Hindawi
Scientific Programming
Volume 2022, Article ID 5799354, 11 pages
https://doi.org/10.1155/2022/5799354

Research Article

Intelligent Processing and Classification of Multisource Health Big Data from the Perspective of Physical and Medical Integration

Haiou Tang [1,2]

[1] *Hunan City University, Yiyang, Hunan 413000, China*
[2] *Dept of Physical Education, Hoseo University, Asan, Chungcheongnam-do 336-795, Republic of Korea*

Correspondence should be addressed to Haiou Tang; tanghaiou@hncu.edu.cn

Received 17 December 2021; Revised 12 January 2022; Accepted 17 January 2022; Published 28 February 2022

Academic Editor: Baiyuan Ding

With the development of computer science and information technology, human society is gradually stepping into the Internet and big data. The medical and health industry can realize the integration and readjustment of existing resources, improve the operation efficiency of the industry, and tap the huge potential of the industry with the support of big data technology. However, the medical data in the new era has the characteristics of massive, high latitude, complex structure, and complex information, which is not conducive to the direct classification of health data. The preprocessing of health data can improve the quality of dataset, reduce the size of data, and improve the efficiency and accuracy of data classification. Based on this and according to the characteristics of health dataset and the existing pretreatment technology, this paper analyzes and improves the algorithm of abnormal data detection and data protocol in the process of reprocessing data cleaning. This paper analyzes and studies feature selection algorithms based on Bayesian inference algorithm and focuses on feature selection algorithms based on random forest. In order to solve the problem that the original algorithm ignored the relationship between the importance degrees of each feature in a single tree, a feature importance degree calculation method based on local importance degree was proposed. Through experimental analysis and comparison, the improved algorithm can select better feature subset and improve the performance of the classification model. Then, TAN classifier, BAN classifier, and MBN classifier were constructed based on preprocessed hypothyroidism data, and the performances of these four classifiers were compared through experiments. The final results show that BAN classifier has the best average classification effect.

1. Introduction

With the development of computer science and information technology, people have more and more opportunities and ways to get in touch with the Internet, and more and more network data are generated [1]. The massive increase and diversity of network data in the new era bring challenges to data analysis. In order to overcome the above difficulties, data mining technology emerges at the historic moment. Data mining is to discover the latent rules or knowledge which is not easy to be obtained directly through data observation from the massive data containing noise information redundancy or information loss data. Data mining has become one of the important directions of the development of contemporary computer science. The development of health and medical informatization is highly related to the development of computer technology [2, 3]. The development of computer technology and the popularization of computers in medical institutions bring a revolution to medical informatization [4–6].

However, medical big data and other types of big data have similar but different problems; that is, there is a large amount of missing data and there are a large number of repeated data outliers in the data, resulting in low data quality and seriously affecting the effect of data mining. In the medical field, there are various types of data, such as basic information, for example, medical treatment information, hospitalization information, physical examination information, and medical insurance information, and the data access mode is changeable [7]. For example, the common medical structure information input is uploaded to the cloud platform through intelligent testing equipment and APP. Due to the different data structures of different information and different input methods and platform design, the structure of medical data is different, and many health data used in data mining have existed for many years; due to the input storage integration process error, which makes the situation of dataset more complex, the original data directly used in data mining will bring a large error, and it is difficult to meet the needs of products and applications [8–12].

Health big data copies the doctors on the infinite resources of high quality, make the limited distribution of medical resources in a more reasonable manner, and promote the grading clinical therapy to make it more reasonable. Through the analysis of large data on health at the same time, government agencies can realize rational pricing of drug products [13], discover the epidemic disease, and take relevant preventive measures [14, 15].

Therefore, semisupervised learning enhances the performance of learning assumptions by using labeled data and unlabeled data at the same time [16–19]. The initial assumptions are usually learned from labeled data and then updated and strengthened by unlabeled data information to complete the improvement of model performance. A semisupervised learning is actually a supervised learning and unsupervised learning in a compromised way; at the same time it combines the advantages of supervised learning and unsupervised learning, uses a lot of unlabeled data to help improve learning model in a small amount of labeled data generalization ability, and has become a current hot spot of machine learning field. More and more researches focus on semisupervised learning (SSL) [20–23].

2. Related Works

In the middle of the 20th century, after medical informatization started, machine learning technology gradually penetrated into the medical industry with the increasing popularity of Internet mobile devices and the increasing demand of social development and certain results were achieved in many aspects such as the development of auxiliary diagnostic drugs and health management. Foreign research on health big data started earlier, and the representative regions are the United States, Europe, and Japan. Their research on health big data mainly focuses on personal health, clinical decision support, medicine, disease prediction, public health, and other fields and has achieved a lot of results. The American Steward healthcare system is a community-based organization that provides basic care for community residents [24]. Every year, it treats more than 1 million patients in Massachusetts in the form of community hospital services. The Korea Biomedical Center plans to run the national DNA management system, which will combine patients' electronic health record data with system biology data, such as biological small molecules, genes, proteins, and other related data, to provide personalized diagnosis, treatment, and health management for patients, relying on the analysis and mining ability of medical and health big data. Google's Flu Trends APP, for example, helps people understand flu outbreaks in different parts of the world by checking health opinion keywords [25]. IBM developed the Healthcare Fraud Prevention and Abuse Management System (FAMS) to help health insurance payers, which can quickly identify healthcare fraud by mining health insurance payment history information. Artificial neural network can simulate the way of thinking of human brain. Considering that it can be as adaptive as the brain when dealing with nonlinear relations, it has very strong practical value [26]. BP neural network algorithm is applied to breast cancer data and improved it with particle swarm optimization algorithm. The results show that the BP neural network has better performance with fewer samples and more attributes. Support vector machine (SVM) maps the sample vector to the high-dimensional space according to the kernel function, and the mapped vector is relatively sparse in the high-dimensional space, which is conducive to finding the best separated hyperplane to complete the classification task [27]. Since it is very effective in the classification of small samples and nonlinear problems, it is also often applied in the medical and health field. In 2002, a variety of classification methods were applied to diagnose skin pigmentation diseases, and the results showed that SVM had the most reliable classification effect.

Although China's information-based medical treatment started late and there is a gap in the application scale of health big data classification technology compared with foreign countries, with the strong support for the development of health big data, the research on health big data classification technology has also received more and more attention [28]. The model was used for auxiliary diagnosis of breast cancer. The results of 5-fold cross verification showed that the detection reached 96.93 in 683 patients. The authors of [29] used SVM to obtain the highest classification ability and classification accuracy and could effectively conduct clinical differential diagnosis for sarcoidosis and *tuberculosis*. Li et al. [30] used artificial neural network (ANN) to perform auxiliary diagnosis of DMD in children with rare leg neuromuscular disease based on magnetic resonance images (MRI) of patients, alleviating the pain caused by traditional diagnosis and detection schemes. Shanghai built a municipal data center in 2018 to share medical data with all 500 public hospitals, with about 16 million pieces of data stored in the core database every day. Liu et al. [31] analyzed the characteristics and content of medical records text, proposed a preprocessing method aiming at these characteristics, and applied this method to coronary heart disease dataset, and the effect of data analysis was significantly improved. Due to the problems of abnormal data, redundant data, and missing data in the original physical examination dataset, it cannot be directly used for data analysis and information mining of diseases. In order to make better use of valuable information in physical examination data, different preprocessing methods are proposed for different purposes: to reduce the time and space complexity of preprocessing, datasets are compressed. Liu et al. [32] realized the consistency and continuity of physical examination data over the years through data transformation based on linear function [34, 33].

From the above analysis, we know that the above methods have studied the intelligent processing and classification of multisource health big data to some extent;

some problem still exists [35, 36]. For example, no scholar has applied the models to this field from the perspective of physical and medical integration till now, so the research here is still a blank, which has great theoretical research and practical application value for intelligent processing and classification of multisource health big data. In addition, almost all classification models have shallow structure framework.

The contributions of this paper are as follows: (a) It introduces the basic theory of Bayesian network, including probability theory, basic principle of Bayesian network, Bayesian network learning, and common Bayesian network classifier. (b) The improved Bayesian network structure learning algorithm in Chapter 3 was used to construct the data classification model of hypothyroidism, and the performances of different Bayesian network classifiers were compared.

This paper consists of five parts. The first and second parts give the research status and background. The third part gives the processing and classification of multisource health big data. The fourth part shows the experimental results and analysis. The experimental results of this paper are introduced and compared and analyzed with relevant comparison algorithms followed. Finally, the fifth part concludes the paper.

3. Processing and Classification of Multisource Health Big Data

3.1. Perspective of Physical and Medical Integration. In fact, before the two terms of sports appeared or became specific nouns, our ancestors have long given us precious historical and cultural heritage: the longevity of the Traditional Chinese guidance method of health preservation represented by the Five Birds Opera and Eight Duan Brocade is the result of historical inheritance. In the new era, people begin to constantly improve their health needs, and sports and medicine are different levels of solutions around health needs.

Health is a complex and multidimensional concept, covering the concept of human physiology, psychology, society, and many other fields. As different branches of physiology, sports technology and medical technology have the same root but different application directions. Simply speaking, medical science is to guarantee the safety of human life, just like food and clothing in life. Solve the problem of human health and sports is the goal of life to a higher level; for example, a well-off standard of living in the life health is also a relatively vague definition; it is difficult to accurately define, having different embodiment in different areas.

Many industries all around health in modern society in the development, such as the primary side of the agriculture and animal husbandry and fisheries, life cannot leave the food processing, such as industrial equipment and quality inspection again. Environmental protection, from this point of view, as mentioned above, sports play a more prominent role in health, while medicine is only to solve the negative impact of disease on health in life, sports method is also more intuitive, and medicine? Without the help of drugs, I believe that the ability of doctors to heal the wounded and save the dying will immediately decline.

3.2. Multisource Health Big Data System. Firstly, this paper designs a multisource health big data management system as shown in Figure 1.

(1) The system uses Bluetooth, network, WIFI, and other technologies for intelligent collection, covering a number of medical and health items such as blood analysis, biochemical analysis, urine analysis, and ECG monitoring.

(2) There is no liquid path or pipeline in the Chinese medicine testing equipment of the system, and it has wireless and wired network automatic data upload function.

(3) Based on B/S and C/S framework structure, the system built a data uploading platform to realize accurate and stable uploading of all medical and health data and a smart signing mobile APP was launched to make it meet the basic public health service requirements.

Based on the computer network communication technology and network technology, the B/S framework structure is used to realize the integration of the detected equipment detection data, and the collected data will generate dynamic health records and connect with other hospital systems to realize the computer monitoring and automatic management of medical examination and test process. The dynamic full-process closed-loop health management mode combining offline and online is adopted to collect medical data in real-time offline and analyze and manage the data in all directions online to achieve prediction, prevention, and personalized health maintenance.

3.3. Feature Extraction Strategy. With the development of the medical big data diagnosis and treatment technology, the realization of a more efficient medical image analysis can be complementary to help the doctor condition analysis, help doctors to determine treatment plan, and reduce the dependence on clinical experience in the diagnosis of misjudgment rate. Therefore, high-efficiency and highly accurate medical diagnosis model can provide quantitative and objective endoscopy diagnosis for doctors. It makes it easier for clinicians to notice suspicious pathological images, reduces the workload of eye screening, and helps doctors to make correct clinical medical decisions.

In order to improve the accuracy of medical diagnosis model and effectively extract medical image features, this chapter proposes an image feature extraction algorithm with rotation invariance, which is named TriZ. TriZ algorithm is improved from image feature extraction algorithm HOG and generated by the HOG algorithm with 378,434 features. In this section, it is proved experimentally that this algorithm has rotation invariance for three gastric diseases, namely, gastric polyp, gastritis, and gastric ulcer, and can achieve effective detection and classification of gastric diseases in the case of 10-fold cross validation. TriZ's classification accuracy reached 87.0% among the four classification problems of the three types of gastric diseases and the healthy control. The specific research process of

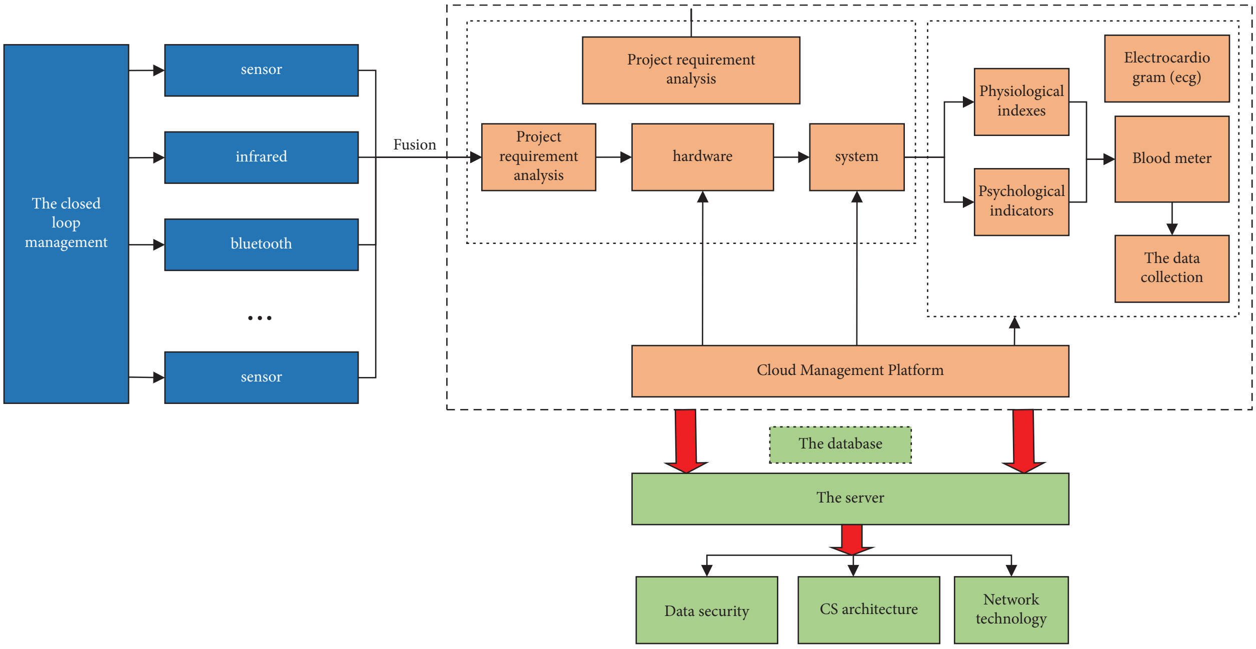


Figure 1: Multisource health big data system process.

extracting the diagnostic model of gastric diseases based on the medical image features of TriZ is shown in Figure 2.

3.4. Application of Naive Bayesian Network Classifier. The Primitive Bayesian network classifier adopts the Attribute Conditional independence assumption, which assumes that all nonclass attributes are independent of each other; that is, each attribute can independently affect the classification results, corresponding to the Bayesian network, each nonclass attribute node only has the category as its parent node, and then formula (1) is obtained.

$$P(F_1, F_2, \ldots, F_n, C) = P(C)\prod_{i=1}^{n} P(F_i|C). \qquad (1)$$

In a similar way,

$$P(F_1, F_2, \ldots, F_n) = \prod_{i=1}^{n} P(F_i|\pi(F_i)) = \prod_{i=1}^{n} P(F_i), \qquad (2)$$

where $F_1, F_2, \ldots, F_n$ denote the n variables and C is the center variable.

By integrating formulas (1) and (2), the posterior formula of calculation in naive Bayes classifier is

$$P(C|F_1, F_2, \ldots, F_n) = \frac{P(C)\prod_{i=1}^{n} P(F_i|C)}{\prod_{i=1}^{n} P(F_i)}. \qquad (3)$$

In the above formula, $P(F_i)$ represents the attributes probability of F_i; $\prod_{i=1}^{n} P(F_i)$ is constant for every category. There are usually several attributes, so class attributes is the a posteriori probability of C and proportional to that of $P(C)\prod_{i=1}^{n} P(F_i|C)$; namely,

$$P(C|F_1, F_2, \ldots, F_n) \propto P(C)\prod_{i=1}^{n} P(F_i|C), \qquad (4)$$

where $P(C)$ represents the prior probabilities of each category and $P(F_i|C)$ denotes the probability of occurrence of attribute F_i under the condition of known class C, which can be directly calculated by sample dataset.

Naive Bayesian network classifier is based on all the class attributes under the premise of mutual independence between complete classification tasks; although in real life it is often difficult to fully meet the conditions of datasets, there are still many researchers who use naive Bayesian network classifier as a kind of commonly used classification model. In this case, even if the dataset does not satisfy the conditional independence hypothesis, it still has good classification performance. Therefore, it is necessary to learn the tree structure between nonclass nodes, and the maximum weighted spanning tree is generally adopted. The weight between two nodes is expressed in conditional mutual information; its calculation formula is as follows:

$$I(X_i, X_j|C) = \sum_{x_i, x_j, c} P(x_i, x_j|c)\log\frac{P(x_i, x_j|c)}{P(x_i|c)P(x_j|c)}. \qquad (5)$$

The main principle of ReliefF algorithm is as follows: firstly, a sample is randomly selected from the dataset as X, and then k samples closest to X are selected as H in the sample set of the same class as X according to Euclidean distance, and k samples closest to X are found in the sample set different from X, and then the above process is repeated M times according to formula (6) to update the weight of each feature and output the final weight of each feature:

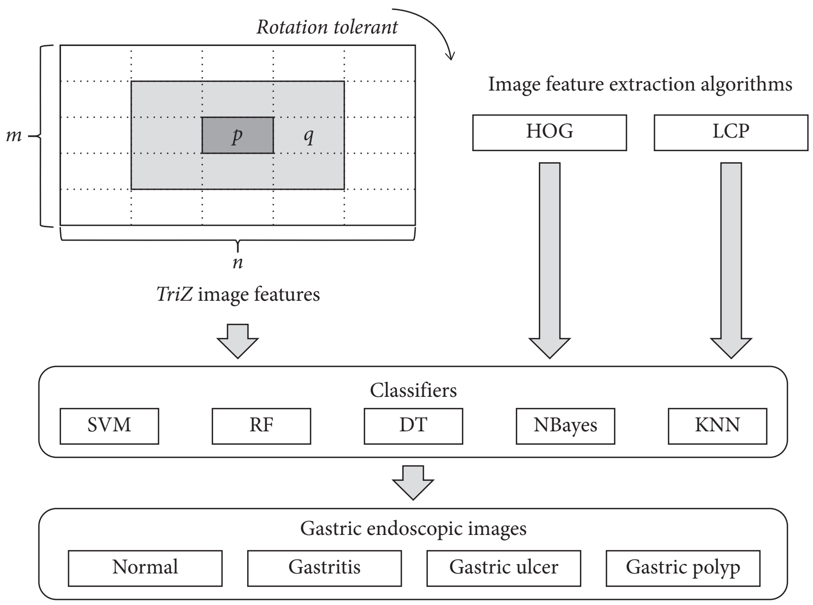


Figure 2: Feature extraction flow chart.

$$W(A) = \frac{\sum_{i=1}^{k} \operatorname{diff}(A, x, H_i)}{mk} + \frac{\sum_{Ce\mathrm{Class}(x)} \left[(P(C)/1 - P(\mathrm{Class}(x))) \sum_{i=1}^{k} \operatorname{diff}(A, x, M_i(C)) \right]}{mk}. \tag{6}$$

In the calculation of feature weight, ReliefF algorithm only considers the correlation between feature and class and ignores the possible redundancy between features. Therefore, this algorithm has certain limitations. In order to eliminate redundant attributes more effectively, this section introduces symmetric uncertainty in information theory based on ReliefF algorithm. Further eliminating redundant features, symmetrical uncertainty (SU) can measure the correlation between two variables. Suppose that the two variables are X and Y, and the formula for calculating the symmetrical uncertainty between two variables is as follows:

$$SU(X, Y) = \frac{2^{*} IG(X; Y)}{H(X) + H(Y)}, \tag{7}$$

where $H(X)$ and $H(Y)$, respectively, represent the information entropies of variables X and Y, and $H(X)$ is defined as follows:

$$H(X) = -\sum p(x) \log p(x), \tag{8}$$

where x represents different values of variable X and $IG(X; Y)$ represents information gain, also known as mutual information, which can be obtained by the following formula:

$$IG(X; Y) = H(X) - H(X|Y), \tag{9}$$

where $H(X|Y)$ denotes the conditional entropy of given variables X and Y, as defined below:

$$H(X|Y) = -\sum_{x \in X, y \in Y} p(x, y) \log p(x|y). \tag{10}$$

By synthesizing the above formulas, the symmetric uncertainty between variables X and Y can be obtained, and symmetry is full due to mutual information. It can be inferred that symmetric uncertainty is also symmetric, and, in order to make the magnitude of symmetric uncertainty comparable, it normalized the mutual information so that the symmetric uncertainty between features is between 0 and 1. When $SU(X, Y) = 0$, it means that variables X and Y are two independent variables; when $SU(X, Y) = 1$, it is indicated that variables X and Y are completely correlated.

$$IG(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = IG(Y; X). \tag{11}$$

Let X be the condition that satisfies Y_v; then variable Y_v satisfies the following equation:

$$p(Y_v | X, Y_u, u \neq v) = p(Y_v | X, Y_u, u{:}\ v), \tag{12}$$

where u and v represent two vertices contained in graph T, and then (X, Y) is a random field of strips.

4. Experimental Results and Analysis

4.1. Introduction to Experimental Environment and Dataset. The purpose of the experiment in this section is to test the effectiveness of the improved algorithm on liver disease detection data, which is mainly reflected from two aspects: the efficiency of the algorithm execution and the detection accuracy of repeated data in the dataset. However, the actual data set accessed does not have a data mark about whether each data is a duplicate, so the performance of the improved algorithm cannot be tested with the original data set. Therefore, in order to measure the efficiency and scalability of the algorithm more comprehensively, the original numbers are standardized according to the centralized data in this paper, and 5000, 10000, and 20000 data points are distributed. The generation rules of repeated data detection for three datasets of different sizes are as follows: The data of each dimension of the original record is standardized. Each original data consists of 0–9 corresponding repeated pieces of data, and the number of repeated pieces of data follows Zipf distribution. Each repeated piece of data has 0–5 changes, and the similarity between the modified data and the original data is greater than or equal to the threshold value. The data in each dataset consists of two parts, 50% of which is the original data, and the other 50% is the repeated data modified according to the original data. All the models in this paper are coded by Python language, and all the experiments in this paper are carried out on a hardware device of NVIDIA 1080Ti GPU.

The standard to measure the performance of the repeated data detection algorithm is whether it is efficient and comprehensive to the repeated data detection in the dataset. According to the setting in this chapter, the detection of repeated data in the experiment in this chapter is essentially a binary algorithm, and the commonly used standards mainly include precision rate, recall rate, consumption time (Time), and AUC area.

4.2. Experimental Results Analysis. In order to verify the performance of classifier SVM, the best values of two parameters C and Gamma need to be filtered. The measured values of classification performance were calculated through 10-fold cross validation. Three heat maps were used to represent the data of four evaluation indicators: S_n, S_p, and Acc. The maximum, average, and minimum values of each measurement were represented in red, yellow, and blue, and the values of the color range were represented in gradient colors, as shown in Figure 3.

Parameter C is set to 20 with step size of 0.125 between 0.125 and 3.000, and parameter Gamma is set to {0.100, 0.178, 0.316, 0.562, 1.00, 1.334, 1.778} for grid search to find the best choice for these two parameters. The results show that when $C = 2.125$ and Gamma = 0.100, the SVM classifier is the best, and its classification accuracy is 97.2%. The algorithm integrates the morphological features of eyes and mouth in the face region and studies and discusses the fatigue detection problem from the aspects of feature number, classifier, and modeling parameters. The algorithm consists of three main steps. First, PCA algorithm is used to calculate the main components. Finally, the SVM model with RBF kernel is trained to classify the images. The experimental results show that the image recognition accuracy of this algorithm reaches 96.07%, and the operation time is only about 21 milliseconds, which can meet the requirements of real-time fatigue monitoring task with 30 frames per second.

In order to verify the stability of the proposed health big data classification method, we can choose the standard deep learning algorithm, whose data processing and parameter setting are roughly the same as the proposed algorithm. At the same time, all the above methods were cross-validated 10 times, and the average result of the test dataset is shown in Figure 4. It can be seen from Figure 4 that the robustness of the proposed method is best correlated with classification. It is worth noting that these experimental results were averaged over 20 times over 80000 datasets for more universality.

In order to verify the validity of the Bayesian network classifier based on the improved ReliefF algorithm, this classifier is compared with the BAN classifier based on ReliefF algorithm (ReliefF-BAN) and three other Bayesian network classifiers (NBC, TAN, and BAN).

Since ReliefF algorithm is needed to calculate the weights during the initial feature screening and the results of the initial feature screening with k larger than the initial value are selected, the difference of k will affect the feature subset finally obtained. If k is too large, for example, 28, or k is too small, it is likely to lead to the deletion of some features that are highly correlated with the class. In this section, $k = 27$, $k = 23$, and $k = 17$ are selected, respectively, as the preliminary screening results, and then further screening of feature subsets is completed according to different thresholds, and the performances of classifiers formed under different feature subsets are compared. The final results are shown in Figure 5.

In this section, Youden index is used to evaluate the Jorden index of each model under different proportions of labeled samples, as shown in Figure 6. After analyzing Figure 6, we can draw the following conclusion. When using the same base classifier to train the classification model, the Youden index of the optimized self-training model is higher than that of the standard self-training model and the supervised learning model. For example, taking naive Bayes as an example, the Youden index of the optimized self-training classification model is 58.90%, and the Youden index of the standard self-training classification model is 49.97%, while the Youden index is 48.84% when only naive Bayes algorithm is used for classification. This is because the self-training algorithm after optimization can learn more unlabeled sample information, and the information learned through repeated labeling strategy is more accurate. Therefore, the comprehensive performance of the optimization algorithm is better, which also proves the effectiveness of the algorithm.

Due to the introduction of mislabeled samples, the Youden index of the standard self-training classification model is not necessarily higher than that of the supervised learning classification model. For example, taking decision tree as an example, the Youden index of the standard self-training classification model is 53.99%. The Jorden index of decision tree classification model is 54.97%, and the comprehensive performance of decision tree classification model is better than that of standard self-training classification model. This illustrates the instability of the standard self-training algorithm.

In conclusion, we can know that the supervised classification algorithm performs well in the results that the training algorithm for standard test data classification does not, which may be due to the low classification performance of the base classifier when selecting unlabeled samples, which leads to the continuous accumulation of errors and weakens the

Sn		C																							
		0.125	0.250	0.375	0.500	0.625	0.750	0.875	1.000	1.125	1.250	1.375	1.500	1.625	1.750	1.875	2.000	2.125	2.250	2.375	2.500	2.625	2.750	2.875	3.000
Gamma	0.100	0.9946	0.9911	0.9902	0.9920	0.9911	0.9902	0.9911	0.9911	0.9911	0.9884	0.9884	0.9875	0.9875	0.9875	0.9875	0.9866	0.9875	0.9875	0.9866	0.9866	0.9866	0.9866	0.9857	0.9848
	0.178	0.9991	0.9929	0.9929	0.9911	0.9902	0.9911	0.9902	0.9911	0.9902	0.9902	0.9902	0.9875	0.9875	0.9848	0.9848	0.9839	0.9821	0.9804	0.9804	0.9795	0.9786	0.9786	0.9777	0.9768
	0.316	1.0000	0.9982	0.9973	0.9946	0.9938	0.9938	0.9920	0.9893	0.9857	0.9830	0.9830	0.9821	0.9821	0.9804	0.9795	0.9795	0.9804	0.9804	0.9795	0.9795	0.9795	0.9786	0.9786	0.9777
	0.562	1.0000	1.0000	0.9982	0.9982	0.9982	0.9964	0.9946	0.9946	0.9938	0.9929	0.9911	0.9911	0.9902	0.9902	0.9893	0.9893	0.9884	0.9884	0.9884	0.9884	0.9884	0.9884	0.9884	0.9884
	1.000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9991	0.9982	0.9973	0.9973	0.9973	0.9964	0.9964	0.9964	0.9964	0.9964	0.9964	0.9964	0.9964	0.9964	0.9964	0.9964	0.9964	0.9964	0.9964
	1.334	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9991	0.9982	0.9982	0.9982	0.9982	0.9982	0.9982	0.9982	0.9982	0.9982	0.9982	0.9982	0.9982	0.9982	0.9982	0.9982	0.9982
	1.778	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9991	0.9991	0.9991	0.9991	0.9991	0.9991	0.9991	0.9991	0.9991	0.9991	0.9991	0.9991	0.9991	0.9991	0.9991

Sp		C																							
		0.125	0.250	0.375	0.500	0.625	0.750	0.875	1.000	1.125	1.250	1.375	1.500	1.625	1.750	1.875	2.000	2.125	2.250	2.375	2.500	2.625	2.750	2.875	3.000
Gamma	0.100	0.632	0.766	0.818	0.852	0.864	0.877	0.9911	0.889	0.891	0.896	0.9884	0.895	0.895	0.896	0.898	0.900	0.907	0.907	0.904	0.905	0.904	0.904	0.904	0.904
	0.178	0.4196	0.6875	0.7750	0.8161	0.8446	0.8554	0.9902	0.8732	0.8750	0.8804	0.9902	0.8875	0.8893	0.8911	0.8911	0.8929	0.8929	0.8911	0.8911	0.8911	0.8911	0.8929	0.8946	0.8982
	0.316	0.1339	0.4518	0.6036	0.6875	0.7571	0.7804	0.9920	0.8179	0.8286	0.8393	0.9830	0.8500	0.8618	0.8500	0.8518	0.8518	0.8518	0.8536	0.8536	0.8536	0.8554	0.8536	0.8554	0.8554
	0.562	0.0000	0.1643	0.3250	0.4500	0.5393	0.5911	0.9946	0.6750	0.6964	0.7089	0.9911	0.7143	0.7143	0.7143	0.7161	0.7161	0.7179	0.7196	0.7196	0.7196	0.7196	0.7196	0.7196	0.7196
	1.000	0.0000	0.0000	0.0500	0.1357	0.2500	0.3196	0.9982	0.4179	0.4518	0.4768	0.9964	0.4821	0.4821	0.4821	0.4821	0.4821	0.4821	0.4821	0.4821	0.4821	0.4821	0.4821	0.4821	0.4821
	1.334	0.0000	0.0000	0.0036	0.0321	0.0946	0.1732	1.0000	0.2786	0.3268	0.3554	0.9982	0.3554	0.3554	0.3554	0.3554	0.3554	0.3554	0.3554	0.3554	0.3554	0.3554	0.3554	0.3554	0.3554
	1.778	0.000	0.000	0.000	0.009	0.039	0.063	1.0000	0.129	0.171	0.200	0.9991	0.204	0.204	0.204	0.204	0.204	0.204	0.204	0.204	0.204	0.204	0.204	0.204	0.204

Acc		C																							
		0.125	0.250	0.375	0.500	0.625	0.750	0.875	1.000	1.125	1.250	1.375	1.500	1.625	1.750	1.875	2.000	2.125	2.250	2.375	2.500	2.625	2.750	2.875	3.000
Gamma	0.100	0.874	0.916	0.933	0.945	0.949	0.952	0.955	0.957	0.958	0.958	0.958	0.957	0.957	0.957	0.958	0.958	0.961	0.961	0.959	0.960	0.959	0.959	0.958	0.958
	0.178	0.8060	0.8911	0.9202	0.9327	0.9417	0.9458	0.9470	0.9518	0.9518	0.9536	0.9548	0.9542	0.9548	0.9546	0.9546	0.9546	0.9524	0.9506	0.9506	0.9500	0.9494	0.9500	0.9500	0.9500
	0.316	0.7113	0.8161	0.8661	0.8923	0.9149	0.9226	0.9280	0.9321	0.9333	0.9351	0.9369	0.9381	0.9387	0.9369	0.9369	0.9369	0.9375	0.9381	0.9375	0.9375	0.9381	0.9369	0.9375	0.9369
	0.562	0.667	0.7214	0.7738	0.8155	0.8432	0.8613	0.8750	0.8881	0.8946	0.8982	0.8988	0.8988	0.8982	0.8982	0.8982	0.8982	0.8982	0.8988	0.8988	0.8988	0.8988	0.8988	0.8988	0.8988
	1.000	0.667	0.6667	0.6833	0.7119	0.7500	0.7726	0.7923	0.8042	0.8155	0.8238	0.8250	0.8250	0.8250	0.8250	0.8250	0.8250	0.8250	0.8250	0.8250	0.8250	0.8250	0.8250	0.8250	0.8250
	1.334	0.667	0.6667	0.6679	0.6774	0.6982	0.7244	0.7405	0.7589	0.7744	0.7839	0.7839	0.7839	0.7839	0.7839	0.7839	0.7839	0.7839	0.7839	0.7839	0.7839	0.7839	0.7839	0.7839	0.7839
	1.778	0.667	0.667	0.667	0.670	0.680	0.688	0.695	0.710	0.724	0.733	0.734	0.734	0.734	0.734	0.734	0.734	0.734	0.734	0.734	0.734	0.734	0.734	0.734	0.734

Figure 3: Optimization of the heat maps of parameters C and Gamma of SVM.

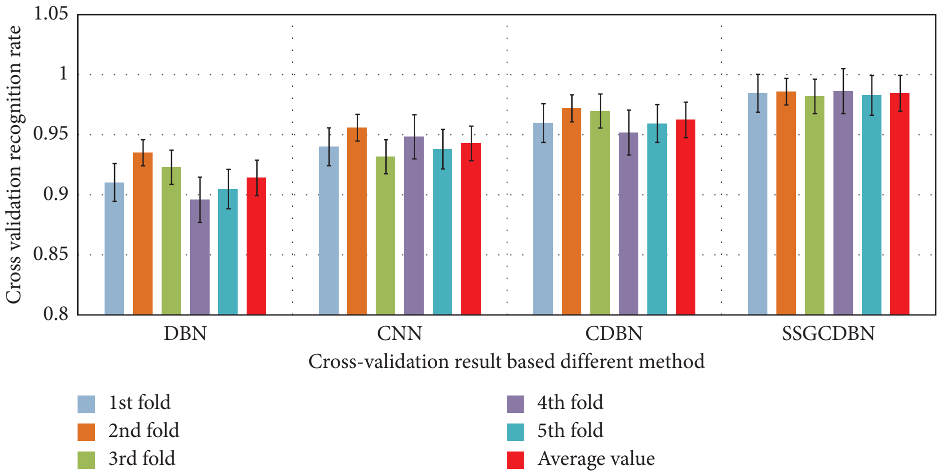


Figure 4: Cross validation based on different classification methods.

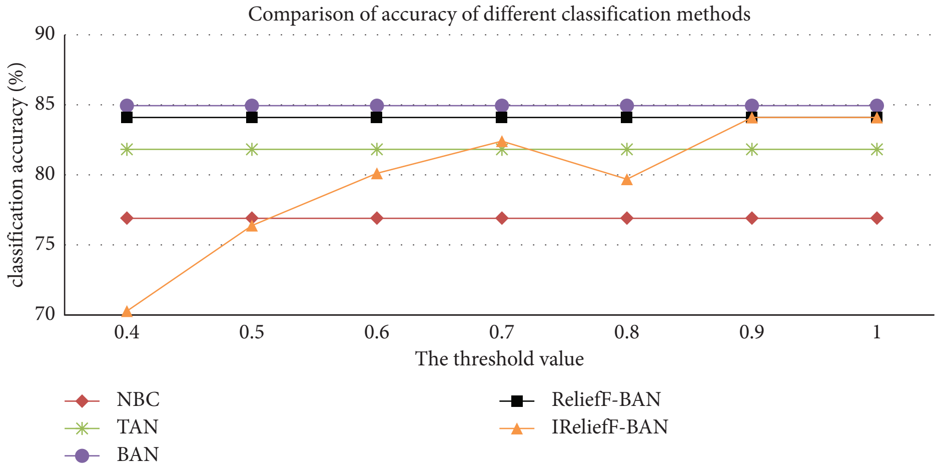


Figure 5: Comparison of the accuracies of different classifiers when $k = 27$.

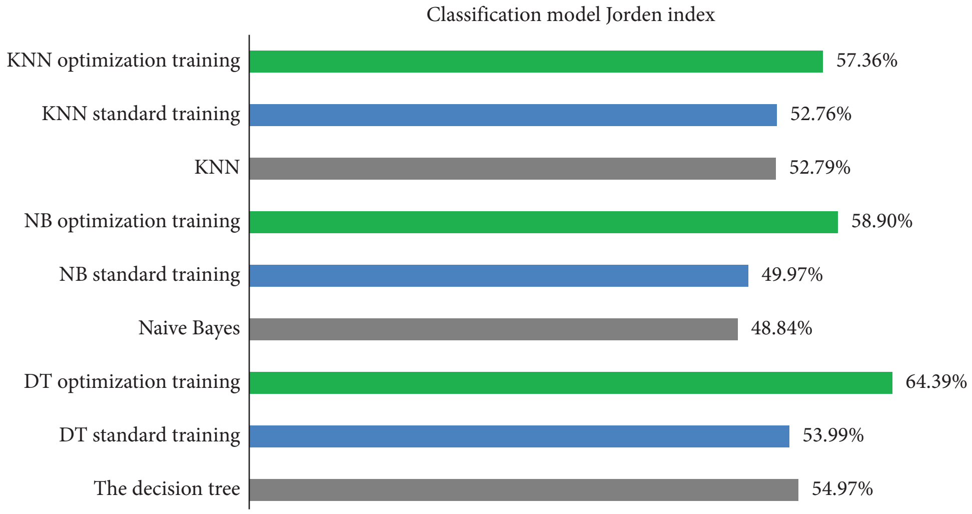


Figure 6: Comparison diagram of Jorden index for classification model.

R2	LR	SVC	KNN	GBDT	DT	RFC	GaussianNB	BernoulliNB	SGDC	PAC
MI	−0.7073	−0.0054	−0.6503	−0.1477	−0.8163	−0.3468	−0.8448	−0.8495	−1.5182	−1.0724
Ftest	−0.2804	0.0136	−0.4701	−0.0244	−0.7547	−0.2520	−0.7452	−0.7310	−1.2858	−0.8638
Ttest	−0.6456	−0.0291	−0.6219	−0.2994	−1.0392	−0.3563	−1.2052	−0.0670	−0.3279	−0.4322
PearsonR	−0.5413	−0.0054	−0.4701	−0.0718	−0.8638	−0.2899	−0.8068	−0.7736	−0.7310	−0.6930
MINE	−0.6883	0.0088	−0.4275	−0.1097	−0.9633	−0.3658	−0.7262	−0.6788	−1.1246	−1.1104
PSR	−0.5413	−0.0054	−0.4701	−0.0338	−0.8638	−0.2899	−0.8068	−0.7736	−0.7310	−0.6930
RankSum	−0.6409	−0.0386	−0.6456	−0.2757	−0.9586	−0.4512	−1.2005	−0.0623	−0.5602	−0.4831
LrRFE	0.2080	−0.0101	−0.5318	−0.2330	−1.0013	−0.4559	−1.1056	−1.2811	−0.6409	−0.3848
LassoRFE	−0.2852	−0.0101	−0.6314	−0.1382	−0.8400	−0.4227	−1.1198	−1.3190	−0.9112	−0.6029
SgdcRFE	−0.0907	−0.0101	−0.6503	−0.1809	−1.0155	−0.3468	−1.0108	−1.2194	−0.3658	−0.3136

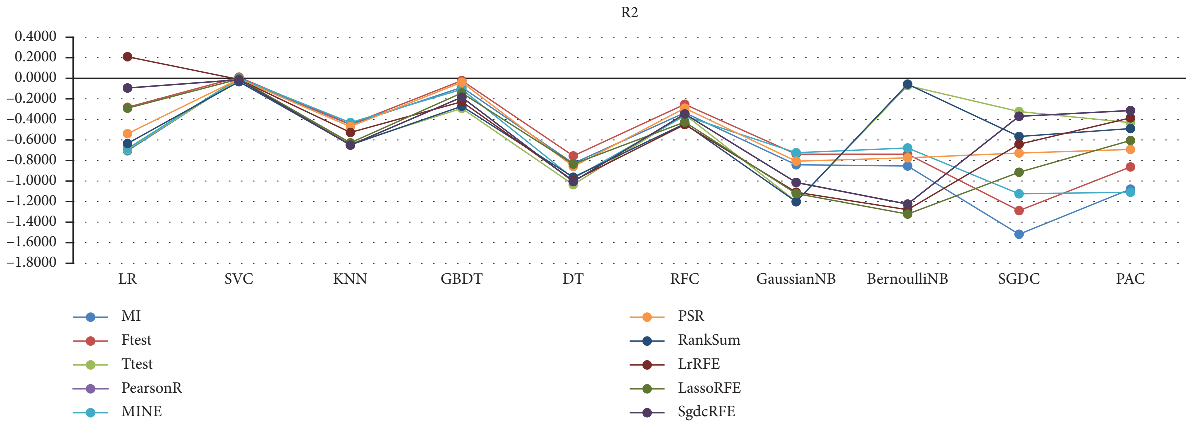


Figure 7: Comparison with the existing feature selection algorithms.

performance of the classifier. However, the classification performance of the self-training algorithm after optimization is generally superior to those of the supervised algorithm and the standard self-training algorithm, which proves the effectiveness of the optimization algorithm.

Figure 7 shows the comparison between the proposed regression biomarker detection algorithm and the 10 existing feature selection algorithms. The figure shows that 10 algorithms of R2 evaluation index calculate the classification accuracy of each feature subset under cross validation for 510 times and mark out the maximum accuracy. The horizontal axis lists the names of 10 classification algorithms.

As shown in Figure 8, the red regular triangle scatter points represent students with excellent physical fitness. At least two of the three physical test datasets of such students are excellent or good. The yellow regular triangle scatter points represent students with average physical fitness. There are few excellent blue inverted triangles in the data, indicating the students with poor physical fitness. The majority of these students are medium and unqualified in the three

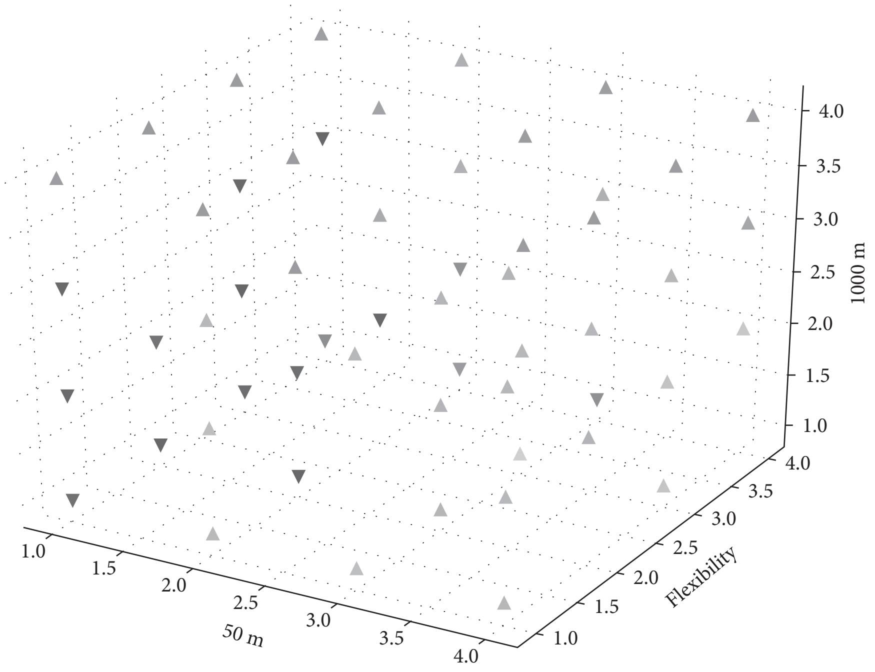
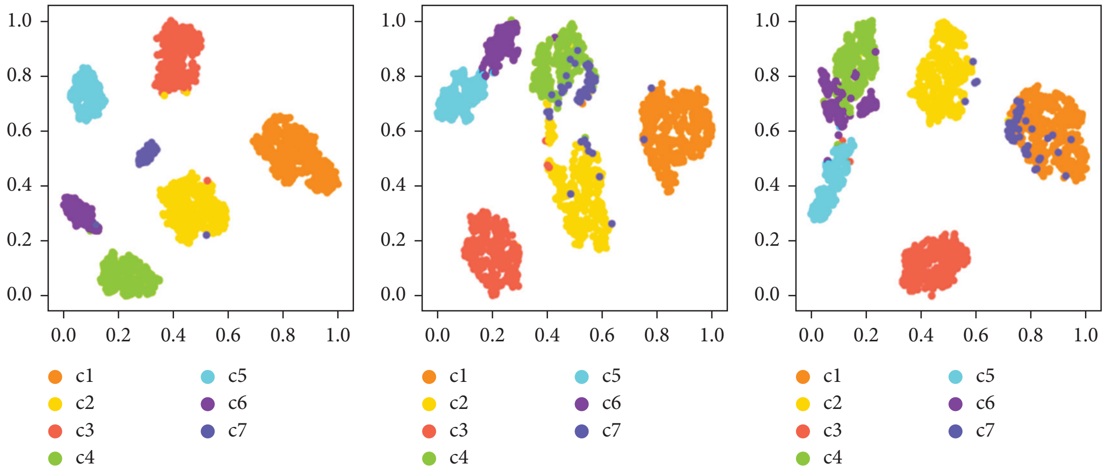


Figure 8: Health data clustering results.



(a)

(b)

(c)

Figure 9: Projection results of the proposed method on different datasets.

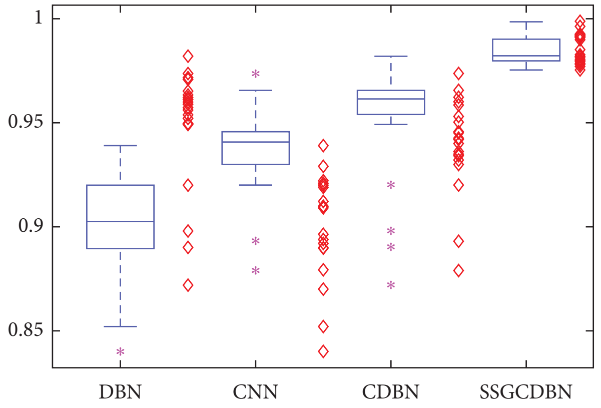


Figure 10: Boxplot of different health big data classification methods.

physical test datasets. Although one of them may be good or excellent, the overall physical fitness needs to be improved.

Figure 9 shows the projection results of the method in this paper on different data sets, including different categories represented by C. The three data sets of each category are classified into seven. It can be seen from the figure that the method proposed in this paper has good classification results on the three data sets, especially in data set 1, which shows that this method can deal with the big health problem of big data. The superiority of the proposed method is proved. In order to better prove the classification effect of the proposed method, 8000 datasets were processed and classified 20 times on average. Figure 8 shows the boxplot and scatter distribution of 20 mean diagnostic results of test samples under different models. As can be seen from Figure 10, the classification performance of the algorithm proposed in this paper is more stable and the classification accuracy is the highest.

5. Conclusion

In this paper, the relevant theories of Bayesian network are studied, and the classifier based on Bayesian network is applied to the data of hypothyroidism. Aiming at the key technologies needed in the application process, the improved ideas on methods are proposed and the specific contents are as follows.

The improved Bayesian network learning algorithm is applied to the classification of hypothyroidism. Firstly, the dataset of hypothyroidism is preprocessed to make it conform to the calculation requirements of the algorithm. Then, four Bayesian network classifiers are constructed for the preprocessed data, namely, naive Bayesian classifier (NBC), TAN classifier, BAN classifier, and Bayesian multi-network classifier. The network structures of different classifiers meet different degrees of dependence. Finally, BAN classifier was found to have the best effect on the classification of hypothyroidism data.

When diversity enters the transition stage, the combination operator method of fast convergence rate and competitive cross mutation can form good species quickly, but when diversity enters the mutation stage, it will be less. In order to avoid population convergence to local optimum, high-precision genetic combination operator and dynamic mutation rate method are used in this stage. Finally, experiments prove that the network structure of the improved algorithm is better.

Data Availability

The dataset can be accessed upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

The author acknowledges General Project of Hunan Philosophy and Social Science Foundation in 2017, Research on the integrated health service model of human medicine for the elderly in poor areas under the background of targeted Poverty Alleviation (no. 17YBA063).

References

[1] J. Sun, H. Wang, Z. Song, J. Lu, P. Meng, and S. Qin, "Mapping essential urban land use categories in Nanjing by integrating multi-source big data," *Remote Sensing*, vol. 12, no. 15, Article ID 2386, 2020.

[2] I. D. Dinov, B. Heavner, M. Tang et al., "Predictive big data analytics: a study of Parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations," *PLoS one*, vol. 11, no. 8, Article ID e0157077, 2016.

[3] F. Li, F. Li, S. Li, and Y. Long, "Deciphering the recreational use of urban parks: experiments using multi-source big data for all Chinese cities," *The Science of the Total Environment*, vol. 701, Article ID 134896, 2020.

[4] F. Lyu and L. Zhang, "Using multi-source big data to understand the factors affecting urban park use in Wuhan," *Urban Forestry and Urban Greening*, vol. 43, Article ID 126367, 2019.

[5] N. Niu, X. Liu, H. Jin et al., "Integrating multi-source big data to infer building functions," *International Journal of Geographical Information Science*, vol. 31, no. 9, pp. 1871–1890, 2017.

[6] J. Zhang, C. Li, Z. Sun, Z. Luo, C. Zhou, and S. Li, "Towards a unified multi-source-based optimization framework for multi-label learning," *Applied Soft Computing*, vol. 76, pp. 425–435, 2019.

[7] Y. Tu, B. Chen, W. Lang et al., "Uncovering the nature of urban land use composition using multi-source open big data with ensemble learning," *Remote Sensing*, vol. 13, no. 21, Article ID 4241, 2021.

[8] W. Hu, "On legal English translation from the perspective of legal linguistics," *Review of Educational Theory*, vol. 2, no. 3, pp. 6–10, 2019.

[9] X. Liu, N. Niu, X. Liu et al., "Characterizing mixed-use buildings based on multi-source big data," *International Journal of Geographical Information Science*, vol. 32, no. 4, pp. 738–756, 2018.

[10] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, pp. 1–9, 2018.

[11] X. Guo, H. Chen, and X. Yang, "An evaluation of street dynamic vitality and its influential factors based on multi-source big data," *ISPRS International Journal of Geo-Information*, vol. 10, no. 3, p. 143, 2021.

[12] D. Viorela-Valentina, "Translation practice–A means for enhancing student employability," *Dialogos*, vol. 22, no. 38, 215 pages, 2021.

[13] J. Prince, F. Andreotti, and M. De Vos, "Multi-source ensemble learning for the remote prediction of Parkinson's disease in the presence of source-wise missing data," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 5, pp. 1402–1411, 2018.

[14] A. R. Martinez, "Classification of covid-19 in ct scans using multi-source transfer learning," arXiv preprint http://arXiv.org/abs/2009.10474, 2020.

[15] Y. Zhang, Q. Li, W. Tu, K. Mai, Y. Yao, and Y. Chen, "Functional urban land use recognition integrating multi-source geospatial

data and cross-correlations," *Computers, Environment and Urban Systems*, vol. 78, Article ID 101374, 2019.

[16] P. Zhang, T. Li, G. Wang et al., "Multi-source information fusion based on rough set theory: a review," *Information Fusion*, vol. 68, pp. 85–117, 2021.

[17] E. M. Lalitha and L. Satish, "Wavelet analysis for classification of multi-source PD patterns," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 7, no. 1, pp. 40–47, 2000.

[18] T. Niu, Y. Chen, and Y. Yuan, "Measuring urban poverty using multi-source data and a random forest algorithm: a case study in Guangzhou," *Sustainable Cities and Society*, vol. 54, Article ID 102014, 2020.

[19] L. Zong, S. He, J. Lian et al., "Detailed mapping of urban land use based on multi-source data: a case study of lanzhou," *Remote Sensing*, vol. 12, no. 12, p. 1987, 2020.

[20] X. He, Y. Cao, and C. Zhou, "Evaluation of polycentric spatial structure in the urban agglomeration of the pearl river delta (PRD) based on multi-source big data fusion," *Remote Sensing*, vol. 13, no. 18, Article ID 3639, 2021.

[21] L. Yuan, Y. Wang, P. M. Thompson, V. A. Narayan, and J. Ye, "Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data," *NeuroImage*, vol. 61, no. 3, pp. 622–632, 2012.

[22] H. Yang, M. Fu, L. Wang, and F. Tang, "Mixed land use evaluation and its impact on housing prices in beijing based on multi-source big data," *Land*, vol. 10, no. 10, Article ID 1103, 2021.

[23] H. Kuai, N. Zhong, J. Chen et al., "Multi-source brain computing with systematic fusion for smart health," *Information Fusion*, vol. 75, pp. 150–167, 2021.

[24] Y. Guo, C. Yin, M. Li, X. Ren, and P. Liu, "Mobile e-commerce recommendation system based on multi-source information fusion for sustainable e-business," *Sustainability*, vol. 10, no. 1, p. 147, 2018.

[25] X. Xu, H. Peng, M. Z. A. Bhuiyan et al., "Privacy-preserving federated depression detection from multi-source mobile health data," *IEEE Transactions on Industrial Informatics*, vol. 1, 2021.

[26] X. Zhou, Z. Guan, J. Xi, and G. Wei, "Public transportation operational health assessment based on multi-source data," *Applied Sciences*, vol. 11, no. 22, Article ID 10611, 2021.

[27] H. Wang, "Marine environment salinity measurement based on data classification system and features of business English translation," *Arabian Journal of Geosciences*, vol. 14, no. 15, pp. 1–14, 2021.

[28] S. Jiang, X. Qian, T. Mei, and Y. Fu, "Personalized travel sequence recommendation on multi-source big social media," *IEEE Transactions on Big Data*, vol. 2, no. 1, pp. 43–56, 2016.

[29] C. Qiu, M. Schmitt, L. Mou, P. Ghamisi, and X. Zhu, "Feature importance analysis for local climate zone classification using a residual convolutional neural network with multi-source datasets," *Remote Sensing*, vol. 10, no. 10, Article ID 1572, 2018.

[30] Y. Li, G. Wen, Y. Hu et al., "Multi-source Seq2seq guided by knowledge for Chinese healthcare consultation," *Journal of Biomedical Informatics*, vol. 117, Article ID 103727, 2021.

[31] S. Liu, L. Zhang, Y. Long, Y. Long, and M. Xu, "A new urban vitality analysis and evaluation framework based on human activity modeling using multi-source big data," *ISPRS International Journal of Geo-Information*, vol. 9, no. 11, p. 617, 2020.

[32] K. Liu, Y. Feng, and X. Xue, "Fault diagnosis of hydraulic retraction system based on multi-source signals feature fusion and health assessment for the actuator," *Journal of Intelligent and Fuzzy Systems*, vol. 34, no. 6, pp. 3635–3649, 2018.

[33] H. Chen, L. Huang, L. Yang, Y. Chen, and J. Huang, "Model-based method with nonlinear ultrasonic system identification for mechanical structural health assessment," *Transactions on Emerging Telecommunications Technologies*, vol. 31, no. 12, Article ID e3955, 2020.

[34] A. B. Stella, M. AjČeviĆ, G. Furlanis et al., "Smart technology for physical activity and health assessment during COVID-19 lockdown," *The Journal of Sports Medicine and Physical Fitness*, vol. 61, no. 3, pp. 452–460, 2021.

[35] S. Fatima, O. Schieir, M. F. Valois et al., "Health assessment questionnaire at one year predicts all-cause mortality in patients with early rheumatoid arthritis," *Arthritis & Rheumatology*, vol. 73, no. 2, pp. 197–202, 2021.

[36] W. Cheng, H. Xi, C. Sindikubwabo et al., "Ecosystem health assessment of desert nature reserve with entropy weight and fuzzy mathematics methods: a case study of Badain Jaran Desert," *Ecological Indicators*, vol. 119, Article ID 106843, 2020.