International Journal of Latest Research in Engineering and Technology (IJLRET)
ISSN: 2454-5031
www.ijlret.com || Volume 06 - Issue 09 || September 2020 || PP. 04-12

Early detection for infectious disease outbreaks using federate learning and blockchain

LI Daoxing[1], LIU Haiqing[1]
[1](*School of Control and Computer Engineering, North China Electric Power University,China)*

Abstract: Early detection of infectious disease outbreak is important to confirm the outbreak of infectious diseases quickly. In recent years, in order to improve the real-time performance of early detection of infectious disease outbreaks, many studies using deep neural networks to extract information from medical data to forecast whether infectious disease outbreaks. However, due to the medical data contains a lot of the patient's privacy information, centralized model training mode will cause patient privacy problems. According to the above problem, we designed an infectious disease early warning model based on federate learning and blockchain, which decentralized trains symptom BERT model to extract symptom vectors by using local electronic medical record data and uses symptom vectors to detect infectious diseases outbreaks. To improve the effect of symptom BERT model training, we also use the reputation blockchain to ensure reliable data.
Keywords: early detection, disease outbreaks, federate learning, blockchain

I. INTRODUCTION

Infectious diseases are disease caused by a pathogen that can spread widely from person to person or from person to animal. Infectious diseases spread quickly and bring great losses to individuals, societies and countries if they break out. [1]In mid-November 2002, Severe Acute Respiratory Syndrome (SARS) broke out in China, causing a total of 8422 cases of infection and 919 deaths. During this period, China's consumption and transportation industries were greatly impacted. In 2014, Ebola broke out in Liberia, Sierra Leone, Guinea and other African countries. The Ebola epidemic has left these countries with stagnant or even negative economic growth. [2]Novel Coronavirus epidemic (COVID-19) broke out in China in 2020. As of 24:00 on March 12, 2020, it had caused 80,981 infections and 3,173 deaths in China, and there were 114 cases of COVID-19 in 114 countries and regions around the world. The COVID-19 epidemic "has the characteristics of a global pandemic". If the outbreak of infectious diseases can be early warned, epidemic prevention measures can be used to control the epidemic in local areas, so as to minimize the negative impact of the epidemic. Therefore, early warning of infectious diseases plays an important role in the process of infectious disease prevention.[3]In this paper, based on federate learning and blockchain technology, a new kind of infectious disease early warning model is proposed, which guarantees the patients privacy while training the neural network model. This model also uses the information in the electronic medical records in hospital to early warning the outbreak of infectious diseases. So that the relevant departments to take relevant measures as early as possible, which may help to reduce the negative effects of epidemic.

II. RELATED WORK

In recent years, the early warning model of infectious diseases has been widely studied. According to the technologies used in the early warning model of infectious diseases, we can divide the early warning model of infectious diseases into two categories: the early warning models of infectious diseases based on statistics and the early warning models of infectious diseases based on big data technology.

2.1 Early warning models of infectious disease based on statistics

Before 2010, the technologies related to big data were not yet mature, statistics methods were widely used in the research related to the surveillance and early warning of infectious diseases. According to the three spread dimensions of infectious diseases, the early-warning model of infectious diseases based on statistics can be divided into three categories: time models, spatial models, and spatio-temporal models. The time models use the information of the time and number of confirmed infectious diseases, such as the regression method [4], the time series method, the statistical process graph method and other methods to detect the outbreak of infectious diseases. [5]The spatial models use the information related to the geographical location of patients to monitor whether there is spatial clustering of infectious disease patients, so as to issue early warning information.[6]For examples, the CUMUS chart of utilization of space[7], spatial scanning statistical method, spatial autocorrelation analysis. The spatio-temporal models use the time and space aggregation detection method to judge whether the number of cases in the fixed area and the space aggregation hotspot area reach the warning

threshold. If so, the system will send an early warning signal. The commonly used methods include Bayesian network[8], PANDA[9], WSARE[10], etc. The biggest problem with traditional early warning models of infectious diseases is that they rely on hospitals to diagnose and report cases of infectious diseases to give early warning of the outbreak of infectious diseases, so they are more effective for known types of infectious diseases. However, for unknown emerging infectious diseases, it is difficult to unify diagnostic criteria within a short period of time, leading to the fact that the number of confirmed cases of new infectious diseases reported by hospitals is likely to be lower than the actual number of patients. A relatively small number of confirmed cases will greatly reduce the real-time performance of the early warning model.

2.2Infectious disease early warning systems based on big data technology

After 2010, the Internet has been widely popularized around the world, and the huge amount of data generated by the Internet has led to the rise of "big data" technology. Big data technology enables the early warning system of infectious diseases to use network information such as website search volume, instead of relying on hospital confirmed cases of infectious diseases to predict the outbreak of infectious diseases, thus improving the real-time performance of the early warning system of infectious diseases. In 2009, Ginsberg et al. [11]used the search data for influenza on Google website to predict the arrival of influenza outbreaks a week in advance of the local CDC. In recent years, with the maturity of deep learning technology, more and more studies have been conducted on the use of social media information for early warning of infectious diseases. Aramaki et al. [12]classified Flu-related Twitter by combining natural language processing technology with support vector machine. SENTINEL, an infectious disease system designed by Serban et al. [13], combines clinical data with Twitter data to give early warning of outbreaks and their severity. Although infectious disease early warning system based on the technology of large data on the early warning and real-time performance is excellent, but because of cultural differences, most people in China after the ill will first choose consult family doctor or directly go to the hospital, very few people in the social media to share their health condition, so the early warning system of infectious diseases based on social media data of in China is not widely used.

2.3 Innovation points of this paper

In this paper, a new kind of infectious disease early detection model is proposed, which uses electronic medical records as data source.This model training BERT model [14]to extract symptoms and the disease related information, and use common adverse events evaluation criteria (CTCAE v5. [15]) to classification symptoms to constitute the infectious disease symptom vector. This method also uses EARS [16]to monitor infectious disease symptoms vector and detect the outbreak of infectious diseases. The innovations of this paper are as follows: 1. BERT can automatically extract information from the electronic medical record to detect the outbreak of infectious diseases with better real-time performance compared with the traditional warning based on the number of confirmed cases. 2. Using of federate learning method to train symptom BERT model to extract symptom vector from electronic medical records can protect the patient's medical data privacy.3. The reputation blockchain is used to select reliable computing nodes to ensure training effect of the symptom BERT model.

III. PRELIMINARY

3.1Medical text information extraction

With the rapid increase of the number of electronic documents related to medicine, the task of extracting medical text information has been paid more and more attention. With the help of advances in natural language processing, many researchers build deep and deep learning models to extract information from medical texts. However, due to the large number of specialized vocabularies in the biomedical field, the corpus distribution is quite different from texts in the general field, so the direct application of natural language processing technology to medical text information extraction is often not very effective. Bio BERT**[17]** pretrained BERT model by using corpus on PubMed and PMC, and fine-turning the model by using NER, RE and QA task data set, and finally improved the effect compared with the most advanced model in these three tasks. However, because this paper intends to use Chinese electronic medical record as the data source, the BioBERT model must be trained in Chinese medical corpus before it can be applied in the early warning system of infectious diseases.

3.2 Federated learning

Since we use electronic medical record that stored in local hospitals as training data to train the BERT model, we adopt the federal learning approach. Federated learning is a new type of distributed privacy-protecting machine learning technology that enables distributed nodes to collaboratively train a global machine learning model without having to upload private local data to a central server. Each node from a central server to

obtain an initial parameter $\boldsymbol{\theta}$ for global machine learning model, in this article for symptomBERT model $\boldsymbol{\Phi_{BERT}(\theta)}$, using local data for training. After the node completes the training with local data, the parameters of the model or gradient will be uploaded to the central server to update the parameters of the global model. Suppose n nodes have local training data set $\boldsymbol{s_n}$, The total of n nodes contains the training data $\sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} \boldsymbol{s_n} = \boldsymbol{S}$. Under the above assumptions, the objective function of federated learning can be written in the following form

$$\min_{\Phi} l(\Phi_{BERT}(\theta)) = \sum_{n=1}^{N} \frac{s_n}{S} l_n(\Phi_{BERT}(\theta)) \quad (1)$$

$$l_n(\Phi_{BERT}(\theta)) = \frac{1}{s_n} \sum_{i \in s_n} f_i(\Phi_{BERT}(\theta)) \quad (2)$$

Where f_i is the loss value generated by sample i in node N.

First t iterations in the global model training, each node uses the local data sets will upload the training parameters $\theta^{(t)}$for global model $\Phi_{BERT}(\theta^{(t)})$ for the gradient of Λ_n, assume that each node using a stochastic gradient method to train the model, the update rate for α_n, the local model parameters as follows: in t + 1 time $\theta_n^{(t+1)} = \theta_n^{(t)} + \alpha_n \Lambda_n$. After the model aggregation stage, the global model will aggregate all the parameters of the local node model t$\theta^{(t)} = \sum_{n=1}^{N} \frac{s_n}{S} \theta_n^{(t)}$. If the nodes use the local data training model with high accuracy and reliability, the convergence rate of the loss function in the model aggregation stage will be accelerated. Therefore, it is very important to select the reliable node training model.

3.3Alliance Chain

In order to find reliable nodes for model training, we plan to use the subjective logic model with multiple weights [18]to calculate the reputation value of each node. The nodes with high reputation value have a greater chance of being selected as the training node of federated learning. We plan to use the alliance chain as the basic technology for the storage, calculation, and management of reputation value. Blockchain is an unmodifiable and anti-jamming distributed bookkeeping book. The most famous application of block chain is the electronic currency bitcoin. Bitcoin uses the public chain as its technical foundation, while we plan to use the alliance chain as the basic technology in the federated learning. Alliance chain is a more efficient and practical blockchain technology. In the alliance chain, only pre-selected computing nodes can participate in the consensus mechanism. Therefore, compared with the public chain, the alliance chain is a more lightweight blockchain technology with faster negotiation speed, and it is also more suitable to use the reputation management learned in the federation. [19]

IV. MODEL DESIGN

4.1Structure of infectious disease warning model

Our detection model for infectious disease is shown in Fig.1 The most important part in the model is extracting symptom vector of infectious disease from the text message in the EMR. The second part is the C2&C3 warning algorithm.

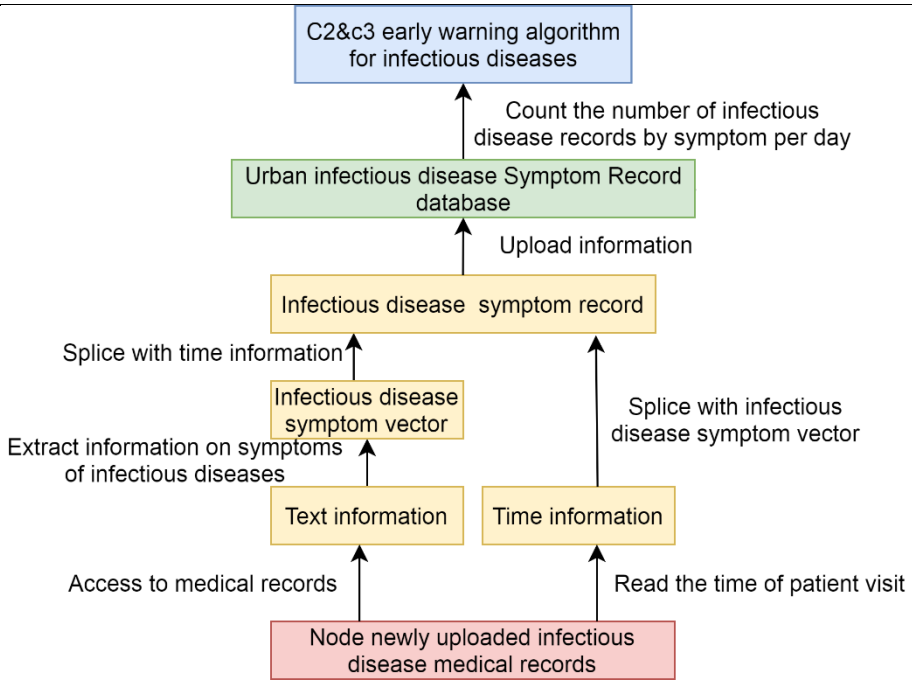

Fig.1 Structure of infectious disease warning model

Since the generation of electronic medical record will precede the infectious disease card (suspected cases will not fill in the infectious disease card before diagnosis), the extraction of infectious disease symptom vector from the text information of electronic medical record for prediction ensures the high real-time performance of our detection model. This has also been improved in part by taking BERT, the latest natural-language processing model, to ensure this information is extracted accurately. The model structure of BERT model [14] is shown in Figure 2. The BERT model apply a multi-layer two-way Transformer[20] as the encoder and decoder. Unlike traditional cyclic neural networks, Transformer is built on a self-attention module, which enables it to better construct dependencies between words that are far apart in a sentence and has a good performance for text-type tasks. In addition, in order to obtain the final infectious disease symptom vector, we added a full join layer at the last layer in the original BERT model to adjust the output dimension. There are 24 Transformer modules with just 1,024 hidden layers each and 16 attention mechanisms all together with nearly 340million parameters. Because the model is large, it needs to be pre-trained to extract the symptom information of medical text and then it is fine-tuned to produce a vector of infectious symptoms. Due to the large amount of data required for pre-training, the medical records we used are stored in the local nodes (including but not limited to the medical records of infectious diseases). The training is completed by MaskLM method [14]. The task of the model to randomly mask 15% of the words in your medical record to read the entire sentence and to predict what the veiled word really is. After the training is over, we fine turning this model to fit the task of extracting the symptom vector for a contagious disease.

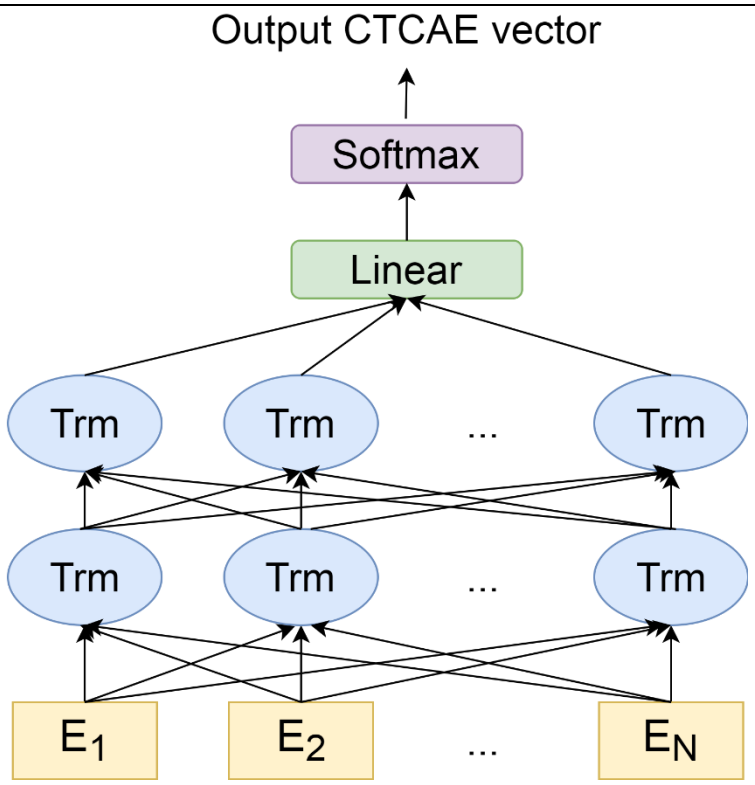

Fig.2 The model structure for Symptom BERT

After uploading infectious disease symptoms of medical records from local nodes, infectious disease symptoms will be recorded in database in the from of daily group according to the symptoms of infectious disease medical record number. The recorded will be used by the C2&C3 algorithm to detect the outbreak of infectious diseases. C2&C3 algorithm is applied in EARS system which is developed by the U.S. centers for disease control and prevention. C2&C3 algorithm is suitable for the early warning surveillance for infectious disease symptoms of comprehensive information. The detail of C2&C3 algorithm is shown below. Assuming that a total of J infectious disease symptoms are monitored, the number of cases with the J symptom uploaded on the first day of the infectious disease symptom record database is N_i^j. The C2&C3 algorithm first calculates the $Median^j$ of the time series of the number of cases with the symptom within 7 days.

$$Median^j = mad(N_i^j, N_{i-1}^j, , , N_{i-6}^j) \tag{3}$$

Where $mad()$ represents the median operation, and then the median standard deviation mad is calculated.

$$MAD^j = \sqrt{\sum_{k=i-6}^{i} (N_i^j - MAD)^2} \tag{4}$$

C2 warning algorithm is an early warning for a certain symptom. Assuming that the symptom is J, the condition for the symptom to trigger C2 warning is:

$$if N_i^j - median^j \geq 3MAD^j, then C2^j = true \tag{5}$$

When 3 or more C2 alerts are triggered, C3 alerts of higher level will be triggered, which can be expressed in the form of formula as follows

$$if count(C2^j \geq 3), then C3^j = true \tag{6}$$

The accuracy of the C2&C3 algorithm depends on the symptom information, but due to the massive amount of data and the privacy problems of EMR, we proposed federated learning method to train the model.

4.2Federated Learning and training Framework for early detection Model of Infectious Diseases based on Blockchain

To train the Sypmtom BERT model using electronic medical record data that containing a large amount of patient privacy data, we have designed a federated learning framework based on the consortium blockchain

for the model designed in [21]. The federated learning framework consists of two layers, the application layer and the blockchain layer, as detailed in Figure 3.

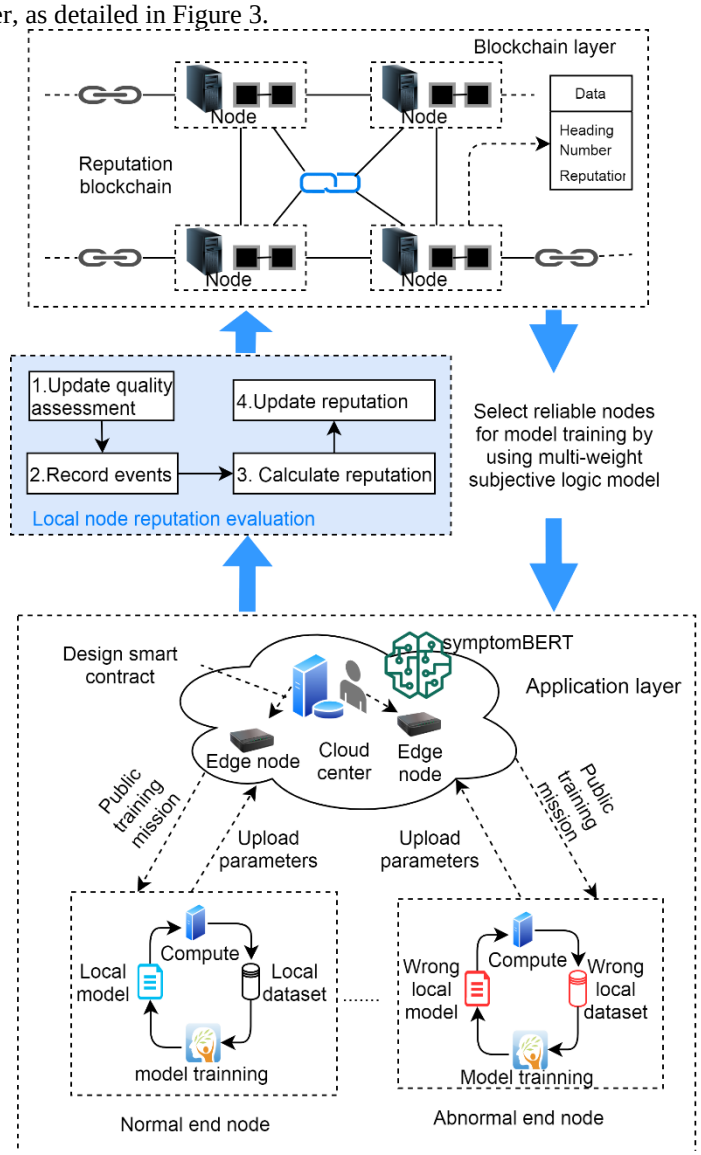

Fig. 3 Federate Learning and Training Framework for Early Warning Models of Infectious Diseases

In the application layer, we consider using Ethernet as the underlying structure, and the whole network adopts a mixed topology structure of ring and star. The terminal node of the network is the deep learning server deployed in the hospital, and the widely deployed switching equipment is used as the side node of the network. The terminal nodes not only have advanced computing power, but also store a large number of electronic medical records that containing patient privacy data. The feature of federated learning is that when the training manager issues the task for training symptom BERT, the endpoint node ensures patient data privacy by training the model directly with local data without uploading it to the centre's server. We create a smart contract to select the node to train for the model because some of them may not be online or temporarily resuscitated by another deep-learning model training task. Each selected node uses local data to train the model to get a new BERT model parameter. When the selected nodes have completed their training the local model parameters, they upload parameters to the training management center to update the global model. The training management center aggregates the parameters of the local model and updates them to the global model. This step is repeated until the model globally converged. The widely distributed edge nodes enable the end nodes to communicate with the training management center in real-time. The training management center will generate the reputation of the end node based on the quality of model update and training. Reputation management is carried out by the blockchain layer, we call this blockchain as reputation blockchain.

In the block chain layer, as the edge nodes are connected with the end nodes and the training management center at the same time, we use the edge nodes to select the eligible end nodes as miners. The criteria for selecting miners is a reputation-based approach. [18] After the reputation of the end node is verified unanimously by edge nodes, the reputation value of the end node will be stored in the data block of the reputation blockchain. Due to the distributed and non-tampering characteristics of the blockchain, the reputation value stored in the data block is still the permanent and public evidence even in the case of disputes and damages. When calculating reputation, we not only consider the reputation value provided by the model training center from this training model of symptom BERT, but also consider the reputation value of end nodes combined with the behaviors of the training model of deep learning.

V. EXPERIMENT

Our experiment is divided into two parts: the first part extracts the symptom vector by the trained symptom BERT model and the second part uses the trained symptom BERT model to detect and give warning of infectious disease by C2&C3. Our experiment was carried out on a virtual machine on the cloud platform. The virtual machine was configured with a NIVIDIA 2080Ti GPU with 16G memory and Ubuntu 16.04LTS operating system.

In the first experiment we imply the model based on PySyft, the federated learning framework with PyTorch kernel. Because of the electronic medical record data contains patient's privacy information, which is difficult to obtain, we only found 529 text type electronic medical record data related to infectious diseases from the Chinese clinical case database. So, we adopt transfer learning method to fine-tuning BERT model in the open source project chinese-clinical-NERto extract symptom vectors. First, we manually annotate 529 textual EMR data, and divide it into two parts: training set and test set. The training set contains 400 EMR samples, which are stored in four federal learning data nodes, with 100 at each node. The test set contains 129 samples of electronic medical records. During this fine-tuning stage, we just propagate back the parameters in the Linear and Softmax layers in the model. The loss value during training is shown in Figure 5.

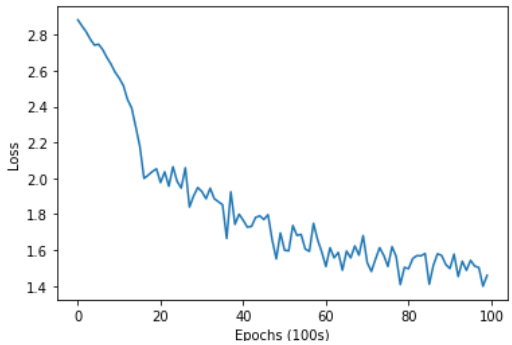

Fig.5 The value of loss during training Symptom BERT model

Figure 5 shows that the loss value decreased along with the increase of the training batch and finally coverageto1.5, indicating that the model converged on the EMR training set. The trained model got 0.80 accuracy and 0.83 precision on test dataset.

In the second part of the experiment, we used 3191 pieces of electronic medical record data related to infectious diseases from 2015 to 2017 provided by a private hospital. 2,956 pieces were structured data which could be directly extracted. 235pieces were unstructured text data that need to be processed by the trained Symptom BERT model. We extract the three symptom information: Cough, Fever and Expectoration and monitor them with the C2&C3 algorithm for detecting the outbreak. The temporal distribution of symptom information is shown in Figure 6.

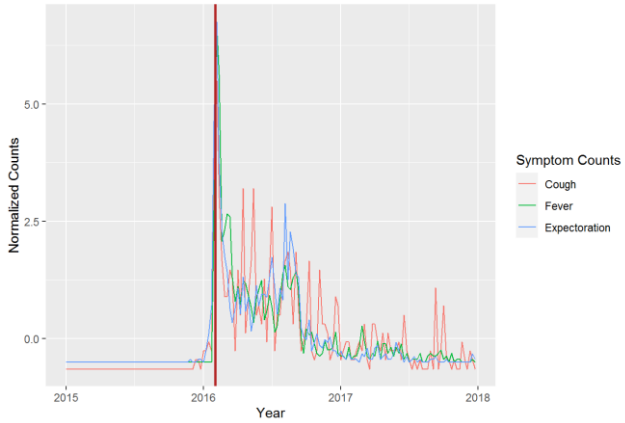

Fig.6 Time distribution of symptom information

From Figure 6, we can see that C2&C3 algorithm produced C2 warning signals of cough, fever, and sputum on February 15, 2016 (red vertical line in Figure 6) and C3 warning signals. On February 16, 2016, the number of coughs, fever and sputum reached its maximum value. This shows that The C2&C3 algorithm can effectively monitor and warn the symptoms of infectious diseases.

VI. CONCLUSION

This paper proposes a new infectious disease warning model based on federated learning and blockchain technology. This model utilizes electronic medical record data stored in hospitals to obtain the symptom information and utilize C2&C3 algorithm to judge whether there is an outbreak of infectious disease. By using the symptom information in the electronic medical records, C2&C3 algorithm can give early warning of infectious diseases outbreak. Therefore, the model is more suitable for detecting emerging infectious diseases outbreak. The main innovation of this model is that the patient privacy in medical record data is protected by the federated model while training symptom BERT and reputation blockchain is used to select reliable node for training symptom BERT.

REFERENCES

[1]. Steele, L., E. Orefuwa, and P. Dickmann, *Drivers of earlier infectious disease outbreak detection: a systematic literature review.* International Journal of Infectious Diseases, 2016. **53**: p. 15-20.
[2]. Bank, W. *Update on the Economic Impact of the 2014 Ebola Epidemic on Liberia, Sierra Leone, and Guinea*. 2015; Available from: https://www.worldbank.org/en/topic/macroeconomics/publication/economic-update-ebola-december.
[3]. Xia, H., et al., *Synthesis of a high resolution social contact network for Delhi with application to pandemic planning.* Artificial Intelligence in Medicine, 2015. **65**(2): p. 113-130.
[4]. Stroup, D.F., et al., *Detection of aberrations in the occurrence of notifiable diseases surveillance data.* Statistics in Medicine, 1989. **8**(3): p. 323-329.
[5]. Unkel, S., et al., *Statistical methods for the prospective detection of infectious disease outbreaks: A review.* Journal of the Royal Statistical Society. Series A: Statistics in Society, 2012. **175**(1): p. 49-82.
[6]. Fitzgibbon, W.E., et al., *Modelling the aqueous transport of an infectious pathogen in regional communities: application to the cholera outbreak in Haiti.* Journal of the Royal Society Interface, 2020. **17**(169): p. 8.
[7]. Yamada, I., P.A. Rogerson, and G. Lee, *GeoSurveillance: A GIS-based system for the detection and monitoring of spatial clusters.* Journal of Geographical Systems, 2009. **11**(2): p. 155-173.
[8]. Wang, Z., et al., *System inference for the spatio-temporal evolution of infectious diseases: Michigan in the time of COVID-19.* Computational Mechanics: p. 24.
[9]. Shen, Y., et al., *Estimating the joint disease outbreak-detection time when an automated biosurveillance system is augmenting traditional clinical case finding.* Journal of Biomedical Informatics, 2008. **41**(2): p. 224-231.
[10]. Carnevale, R.J., et al., *Evaluating the utility of syndromic surveillance algorithms for screening to detect potentially clonal hospital infection outbreaks.* Journal of the American Medical Informatics Association, 2011. **18**(4): p. 466-472.

[11]. Ginsberg, J., et al., *Detecting influenza epidemics using search engine query data.* Nature, 2009. **457**(7232): p. 1012-1014.
[12]. Aramaki, E., S. Maskawa, and M. Morita. *Twitter catches the flu: Detecting influenza epidemics using Twitter*. in *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. 2011.
[13]. Şerban, O., et al., *Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification.* Information Processing and Management, 2019. **56**(3): p. 1166-1184.
[14]. Devlin, J., et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv e-prints, 2018. arXiv:1810.04805.
[15]. Puzanov, I., et al., *Managing toxicities associated with immune checkpoint inhibitors: consensus recommendations from the Society for Immunotherapy of Cancer (SITC) Toxicity Management Working Group.* Journal for ImmunoTherapy of Cancer, 2017. **5**(1): p. 95.
[16]. Hutwagner, L., et al., *The Bioterrorism Preparedness and Response Early Aberration Reporting System (EARS).* Journal of Urban Health, 2003. **80**(2 SUPPL. 1): p. i89-i96.
[17]. Lee, J., et al., *BioBERT: a pre-trained biomedical language representation model for biomedical text mining.* Bioinformatics (Oxford, England), 2020. **36**(4): p. 1234-1240.
[18]. Kang, J., et al., *Toward Secure Blockchain-Enabled Internet of Vehicles: Optimizing Consensus Management Using Reputation and Contract Theory.* IEEE Transactions on Vehicular Technology, 2019. **68**(3): p. 2906-2920.
[19]. Kang, J., et al., *Enabling Localized Peer-to-Peer Electricity Trading Among Plug-in Hybrid Electric Vehicles Using Consortium Blockchains.* IEEE Transactions on Industrial Informatics, 2017. **13**(6): p. 3154-3164.
[20]. Vaswani, A., et al. *Attention Is All You Need*. arXiv e-prints, 2017. arXiv:1706.03762.
[21]. Kang, J., et al., *Incentive Mechanism for Reliable Federated Learning: A Joint Optimization Approach to Combining Reputation and Contract Theory.* IEEE Internet of Things Journal, 2019. **6**(6): p. 10700-10714.