Approaching High-Accuracy Side Effect Prediction of Traditional Chinese Medicine Compound Prescription Using Network Embedding and Deep Learning

ZHENG WANG1,2 (Student Member, IEEE), LIANG LI1, JING YAN3, AND YUANZHE YAO1

1School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China
2School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, NSW 2052, Australia
3First Clinical Medical College, Zhejiang Chinese Medicine University, Hangzhou 310006, China

Corresponding authors: Zeheng Wang (zenwang@outlook.com) and Yuanzhe Yao (yzyao@tsinghua.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61370202.

ABSTRACT In this paper, we realize high-accuracy side-effect prediction of Traditional Chinese Medicine Compound Prescription by introducing network embedding and deep learning. A random walk network that could efficiently interpret the information in the prescription is established from a conventional Bag-of-Word network. After the validation of this random walk network, the highest prediction accuracy reaches 0.908 where a simple five-layer artificial neural network is implemented, rendering this method is promising for Traditional Chinese Medicine side-effect prediction and other medicines with a similar structure such as the compound drugs.

INDEX TERMS Side effect, network embedding, deep learning, traditional Chinese medicine.

I. INTRODUCTION

It has been proven by clinical experiments that Traditional Chinese Medicine (TCM) could be considered as a supplementary treatment in many diseases, such as cancers and mental disorders [11]–[4]. However, the mechanism behind the treatment remains elusiveness [5], [6]. Typically, the TCM treatment is implemented in the form of a compound prescription (CP) containing several herbs and/or minerals. Some researchers believe that the functionality of TCM CP comes from certain latent attributions, which could be supported by the traditional Chinese philosophy [7], [8]; others reckon that it is because of the chemical ingredients [9]–[11].

That contradiction suggests that there is a sort of uncertainty in TCM theory, which has been more or less proven by frequent reports of unexpected side effects (SE) after the use of TCM CP [12]. In radical cases, the application of TCM in the same disease may cause opposite results [13]–[17]. Therefore, the SE of TCM CP, which is an unavoidable challenge, has hindered the further application of TCM. Whereas, recent studies of SE of TCM focus on purely reporting the SE cases with some statistical analysis [18], [19], which is of great insufficiency to clarify the source of the SE and then to predict the SE.

On the other hand, Artificial Intelligence (AI) is an emerging technology that could be used in learning features or interactions among lots of relationship-unknown objects [20], [21]. In this field, the Artificial Neural Network (ANN) is the most representative technique. According to the universal approximation theorem, the most important advantage of ANN is in fitting the laws of paired data where the relationships between these two parts are hard to describe in traditional statistical models. Recently, some novel AI techniques in Representation Learning and Graph Neural Networks, for example, the widely used model Node2vec, have been proposed and investigated thoroughly [22]–[25]. Implemented with that, many strategies have already been adopted and obtained success in the research of disease analysis [26], medical imaging [27], [28]. In this sense, introducing AI into TCM research has attracted extreme attention as well, and hereby, has resulted in many achievements [29]–[33]. Especially, some novel concepts, such as using the convolutional neural network to classify the herbs [34] and exploring the relationship between the treatment action and herbal
property by deep learning [35], have paved the way to extend the research of SE prediction of TCM CP.

In this work, a Node2vec-based SE prediction strategy of TCM CP is proposed and investigated. A model developed in our previous work is used to classify the key attribution of each herb in the TCM CP into cold and hot [12]. Before the prediction of ANN, we also exam the model’s stability and effectivity by the ICC coefficient and Kendall’s correlation coefficient respectively. Taking advantage of these techniques, the critical information, which is the attribution-to-SE relationship, is effectively learned by ANN. And therefore, the SE of the CP could be predicted with very high accuracy, indicating the presented strategy is promising in the TCM clinic decision support, new drug development, and other similar applications.

II. METHOD

The data used in this paper are gathered from the prestigious ancient TCM book Shanghanzabinglun and recently reported papers recording CP that cause SE frequently [12]. The 150 CP in the book Shanghanzabinglun are adopted as the non-SE prescriptions, and, the 73 SE-causing CP are introduced as complementary data.

Before the training and the prediction, all the data are cleaned up properly. Because a Chinses herb usually has several different names in different prescriptions due to historical and cultural reasons. Besides, the dosage unit does not meet international standards. More importantly, there are plenty of cases that are missing information such as dosage in the collected data. Therefore, the data cleaning process is as follows: first, the toxic prescriptions or the prescriptions missing information are all removed. Secondly, all Chinese herbs names and non-standard units are standardized. Finally, the attribution of each Chinese herb is classified into Hot or Cold according to pharmacopeia published by the official institute.

A. NETWORKS CONSTRUCTION

As illustrated in Fig. 1, different TCM CPs are firstly represented by vectors based on the Bag of Word (BOW) model. This representation contains the information of the ingredients with dosages as well as their attributions. Two main attributions, hot and cold, are chosen here because they may be important indicators for SE prediction, according to our previous study [12]. Based on BOW, the cosine similarity among these vectors is calculated, and then, a complex network is constructed (see green circle in Fig. 1) [36]. In this complex network, different nodes represent different CPs, and, each edge’s length represents the similarity between two different CPs.

To further clarify the network construction process, two well-known prescriptions, the Mahuang Decoction and the Guizhi Decoction are processed step by step as in Fig. 2. They are both typical prescriptions developed in Shanghanzabinglun during the Han Dynasty in China, which have been widely adopted to treat many diseases clinically for thousands of years [37]. In the figure, the number 0 or ±1 in the BOW vector indicates the presence or absence of the ingredient in the dictionary that contains all the ingredients used in this research. The sign of the number represents the attribution, namely hot or cold in this paper, of the corresponding herb. Table 1 shows the details of these two prescriptions, where it can be seen that the ingredients, as well as their dosages and attributions, are very similar [38], resulting in an adjacent location (the edge is short) in the ego network.
The equation of cosine similarity above is shown as follows:

\[ S(X, Y) = \frac{\left| \sum_{k=1}^{n} x_k y_k \right|}{\sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}} \]  

(1)

where \( X \) and \( Y \) are vectors of two TCM CP in the BOW model. The component \( x_k \) and \( y_k \) is the \( k^{th} \) component in vector \( X \) and \( Y \) respectively, where \( n \) is the dimension of BOW vectors. Based on the BOW network, as shown in Fig. 1 and Fig. 3, the random walk network embedding, namely node2vec, is implemented, where a group of new vectors is generated (see red rectangle in Fig. 1) to capture the high-level topological information of CPs in the BOW complex network. New vectors generated from node2vec contain the information associated with all CPs along the whole random walking path, shown in Fig. 3.

**B. NETWORKS VALIDATION**

Before verifying the prediction, the evaluation of stability and effectivity of this network embedding method is carried out to validate the network structures.

First, we repeat the construction procedure of the node2vec network for 100 times, in which each construction generates 337 embedding vectors. Using cosine of two adjacent vectors to estimate the similarity of these two vectors, some new lists (for instance blue and orange region in Fig. 4) that each contains a group of similarities associated with one vectors.
certain vector is generated. We use the Intraclass Correlation Coefficient (ICC) to quantify the stability of the relationship between one embedding vector and other vectors located around. In other words, ICC is used to evaluate the similarity of the environment where one vector locates. Here, one list outputs one ICC component, and the total ICC coefficient is achieved by average all same ICC components in 100 times construction.

Detailly, the coefficient ICC could be achieved by the equation below:

\[
ICC = \frac{MS_p - MS_e}{MS_p - (d - 1) MS_e}
\]

where \(MS_p\) is the mean square for rows, \(MS_e\) is mean square for error, \(d\) is the number of simulation times. Each ICC value is evened from 100 times simulations. For example, \(ICC_a\) and \(ICC_d\), respectively contain the pairwise information between node a and other nodes, and, between the node d and other nodes in all 100 simulations.

Then, network embedding effectivity is evaluated by weighted Kendall’s correlation coefficient \(\tau_w\). In other words, the information fidelity between the original BOW network and the proposed random walk network is explored. The value of \(\tau_w\) is between -1 and 1: \(\tau_w > 0\) means the two are positive correlation, \(\tau_w < 0\) means the two are negative correlation. The greater the absolute value of \(\tau_w\) is, the stronger the correlation is. The calculation formula of \(\tau_w\) is as follows:

\[
\tau_w = \frac{2}{m(m-1)} \sum_{1 \leq j < i \leq m} sgn(R_i - R_j)
\]

where \(w_{ij} = v_i v_j, v_i = I[i \leq [(n - 1) p]]\), \(m = [(n + 1) p]\), \(n\) is the number of samples, \(R_i\) is the ordinal number of weight, \(I\) is the indicator function, \(i\) and \(j\) is the index of the value in two compared list, and, \(p\) is the quantity percentage of the samples that agree in two datasets used for comparison [39]. The whole equation could be simplified as follows:

\[
\tau_w = \frac{2}{m(m-1)} \sum_{1 \leq j < i \leq m} sgn(R_i - R_j)
\]

After the network construction and validation, embedding vectors that represent the information of a prescription and its similar derivatives are gathered to construct the deep learning dataset. Then, an ANN model is built to predict the SE of prescriptions with this dataset. The ANN contains five layers with one input layer for inputting the embedding vector and one output layer for the safety prediction. The detail is as shown in Fig. 5.

Detailly, this ANN contains 5 layers including an input layer, three hidden layers, and an output layer, where the input layer has 512 units for receiving the vector derived from the Node2vec model, and the output layer contains two units for generating the evaluation of SE. The three hidden layers that have more than 60 units are designed to fit the complex relationships between latent attributions, such as Cold and Hot, and their side effects in this TCM CP.
III. RESULT AND DISCUSSION

As it is shown in Fig. 6, the BOW network and the random walk network are constructed by the method depicted in the Method section. Fig. 6 (left figure) shows the BOW network where each edge represents a similarity between two prescriptions (nodes). In this figure, the red node represents safe prescriptions, and the green node represents side effect prescriptions. The node size is determined by the degree of a node. Fig. 6 (right figure) shows the random walk network, in which the node is instead a vector that generated from a random walk path beginning with a certain prescription in the BOW network, as it is depicted in the Method section. And, red and green node denotes the initial node in the random walk is safe prescription or not, where the node size represents the degree of the node.

To be specific, the random walk network contains the information of a group of similar prescriptions within one path. In comparison, the BOW network only considers one prescription as one node. The difference is that in the one-prescription-to-one-node network the deep learning does not consider the comprehensive effects of similar prescriptions. However, in the one-path-to-one-node network, the related prescription group is entirely learned in one-shot learning, and hereby, the error coming from the conflicting effects of similar prescriptions is suppressed efficiently.

As a result, the high-accuracy side-effect prediction of Traditional Chinese Medicine CP is realized using network embedding and deep learning. A 5-fold-cross-validation is used to train the five-layer artificial neural network model described in the Method section. This process is repeated 5 times with the highest mean prediction rate of 0.908, shown in Table 2. Five prescriptions, Gancaofuzi Decoction, Xiexin Decoction, Jinkuishenqi Pills, Xiaochaihu Granules, and Angong Niuhuang Pills, where three of them are safe prescriptions and other two prescriptions are unsafe, are chosen to show this process in detail in Fig. 7. Specifically, after the ANN-based evaluation, two indicators (see the right-hand figure), which are the probabilities of the safety and unsafety respectively, are given as a result of the statistics of 5-fold cross-validation. The symbols at the right end indicate whether the prediction results are correct regarding the real labels. By repeating this process five times in each round, the total mean accuracy of prediction is obtained as listed in Table 2.

The good performance of our classifier suggests there may be a non-linear relationship between the topological structures of one prescription and its SE, which could be effectively learned by our model. In our construction of the random walk network, the similar prescriptions are associated by a path. In this way, the topological structure of this network could interpret the SE potential with regards to some fractions of the prescription composition. Thus, this topological
structure could not only reflect SE of one prescription but also further reflect the fact that if a prescription may cause SE a prescription with similar fractions can also cause side effects.

Fig. 8 shows the ICC of the random walk network regarding to the original BOW network. There are 337 nodes in this complex network with an average ICC of 0.604, and each of them is bigger than 0.5. According to a previous study [40], those values of ICC suggest the random walk network is moderately stable. Meanwhile, the calculated Kendall’s correlation coefficient is 0.57, it is a moderate correlation between BOW network and random walk network[41], [42]. This Kendall’s correlation coefficient also indicates that the embedding vectors from node2vec could also effectively interpret the SE information derived from the BOW network.

IV. CONCLUSION
In this paper, a high-accuracy side-effect prediction of Traditional Chinese Medicine CP is realized using network embedding and deep learning. A random walk network that abstractly interprets the similar prescriptions within a path into one node is established from a conventional BOW network. After the validation of this random walk network, the prediction results with the highest mean prediction rate of 0.908 are achieved by implementing a five-layer artificial neural network, rendering this method is promising in SE prediction in TCM and other medicines with similar structure such as the compound drugs.

ACKNOWLEDGMENT
(Zeheng Wang and Liang Li contributed equally to this work.)

REFERENCES
[1] Y. Xiang, Z. Guo, P. Zhu, J. Chen, and Y. Huang, “Traditional chinese medicine as a cancer treatment: Modern perspectives of ancient but advanced science,” Cancer Med., vol. 8, no. 5, pp. 1958–1975, May 2019.
[2] Y.-H. Liao, C.-I. Li, C.-C. Lin, J.-G. Lin, J.-H. Chiang, and T.-C. Li, “Traditional chinese medicine as adjunctive therapy improves the long-term survival of lung cancer patients,” J. Cancer Res. Clin. Oncol., vol. 143, no. 12, pp. 2425–2435, Dec. 2017.
[3] A. Singh and K. Zhao, Treatment of Insomnia With Traditional Chinese Herbal Medicine, vol. 135, 1st ed. Amsterdam, The Netherlands: Elsevier, 2017.
[4] F. Ren, Z. Ma, Y. Shen, G. Li, Y. You, X. Yu, Z. Li, D. Chang, and P. Zhang, “Effects of chai-hua-shan-san capsule for psychogenic erectile dysfunction: Study protocol of a randomized placebo-controlled trial,” Medicine, vol. 98, no. 46, Nov. 2019, Art. no. e17925.
[5] P. Hao, F. Jiang, J. Cheng, L. Ma, Y. Zhang, Y. Zhao, and A. Wiebrocht, “Traditional chinese medicine for cardiovascular disease: Evidence and potential mechanisms,” Deutsche Zeitschrift für Akupunktur, vol. 60, no. 3, pp. 2952–2966, 2017.
[6] C. Liu, C. Zhang, T. He, L. Sun, Q. Wang, S. Han, W. Wang, J. Kong, F. Yuan, and J. Huang, “Study on potential toxic material base and mechanisms of hepatotoxicity induced by dysosma versipellis based on toxicological evidence chain (TEC) concept,” Ecotoxicology Environment Saf., vol. 190, Mar. 2020, Art. no. 110073.
[7] M. Zhou, Y. Hong, X. Lin, L. Shen, and Y. Feng, “Recent pharmaceutical evidence on the compatibility rationality of traditional chinese medicine,” J. Ethnopharmacol., vol. 206, pp. 363–375, Jul. 2017.
[8] P. Li, W. Su, S. Yun, Y. Liao, Y. Liao, H. Liu, P. Li, Y. Wang, W. Peng, and H. Yao, “Toward a scientific understanding of the effectiveness, material basis and prescription compatibility of a chinese herbal formula dan-hong injection,” Sci. Rep., vol. 7, no. 1, May 2017, Art. no. 46266.
Knowledge graph for TCM health preservation: Design, construction, and applications,” *Artif. Intell. Med.*, vol. 77, pp. 48–52, Mar. 2017.

C. Y. Ung, H. Li, Z. W. Cao, Y. X. Li, and Y. Z. Chen, “Are herb-pairs of traditional chinese medicine distinguishable from others? Pattern analysis and artificial intelligence classification study of traditionally defined herbal properties,” *J. Ethnopharmacol.*, vol. 111, no. 2, pp. 371–377, May 2007.

K. Han, L. Zhang, M. Wang, R. Zhang, C. Wang, and C. Zhang, “Prediction methods of herbal compounds in chinese medicinal herbs,” *Molecules*, vol. 23, no. 9, p. 2303, 2018.

X. Sun and H. Qian, “Chinese herbal medicine image recognition and retrieval by convolutional neural network,” *PLoS ONE*, vol. 11, no. 6, 2016, Art. no. e0156327.

Z. Chen, Y. Cao, S. He, and Y. Qiao, “Development of models for classification of action between heat-clearing herbs and blood-activating stasis-resolving herbs based on theory of traditional chinese medicine,” *Chin. Med.*, vol. 13, no. 1, pp. 1–11, Dec. 2018.

M. Gan, “Walking on a user similarity network towards personalized recommendations,” *PLoS ONE*, vol. 9, no. 12, 2014, Art. no. e114662.

H. Wan, L. Pan, Y. Wang, C. Li, L. Yu, H. Zhou, H. Wan, and Y. He, “Pharmacokinetics of seven major active components of Mahuang decoction in rat blood and brain by LC–MS/MS coupled to microdialysis sampling,” *Naunyn-Schmiedeberg’s Arch. Pharmacol.*, 2020, doi: 10.1007/s00210-019-01786-0.

H. Luo, Q.-L. Tang, Y.-X. Shang, S.-B. Liang, M. Yang, N. Robinson, and J.-P. Liu, “Can chinese medicine be used for prevention of corona virus disease 2019 (COVID-19)? A review of historical classics, research evidence and current prevention programs,” *Chin. J. Integrative Med.*, vol. 26, no. 4, pp. 243–250, Apr. 2020.

G. S. Shieh, “A weighted Kendall’s tau statistic,” *Statist. Probab. Lett.*, vol. 39, no. 1, pp. 17–24, Jul. 1998.

T. K. Koo and M. Y. Li, “A guideline of selecting and reporting intraclass correlation coefficients for reliability research,” *J. Chiropractic Med.*, vol. 15, no. 2, pp. 155–163, Jun. 2016.

L. Douw, M. Quaak, S. M. D. Fitzsimmons, S. J. de Wit, Y. D. van der Werf, O. A. van den Heuvel, and C. Vriend, “Static and dynamic network properties of the repetitive transcranial magnetic stimulation target predict changes in emotion regulation in obsessive-compulsive disorder,” *Brain Stimulation*, vol. 13, no. 2, pp. 318–326, Mar. 2020.

M. von Bernstorff, J. Rapp, F. Bausenhart, M. Feierabend, I. Ipach, and U. K. Hofmann, “Estimating braking performance in osteoarthritis of the knee or hip with a reaction timer,” *Orthopaedic Surgery*, vol. 11, no. 2, pp. 248–254, Apr. 2019.