithm 1. For the purpose of recommending illness to physicians, the step of 'Generating suggested label vector' is changed to 'Choose N labels whose probabilities are ranked top N in the probabilities of all labels'.

IV. ANALYSIS OF HEALTH DATA

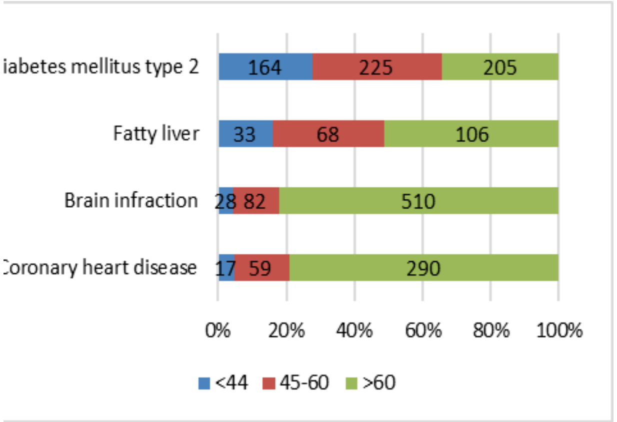
A. Health Medical Data

The authors analyzed laboratory test data and some health medical records of a randomly selected patient, which were collected by a local hospital from 18th May to 18th October. Laboratory test data was stored in the databases in a structured format, while basic information, medical history and

B. Correlation between Information and Diseases

1) Laboratory Test Data: A patient, who was diagnosed with hyperlipidemia, was selected from the health medical database. This patient's exceptions of laboratory test results and basic information are listed in Fig. 4. This patient is a female and 53 years old, and her condition reflects the fact that the incidence of hyperlipidemia is higher in menopausal women [47]. Triglyceride (TG), Total Cholesterol (TC), Total Lipids (TL) and High Density Lipoprotein Cholesterol (HDL-C) are shown as the exceptions of her lab tests. In the patients with hyperlipidemia, the results of TG, TC and TL are usually higher than their biological reference intervals, whereas the result of HDL-C is lower [48]. Hence, physicians require laboratory test results to diagnose some diseases.

2) Gender: Fig. 5 shows the number of patients, in terms of gender, suffering from coronary heart disease, brain infarction, fatty liver and diabetes mellitus type 2. For coronary heart



raction, the
males. The
higher than
re, men are
owever, the
han that of
women than
diseases are

Fig. 6. The Statistics of Different Diseases, in terms of age

. 7. The

Fig. 8.

3) Age: The numbers of patients, in terms of age, suffering from coronary heart disease, brain infraction, fatty liver and diabetes mellitus type 2 are shown in Fig. 6. According to the standards of World Health Organization (WHO) [50], age is divided into 3 groups, namely young group (<44), middle-age group (45-60) and old-age group (>60). From the figure, it can be seen that the incidence of coronary heart disease and brain infraction is higher in the old-age group than in other groups, implying thereby that people of over-60 years of age are more vulnerable to be affected by those illnesses. The incidence of fatty liver in the young and middle age groups is almost the same as that in the old-age group. Furthermore, the percentage of people from the middle age group affected by diabetes mellitus type 2 is the largest, almost 65%. This implies that fatty liver and diabetes mellitus type 2 are two common diseases of middle age group. These statistics again demonstrate that some diseases are correlative with age.

C. *Correlation between Diseases*

In the medical context, complication is a negative evolution or consequence of a disease or a physical condition. Once the physician diagnoses the disease of a patient, he or she will make a list of the disease's most common complications and recommend the treatment required for recognizing or preventing those complications, easily and rapidly.

Diabetes mellitus type 2 and hyperlipemia are the two most common diseases in the world. Diabetes mellitus type 2 can give rise to more than 100 complications, and hyperlipemia has the propensity of inducing some heart diseases. The diagnostic results of 495 patients with diabetes mellitus type 2 and those of 117 patients with hyperlipemia were analyzed and the results are shown in Figs. 7 & 8, in terms of the percentages of patients affected by associated complications.

In Fig. 7, the percentages of brain infarction, coronary heart disease, fatty liver and hyperlipemia are, respectively, 34.7%, 24%, 18.8% and 17.4%. These figures suggest that patients with diabetes mellitus type 2 are more likely to get those complications. In Fig. 8, the percentages of six diseases have exceeded 19%, especially those of diabetes mellitus type 2 and brain infarction, which have reached 73.5% and 65% respectively. This demonstrates that hyperlipemia is closely related to diabetes mellitus type 2 and brain infarction.

V. Model Validation

A. *Multi-label Dataset*

Patients with one or more of the nine common diseases, namely diabetes mellitus type 2, hyperlipemia, fatty liver, kaliopenia, diabetic nephropathy, brain infarction, coronary heart disease, hypoalbuminemia and osteoporosis, were chosen from Haikou people's hospital as experimental samples. Their laboratory test reports and basic information were collected as the input features. After data cleaning, 459 cases with five basic attributes of patients and 278 items of laboratory test reports were obtained. Gender, age, temperature, height and weight were extracted as the basic attributes of patients from the textual medical health records by text processing methods. The value of gender is binary, i.e.,

STATISTICS OF INPUT FEATURES

nput Features		Number	Mean
ic Information			
nder	Male	279	
	Female	180	
e			64.64
nperature			36.5
ght			167.81
ight			67.75
test results			
ns		278	

TABLE II
STATISTICS OF OUTPUT LABELS

Output labels	The number of patients
betes mellitus type 2	168
erlipemia	34
ty liver	138
iopenia	36
oalbuminemia	84
betic nephropathy	17
in infarction	147
onary heart disease	127
eoporosis	7

TABLE III
EXPERIMENTAL RESULTS OF MULTI-LABEL LEARNING ALGORITHMS

orithm	Ham loss ↓	Precision ↑	Recall ↑	F1 score ↑
IL-*k*NN	**0.2117**	**0.2360**	**0.3793**	**0.2915**
-*k*NN	0.2594	0.2133	0.3366	0.2611
-*k*NN	0.2622	0.2049	0.3157	0.2485

ne smaller the value is, the better the performance is. ↑ : The larger the value better the performance is.

male is 0 and female 1. Age, temperature, height and weight were recorded as numerical values, and hence their true numerical values were kept in the vector of attributes. The lab test values include both numerical values and textual description values. For an item with numerical value, the value was set to 0, when it fell in the reference range, and to true numerical value when the value fell beyond or below the reference range. For an item with textual description value, different textual description values of that item were collected and arranged as a list with the help of a clinical expert. If the textual description value was the same as that of the reference, it was set to 0 in the vector; otherwise, it was set to its corresponding index number in the list. The statistics of the final features and those of the final labels are shown in TABLES I & II. Sixty percent of the patients were males and remaining patients were females. The average age, temperature, height and weight of the patients were 64.64, 36.5, 167.81 and 67.75 respectively. From the statistics of labels (illnesses), diabetes mellitus type 2 and brain infarction were found to be the two most common illnesses among the illnesses under consideration. Actually, these diseases are common among the old people.

B. Evaluation Metrics

The evaluation metrics of multi-label classification problems are divided into two groups: (1) Rank-based evaluation metrics, whose purpose is to rank relevant cases before irrelevant cases; (2) Binary prediction measures, whose purpose is to make a strict yes/no classification about each target sample. In this study, Hamming Loss, Precision, Recall and F1-score were employed.

Hamming Loss is defined as the average difference between the suggested and true labels of test samples, which is assessed thus:

$$Hamloss = \frac{1}{p}\sum_{i=1}^{p} | h(x_i) \Delta Y_i | \qquad (5)$$

where $h(x_i)$ is the set of suggested labels of test sample x_i ; p is the number of test sets; Y_i is the set of true labels of test sample x_i ; and Δ is the symmetric difference.

Precision is defined as the ratio of the number of hit labels in the recommended list and the number of all suggested labels in the recommended list._The Precision is calculated as follows:

$$\text{Precision} = \frac{1}{p}\sum_{i=1}^{p} | \frac{Y_i \cap h(x_i)}{h(x_i)} | \qquad (6)$$

Recall is defined as the ratio of the number of hit labels in the recommended list and the number of all true labels from a target sample. Recall is calculated as follows:

$$\text{Recall} = \frac{1}{p}\sum_{i=1}^{p} | \frac{Y_i \cap h(x_i)}{| Y_i |} | \qquad (7)$$

F1 score synthesizes Precision and Recall for taking these evaluation metrics into account. In other words, this score considers both false positives and false negatives. The following is the formula for obtaining F1 score:

$$\text{F1} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \qquad (8)$$

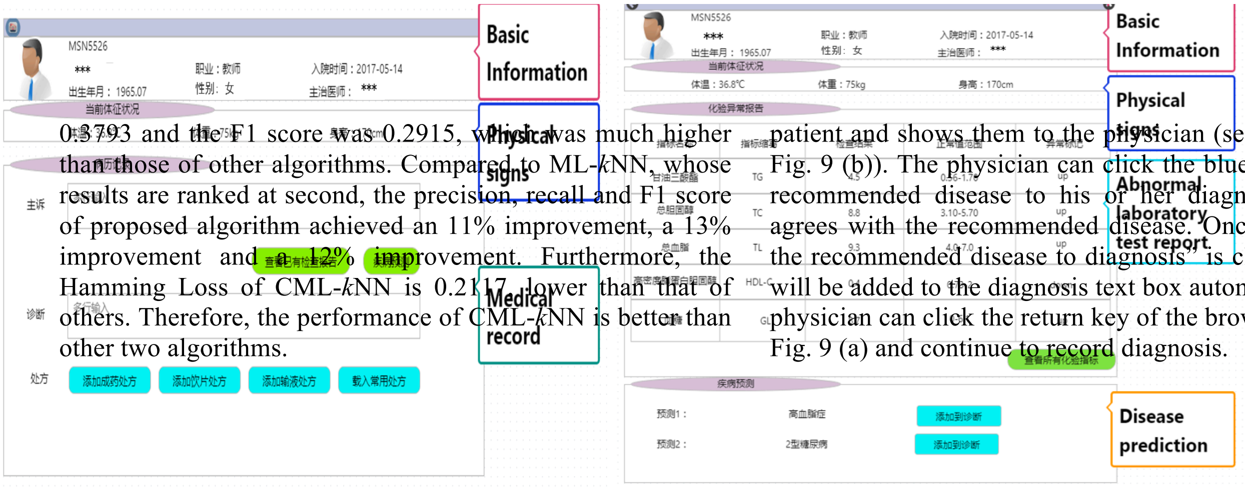
C. Model Performance

For the purpose of evaluating the performance of the proposed model, the following state-of-art methods are chosen for comparison.

1) BR*k*NN: By combining BR method and *k*NN algorithm, Spyromitros et al. [34] proposed the BR*k*NN method, in which the labels of k neighbors of a test sample are used to estimate the possible labels of the test sample. The number of neighbors is set to 10.
2) ML-*k*NN: A lazy learning approach to multi-label learning, proposed by Zhang et al. [13], combines k-nearest neighbor algorithm with Bayesian theory into multi-label learning method. Through learning label information from k-nearest neighbors of each unknown sample, it estimates the possible labels based on maximum a posteriori principle. The number of neighbors is set to 10, and the smoothing factor to 1.

In CML-*k*NN, the number of neighbors and smoothing factor are set as ML-*k*NN. In all the algorithms, the number of suggested labels was 2. Seventy percent of 459 cases were included in the training set and the rest in the test set. Ten-fold cross validation was exploited to perform the experiments, and the final result is the average value of 10 experiments' results. The experimental results of different algorithms are shown in TABLE III.

For CML-*k*NN, its precision was 0.236, the recall was



(a)

(b)

Fig. 9. Two Screenshots of the Clinical Decision Support System for Heterogeneous Data Sources Prototype

0.3793 and the F1 score was 0.2915, which was much higher than those of other algorithms. Compared to ML-*k*NN, whose results are ranked at second, the precision, recall and F1 score of proposed algorithm achieved an 11% improvement, a 13% improvement and a 12% improvement. Furthermore, the Hamming Loss of CML-*k*NN is 0.2117, lower than that of others. Therefore, the performance of CML-*k*NN is better than other two algorithms.

D. System Implementation

The proposed clinical decision support system prototype was so developed that it can be run on the web browser. Python was used to process health data and implement CML-*k*NN, and Mysql was used to manage the structured data. Fig. 9 shows two screenshots of the clinical decision support system prototype; Fig. 9 (a) is physician's main work interface and Fig. 9 (b) is the lab test results report of the patient. The work interface (see Fig. 9 (a)) shows the basic information of the patient (see the pink region), some physical signs of the patient (see the purple region) and medical record (see the green region). The patient's basic information includes name, gender (Female), birth day (1965), occupation, visiting time and physician's name; the patient's temperature (36.8), height (170cm) and weight (75kg) are shown against the physical signs. To safeguard the patient's privacy, only those of his or her attributes, which are considered necessary for the proposed method, are described. Medical record requires that the physician fill in the chief complaint of the patient, diagnosis of patient's disease(s) and prescriptions, including the medical prescription, decoction pieces prescription, transfusion prescription and common prescription. Sometimes, the physician needs to check the correctness of his or her diagnosis for which he or she needs the laboratory test results of the patient. Once these results are uploaded into the system, the physician can click the left green button to access the laboratory test results (see Fig. 9 (b)). Also, the physician can see the abnormal laboratory test results (see the blue region in Fig. 9 (b)) and review all the laboratory test results by clicking the green button. Based on the laboratory test results and basic attributes, the system identifies two possible diseases of the patient and shows them to the physician (see orange region in Fig. 9 (b)). The physician can click the blue button to add the recommended disease to his or her diagnosis if he or she agrees with the recommended disease. Once the key of "add the recommended disease to diagnosis" is clicked, the disease will be added to the diagnosis text box automatically. Then the physician can click the return key of the browser to get back to Fig. 9 (a) and continue to record diagnosis.

VI. Conclusions and Future Work

Promoted by Industry 4.0 and Health 4.0, Health data from massive medical infrastructures (Cyber-Physical Systems, the Internet of Things) is collected continually to generate "Big data" in the health domain. To fully utilize the health data and support smart health services, the authors propose here a clinical decision support framework, which integrates heterogeneous health medical data from different sources, such as laboratory test results, basic information of patients, health medical records and monitoring data, including structured and unstructured data, to construct an integrated representation of features for all the patients. An improved multi-label classification was applied to this representation of features to recommend suggested diseases to physicians in the framework. After the physician diagnoses the disease of a visiting patient, he or she has to consider the complications of that disease for recognizing possible other diseases. It is an accepted fact that correlations do exist among some diseases. Therefore, a k-nearest neighbors algorithm was improved for multi-label learning by using correlations among labels (CML-*k*NN), which can be applied it in the proposed framework to recommend diseases to physicians.

Some experiments were performed on real medical data to evaluate the proposed framework. For conducting the experiments, patients with 9 common diseases were selected as samples. Five kinds of basic information and 278 laboratory test results of the patients were combined to generate integrated features and carry out multi-label learning methods. The experimental results show that the improved multi-label classification method performs better than the existing methods. Based on the design of the proposed framework, the clinical decision support system prototype was developed as well.

For their future work, the authors propose to continue integrating textual and monitoring data to generate more comprehensive integrated features for each patient. The increasing diversity in data types calls for an appropriate method to decrease the number of integrated features for ensuring the efficiency of the clinical decision support system. Because of the scale of labels, the processing of improved multi-label algorithm will be a little slow. Therefore, a more appropriate and efficient method to correlate labels will have to be developed.

References

[1] M. Hermann, T. Pentek, and B. Otto, "Design principles for industrie 4.0 scenarios," in Proc. 49th Hawaii Int.Conf. Syst. Sci., 2016, pp. 3928–3937.

[2] C. Thuemmler and C. Bai, Health 4.0: How Virtualization and Big Data are Revolutionizing Healthcare. New York, NY, USA: Springer, 2017.

[3] P. Cao, "The impact of big data on the development of Chinese health service industry," J. Electron. Test., vol. 21, no. 11, pp. 111–112, 2014.

[4] M. Akay, D. I. Fotiadis, K. S. Nikita, and R.W.Williams, "Guest editorial: Biomedical informatics in clinical environments," IEEE J. Biomed.Health Informat., vol. 19, no. 1, pp. 149–150, Jan. 2015.

[5] A. D. Black et al., "The impact of eHealth on the quality and safety of health care: A systematic overview," PLoS Med., vol. 8, no. 1, 2011, Art. no. e1000387.

[6] J. Horsky et al., "Methodological review: Interface design principles for usable decision support: A targeted review of best practices for clinical prescribing interventions," J. Biomed. Informat., vol. 45, no. 6, pp. 1202–705 1216, 2012.

[7] A. C. Valdez et al., "Recommender systems for health informatics: State-of-the-art and future perspectives," in Machine Learning for Health Informatics. New York, NY, USA: Springer, 2016.

[8] W. D. Yu et al., "A modeling approach to big data based recommendation engine in modern health care environment," in Proc. IEEE 39th Annu. Comput. Softw. Appl. Conf., 2015, pp. 75–86.

[9] U. A. Bhatti et al., "Research on the smartphone based eHealth systems for strengthening healthcare organization," in Smart Health. New York, NY, USA: Springer, 2017.

[10] D. Vancampfort et al., "The prevalence of diabetes mellitus type 2 in people with alcohol use disorders: A systematic review and large scale meta-analysis," Psychiatry Res., vol. 246, pp. 394–400, 2016.

[11] M. H. Drazner, "The progression of hypertensive heart disease," Circulation, vol. 123, no. 3, pp. 327–334, 2011.

[12] Y.Wang, P. F. Li, Y. Tian, J. J. Ren, and J. S. Li, "A shared decision making system for diabetes medication choice utilizing electronic health record data," IEEE J. Biomed. Health Informat., vol. 21, no. 5, pp. 1280–1287, Sep. 2017.

[13] M. L. Zhang and Z. H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," Pattern Recog., vol. 40, no. 7, pp. 2038–2048, 2007.

[14] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," IEEE Access, vol. 2, pp. 652–687, 2017.

[15] Z. Yu et al., "Incremental semi-supervised clustering ensemble for high dimensional data clustering," IEEE Trans. Knowl. Data Eng., vol. 28, no. 3, pp. 701–714, Mar. 2016.

[16] M. Mukaka, "A guide to appropriate use of correlation coefficient in medical research," Malawi Med. J., vol. 24, no. 3, pp. 69–71, 2012.

[17] S. Rallapalli, R. R. Gondkar, and U. P. K. Ketavarapu, "Impact of processing and analyzing healthcare big data on cloud computing environment by implementing hadoop cluster," Procedia Comput. Sci., vol. 85, pp. 16–22, May 2016.

[18] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, "What can natural language processing do for clinical decision support?" J. Biomed. Informat., vol. 42, no. 5, pp. 760–772, 2009.

[19] W. Li et al., "A study on the relation between the age and gender factors for fatty liver, hyperlipidemia and the insulin resistance," Gen. Med. J.,vol. 3, no. 5, pp. 355–356, 2000.

[20] H. R. Brian and N. L.Wilczynski, "Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: Methods of a decision-maker-researcher partnership systematic review," Implementation Sci., vol. 5, no. 1, 2010, Art. no. 12.

[21] M. Viceconti, P. Hunter, and R. Hose, "Big data, big knowledge: Big data for personalized healthcare," IEEE J. Biomed. Health Informat., vol. 19, no. 4, pp. 1209–1215, Jul. 2015.

[22] K. Zarkogianni and K. S. Nikita, "Personal health systems for diabetes management, early diagnosis and prevention," in Handbook of Research on Trends in the Diagnosis and Treatment of Chronic Conditions.Hershey, PA, USA: IGI Global, 2015, p. 465.

[23] E. C¸ omak, A. Arslan, and I˙. Tu¨rkog˘lu, "A decision support system based on support vector machines for diagnosis of the heart valve diseases," Comput. Biol. Med., vol. 37, no. 1, pp. 21–27, 2007.

[24] M. Anthimopoulos et al., "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," IEEE Trans.Medical Imaging, vol. 35, no.5, pp. 1207–1216, 2016.

[25] L. Song,W. Hsu, J. Xu, and M. van der Schaar, "Using contextual learning to improve diagnostic accuracy: Application in breast cancer screening," IEEE J. Biomed Health Informat., vol. 20, no. 3, pp. 902–914, May 2016.

[26] P. Anooj, "Clinical decision support system: Risk level estimation of heart disease using weighted fuzzy rules," J. King Saud Univ., Comput. Inf. Sci., vol. 24, no. 1, pp. 27–40, 2012.

[27] J. Xu, D. Sow, D. Turaga, and M. van der Schaar, "Online transfer learning for differential diagnosis determination," in Proc. AAAI Workshop World Wide Web Public Health Intell., 2014, pp. 25–29.

[28] S. Molinaro, S. Pieroni, F. Mariani, and M. N. Liebman, "Personalized medicine: Moving from correlation to causality in breast cancer," NewHorizons in Translational Med., vol. 2, no. 2, 2015.

[29] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining multi-label data," in Data Mining and Knowledge Discovery Handbook. New York, NY, USA: Springer, 2010, pp. 667–685.

[30] M. R. Boutell et al., "Learning multi-label scene classification," Pattern Recog., vol. 37, no. 9, pp. 1757–1771, 2004.

[31] A. Clare and R. D. King, "Knowledge discovery in multi-label phenotype data," Lecture Notes Comput. Sci., vol. 2168, no. 2168, pp. 42–53, 2001.

[32] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in Proc. Int. Conf. Neural Inf. Process. Syst., Natural Synthetic, 2001, pp. 681–687.

[33] M.-L. Zhang and Z.-H. Zhou, "Multilabel neural networks with applications to functional genomics and text categorization," IEEE Trans. Knowl. Data Eng., vol. 18, no. 10, pp. 1338–1351, Oct. 2006.

[34] E. Spyromitros, G. Tsoumakas, and I. Vlahavas, "An empirical study of lazy multilabel classification algorithms," in Artificial Intelligence: Theories, Models and Applications. Berlin, Germany: Springer, 2008, pp. 401–406.

[35] J. Xu et al., "Classifying syndromes in Chinese medicine using multi-label learning algorithm with relevant features for each label," Chin. J. Integr. Med., vol. 22, no. 11, pp. 867–871, 2016.

[36] A. Frazer-Abel et al., "Overview of laboratory testing and clinical presentations of complement deficiencies and dysregulation," Adv. Clin. Chem., vol. 77, no. 1, pp. 1–75, 2016.

[37] C. Shivade et al., "A reviewof approaches to identifying patient phenotype cohorts using electronic health records," J. Amer. Med. Informat. Assoc., vol. 21, no. 2, pp. 221–30, 2014.

[38] M. D. Hoffman, D. M. Blei, and F. R. Bach, "Online learning for latent Dirichlet allocation," Adv. Neural Inf. Process. Syst., vol. 23, pp. 856–864, 2010.

[39] L. S. C. De Oliveira, R. V. Andre˜ao, and M. Sarcinelli-Filho, "Premature ventricular beat classification using a dynamic Bayesian network in engineering," in Proc. Ann. Int. Conf. IEEE Med. Biol. Soc., 2011, pp. 4984–4987.

[40] H. F. Huang, G. S. Hu, and L. Zhu, "Sparse representation-based heart beat classification using independent component analysis," J. Med. Syst., vol. 36, no. 3, pp. 1235–1247, 2012.

[41] S. Banerjee and M. Mitra, "Application of cross wavelet transform for ECG pattern analysis and classification," IEEE Trans. Instrum. Meas. vol. 63, no. 2, pp. 326–333, Feb. 2014.

[42] MIT-BIH, "Arrhythmia database directory." [Online]. Available: http:// www.physionet.org/physiobank/database/html/mitdbdir/mitdbdir.html. Accessed on: Oct. 18, 2017

[43] L. I. Smith, "A tutorial on principal components analysis," Inf. Fusion, vol. 51, no. 3, 2002, Art. no. 52.

[44] A. C. Kak and A. M. Mart´ınez, "PCA versus LDA," IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, nos. 3–4, pp. 228–233, Feb. 2001.

[45] F. Kang, R. Jin, and R. Sukthankar, “Correlated label propagation with application to multi-label learning,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog., 2006, pp. 1719–1726.

[46] G. Weikum, “Foundations of statistical natural language processing,” Inf. Retrieval J., vol. 4, no. 1, pp. 80–81, 2001.

[47] W. Li et al., “A study on the relation between the age and gender factors for fatty liver, hyperlipidemia and the insulin resistance,” Gen. Med. J., vol. 3, no. 5, pp. 355–356, 2000.

[48] D. Hu and Z. Xiang, “How to correctly diagnose and treat hyperlipidemia?” Chin. J. Med., vol. 24, no. 4, pp. 25–27, 2000.

[49] T. H. Lin et al., “Association between periodontal disease and osteoporosis by gender: A nationwide population-based cohort study,” Medicine, vol. 94, no. 7, 2015, Art. no. e553.

[50] [Online]. Available: http://www.who.int/healthinfo/survey/ageing_mds_report_en_daressalaam.pdf.