A microcosmic syndrome differentiation model for metabolic syndrome with multi-label learning

Shujie Xia
Fujian University of Traditional Chinese Medicine
Jia Zhang
Xiamen University
Guodong Du
Xiamen University
Shaozi Li
Xiamen University
Zhaoyang Yan
Fujian University of Traditional Chinese Medicine
Jiliang Xin
Fujian University of Traditional Chinese Medicine
Long Zhu
Fujian University of Traditional Chinese Medicine
Bizhen Gao
Fujian University of Traditional Chinese Medicine
CanDong Li (warm_wind888@163.com)
https://orcid.org/0000-0002-6302-2675

Research

Keywords: metabolic syndrome, syndrome elements, microcosmic syndrome differentiation, ML-kNN, diagnostic model, machine learning

Posted Date: June 15th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-35201/v1

A microcosmic syndrome differentiation model for metabolic syndrome with multi-label learning

ShuJie Xia[1], Jia Zhang[2], GuoDong Du[2], ShaoZi Li[2], ZhaoYang Yang[1], JiLiang Xin[1], LongZhu[1] , BiZhen Gao[1*],CanDong Li [1*]

1. Research Base of Traditional Chinese Medicine Syndrome, Fujian University of Traditional Chinese Medicine, Fuzhou 350122, China
2. Department of Artificial Intelligence, Xiamen University, Xiamen 361005,China

ORCID: CL:0000-0003-1799-4315; GB: 0000-0002-3792-7761

Corresponding Authors:

Professor Candong Li MD., Ph.D.

Research Base of Traditional Chinese Medicine Syndrome,

Fujian University of Traditional Chinese Medicine, Fuzhou 350122, China

No.1 Qiuyang Road, Shangjie town, Minhou District, Fuzhou 350122, China

Tel: +86-591-2286-1989; Fax: +86-591-2286-1989

E-mail: fjzylcd@126.com

Or

Professor Bizhen Gao, Ph.D.

Research Base of Traditional Chinese Medicine Syndrome,

Fujian University of Traditional Chinese Medicine, Fuzhou 350122, China

No.1 Qiuyang Road, Shangjie town, Minhou District, Fuzhou 350122, China

Tel: +86-591-2286-1989; Fax: +86-591-2286-1989

E-mail: gbz688@163.com

ABSTRACT

Background: Metabolic Syndrome (MS) is a complex multi-system disease. Traditional Chinese medicine (TCM) is satisfactory in preventing and treating MS. Syndrome differentiation is the basis of TCM treatment, which is composed of location and/or nature syndrome elements. Many studies have found TCM syndromes are correlated with biological indicators. However, there are various data that mainly come from point-to-point studies, which can't fully summarize the many to many relationship of TCM syndromes and biological indicators. Thus, it is also hard to deal with the problem that several types of syndromes may possibly happen to a patient at once in the real world. The purpose of this study is to find out the potential relationship between microcosmic index and syndrome elements from a holistic view by means of multi-label learning (MLL) technique, and to provide a multi-label model for TCM syndrome differentiation.

Methods: The standardization scale on TCM four diagnostic information for MS is designed, which is used to obtain the results of TCM diagnosis. The model of microcosmic syndrome differentiation is constructed based on 39 physicochemical indexes, such as BMI, abdominal circumference, blood pressure，platelet，fasting blood sugar, insulin, blood lipid，by MLL techniques, called ML-kNN. First, the multi-label learning method is compared with three commonly used single learning algorithms. Then, the comparison of results of ML-kNN between physicochemical indexes and TCM information is also made. Next. the influence of parameter k of ML-kNN to the diagnostic model is investigated and then the best k-value is chosen for the TCM diagnosis.

Results: A total of 698 cases are collected for the modeling of the microcosmic diagnosis of MS. The comprehensive performance of the ML-kNN model works obviously better than the others, whose average precision of diagnosis reach 71.4%. The results of ML-kNN based on microcosmic

indexes are close to the results based on TCM information. The k value has less influence on the prediction results of ML- kNN.

Conclusions: The MLL techniques facilitate building microcosmic syndrome differentiation model in MS and the experiments show this is a practical approach to solve the problem of labeling multiple syndromes simultaneously. Besides, it is also suggested that there is many to many relationships between TCM syndrome elements and physicochemical indexes, which will be conducive to the future objective and comprehensive study of syndrome differentiation in MS.

Keywords: metabolic syndrome; syndrome elements; microcosmic syndrome differentiation; ML-kNN; diagnostic model; machine learning

Background

Metabolic Syndrome (MS) is a metabolic disorder syndrome characterized by obesity, hyperglycemia, hypertension, dyslipidemia and hyperuricemia, which seriously endanger patient's health[1]. This is a complex multi-system disease, and the difficulty lies in its early intervention and multi-target therapy[2]. TCM has the characteristics of multi-target, small side effects and overall regulation, so its advantages in the prevention and treatment of MS have become more obvious. There are many studies showing that the therapeutic effects of MS in TCM are fairly satisfying[3]. Syndrome differentiation is the basis of TCM treatment, which is based on TCM theories and the doctor's experiences according to the patient's four diagnosis information of TCM[4]. However, the traditional way of syndrome differentiation inevitably exists the problems of subjectivity and fuzziness, which actually hinders the application prospects of TCM[5].

In order to achieve an effective and objective standard of syndrome differentiation, many researchers have searched the inherent relationship between symptoms and syndromes by using machine learning and data mining methods. For example, to deal with the problems of high nonlinearity and complex interaction of different symptoms[6,7], many machine learning methods, such ask nearest neighbor (kNN), support vector machine(SVM), neural networks(NNs), bayesian networks(BN), and decision tree(DT), are applied to the TCM diagnosis. To be specific, some authors introduced the method of SVM for the hypertension diagnosis in TCM, and the experimental results demonstrated that using the SVM algorithm to model TCM syndrome diagnosis not only can obtain high accuracy, but also has the methodological feasibility [8]. Some authors used BN for the clinical analysis and find that it constructed an association mode of symptoms and phlegm-heat congesting lung syndrome. Given that TCM Syndromes can be divided into location and nature syndrome elements[10]. Various mathematical models are employed to quantify the four diagnostic information[11], thus achieving syndrome element diagnostic task, such as factor analysis and logistic regression analysis in the TCM diagnosis of ulcerative colitis [12].

With the development of modern medicine, many studies have found that TCM syndromes are confronted with the multidimensional complexity and also correlated with multiple microcosmic indexes, and these indexes can make the process of syndrome differentiation more objective and standardized[13,14]. More importantly, microcosmic indexes can be used to assist differentiation syndrome, especially when the symptoms are not discernible. There are some works which consider to exploring the relationship between different syndromes and physicochemical indexes. For example, the CHAID decision tree was used to build a recognition mode of phlegm-heat stasis syndrome according to the indexes of clinical routine examinations in unstable angina (UA). Its

results showed that the CHAID decision tree model has certain advantages for the disease recognition[15]. Another study investigated whether these biomedical indexes could be beneficial to TCM syndrome diagnostics in Chronic Hepatitis B (CHB) patients. Its results showed that the performance of syndrome classification based on proper integration of TCM and modern clinical indexes was significantly higher than those based on one view of parameters only. Besides, through the correlation analysis and clinical verification test, they found potential associations between symptoms and clinical indicators, which hinted to the mitogen-activated protein kinase（MAPK） signaling pathway[16].

However, the aforementioned learning methods are aimed at solving the problem of single syndrome diagnosis, i.e., single-label learning, which can't deal with the problem of multiple syndromes occur at the same time, i.e., multiple labels simultaneously. In clinical practice, many symptoms are associated with various syndromes[17]. The previous studies[18, 19] have shown that many MS patients always have two types even more than two types of syndrome elements. Compared with conventional learning methods, multi-label learning is more capable to identify syndrome information in TCM, i.e., it can solve the learning problem of one patient with several syndromes. The research idea of this article can be seen in Fig.1. In this paper, the TCM informations was obtained to confirming syndrome diagnosis according to syndrome element differentiation and 39 physicochemical indexes are collected as predictors for diagnostic task. Based on the collected physicochemical data set, the microcosmic syndrome differentiation model for multiple syndromes is constructed by multi-label learning, which has proved to be feasible and effective. Besides, the influence of the location and the nature syndrome elements on modeling is also assessed.

The rest of this paper is organized as follows. In the Method Section, we describe the dataset,

syndromes selection method, and the multi-label learning method. In the Results Section, the results of all the models based on ML-kNN，kNN，DT，SVM are analyzed and the effect of these models is assessed. In the Discussion Section, the reason why multi-label learning could improve the results is clarified. Eventually, we summarize the paper in the Conclusion Section.

Materials and Methods

Patients

The inpatients and outpatients with MS were selected in the Second People's Hospital Affiliated to the Fujian University of TCM, the Third People's Hospital Affiliated to Fujian University of TCM, the Fuzhou Second Hospital and the Jinjiang Hospital of TCM from 2015 to 2019. All participants signed consent forms. Ethics approval for the present study was given by the Medical Ethics Committee of Fujian University of Traditional Chinese Medicine.

According to the "syndrome differentiation significance of 600 common symptoms" and the MS common symptoms standard in the Guideline of Clinical Research of TCM New Drugs, the four diagnosis information collection scale is established. The symptoms and signs were classified as none, mild, moderate and severe with 0, 1, 2 and 3 points respectively. The four diagnostic data are collected by two qualified professionals of TCM. The physicochemical indexes include the following indexes: body weight, height, blood pressure, abdominal circumference, blood routine, fasting blood sugar, insulin, blood lipid, liver function, kidney function, etc.

Diagnostic criteria

Diagnostic criteria of Western medicine: According to the MS diagnostic criteria issued by the IDF and AHA and Diabetes Society of the Chinese Medical Association, the diagnosis can be made if the following 3 items or more are met: (1) abdominal obesity (waist circumference: male ≥90 cm, female ≥85 cm); (2) hyperglycemia: FPG ≥ 6.1 mmol/L or 2 HPG ≥ 7.8 mmol/L and/or diagnosed as diabetes; (3)hypertension: BP≥130/85 mmHg and/or diagnosed as hypertension (4) TG≥ 1.70mmol/L; (5) HDL-C < 1.04mmol/L. TCM syndrome elements diagnosis standard: According to the Syndrome Part of TCM Clinical Diagnosis and Treatment Terminology (the National Standard of the People's Republic of China), Syndrome Elements Differentiation[10] and combined with the expert consultation.

Inclusion and Exclusion criteria

Inclusion criteria of the patients are 1) patients who meet the diagnostic criteria of MS; 2) The patients who are given the informed consent; 3) patient's age ranges from 18 to 75.

Exclusion criteria are: 1) patients with mental diseases or other severe diseases; 2) The patients who could not express their feelings clearly; 3) patients who refuse to participate in our study or without informed consent.

Data processing

Selection of syndrome elements: We put the four diagnostic information into the TCM identification system to obtain the location and nature of syndrome elements, and the results are examined by three professional doctors. Combining with the literature research, 15 of the most common syndrome elements of MS are screened out, among which the location syndrome elements

include spleen, liver, kidney, lung, heart and stomach, and the nature syndrome elements include phlegm, dampness, Yin deficiency, Yang deficiency, Qi stagnation, Qi deficiency, blood stasis and blood deficiency.

Data normalization: First, dimensionality reduction of the original data is done to eliminate some of the diagnostic information that have never appeared in the data set. Second, we fill up individual missing data, and to eliminate some cases that has too much incomplete data. Finally, in order to facilitate data processing and ensure the faster convergence of the running program, the unified normalization of the data is carried out.

Computational methods: We construct the model to map the relationship between microcosmic indexes and syndromes elements by means of ML-kNN algorithm. ML-kNN is the extension of the kNN algorithm for multi-label learning, which consider searching for k nearest instances with their labels for each test instance, and thus predicting the labels of the test instance. KNN is only suitable for processing the case with single labels, whereas the data set for differentiation diagnosis of MS has the characteristic of the multi-label. For this reason, the multi-label learning algorithm, such as ML-kNN [20], is developed so as to better reveal the correlation of the labels. Suppose there are p samples in the test set, the ML-kNN algorithm is shown below.

Step 1: Suppose *t* is an instance in the test set, calculate the distances of test instance *t* to all training instances, and find the k instances of training data with the smallest distance;

Step 2: According to the labels of the k nearest instances, calculate the number of each label in the k nearest instances for test instance *t*;

Step 3: According to the statistical result, predict the *t*'s posterior probability on each label;

Step 4: Output the posterior probability of all the labels with naive bayes method;

Step 5: Repeat Step1 to Step4 until the prediction of all the test instances are finished;

Step 6: Assess the predicted results according to multi-label evaluation criteria.

Experimental design and evaluation

Specifically, this work uses the multi-label learning approach to establish the microcosmic syndrome differentiation for metabolic syndrome in TCM. To attain this purpose, we utilize the collected TCM data and physicochemical data from three TCM hospitals in China. The TCM data is to get the diagnosis of syndrome elements and the physicochemical indexes to training the diagnosis model. Furthermore, effective data pre-processing methods are performed, including missing data processing, data normalization, and data combination. Then, we use the ML-kNN algorithm to explore the potential relationship and rules between physicochemical indexes and syndrome elements and thus establish a microcosmic syndrome diagnosis model. Finally, we use four evaluation metrics to assess the effectiveness of our proposed method with other state-of-the-art approaches (Fig. 2).

In order to ensure the adequacy of training, 10-fold cross-validation is used for evaluating the performance systematically. To be specific, 90% of the samples are randomly selected as the training set and the remaining 10% as the test set. As the validation is iterated 10 times, the average value is taken as the final prediction. All the predicted results will be compared with three traditional algorithms, namely KNN, DT, and SVM. Note that the evaluation metrics involved in the experiments include Hamming loss (HL), Coverage, Ranking loss (RL), Average precision (AVP), and the performance of multi-label algorithm can obtain the objective evaluation data by these metrics[21, 22]. For AVP, the higher the average accuracy is, the better the model is, and for the others the smaller the evaluation indexes are, the better the model is.

Results

Basic information about MS patients

A total of 767 cases are obtained in the study. Among the patients, 396 patients are male (51.6%) and 371 patients are female (48.4%), and the average age is 44.95 years. After deleting data with too much missing data and obvious errors, there are 698 samples selected to conduct the experimental design, each of them contains four diagnostic information, syndrome identification results and related microcosmic indexes. The score distribution of syndrome elements according to syndrome element differentiation can be seen in Fig. 3 and the baseline characteristics of those MS patients in table 1.

Table 1. Baseline Characteristics of MS Patients

Main index	**Male(365)**		**Female(333)**		**Total(698)**	
	Mean	SD	Mean	SD	Mean	SD
Age (year)	42.89	10.85	47.39	11.13	45.04	11.21
BMI	30.80	8.77	30.56	9.72	30.69	9.23
WC (cm)	97.67	6.76	95.46	8.16	96.61	7.54
SBP (mmHg)	133.74	17.28	133.88	19.53	133.81	18.38
DBP (mmHg)	87.70	11.60	84.49	11.72	86.17	11.76
TG(mmol/L)	3.13	3.32	2.36	2.50	2.76	2.98
HDL(mmol/L)	1.20	0.51	1.60	4.33	1.39	3.02
FBG(mmol/L)	10.95	4.75	10.37	4.30	10.67	4.55

WC, waist circumference; SBP, systolic blood pressure; DBP,diastolic blood pressure; TG, triglyceride; HDL, high density lipoprotein; FBG, fasting blood glucose.

Performance comparison of microcosmic syndrome models

We use the model of ML-kNN to conduct the experiment on the collected data set, which is built as described in the Methods Section, and we set the parameter K of ML-kNN to be 5. Furthermore, we compare ML-kNN with kNN, DT and SVM algorithm. Each prediction model is verified by ten folds cross validation, and the results are shown in Fig. 4, where the horizontal coordinate represents the number of cross validation, and the longitudinal coordinate stands for the average precision as 100% is the highest value.The comparison results of ML-kNN, kNN, DT and SVM on all the five evaluation metrices are shown in Table 2.

From Fig. 4 and Table 2, we can get: (1) Overall, the average precision of ML-kNN is always higher than kNN, DT and SVM in the process of ten fold cross validation; (2) In the view of Average Precision, ML-kNN is 0.714, which is higher than the comparing methods (kNN, DT and SVM are 0.217, 0.226 and 0.16, respectively). (3) In consideration of Hamming Loss, ML-kNN is 0.233, and has the lower classification error than kNN, DT and SVM. Although the difference is small, ML-kNN is the smallest one; (4) Considering the other evaluation metrics, i.e., Ranking Loss and Coverage, ML-kNN is still superior to kNN, DT and SVM.

Table 2 Evaluation of Prediction Results of ML-kNN, kNN, DT and SVM using microcosmic indexes and TCM information respectively

Evaluation criteria	ML-kNN	kNN	DT	SVM
Average Precision	0.714±0.024*	0.497±0.028	0.488±0.036	0.554±0.039
Hamming Loss	0.233±0.021*	0.297±0.030	0.308±0.020	0.236±0.028
Ranking Loss	0.169±0.012*	0.698±0.053	0.678±0.044	0.706±0.046
Coverage	5.123±0.476*	7.512±0.894	7.866±0.796	7.648±0.743

* Representing the index in this model is the best compared with others.

The comparision of forecast results bewteen microcosmic indexes and TCM information

Generally, the results of TCM syndrome differentiation are obtained through the comprehensive analysis of the TCM information, namely inspection, auscultation, interrogation and palpation. In order to further illustrate the diagnostic value of microscopic indexes, we made a comparision of prediction results of ML-kNN between those physicochemical indexes (PI) and TCM information (TI) .

From Fig. 5, it can be concluded that: (1) On the whole, although the prediction results of PI were slightly lower than those of TI, they were generally close to each other. (2) To be specific, for Average Precision(Fig. 5(A)), the result of PI and TI are 0.714 and 0.753 respectively. For Hamming Loss(Fig. 5(B)), the result of PI and TI are 0.233 and 0.188 respectively. For Ranking Loss(Fig. 5(C)), the result of PI and TI are 0.169 and 0.144 respectively. For Coverage(Fig. 5(D)), the result of PI and TI are 5.123 and 4.731 respectively. From the above, the results of microcosmic indexes using ML-kNN algorithm is close to that of TCM information.

The effect of different K values on the results

In order to determine whether the K value affects the predicted results of ML-kNN, we construct the model with k values of 1, 3, 5, 7, 9, 11, 13, 15, 17 and 19 respectively. Then the predicted results are shown in Fig. 5. Among them, the horizontal coordinate represents the K value, and the longitudinal coordinate stands for the average results of Average Precision, Hamming Loss, Ranking Loss and Coverage respectively.

From Fig. 6, it can be seen that: (1) The prediction results of ML-kNN fluctuate with the change of K values, but the fluctuation is small. (2) Generally speaking, when k value is smaller, Average

Precision is higher, accompanied with lower Hamming Loss, Ranking Loss, and Coverage. (3) For this model, when the k value is 5, Average Precision is higher and Error Precision get the best results. When k is 1 or 3, the result of Hamming Loss is optimal. When k is 1 or 7, Ranking Loss or Coverage get the best results.

The influence of syndrome elements selection on forecast results

Different types of labels may lead to different predicted results. In order to further explore the influence of label selection on the micro-diagnosis model of syndrome elements, the average precision of the muti-label model of the location and the nature of syndrome elements is analyzed.

From Fig. 7, it can be concluded that: (1) On the whole, ML-kNN model has the highest prediction precision for the location and the nature of syndrome elements, which are 0.728 and 0.776 respectively, while kNN, DT and SVM have relatively lower prediction performance. (2) For ML-kNN, the location syndrome elements are better than the nature syndrome elements on predicting results, while for kNN and SVM, the nature syndrome elements are better than the location syndrome elements on predicting results, but for DT, the difference between the location syndrome elements and the nature of syndrome elements is not obvious in terms of the average precision.

Discussion

Application of multi-label learning model in TCM diagnosis

Traditional Chinese Medicine is an empirical medicine, and the discovery of the underlying objective law can contribute to the development of Chinese medicine. The law is often contained in a large number of information on TCM diagnoses and treatments, therefore, how to extract the true and potential law from considerable cases becomes an important problem in the field of TCM. With the introduction of machine learning and data mining technologies, people have found a new way for TCM diagnosis research. TCM syndrome takes complex life system as its object and has high-dimensional characteristics[23]. For example, it may correspond to multiple syndrome elements (such as phlegm, Qi deficiency, spleen, etc.). Furthermore, there is a primary-secondary relationship between the syndrome elements, so it is unreasonable to use a single name for TCM diagnosis in MS. From the view of computer science, this is a typical multi-label learning problem. In this paper, the multi-label algorithm is applied to TCM in order to establish a diagnostic model, which is suitable for syndrome differentiation. Through learning based on the training data, it find the potential correlative rules between multiple TCM information and their syndrome elements as well as rules between physicochemical indexes and syndrome elements . By judging the probability of multiple predicted syndromes, the problem of identificating the primary and secondary syndrome elements at the same time can be solved.

ML-kNN model is a typical multi-label learning algorithm in recent years, which is originally proposed by Zhang and Zhou[20], and developed based on the traditional single-label kNN model combined with Bayesian algorithm. It has the advantages in consideration of the simplicity, feasibility and low error rate[24,25]. Considering that symptoms and signs or microcosmic indicators

often do not appear singly, and the syndrome elements are also related to each other, ML-kNN model estimates the posterior probability of multiple labels related to the final diagnosis as a whole[26], so it is more in accord with the core of TCM holistic concept. This study shows that the prediction results of ML-kNN are superior to the traditional single label algorithms kNN, DT and SVM, whose average precision is 0.714, Hamming loss 0.233, Ranking loss 0.169 and Coverage 5.123. Among them, the K value has little influence on the prediction results, which indicates that the model is stable.

Microcosmic syndrome differentiation of TCM

Microcosmic syndrome differentiation is the deepening and expansion of traditional syndrome differentiation of TCM. Nowadays, many biology researches of TCM syndromes often adopt the theory of reduction analysis. Starting from the correlation between some biological indicators and TCM syndromes, the syndromes are attributed to the abnormality of a certain substance. Some researchers expected to find a scientific basis or specificity index by using point-to-point research methods. However, in practice, it is difficult to find specific indicators for syndromes [27]. The reason is that TCM syndromes put a high value on interrelation and aim to grasp one's vital signs from the macro level. Its material basis is hard to be interpreted by single or specific substances. Therefore, in order to explore its inherent law and biological significance of TCM, we should follow the principle of holism and systematicness, and conduct the analysis comprehensively from the relationship and combination of different substances.

From this perspective, the multi-label algorithm can effectively explore the relationship between macroscopic syndromes and microcosmic indicators. The results also show that the average precision

of predicting common syndromes of MS can reach 71.4% by using physicochemical indexes. What is more, the prediction performance of physicochemical indexes is close to that of TCM information when using ML-kNN. It indicates that using microcosmic indexes is also helpful to syndrome differentiation for MS. Besides, the average precision of predicting location syndromes can reach 72.8%. The average precision of predicting nature syndromes can reach 77.6%. This shows that physicochemical indexes have better performance in predicting the possibility of nature syndrome elements.

In short, this study promotes the development of microcosmic syndrome differentiation of TCM from a new perspective. In future work, it is supposed to expand the sample size and biomedical indexes for microcosmic differentiation. However, the core of TCM syndrome differentiation is inseparable from the macroscopic symptoms and signs. It is the most reasonable to combine the microscopic indexes with the macroscopic symptoms and signs to realize the accurate and comprehensive differentiation of syndromes. So, this study also provides the research basis for the combination of multiple categories of indicators to syndrome differentiation

Conclusion

This study deals with the problem of microcosmic syndrome differentiation of MS by using multi-label learning. Through the comprehensive analysis of the evaluation metrics and the comparison of the four diagnostic models, i.e., ML-kNN, kNN, DT, and SVM, it is found that ML-kNN has better prediction results than all the single-label models. Thus, we can conclude that the multi-label learning method like ML-kNN helps to obtain good results in terms of MS syndrome

differentiation. It overcomes the defect of conventional single-label learning methods and turns out to be an effective technique for solving the problems in the clinical practice of TCM. The preferable prediction performance of microcosmic indexes suggests that there is a potentially complex connection between TCM syndromes and physicochemical indexes is conducive to the future study of objective and comprehensive syndrome differentiation in MS .

Declarations

Abbreviations

MS:Metabolic Syndrome; TCM: Traditional Chinese medicine; MLL: multi-label learning; kNN: k nearest neighbour; SVM:support vector machine; NNs:neural networks; BN:bayesian networks; DT: decision tree; UA: unstable angina;CHB:Chronic Hepatitis B;MAPK:mitogen-activated protein kinase; HL:Hamming loss; RL:Ranking loss; AVP:Average precision; PI: Physicochemical index; TI: TCM information

Acknowledgements

Not applicable.

Authors'contributions

All authors contributed to this study.SJX, CDL and BZG conceived anddesigned the experiments;JLX and PZ performed the experiments; JZ, GDD and SZL contributed to statistical analysis of the data; and SJX, ZYY, JZ wrote the manuscript. All authors have read and approved the manuscript.

Funding

This work is partially supported by Joint Fund Project of National Natural Science Foundation of China (U1705286), National Natural Science Foundation of China(81873234),Major Project of Standardization of State Administration of Traditional Chinese Medicine of the PRC(GZY-FJS-2017-64). The funding organizations for this study were not

involved in any aspect of the study design, enrollment of participants, collection and analysis of data,interpretation of the results, nor the writing of the manuscript.

Availability of data and materials

The data used to support the findings of this study are available from the corresponding author upon request.

Ethics approval and consent to participate

The study was approved by the Ethics Committee of Fujian university of traditional Chinese medicine [FTCMER 2012. 004] and all patients signed informed consent.

Consent for publication

All authors have consent for publication.

Competing interests

All authors declare that they have no competing interests.

Author details

The author is a medical doctor of internal medicine of TCM in Fujian university of Traditional Chinese Medicine. She won the first place in national professional competitions five times. Her research direction is the prevention and treatment of endocrine and metabolic diseases by Traditional Chinese Medicine.

Reference

1. Alberti K, Zimmet P, Shaw J: **The metabolic syndrome - a new worldwide definition**. *Lancet* 2005, **366**(9491):1059-1062.

2. Zhang YY, Li YC, Ni Q: **Diagnosis and treatment of metabolic syndrome**. *Chinese Clinical Doctor* 2018, **46**(11):15-18.

3. Wang Q, Liu L: **Advances in the treatment of metabolic syndrome by traditional Chinese medicine**. *Information on Traditional Chinese Medicine* 2017, **34**(06):142-145.

4. Li CD: **Chinese Diagnostics**. Beijing: China Press of Traditional Chinese Medicine; 2016.

5. Xia SJ, Li CD: **Traditional Chinese Medicine Diagnosis Modernization and Health Status Identification**. *Journal of Traditional Chinese Medicine* 2019, **60**(04):3.

6. Zhang X, Shi QL, Wang B, et al: **Review of Machine Learning Algorithms in Traditional Chinese Medicine**. *Computer Science* 2018, **45**(11A):5.

7 Du G, Ma L, Hu J S, et al. Prediction of 30-Day Readmission: An Improved Gradient Boosting Decision Tree Approach[J]. Journal of Medical Imaging and Health Informatics, 2019, 9(3): 620-627.

8. Xu M, Ma X, Wen Z, et al: **Application of support vector machine in the diagnosis of hypertension in TCM syndrome**. *China Journal of Traditional Chinese Medicine and Pharmacy* 2017, **32**(6):4.

9. Wang ZW, Li JS, Chen CX, et al: **Constructing Symptom and Syndrome Association Mode of Acute Exacerbation of Chronic ObstructivePulmonary Disease with Phlegm-heat Congesting Lung Syndrome: Based on Bayesian Network**. *Journal of Traditional Chinese Medicine* 2018, **59**(3):4.

10. Zhu WF: **syndrome elements differentiation**. Beijing: People's Health Publishing House; 2008.

11. Wang ZW, Ai J, Chen S, et al: **Mathematical Methods and Analysis of Science in TCM**. *Chinese Archives of Traditional Chinese Medicine* 2017, **35**(4):94-96.

12. Li Y, Liu Y, Xu YP, et al: **Study on Syndrome of Ulcerative Colitis Based on Factor Analysis and Logistic Regression Analysis**. *Liaoning Journal of Traditional Chinese Medicine* 2018, **45**(2):4.

13. Fang H, Luo TJ: **Research progress about relationship between traditional Chinesemedicine distinction of MS and lab examinations**. *Journal of Changchun University of Chinese Medicine* 2015, **31**(06):4.

14 Li S , Zhang Z Q , Wu L J , et al. **Understanding ZHENG in traditional Chinese medicine in the context of neuro-endocrine-immune network**[J]. *IET SYSTEMS BIOLOGY,* 2007, 1(1):1.

15. Shi Q, Zhao HH, Chen JX, et al: **Study on Recognition Modes of Phlegm-Heat Stasis Syndrome Decision Tree of Patients with Coronary Heart Diseases**. *World Chinese Medicine* 2018, **13**(9):12-18.

16. Kang H, Zhao Y, Li C, *et al*: **Integrating clinical indexes into four-diagnostic information contributes to the Traditional Chinese Medicine (TCM) syndrome diagnosis of chronic hepatitis B**. *Scientific Reports* 2015, **5(1)**:9395.

17 Wu, Y., Zhang F., Yang K., et al:. SymMap: an integrative database of traditional Chinese medicine enhanced by symptom mapping. *Nucleic Acids Res* 47: D1110-D1117, 2019.

18. Zhuang YD,Gao BZ, Wang YF, et al: **Study on correlations between disease location, concurrent nature of disease, insulin resistance and biochemical indicator in metabolic syndrome with phlegm syndrome**. *China Journal of Traditional Chinese Medicine and Pharmacy* 2017, **32**(02):3.

19. Yan X: **A study on the correlation between TCM syndrome elements, Heath-Carter body shape characteristics and the resistance of aishin in metabolic syndrome**. *doctor.* Beijing University of Chinese Medicine; 2016.

20. Zhang ML, Zhou ZH: **ML-KNN: A lazy learning approach to multi-label learning**. *Pattern Recognition* 2007, **40**(7):2038-2048.

21. Zhang J, Luo ZM, Li CD, et al: **Manifold regularized discriminative feature selection for multi-label learning**. *Pattern Recognition* 2019, 95: 136-150.

22. Zhang J, Li CD, Sun ZQ, et al: **Towards a unified multi-source-based optimization framework for multi-label learning**. *Applied Soft Computing* 2019, 76: 425-435.

23. Ma T, Tan C, Zhang H, et al: **Bridging the gap between traditional Chinese medicine and systems biology: the connection of Cold Syndrome and NEI network**. *Mol Biosyst* 2010, **6**(4):613-619.

24. Sun ZQ, Zhang J, Dai L, et al: **Mutual information based multi-label feature selection via constrained convex optimization**. *Neurocomputing* 2019, 329: 447-456.

25. Dai L, Zhang J, Li CD, et al:**Multi-label feature selection with application to TCM state identification.** *Concurrency and Computation: Practice and Experience* 2018, e4634.

26. Du G, Zhang J, Luo Z, et al. **Joint imbalanced classification and feature selection for hospital readmissions**[J]. *Knowledge-Based Systems,* 2020, 200: 106020.

27. Zhang YG, Zhang W, Yang YQ: **Application of Systematic Biology in TCM Research**. Beijing: Science Press; 2014.

Figure Legends

Fig.1 Process of TCM syndrome differentiation

Fig.2 Paradigm of the proposed method

Fig.3 The score distribution of syndrome elements

Fig.4 The fluctuation on average precision of four machine learning in the process of cross validation

Fig.5 The comparison of the prediction performances of ML-kNN using microcosmic indexes and TCM information respectively

Fig.6 The influence of ML-kNN algorithm using microcosmic indexes on prediction results with different k values

Fig.7 The influence of different syndrome elements on the average precision of ML-kNN using microcosmic indexes

Figures

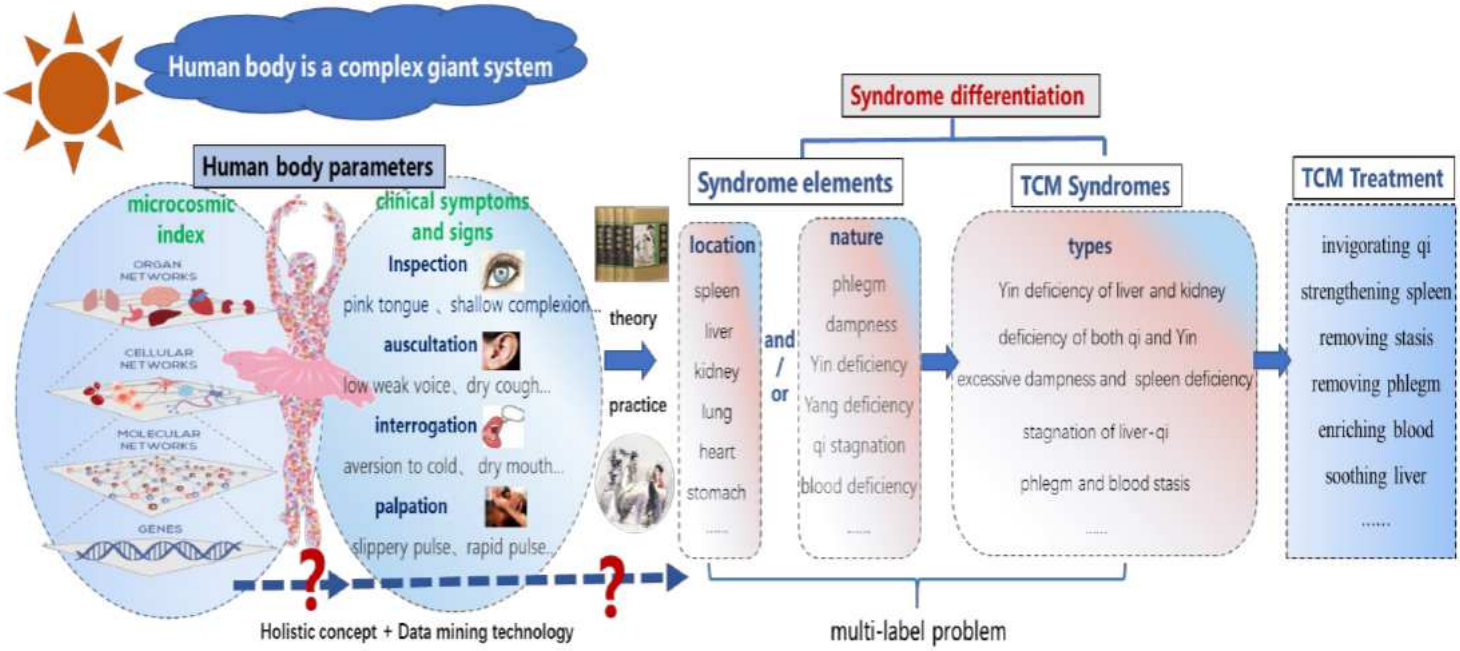

Figure 1

Process of TCM syndrome differentiation

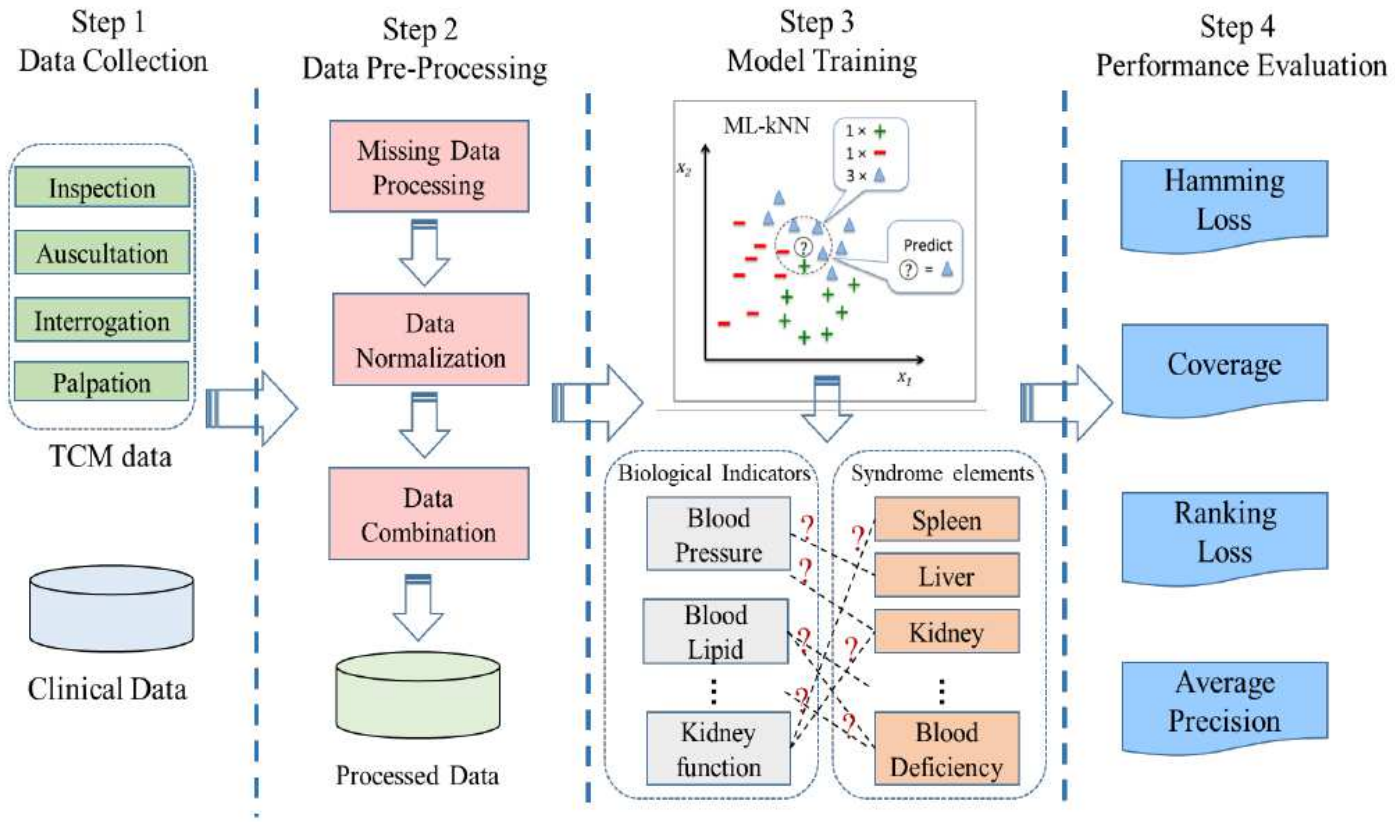

Figure 2

Paradigm of the proposed method

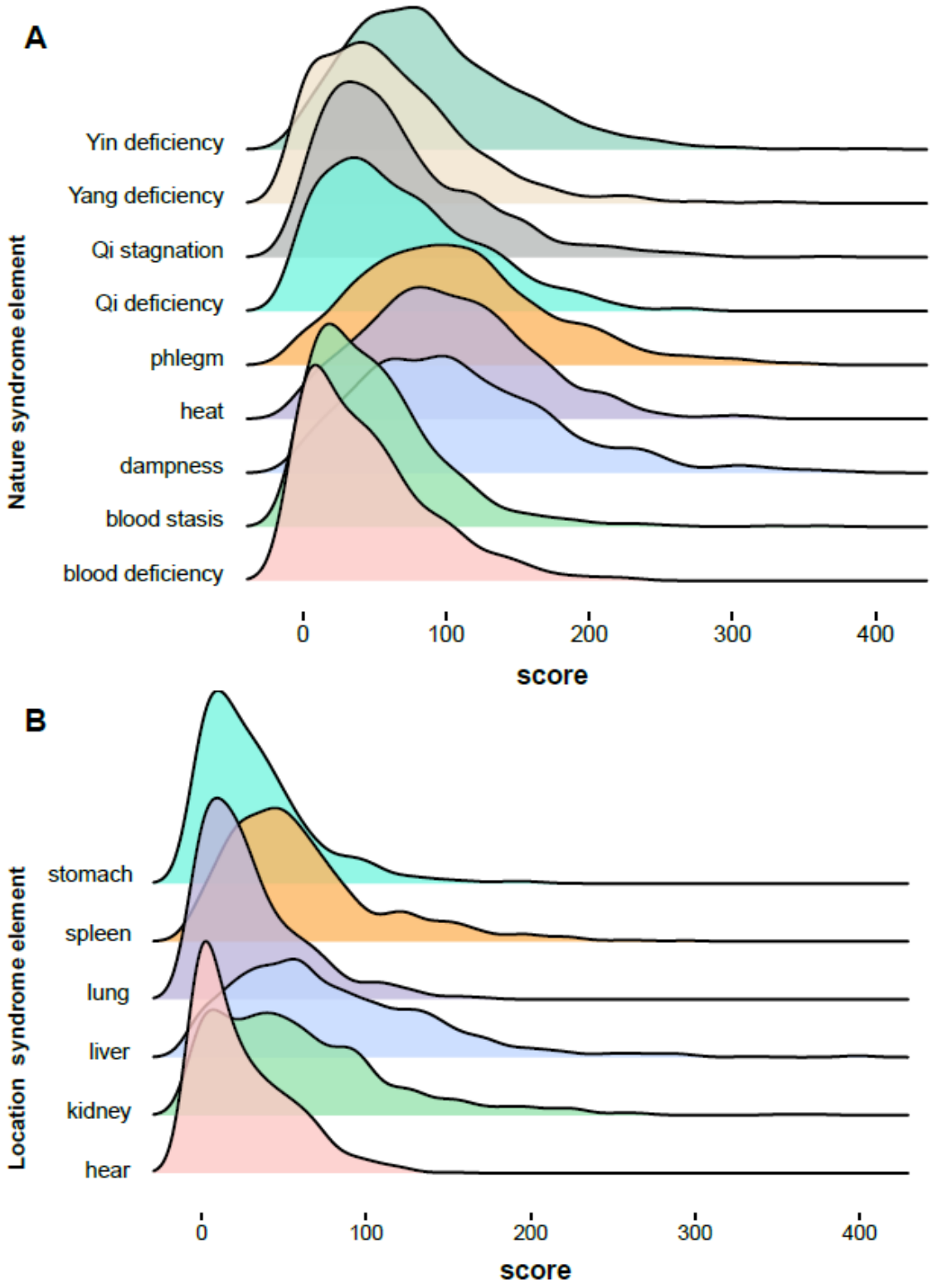

Figure 3

The score distribution of syndrome elements

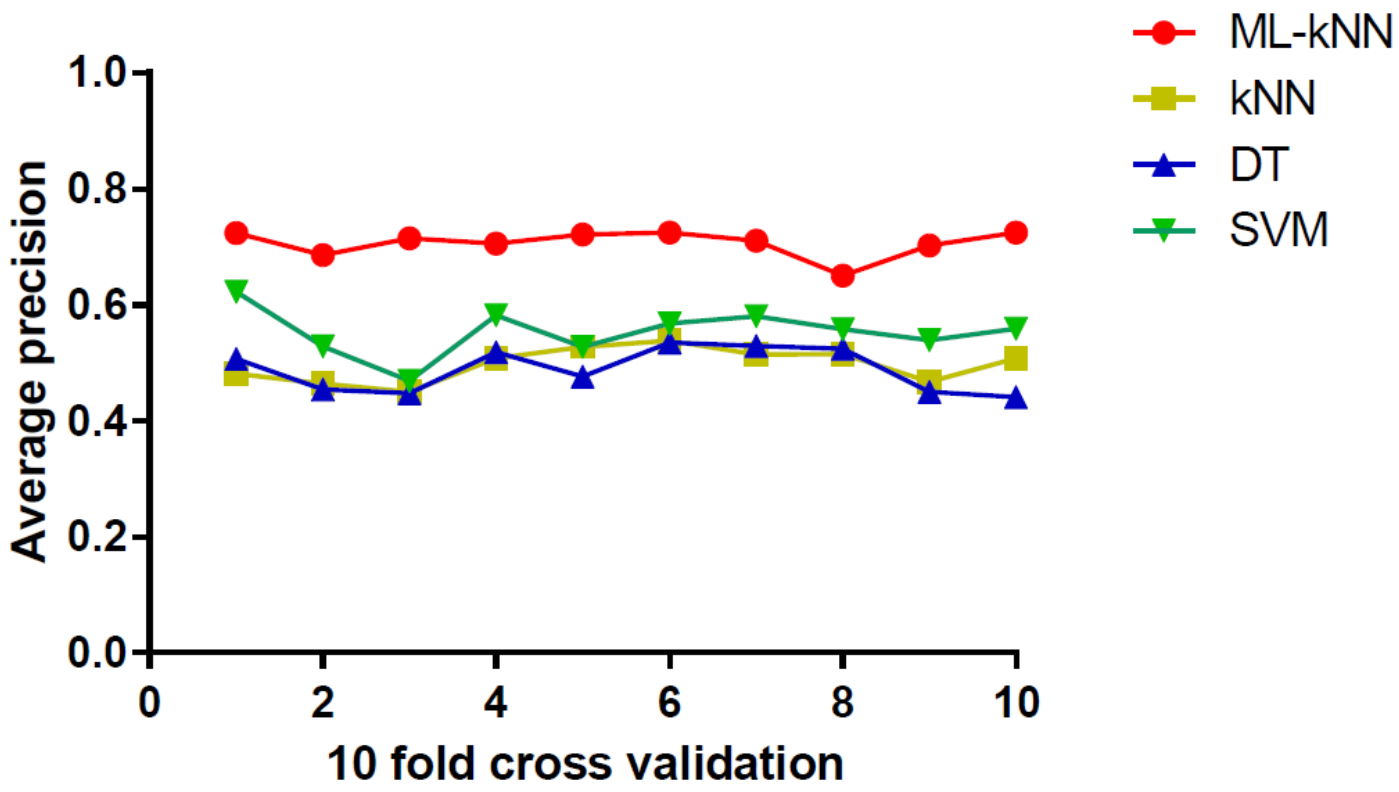

Figure 4

The fluctuation on average precision of four machine learning in the process of cross validation

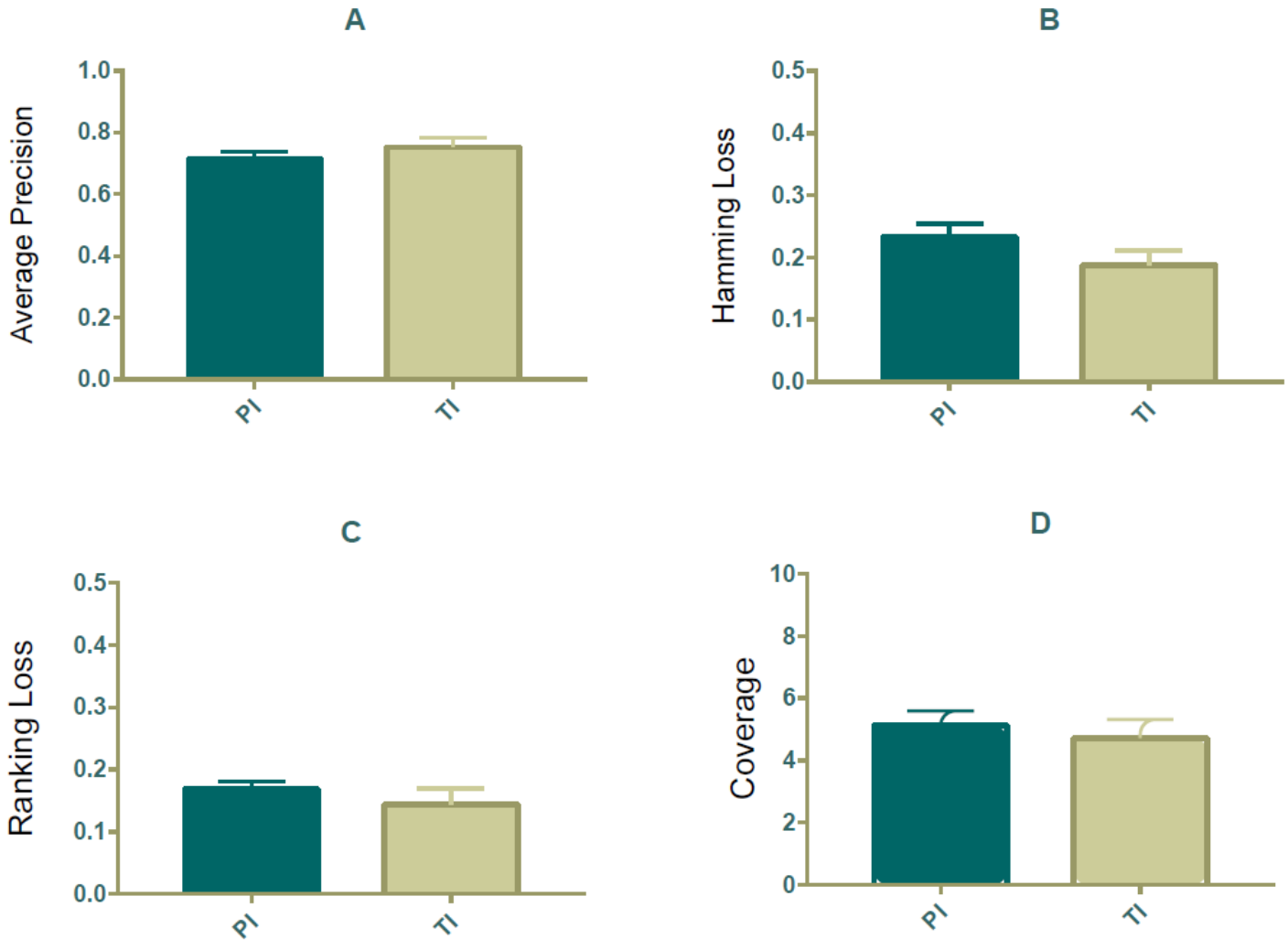

Figure 5

The comparison of the prediction performances of ML-kNN using microcosmic indexes and TCM information respectively

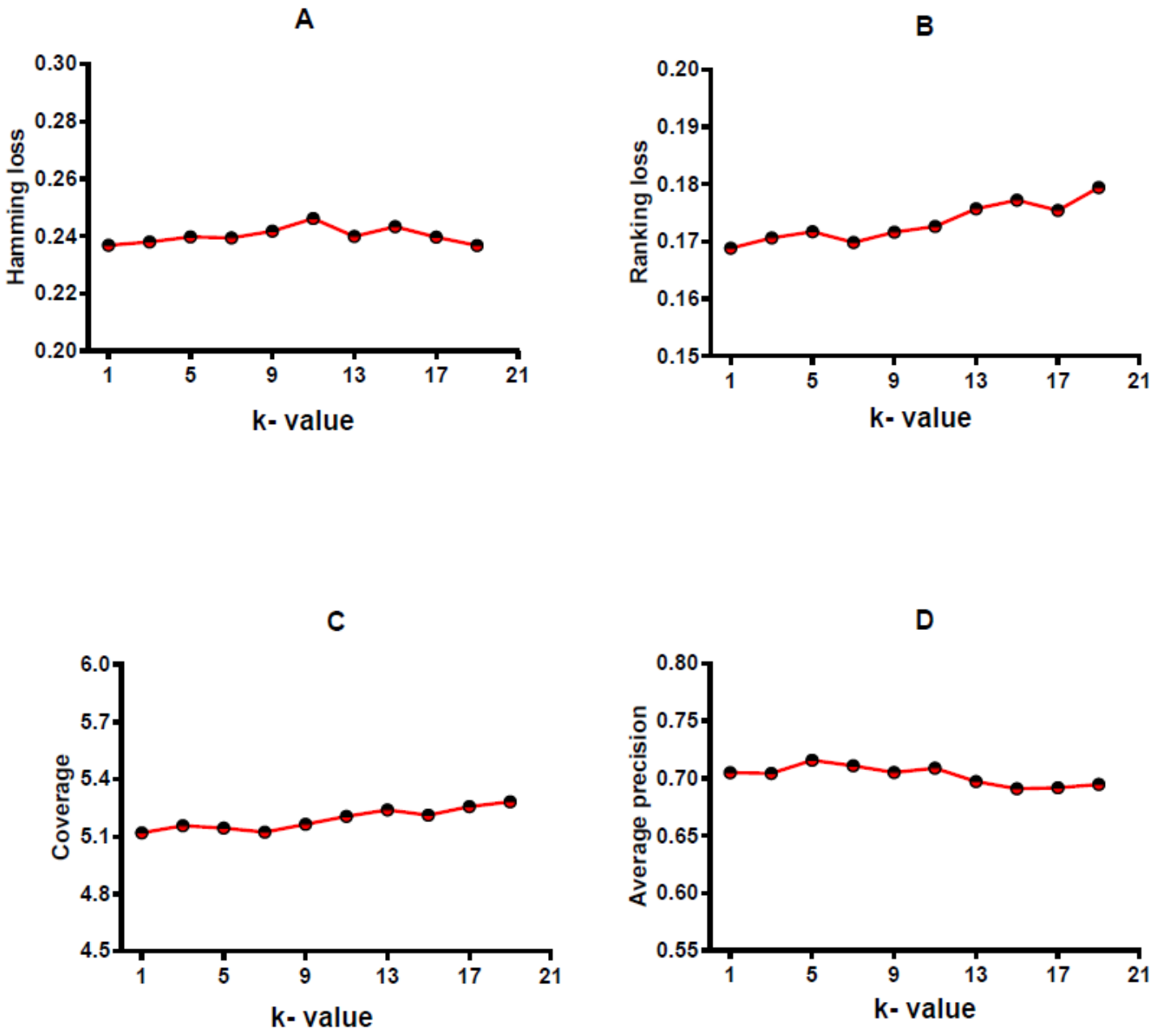

Figure 6

The influence of ML-kNN algorithm using microcosmic indexes on prediction results with different k values

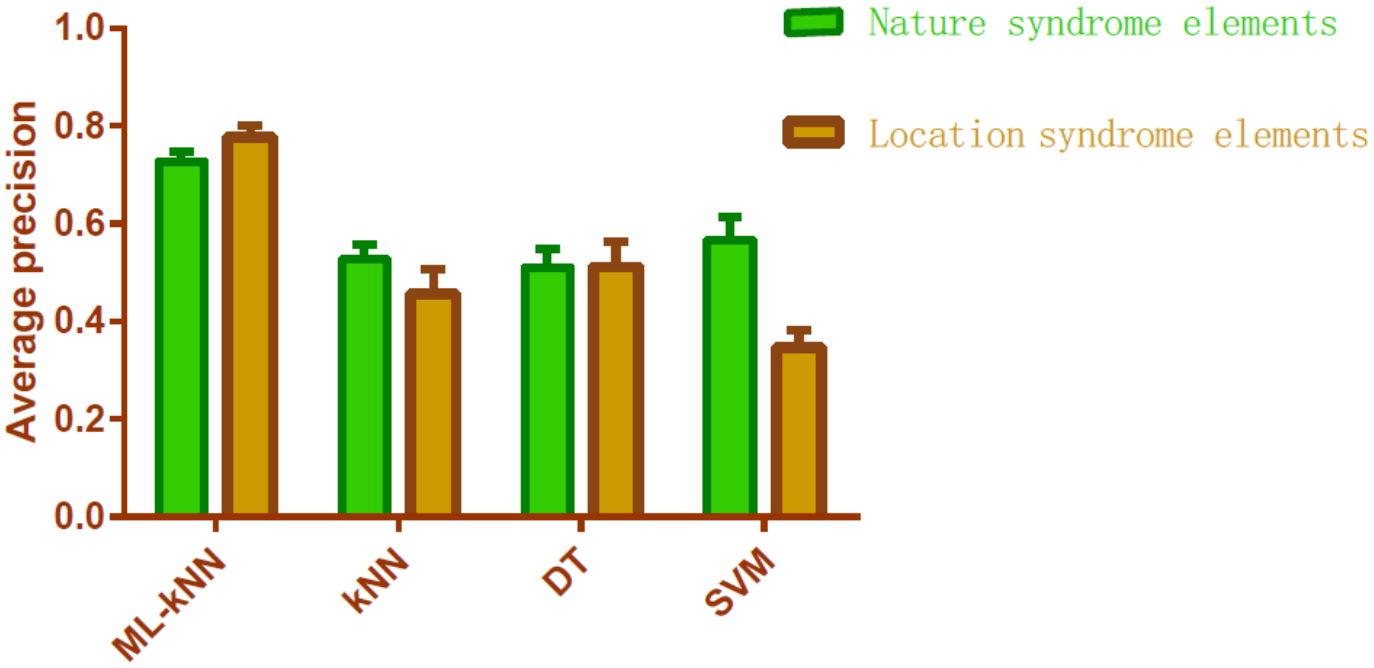

Figure 7

The influence of different syndrome elements on the average precision of ML-kNN using microcosmic indexes