Future Generation Computer Systems 102 (2020) 534–548

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

A novel data-driven robust framework based on machine learning and knowledge graph for disease classification

Zhenfeng Lei [a], Yuan Sun [b], Y.A. Nanehkaran [a], Shuangyuan Yang [a,*], Md. Saiful Islam [b], Huiqing Lei [c], Defu Zhang [a,*]

[a] School of Informatics, Xiamen University, Xiamen 361005, China
[b] School of Electrical & Electronic Engineering, The University of Adelaide, Adelaide 5005, Australia
[c] Department of Breast Surgery, The First Affiliated Hospital of Zhengzhou University, Zhengzhou 450000, China

ARTICLE INFO

Article history:
Received 23 March 2019
Received in revised form 27 June 2019
Accepted 27 August 2019
Available online 3 September 2019

Keywords:
Disease classification
NCDs
Data fusion
Machine learning
Knowledge graph

ABSTRACT

As Noncommunicable Diseases (NCDs) are affected or controlled by diverse factors such as age, regionalism, timeliness or seasonality, they are always challenging to be treated accurately, which has impacted on daily life and work of patients. Unfortunately, although a number of researchers have already made some achievements (including clinical or even computer-based) on certain diseases, current situation is eager to be improved via computer technologies such as data mining and Deep Learning. In addition, the progress of NCD research has been hampered by privacy of health and medical data. In this paper, a hierarchical idea has been proposed to study the effects of various factors on diseases, and a data-driven framework named d-DC with good extensibility is presented. d-DC is able to classify the disease according to the occupation on the premise where the disease is occurring in a certain region. During collecting data, we used a combination of personal or family medical records and traditional methods to build a data acquisition model. Not only can it realize automatic collection and replenishment of data, but it can also effectively tackle the cold start problem of the model with relatively few data effectively. The diversity of information gathering includes structured data and unstructured data (such as plain texts, images or videos), which contributes to improve the classification accuracy and new knowledge acquisition. Apart from adopting machine learning methods, d-DC has employed knowledge graph (KG) to classify diseases for the first time. The vectorization of medical texts by using knowledge embedding is a novel consideration in the classification of diseases. When results are singular, the medical expert system was proposed to address inconsistencies through knowledge bases or online experts. The results of d-DC are displayed by using a combination of KG and traditional methods, which intuitively provides a reasonable interpretation to the results (highly descriptive). Experiments show that d-DC achieved the improved accuracy than the other previous methods. Especially, a fusion method called RKRE based on both ResNet and the expert system attained an average correct proportion of 86.95%, which is a good feasibility study in the field of disease classification.

1. Introduction

With the acceleration of social modernization, the pace of life is also speeding up. Meanwhile, many people suffer from chronic diseases such as diabetes, hypertension and other non-communicable diseases (NCDs), which often cannot be completely cured and affects daily life and work of patients seriously [1,2]. According to the report [3], the total deaths of NCDs were around 40 million, which is estimated to account for 70% of all deaths. *Report on Cardiovascular Diseases in China (2017)* released by National Center for Cardiovascular Diseases showed that the death of cardiovascular diseases took up more than 40%, which is much higher than tumors and other diseases. Besides, the death of cardiovascular diseases in Chinese rural areas has increased significantly in recent years, and its mortality has continued to be higher than that in urban areas. Fortunately, with a large number of studies proving that chronic diseases can be alleviated by diet, exercise, etc., people are gradually shifting their focus to prevention [4]. A number of facts have also been proved that the prevention is the best solution for those suffering chronic diseases nowadays. For example, a well-known cardiovascular disease expert, Hong, said that human health and longevity depend on heredity of 15%, social conditions of 10%,

* Corresponding authors.
E-mail addresses: yangshuangyuan@xmu.edu.cn (S. Yang), dfzhang@xmu.edu.cn (D. Zhang).

https://doi.org/10.1016/j.future.2019.08.030
0167-739X/

medical conditions of 8% and the natural environment of 7%. Most importantly, they mainly lie on people's lifestyle accounting for 60%, which has reached a consensus in the medical community (map 1 or map 2 of cardiovascular deaths). Although environmental factors cannot be controlled, if about 60% of the living habits are properly adjusted, there will be no major problems with disease prevention. For example, by increasing awareness to the people via Japanese government-led reduced intake of salt, which decreased the morbidity of cerebral apoplexy by more than 70% [5]. In America, the behavioral intervention reduced the prevalence rate of coronary heart disease by 35% and the morbidity of cerebrovascular disease by 48% [6]. However, since their occurrence is often affected by the uncontrollable natural environment and social conditions, there are still the non-negligible deaths and the morbidity in every region of the world. Therefore, the hierarchical research such as according to regional distribution, age, occupation, etc., may provide more clues to analyze the etiology and epidemic factors of the disease for formulating prevention measures [7]. In recent years, it has increasingly become a research hotspot [8].

Prevention of diseases requires a multi-pronged approach due to the diversity of disease influencing factors. A certain disease can occur in different populations and various diseases may be found in a certain region. Disease outbreaks are mainly concentrated in age, gender, timeliness, regionalism, occupation, etc. For instance, *Age*: a study in Nigeria showed that 17% of 14-year-old pregnant women catches pregnancy-induced hypertension, compared with only 3% of pregnant women aged 20–34. In developing countries, cardiovascular and cerebrovascular diseases are the leading cause of deaths for women over the age of 45 [9]. *Timeliness or Seasonality*: both infectious and non-infectious diseases change over time. For example, the three months, August, September and October, are the peak season of epidemic encephalitis B in northern China. The disease causes suffering from epidemic encephalitis B accounted for more than 40% of the total disease causes in some Chinese regions in those three months [10]. A number of studies have shown significant differences in acute mountain sickness (AMS) mortality in various seasons. *Regionalism*: people from various regions suffer from different diseases, which is not only related to eating habits and genes, but also affects by the regional climate. Some diseases with high morbidity in specific areas are often directly related to the local geographical situation. For example, breast cancer is the most common disease in North America, Northern and Western Europe, but less in Asia and Africa [11]. The E1+Tor cholera used to occur only in the Sulawesi region of Indonesia, has invaded Africa and Europe without cholera for more than 20 years since 1970 [12]. According to statistics, about 40% of nasopharyngeal carcinoma (NPC) happened in southern China, especially in Guangdong province. Therefore, NPC is also known as "*Guangdong cancer*". In addition, some well-known diseases like Kashin-Beck disease, endemic goiter and endemic fluorosis are caused by the lack or excessive existence of trace elements in the local geographical environment, which highlight the feature of regionalism [13].

In addition to the above factors, there is also an important factor: *Occupation*, i.e., people with various occupations may suffer from different diseases and certain rules can be discovered between them. According to the report provided by [14], the suicide rate of dentists is twice higher than people in other professions. In 2002, the BBC News reported that mercury concentrations of dentists were four times workers in other industries by comparing urine, hair and nail occupations [15]. Studying workers who have been exposed to beryllium smog for a long time revealed that the morbidity of lung cancer has increased significantly, which was obtained by the Wuhan CDC in China for four months [16]. In fact, experts from the World Health Organization's International Agency for Research on Cancer have also discovered the phenomenon that beryllium can cause the lung cancer.

The above evidence only indicates that the factors are more intuitively accepted by everyone, but there may be other factors including those that have not yet been discovered and cannot be explained by science. Since the outbreak of the disease is affected by many factors, their prevention has also increased difficulties [17]. As to the above research on multi-factor problems, some results have been made in certain studies. However, their achievements, including medical and computer-based results, are currently difficult to learn and reuse through Deep Learning (DL) and data mining methods. Because they only deliver the experimental results without specific intermediate details. For example, the international classification of diseases [18] only provided the results of disease classification, and it was difficult for researchers to take advantage of them to conduct multi-level learning to perform effective prevention of diseases and to explain how to effectively prevent the outbreak of diseases. In this paper, there are main contributions:

- We classified diseases based on hierarchical thoughts by adopting the effects of various factors on the disease, and a data-driven robust framework with good extensibility, called d-DC, is proposed;
- We designed a data collection model for structured data, unstructured data such as plain texts, images or videos, which can achieve automatic replenishment of data and is able to effectively tackle the cold start problem of the prediction model without much data;
- We first combined knowledge graph (KG) with Deep Learning to classify diseases. Moreover, the vectorization of medical texts by using knowledge embedding (KE) showed a new consideration in the field of disease classification;
- When the results are singular, the expert system built through knowledge bases or on-line experts is proposed to address inconsistencies, and their final results are presented by the knowledge graph technology;
- The experimental results show that our proposed framework can achieve higher accuracy over previous methods, especially the fusion method RKRE achieved an average correct proportion of 86.95%.

The rest of the paper is organized as follows: In Section 2, we provide an introduction to related work, mainly including diseases, multi-factor research and their relationship, and techniques used in this paper such as DL, KG and so on. In Section 3, the proposed framework and the proof of important basis used by d-DC have been presented. Next, the methods, involving data filling methods, similarity measures, feature extraction, attribute reduction methods, and classification methods are illustrated in detail in Section 4. Section 5 mainly introduces the experimental content, containing the dataset, the verification method and evaluation metrics whereas Section 6 shows detailed experimental results and analyses. Finally, the summary of the research and future work prospects are presented in Section 7.

2. Related work

It has been reported in recent works that the risk of chronic diseases can be reduced or even prevented by regulating nutrients, exercise, etc. [19]. The nature and intensity of work are closely related to the occurrence of chronic diseases. In general, the morbidity of physical workers is lower than that of brain-workers. The morbidity of chronic diseases in individual businesses, service personnel and cadres in China is snowballing and has attracted increasing attention. Early detection of the

disease symptoms and medical drug service status of different occupations is conducive to further deepening the reform of the medical and health system, and to making certain contributions to the establishment of the medical security system. Furthermore, it also can provide a reference for the decision-making of the government's medical and health departments, which is of great significance for preventing NCDs [20].

It is a more effective way to discover the connection between occupations and diseases by mining people's lifestyles. Some research indicated that western-style diets are rich in fats and monosaccharides but lack specific micronutrients, which often increases the prevalence of autoimmune diseases or immune-related diseases such as allergies, atopic dermatitis and the obesity, etc. [21]. A study surveyed health, diet and lifestyle for more than 1000 Chinese for five years, and found that the proportion of people with more than one type of chronic disease using western-diet has increased from 14% to 34%. According to another study, it is observed that higher fruit intake helps prevent the onset of a kind of chronic disease, while higher vegetable intake benefits prevent more than one chronic disease [22]. Patients with complex regional pain syndrome (CRPS) will experience abnormalities in visual signals on one side of the injured limb. This finding may help develop a new treatment for patients with CRPS [23,24]. Doctors have accumulated data about the association between low levels of vitamin D and high-risk of chronic disease [25]. It is found that low levels of vitamin D in the body are directly related to the increased risk of chronic headaches [26]. Many related-disease classification conclusions have been drawn through clinical medicine, and the usage of computers has also yielded significant results in this field. According to the *Report on cardiovascular disease in China (2015)*, the number of cardiovascular patients in China has reached 290 million, including 13 million strokes, 11 million coronary heart disease, and 270 million hypertension. In general, the morbidity and mortality of cardiovascular disease in China are still on the rise due to the aging of society, the acceleration of urbanization and the prevalence of unhealthy lifestyles [27,28].

In Deep Learning, the representation of samples is generally learned in an unsupervised way in unlabeled data, and it is to be refined-tuned through supervised learning according to the final task [29,30]. Representation learning based on Deep Learning has received extensive attention in the study of large-scale knowledge graph and has made significant progress. Knowledge embedding in KG aims to map all entities and relationships in a graph into a low-dimensional, continuous and real-valued vector space [31]. By projecting the entity or relationship into a low-dimensional vector space, the representation of their semantic information can be realized and their complex semantic associations can be efficiently calculated [32]. The embeddings learned by knowledge representation can be applied to the following applications: similarity calculation, relationship extraction, automatic question and answer, entity link, link prediction, etc. Currently, the knowledge embedding methods mainly include the following models: structured embedding (SE) [33], latent factor model (LFM), matrix decomposition model, translation model and neural network model [34]. Since the latter two are more flexible, suitable for a large amount of data and even can provide a weak link between entities and relationships, they are highly respected nowadays. For example, the TransE model [35] works well on large-scale knowledge graphs, which is considered as the most classic one. TransH model [36] proposed to represent the relationship r as the normal vector wr and the translation vector dr on a hyperplane. TransD [37] set two projection matrices that respectively project the head and tail entities onto the relational space to solve the problem of redundant model parameters. The integration of KG and DL has become one of the most important ideas to enhance the effect of models further. There are two main ways of knowledge graph based on Deep Learning. Firstly, the semantic information in KG is input into the DL model; the discretized knowledge graph is expressed as a continuous vector so that the prior knowledge in KG can be entered into Deep Learning. The second is to use the knowledge as the constraint of optimization objectives to guide the learning of models. The knowledge in KG is the posterior regular term of the optimization goal [38]. The method adopted in this paper belongs to the second one, it used the knowledge based on occupations and diseases as the optimization criterion to perform our classification tasks through classical traditional models, knowledge graph methods and Deep Learning.

3. Proposed framework

This section mainly introduces our proposed framework. Considering that diseases vary from region to region, whether they are classified by occupation or other attributes, it is necessary first to determine the region before classification [39]. Since this paper takes advantage of the important basis or conclusion that morbidity or mortality varies by region, it is important to briefly prove its existence before introducing our proposed framework in this section.

3.1. Proof of data for the proposed method

It is more scientific to conduct the more specific classification of diseases after consideration of regionality because an occupation will have different types or symptoms in various countries or regions. In order to elaborate the geographical differences of the diseases, we give a distribution of the total deaths of 195 countries in 2016,[1] called WordMap (left side in Fig. 1), and a distribution of the total deaths of 34 provinces in China,[2] called ChinaMap (right side in Fig. 1). The brightness difference of the color in the map indicates the death severity, that is, the darker the color, the higher the death or mortality in the area. Conversely, the brighter the color, the lower the death or mortality of that area. It is worth noting that, although the total number of people in each country is different, resulting in a different number of deaths, here we only emphasize the difference in disease severity among regions, not the specific figures. Therefore, we can use those numbers to reflect the problem we want to highlight. We can see from WordMap that there are so many differences in deaths. For example, the death toll is very few in Chile, Argentina and Australia, while China, India and Canada have reached a large scale in the death toll. We can also find the similar conclusion in ChinaMap, that is, the death varies widely among provinces. For example, the death toll in northwestern China is much larger than that of coastal cities such as Fujian province and Shandong province, which may be related to the climate of the region.

The current solutions or methods cannot classify diseases with more objectivities and extensibilities. For example, many experts have always used their own experience to judge that there may be a particular correlation between occupations and diseases, but they are unable to prove that correlation objectively [40]. Moreover those solution's applications are rare but needed urgently such as in the field of recommendation, etc. Therefore, we should try to come up with a workable plan as much as possible, such as classifying diseases according the occupation feature.

[1] https://www.who.int/nmh/countries/en/.

[2] http://ncncd.chinacdc.cn/xzzq/201611/t20161101_135246.htm.

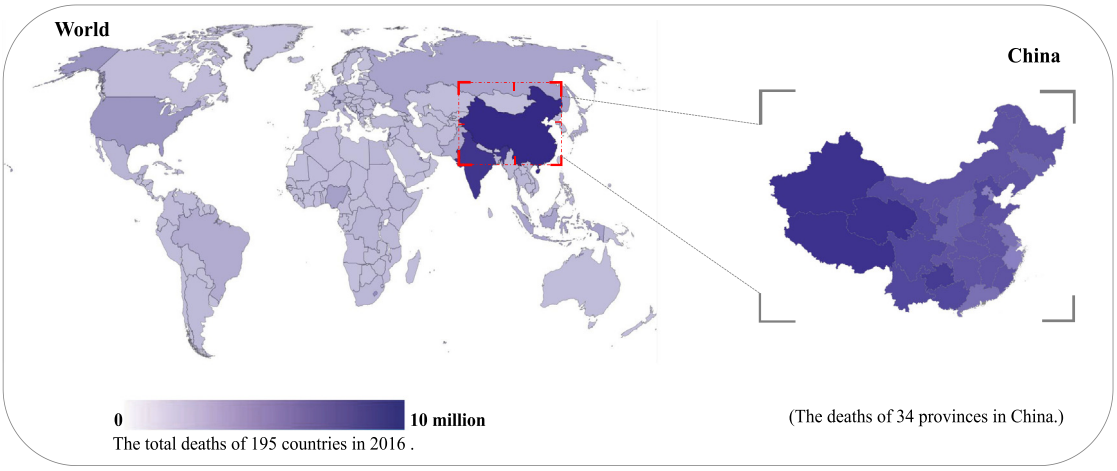


Fig. 1. Distribution of deaths in different countries or cities in China.

3.2. Framework description

After giving the relevant proofs above, we will present the framework proposed in this paper, which mainly classifies diseases according to the occupational attribute. That we are doing has good application values, which is also an essential meaning of this paper. Our framework can not only be used for the recommendation, but also can provide a reasonable explanation, which is one of the hotspots of the current recommendation system research and machine learning (being interpretable) [41]. It is worth pointing out that the proposed framework can be used to classify diseases according to other disease-related attributes, not just occupations. If it is necessary to perform disease classification based on other attributes, the proposed framework will replace the occupation with the data what you want. In other words, the proposed framework is very expandable and robust. The realization of the framework benefits from the classification of diseases in a certain specific region, which is mainly reflected in the collection of data. If the data collected by the framework is mainly from China, it can be stated that the framework is to do a classification of diseases based on the Chinese region.

Fig. 2 represents the data-driven disease classification framework called d-DC, which consists of four main components: the data fusion layer, the feature extraction layer, the method layer, and the result presentation layer. The collection of data is designed with a model that enables automatic replenishment. For humans, the current type of diseases and occupations are only one part while we believe that there will be others as time goes by. Mainly because we still have a lot to explore in the future, and we humans cannot prove that what we know now is all. In addition to the automatic replenishment, we design the data model to effectively prevent the cold start problem of the model when there are no more data. The names of diseases and occupations are primarily added manually as they must be clinically or experimentally proven. Those data is generally given by relevant authorities and is recognized worldwide. The disease names and occupation names used herein are given by WHO and CLSSPH5[3] respectively, which will retain the cascading settings in this framework. It can be shown in Fig. 3. The advantages of retaining cascading settings are (1) for easy retrieval; (2) fuzzy queries can be performed with large classes when small classes have no way to deal with them. The data samples used in this paper are mainly texts (including numbers) and images or videos. There are three ways to collect text data: question and answer (Q & A) forms, automatic uploads, and case (clinical history) associations. The Q & A form is designed not by the user or the patient, but by the hospital, which enables the acquisition of specific structured data. Automatic uploading provides semi-structured data primarily through users or patients, and its format is not much required. This way is mainly to supplement the deficiency of Q & A forms, so as to avoid losing the patient's full personal data as much as possible. With the advent of the digital age, medical institutions such as hospitals have established personal or family medical records history (cases). Therefore, data from the third method is provided by cases. Since the Q & A form is professionally designed, the acquired data is directly pre-processed without the auditing, and data of the other two methods must be reviewed before they can be pre-processed. One of the biggest bases for disease prevention or clinical intervention is based on medical images or videos. Therefore, the design of this paper is very reasonable to analyze the relationship between diseases and occupations by them. After preprocessing, in addition to necessary attributes such as height, weight, region, there are 210 other factors, such as various carcinoembryonic antigens indexes, different glycated hemoglobin Alc, glutamyl transpeptidases, etc. [42].

The feature expression layer is designed mainly because of the different data formats obtained from the data fusion layer, and the unit and normal range of each attribute are inconsistent. If it becomes impossible to make an unified expression, then it will be difficult to be used in the next process. Due to a large number of samples and attributes, we will carry out the preliminary attribute reduction in this layer. Another reason for reduction of attributes is that some of them are very poorly correlated with what we are studying. The correlation analysis used here is implemented by SPSS [43]. Since some of the above data are difficult to express by numbers, for example, some cases use text descriptions to illustrate a result, another module (knowledge representation) in this layer can make up for the above shortcomings. Knowledge representation is primarily achieved through the classic TransE model and the machine learning models.

The method layer is chiefly implemented by traditional methods, machine learning classification methods, and knowledge graph. Traditional methods are mainly composed of simple statistical methods and information entropy methods. Machine learning classification methods include classical models, such as the support vector machine (SVM), DNN, ANN and their combinations [44,45]. The knowledge graph method used here has no specific classification algorithm, and it is mainly to classify diseases by using the association between data. The proposed

[3] http://www.class.com.cn/.

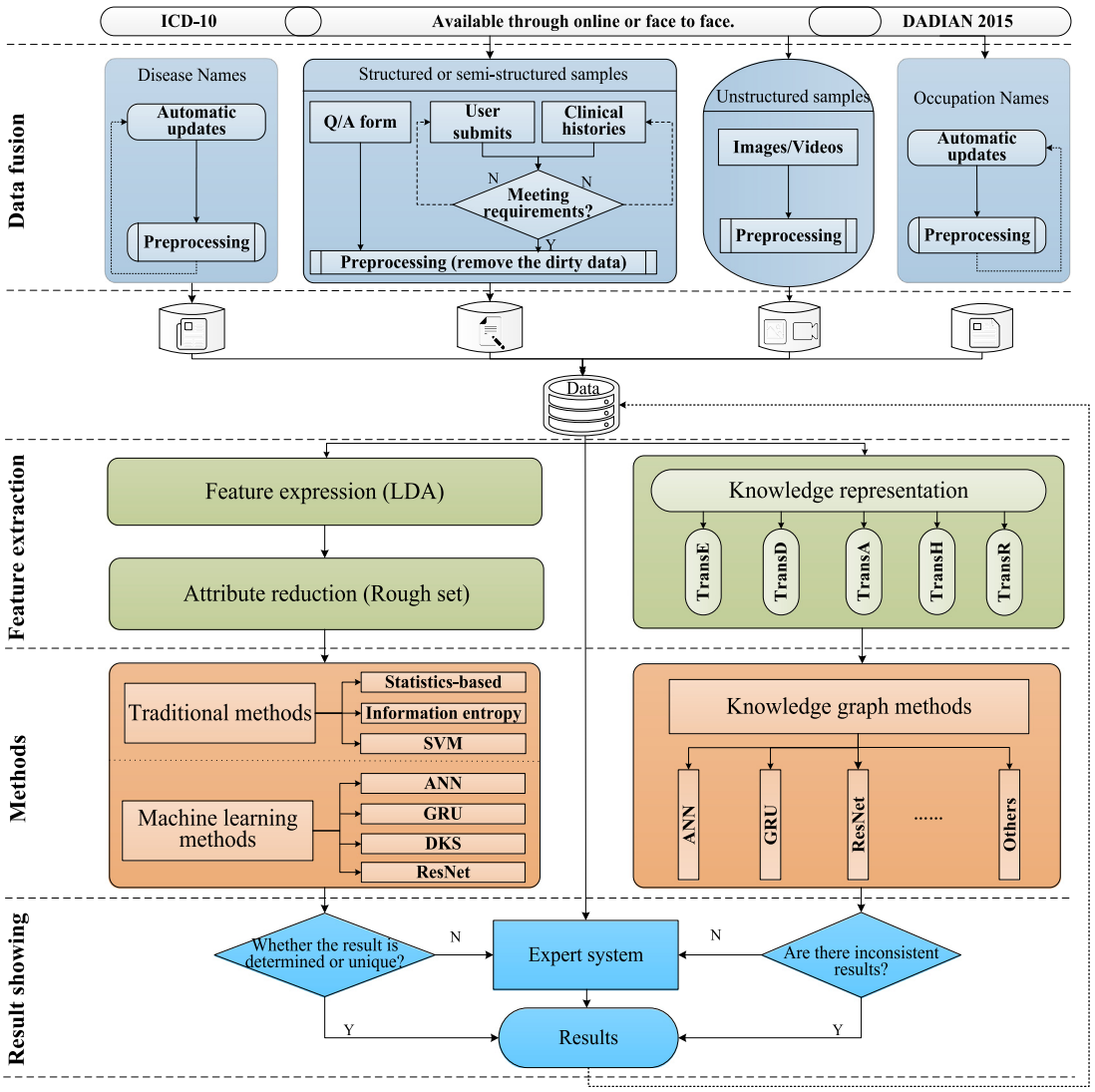


Fig. 2. Occupation-based framework for disease classification.

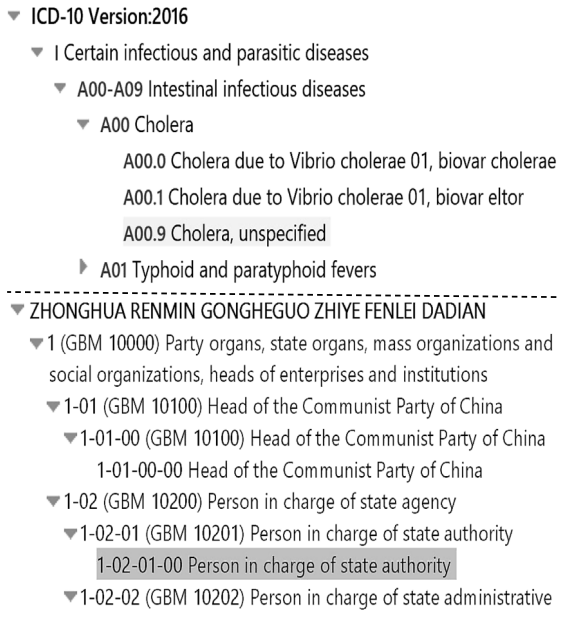


Fig. 3. Tree structure storage for disease names and occupation names.

framework does not give a specific introduction of methods here, and they will be explained in detail in Section 4. The intermediate results from the method layer will be verified by the medical expert system before being output as a final result. For example, there are too many classification results for a certain occupation, and it is necessary to give a simplified result among them through the expert system. The establishment of the expert system may require more online in the early stage, similar to the medical online consultation. The result presentation layer is not just a simple showing of the results, but also gives a certain explanation for better presentation. The final results will be expressed using KG-related techniques. For example, Fig. 4 directly explains which occupations will occur in breast cancer, and why the teaching profession is associated with periarthritis.

4. Methodology

4.1. Data filling

In previous studies, we found that some samples are always incomplete but inevitable due to objective reasons. Since the samples we have involve many attributes and autonomy in collecting data, some attribute values may be missing in many samples. In order to fill those missed values, the suitable method of filling

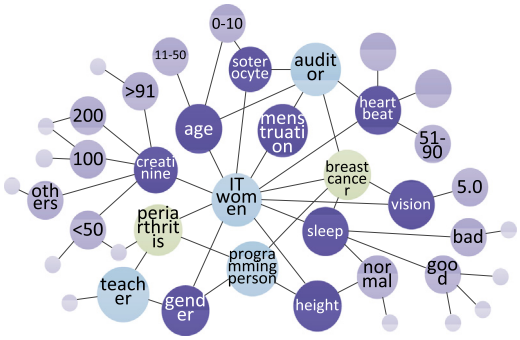


Fig. 4. An example of the result showing.

data is needed. The filling methods used in this paper are mainly as follows [46]:

- *Mean replacement method (MRM)*. It is the decision to fill an average value calculated in the class or the entire dataset. For example, considering we are measuring the student height of a certain class and if a student in the class remain absent, then, the height of the student can be estimated from the mean of height of the existing students. This method is applicable when the difference in the dataset is small, but it is not recommended to use this method to fill the missing value when the samples in the dataset differ largely.
- *Regression replacement method (RRM)*. It is based on the regression equation. First, it is necessary to give the predicted value of the filling data to establish the corresponding regression equation, and then to use the conditional expectation as the filling value.
- *Expert experience method (EEM)*. It is based on the experience of people in the field of filling data to estimate a or some values, which is considered better than the above two filling methods. Since the results obtained by EEM method vary from person to person, the accuracy of filling data is not guaranteed to be high. In this paper, this method is based on the normal range and standard of various medical indicators.

4.2. Similarity measure

When the classification or prediction is performed, there may be multiple results. The framework presented in this paper is no exception. Similarity measures are particularly important in this paper, and we will introduce them in the following section.

- *Cosine value* evaluates their similarity by calculating the cosine of two vectors, a and b, which ranges from $[-1,1]$. The larger the cosine value, the smaller the angle θ between the two vectors, representing the orientation of the two vectors is relatively close. Conversely, the smaller the cosine value, the larger their angle θ. When the two vectors are orthogonal (a 90 degree angle), the cosine value takes a minimum of -1.

$$C_{oss}(\theta) = \frac{a \cdot b}{\|a\|\|b\|} = \frac{\sum_{i=1}^{n} x_{1i}x_{2i}}{\sqrt{\sum_{i=1}^{n} x_{1i}^2}\sqrt{\sum_{i=1}^{n} x_{2i}^2}}. \quad (1)$$

- *The correlation coefficient* is a statistical indicator used to reflect the close relationship between variables. It is based on the dispersion between two variables and their respective mean values. The calculated two dispersions can reflect the degree of correlation between the two variables. Here, the larger the correlation coefficient, the smaller the difference between the vectors. $C_{orc}(a, b) =$

$$\frac{n\sum_{i=1}^{n} x_{1i}x_{2i} - \sum_{i=1}^{n} x_{1i} \cdot \sum_{i=1}^{n} x_{2i}}{\sqrt{n\sum_{i=1}^{n} x_{1i}^2 - (\sum_{i=1}^{n} x_{1i})^2} \cdot \sqrt{n\sum_{i=1}^{n} x_{2i}^2 - (\sum_{i=1}^{n} x_{2i})^2}}. \quad (2)$$

- *Membership degree* indicates that the proximity of an object belongs to a certain set. We know that there are only two values, 0 or 1, in the relationship between an object and a set in the general set concept. However, it is not defined like that in a fuzzy set. It considers that the object's degree of membership on the set can choose any value in the range [0,1]. Therefore, the membership function may represent well the ambiguity of the category, which can be expressed as follows [47]:

$$M_{ik} = \frac{1}{\sum_{j=1}^{c} (\frac{d_{ik}}{d_{ij}})^{\frac{2}{m-1}}} \ (k = 1, 2, \ldots, c.), \quad (3)$$

where M_{ik} represents the membership between the ith sample and the kth class, and d_{ik} presents the Euclidean distance between the ith sample and the kth class. c denotes the number of classes.

4.3. Information entropy

Information entropy is used to measure the disorder degree of data, and also to measure the amount of valid information provided by the data. The smaller the information entropy, the more effective the information it provides. Conversely, the larger the information entropy, the less effective the information it includes [48]. We use information entropy to calculate the weight of each factor, in other words, the effectiveness of the component is calculated through attribute information. The greater the effectiveness, the greater importance of its correspondence. According to the definition of entropy, the entropy of the jth component can be expressed as follows:

$$E_j = -\frac{1}{\ln n}\sum_{i=1}^{n} f_{ji} \ln f_{ji}, \quad (4)$$

where f_{ji} should be understood as the probability that the jth factor will appear in the ith disease. Since our samples are all de-emphasized, there is no identical disease, that is, the probability of each factor occurring is 1, which is not suitable for our data. In this paper, we use the value of the factor to indicate the probability of occurrence. If the value is larger, the probability of occurrence is large, which is also true. If $f_{ji} = 0$, we specify $f_{ji} \ln f_{ji} = 0$ in order to make the above expression $\ln f_{ji}$ to be meaningful. Then, the effectiveness $\check{E}_j$ of the component jth can be defined as follows:

$$\check{E}_j = \frac{1 - E_j}{m - \sum_{j=1}^{m} E_j}, \quad (5)$$

where m represents the number of factors according to a certain disease. Moreover, the probability of the jth factor can be redefined as $f_{ji} = \frac{x_{ji}}{\sum_{i=1}^{n} x_{ji}}$.

4.4. Rough set

Rough set theory can be viewed as a mathematical tool developed by Pawlak in 1982 to that uncertain and inexact knowledge. It has been successfully applied in the fields of artificial intelligence (AI), data mining, pattern recognition and intelligent information processing [49]. Attribute reduction is one of the important research contents of rough set theory. It is to eliminate

duplicate and redundant attributes to achieve the compression and re-refining of knowledge while preserving the basic information, and ensuring that the classification ability of objects is not reduced. At present, a variety of attribute reduction algorithms based on rough sets have been proposed, such as those based on positive domain, distinguishing matrix, information entropy and decision tree. The form of decision table S in a rough set can be represented by a four-tuple: $S = (U, A, V, f)$. Usually, we use $S = (U, A)$ to represent $S = (U, A, V, f)$. U represents a non-empty finite set of objects, called the universe. A denotes a non-empty finite set of attributes, $A = C \cup D$. C is a conditional attribute set, and D is a decision attribute set. If $D = \phi$, then, S represents a decision table with no decision attributes. $V = \cup_{a \in A} V_a$ is a set of attribute values, and V_a is the value range of the attribute a. $f : U \times A \to V$ is an information function that assigns an information value to each attribute in each object, namely $\forall a \in A, x \in U, f(x, a) \in V_a$. Let P be a subset of the A attribute ($P \subseteq A$), then, the indistinguishable relationship $ind(P)$ of the attribute set P can be $ind(P) = \{(x, y) \in U \times U \mid \forall a \in P, f(x, a) = f(y, a)\}$. $ind(P)$ is the equivalence relation on U, and the symbol $U/ind(P)$ ($or\ U/P$) indicates that the indistinguishable relationship $ind(P)$ is derived on U. P for $x \in U$, the equivalent class containing x on the P attribute can be denoted as $[x]_P = \{y \mid y \in U, (x, y) \in ind(P)\}$. The dependency of the decision attribute D on the condition attribute P is presented as $|POS_P(D)|/|U|$, which represents the positive domain of the decision attribute under the condition attribute. In other words, it means all the elements in U can be uniquely categorized into subsets of the subset U/D by P. The importance $I(a)$ of the attribute a in the attribute set P can be defined as follows:

$$I(a) = |POS_P(D) - POS_{P-a}(D)|/|U|, \tag{6}$$

$$C(S) = |POS_C(D)|/|U|, \tag{7}$$

where $C(S)$ refers to the compatibility of the decision table S. If $|POS_C(D)| = |U|$, then, S is called a compatible decision table, otherwise it is named an incompatible decision table. Notably, the necessary and sufficient condition for a decision table to be a compatible decision table is that its compatibility is equaled to one.

4.5. Linear discriminant analysis (LDA)

Linear discriminant analysis (LDA) is a supervised dimensionality reduction algorithm that is linear. Compared with PCA, LDA maintains different data information [50]. It requires samples within the same category to be as close as possible, and those samples that are not within the same class are to be separated as much as possible. Its goal is to find a subspace that distinguishes different categories better. The subspace is obtained by minimizing the rank of the discrete matrix S_W within the class and maximizing the rank of the discrete matrix S_B between classes. According to its definition, S_W and S_B can be expressed as follows:

$$S_W = \sum_{i=1}^{c} S_i, \ \ (S_i = \sum_{j=1}^{n_i} (x_j^{(i)} - \mu_i)(x_j^{(i)} - \mu_i)^T), \tag{8}$$

$$S_B = \sum_{i=1}^{c} n_i (\mu - \mu_i)(\mu - \mu_i)^T, \tag{9}$$

where c represents the number of categories, n_i shows the number of samples of the ith class, and $\mu = \bar{x}$ denotes the mean of all samples. $\mu_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_j^{(i)}$ presents the mean of the samples for each class. $x_j^{(i)}$ represents the jth object in the ith class. Fisher linear discriminant function can be expressed as follows:

$$F(W) = \frac{|W^T S_B W|}{W^T S_W W}. \tag{10}$$

4.6. Support vector machine (SVM)

Support vector machine (SVM) proposed in 1995 is a supervised learning model based on the principle of structural risk minimization. The ultimate goal of the SVM is to obtain a hyperplane that makes samples as separable as possible. Moreover, the hyperplane can not only completely separate the different types of samples but also make the distances of different categories as large as possible [51]. With a linearly separable dataset $(x_i, y_i), x_i \in R^d, y_i \in 0, 1$, there is a function $g(x) = w^T x + b$. After normalization, $|g(x)| \geqslant 1$ is satisfied for all samples, and its classification equation is $w^T x + b = 0$. Substituting the data closest to the hyperplane into the equation, which satisfies $|g(x)| = 1$, then, the distance between the two classes can be calculated as $\frac{2}{\|w\|}$. If it is necessary to completely separate all samples, then it needs to make $y_i[(w^T x_i) + b] - 1 \geqslant 0$ and the hyperplane is $y_i[(w^T x_i) + b] - 1 = 0$. If the hyperplane maximizes the distance between the two categories, we need to minimize $\|w\|$. Lagrangian function could be defined as follows:

$$L(w, b, \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^{n} y_i[(w^T x_i) + b] - 1, \tag{11}$$

where $\alpha > 0$ represents the Lagrangian coefficient. The solution of the largest hyperplane can be obtained by:

$$\begin{cases} \max \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ s.t\ \alpha_i \geqslant 0,\ i = 1, 2, \ldots, N \\ \sum_{i=1}^{N} \alpha_i y_i = 0 \end{cases}. \tag{12}$$

If α_i^* is the optimal solution, then $w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$. According to the Kuhn–Tucker conditional theorem, α_i^* has a unique solution and $\alpha_i[y_i(w^T x_i + b) - 1] = 0$. The obtained α_i^* is zero for most samples x_i, but it has exceptions. The hyperplane $y_i[(w^T x_i) + b] - 1 = 0$ is the support vector, and the optimal classification function is:

$$\begin{aligned} f(x) &= sgn\{(w^*)^T x + b\} \\ &= sgn\{(\sum_{i=1}^{N} \alpha_i^* y_i x_i)^T x + b^*\}. \end{aligned} \tag{13}$$

4.7. Knowledge embedding model (TransE)

Knowledge graph is generally represented by a triple, such as (*subject, predicate, object*) or (*h, r, t*). The *subject* and *object* are called head node and tail node (or entity), which are expressed by h and t (or e). The *predicate* is named edge or relation, which is expressed by r. Each instance (triple) denotes a fact. The bold **e** or **r** represents the corresponding embedded vector. Δ represents a set of triples in KG, and Δ' shows a set of corruptions. e', h', t' and r' present the corrupt entity and the corrupt relationship. E and R denote the entity set and the relationship set. $\Delta = \{\langle h, r, t\rangle \mid h, t \in E, r \in R\}$ and $\Delta' = \{\langle h', r, t\rangle \mid h' \in E\} \cup \{\langle h, r, t'\rangle \mid t' \in E\}$ [35]. TransE expresses the relationship r as a translation vector **r**, so that the entity pair in the triple (h, r, t) can be connected by r and the L_2 error is small enough.

$$f_r(\mathbf{h}, \mathbf{t}) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_{\ell_1/\ell_2}, \tag{14}$$

$$\mathcal{L} = \sum_{(h,r,t)\in\Delta} \sum_{(h',r',t')\in\Delta'_{(h,r,t)}} [f_r(\mathbf{h},\mathbf{t}) + \gamma - f_r(\mathbf{h}',\mathbf{t}')]_+, \tag{15}$$

where γ is the edge hyperparameter and $[\star]_+$ is the negative sample distribution associated triples. The objective function $\mathcal{L}$ is to guarantee correct triples possess the lower scores, and the wrong triples have the higher scores. The marginal value can be seen as the minimum threshold between the two kinds of triples. We usually use the stochastic gradient descent (SGD) or Adam to optimize $\mathcal{L}$ [52].

4.8. Classifier for extracting the domain-related knowledge-rich sentence (DKS)

The classifier for extracting the domain-related knowledge-rich sentence (DKS) aims to determine whether a sentence in the text corpus is a DKS or not [53]. DKS classifier mainly composes of four layers (sentence layer, embedding layer, LSTM layer and output layer) wherein the middle two layers are formed by three parallel networks. DKS considers the target sentence and the other two sentences (its precursor sentence and successor sentence) as features of the classification. It constructs three parallel networks (one embedding layer and one LSTM layer) with similar structures for three sentences. A total score for their outputs in the output layer is generated. The sentence layer distributes embedding matrices for each sentence. In other words, for a target sentence s_1 with a precursor sentence s_2 and a successor sentence s_3, the word w in the sentence s_i is vectorized using the word vector matrix M_i ($e_w = M_i r(w)$). $r(w) \in \{0,1\}^{|V|}$ is the $one-hot$ expression of w, and $|V|$ is the size of the vocabulary list. The result from sentence layer is passed to the LSTM layer, which is used to model for the long dependence of the word sequence. The LSTM layer consists of a sequence of memory cells, each of which attains the input from the embedding layer and the precursor unit. The memory unit has four essential components: an input gate, a forget gate, a state storage unit, and an output gate. The parameters involved in LSTM are updated as follows [54]:

$$\begin{cases} i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \\ j_t = \sigma(W_{xj}x_t + W_{hi}h_{t-1} + b_j) \\ f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \\ o_t = \tanh(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \\ o_t = \tanh(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \\ c_t = c_{t-1} \odot f_t + i_t \odot j_t \\ h_t = \tanh(c_t) \odot o_t \end{cases}, \tag{16}$$

where σ is a sigmoid function, and i, f, o and c are the input gate, the forget gate, the output gate and the unit activation vector, respectively. j is used to calculate the new c. W and b are parameters of LSTM. The output layer connects the results of LSTM and uses the sigmoid function to determine the total score (TS) of the target sentence to judge whether it is a DKS.

$$TS(x_i;\theta) = \sigma(W[h_p, h_i, h_a] + b), \tag{17}$$

where h_p, h_i and h_a are the three outputs of the LSTM layer. The model training uses the seed DKS marked by the marker module as positive samples. Also, the meaningless sentences are used as negative samples for training the model. The binary cross entropy is viewed as a loss function during training. If $TS(x_i;\theta) \geq 0.5$, the prediction label of x_i satisfies $\widehat{f}(x_i;\theta) = 1$, otherwise $\widehat{f}(x_i;\theta) = 0$.

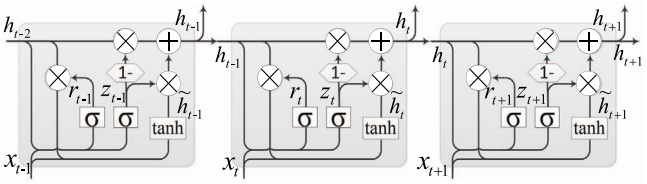


Fig. 5. Gated recurrent unit.

For the parameter θ, the corresponding objective function can be viewed as:

$$L(\theta) = -\sum_{i=1}^{n} y_i \log\widehat{f}(x_i;\theta) + (1-y_i)\log(1-\widehat{f}(x_i;\theta)). \tag{18}$$

4.9. Gated recurrent unit (GRU) network

Gated recurrent unit (GRU) network [55] is a variant of LSTM, which simplifies the LSTM structure. Compared to LSTM, GRU is simpler in construction, easier to be trained and has fewer parameters. A specific GRU replaces the forget gate and the input gate of LSTM unit with an update gate. Moreover, it does not build linear self-updates on additional memory cells but on a hidden state to adjust the information using the control gate. Its structure is shown in Fig. 5.

The computational iteration equations of GRU are as follows:

$$\begin{cases} r_t = \sigma(W_r x_t + U_r h_{t-1}) \\ z_t = \sigma(W_z x_t + U_z h_{t-1}) \\ \widetilde{h}_t = \tanh(W_h x_t + U(r_t \cdot h_{t-1})) \\ h_t = z_t \widetilde{h}_t + (1-z_t)h_{t-1} \end{cases}, \tag{19}$$

where r_t represents the output value of the reset gate in GRU, that determines whether the current state is resetted according to the state of $t-1$ or not. z_t indicates the output value of the update gate. It determines whether the current state is updated according to the state $t-1$. h_t represents the output of the GRU unit, and $\widetilde{h}_t$ is the candidate output of the GRU, which is controlled by the reset gate. Both W and U are coefficient matrices. Here we introduce the critical parts of GRU. Its basic details of hidden layers and how to train the GRU network are similar to other Deep Learning models.

5. Experiments

5.1. Description of datasets

The data used in this paper is mainly composed of three parts, that is, names of diseases, occupational names, and disease samples. The disease names mainly derive from the *International Statistical Classification of Diseases and Related Health Problems 10th Revision* (ICD-10),[4,5] which is an international and unified disease classification results developed by WHO. It classifies diseases according to the etiology, pathology, clinical manifestations and anatomical location [56]. Occupational names are mainly provided by the *ZHONGHUA RENMIN GONGHEGUO ZHIYE FENLEI DADIAN 2015 Edition*.[6] It mainly consists of eight major categories, 434 subcategories, and 1481 occupation names. Disease samples are almost from multiple hospitals, disease research centers and

4 https://icd.who.int/browse10/2016/en#/XXII.
5 http://www.a-hospital.com/w/ICD-10.
6 https://product.suning.com/0070167435/647168266.html.

Table 1
Details of data (names of diseases, occupation names and disease samples).

No.	Names of diseases		Occupation names	Disease samples
1	Certain infectious and parasitic diseases	Diseases of the skin and subcutaneous tissue	Staff and related personnel	Text data
2	Mental and behavioral disorders	Diseases of the nervous system	Professional technical staff	Picture data
3	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	Congenital malformations, deformations and chromosomal abnormalities	Party, state organs, mass organs, leaders organs, enterprises and institutions	Video data
4	Endocrine, nutritional and metabolic diseases	Pregnancy, childbirth and the puerperium	Manufacturing and related personnel	
5	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	Injury, poisoning and certain other consequences of external causes	Social production service and life service personnel	
6	Diseases of the musculoskeletal system and connective tissue	Certain conditions originating in the perinatal period	Agriculture, forestry, animal husbandry, fishery production, supporting personnel	
7	Diseases of the eye and adnexa	Diseases of the genitourinary system	Soldiers	
8	Factors influencing health status and contact with health services	Codes for special purposes	Other practitioners who are inconvenient to classify	
9	Diseases of the circulatory system	External causes of morbidity and mortality		
10	Diseases of the respiratory system	Diseases of the ear and mastoid process		
11	Diseases of the digestive system	Neoplasms		
Total	24,000		1481	48,952

communities. Most of them are collected from the first affiliated hospital of Zhengzhou University and others from 21 community hospitals in a wide area surrounding Henan province and Fujian province. The details of the data are as follows (see Table 1):

5.2. *Verification approach* & *evaluation metrics*

Currently, general methods for verifying model performance include the re-substitution test (RT) and the cross-validation test (CVT). RT verifies the ability of a classifier to be self-inclusive. Since RT uses the same data during verification and causes all the information of test samples to be included in the entire dataset, its value will be high. CVT is capable of verifying the adaptability of the classifier, which uses different data. The Jackknife is a classic CVT method, which is used in this paper.

The performance of the algorithm requires some metrics to measure how it is. There are many indexes for evaluating the quality of classifiers, and we can summarize them as the sample-based ones and the category-based ones. The former is to compare the prediction results of the algorithm on each test sample, and then to calculate their average value in the overall test set as the final result. The latter is to calculate a mean of all categories as the final result. Suppose the examples were divided into positive and negative classes, then, there would be four cases in the actual classification. Their relationship[7] can be represented by Fig. 6.

TP means true positive, that indicate a positive sample is predicted in a positive class; FP indicates false positive, that is, a negative sample is predicted in a positive class; FN means false negative, i.e., a positive sample is predicted in a negative class; TN indicates true negative, i.e., a negative sample is predicted in a negative class. According to the relationship of the above four cases, we can get the following five indicators, i.e., precision (Pre), recall (Rec), accuracy (Acc), F1-score (F1) and Matthew Correlation Coefficient (MCC). We use P and N to represent actual positive results and actual negative results respectively, i.e., $P = TP + FN$, $N = FP + TN$. Then, all cases can be expressed as $P + N$ or $TP + TN + FP + FN$. Those indexes be expressed as follows.

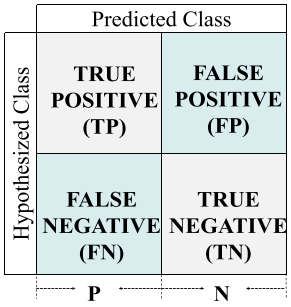


Fig. 6. Relationship between the elements involved in evaluation indicators.

$$Pre = \frac{TP}{TP + FP}, \tag{20}$$

$$Rec = \frac{TP}{P}, \tag{21}$$

$$Acc = \frac{TP + TN}{P + N}, \tag{22}$$

$$F1 = \frac{2}{\frac{1}{Pre} + \frac{1}{Rec}} = \frac{2TP}{2TP + FP + FN}, \tag{23}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{P \cdot (TP + FP) \cdot N \cdot (TN + FN)}}. \tag{24}$$

Note that *Rec* reflects the ability of the classifier to identify positive samples. The higher the recall, the stronger the ability of

[7] https://en.wikipedia.org/wiki/Sensitivity_and_specificity.

the classifier to identify positive samples. *Pre* means the ability of the classifier to distinguish between negative samples. The higher the precision, the stronger the ability of the classifier to distinguish negative samples. F1-score is a combination of *Rec* and *Pre*. The higher the F1-score, the more robust the classifier is. In this paper, there may be multiple classification results that are all correct during prediction. For example, for a teacher occupation, there may be two or more diseases that occur such as sore throat and varicosity or others. If the classifier considers only one as the correct result, then it is not in line with the reality or the actual situation. Therefore, we proposed the sixth indicator. For a certain occupation (or disease), we first calculate a proportion that refers to how many classified diseases (or occupations) are correct, and then find an average for all proportion as the final result of the sixth index, which is called the correct proportion (Corp).

6. Experimental results

In this paper, experimental results will be presented in two parts. The first is about nine diseases with different deaths across 33 provinces of China. It is given because we use a criterion that diseases can be distinguished according to different regions. So, it can be concluded that the classification of diseases can be carried out based on the hierarchical idea such as according to certain attributes. As described in the previous section, we only gave the differences in the deaths of countries around the world. In order to further prove that the hierarchical idea can be used in our proposed framework in details, we give data about the specific death toll in each province of China. The second is the classification results based on the proposed framework using different algorithms. It is mainly divided into four small parts: the results of different classification methods are compared; the classification results are compared with the data obtained by different filling methods; the results of the classification methods using different similarity measures are compared; the before-and-after results of the expert system are compared. Since the focus of this paper is not to how to improve algorithms, the parameter configuration or other settings of the algorithm are not detailed here. They can be attained by the literature we referred to. In this paper, the dataset is divided into three subsets, a training set, a verification set, and a test set according to a proportion of three:one:two.

6.1. Deaths of nine diseases

In order to explain more scientifically the hierarchical and local characteristic of diseases, we have calculated the total number of deaths of nine diseases of 33 provinces of China in 2016. Those nine diseases are selected because their deaths are more than others. We can see from Fig. 7 that the number of deaths caused by different diseases is obviously different in various regions. For example, the diabetes mellitus (g) and the diarrheal diseases (h) are significantly less than the ischemic heart disease (a), chronic obstructive pulmonary disease (c) and ischemic stroke (b). In addition, we can also find that the deaths vary largely in a certain disease. For example, the deaths of tuberculosis (i) in Tibet and Xinjiang province are far greater than in other provinces. The tuberculosis is likely to be called the endemic. Finally, we can also see that gender also has a certain impact on the number of deaths. The deaths of lower respiratory infections (d) are very different in a certain province, and it can be seen that the proportion of male deaths is larger than that of women. The similar conclusion can be completely reflected by other diseases. Therefore, we can conclude that the death of diseases has a strong correlation with factors such as region and age, and our proposed framework based on the hierarchical idea is feasible.

Table 2
The results of ten classification methods (%).

Method	Index					
	Pre	Rec	MCC	F1	Acc	Corp
Sat	38.32	62.16	3.38	47.42	49.10	43.67
Sat_Inf	40.01	80.02	26.73	53.33	58.00	49.23
ANN	68.33	74.55	32.82	71.30	67.05	67.67
SVM	70.10	89.36	56.44	78.50	77.15	73.50
GRU	58.33	77.78	32.82	66.67	65.10	61.67
DKS	76.67	79.31	46.32	77.97	74.21	75.33
ResNet	83.33	87.72	65.14	85.47	83.30	83.17
KG_ANN	38.33	76.67	22.27	51.11	56.04	47.17
KG_GRU	76.67	69.71	27.58	73.02	66.18	71.33
KG_ResNet	85.20	89.47	69.27	87.18	85.06	85.03

6.2. Comparison results from different classification methods

In our proposed framework, the classification methods we used are mainly divided into three parts, the traditional statistical method, the machine learning method and the KG method based on machine learning. Therefore, the comparison results are derived from those methods, and a total of ten methods are compared here. The filling method of data used in this part is the mean replacement. Sat represents results of the statistical method, and Sat_Inf denotes the result of introducing information entropy in statistical methods. ANN presents the result of a traditional neural network with four hidden layers. KG_ANN represents the KG method by using the neural network. Other letters indicate the corresponding methods or fusion methods. Their specific results are given by six indexes, which are shown in Table 2.

From the Pre indicator in Table 2, we see that the results of the statistical method are significantly lower than other methods. In particular, the Sat method has a minimum value, indicating that the statistical method used in our framework is not suitable for the classification of diseases. However, from the Rec index, we can clearly know that the statistics-based approach is not the worst. For example, 80.02% of Sat_Inf method is more than some machine learning methods such as ANN, GRU, etc. In general, we can see that the results of machine learning-based methods are superior to traditional statistical methods. This conclusion can be seen by the Acc indicator, which can represent the overall performance of classification methods. It is worth pointing out that, the results of the SVM method are acceptable. As can be seen from the Rec indicator, 89.36% of the SVM method is second only to the KG_ResNet method. Similar conclusions can also be drawn through the Corp indicator. On the other hand, we can know that the ResNet method shows the most satisfactory results, and we can draw that conclusion in any one of the indicators. Due to the combination between this method and KG, the best results were obtained in KG_ResNet. The 85.06% of Acc and 85.03% of Corp are sufficient to show that it is very feasible. ResNet and KG_ResNet can produce positive outcome probably because they used more hidden layers. In some cases, we can think that the network model with more hidden layers will have a strong performance. Based on this basis, the unpleasant results of both ANN with only four hidden layers and ANN-based KG methods are entirely explainable.

6.3. Classification results using different data filling ways

The disease samples attained in this paper derived from multiple organizations. Therefore, the data with missing attributes will lead to various classification results if they are pre-processed by different filling methods. The expert experience method indicates that the missing attributes are filled according to medical standards. Thus they may not be unique and have a lot of randomness.

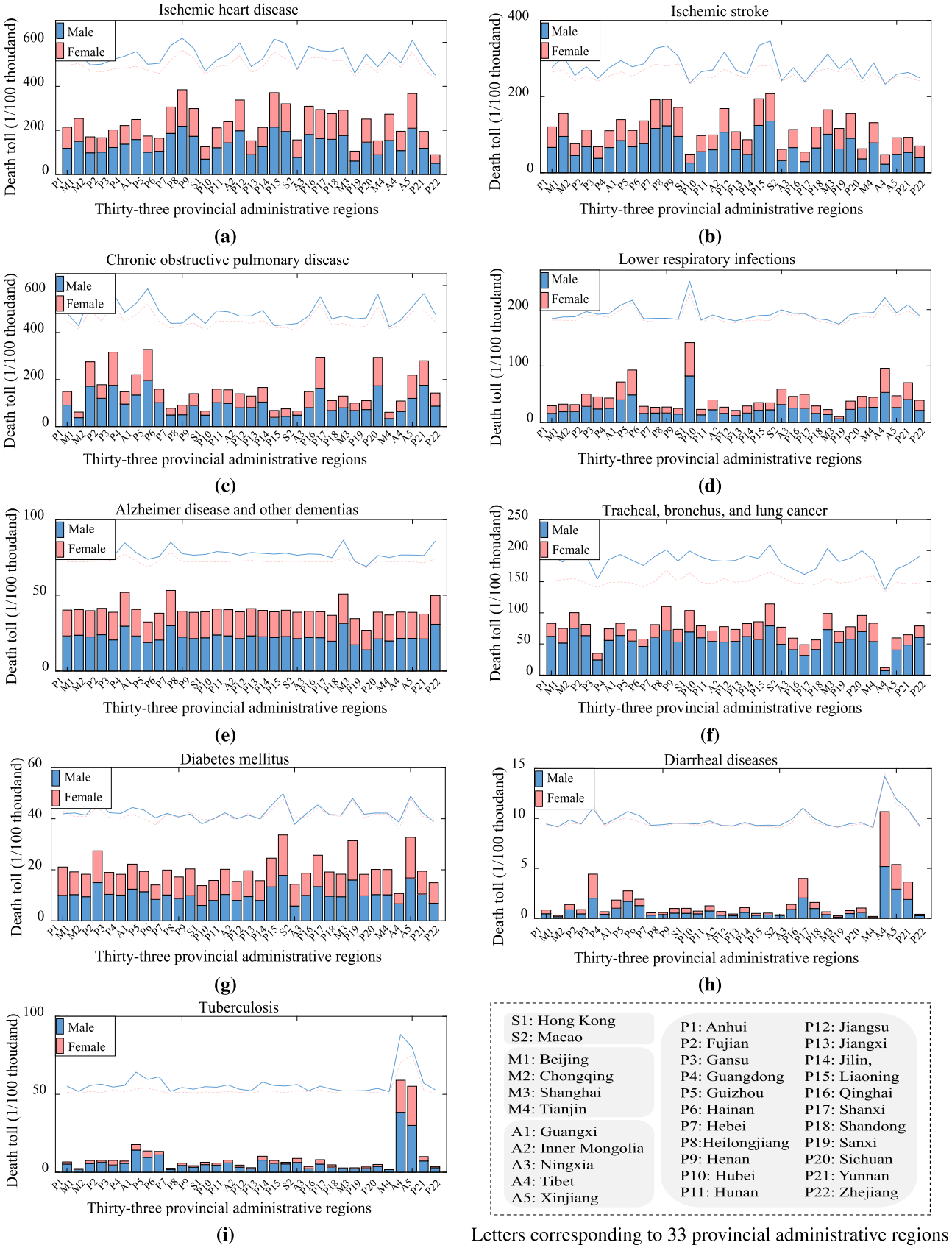


Fig. 7. Deaths of nine diseases in thirty-three provincial administrative regions of China.

Table 3
The comparison results of different data filling methods (%).

Filling way	Method	Index					
		Pre	Rec	MCC	F1	Acc	Corp
RRM	Sat	40.15	58.54	−2.49	47.52	47.03	43.50
	Sat_Inf	43.33	89.66	38.69	58.43	63.20	53.17
	ANN	53.33	68.09	15.54	59.81	57.31	55.17
	SVM	75.18	90.05	61.24	81.82	80.05	77.51
	GRU	60.12	81.82	39.48	69.23	68.08	64.18
	DKS	73.33	78.57	42.77	75.86	72.14	72.67
	ResNet	86.67	89.66	71.14	88.14	86.13	86.33
	KG_ANN	43.33	74.29	21.41	54.74	57.08	50.17
	KG_GRU	76.67	73.02	34.67	74.80	69.31	72.83
	KG_ResNet	83.33	89.29	67.44	86.21	84.10	83.67
EEM	Sat	58.33	77.78	32.82	66.67	65.00	61.67
	Sat_Inf	58.33	81.41	37.93	67.96	67.10	62.67
	ANN	66.67	75.47	33.54	70.82	67.12	66.83
	SVM	58.33	94.59	54.12	72.16	73.13	65.67
	GRU	51.67	86.11	39.97	64.58	66.05	58.83
	DKS	78.33	83.93	55.10	81.03	78.51	78.17
	ResNet	83.33	83.33	58.33	83.33	80.16	81.67
	KG_ANN	43.33	81.25	29.76	56.52	60.28	51.67
	KG_GRU	75.53	76.27	39.84	75.63	71.16	73.20
	KG_ResNet	86.67	91.23	73.39	88.89	87.27	86.83

For example, the human creatinine value is normally in the range of [80, 120]. When the creatinine of a certain sample is missed, we will use 85 and 90 or other close values to fill it. As can be seen from Table 3, different filling methods have significant impacts on the results. Combined with Table 2, the same conclusion is more clearly visible. Although different filling methods have a significant impact on classification methods, the conclusions drawn in Section 6.2 are similar, that is, the results of statistical methods are worse than that of other methods. Also, the ResNet method yields the best results. Overall, the results obtained by the expert experience method are mostly better than the regression replacement method. From the MCC indicator, the Sat method based on RRM is actually a negative value, indicating that the ability to classify diseases incorrectly is much higher than that of the correct classification. In other words, in some data, Sat has no ability to classify diseases. It can be seen from F1 and Corp that the results of KG_ResNet obtained by EEM are 88.89% and 86.83%, which are very acceptable. Therefore, our framework d-DC is utterly applicable to the classification of diseases. In particular, the expert experience approach is the right choice to fill the missing values. The reason EEM can get good results is that it considers not just the integrity of data, but also the practical significance that could have obtained the actual values in real life.

6.4. Classification results of different similarity measures

The classification methods used in this paper are sensitive to similarity measures. Notably, our framework involves the integration of multiple methods and an expert system. When the performance of the model cannot be changed, the quality of the similarity measure will directly affect the final effect of methods. Therefore, this section will compare the differences between the various methods through different similarity measures. We gave comparison results of the Corp index using four different similarity measures on the classification method in Fig. 8.

Fig. 8 shows a comparison of results given by those methods based on different similarity measures. It can be seen that various similarity measures have an impact on certain methods whether they are in statistical methods or machine learning methods. In the ANN method, there are not many changes according to four similarity measures, which may be related to the structure of the neural network. Because we know that the neural network involves the similarity measure only when determining the final result. In the KG_ANN method, we find that the similarity measure has a great impact. Maybe this is because the ANN-based knowledge graph method uses similarity measures for the many times in the knowledge representation, especially in the process of training embeddings. On the other hand, the use of the correlation coefficient is better than the Euclidean distance in each method. In general, we can see that different similarity measures produce various results when our proposed framework is used. Therefore, it is necessary to consider the impact of the similarity measure when adopting different methods for disease classification.

Table 4
The before-and-after comparison results of introducing the expert system based on three indexes (%).

	SVM	ResNet	KG_ResNet	SKRE	RKRE
Pre	58.33	83.33	86.67	**82.93**	88.31
Acc	73.13	80.16	87.27	81.65	88.57
Corp	65.67	81.67	86.83	81.75	86.95

6.5. The before-and-after comparison results of introducing the expert system

In our proposed framework, different classification methods may have inconsistent results for a certain disease. If it happens, we will use an expert system to make the final result. We will combine better methods that can achieve good results. In other words, the SVM method, the ResNet method and the KG_ResNet method all attained the better results in the above experiments. Those three methods are introduced into the expert system to achieve more satisfying results. SKRE means the fusion of SVM, KG_ResNet and our expert system in Table 4. RKRE represents the combination of ResNet, KG_ResNet and our expert system. In this section, we used three indexes to compare their results. We can know that the methods of achieving good results in previous experiments have been improved after they are combined with the expert system. Both in SKRE and RKRE, the values of the three indexes are higher than before except for the Pre. In RKRE method, the result of Corp reached 86.95%, which is very satisfactory. Overall, the introduction of an expert system is very feasible for being used in our framework for disease classification.

7. Conclusion and future work

Classification of diseases through different factors such as occupations can reveal trends of the morbidity and the mortality. Correct analysis of the disease tendency in a certain region can clarify the focus of corresponding prevention and control. In this paper, we proposed a data-driven robust framework with improved scalability called d-DC, that is mainly based on hierarchical ideas for disease classification. In our proposed framework, the establishment of the data collection model can not only achieve automatic replenishment of data, but also effectively tackle the cold start problem of the prediction model without much data. We collected multiple types of data, including structured data, unstructured data such as plain texts, images and videos. Multi-type fusion of data provides conditions for mining more new knowledge and improving the performance of the model, and has even contributed to the establishment of a public medical knowledge base. d-DC combined Deep Learning methods with the knowledge graph, which is the first attempt in the field of classifying diseases. Vectorizing medical texts through knowledge embedding provided a new consideration for the analysis of diseases. When the classification results are singular, our expert

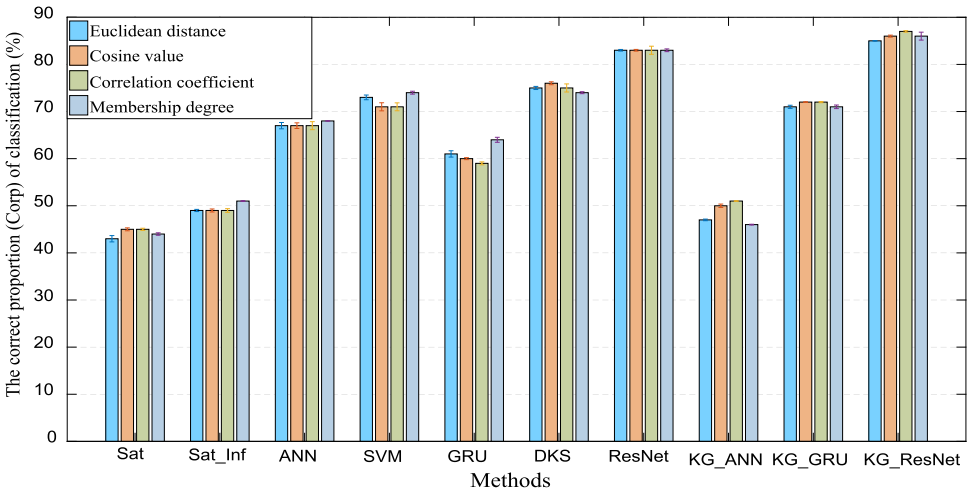


Fig. 8. The classification Corp results based on different similarities (%).

system built through knowledge bases or online experts in d-DC were proposed to address the inconsistencies. The classification was displayed by using a combination of traditional methods and the knowledge graph, which intuitively provides a reasonable basis for explaining why they are like this. Moreover, the experimental results show that the proposed framework achieved better accuracies, which could be used by professional medical institutions to improve healthcare services in terms of developing preventive measures. Although we collected the large number of data with many types, there is not much given about the selection of preprocessing methods, which will be our future work. Besides, for personal medical records, we only used structured data. In future research, we will consider combining other types of samples such as tongue coating data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

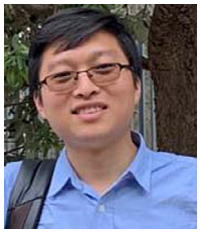
Acknowledgments

This work is partially supported by the National Natural Science Foundation of China under Grant 61672439.

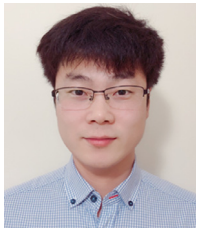
References

[1] H. Wang, M. Naghavi, C. Allen, R.M. Barber, Z.A. Bhutta, et al., Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2015: a systematic analysis for the global burden of disease study, Lancet 388 (10053) (2016) 1659–1724.
[2] L.M. Baxter, M.S. Eldin, A. Al Mohammed, M. Saim, F. Checchi, Access to care for non-communicable diseases in mosul, iraq between 2014 and 2017: a rapid qualitative study, Conflict Health 12 (1) (2018) 1–4.
[3] WTO, World Health Statistics 2017, World Health Organization, 2017.
[4] K. Aleksovska, A. Puggina, L. Giraldi, C. Buck, et al., Biological determinants of physical activity across the life course: a determinants of diet and physical activity (dedipac) umbrella systematic literature review, Sports Med. 5 (1) (2019) 1–18.
[5] H. Wang, J. Zuo, S. Chen, Q. Mei, The status of global chronic non-communicable diseases and prevention and control strategies, Foreign Med. Sci. Soc. Med. Sect. (1) (2005) 10–14.
[6] L. Kong, Prevalence, development trend and prevention strategies of chronic non-communicable diseases, Chin. Gen. Pract. 4 (12) (2001) 927–929.
[7] P. Balakumar, U. K. Maung, G. Jagadeesh, Prevalence and prevention of cardiovascular disease and diabetes mellitus, Pharmacol. Res. 113 (2016) 600–609.
[8] O. Saidi, M. O'Flaherty, N.B. Mansour, W. Aissi, O. Lassoued, S. Capewell, J.A. Critchley, D. Malouche, H.B. Romdhane, Forecasting tunisian type 2 diabetes prevalence to 2027: validation of a simple model, BMC Public Health 15 (1) (2015) 1–8.
[9] W. Chen, R. Gao, L. Liu, M. Zhu, et al., Summary of report on cardiovascular diseases in china (2017), Chin. Circul. J. 32 (6) (2017) 521–530.
[10] X. Wang, Y. Wang, Y. Li, Great Dictonary of Hygiene, Qingdao Publishing House, 2000.
[11] M. Lu, S. Moritz, D. Lorenzetti, L. Sykes, S. Straus, H. Quan, A systematic review of interventions to increase breast and cervical cancer screening uptake among asian women, BMC Public Health 12 (2012) 1–16.
[12] G. Fraser, Preventive Caricology, Oxford University PressInc, 1986.
[13] B. Du, J. Zhou, J. Zhou, Selenium status of children in kashin–beck disease endemic areas in shaanxi, china: assessment with mercury, Environ. Geochem. Health 40 (2) (2018) 903–913.
[14] R. Rattehalli, S. Deshpande, Mental health and its relevance to dentistry, Textb. Oral Med. (2018).
[15] Questia, Your money: Critical illness Can seriously damage your financial health, The News Lett. (2004).
[16] C. Wang, Major Illness Insurance: Protect Your Financial Health (3), Sina Weibo, 2011.
[17] W. Chiu, K. Tsai, Celiac discase: an unrecognized predisposing factor for hypocalcemia in taiwan, Osteoporos. Int. 27 (2016) 759–771.
[18] WHO, ICD-10:international statistical classification of diseases and related health problems:tenth revision, World Health Organization, 2016.
[19] I. Trouwborst, A. Verreijen, R. Memelink, P. Massanet, Y. Boirie, P. Weijs, M. Tieland, Exercise and nutrition strategies to counteract sarcopenic obesity, Nutrients 10 (5) (2018) 1–21.
[20] C. Calitz, K.M. Pollack, C. Millard, D. Yach, National institutes of health funding for behavioral interventions to prevent chronic diseases, Am. J. Prev. Med. 48 (4) (2015) 462–471.
[21] M.M. Anne, V. Sophie, R.R. Wendy, et al., Low-grade inflammation, diet composition and health: current research evidence and its translation, Br. J. Nutr. 114 (7) (2015) 999–1012.
[22] G. Ruel, Z. Shi, S. Zhen, H. Zuo, E. Kröger, C. Sirois, J.F. Lévesque, W.T. Anne, Association between nutrition and the evolution of multimorbidity: The importance of fruits and vegetables and whole grain products, Clin. Nutr. 33 (3) (2014) 513–520.
[23] Y. Zhong, Clinical analysis of complex regional pain syndrome, in: The General Hospital of the People's Liberation Army, 2018.
[24] H.B. Janet, W. Ian, S. Charles, Space-based bias of covert visual attention in complex regional pain syndrome, Brain 140 (9) (2017) 2306–2321.
[25] H.E. Meyer, K. Holvik, P. Lips, Should vitamin d supplements be recommended to prevent chronic diseases?, Br. Med. J. 350 (2015) 1–4.
[26] K. Jyrki, G. Rashid, M. Pekka, V. Sari, N. Tarja, M. Jaakko, K. Jussi, T.P. Tuomainen, Low serum 25-hydroxyvitamin d is associated with higher risk of frequent headache in middle-aged and older men, Sci. Rep. 7 (2017) 39697.

[27] G. Hulsegge, M. Looman, H.A. Smit, M.L. Daviglus, Y.T. van der Schouw, W.M.M. Verschuren, Lifestyle changes in young adulthood and middle age and risk of cardiovascular disease and all-cause mortality : The doetinchem cohort study, J. Amer. Heart Assoc. 5 (1) (2016) 1–11.
[28] W. Lee, K. Choi, R. Yum, D. Yu, S. Chair, Effectiveness of motivational interviewing on lifestyle modification and health outcomes of clients at risk or diagnosed with cardiovascular diseases: A systematic review, Int. J. Nursing Stud. 53 (2016) 331–341.
[29] I. Goodfellow, Deep learning / ian goodfellow, yoshua bengio and aaron courville, Adapt. Comput. Mach. Learn. (2016).
[30] S. Li, Y. Fu, Robust representation for data analytics : models and applications, Adv. Inf. Knowl. Process. (2017).
[31] S. Luo, W. Fang, Potential probability of negative triples in knowledge graph embedding, in: Proceedings of the International Conference on Neural Information Processing, 2018, pp. 48–58.
[32] Y. Lin, Z. Liu, M. Sun, Y. Liu, X. Zhu, Learning entity and relation embeddings for knowledge graph completion, in: Proceedings of the 29th AAAI Conference on Artificial Intelligence, 2015, pp. 2181–2187.
[33] A. Bordes, J. Weston, R. Collobert, Y. Bengio, Learning structured embeddings of knowledge bases, in: Proceedings of the 25th AAAI Conference on Artificial Intelligence, 2011, pp. 1–6.
[34] B. Yang, W.-t. Yih, X. He, J. Gao, L. Deng, Embedding entities and relations for learning and inference in knowledge bases, in: Proceedings of the 3rd International Conference on Learning Representations, 2015, pp. 1–13.
[35] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: Proceedings of the 27th Annual Conference on Neural Information Processing Systems, 2013, pp. 2787–2795.
[36] Z. Wang, J. Zhang, J. Feng, Z. Chen, Knowledge graph embedding by translating on hyperplanes, in: Proceedings of the 28th AAAI Conference on Artificial Intelligence, 2014, pp. 1112–1119.
[37] G. Ji, S. He, L. Xu, K. Liu, J. Zhao, Knowledge graph embedding via dynamic mapping matrix, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, 2015, pp. 687–696.
[38] Z. Hu, X. Ma, Z. Liu, E. Hovy, E. Xing, Harnessing deep neural networks with logic rules, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, pp. 1–11.
[39] L.E. Hinkle, L.H. Whitney, E.W. Lehman, J. Dunn, B. Benjamin, R. King, A. Plakun, B. Flehinger, Occupation, education, and coronary heart disease, Science 161 (3838) (1968) 238–246.
[40] H. Yan, Y. Jiang, J. Zheng, B. Fu, S. Xiao, C. Peng, The internet-based knowledge acquisition and management method to construct large-scale distributed medical expert systems, Comput. Methods Programs Biomed. 74 (1) (2004) 1–10.
[41] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, et al., Wide & deep learning for recommender systems, in: Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, 2016, pp. 7–10.
[42] M. Tomizawa, F. Shinozaki, R. Hasegawa, Y. Shirai, Y. Motoyoshi, T. Sugiyama, S. Yamamoto, N. Ishige, Patient characteristics with high or low blood urea nitrogen in upper gastrointestinal bleeding, World J. Gastroenterol. 21 (24) (2015) 7500–7505.
[43] T. Morelli, C. Shearer, A. Buecker, IBM SPSS predictive analytics: Optimizing decisions at the point of impact, IBM redpaper, 2010.
[44] V.B. Semwal, J. Singha, P.K. Sharma, A. Chauhan, B. Behera, An optimized feature selection technique based on incremental feature analysis for biometric gait data classification, Multimedia Tools Appl. 76 (22) (2017) 24457–24475.
[45] C. Zhang, Z. Zheng, Task migration for mobile edge computing using deep reinforcement learning, Future Gener. Comput. Syst. 96 (2019) 111–118.
[46] E. Liberty, S. Zucker, Y. Keller, M. Maggioni, R. Coifman, F. Geshwind, Methods for filtering data and filling in missing data using nonlinear inference, United States Patent Application, 2007.
[47] W. Au, K.C. Chan, Classification with degree of membership: A fuzzy approach, in: Proceedings of the 2001 IEEE International Conference on Data Mining, 2001, pp. 35–42.
[48] M.N. Banu, B. Gomathy, Disease forecasting system using data mining methods, in: Proceedings of the 2014 International Conference on Intelligent Computing Applications, 2014, pp. 130–133.
[49] C. Cheng, W. Liu, Identifying degenerative brain disease using rough set classifier based on wavelet packet method, J. Clin. Med. 7 (6) (2018) 1–12.
[50] D. Giri, U. Rajendra Acharya, R. Martis, S. Vinitha Sree, T. Lim, T. Ahamed, J. Suri, Automated diagnosis of coronary artery disease affected patients using lda, pca, ica and discrete wavelet transform, Knowl.-Based Syst. 37 (2013) 274–282.
[51] Z. Sun, Y. Qiao, B. Lelieveldt, M. Staring, Integrating spatial-anatomical regularization and structure sparsity into svm: Improving interpretation of alzheimer's disease classification, NeuroImage 178 (2018) 445–460.
[52] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Proceedings of the 3rd International Conference on Learning Representations, 2015, pp. 1–15.
[53] W. Cui, Research of Key Technologies for Question Answering over Knowledge Graphs, Fudan University Press, 2017.
[54] K. Greff, R.K. Srivastava, J. Koutní k, B.R. Steunebrink, J. Schmidhuber, Lstm: A search space odyssey, IEEE Trans. Neural Netw. Learn. Syst. 28 (10) (2017) 2222–2232.
[55] R. Zhao, D. Wang, R. Yan, K. Mao, F. Shen, J. Wang, Machine health monitoring using local feature-based gated recurrent unit networks, IEEE Trans. Ind. Electron. 65 (2) (2018) 1539–1548.
[56] WHO, ICD-11: Classifying disease to map the way we live and die, World Health Organization, 2018.



Zhenfeng Lei received his B.S. degree of computer science from the Zhengzhou University in June 2015 and his MA.Eng (honors) degree of computer science in Yunnan University in 2017. He is currently pursuing the Ph.D. degree in School of Informatics, Xiamen University, Xiamen, China. He won the first prize in China Graduate Contest on Application, Design and Innovation of Mobile-Terminal, which is held at Xidian University in 2018. He had special insights in the field of protein sub-cellular localization. His current research interests include data mining and Deep Learning techniques, knowledge graph and recommended system.



Yuan Sun received the B.E. degree from Zhengzhou University, Henan, China, in 2013 and the M.E. degree in Electrical Engineering from the University of Adelaide, Adelaide, SA, Australia, in 2016. He is currently pursuing the Ph.D. degree in electrical and electronic engineering with the University of Adelaide, Adelaide, SA, Australia. Currently, his research interests include cooperative control, multi-agent systems and machine learning.

Yaser Ahangari Nanehkaran received the B.E. degree from IAU of Ardabil Branch, Ardabil, Iran, in Power Electrical Engineering and M.Sc. degree in IT from Cankaya University, Ankara, Turkey. He is currently pursuing the Ph.D. degree in the Department of Computer Science at Xiamen University, Xiamen city, in China. His research area mainly includes data mining, big data and Deep Learning techniques.

Shuangyuan Yang received the Ph.D. degree from the Computer School, Huazhong University, Wuhan, China, in 2004. He is currently an Associate Professor with the School of Informatics, Xiamen University, Xiamen, Fujian, China. In the past five years, he has hosted or participated in many projects, including two National Natural Science Foundation Projects, one National 863 Project, one Major Science and Technology Project in Fujian Province, five Key Science and Technology Projects in Fujian Province and Xiamen City, and over 10 Enterprise Cooperation Projects. Among them, as the person in charge of the project, nine projects were undertaken, with a total amount of 7.93 million RMB. He has published over 30 journal or conference papers. His research topics include Internet of Things, SAAS service, image processing, and machine learning.

Md. Saiful Islam (M'18) is conducting Ph.D. in the School of Electrical and Electronic Engineering at The University of Adelaide, Australia. Mr. Islam is a member of IEEE, IEEE Young Professionals, Optical Society of America (OSA) and Institute for Photonics & Advanced Sensing (IPAS). His research interests include optical fiber communication, PCF based terahertz waveguides, terahertz sensors, surface plasmon resonance biosensors, topological insulators, metamaterials for sensing applications. Mr. Islam published 33 peer-reviewed articles and actively reviews for IEEE Journal of

Lightwave Technology, IEEE Photonics Journal, IEEE Sensors Journal, IEEE Photonics Technology Letters, Optics Express, Applied Optics, Optical Material Express, and Optical Fiber Technology.

Huiqing Lei got the associate degree diploma of nursing from The Nursing College Of Zhengzhou University in China in 2009. She received her B.S.Nurse degree of The Nursing College Of Zhengzhou University in 2013. And now she is pursing master degree of nursing management. She has been working in the department of Breast Surgery, The First Affiliated Hospital of Zhengzhou University since March 2009. Her research interests include nursing management and medicine-based teaching.

Defu Zhang received his bachelor degree in computational mathematics in 1996, and master degree in computational mathematics in 1999, both from Xiangtan University, and his Ph.D. degree in computer software and its theory from the school of computer science in Huazhong University of Science & Technology. He was a senior researcher of Shanghai Jinxin financial engineering academe from Jun. 2002 to Apr. 2003. Now he works in the department of Computer Science at Xiamen University as a professor. He supervised ACM/ICPC team in Xiamen University and obtained 3 gold medals, 8 silver medals from 2004 to 2009, and took part in the world final contest in 2007. He worked as a PostDoc at the Longtop for financial data mining group from 2006 to 2008. From 2008 to 2016, he visited Hong Kong City University, University of Wisconsin-Madison, Macau University. Besides, he developed an Internet + big data platform (http://www.pzcnet.com). He has published over 40 journal articles and his research interests include computational intelligence, data mining, big data, cloud computing and online decision optimization.