Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Information Fusion 73 (2021) 11–21

Contents lists available at ScienceDirect

Information Fusion

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

An analysis model of diagnosis and treatment for COVID-19 pandemic based on medical information fusion

Fang Hu [a,b], Mingfang Huang [a], Jing Sun [c], Xiong Zhang [d], Jifen Liu [c,*]

[a] College of Information Engineering, Hubei University of Chinese Medicine, Wuhan 430065, PR China
[b] Department of Mathematics and Statistics, University of West Florida, Pensacola 32514, USA
[c] Department of Data Center, Hubei Provincial Hospital of Traditional Chinese Medicine, Wuhan 430060, PR China
[d] Department of Geriatrics, Hubei Provincial Hospital of Traditional Chinese Medicine, Wuhan 430060, PR China

ARTICLE INFO

Keywords:
Medical information fusion
Coronavirus Disease 2019 (COVID-19)
Diagnosis and treatment
Analysis model

ABSTRACT

Exploring the complicated relationships underlying the clinical information is essential for the diagnosis and treatment of the Coronavirus Disease 2019 (COVID-19). Currently, few approaches are mature enough to show operational impact. Based on electronic medical records (EMRs) of 570 COVID-19 inpatients, we proposed an analysis model of diagnosis and treatment for COVID-19 based on the machine learning algorithms and complex networks. Introducing the medical information fusion, we constructed the heterogeneous information network to discover the complex relationships among the syndromes, symptoms, and medicines. We generated the numerical symptom (medicine) embeddings and divided them into seven communities (syndromes) using the combination of Skip-Gram model and Spectral Clustering (SC) algorithm. After analyzing the symptoms and medicine networks, we identified the key factors using six evaluation metrics of node centrality. The experimental results indicate that the proposed analysis model is capable of discovering the critical symptoms and symptom distribution for diagnosis; the key medicines and medicine combinations for treatment. Based on the latest COVID-19 clinical guidelines, this model could result in the higher accuracy results than the other representative clustering algorithms. Furthermore, the proposed model is able to provide tremendously valuable guidance and help the physicians to combat the COVID-19.

1. Introduction

As the continuous growth of Coronavirus Disease 2019 (COVID-19) cases worldwide, the early diagnosis and supportive treatments are crucial to cure the patients [1]. Based on the various patients' manifestations, the physicians always need to find the representative symptoms from the complicated information to support their diagnosis for the different individuals [2]. After giving the accurate syndrome diagnosis based on the clinical symptoms, a serial of medicines should be combined to efficiently treat the patients. Therefore, the exploration of the effective diagnosis and therapy strategies referring to the key factors, symptom distribution, and medicine combination is of the greatest significance to diagnose and treat the COVID-19 patients.

Researches in the artificial intelligence (AI) or network analysis [3–5] make great contributions to fight against COVID-19 for the aspects of diagnosis and prognosis, treatments and cures [6,7]. Researchers have already developed numerous algorithms or models to explore the therapeutic schemes for COVID-19 [8,9]. Although these studies are relevant to discover the relationships of medical information, higher quality researches are needed to supply effective and reliable approaches to manage the COVID-19 pandemic [10]. Moreover, there is a broad range of potential applications of AI covering diagnosis and treatment challenges posed by the COVID-19.

The most challenges of the clinical diagnosis and treatment methods for the COVID-19 are the following two aspects:

- Most current researches identify the key factors (symptoms, medicines) in diagnosis and treatment by the simple statistical methods [6,11], such as frequency count, percentage, etc., which may be trapped into the restriction and one-sidedness without considering the multi-dimensions or multi-aspects of clinical information. Besides, the inherent relationships underlying the symptoms and medicines cannot be effectively discovered, which also leads to the inaccuracy of the analysis results [12,13].
- Despite a high volume of analysis models are constructed to support the clinical diagnosis or treatment for COVID-19, most of them only focus on one of the aspects: diagnosis or treatment [8,

* Corresponding author.
E-mail addresses: naomifang@hbtcm.edu.cn (F. Hu), mfhuang924@stmail.hbtcm.edu.cn (M. Huang), hbtcmsj@sina.com (J. Sun), zhangxiong2020@sina.com (X. Zhang), hbtcmljf@sina.com (J. Liu).

https://doi.org/10.1016/j.inffus.2021.02.016
Received 28 August 2020; Received in revised form 18 November 2020; Accepted 21 February 2021
Available online 1 March 2021

14]. Moreover, most of the diagnosis models just consider the X-ray or computed tomography (CT) image data only without referring to the key factors (symptoms) in the diagnosis process, which will also lead to the inaccuracy of the analysis results [15, 16].

In this paper, we build a diagnosis and treatment analysis model for the COVID-19 to address the challenges of one-sidedness and inaccuracy of clinical analysis results. Through the patients' manifestations, we hope this model can give the accurate diagnosis of syndromes and provide the corresponding therapies. In order to improve the comprehensiveness of analysis, we use six centrality evaluation metrics to identify core nodes in symptom and medicine networks. Then, based on the skip-gram model [17,18], we generate the symptom and medicine embeddings with conserving the hidden relationships underlying clinical information. Finally, we use spectral clustering (SC) [19,20] algorithm to improve the division accuracy through similarity calculation between any two symptom (medicine) embeddings. Through information fusion, we summarize the regularity of the main COVID-19 syndromes and their core symptoms and symptom distribution, core medicines and medicine combination. The following lists the contributions of the proposed analysis model of diagnosis and treatment for COVID-19 pandemic:

- This analysis model identifies the key symptoms and medicines based on six centrality evaluation metrics. It uses the skip-gram model to train and generate the symptom and medicine embeddings conserving the inherent information. Furthermore, combining the node embedding and spectral clustering, the performance of symptom and medicine division will be greatly improved compared to other contrast algorithms.
- The analysis model can comprehensively realize the diagnosis and treatment for COVID-19 by combining with the medical information fusion of the syndromes, symptoms, and medicines together. It can effectively identify the significant syndromes, their core symptoms and symptom distribution, core medicines and medicine combinations. By the information fusion, the accuracy of diagnosis and treatment of COVID-19 will be greatly improved.

The remaining parts of the article are organized as follows. We summarize the related work in Section 2. In Section 3, we depict the architecture of diagnosis and treatment analysis model, data preparation, and model realization. We present the experiment design and process, then give the experimental results and the discussion in Section 4. Finally, we summarize the conclusion and forecast the future works in Section 5.

2. Related work

2.1. Symptom and medicine analysis

Recently, most of the analyses for manifestations and diagnosis, drug use and treatment of COVID-19 patients focus on the simple statistical analysis methods (such as frequency count, percentage) of medical information. For the early manifestations of COVID-19, Adhikari et al. [6] reported that the classical symptoms of COVID-19 include cough, fever, headache, fatigue, pneumonia, hemoptysis, diarrhea, and dyspnea following a methodological framework. Gautier et al. [12] presented the new symptoms of taste and smell loss, excepting for the major symptoms including cough, fever, trouble breathing, etc. Based on 2,450,569 UK and 168,293 US individuals, Menni et al. [11] showed the rank of significant symptoms like loss of smell and taste, shortness of breath, fever, fatigue, persistent cough, skipped meals, diarrhea, hoarse voice, abdominal pain, delirium, chest pain. Song et al. [13] summarized that the most common symptoms included dry cough, fatigue, fever, myalgia and dyspnea, the less common symptoms involved abdominal pain, headache, nausea, diarrhea, and vomiting, and few patients have the manifestations of gastrointestinal symptoms. Medicine therapy is one of the significant treatments for COVID-19 patients. Statistical methods are also applied in medicine analysis. Dong et al. [21] discovered that some medicines, such as remdesivir, chloroquine, favipiravir, and arbidol are undergoing clinical researched to test their safety and efficiency for treating the COVID-19. Romo et al. [22] presented that the antimalarial drugs chloroquine and hydroxychloroquine may have activity against COVID-19. Ren et al. [23] presented that some medicines, such as Gypsum Fibrosum, Poria, etc., play a significant role in curing COVID-19 patients. Luo et al. [24] presented that the clinical drugs used frequently most include Armeniacae Semen Amarum, Scutellariae Radix, and Glycyrrhizae Radix et Rhizome.

Most of these aforementioned studies can effectively discover the key symptoms or medicines in the manifestations and treatments of COVID-19. However, most of the key factors are identified based on the simple statistical methods, such as frequency count, percentage, etc., which cannot be considered from the multiple inherent dimensions. Moreover, the relationships underlying symptoms, medicines, or these two together are not be considered. Therefore, an effective method to explore the key factors and their relationships is still needed to be researched in-depth.

2.2. Diagnosis and treatment model

During the period of COVID-19 pandemic, AI presents a powerful ability to support COVID-19 diagnosis and treatment based on multiple clinical data extracted from electronic medical records (EMRs) [10]. For COVID-19 detection, most of the diagnostic models are designed based on the X-ray or CT image data [7,14]. For example, based on the X-ray and CT images, Alom et al. [25] introduced the transfer learning method and presented an inception residual recurrent convolutional neural network for COVID-19 diagnosis, which can effectively segment the COVID-19 infected regions and get a highly accurate result. Gozes et al. [26] developed an AI-based image analysis approach to realize the COVID-19 automated detection and patient monitoring, which can achieve high accuracy. Through collecting a huge amount of CT slices from various hospitals, Kumar et al. [2] trained the deep learning model by a decentralized network to detect the COVID-19. AI is also applied for discovering efficient drugs to combat the COVID-19. Gao et al. [8] applied a generative network complex for drug discovery and identified 15 novel candidate drugs and two proposed HIV drugs for curing of COVID-19. Magar et al. [9] took Extreme Gradient Boost (XGBoost) and graph embeddings to discover new therapies in which they searched for antigen-neutralizing antibodies, and proposed 8 antibodies as potentially effective COVID-19 treatments. Bung et al. [27] used a reinforcement learning approach to generate drug compounds with desirable properties and proposed 31 candidate inhibitors for COVID-19.

Despite the diagnosis methods have good performance for COVID-19 detection, most of the diagnosis analysis data just includes the X-ray or CT images, which cannot reflect the inherent and significant characteristics, symptoms, referring to the diagnosis. Moreover, these analysis models of COVID-19 always focused on one aspect of diagnosis or treatment. It is more meaningful if the model research refers to these two aspects. Therefore, the diagnosis and treatment analysis model combining symptoms with medicines together should be further studied.

2.3. Information fusion

Recently, with the development of AI, complex network, and data mining, the utilization of converged various data, such as syndromes, symptoms, and medicines, from clinical EMRs, is also becoming popular [28,29]. Ang et al. [30] performed a network analysis to find 9 guidelines that provide medical formula for medical observation

based on clinical manifestation and showed that the Citri Reticulatae Pericarpium strongly paired with the Glycyrrhizae Radix et Rhizoma. Zhai et al. [1] reported that COVID-19 infection can cause a serial of severe respiratory diseases, and the treatments involving chloroquine and hydroxychloroquine, corticosteroids, antiviral agents, antibodies, convalescent plasma transfusion, and vaccines may work efficiently for COVID-19. Chan et al. [16] presented that the manifestations include dry cough, fever, upper airway congestion, fatigue, shortness of breath, sputum production, myalgia/arthralgia with lymphopenia, prolonged prothrombin time. Some conventional medicines, such as ribavirin, lopinavir/ritonavir, glucocorticoid, beta-interferon, etc., are applied in clinical trials. Jamshidi et al. [15] constructed a diagnosis and treatment platform to combat the COVID-19 through some deep learning methods, which integrates various aspects of information from the structured or unstructured data sources, such as medical imaging, medicine data, etc., and gave a good performance of service.

Although these studies can successfully realize the tentative exploration of the relationships between symptoms and medicines, the analysis model is still considering more medical information referring to the various factors with strong associations. Hence, information fusion, such as syndromes, symptoms, and medicines combined, is effective to improve the performance of the COVID-19 diagnosis and treatment model.

3. Diagnosis and treatment analysis model design

3.1. Model architecture

Fig. 1 shows the architecture of the COVID-19 diagnosis and treatment analysis model based on clinical EMRs, which consists of by the four components as follows:

- **Clinical Electronic Medical Records:** COVID-19 clinical data are extracted from the EMRs in the hospital information system. Based on the research theme, we focus on the diagnosis and treatment data referring to the different databases.
- **Information Fusion:** Each EMR corresponding to a COVID-19 patient contains the whole diagnosis and treatment information while in hospital. The data sets extracted from EMRs are desensitized, cleaned, and standardized, then, are constructed as three categories: syndrome data, symptom data, and medicine data. The relationships among homogeneous data or heterogeneous data will be conserved in each record.
- **Analysis Model:** The three category data sets and their hidden relationships will be abstracted as the nodes and edges in a heterogeneous network. We use a serial of methodologies, including the network analysis, node centrality, node embedding [31], clustering analysis, etc., to construct the overall analytical framework. By using this framework, we can discover the core symptoms and medicines for COVID-19, and summarize the regularity of symptom distribution and medicine combination in terms of the analysis results.
- **Applications:** In the period of COVID-19 pandemic, the clinicians can use this analysis model to realize the diagnosis and treatment: (i) Clinical diagnosis: the output results of core symptoms and common symptom distribution can offer the effective reference for the diagnosis. Based on them, clinicians can give the accurate judgment. (ii) Clinical treatment: the model will give the medicine analysis results referring to the core medicines and general medicine combination. In terms of them, the clinicians can provide the reasonable therapeutic schedule.

3.2. Data preparation

The analysis data of 570 COVID-19 inpatients is extracted from the hospital information system in Hubei Provincial Hospital of Traditional Chinese Medicine, Wuhan, China, one of the COVID-19 designated hospitals, from January 15th to March 13th, 2020. The data preparation includes two parts (shown in Fig. 2) as follows.

3.2.1. Data cleaning and standardization

The three categories of medical information is extracted from the "computerized physician order entry system": the syndrome information is from the diagnosis conclusion in the "discharge note table"; the symptom information is from the current symptom description in the "admission note table"; the medicine information is from the "medicine order table". Base on "International Classification of Diseases 11th Revision (ICD-11)",[1] the syndromes, symptoms, and medicines have been cleaned and standardized, of which the key processes include the term extraction, modifier deletion, terminology standardization, synonym combination, etc. Subsequently, the dictionaries of syndromes, symptoms, and medicines and their corresponding standardized data sets of COVID-19 have been constructed.

3.2.2. Information fusion

For each inpatient, the information from the EMRs has the inherent relationships among them, and the different category data should be analyzed synthetically. Hence, we fuse the syndromes, symptoms, and medicines together and try to discover the regularity of diagnosis and treatment for COVID-19.

3.3. Model realization

We used four approaches to realize the analysis model of the diagnosis and treatment for COVID-19 (Fig. 2) as follows.

3.3.1. Construction regularity of medical network

We used the heterogeneous information network (HIN) to abstract the hidden relationships underlying the medical information of COVID-19. HIN, a graph data model, can capture the relationships among entities (nodes), in which nodes and edges are annotated with class and relationship labels [32–34]. Through abstracting the syndromes, symptoms, medicines and their relationships underlying the EMRs, we constructed a heterogeneous information network model of COVID-19. In the COVID-19 HIN model, the different nodes represent the syndromes, symptoms, and medicines; the different edges denote the co-occurrence between the syndromes, symptoms, or medicines in the same EMR; the edge weights express the co-occurrence frequencies. Then, based on the theory of complex network [35,36], we projected the nodes and edges of the COVID-19 heterogeneous information network into the symptom and medicine networks, respectively. The construction regularity of the network is set as follows: take each symptom (medicine) in the records as a node; extract the co-occurrence relation between any two symptoms (medicines) in a diagnosis (prescription) as an edge; denote the co-occurrence frequency of two symptoms (medicines) as the edge weight. Based on it, we define the undirected weighted graph $G(V, E, W)$ of symptom or medicine, where V, E, and W denote the set of nodes (symptoms or medicines), edges (co-occurrence relations), and weights (frequencies of co-occurrences), respectively.

3.3.2. Evaluation metrics of node centrality

In complex networks, researchers always identify the important nodes using several evaluation metrics [37]. The representative metrics includes degree centrality (BC), closeness centrality (CC), betweenness centrality (BC) [38], eigenvector centrality (EC) [39], current flow closeness centrality (CCC) [40], and load centrality (LC) [41]. In general, the DC, BC, CC, and EC are the conventional metrics to evaluate the node centrality. Considering the complexity of the medical networks, we introduced the CCC metric with the anti-noise feature and the LC metric with the characteristic of local information comparison to further evaluate. The description about these six evaluation indices are shown in Table 1. For DC, $deg(v)$ denotes the degree of node v, n represents the number of nodes; for BC, δ_{st} indicates the number of

[1] https://icd.who.int/en.

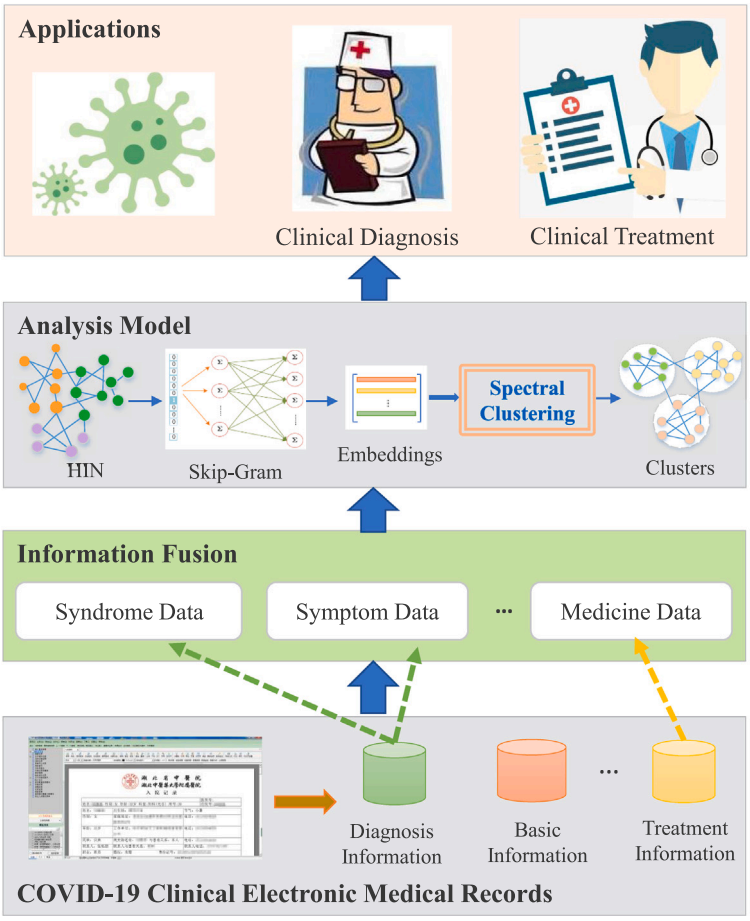

Fig. 1. The architecture of diagnosis and treatment analysis model.

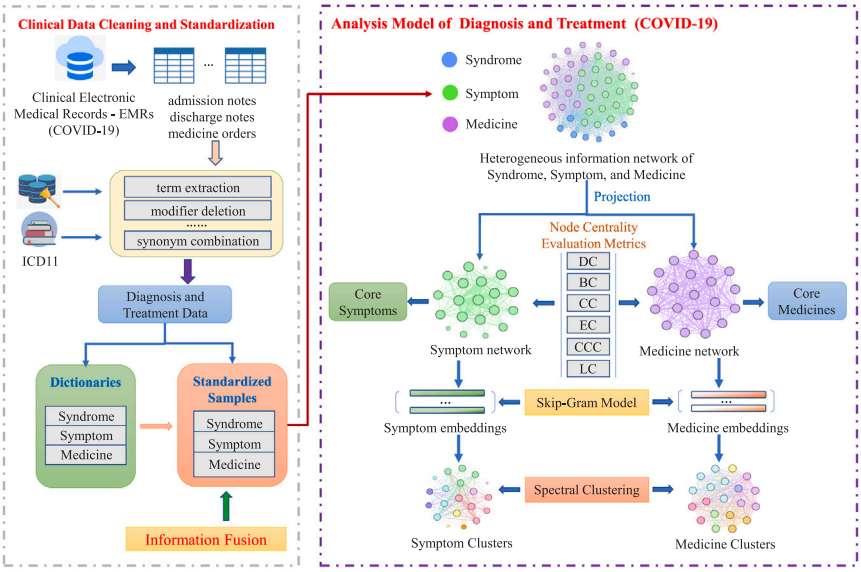

Fig. 2. Flowchart of data processing.

the shortest paths from nodes s to t, $\delta_{st}(v)$ denotes the number of the shortest paths through node v; for CC, $d_G(v,t)$ represents the shortest paths from v to t; for EC, $\lambda_1, \lambda_2, \ldots, \lambda_n$ indicates the eigenvalues of the adjacent matrix A, and e_t is the corresponding eigenvector of λ_t; for CCC, $n-1$ is a normalizing factor, $p_{st}(v)$ is closeness index of node v based on shortest paths, and $p_{st}(v) - p_{st}(t)$ corresponds to the distance between s and t; for LC, $\theta_{s,d}$ is a quantity of the information that is sent from s to d, and $\theta_{s,d}(v)$ denotes the overall information forwarded by v.

Table 1
Evaluation metrics of node centrality.

Names	Abbreviations	Equations
Degree centrality	DC	$C_d(v) = \frac{deg(v)}{n-1}$
Betweenness centrality	BC	$C_b(v) = \frac{\sum_{s\neq v\neq t\in V} \frac{\delta_{st}(v)}{\delta_{st}}}{(n-1)(n-2)/2}$
Closeness centrality	CC	$C_c(v) = \frac{\sum_{t,v\in V} d_G(v,t)}{n-1}$
Eigenvector centrality	EC	$C_e(v) = \lambda^{-1}\sum_{t=1}^{N} a_{vt}e_t$
Current flow closeness centrality	CCC	$C_{cc}(v) = \frac{n-1}{\sum_{v\in V\wedge v\neq t}(p_{st}(v)-p_{st}(t))}$
Load centrality	LC	$LC(v) = \sum_{s,d\in V}\theta_{s,d}(v)$

3.3.3. Training approach of node embedding

Based on the symptom or medicine network, we used the random walk probability and the Skip-Gram model [17] to generate the symptom embeddings or medicine embeddings [42], respectively. We show the training steps as follows:

Step 1: Initial the unnormalized transition probability for a symptom (medicine) node in the network. Traverse a selected node by the following rules: store the weights between this node and its neighbor nodes; summarize and normalize the weights of the current symptom (medicine) node; obtain the transition probabilities from this node to its neighbors.

Step 2: Based on Eqs. (1) and (2), set the unnormalized transition probabilities of each edge.

$$\pi_{tj} = \alpha_{pq}(i,j)\cdot w_{tj} \tag{1}$$

$$\alpha_{pq}(i,j) = \begin{cases} \frac{1}{p}, & if\ d_{ij}=0 \\ 1, & if\ d_{ij}=1 \\ \frac{1}{q}, & if\ d_{ij}=2 \end{cases} \tag{2}$$

where i and j indicate the nodes v_i and v_j, separately; w_{tj} denotes the weight between v_t and v_j; v_t is an intermediate node from v_i to v_j; $\alpha_{pq}(i,j)$ expresses the search bias between v_i and v_j, p is the breadth weight and q is the depth weight; π_{tj} indicates the unnormalized transition probability from v_t to v_j; d_{ij} is the shortest path from v_i to v_j, and $d_{ij}\in\{0,1,2\}$.

Step 3: Normalize the weight of each edge; obtain the transition probability from the current node to its neighbors.

Step 4: Acquire each node's walk paths ($walks$). Calculate the transition probabilities from the current node to its neighbors based on Eq. (3).

$$P\left(c_l = j \mid c_{l-1} = t\right) = \begin{cases} \frac{\pi_{tj}}{Z}, & if\ \ (t,j)\in E \\ 0, & otherwise \end{cases} \tag{3}$$

where c_l indicates the lth node in a walk path; t and j demonstrate the nodes v_t and v_j; Z indicates a constant for the normalization.

Step 5: Construct a node list for all the nodes and their paths. The walk regularity is defined as: shuffle the sequence of nodes and randomly select a node; the selected node walks to the neighbor using $P\left(c_l = j \mid c_{l-1} = t\right)$.

Step 6: Use the Skip-gram model to train the walk paths of a node, and obtain the embedding for each symptom (medicine) node.

3.3.4. Clustering approach of node embedding

The combination of graph embedding and spectral clustering (SC) algorithm works well on 20,000 node network in [20]. According to the trained symptom or medicine embeddings, we also selected the SC algorithm [19] to obtain the clustering results of symptoms or medicines, respectively. We show the clustering steps as follows:

Step 1: Get the weighted matrix W (similarity matrix S) by Eq. (4). Present w_{ij} as the edge weight between v_i and v_j, and s_{ij} as the similarity of these two symptom (medicine) node embeddings. The weighted matrix is denoted as $W = \left\{w_{ij} \mid 1\le i\le n, 1\le j\le n\right\}$, here, n is the number of nodes. Obtain the Euclidean distance $\|v_i - v_j\|_2^2$ between v_i and v_j, then get w_{ij} and s_{ij} as follows:

$$w_{ij} = s_{ij} = exp\left(-\frac{\|v_i - v_j\|_2^2}{2\sigma^2}\right) \tag{4}$$

where σ is a scaling parameter to control the descending speed of w_{ij} as the distance descending between v_i and v_j. In [20], the SC algorithm can acquire the best performance when σ is set as 1, here, we also set σ as 1.

Step 2: Based on Eq. (5), obtain the degree matrix D, d_i indicates the sum of weighted edges connecting to v_i.

$$d_i = \sum_{j=1}^{n} w_{ij} \tag{5}$$

Step 3: Denote the Laplacian matrix as $L = D - W$, normalize L as $L' = D^{-1/2}LD^{1/2}$. Get the first k minimum eigenvalues of L' and their corresponding eigenvectors. Reconstruct the normalized eigenvectors to an new eigenmatrix F of size $n\times k$, here $k \ll n$.

Step 4: Set the cluster number as m, use the K-means algorithm to divide the eigenmatrix F into the symptom (medicine) clusters $C = \{C_1, C_2, \ldots, C_m\}$.

4. Experiment

4.1. Experiment design

We designed four experiments and conducted them on a single 16 GB RAM 3.6 GHz Intel (R) Core (TM) CPU. Firstly, in Section 4.2, we constructed the heterogeneous information network of syndromes, symptoms and medicines, and then projected it into the symptom and medicine networks, respectively. Then, we identified the core symptom and core medicine nodes by six evaluation metrics of node centrality in Section 4.3. Subsequently, in Section 4.4, we obtained the symptom and medicine embeddings using the skip-gram model. Then, we acquired the symptom groups and medicine combinations through the SC algorithm, respectively. Finally, we gave the experimental results and the relative analysis compared SC with other representative algorithms: K-means and hierarchical clustering algorithms in Section 4.5.

4.2. Network model construction

Based on the regularity of network construction in Section 3.3.1, We constructed the HIN model (Fig. 3), in which the different types of nodes represent the entities: syndromes, symptoms, and medicines with three colors; the various edges show the different relationships underlying these entities.

Then, we show the construction processes of COVID-19 symptom network, and medicine network in Figs. 4 and 5, respectively. We initialized a network of two symptom nodes *fever* and *cough* with the relative edges. In development, the other two symptom nodes *weakness* and *expectoration* with their edges were extended into the network. Finally, we obtained a COVID-19 undirected weighted symptom network with 83 nodes and 10,126 edges in Fig. 4. By the similar way, we also initialized a network with two medicine nodes *Radix Glycyrrhizae* and *Rhizoma Pinelliae Preparatum*, and then, extended other two medicine nodes *Agastache rugosa* and *Fructus Aurantii Immaturus* into this network. Ultimately, we acquired the medicine network with 316 nodes and 216,327 edges in Fig. 5. We show the statistical summary of the symptom and medicine networks in Table 2, including the name, number of nodes and edges, minimum/maximum/average weights.

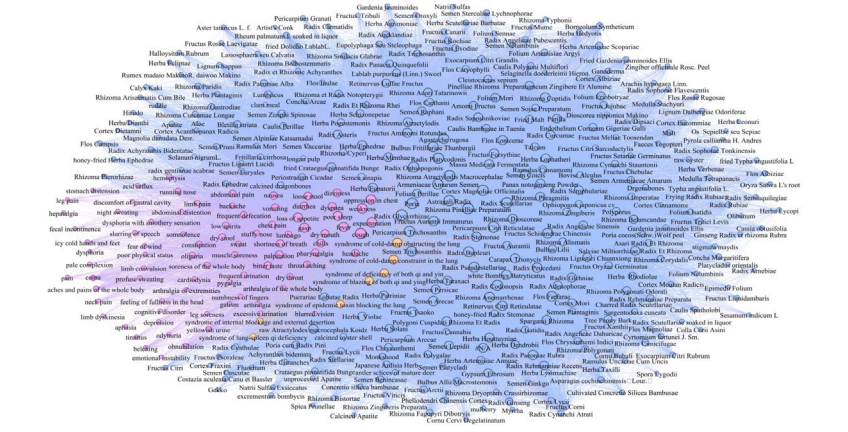

Fig. 3. Heterogeneous information network of COVID-19 with syndromes (yellow nodes), symptoms (purple nodes), and medicines (blue nodes). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

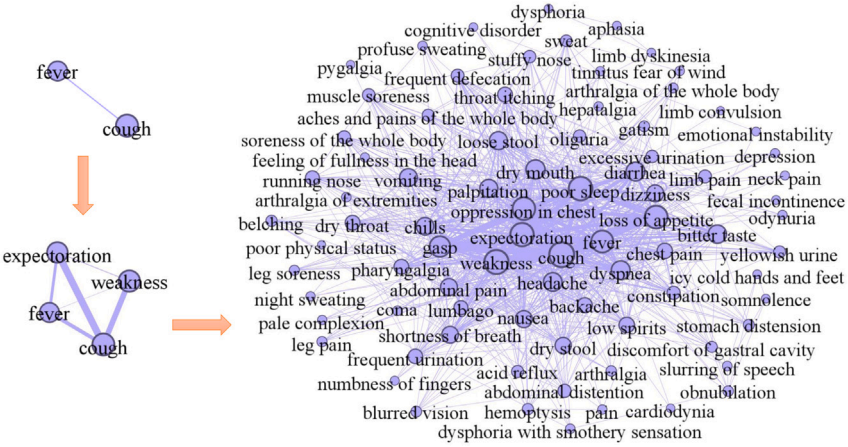

Fig. 4. Construction of symptom network.

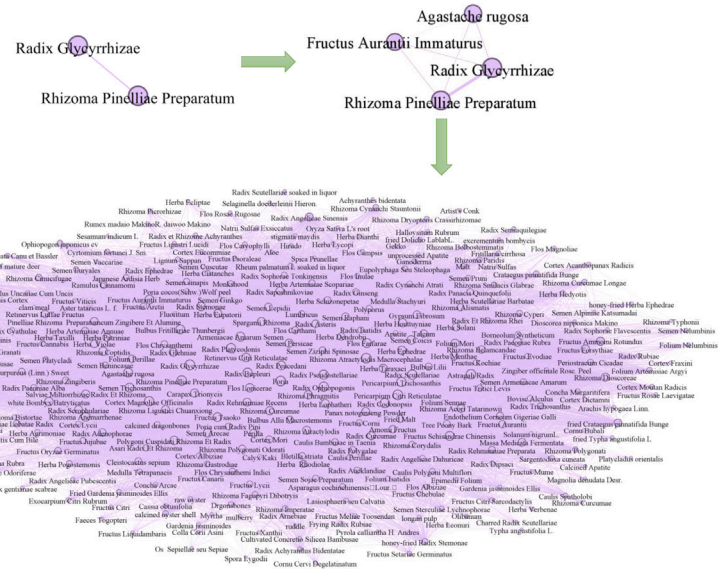

Fig. 5. Construction of medicine network.

4.3. Core node evaluation

Based on these two networks, we used six evaluation metrics, including DC, BC, CC, EC, CCC, and LC, to get the various centrality values for the nodes, and just show the top 20 significant symptoms and medicines in Tables 3 and 4 (the digits of the top 10 symptoms and medicines are bolded, respectively). From Table 3 in Appendix, we show that these six metrics can identify the same top 8 symptoms,

Table 2
Statistical summary of the symptom and medicine networks.

Names	Number of nodes	Number of edges	Minimum weights	Maximum weights	Average weights
Symptom network	83	10,126	1	272	13
Medicine network	316	216,327	1	401	11

including *cough*, *weakness*, *poor sleep*, *expectoration*, *loss of appetite*, *oppression in chest*, *fever*, and *gasp*; the DC, CC, EC, and CCC can find the same top 9 important symptom *dry mouth*; the DC, CC, BC, CCC, and LC, excepting for EC, can discover the top 10 symptom *dyspnea*. Based on Table 4 in Appendix, we show that these six metrics can identify the same top 10 medicines, including *Radix Glycyrrhizae*, *Poria*, *Pericarpium Citri Reticulatae*, *Rhizoma Atractylodis Macrocephalae*, *Rhizoma Pinelliae Preparatum*, *Astragali Radix*, *Radix Scutellariae*, *Cortex Magnoliae Officinalis*, *Radix Codonopsis*, and *Radix Ophiopogonis*; excepting for the EC identified the top 10 medicine as *Radix Platycodonis*, however, the centrality value 0.0912 is similar to the *Radix Ophiopogonis* with 0.0899.

We compared the identified top 20 significant symptoms and medicines to latest clinical guidelines of COVID-19 "Diagnosis and treatment of corona virus disease-19 (7th trial edition)".[2] We show that 100% of the top 20 symptoms (bolded in Table 3) and 80% of the top 20 medicines (bolded in Table 4) are presented in the latest guidelines, respectively.

4.4. Symptom and medicine embedding training and clustering analysis

According to the matrices of symptom and medicine networks, we used the realization steps of node embeddings described in Section 3.3.3 to train the COVID-19 symptom and medicine embeddings (also called vectors), respectively. By transferring the one-hot vectors of symptoms and medicines into numerical vectors, we got 83 symptom embeddings with 128 dimensions and 316 medicine embeddings with 128 dimensions.

Community division [43] is introduced to discover the closely connections underlying the symptoms or medicines, which can effectively explore the special clinical syndromes of COVID-19. We used the SC algorithm to respectively divide 83 symptom embeddings and 316 embeddings into the their corresponding syndrome communities. The SC algorithm is well-known that the community number is confirmed as an input parameter. As the aforementioned latest clinical guidelines of COVID-19. The main syndromes of COVID-19 have been divided into seven categories, therefore, we set the input community number of SC algorithm as seven.

As shown in Figs. 6 and 7, we divided the COVID-19 symptom network (Fig. 4) and medicine network (Fig. 5) into seven communities by the SC algorithm. The clinical syndromes (denoted as seven communities with different colors) and their corresponding symptoms and medicines are presented as follows:

- Syndrome of cold-damp constraint in the lung (colored by rose pink) includes the core symptoms *weakness*, *expectoration*, *cough*, *fever*, and *loss of appetite* and other symptoms *palpitation*, *headache*, *dizziness*, *abdominal pain*, *loose stool*, etc. The effective treatment for this syndrome refers to the core medicine *Radix Scrophulariae* and other medicines *Astragali Radix*, *Herba Taraxaci*, *Rhizoma Anemarrhenae*, *Herba Ephedrae*, *Rhizoma Zingiberis*, etc.
- Syndrome of deficiency of both qi and yin (colored by wisteria) includes the core symptom *chills* and other symptoms *lumbago*, *constipation*, *leg soreness*, *emotional instability*, *fecal incontinence*, etc. The effective treatment for this syndrome refers to the core medicine *Radix Adenophorae* and other medicines *Fructus Schisandrae Chinensis*, *Salviae Miltiorrhizae Radix Et Rhizoma*, *Semen Arecae*, *Herba Lophatheri*, *Armeniacae Amarum Semen*, etc.
- Syndrome of epidemic toxin blocking the lung (colored by light blue) includes the core symptom *gasp* and other symptoms *bitter taste*, *nausea*, *soreness of the whole body*, *frequent urination*, *abdominal distention*, etc. The effective treatment for this syndrome refers to the core medicine *Bulbus Fritillariae Thunbergii* and other medicines *Fructus Forsythiae*, *Pericarpium Trichosanthis*, *Poria*, *Exocarpium Citri Grandis*, *Semen Lepidii*, etc.
- Syndrome of lung–spleen qi deficiency (colored by orange) includes the core symptoms *acid reflux* and *excessive urination* and other symptoms *aches and pains of the whole body*, *arthralgia*, *icy cold hands and feet*, *odynuria*, *gatism*, etc. The effective treatment for this syndrome refers to the core medicine *Rhizoma Dioscoreae* and other medicines *Pericarpium Citri Reticulatae*, *Polygoni Cuspidati Rhizoma Et Radix*, *Radix Paeoniae Rubra*, *Radix Glycyrrhizae*, etc.
- Syndrome of cold-damp obstructing the lung (colored by light green) includes the core symptom *oppression in chest* and other symptoms *dry mouth*, *vomiting*, *diarrhea*, *shortness of breath*, *low spirits*, etc. The effective treatment for this syndrome refers to the core medicine *Rhizoma Atractylodis* and other medicines *Radix Platycodonis*, *Gypsum Fibrosum*, *Fried Malt*, *Radix Peucedani*, *Fructus Aurantii Immaturus*, etc.
- Syndrome of internal blockage and external desertion (colored by lilac) includes the core symptom *dyspnea* and other symptoms *limb dyskinesia*, *cognitive disorder*, *slurring of speech*, *somnolence*, *dysphoria*, etc. The effective treatment for this syndrome refers to the core medicine *Tree Peony Bark* and other medicines *Radix Saposhnikoviae*, *Rhizoma Phragmitis*, *Flos Farfarae*, *Amomi Fructus*, *Agastache rugosa*, etc.
- Syndrome of blazing of both qi and ying (colored by yellow) includes the core symptom *poor sleep* and other symptoms *dry stool*, *coma*, *limb convulsion*, etc. The effective treatment for this syndrome refers to the core medicine *Radix Pseudostellariae* and other medicines *Radix Ophiopogonis*, *Rhizoma Coptidis*, *Folium Perillae*, *Massa Medicata Fermentata*, *Periostracum Cicadae*, etc.

4.5. Experimental result comparison analysis

In general, for the clinical diagnosis and treatment in real world, the numbers of symptom representations and medicines may be greater than the medical information in the latest clinical guidelines of COVID-19. Thus, we summarized the all specific 7 syndromes and their corresponding 63 primary symptoms and 78 primary medicines as the evaluation standard to verify the accuracy and performance of the analysis model.

As the previous studies, we found that the non-adaptive clustering algorithms can acquire the better results than the adaptive algorithms on the graph embeddings [44]. Therefore, we selected the representative non-adaptive algorithms: K-means and hierarchical clustering algorithms as the comparison. We have used the representative evaluation metrics: accuracy, modularity [45], normalized mutual information (NMI) [46], Fowlkes–Mallows index (FMI) [47], adjusted rand index (ARI) [48], and adjusted mutual information (AMI) [49], to evaluate the quality of community division. We show the evaluation results of different symptom communities in Fig. 8, and various medicine communities in Fig. 9, respectively.

After comparative analysis, the experimental results show that the SC algorithm works better than the other two algorithms and gets more successful runs on the all metrics. From Fig. 8 of symptom division

[2] General Office of National Health Commission of the People's Republic of China, Office of National Administration of Traditional Chinese Medicine. Diagnosis and treatment of corona virus disease-19 (7th trial edition) [J]. China Medicine, 2020, 15(6): 801–805. DOI: 10. 3760/j. issn. 1673–4777. 2020. 06. 001.

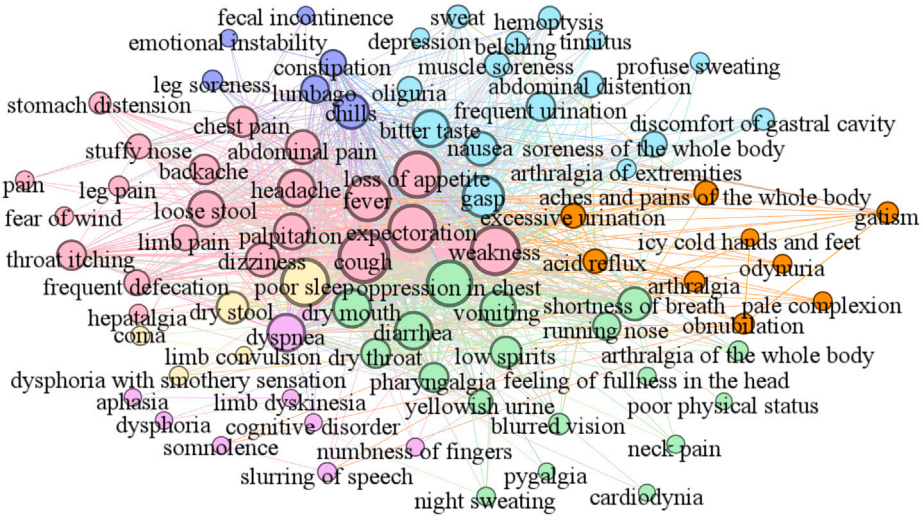

Fig. 6. Symptom communities. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

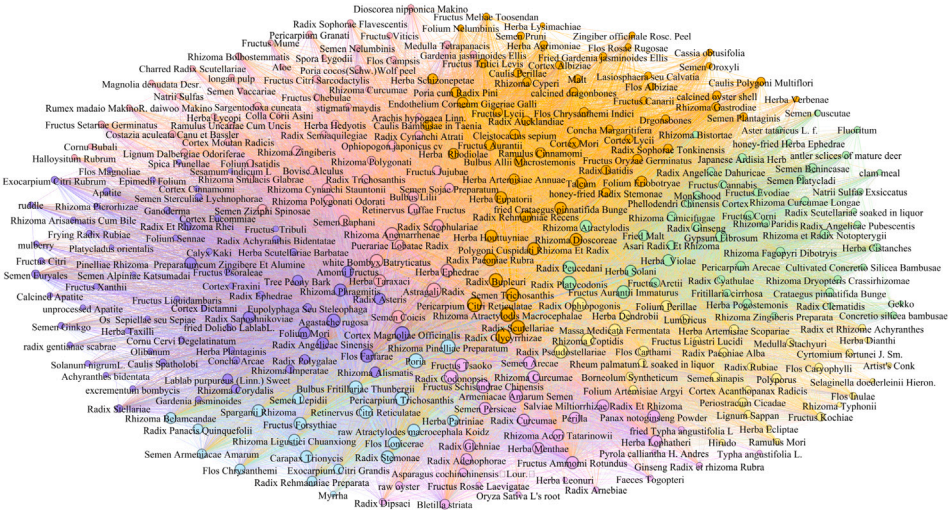

Fig. 7. Medicine communities. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

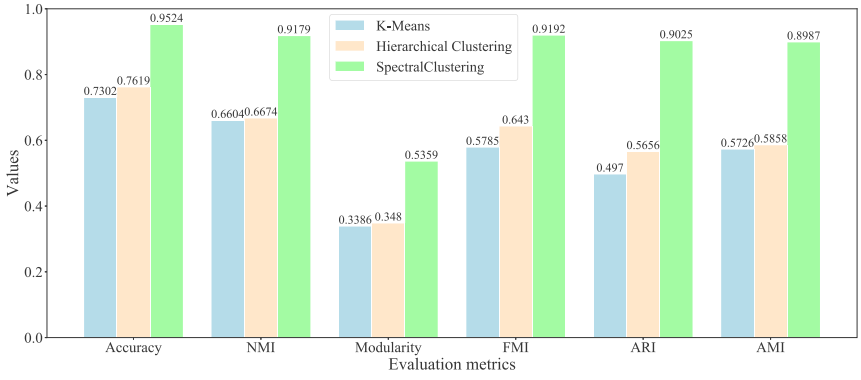

Fig. 8. Evaluation of different clustering algorithms on the primary symptom network.

evaluation, we show that the SC algorithm can get the highest accuracy 0.9524 than other algorithms: K-means with 0.7302 and hierarchical clustering with 0.7619. For the classical metric modularity, the SC algorithm can acquire the higher value 0.5359 than K-means with

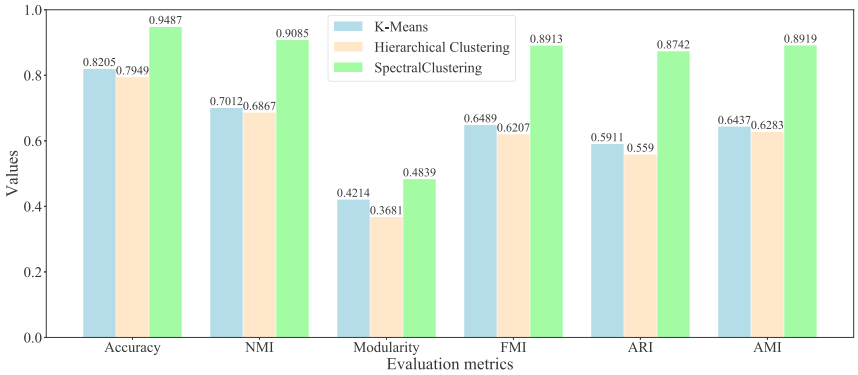

Fig. 9. Evaluation of different clustering algorithms on the primary medicine network.

Table 3
Node centrality analysis of symptom network.

No.	Symptoms	Degree	Closeness	Betweenness	Eigenvector	Current flow closeness	Load
1	**cough**	**0.7683**	**0.8039**	**0.1117**	**0.2147**	**0.0689**	**0.1121**
2	**weakness**	**0.7317**	**0.7810**	**0.0987**	**0.2121**	**0.0686**	**0.0992**
3	**poor sleep**	**0.7073**	**0.7593**	**0.1319**	**0.1998**	**0.0684**	**0.1315**
4	**expectoration**	**0.6951**	**0.7523**	**0.0619**	**0.2106**	**0.0681**	**0.0619**
5	**loss of appetite**	**0.6707**	**0.7455**	**0.0645**	**0.2050**	**0.0679**	**0.0624**
6	**oppression in chest**	**0.6585**	**0.7257**	**0.0442**	**0.2047**	**0.0677**	**0.0441**
7	**fever**	**0.6341**	**0.7257**	**0.0479**	**0.1971**	**0.0675**	**0.0482**
8	**gasp**	**0.5732**	**0.6833**	**0.0207**	**0.1946**	**0.0666**	**0.0206**
9	**dry mouth**	**0.5000**	**0.6508**	0.0149	**0.1787**	**0.0654**	0.0150
10	**dyspnea**	**0.4878**	**0.6560**	**0.0259**	0.1718	**0.0653**	**0.0260**
11	**diarrhea**	**0.4878**	0.6457	0.0157	0.1740	0.0653	0.0156
12	**headache**	0.4756	0.6406	0.0092	**0.1765**	0.0650	0.0092
13	**bitter taste**	0.4634	0.6357	0.0075	0.1717	0.0648	0.0075
14	**palpitation**	0.4512	0.6308	0.0046	0.1753	0.0645	0.0046
15	**loose stool**	0.4512	0.6308	0.0092	0.1653	0.0645	0.0093
16	**vomiting**	0.4390	0.6308	**0.0874**	0.1510	0.0642	**0.0875**
17	**chills**	0.4268	0.6212	0.0070	0.1615	0.0640	0.0070
18	**dizziness**	0.4024	0.6029	0.0059	0.1517	0.0628	0.0059
19	**nausea**	0.4024	0.6119	0.0030	0.1615	0.0634	0.0030
20	**abdominal pain**	0.4024	0.6119	0.0057	0.1569	0.0634	0.0056

0.3386 and hierarchical clustering with 0.348. For the evaluation of community detection, it is widely believed that it is a good division when the modularity is greater than 0.3. For the other four metrics: NMI, FMI, ARI, and AMI, the SC algorithm can also obtain the highest values 0.9179, 0.9192, 0.9025, and 0.8987 than K-means with 0.6604, 0.5785, 0.497, and 0.5726, and hierarchical clustering with 0.6674, 0.643, 0.5656, and 0.5858.

Based on Fig. 9 of medicine division evaluation, the experimental results demonstrate that the SC algorithm obtain the highest accuracy 0.9487 than other algorithms: K-means with 0.8205 and hierarchical clustering with 0.7949. For the modularity, the SC algorithm can acquire the higher value 0.4839 than K-means with 0.4214 and hierarchical clustering with 0.3681. For the other four metrics: NMI, FMI, ARI, and AMI, the SC algorithm can also obtain the highest values 0.9085, 0.8913, 0.8742, and 0.8919 than K-means with 0.7012, 0.6489, 0.5911, and 0.6437, and hierarchical clustering with 0.6867, 0.6207, 0.559, and 0.6283.

5. Conclusion

In this study, we have explored an effective model to support the diagnosis and treatment of COVID-19. Through the combination of complex network and machine learning techniques, the proposed model is able to find the key factors and the clinical regularity in the process of diagnosis and treatment accurately and effectively. Firstly, we designed the symptom and medicine networks to represent the complex relationships underlying the symptoms and medicines, respectively. Secondly, we utilized six evaluation metrics to identify the core symptom and medicine nodes based on the network topology. Thirdly, we trained each symptom or medicine node using the skip-gram model to generate the numerical symptom or medicine embedding (or called a vector), which conserves the similarity relationship between any two symptoms or medicines. In order to find the symptom group or the medicine combination, we used the SC algorithm to divide the symptoms or medicines into seven communities (clusters) with the corresponding to the specific syndromes. Through checking up on the latest clinical guidelines of COVID-19 and compared to the representative clustering algorithms, we show that our model can accurately and effectively discover the symptom groups and medicine combinations, and their corresponding syndromes.

In the clinical practice of COVID-19, this model can filter the irrelevant information and acquire the key factors and important clinical regularity. Through introducing the medical information fusion, the model discovers the relative effects among the syndromes, symptoms, and medicines, which provides tremendously valuable guidance and helps physicians to give the effective diagnosis and treatment strategies for the COVID-19 patients. Furthermore, the dictionaries and embedding sets of syndromes, symptoms, and medicines will be supplied as the basic data sets for the COVID-19 researchers. The efficiency of the prediction of disease evolution for COVID-19 remains a challenging area. We will use the efficient solver such as iterative solvers to find the eigenvectors in the Spectral Clustering method in the future to improve the algorithm's efficiency.

Table 4
Node centrality analysis of medicine network.

No.	Medicines	Degree	Closeness	Betweenness	Eigenvector	Current flow closeness	Load
1	**Radix Glycyrrhizae**	**0.9714**	**0.9722**	**0.0200**	**0.0936**	**0.1805**	**0.0200**
2	**Poria**	**0.9524**	**0.9545**	**0.0165**	**0.0931**	**0.1798**	**0.0165**
3	**Pericarpium Citri Reticulatae**	**0.9429**	**0.9459**	**0.0173**	**0.0927**	**0.1795**	**0.0173**
4	**Rhizoma Atractylodis Macrocephalae**	**0.9429**	**0.9459**	**0.0211**	**0.0927**	**0.1795**	**0.0211**
5	**Rhizoma Pinelliae Preparatum**	**0.9365**	**0.9403**	**0.0158**	**0.0926**	**0.1792**	**0.0158**
6	**Astragali Radix**	**0.9175**	**0.9238**	**0.0140**	**0.0916**	**0.1785**	**0.0140**
7	**Radix Scutellariae**	**0.9143**	**0.9211**	**0.0132**	**0.0920**	**0.1784**	**0.0132**
8	**Cortex Magnoliae Officinalis**	**0.9143**	**0.9211**	**0.0134**	**0.0918**	**0.1784**	**0.0134**
9	**Radix Codonopsis**	**0.8984**	**0.9078**	**0.0125**	**0.0910**	**0.1778**	**0.0125**
10	**Radix Ophiopogonis**	**0.8889**	**0.9000**	**0.0129**	0.0899	**0.1774**	**0.0129**
11	**Radix Platycodonis**	**0.8889**	**0.9000**	0.0109	**0.0912**	0.1773	0.0109
12	**Pericarpium Trichosanthis**	**0.8889**	**0.9000**	0.0113	0.0908	**0.1774**	0.0113
13	**Radix Pseudostellariae**	0.8794	0.8924	0.0108	0.0904	0.1770	0.0108
14	**Radix Bupleuri**	0.8635	0.8799	0.0109	0.0898	0.1763	0.0109
15	**Armeniacae Amarum Semen**	0.8540	0.8726	0.0100	0.0889	0.1759	0.0100
16	**Fructus Schisandrae Chinensis**	0.8476	0.8678	0.0103	0.0883	0.1756	0.0103
17	Flos Farfarae	0.8413	0.8630	0.0080	0.0892	0.1753	0.0080
18	**Agastache rugosa**	0.8381	0.8607	0.0087	0.0889	0.1752	0.0087
19	Semen Trichosanthis	0.8349	0.8583	0.0084	0.0885	0.1751	0.0084
20	Bulbus Fritillariae Thunbergii	0.8349	0.8583	0.0086	0.0889	0.1751	0.0086

CRediT authorship contribution statement

Fang Hu: Methodology, Formal analysis, Writing - original draft, Writing - review & editing. **Mingfang Huang:** Methodology, Software, Visualization. **Jing Sun:** Data curation, Validation. **Xiong Zhang:** Conceptualization, Investigation. **Jifen Liu:** Resources, Conceptualization, Supervision.

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.

Appendix

See Tables 3 and 4.

References

[1] P. Zhai, Y. Ding, X. Wu, J. Long, Y. Zhong, Y. Li, The epidemiology, diagnosis and treatment of COVID-19, Int. J. Antimicrob. Ag. 55 (5) (2020) 105955.

[2] R. Kumar, A.A. Khan, S. Zhang, W. Wang, Y. Abuidris, W. Amin, J. Kumar, Blockchain-federated-learning and deep learning models for COVID-19 detection using CT imaging, 2020, arXiv preprint arXiv:2007.06537.

[3] K.-H. Yu, A.L. Beam, I.S. Kohane, Artificial intelligence in healthcare, Nat. Biomed. Eng. 2 (10) (2018) 719–731.

[4] T. Davenport, R. Kalakota, The potential for artificial intelligence in healthcare, Future Healthc. J. 6 (2) (2019) 94–98.

[5] F. Hu, L. Li, X. Huang, X. Yan, P. Huang, Symptom distribution regularity of insomnia: Network and spectral clustering analysis, JMIR Med. Inform. 8 (4) (2020) e16749.

[6] S.P. Adhikari, S. Meng, Y.-J. Wu, Y.-P. Mao, R.-X. Ye, Q.-Z. Wang, C. Sun, S. Sylvia, S. Rozelle, H. Raat, et al., Epidemiology, causes, clinical manifestation and diagnosis, prevention and control of coronavirus disease (COVID-19) during the early outbreak period: a scoping review, Infect. Dis. Poverty 9 (1) (2020) 1–12.

[7] A.A. Ardakani, A.R. Kanafi, U.R. Acharya, N. Khadem, A. Mohammadi, Application of deep learning technique to manage covid-19 in routine clinical practice using ct images: Results of 10 convolutional neural networks, Comput. Biol. Med. 121 (2020) 103795.

[8] D. Nguyen, K. Gao, J. Chen, R. Wang, G. Wei, Potentially highly potent drugs for 2019-ncov, BioRxiv (2020) 1–13.

[9] R. Magar, P. Yadav, A.B. Farimani, Potential neutralizing antibodies discovered for novel corona virus using machine learning, 2020, arXiv preprint arXiv:2003.08447.

[10] J. Bullock, K.H. Pham, C.S.N. Lam, M. Luengo-Oroz, et al., Mapping the landscape of artificial intelligence applications against covid-19, 2020, arXiv preprint arXiv:2003.11336.

[11] C. Menni, A.M. Valdes, M.B. Freidin, C.H. Sudre, L.H. Nguyen, D.A. Drew, S. Ganesh, T. Varsavsky, M.J. Cardoso, J.S.E.-S. Moustafa, et al., Real-time tracking of self-reported symptoms to predict potential covid-19, Nat. Med. 26 (2020) 1037–1040.

[12] J.-F. Gautier, Y. Ravussin, A new symptom of covid-19: Loss of taste and smell, Obesity 28 (5) (2020) 848.

[13] Y. Song, P. Liu, X. Shi, Y. Chu, J. Zhang, J. Xia, X. Gao, T. Qu, M. Wang, Sars-cov-2 induced diarrhoea as onset symptom in patient with covid-19, Gut 69 (6) (2020) 1143–1144.

[14] J.P. Kanne, B.P. Little, J.H. Chung, B.M. Elicker, L.H. Ketai, Essentials for radiologists on covid-19: an update—radiology scientific expert panel, Radiology 296 (2) (2020) E113–E114.

[15] M. Jamshidi, A. Lalbakhsh, J. Talla, Z. Peroutka, F. Hadjilooei, P. Lalbakhsh, M. Jamshidi, L. La Spada, M. Mirmozafari, M. Dehghani, et al., Artificial intelligence and covid-19: Deep learning approaches for diagnosis and treatment, IEEE Access 8 (2020) 109581–109595.

[16] K.W. Chan, V.T. Wong, S.C.W. Tang, Covid-19: An update on the epidemiological, clinical, preventive and therapeutic evidence and guidelines of integrative chinese–western medicine for the management of 2019 novel coronavirus disease, Amer. J. Chin. Med. 48 (03) (2020) 737–762.

[17] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 2013, arXiv preprint arXiv:1301.3781.

[18] A. Grover, J. Leskovec, Node2vec: Scalable feature learning for networks, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 855–864.

[19] M. Fiedler, Algebraic connectivity of graphs, Czechoslovak Math. J. 23 (2) (1973) 298–305.

[20] U. Von Luxburg, A tutorial on spectral clustering, Stat. Comput. 17 (4) (2007) 395–416.

[21] L. Dong, S. Hu, J. Gao, Discovering drugs to treat coronavirus disease 2019 (covid-19), Drug Discov. Ther. 14 (1) (2020) 58–60.

[22] B.N. Rome, J. Avorn, Drug evaluation during the Covid-19 pandemic, New Engl. J. Med. 382 (2020) 2282–2284.

[23] J.-l. Ren, A.-H. Zhang, X.-J. Wang, Traditional chinese medicine for COVID-19 treatment, Pharmacol. Res. 155 (2020) 104743.

[24] L. Luo, J. Jiang, C. Wang, M. Fitzgerald, W. Hu, Y. Zhou, H. Zhang, S. Chen, Analysis on herbal medicines utilized for treatment of COVID-19, Acta Pharm. Sin. B 10 (7) (2020) 1192–1204.

[25] M.Z. Alom, M. Rahman, M.S. Nasrin, T.M. Taha, V.K. Asari, Covid_Mtnet: covid-19 detection with multi-task deep learning approaches, 2020, arXiv reprint arXiv:2004.03747.

[26] O. Gozes, M. Frid-Adar, H. Greenspan, P.D. Browning, H. Zhang, W. Ji, A. Bernheim, E. Siegel, Rapid ai development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis, 2020, arXiv preprint arXiv:2003.05037.

[27] N. Bung, S.R. Krishnan, G. Bulusu, A. Roy, De novo design of new chemical entities (nces) for sars-cov-2 using artificial intelligence, ChemRxiv (2020).

[28] M. Muzammal, R. Talat, A.H. Sodhro, S. Pirbhulal, A multi-sensor data fusion enabled ensemble approach for medical data from body sensor networks, Inf. Fusion 53 (2020) 155–164.

[29] Y. Zhang, R. Gravina, H. Lu, M. Villari, G. Fortino, Pea: Parallel electrocardiogram-based authentication for smart healthcare systems, J. Netw. Comput. Appl. 117 (2018) 10–16.

[30] L. Ang, H.W. Lee, A. Kim, M.S. Lee, Herbal medicine for the management of covid-19 during the medical observation period: A review of guidelines, Integr. Med. Res. 9 (3) (2020) 100465.
[31] W.L. Hamilton, R. Ying, J. Leskovec, Representation learning on graphs: Methods and applications, 2017, arXiv preprint arXiv:1709.05584.
[32] C. Meng, R. Cheng, S. Maniu, P. Senellart, W. Zhang, Discovering meta-paths in large heterogeneous information networks, in: Proceedings of the 24th International Conference on World Wide Web, 2015, pp. 754–764.
[33] Y. Zhang, R. Wang, M.S. Hossain, M.F. Alhamid, M. Guizani, Heterogeneous information network-based content caching in the internet of vehicles, IEEE Trans. Veh. Technol. 68 (10) (2019) 10216–10226.
[34] C. Shi, B. Hu, W.X. Zhao, S.Y. Philip, Heterogeneous information network embedding for recommendation, IEEE Trans. Knowl. Data Eng. 31 (2) (2018) 357–370.
[35] M.E. Newman, The structure and function of complex networks, SIAM Rev. 45 (2) (2003) 167–256.
[36] F. Hu, Y. Zhu, J. Liu, Y. Jia, Computing communities in complex networks using the dirichlet processing gaussian mixture model with spectral clustering, Phys. Lett. A 383 (9) (2019) 813–824.
[37] F. Hu, Y. Liu, Multi-index algorithm of identifying important nodes in complex networks based on linear discriminant analysis, Modern Phys. Lett. B 29 (03) (2015) 1450268.
[38] L.C. Freeman, Centrality in social networks conceptual clarification, Social Networks 1 (3) (1978) 215–239.
[39] P. Bonacich, Some unique properties of eigenvector centrality, Social Networks 29 (4) (2007) 555–564.
[40] U. Brandes, D. Fleischer, Centrality measures based on current flow, in: Annual Symposium on Theoretical Aspects of Computer Science, Springer, 2005, pp. 533–544.
[41] S. Dolev, Y. Elovici, R. Puzis, Routing betweenness centrality, J. ACM 57 (4) (2010) 1–27.
[42] F. Hu, J. Liu, L. Li, J. Liang, Community detection in complex networks using node2vec with spectral clustering, Physica A 545 (2019) 123633.
[43] M.E. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74 (3) (2006) 036104.
[44] N. Zhang, S. Deng, Z. Sun, G. Wang, X. Chen, W. Zhang, H. Chen, Long-tail relation extraction via knowledge graph embeddings and graph convolution networks, 2019, arXiv preprint arXiv:1903.01306.
[45] M.E. Newman, Modularity and community structure in networks, Proc. Natl. Acad. Sci. 103 (23) (2006) 8577–8582.
[46] L. Danon, A. Diaz-Guilera, J. Duch, A. Arenas, Comparing community structure identification, J. Stat. Mech. Theory Exp. 2005 (09) (2005) P09008.
[47] E.B. Fowlkes, C.L. Mallows, A method for comparing two hierarchical clusterings, J. Amer. Statist. Assoc. 78 (383) (1983) 553–569.
[48] A. Feizollah, N.B. Anuar, R. Salleh, F. Amalina, Comparative study of k-means and mini batch k-means clustering algorithms in android malware detection using network traffic analysis, in: International Symposium on Biometrics and Security Technologies, IEEE, 2014, pp. 193–197.
[49] S. Romano, N.X. Vinh, J. Bailey, K. Verspoor, Adjusting for chance clustering comparison measures, J. Mach. Learn. Res. 17 (1) (2016) 4635–4666.