Expert Systems With Applications 151 (2020) 113361

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Modeling methodology for early warning of chronic heart failure based on real medical big data[☆]

Chunjie Zhou[a,c,*], Ali Li[a,*], Aihua Hou[b], Zhiwang Zhang[a], Zhenxing Zhang[a], Pengfei Dai[c], Fusheng Wang[d]

[a] *Department of Information and Electrical Engineering, Ludong University, Shandong, China*
[b] *Department of Oncology, Yantai City Hospital of Chinese Medicine, Shandong, China*
[c] *Yantai Cloud Software Co., Ltd., Shandong, China*
[d] *Department of Biomedical Informatics and Computer Science, Stony Brook University, New York, USA*

ARTICLE INFO

Article history:
Received 4 August 2019
Revised 20 February 2020
Accepted 5 March 2020
Available online 6 March 2020

Keywords:
Heart failure
Early warning
Social network
Risk factors
Medical big data

ABSTRACT

Heart failure (HF) is among the most costly diseases to our society, and the prevalence keeps on increasing these days. Early detection of HF plays a vital role in saving lives through adjusting lifestyles and drug interventions that can slow down disease progression or prevent HF. There are many cardiovascular risk factors associated with HF, and they often coexist. In this paper, we assess the predictive value of pathological factors for early HF detection through a social network based approach. We use electronic health records (collected from the project HeartCarer) and compute the similarity of risk factors. The similarity values are used to construct an unweighted and a weighted medical social network. The constructed medical social network is further divided into a HF high-risk group and HF low-risk group using a group division algorithm. Patients in the high-risk group will be suggested for early screening. To evaluate the prediction value of our method, we perform four experiments based on real world data. The results demonstrate the high effectiveness of our method on heart failure risk assessment, with the best accuracy close to 90%.

1. Introduction

Heart failure (HF), a common condition causing severe morbidity and mortality, is increasing in prevalence in many areas of the world (Delanaye, Guerber, & Scheen, 2017; Hiragi, Tamura, & Goto, 2018; Ravizza, Huschto, & Adamov, 2019). More than 26 million people worldwide are affected by HF, and this number is growing rapidly every year (Bekhet, Yonghui, & Ningtao, 2018; Miao, Cai, & Zhang, 2018). There are at least 10 million patients with heart failure in China, and it has become one of the countries with the largest numbers of heart failure patients (Meystre, Kim, & Gobbel, 2017). The annual mortality rate of hospitalized patients with heart failure is 35% to 45%, and the 5-year survival rate is comparable to malignant tumors (Maddalena, Salvini, & Bardoni, 2019; Ralph, 2018). Fortunately, heart failure is a chronic progressive disease. If risk factors for heart failure can be controlled at an early stage, heart failure can be delayed or prevented.

However, due to the diversity of alternative interpretations of symptoms, HF is a heterogeneous and complex disease that is difficult to detect in routine cares (Bekhet et al., 2018; Li, Wang, & Liu, 2019). The existing research works mainly consider the analysis of ECG, such as Auto-regressive (AR) spectral estimation and Fourier transformation (Bohacik, Matiasko, & Benedikovic, 2015; Vikram, Esther, & Wiro, 2019). Unfortunately, when an ECG abnormality is detected, the patient may already experience heart failure. In fact, the long-term existence of certain risk factors increases the likelihood of heart failure. If we can analyze and track these risk factors in advance, and predict those high-risk groups, we can effectively prevent heart failure by targeted drug intervention or lifestyle changes.

Interdisciplinary collaboration in disease prediction, etiology analysis and diagnosis has become a major trend, with the improvement and enrichment of biological networks and clinical big data (Fabian, Paul, & Peter, 2017; Kalid, Zaidan, & Zaidan, 2018; Leonardo, Jose, & Ruben, 2018; Sengul & Ibrahim, 2018). With the development of wearable computing (Abdelaziz, Elhoseny, & Salama, 2018) and deep learning (Chen, Ma, & Li, 2017), more re-

[☆] This research was partially supported by the grants from the National Natural Science Foundation of China (No. 61202111, 61472141, 61273152, 61303017, 61877061); the Natural Science Foundation of Shandong Province China (No.ZR2011GQ001).

[*] Corresponding author.
E-mail addresses: luckyzcj@163.com (C. Zhou), allii2000@sohu.com (A. Li).

https://doi.org/10.1016/j.eswa.2020.113361
0957-4174/© 2020 Elsevier Ltd. All rights reserved.

search uses big data analysis to predict disease risks (Lin, Xia, & Li, 2019). HF prediction using traditional disease risk models typically involves machine learning algorithms, e.g., regression analysis and logistic regression. To improve the accuracy and reliability, existing related research considers as many factors or features as possible (Khalifa & Meystre, 2015) and adds them to the regression model, but this approach can lead to low efficiency and poor interpretability. Khalifa et al. recommend the use of clinical records to assess cardiovascular risk factors (Khalifa & Meystre, 2015). However, multimodel disease risk assessments are considered by only few works, and most existing works only consider unstructured data or structured data. Based on a convolutional neural network, literature (Chen et al., 2017) proposed a multimodel disease risk prediction algorithm. Based on a recursive neural network, structured data and unstructured data of cerebral infarction patients were considered in Hao, Usama, and Yang (2019). Big data technologies can help risk prediction and early diagnosis of chronic kidney disease (Fu, Zeng, & Liu, 2018). In order to process time series data with high dimensionality, Zhou et al. proposed a unified solution that comprehensively considers time granularity, time interval, and time chaos (Zhou, Wang, & Zhang, 2019).

Cardiovascular risk factors are responsible for the highest percentage of cases of HF. For example, smoking and drinking may cause hypertension, which is a common cardinal causes of heart failure (Chang, Wang, & Jiang, 2011; Rani, Kannan, & Vasantha, 2012). Hypertension should not be considered in isolation because other risk factors such as infection, diet, arrhythmia, personal disease history, mental stress, long-term smoking and alcohol abuse, obesity, excessive physical exertion, family history of heart failure, motion, environment have also been a considerable impact on the progression of heart failure Kim (2015), Acharya, (2017), Khalifa, (2018). These risk factors can regulate the cellular and molecular roles in pathological and physiological settings Acharya, (2017). By undertanding these risk factors, proper outpatient management and targeted treatment can reduce the severity and frequency of exacerbations.

Social networks have helped to solve some questions related to people's personal heath and public health. In the research of social networks, scientists discovered that some social networks in the real world have community structure characteristics (Zhou, Cao, & Dong, 2013). A community discovery algorithm can divide a social network into different groups based on connection strength. The connections among internal nodes in a community are relatively tight, but the connections between the various communities are relatively sparse. The community structure reveals the common interests, hobbies or backgrounds of the social groups in the social network and serves as the inspiration for the proposed approach for heart failure risk assessment. In this paper, to tackle the complexity and interpretability of HF, we propose a social network-inspired method for predicting the risk of HF. In order to better carry out heart failure assessment, we consider risk factors that can easily cause heart failure. An unweighted and a weighted medical social network are constructed respectively according to the similarity values among risk factors based on the electronic health records. Based on this, a person's health can be denoted as a probabilistic combination of risk factors. The model will also provide HF assessment of patients at different time. The constructed medical social network is divided into a HF high-risk group and a HF low-risk group using a group division algorithm, and those falling into the HF risk groups are suggested for early screening.

We evaluated our methods based on a real world dataset provided by the research project *HeartCarer*, a home-based remote monitoring system based on a cloud platform designed to monitor HF patients for timely interventions. Our experiments demonstrate the best accuracy can reach close to 90%.

2. Preliminaries

We first introduce the dataset and provide an overview of our methods.

2.1. Data set

The HeartCarer system collects various physiological signs and risk factor information through wearable devices, inquiry or statement, and uploads it to the cloud through a mobile terminal or a telephone line. The architecture of the system is shown in Fig. 1, that allows third-party systems to access to prediction services and provide monitoring data. Data collection is the first part, which is mainly from our project-HeartCarer, also involves inspection hospital records or other methods. The collected data can be used both in real-time and off-line to derive multiple inferences about the patients condition. Constructing the medical social network is the second part. Based on the collected data, the unweighted or weighted medical social network is constructed by calculating the similarity values among rf_is. The group division is the third part. The medical social network is divided into different groups using the group division algorithm. Then through applying and calculating the necessary threshold value, two groups (high-risk group and low-risk group) are discovered.

There was a cohort of 1026 HF patients distributed in six medical institutions in China, and these patients received care between 2015 and 2018. From these patients, we selected 800 (78%) of them with telemonitoring measurements for at least 30 days. The range of age is 63.8 ± 12 years and (70%) of them were male. For this project, family members can use the devices as well. As a result, we also included 7500 family members for the study. The ages of family members were between 10 and 90. In addition, we also included a total of 105,239 healthy people from the provinces the branch offices of the sponsoring company were located in.

In a summary, our study includes a total of 1026 heart failure patients who were identified with HF, 7500 family members closely related to the patients, and 105,239 healthy people. The size of the data set is more than 100GB. Detailed presence of contributing cardiovascular risk factors were collected for each person and recorded in the data collection forms designed exclusively for the study. In our research, we collected eleven candidate risk factors, including infection (e.g., respiratory infection, lung infection), diet (e.g., poor control of water and sodium), arrhythmia(e.g., atrial fibrillation, anemia, renal impairment), personal disease history (e.g., hypertension, coronary heart disease, diabetes, hyperlipidemia, atherosclerosis), mental stress (eg. emotional, fury), long-term smoking and alcohol abuse, obesity, excessive physical exertion, family history of heart failure, motion and environment (eg. air pollution, forced secondhand smoking). Cardiovascular risk factors were defined as follows: Smoking: current smoking of $\geq$ 1 pack/day; obesity: if the BMI $>$ 30 kg/m^2, family history: history of heart failure or sudden cardiac death before the age of 55 years in the father or any other first-degree male relative or before the age of 65 years in the mother or any other first degree female relative; Hypertension: having a BP of $>$ 140/90 mmHg on two separate examinations or usage of antihypertensive agents; and diabetes mellitus: fasting glucose $\geq$ 126 mg/dl, postprandial glucose $\geq$ 200 mg/dl and random blood sugar $\geq$ 200 mg/dl.

2.2. Overview of methods

Previous work shows that people with infected first-degree relatives are at high risk of heart failure (Khalifa & Meystre, 2015). Additionally, diet, arrhythmia, personal disease history, among others, are discovered to be related to heart failure (Kalid et al., 2018; Leonardo et al., 2018). So the risk of developing heart failure in the

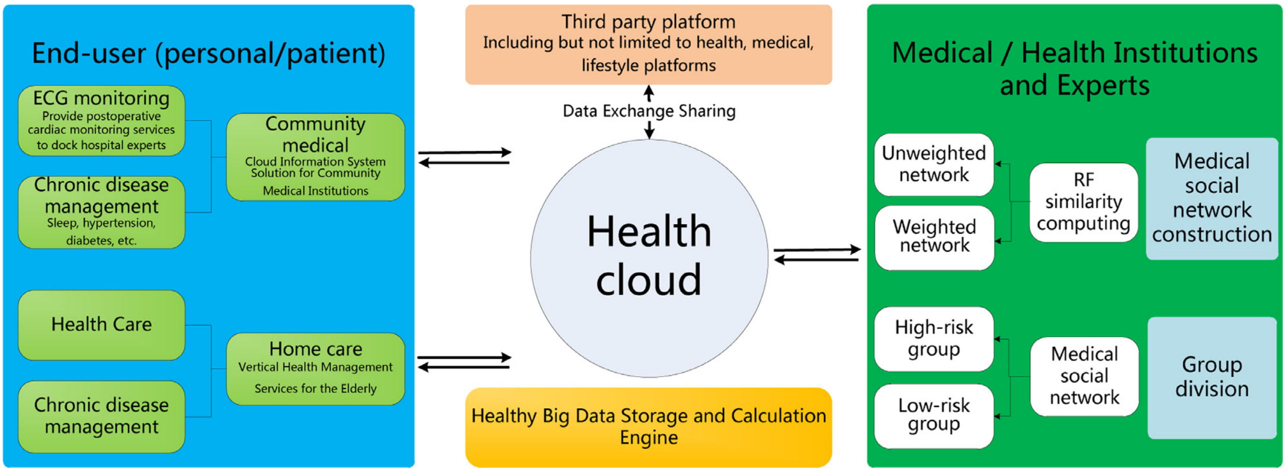

End-user (personal/patient)
ECG monitoring
Provide postoperative cardiac monitoring services to dock hospital experts
Chronic disease management
Sleep, hypertension, diabetes, etc.
Community medical
Cloud Information System Solution for Community Medical Institutions
Health Care
Chronic disease management
Home care
Vertical Health Management Services for the Elderly
Third party platform
Including but not limited to health, medical, lifestyle platforms
Data Exchange Sharing
Health cloud
Healthy Big Data Storage and Calculation Engine
Medical / Health Institutions and Experts
Unweighted network
Weighted network
RF similarity computing
Medical social network construction
High-risk group
Low-risk group
Medical social network
Group division

Fig. 1. The architecture of heart failure risk prediction method.

future are linked to these physiological factors, and might be predicted from these factors. Here, each physiological factor related to heart failure can be defined as a related risk factor (*rf*). $RF_N = (rf_1, rf_2, rf_3, \cdots, rf_N)$ represents the set of all *rf*s, where N is the total number of *rf*s. As mentioned above, the number of candidate risk factors is eleven in this research. RF_i includes i risk factors, $1 \le i \le 11$. For example, RF_2 refers to any two risk factors from the above eleven ones. R_i^j is a set of i different combinations of risk factors, which are randomly taken from *rf*. $\tilde{R_i^j}$ is a set of R_i^j $(1 \le j \le (i+1))$. Suppose there are totally three risk factors (infection, diet, personal disease history), then $\tilde{R_2^j}$ $(1 \le j \le 3)$ can be shown in Eq. (1). The computation of similarity will be discussed in Section 3.2. RF_i is the average of R_i^j $(1 \le j \le (i+1))$, as shown in Eq. (2).

$$\begin{aligned} \tilde{R_2^j} &= \{R_2^1, R_2^2, R_2^3\} \\ R_2^1 &= \{infection, diet\} \\ R_2^2 &= \{infection, personal : disease : history\} \\ R_2^3 &= \{diet, personal : disease : history\} \end{aligned} \tag{1}$$

$$RF_i = \frac{\sum_{j=1}^{(i+1)} R_i^j}{i+1} \tag{2}$$

The role of similarity in diseases has been studied in Li et al. (2019). Based on this, we can infer that the higher the similarity among *rf*s that a person may have with HF patients, the higher risk of the person to develop into the disease. A high-risk group includes all people whose risk factors are similar to heart failure patients'. Different groups are constructed according to the similarity among *rf*s of people. The similarity can be computed according to two people's *rf*s. For instance, the similarity value of two people is high if they have the same HF history. Thus, a network can be constructed based on people's similarity values among *rf*s. The constructed network in this paper is named as the medical social network, which can be used to make medical and health analysis.

Definition: The medical social network can be regarded as a graph based problem, in which nodes are people (labeled with names), and edges are the similarities between these people, i.e., a graph $G(O, \xi)$ will be built with n nodes $P = \{p_1, \ldots, p_n\}$ and s edges $\xi = \{S_{rf}(p_1, p_2), \ldots, S_{rf}(p_{n-1}, p_n)\}$.

In this paper, the constructed medical social network includes an unweighted medical social network and a weighted medical social network. The unweighted social network is used to make a rough assessment of the relationship among people and give a simple classification based on risk factors. The weighted social network is used to calculate the degree of association among people in refined details. The detailed description can be found in Section 3.2.

The similarities among people's *rf*s are described using the constructed medical social network. People are divided into different groups, which are involved in the medical social network, e.g., the group with heart failure. This will enable us to screen patients with high risk early. We use a matrix to represent the weighted medical social network. The similarity values among people's *rf*s are the weights. The goal of this paper is to reflect personal health using a set of related *rf*s. RF_k is a set of k risk factors which can express the health indexes of the corresponding people and associate them with a probabilistic combination of risk factors. Because every risk factor of a subject has a timestamp, we can further represent different heart failure assessments of the same people at different timestamps.

3. The heart failure risk prediction model

Before we discuss the implementation of the prediction model, we first summarize in Table 1 symbols to be used in the explanation.

Table 1
Description of symbols.

S.no	Symbol	Description
1.	F	The number of heart failure patients
2.	H	The number of healthy people
3.	M	The total number of RFs
4.	N	The total number of rfs
5.	P	a set of all people $\{p_1, \ldots, p_{H+F}\}$
6.	RF	a set of all *RF*s $\{RF_1, \ldots, RF_M\}$
7.	rf	a set of all *rf*s $\{rf_1, \ldots, rf_N\}$
8.	$W^{x,y}$	the relationship of x and y using adjacency matrix
9.	GF	The group involved the patients
10.	GH	The group involved the healthy people
11.	$S_{rf_m}(i, j)$	The rf_m similarity of persons i and j
12.	ΔZ_{GP}	When the node n is in the GP, the variation of the modular (Z) measures the quality of the group partition
13.	ΔZ_{GH}	When the node n is in the GH, the variation of the modular (Z) measures the value of the group partition
14.	PR	The risk probability of each group
15.	L	The threshold of high-risk

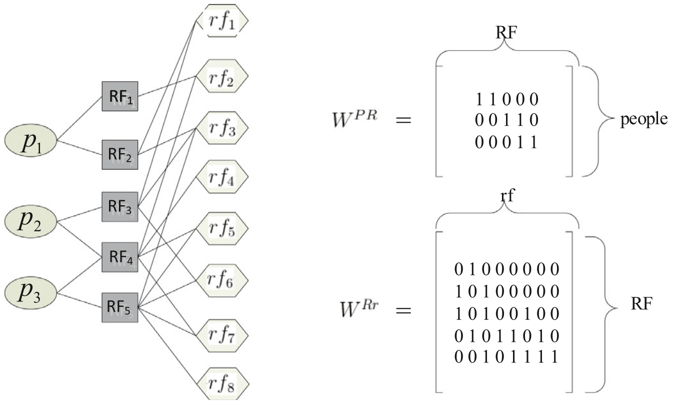


Fig. 2. The relationship expression.

3.1. The construction of the medical social network

The data set includes a large group of individuals, and each of them has several *RF*s. Every RF_m is associated with a set of $rf_i = \{rf_1, rf_2, \ldots, rf_m\} \subseteq RF_m$. Each rf_i is associated with a timestamp $t(rf_i)$, which represents the timestamp to evaluate the risk factor. The risk factor $rf_j \in RF_m$ $(j < m)$ can be used to label a particular risk factor of the subject p_j. $P(rf_j)$ is used to represent the collection of everyone marked by the risk factor rf_j, so we have $P(rf_j) = \{p_1, p_2, \ldots, p_y\} \subseteq P$.

In Fig. 2, we demonstrate the medical social network of three people (p_1, p_2, p_3). These people are related with five *RF*s ($RF_1, \ldots, RF_5$), and each *RF* involves several risk factors *rf*s. Here, these five *RF*s totally involve eight *rfs* ($rf_1, \ldots, rf_8$). To simplify the description, 'people', '*RF*' and '*rf*' are regarded as different types of objects. Moreover, we use the adjacency matrixes ($W^{x,y}$) to model the relationships among different types of objects. For example, the relationship between the 'people' object and the '*RF*' object can be modeled as $W^{P,R}$, where $W^{P,R}(i, j) = 1$ iff the *jth RF* is related with the *i*th people; otherwise, $W^{P,R}(i, j) = 0$. Similarly, the relationship between the '*RF*' object and the '*rf*' object can be modeled as $W^{R,r}$, where $W^{R,r}(i, j) = 1$ iff the *jth* instance of risk factor rf_j is involved in the *ith RF*; otherwise, $W^{R,r}(i, j) = 0$.

Given a particular *RF* and a corresponding set of risk factors *rfs*, the problem we will address is to find a set of related *rfs* to reflect one's personal health. RF_k is a set of *k* risk factors which can express the health indexes of the corresponding people and associate them with a probabilistic combination of risk factors. Because every risk factor of a certain person has a timestamp, we can further represent different heart failure assessment of the same person at different time.

3.2. The similarity computation method

The value of the similarity between two people p_x and p_y is expressed as $S_{rf}(p_x, p_y)$, which is the average of $S_{rf_i}(p_x, p_y)$ $(0 \le i \le N)$, as shown in Eq. (3).

$$S_{rf}(p_x, p_y) = \frac{\sum_{i=1}^{N} S_{rf_i}(p_x, p_y)}{N} \tag{3}$$

Here, $S_{rf_i}(p_x, p_y)$ is the similarity between these two people based on a particular risk factor rf_i. The closeness degree between two people is represented by the similarity value in terms of rf_i. In order to calculate the similarity of document data for recommendation systems, some existing methods have been proposed, such as Euclidean distance and cosine similarity measure (Li et al., 2019). These methods calculate the similarity based on vectors. Here, we use a simple method to calculate the similarity because rf_i has only one value in a certain situation. Divide the absolute difference between the two numbers with their maximum value, and then subtract this value from 1. For example, Eq. (4) expresses the similarity between persons p_x and p_y in terms of rf_i. The value of $S_{rf_i}(p_x, p_y)$ can be 0, 1 or $0 < S_{rf_i}(p_x, p_y) < 1$.

$$S_{rf_i}(p_x, p_y) = 1 - \frac{abs\left(p_{x_{rf_i}} - p_{y_{rf_i}}\right)}{max\left(p_{x_{rf_i}}, p_{y_{rf_i}}\right)} \tag{4}$$

According to Eq. (4), $S_{rf_i}(p_x, p_y)$ is 0 or 1, if rf_i has only two values (i.e., 0 and 1). However, it can be other values in actual cases. Fig. 3 shows an obtained unweighted medical social network, in which, people with the same value of rf_i belong to the same group. Two fully connected sub-graphs are involved in the network. For example, based on personal disease history, a medical social network can be constructed. For the data set in this paper, the history of a disease of a person can be represented as 0 (no existence) or 1 (existence). Therefore, when two people both have the same personal disease history, $S_{rf_i}(p_x, p_y) = 1$. This means that there is an edge connecting the two nodes in the unweighted network. If the personal disease history is different for the two people, $S_{rf_i}(p_x, p_y) = 0$. This means that the two people are not connected by an edge in the network.

In fact, each risk factor in *RF* has a specific weight for each person. $S_{rf_i}(p_x, p_y) \in [0, 1]$, and a weighted medical social network can be obtained, which corresponds to a matrix. The similarity values are the weights. The weighted network and the corresponding matrix are shown schematically in Fig. 4. In this schematic, $G = (6, 15)$. Based on a particular risk factor rf_i, $S_{rf_i}(1, 2) = 0.8$, $S_{rf_i}(1, 3) = 0.2$, $S_{rf_i}(1, 4) = 0.6$, $S_{rf_i}(1, 5) = 0.1$, and $S_{rf_i}(1, 6) = 0.3$. The network is undirected, so $S_{rf_i}(1, 2) = S_{rf_i}(2, 1) = 0.8$. Note that the numbers used in the example are pseudo numbers only for illustration purpose.

3.3. Heart failure expression and detection

To provide a knowledge base of people and risk factors, the first objective is to build the heart failure database. Our expression approach involves three steps. Firstly, we analyze risk factors and express *RF*s as a probabilistic combination of risk factors. As described above, we aim to discover *RF*-related risk factors based on some experience values. Secondly, logical relationships between *rfs* and people are analyzed using a probabilistic topic model (Blei, Ng, & Jordan, 2003), in order to find a combination of risk factors to assess heart failure of people, such as infection, diet, arrhythmia, personal disease history, and so on. Given a set of risk factors, we should classify these *rfs* before assessing heart failure. Finally, from the aspect of time, we can find the difference of heart failure assessment at different time according to risk factors assessed at different timestamps.

Based on the topic model, people's heart failure assessment and their difference in different time can be detected automatically without any user annotation or intervention. To illustrate how this can be achieved, we describe the process of expressing a heart failure. Suppose that different *RF*s have different impact factors for the same people. *RF*s are modeled as a probability distribution $p(RF|P)$ over people *P*. Similarly, the importance of each risk factor for each *RF* is also modeled as a probability distribution $p(rf|RF)$ over the *RF*. Given these two distributions, we can compute the probability of risk factors *rf* occurring in people *P* in Eq. (5).

$$p(rf|P) = \sum_{i=1}^{M} p(rf|RF)p(RF|P) \tag{5}$$

This probability distribution $p(rf|P)$ does not include any notion of *RF*s any more. Having many people, we observe a data matrix

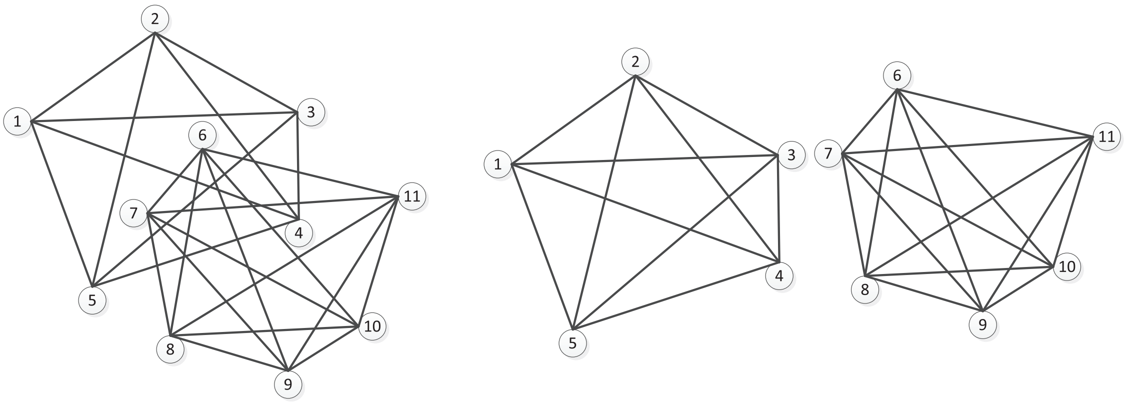
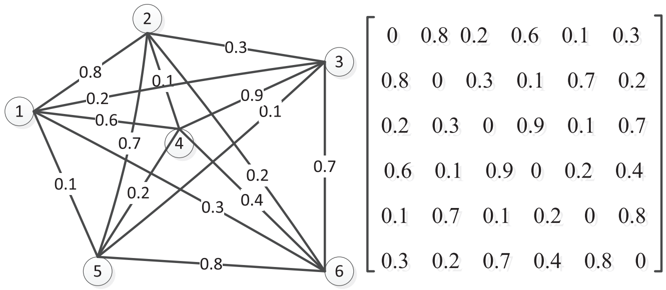


Fig. 3. Unweighted medical social network.



$$\begin{bmatrix} 0 & 0.8 & 0.2 & 0.6 & 0.1 & 0.3 \\ 0.8 & 0 & 0.3 & 0.1 & 0.7 & 0.2 \\ 0.2 & 0.3 & 0 & 0.9 & 0.1 & 0.7 \\ 0.6 & 0.1 & 0.9 & 0 & 0.2 & 0.4 \\ 0.1 & 0.7 & 0.1 & 0.2 & 0 & 0.8 \\ 0.3 & 0.2 & 0.7 & 0.4 & 0.8 & 0 \end{bmatrix}$$

Fig. 4. Illustration of a weighted medical social network and matrix.

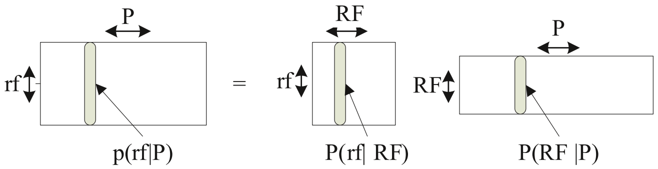


Fig. 5. Expression of heart failure decomposition.

of observed $p(rf|P)$ as depicted on the left hand side of the equation in Fig. 5. According to the equation of the topic model, the data matrix can be reconstructed by a matrix product of the risk factor relevances for each *RF* and a mixture of *RF*s $p(RF|P)$ for each person. Estimating the topic model means a reverse operation. The data matrix on the left-hand side is decomposed into two matrixes on the right-hand side. Based on this, we can recover the characteristic risk factors for each *RF* and the mixture of *RF*s for each person.

Here, Latent Dirichlet Allocation (LDA)-a particular instantiation of topic models is used. β is the control parameter of probability distribution between people and *RF*s, while γ is the control parameter of *RF*s and risk factors. Fitting the model is equivalent to finding the parameter β for the dirichlet distribution and the parameter γ for the $RF - rf$ distributions $p(rf|RF, \gamma)$ that maximizes the likelihood in Eq. (6). Here M is the number of *RF*s, N is the number of risk factors, and $H + F$ is the number of all people.

$$p(rf|\beta,\gamma) = \prod_{r=1}^{H+F} \int p(\theta_r|\beta)\left(\prod_{j=1}^{N}\sum_{i=1}^{M} p(rf_j|RF_i,\beta)p(RF_i|\theta_r)\right)d\theta_r \tag{6}$$

3.4. Group division

A group partitioning algorithm is used to process the medical social network constructed above. Based on the similarity, the medical social network will be divided into a high-risk group and a low-risk group using the group division algorithm. There are healthy people and heart failure patients in the constructed medical social network. Here, the people who have not established a state of heart failure through clinical diagnosis are the healthy people, while the confirmed cases are the patients. The methods of network construction and division are the same for different *RF*s. Many *RF*s will be iterated in the algorithm. We need to initialize the numbers of heart failure patients, healthy people and *RF*s at the beginning of the algorithm. Each execution has different input parameter initialization during the iteration process. The group division outputs of the $(n-1)th$ execution are the inputs of the nth execution. Only one RF_m is used to describe the construction of medical social network and the method of group division is explained in the pseudo-code description below.

From step 1 to 5, the rf_m similarity between people is calculated. In step 6, we construct the network based on rf_m. From step 7 to 19, the constructed network is divided into two medium groups. The division process will continue until all *rf*s have been used. Then the algorithm generates many final groups.

The set H includes all healthy people. The set F includes all heart failure patients. GH stands for the low-risk group, and GF stands for the heart failure high-risk group. We should first initialize GF and GH. $GF = \{p_x \mid$ in order to make sure F is a representative sample, choose only one p_x from F based on medical knowledge$\}$. Here, the medical knowledge is regarded as the general method, including infection, diet, arrhythmia, and other pathological features of the patient. GH can be expressed as Eq. (7) after confirming the GF nodes. The less similarity the more probability that p_x and p_y are in different groups, eg. GF and GH.

$$GH = \left\{y|S_{rf_i}(p_x, p_y) is\ minimum, p_x \in GF, p_y \in H\right\} \tag{7}$$

Firstly, GF or GH is a group based on group initialization, and each other node stands for a group. We use a parameter to measure the quality of group division. If the edges are randomly distributed, the score of the inner edge of the given group minus the expected score is denoted as Z, which is shown in Eq. (8).

$$Z = \sum_{i=1}^{C}\left(E_{ij} - e_i\right)^2 \tag{8}$$

In Eq. (8), the number of edges connecting groups i and j is denoted as E_{ij}. The number of edges connecting group i is denoted as e_i.

Secondly, when a node p_i moves to the group *GF* or *GH*, for all nodes except p_x and p_y, the *Z* gain ΔZ is expressed as Eq. (9).

$$\Delta Z = \left[\frac{\sum in + 2W_{i,in}}{2W} - \left(\frac{\sum out + W_i}{2W} \right)^2 \right] - \left[\frac{\sum in}{2W} - \left(\frac{\sum out}{2W} \right)^2 - \left(\frac{W_i}{2W} \right)^2 \right] \quad (9)$$

The *Z* value of a node has two different situations. One term is the node moving into group *i* (*GF* or *GH*). Another term is the node not being in group *i*, that is to say, it is a group by itself. In Eq. (9), the sum of the links' weights in the group *i* is expressed as Σin, the sum of the links' weights incident to nodes in *i* is expressed as Σout, the sum of the links' weights incident to node p_i is expressed as W_i, the sum of the links' weights from node p_i to nodes in *i* is expressed as $W_{i,in}$, and the sum of all links' weights in the network is expressed as *W*, with value expressed in Eq. (10).

$$W = \frac{1}{2} \sum_{ij} S_{sa}\left(p_i, p_j\right) \quad (10)$$

In Eq. (10), the total weight of all links that incident to node p_i is expressed as $\Sigma_{ij} S_{sa}(p_i, p_j)$. The weight of the edge between p_i and p_j is expressed as $S_{sa}(p_i, p_j)$. Here, the sum of the weights is half of $\Sigma_{ij} S_{sa}(p_i, p_j)$, because the weight of each link is calculated twice in the undirected graph.

ΔZ is represented as ΔZ_{GF} when the node moves to the *GF* group. ΔZ is represented as ΔZ_{GH} when the node moves to the *GH* group. The node joins the *GH* group, if $\Delta Z_{GF} < \Delta Z_{GH}$; otherwise, it joins the *GF* group. *GF* and *GH* are just intermediate groups in the entire model. Each node is involved either in *GF* or in *GH*.

RF_o is a suitable set of risk factors, which will be introduced in detail in Section 4.2.2. rf_o is one of risk factors in RF_o. *PT* stands for the risk probability value, which is calculated by dividing the number of patients by the total number of people. Terminal groups are defined as either the value of *PT* is 1, or the groups generated using rf_o. In the beginning, both *GF* and *GH* are used as inputs, and another *rf* is used to continue grouping. Then, *GH* is a terminal group when the *PT* value of *GH* is equal to 1, and only *GF* is treated as input and joins in the next grouping. After each *rf* in RF_o has been used, the group division will finish, and many terminal groups will be constructed.

The high-risk probability threshold value *L* can be confirmed through the *ROC* curve. Then, a terminal group is a low-risk group if its *PT* value is less than *L*; otherwise, it is a high-risk group. Consequently, we can confirm the heart failure high-risk group.

4. Experiments

In this section we first describe the method for establishing the experiment and then discuss the experimental results.

4.1. Experimental setup

As introduced in Section 2.1, the method is tested using physiological data and risk factor information collected from a real research project-HeartCarer, which includes 1026 heart failure patients, 7500 heart failure closely related cases, and 105,239 healthy people. Detailed presence of contributing cardiovascular risk factors was collected for each person and recorded in the data collection forms designed exclusively for the study. All data were entered prospectively in a computerized database. Analysis was done with the SPSS Statistical software. All categorical values were presented as mean ± standard deviation. A multivariate logistic regression analysis and odds ratio (OR) (95% confidence interval [CI]) was performed to identify the contribution of major cardiovascular risk factors in the progression of HF in the study population.

A large matrix is required to store information when constructing a medical social network. Here we use a OrientDB in a cluster to store the large-scale matrix graph, HBase to store the vertex properties, and Hadoop MR for data analysis and calculation. The scale of the constructed medical social network will grow larger and larger, as the number of people increases. This will lead to high load of CPU and memory demands, which will slow the processing speed. Therefore, we divide the data into several groups. The cluster includes 8 servers which are running the CentOS 7.4 OS, equipped with an 12-core(24-thread) Intel Xeon CPU running at 2.80 GHz and 64 GB of memory.

4.2. Experimental results

Four tests are included in our experiments. Test 1 shows the heart failure expression and detection; Test 2 discovers suitable RF_o to be used by our method. The newly discovered patients in the high-risk group are identified in Test 3, by using different threshold values (*L*) to compute the high-risk group. Test 4 proposes a Mean Average Precision-inspired evaluation method to measure the effectiveness of our method.

In the medical field, in order to evaluate a risk prediction method, researchers often use the terms of sensitivity, specificity, FNR (False-negative rate, equals 1-sensitivity), FPR (False-positive rate, equals 1-specificity), F-measure and ROC curve.

It is necessary to maximize sensitivity while minimizing FNR, when the specificity and sensitivity are used in a medical diagnosis. With the higher of the sensitivity, the probability that actual heart failure patients can be diagnosed will be higher. In order to ensure most healthy people are diagnosed correctly, the specificity needs to be as large as possible. However, specificity and sensitivity are conflicting. Increasing the specificity will reduce the sensitivity. Therefore, ensuring that all indicators are optimal is difficult. It is well known that the cost associated with the FPR is much smaller than the cost associated with the FNR. For this reason, it is more important to maximize the sensitivity of the method than maximize its specificity. However, for the heart failure risk prediction method, the target of getting the HF high-risk group will not be achieved if the value of specificity is too large (e.g., close to 1). Otherwise, if the value of specificity is too small, most people will be at high risk, which will make the method meaningless. It is suitable to obtain a medium specific value, and make the value of sensitivity as large as possible.

Based on sensitivity and specificity, another medical evaluation criterion-ROC curve can be used. The value of a method is assessed by the area under the ROC curve (AUC). The assessment value of the method will be better and better as the AUC increases. No equilibrium data has no effect on this indicator. Here, due to the lack of no equilibrium data, we use AUC, instead of the accuracy index. In addition, 113,765 people are divided into 50 groups for testing, each of which is about 2275 people, including both heart failure patients and healthy people.

4.2.1. The experiments of heart failure expression and detection

The distribution of *rf*s and *RF*s are shown in Table 2, which only includes risk factors with probability $p(rf_i|RF_j) \geq 0.1$, where $i \in [1, N]$, $j \in [1, M]$. The table shows that many *rf*s with different probabilities may belong to the same *RF*. Meanwhile, the same *rf* may belong to different *RF*s (e.g., rf_1). Therefore, only risk factors with larger probabilities are selected for probabilistic combination.

Fig. 6 shows the ratio relationships of different *RF*s associated with people p_i. As can be seen from the figure, different *RF*s have different influencing factors for people' heart failure expression and detection.

Table 2
Distribution of *RF*s and *rf*s.

RF	rf	Probability
RF_1	rf_1	0.4402
RF_1	rf_2	0.1682
RF_1	rf_3	0.1153
RF_2	rf_1	0.2346
RF_2	rf_4	0.3427
RF_2	rf_5	0.1208

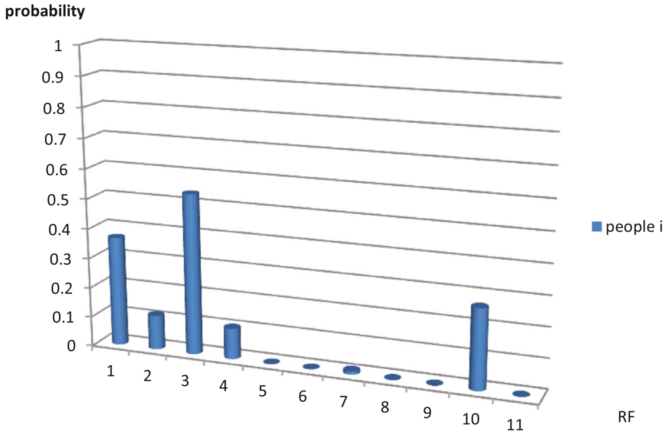


Fig. 6. Distribution of people and *RF*s.

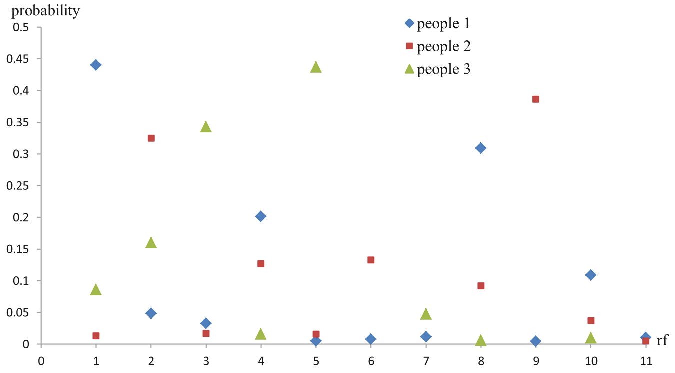


Fig. 7. Distribution of people and risk factors.

As shown in Fig. 5, for the expression of heart failure decomposition, the decomposition matrixes $p(rf|RF)$ and $p(RF|P)$ on the right hand of the equation can be obtained now. The matrix $p(rf|P)$ on the left hand of the equation will be analyzed in the next step. The distribution of people and risk factors is shown in Fig. 7. As shown in Section 3.1, the number of candidate risk factors is eleven in our research. The probability of the same risk factor in different people shows significant differences. Different people also have different probabilistic combinations of risk factors. For example, risk factor rf_1 can better represent the heart failure of people p_1, while rf_5 is more proper to p_3.

Different distributions of people p_i in different time are shown in Fig. 8. This figure shows that risk factor rf_4 can best represent the heart failure in time1 (e.g., October). In time2 (e.g., April) and time3 (e.g., December), the probability of rf_2, rf_6, rf_9 and rf_{10} increases, while the probability of rf_4 decreases.

These experiments show that: 1) many risk factors may belong to the same *RF*, and the same risk factor may be included in different *RF*s; 2) different *RF*s have different influencing factors on people' heart failure; 3) different people have different risk factor combinations; 4) different people have different heart failure performances in the same period; and 5) the same people has different heart failure expression in different time.

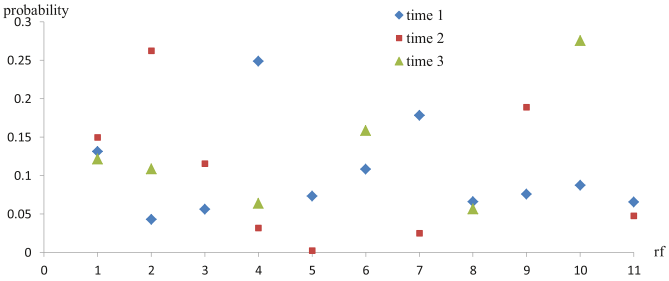


Fig. 8. Distribution of people's heart failure in different time.

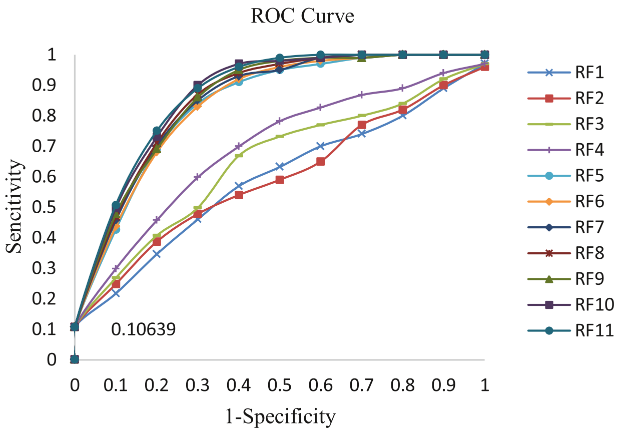


Fig. 9. ROC curves of different *RF*s.

4.2.2. *The influences of different* RF *for heart failure risk prediction method*

In this test the influences of different *RF* for heart failure risk prediction method will be determined, and then discover a suitable RF_o. Some *RF*s will be selected, and then a suitable RF_o will be discovered using our method. As time changes, the discovery of *RF* is always in progress. The sensitivity and 1-specificity will be calculated using different threshold values, and then the ROC curve is draw separately.

The method presented in this research includes three steps.

Step 1: As mentioned in Section 2.2, RF_i is the average of R_i^j $(1 \le j \le (i+1))$. The sensitivity and 1-specificity will be calculated using different threshold values. With different RF_i, all 50 test data groups are assessed using our heart failure risk prediction method. Finally, many terminal groups will be generated through the process of group division. We draw the ROC curve based on the calculated sensitivity and 1-specificity in these groups. The ROC curves are shown in Fig. 9. From the figure, we can see the ROC curves of RF_1, RF_2, RF_3 and RF_4 are much smaller. Compared to the above four ones, the assessment result of RF_5 is significantly improved. In addition, the ROC curves of RF_6, RF_7, RF_8, RF_9, RF_{10} and RF_{11} are closer to the ROC curve of RF_5. Further increased factors will not improve the assessment results. Because matrix calculations are very time consuming, RF_5 is chosen as the benchmark combination of risk factors.

Step 2: From Eq. (2), we know RF_5 is the average of R_5^j $(1 \le j \le 6)$. The ROC curves of R_5^1, R_5^2, R_5^3, R_5^4, R_5^5, and R_5^6 are shown in Fig. 10. Obviously, R_5^2 has the largest ROC curve, so R_5^2 is chosen as the best combination. That is to say, compared with other *rf*s, infection, diet, arrhythmia, personal disease history, and mental stress have higher impacts on heart failure. Therefore, these five factors are chosen to form RF_5.

Table 3
The selection of RF_i.

Symbol	Concrete related risk factors
RF_5	infection, diet, arrhythmia, personal disease history, mental stress
RF_6	infection, diet, arrhythmia, personal disease history, mental stress, long-term smoking and alcohol abuse
RF_7	infection, diet, arrhythmia, personal disease history, mental stress, long-term smoking and alcohol abuse, obesity
RF_8	infection, diet, arrhythmia, personal disease history, mental stress, long-term smoking and alcohol abuse, obesity, excessive physical exertion
RF_9	infection, diet, arrhythmia, personal disease history, mental stress, long-term smoking and alcohol abuse, obesity, excessive physical exertion, family history of heart failure
RF_{10}	infection, diet, arrhythmia, personal disease history, mental stress, long-term smoking and alcohol abuse, obesity, excessive physical exertion, family history of heart failure, motion
RF_{11}	infection, diet, arrhythmia, personal disease history, mental stress, long-term smoking and alcohol abuse, obesity, excessive physical exertion, family history of heart failure, motion, environment

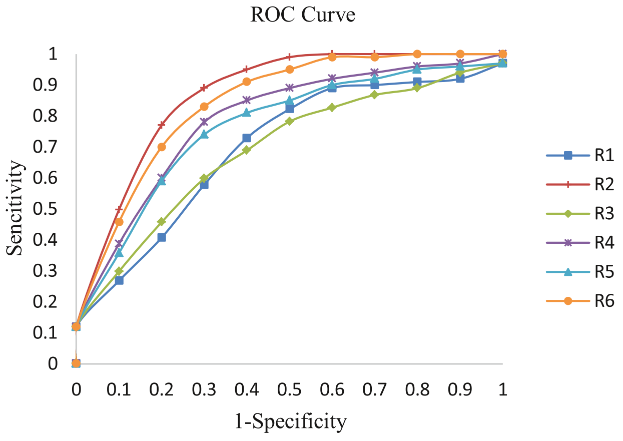


Fig. 10. ROC curves of different factors in RF_5.

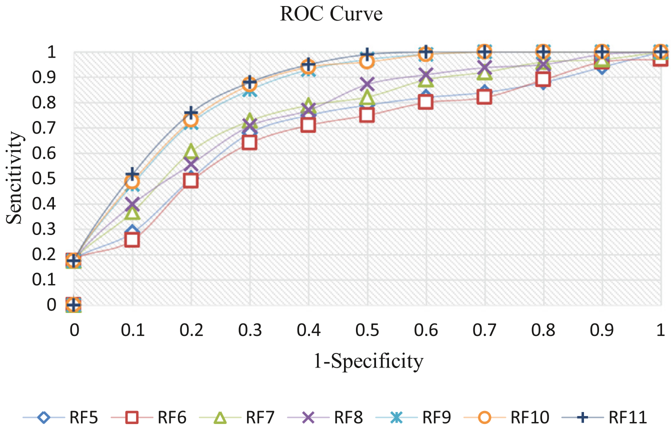


Fig. 11. The log value of ROC curves for a suitable RF_o.

Step 3: Based on the RF_5 in step 2, in order to make the method with the best assessment value, other *rfs* are added to get different choices of RF_i. As shown in Table 3, we add other *rfs* to RF_5 one by one, in order to discover suitable RF_o. In Fig. 11, we calculate the log value of ROC curves, which can more clearly see the difference in results.

From the figure, we can see that the assessment result of our heart failure risk prediction method is the worst when it is used with RF_5. However, when RF_9 is used, the assessment result is the best. It can be seen that the assessment result with RF_9 is significantly better than RF_5, which indicates that the assessment result can be improved by adding factors. However, the assessment results does not always improve when further factors are constantly added, e.g., the log value of ROC curves with RF_{10} and RF_{11} are closer to RF_9. From the log value of ROC curves, we can see that RF_9 is the best choice for heart failure risk prediction.

In the medical field, researches have agreed on some *RF*s. For example, people have a higher risk of heart failure if he is a first-degree relative with heart failure patients. Here, the logistic method is used to obtain *RF*s based on the same data to prove the credibility of the result. The specific process is as follows: 1) the single-factor conditional logistic regression analysis is used to analyze all study variables with heart failure one by one. We analyze all statistically significant factors when a=0.02; 2) based on the single factor analysis, variables are further screened using multi-factor conditional logistic regression analysis and using step wise regressive method; 3) then we discover nine *RF*s: infection, diet, arrhythmia, personal disease history, mental stress, long-term smoking and alcohol abuse, obesity, excessive physical exertion, family history of heart failure. Except miscarriages, the logistic result includes the factors of RF_9 in two results, which explains the credibility of RF_9. However, a corresponding risk assessment will be used to consider which is better of the two results.

4.2.3. The early-warning value of the heart failure risk prediction method

In order to identify the high-risk group, 13,970 healthy people are assessed using our heart failure risk prediction method in this test. Each test group includes 4000 people (260 heart failure patients and 3740 healthy people). Three test groups are involved in the test, and the average of the three test groups is the final result. Fig. 12 shows the ROC curves. The Youden index = sensitivity +specificity-1 in the ROC curve. The larger the Youden index, the better the assessment result of the method. As shown in Fig. 12, the best assessment value can be achieved when 1-specificity=0.25, sensitivity=0.775, and the Youden index is the largest. Since a certain threshold value *L* determines the values of sensitivity and 1-specificity, the Youden index is the largest when *L* is 0.068.

We confirm 28 people to be heart failure patients in the follow-up data, although they are shown as healthy in the initial data. If the accuracy of the initial data assessment is not considered, the high-risk group can be calculated using a different high-risk threshold value *L*. Then in the high-risk group, we can identify the newly discovered patients, as shown in Fig. 13.

Different high-risk groups will be calculated with different values of *L*. The high-risk group includes more newly discovered pa-

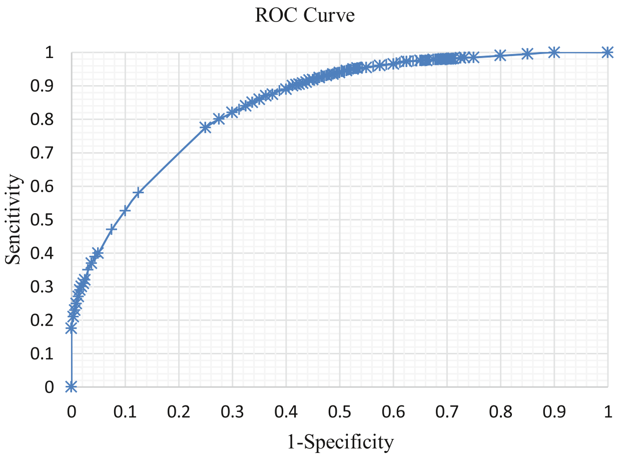


Fig. 12. ROC curve of our method based on 13,970 data points.

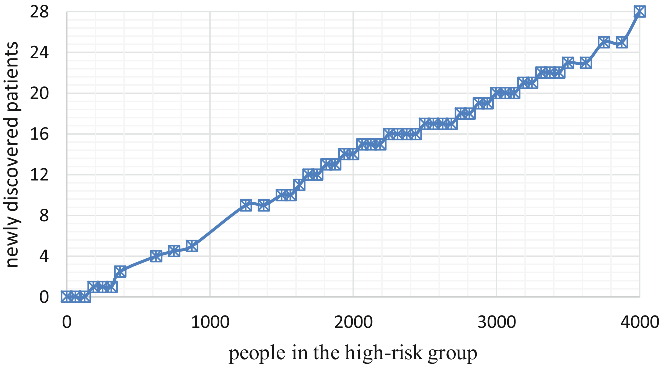


Fig. 13. Statistics of high-risk group and newly discovered patients.

tients when the value of L is smaller. For example, the high-risk group involves 1250 people when L=0.068, and in the follow-up data, 9 people are identified to be newly discovered patients. The method can provide early warning for 32.1% of the newly discovered patients. From this we can see that the value of early warning is not very high when L=0.068, although the assessment result is the best. From the perspective of early-warning, lower L value should be chosen to get larger high-risk group. The high-risk group involves 2562 people when L=0.02676, and in the follow-up data, 17 people are identified to be newly discovered patients. The high-risk group involves 3312 people when L=0.01297, and in the follow-up data, 22 people are identified to be newly discovered patients. That is to say, we can provide early warning for 87.9% of the newly discovered patients and sent them for screening. Meanwhile, the method can identify 28.16% of healthy people without a need for screening, which avoids the cost of screening. In the follow-up data, 25 people are identified to be newly discovered patients when L=0.0075, and the method can provide early warning for 89.28% of the newly discovered patients. The high-risk group involves 3875 people when L=0.005. In the follow-up data, 25 people are identified to be newly discovered patients, and the method can provide early warning for more patients. That is to say, we can provide early warning for 92.86% of the newly discovered patients and sent them for screening. Meanwhile, the method can identify 12.68% of healthy people without a need for screening, which avoids the cost of screening. Based on the above results, all healthy people and all 28 newly discovered patients in the follow-up data are involved in the high-risk group when L=0. If so, the heart failure risk prediction method has no effect, because all people need heart screening. If $L > 0$, the high-risk group will not include all 28 newly discovered patients in the follow-up data, and the low-risk group will involve some of them. In other words, the heart failure risk prediction method cannot provide early warning for 100% patients who will have heart failure in the future, because not all patients have significant risk factors. However, the method can provide early warning for a high percentage of newly discovered patients in the future, and reduce the screening range using the low threshold L.

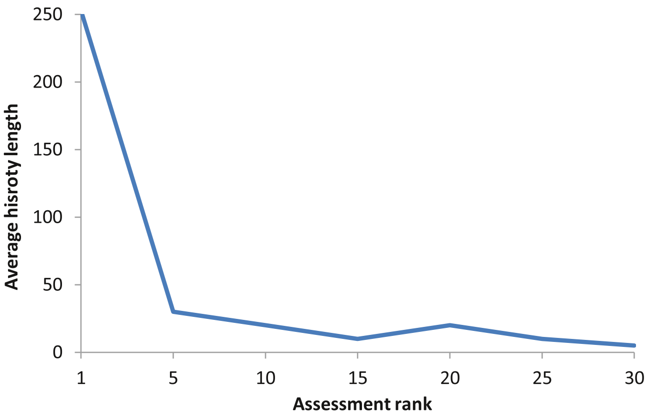


Fig. 14. History length vs. next admission assessment rank.

All experimental results demonstrate that the heart failure risk prediction method is an effective approach to prevent and control of heart failure.

4.2.4. The evaluation of the heart failure risk prediction method

We use the length of the partial history as input, and calculate the average rank of the proposed method's assessment. The dependency between them is shown in Fig. 14. In particular, we notice that short histories are associated with poor performance (i.e., higher ranks). In other words, the patient's future disease may be more uncertain when there is little information about a patient's medical history. These results show that it is more meaningful to use and represent the accumulating evidence. Allowing for the learning of complex interactions between conditions and their development strongly supports the effectiveness of our method.

By deleting a randomly chosen RF^* of each people p_j, another approach is proposed to obtain a set of validation data. In order to evaluate the method, the predicted risk factors are compared with the golden standard (e.g., RF^*). Our method predicts a list of risk factors ranked according to Eq. (11). A new evaluation method, which is inspired by the idea of Mean Average Precision evaluation method in Information Retrieval domain (Sanderson, 2005), is also proposed to measure the effectiveness of our method. Each risk factor is assigned a relevancy weight, and all the risk factors are ranked in the database as list $RF = \{rf'_{1,j}, \ldots rf'_{k,j}\}$. Our method is more likely to choose a risk factor with the higher relevancy weight. In order to evaluate our method, the following measurement is proposed.

Definition: The *CorrectRate* evaluates the accuracy of our method. It is the number of desired people divided by the total number of the people. The desired people are those whose risk factors matching one of the top M risk factors generated by our method.

$$CorrectRate_M = \frac{\sum_{j=1}^{H+F}\sum_{i=1}^{M} TOP_{j,i}}{(H+F)*R} \tag{11}$$

where

$$\forall TOP_{j,i} = \begin{cases} 1 : if RF^* matches\, a\, risk\ factor\, in \left\{rf'_{1,j}, \ldots rf'_{M,j}\right\} \\ 0 : otherwise \end{cases} \tag{12}$$

Table 4
An example of CorrectRate.

	Demanded factor RF^*	Ranked recommendation factors list	M=1	M=3	M=5
p_1	rf_9, rf_7, rf_3	$rf_9 > rf_7 > rf_{11} > rf_3 > rf_6 > \ldots$	$TOP_{1,1} = 1$	$TOP_{1,3} = 2$	$TOP_{1,5} = 3$
p_2	rf_5, rf_1, rf_3	$rf_{11} > rf_8 > rf_5 > rf_1 > rf_4 > \ldots$	$TOP_{2,1} = 0$	$TOP_{2,3} = 1$	$TOP_{2,5} = 2$
p_3	rf_2, rf_6, rf_1	$rf_7 > rf_9 > rf_3 > rf_5 > rf_1 > \ldots$	$TOP_{3,1} = 0$	$TOP_{3,3} = 0$	$TOP_{3,5} = 1$
All	–	–	$CorrectRate_1 = 0.11$	$CorrectRate_3 = 0.34$	$CorrectRate_5 = 0.67$

Table 5
The performance of our method.

Training Data Percentage	$Correct - Rate_1$	$Correct - Rate_3$	$Correct - Rate_5$	$Correct - Rate_9$	$Correct - Rate_{11}$
60%	0.7180	0.7641	0.7953	0.8391	0.8394
50%	0.7069	0.7532	0.7832	0.8027	0.8046
40%	0.6943	0.7346	0.7803	0.7986	0.7992

The number of all people is denoted as $H + F$. The number of top risk factors with comparison to the set of golden standard risk factors RF^* is denoted as M. R is the number of golden standard risk factors of each person.

Formula (12) gives a rate of correctly predicating RF^* in the top M risk factors of the ranked list RF. When $M >= 1$, we regard the assessment correct as long as RF^* is within the top M predicted laboratory risk factors. Since each person has more than one desired risk factor, other measurements can be deducted directly from $CorrectRate_M$. We adopt $M = 1$, $M = 3$ and $M = 5$, which means the proportion of people that the golden standard risk factors rank in the top 1, the top 3 or the top 5, namely $CorrectRate_1$, $CorrectRate_3$ and $CorrectRate_5$ respectively.

To show how the proposed measurement works, we give an example in Table 4. Assume there are totally 3 people and 11 laboratory risk factors. Using our method, each person has a ranked list, which are compared with his/her demanded risk factor. We use an symbol $'' < ''$ to indicate that the relevancy weight of the left part is smaller than that of the right part. Table 4 shows that RF^* has different ranks for different people. $patient_1$ ranks in the 1st, 2nd and 4th, $patient_2$ ranks in the 3rd and 4th, $patient_3$ ranks only in the 5th within top 5. $TOP_{j,i}$ is shown in the table for given M. The $CorrectRate_1$, $CorrectRate_3$ and $CorrectRate_5$ have a value of 0.11, 0.34 and 0.67 accordingly. $CorrectRate_1$ and $CorrectRate_3$ are always smaller or equal to $CorrectRate_5$, because the people with RF^* ranked in the top 5 will include the people with RF^* ranked in the top 1 and 3.

Table 5 shows the overall performance of our method. By resampling with different training validation proportions, the experiments are conducted on different training sets and validation sets. The percentage of training data we used are 40%, 50%, and 60% respectively. Using 60% data for training has the highest $CorrectRates$: are 0.7180, 0.7641, 0.7953, 0.8391 and 0.8394. Using 50% data for training has the median $CorrectRates$: 0.7069, 0.7532, 0.7832, 0.8027 and 0.8046. The $CorrectRates$ is the lowest when using 40% data for training, which are 0.6943, 0.7346, 0.7803, 0.7986 and 0.7992. $CorrectRate_{11}$ is larger than $CorrectRate_1$, $CorrectRate_3$, $CorrectRate_5$ and $CorrectRate_9$, but it is similar to $CorrectRate_9$. According to Section 5.2.2, we choose $CorrectRate_9$ as the suitable one. The results show that accurate recommendations of laboratory risk factors could be presented.

Furthermore, we compare the efficiency of our method with existing traditional methods such as SVM and ANN. From Fig. 15, we can see, when the training data percentage is almost 0, the efficiency of all methods is very low. Meanwhile, with the increase of training data percentage, the efficiency of three methods all increase. When the training data percentage is smaller than 50%, the efficiency increase fast; however, when it is more than 50%, the efficiency grows slowly. When the training data percentage is smaller than 50%, the efficiency of SVM is larger than ANN; otherwise, it is smaller. That is because SVM is suitable for small sample learning. When it comes to the law of large numbers, the computational and storage performance requirements of SVM are too high. However, the efficiency of our method is always the best, which is about 90% with the increase of training data percentage.

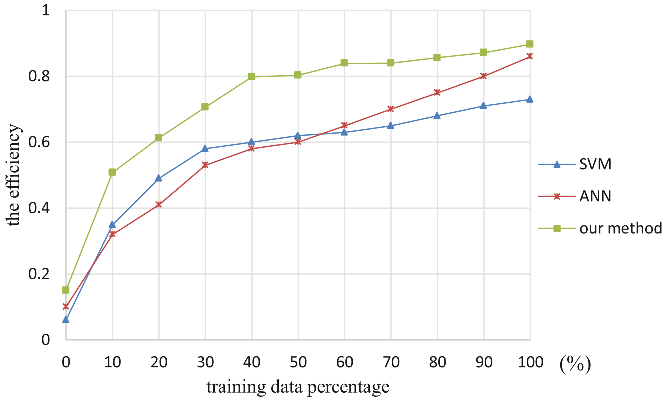


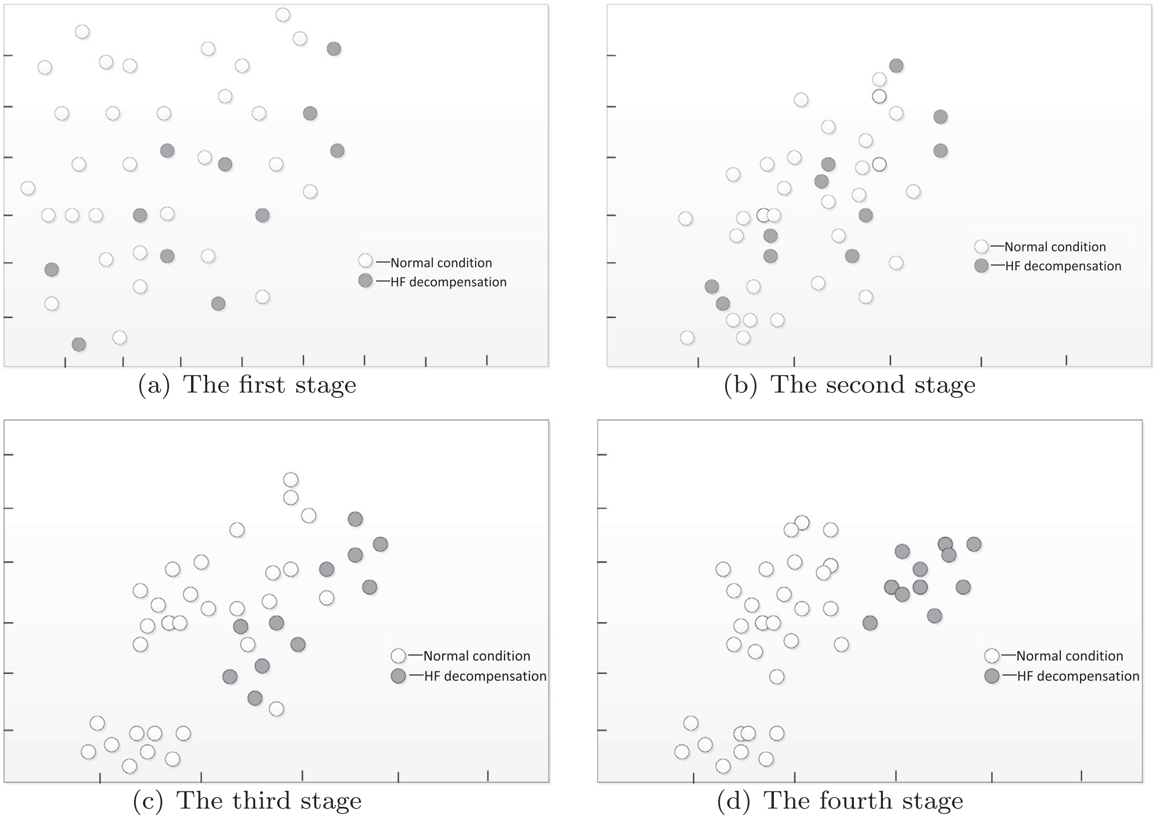
Fig. 15. The comparison of existing methods.

Finally, the decision of the assessment process is illustrated in Figs. 16 and 17. These numbers use the nearest neighbor method to describe the classification of chosen 41 cases. As described above, this result supports similarity metrics calculated between the most similar examples determined in each case (template) and historical data sets. In this particular example, the number of modes used in the assessment is $M = 9$, which is based on the analysis of Section 5.2.2.

5. Discussion

We propose a social medical network-based method for predicting the risk of heart failure. A heart failure high-risk group in the network is constructed through assessing whether a person is at high risk of developing heart failure. Further screening of people in the high-risk group can help to achieve early detection, early diagnosis, and early treatment, which can eventually improve the survival rate of heart failure and reduce the mortality. It will also have the potential to reduce the cost of medical visits, and effectively relieve medical resources.

Four tests are included in our experiments. Test 1 shows the heart failure expression and detection. These experiments show that: 1) many risk factors may belong to the same RF, and the same risk factor may be included in different RFs; 2) different RFs have different influencing factors on heart failure; 3) different people have different risk factor combinations; 4) different people



(a) The first stage

(b) The second stage

(c) The third stage

(d) The fourth stage

Fig. 16. The assessment process.

have different heart failure performances in the same period; and 5) same people have different heart failure expressions at different times.

Test 2 tested the influences of different *RF* for heart failure risk prediction method, and discovered a suitable RF_0. All 50 test data groups are assessed. As proved in the medical field, the larger the area under the ROC curve, the better the assessment value of the method. Here, the logistic method is used to obtain *RFs* based on the same data to prove the credibility of the result. The specific process is as follows: 1) the single-factor conditional logistic regression analysis is used to analyze all study variables with heart failure one by one. We analyze all statistically significant factors when $\alpha = 0.02$; 2) based on the single factor analysis, variables are further screened using multi-factor conditional logistic regression analysis and stepwise regressive method; 3) we discover nine *RFs*: infection, diet, arrhythmia, personal disease history, mental stress, long-term smoking and alcohol abuse, obesity, excessive physical exertion, family history of heart failure. Except miscarriages, the logistic result includes the factors of RF_9 in two results, which proves the credibility of RF_9. However, a corresponding risk assessment will be used to consider which is better of the two results. In future work, we will explore the risk factors in higher detail, including the best choice of risk factor (or a combination), the time influence, and the validation from a clinical perspective.

By using different threshold values (L) to compute the high-risk group, the newly discovered patients in the high-risk group are identified in Test 3. The high-risk group includes more newly discovered heart failure patients when the value of L is smaller. However, when $L = 0$, the heart failure risk prediction method has no effect, because all people need heart screening. So we tried to find the threshold value of L, in which the method can provide early warning for a high percentage of newly discovered patients in the future, and reduce the screening range. However, the heart failure risk prediction method cannot provide early warning for 100% patients who will have heart failure in the future, because not all patients have significant risk factors.

Test 4 proposes a Mean Average Precision-inspired evaluation method to measure the effectiveness of our method. Since each person has more than one demanded risk factor, other measurements can be deducted directly from $CorrectRate_M$. We adopt $M = 1$, $M = 3$, $M = 5$, and $M = 9$, which means the proportion of people that the golden standard risk factors rank in the top 1, the top 3, the top 5 or the top 9. By re-sampling with different training validation proportions, the experiments are conducted on different training sets and validation sets. From the experimental results of Tables 4 and 5, we choose $CorrectRate_9$ as the suitable one. The best accuracy can reach almost 90%. Furthermore, we compare the efficiency of our method with the existing traditional methods-SVM and ANN in Fig. 15. The efficiency of our method is always the best, which is about 90% with the increase of training data percentage.

In short, the method proposed in this paper has several advantages: 1) The data set contains records of different types of people, thus the experimental results are very representative. 2) The unique design for selecting risk factors related to heart failure makes the method generalizable for other countries and regions. 3) It can provide early warning to a high proportion of patients who are discovered to have heart failure in the future, which is meaningful for the prevention and control of heart failure.

However, there are also some limitations of our method: 1) we only prove which combination of risk factors is better suited to diagnose heart failure based on the tested data set, but not considering the time influence, and the validation from a clinical perspective. 2) our heart failure risk prediction method cannot provide early warning for 100% patients who will have heart failure in the future, because not all patients have significant risk factors.

Algorithm 1 Heart Failure Risk Prediction Method.

```
Input:
    The number of HF patients, $F$;
    The number of healthy people, $H$;
    The total number of $RF$s, $M$;
    $rf_m \in RF_M$;
Output:
    GF, GH;
for $j = 1;\ j < H + F;\ j++$ do
    for $i = j + 1;\ i <= H + F;\ i++$ do
        Compute $S_{rf_m}(j, i)$;
    end for
end for
    Network G is constructed based on $S_{rf_m}$
    Based on the rules to initialize the group
    Select two nodes $h_i$ and $f_j$, $h_i \in GH$, $f_j \in GF$
for $k = 1;\ k <= H + F;\ k++$ do
    if $k! = i$ and $k! = j$ then
        k is involved in GH;
        calculate $\Delta Z_{GH}$;
        k is involved in GF;
        compute $\Delta Z_{GF}$;
        if $\Delta Z_{GH} > \Delta Z_{GF}$ then
            k joins GH;
        else
            k joins GF;
        end if
    end if
end for
```

6. Conclusions

The existing research works mainly consider the analysis of ECG. However, when an ECG abnormality is detected, the person has already had heart failure. In fact, the long-term existence of certain risk factors has long increased the likelihood of heart failure of the persons. If we can analyze and track these risk factors in advance, and predict high-risk groups and low-risk groups, we can effectively prevent heart failure by targeted drug intervention or lifestyle changes. In order to better carry out heart failure assessment, this paper considers risk factors that can easily cause heart failure, and proposes a social medical network-based method for predicting the risk of heart failure. Through early detection, early diagnosis, and early treatment, our goal is to improve the survival rate of heart failure and reduce the mortality. The experiments using physiological data and risk factor information collected from a real research project-HeartCarer, show the efficiency of our method.

The future works which may be researched in higher detail are listed as follows: 1) Here, we use a simple method to calculate the similarity among people. Based on Eq. (4), we first compute the similarity between two people for a particular risk factor rf_i, and obtain an unweighted medical social network. Then based on Eq. (3), we compute the similarity between two people for several risk factors, and obtain a weighted medical social network. However, comparing with this method, other methods may be researched in the future work. 2) The risk factors may be discussed in higher detail in the future work, including which risk factor (or a combination) is better suited to diagnose heart failure, how do they change over time, and how to validate these findings from a clinical perspective. 3) Different threshold value L has different influence on the calculation of high-risk group. The high-risk group includes more newly discovered heart failure patients when the value of L is smaller. However, when L=0, the heart failure risk prediction method has no effect, because all people need heart screening. So we tried to find the threshold value of L, in which the method can provide early warning for a high percentage of newly discovered patients in the future, and reduce the screening range. In the future work, the computation of L threshold and posterior choice of values may be analysed in higher detail. 4) Up to now, there are no very relevant researches which aim directly at predicting the risk of heart failure based on the analysis of cardiovascular risk factors. The purpose of this manuscript is to provide early warning to patients with suspected heart failure, not necessarily diagnosed heart failure patients. However, in order to prove the efficiency of our method, we compare with the traditional methods-SVM and ANN, as shown in Fig. 15. By comparing, we can see the efficiency of our method is always the best, which is about 90% with the increase of training data percentage. In addition, we will do some further comparison in our future work, such as space cost, time cost, parameter influence, and so on.

Declaration of Competing Interest

There is no conflict of interest form in this paper.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.eswa.2020.113361.

Credit authorship contribution statement

Chunjie Zhou: Conceptualization, Methodology, Validation, Formal analysis, Writing - original draft, Funding acquisition. **Ali Li:** Methodology, Validation, Visualization. **Aihua Hou:** Investigation, Resources, Data curation, Formal analysis. **Zhiwang Zhang:** Validation, Funding acquisition. **Zhenxing Zhang:** Resources, Project administration. **Pengfei Dai:** Software, Supervision, Project administration. **Fusheng Wang:** Writing - review & editing.

References

Abdelaziz, A., Elhoseny, M., Salama, A., et al. (2018). A machine learning model for improving healthcare services on cloud computing environment. *Measurement, 119*, 117–128.

Bekhet, L. R., Yonghui, W., Ningtao, W., et al. (2018). A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set. *Journal of Biomedical Informatics, 84*, 11–16.

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 993–1022.

Bohacik, J., Matiasko, K., Benedikovic, M., et al. (2015). Algorithmic model for risk assessment of heart failure patients. In *IEEE international conference on intelligent data acquisition and advanced computing systems: Technology and applications* (pp. 177–181). IEEE.

Chang, C., Wang, C., & Jiang, B. (2011). Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors. *Expert Systems with Applications, 38*(5), 5507–5513.

Chen, M., Ma, Y., Li, Y., et al. (2017). Wearable 2.0: Enabling human-cloud integration in next generation healthcare systems. *IEEE Communications Magazine, 55*(1), 54–61.

Delanaye, P., Guerber, F., Scheen, A., et al. (2017). Discrepancies between the cockcroft-gault and chronic kidney disease epidemiology (CKD-EPI) equations: Implications for refining drug dosage adjustment strategies. *Clinical Pharmacokinetics, 56*(2), 193–205.

Fabian, I., Paul, F., Peter, M., et al. (2017). Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features. *Statistical Atlases and Computational Models of the Heart*, 120–129.

Fu, P., Zeng, X., Liu, J., et al. (2018). Big data research in chronic kidney disease. *Chinese Medical Journal, 131*(22), 2647–2650.

Hao, Y., Usama, M., Yang, J., et al. (2019). Recurrent convolutional neural network based multimodal disease risk prediction. *Future Generation Computer Systems, 92*, 76–83.

Hiragi, S., Tamura, H., Goto, R., et al. (2018). The effect of model selection on cost–effectiveness research: A comparison of kidney function-based microsimulation and disease grade-based microsimulation in chronic kidney disease modeling. *BMC Medical Informatics and Decision Making, 18*(1), 1–11.

Kalid, N., Zaidan, A., Zaidan, B., et al. (2018). Based real time remote health monitoring systems: Areview on patients prioritization and related "big data" using body sensors information and communication technology. *Journal of Medical Systems, 42*(2), 30.

Khalifa, A., & Meystre, S. (2015). Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. *Journal of Biomedical Informatics, 58*, S128–S132.

Leonardo, A., Jose, D., Ruben, O., et al. (2018). Classification of neuron sets from non-disease states using time series obtained through nonlinear analysis of the 3d dendritic structures. *International Journal of Engineering Research and Technology, 2*(1), 17–25.

Li, A., Wang, R., Liu, L., et al. (2019). BCRAM: A social-network-inspired breast cancer risk assessment model. *IEEE Transactions on Industrial Informatics, 15*(1), 366–376.

Lin, K., Xia, F., Li, C., et al. (2019). Emotion aware system design for the battlefield environment. *Information Fusion, 47*, 102–110.

Maddalena, C., Salvini, R., Bardoni, A., et al. (2019). Searching for biomarkers of chronic obstructive pulmonary disease using proteomics: The current state. *Electrophoresis, 40*(1), 151–164.

Meystre, S. M., Kim, Y., Gobbel, G. T., et al. (2017). Congestive heart failure information extraction framework for automated treatment performance measures assessment. *Journal of the American Medical Informatics Association, 24*(e1), e40–e46.

Miao, F., Cai, Y., Zhang, Y., et al. (2018). Predictive modeling of hospital mortality for patients with heart failure by using an improved random survival forest. *IEEE Access, 6*, 7244–7253.

Ralph, B. (2018). Illness-death model in chronic disease epidemiology: characteristics of a related, differential equation and an inverse problem. *Computational and Mathematical Methods in Medicine, 5091096*, 1–6.

Rani, N. V., Kannan, G., Vasantha, J., et al. (2012). Risk assessment for congestive heart failure in a south indian population: A clinical pharmacist perspective. *Indian Journal of Clinical Practice, 22*(9), 431–436.

Ravizza, S., Huschto, T., Adamov, A., et al. (2019). Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. *Nature Medicine, 25*, 57–59.

Sanderson, M. (2005). Information retrieval system evaluation: Effort, sensitivity, and reliability. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 162–169).

Sengul, D., & Ibrahim, T. (2018). Diagnosing hyperlipidemia using association rules. *Mathematical and Computational Applications, 13*(3), 193–202.

Vikram, V., Esther, E., Wiro, J., et al. (2019). Disease progression timeline estimation for alzheimer's disease using discriminative event based modeling. *NeuroImage, 186*, 518–532.

Zhou, C., Wang, X., Zhang, Z., et al. (2019). The time model for event processing in internet of things. *Frontiers of Computer Science, 13*(3), 471–488.

Zhou, J., Cao, Z., Dong, X., et al. (2013). Securing healthcare social networks: Challenges, countermeasures and future directions. *IEEE Wireless Communation, 20*(4), 12–21.