A risk assessment approach of hypertension based on mobile crowd sensing

Huan-Huan Zhao¹²³, Zu-Chang Ma¹ AND Yi-Ning Sun¹

¹Institute of Intelligent Machines, Chinese Academy of Sciences
Hefei, China
²University of Science and Technology of China
Hefei, China
³School of Computer and Information Engineering, Chuzhou University, China
E-mail: {zhhh206; zuddrj; suddrj}@126.com

Abstract: Mobile crowd sensing (MCS) makes full use of the sensing and wireless communication capabilities of smart mobile devices to effectively collect real-time information. It makes it possible for monitoring people’s health condition in real time. Our health information collected through MCS can be used to improve the healthcare service. Hypertension is a widespread chronic disease, and preventing hypertension can effectively reduce the incidence of cardiovascular disease. In this paper, we propose a hypertension risk assessment approach based on mobile crowd sensing, which allows for real time health monitoring and warning. In order to stimulate the enthusiasm of MCS volunteers, optimized communication model is used to reduce the communication cost of non-data-users. Additionally, the current hypertension risk status of patients will be feed back to them in real time. In our approach, binary logistic regression is used to select risk factors of hypertension, and then the risk factors are used as the inputs of BP neural network to construct the risk prediction model. Furthermore, the hypertension risk is further divided into low risk, medium risk and high risk through cumulative distribution function. 4498 samples from a community health service center in Hefei area were used to evaluate the performance of the proposed approach. The experimental results show that the proposed approach can provide real-time, effective monitoring and dynamic feedback of the hypertension risk, offering a novel clinical tool for the early warning of hypertension. The proposed approach also provides a general framework for risk assessment of other chronic diseases.

Keywords-Mobile crowd sensing; Hypertension; Risk assessment; BP neural network

1. INTRODUCTION

Nowadays, with the aging of population and the popularization of sub-health, the traditional interaction between doctors and patients has been unable to meet the increasing demands of physical health and medical care of individuals. With the popularity of smart phones and the development of wireless technology and sensors, a new sensing structure named mobile crowd sensing (MCS) has emerged [1]. Mobile crowd sensing refers to people sensing information about people and surrounding environment using smart phones and other mobile terminals carried around, and then providing useful information and services to users based on the collected information. MCS provides a new way for people to sense the world. It enables anyone who performs sensing tasks through mobile terminals to participate in the sensing process and provides
universal services. As an important application branch, mobile health crowd sensing is a major component of MCS. It combines mobile crowd sensing with telemedicine. Participants upload health data collected by mobile terminals to cloud server and enjoy various services provided by medical institutions. Among the applications of health service, developing and constructing an effective risk prediction model for chronic diseases is of great value in chronic disease management. At present, the service of disease risk assessment is mostly implemented according to the experience of doctors, which takes a lot of time and resources of one-on-one interventions and on-going assessment by clinicians or other health care providers. Therefore, there is an increasing demand for an automated disease risk assessment system, which can make full use of the collected information and ultimately provide medical decision support. The system is of great value for disease prevention. In order to realize this requirement, we need to collect large amount of effective data in real time, at the same time, an effective disease risk assessment approach is also required.

Chronic diseases are responsible for 70% of all deaths worldwide. Hypertension is a widespread chronic disease. At present, the total number of hypertensive patients can reach 972 million worldwide and this number continues to grow [2]. Hypertension is also an important risk factor of cardiovascular and cerebrovascular diseases [3, 4]. Studies have shown that hypertension is the first risk factor leading to death and the third factor of economic burden of disease [5]. According to the latest investigation [6], from 2012 to 2015, the prevalence rate for hypertension of Chinese adults had reached 27.9%, and keeps increasing continuously. Especially for adolescents, the situation is also unoptimistic [7]. The awareness, treatment and control rates of hypertension for Chinese adults were 51.6%, 45.8% and 16.8% respectively, which indicates that prevention and control of hypertension is unsatisfactory and needs to be improved in China at present.

The best way to improve the situation of hypertension is to prevent the occurrence of hypertension. Researches have shown the occurrence of hypertension is affected by Heredity, lifestyle etc. Improving unhealthy lifestyle before hypertension can reduce the risk of hypertension and prevent the occurrence of hypertension effectively [8]. Therefore, the risk assessment of hypertension plays a vital role in preventing hypertension. Currently, the risk assessment of hypertension for Chinese people is still a challenge. For one thing, the data used to build the model are mainly from hospitals or specialized medical institutions. Most of the data are non-real-time with a long update cycle, and it is difficult to feed the assessment result back to individuals in real time. For another, the selection of hypertension risk factors and modeling methods are of vital importance on the performance of risk assessment model. In order to provide a basis for the primary prevention of hypertension for Chinese population, a hypertension risk assessment approach based on mobile crowd sensing was proposed in this paper. First, hypertension relative information including medical data, lifestyle data, individual characteristic data were collected through mobile crowd sensing network, then binary logistic regression was used to analysis of risk factors of hypertension from the collected information, and then hypertension prediction model was constructed based on BP neural network training by selecting the optimal threshold to solve the unbalanced problem of training sample. Based on the risk value, cumulative distribution function was used to classify hypertension risk by controlling error rate. The remainder of this paper is organized as follows. Section 2 summarizes related research work. Section 3 illustrated our proposed
Section 4 reports the experimental results on real world medical data set and presents the performance study. Finally, Section 5 gives conclusions and future work of this study.

2. RELATED WORK

A large number of researches have constructed considerable work in investigating remote health monitoring based on wireless sensor network (WSN). For example, in a multi-source healthcare architecture called MSHA [9], multi-source data are from sensors in wireless body domain networks and text-based input anywhere. Kaur et al. [10] proposed a general multi-sensor fusion method to realize continuous monitoring of remote patients using wireless body sensor network. Health risk assessment and decision making algorithm were used to make reliable health management decisions. However, the location of WSN is fixed and cannot monitor health information of individuals in various situations in real time. Mobile crowd sensing provides new ideas to solve these problems. Mehdi et al. [11] proposed a general MCS reference architecture for healthcare. Collection, networking and transmission of healthcare information in MCS were further studied [12, 13]. Yang et al.[14] and Wen et al. [15] explored how to take effective measures to encourage individuals to participate in MCS actively. Pryss et al. [16] taken a specific disease as an example, MCS was used in conjunction with data anonymization to provide a large number of daily data sets of patients at a low cost, making it possible for clinical trial anywhere. In order to create an experimental database using cloud service platform as a service through MCS, Jovanovic et al. [17] proposed a machine learning method to determine the current hypertension status through ECG signals and to collect data of hypertensive patients. It indicated that it is feasible to determine hypertension status according to information collected from MCS. However, it only considered ECG signals to determine hypertension or not, in fact, there are many risk factors and indicators of hypertension.

Many researchers have studied the risk factors and prediction models of hypertension, which indicate that different populations have different characteristics of hypertension. Age, gender, BMI, diabetes and blood pressure are the most common predictors [18]. Logistic regression is used in [19, 20] and neural network is used in [21-23] to establish corresponding hypertension prediction models for different populations. Ture et al. [24] compared the performance of hypertension prediction models established by decision tree, statistical method and neural network, and found that neural network performed the best. The Framingham heart research center calculated the risk of hypertension in the next 1, 2, 4 years based on cohort studies [25]. Muntner et al. [26] and Mika et al.[27] researched the validation of Framingham prediction model for local residents, and the results showed that the model is validate for European population. But a South Korean study [28] showed that the Framingham hypertension risk prediction model underestimated the risk of South Koreans. Researchers from Japan, India, Iran and Swedish have developed risk prediction models of hypertension for local residents [29-32]. The research of Zheng et al. [33] showed that the Framingham hypertension risk prediction model couldn't assess hypertension risk of Chinese rural population effectively. Thus it is urgent to develop a risk prediction model of hypertension for Chinese people. Chien et al. [34] established a hypertension risk assessment model for Chinese-Taiwanese, and Li et al. [35] established
a 15-year hypertension risk prediction model based on a cohort study of 3,899 people in 11 provinces of China, but the lifestyle risk factors were not included in both the models, which limits their application in primary prevention of hypertension. In order to evaluate the hypertension risk of Chinese steel workers, Wu et al. [36] established a hypertension risk score model for steel workers using LVQ neural network, which has a high accuracy. However, this model is only applicable to steel workers.

Although some progress has been made in the research of hypertension risk assessment, there are still some difficulties and limitations in the practical application of the present study in the prevention of hypertension for Chinese population. As for the collection of modeling data and the feedback of assessment results, telephone follow-up or face-to-face questionnaire are the most common approaches for collecting information, these approaches may cost lots of time and resources, which make them difficult to be applied to a large population. Mobile devices can monitor activity and behaviors effectively anytime and anywhere, and provide real time, dynamic feedback in response to collected information. Refer to the assessment of hypertension risk, it is challenging to predict the risk of hypertension using risk factors. Improper selection of risk factors and modeling methods may affect the performance of the model. Research has showed that neural network has good classification performance, but now it is mostly used in predicting the occurrence of hypertension or not, but in fact, risks assessment of hypertension is what we need. In order to solve the problems above, this study proposes a risk assessment approach for hypertension based on mobile crowd sensing. Experimental results show that the risk assessment approach can identify the risk of hypertension effectively and give real time feedback to users, providing novel guidance for primary prevention of hypertension.

3. PROPOSED APPROACH

In this section, we first present the framework and the flow of hypertension risk assessment approach. And then introduce the components of the framework. In the approach, a large amount of hypertension relative data, including medical data, lifestyle data and individual characteristic data can be collected via mobile equipment, which establishes a strong foundation for hypertension risk assessment. Such risk assessment system learns medical knowledge from the collected information, and simulated the doctors’ diagnosis using machine learning to provide a reliable risk assessment and diagnosis. As illustrated in Figure 1, the entire framework of the proposed approach can be divided into three modules: data collection module, communication network module and data processing module.
The specific flow of the approach is shown in Figure 2. When starting the task of building the model, in first step, health-relative information is collected by participants through smartphone. In the second step, the collected information is transmitted to cloud server for data storage through communication network. Third, the collected information is processed in the cloud server. Data preprocessing is used to deal with missing values and finish data standardization. For data processing, first, logistic regression is used to analyze the risk factors of hypertension. Second, BP neural network is used to construct a model for outputting the risk value of hypertension. Third, the cumulative distribution function was used to stratify the risk of hypertension by controlling the error rate. Thus, we construct the hypertension risk assessment model. The model can provide real-time service for participants and some institutes.
3.1 Data collection module

Health-related information is multi-dimensional, physiological indexes during working, exercising and even travelling are important components of health-related information. This kind of dynamic health information cannot be obtained from traditional medical system. Mobile crowd sensing provides a solution to this problem. Mobile crowd sensing is a kind of voluntary action, participants volunteered to submit their health-related information for the sake of common interests and benefits. Therefore, appropriate incentive mechanism should be taken to maintain the enthusiasm of participants. In this study, each time participants submit sensing information, the risk level of hypertension will be assessed and feedback to them as incentives. The hypertension relative information are gathered by smartphones, medical equipment and then sent to the cloud server center. Data collection module is responsible for collecting medical data, lifestyle data, and individual characteristic data through mobile equipment. For example, body fat calculator can measure weight and body fat, blood pressure can be measured by a blood pressure machine, blood fat can be measured by a blood lipid machine, blood glucose can be measured by a blood glucose machine, ECG can be measured by a single derivation electrocardiography sensor, temperature probe can measure temperature, the band can collects information such as heart rate, sleep and physical activity, lifestyle data can be collected through questionnaire on the APP in the
smartphone. Therefore, a large number of hypertension relative information can be relayed from the device to the smartphone and then to the Internet.

3.2 Communication network module

The data collected from the information collection module is transmitted to the cloud server center through wireless access communication network and IP-based core network. Among them, the wireless access communication network is the network that connects the customer and service provider directly. It varies from a few hundred meters to a few miles in diameter and includes everything between the core network and the user terminal. The core network adopts fiber structure with high transmission rate. It provides a variety of services for users accessing communication networks.

Wireless communication protocols for access communication networks include ZigBee (IEEE 802.15.4), Bluetooth (IEEE 802.15.1), Wi-Fi (IEEE 802.11b) and 3G/4G/5G. In this study, ZigBee is excluded because it is not supported by smart phones at present. In order to meet the communication accuracy requirements, the distance required for Bluetooth between communication devices is between 10 meters and 100 meters. Therefore, in this system, the detection results of medical devices are transmitted to smart phones through Bluetooth. Wi-Fi is an important communication technology that can be used in mobile devices. Compared with Bluetooth, the transmission speed of Wi-Fi is faster and it can spread over a longer distance. Therefore, mobile users who have access to Wi-Fi can use Wi-Fi for data transmitting. For some scenarios where Wi-Fi is not available, mobile devices can transmit data over cellular network, which is the most widely used technology because of easy access to mobile phones despite their high cost. At present, full Wi-Fi coverage has not been achieved, extra costs may be incurred when using cellular network for data transmitting. Some users may not will to participate in the mobile crowd sensing due to the extra cost, especially for non-data-plan users. Therefore, inspired by the research [37, 38], the communication model we used in cellular network is shown in figure 3, the specific communication process is as follows:

Step 1: deploy and initialize the base station of data collection point;

Step 2: participants who are data-plan users can apply to become temporary base station nodes (in this way, they will be rewarded accordingly);

Step 3: among the base stations within the allowed communication range, participants select the base station with the shortest distance from themselves to submit the sensing data, in this case, Bluetooth and Wi-Fi are mostly used communication media;

Step 4: base station nodes puts all collected data forward to the cloud service center through one or multiple hop.

This communication model can reduce the energy consumption of data transmission and balance the load of cellular network while encouraging users to participate in mobile crowd sensing.
3.3 Data processing module

Cloud service center will process lots of health information, many kinds of analysis with health information will be implement to provide medical and health management decision support for health service providers, and also provide real-time health care services for the community and individuals. Our approach of hypertension risk assessment consists of three steps. Firstly, binary logistic regression was used to analyze the risk factors of hypertension. Secondly, BP neural network is used to construct hypertension prediction model which can output the risk value of hypertension. Third, the cumulative distribution function was used to stratify the risk of hypertension by controlling the error rate.

(A) Logical regression analysis

Logistic regression is a statistical method, which can explore the relationship between a categorical dependent variable and several independent variables. It has been widely used in epidemiological risk factor analysis. According to the number of possible values of dependent variables, logistic regression can be divided into binary logistic regression and multivariate logistic regression. Taking the risk factors for hypertension as an example, binary logistic regression analysis is required.

Dependent variable \( y \) stands for hypertension or not, independent variables \( x_1, x_2, \cdots, x_n \) stand for \( n \) diseases associated with hypertension. Logistic regression model formula is shown in (1), where \( \beta \) stands for the coefficient of \( x_i \).

\[
\ln \left( \frac{p}{1 - p} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n
\]  

(1)
We can express the predicted probability of \( y \) as formula (2). Logistic prediction model can be used to analyze \( x_1, x_2, \ldots, x_n \) one by one, and finally obtain influential factors from the \( n \) independent variables.

\[
p = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)]}
\]  

(2)

In this paper, binary logistic regression analysis is used to explore the correlation between independent variables (medical data, lifestyle data, individual characteristic data) and dependent variable (hypertension or not), so as to explore the risk factors of hypertension.

(B) BP neural network

Neural network is a supervised learning method, which can learn the law of the training data, thus can transform input data into the appropriate output, because the nerve can solve the problem of collinearity and interaction among variables and have strong tolerance of fault, the neural network has been applied widely to forecasting diseases [39, 40].

BP neural network is a feed forward neural network model for classification and pattern recognition. The structure of typical neural network includes input layer, hidden layer and output layer, full connection is between one layer and another, that is, any neuron of each layer has connections to all neurons of the layer before, so that the relationship from input to output can be learned. The network is developed by updating weight iteratively using the back propagation algorithm and optimization algorithm, which contributes to reducing the error between the desired output and the actual output progressively. Algorithm 1 presents the pseudo-code of BP neural network.

**Algorithm 1:** Neural network learning with the back propagation algorithm

**Input:** \( N \) train samples \((x_i, y_i), (x_2, y_2), \ldots, (x_n, y_n)\), with inputs \( x_i \) and corresponding outputs \( y_i \), \( x_i = (x_{i1}, x_{i2}, \ldots, x_{ik}) \), where \( x_{i1}, x_{i2}, \ldots, x_{ik} \) is \( k \) features of \( x_i \).

**Output:** a BP neural network

1: Initializing weight \( W(W_0) \) and threshold value \( B(b_j) \) of network
2: Input a training sample \((x_i, y_i)\), \( i = 1, 2, \ldots, n \).
3: While \( E \) not satisfying {
   Calculate the output value of each node \( j \) of the hidden layer and output layer:
   \[
   L_j = \sum_{i=1}^{m} w_{ij} \cdot z_i + b_j 
   \]
   \[
   z_j = \frac{1}{1 + e^{-L_j}}
   \]
   Where \( m \) is the number of nodes in the previous layer
4: Calculate the error of each node \( j \) in the output layer:
   \[
   err_j = z_j \cdot (1 - z_j) \cdot (y_i - z_i)
   \]
   For (the last hidden layer: the first hidden layer)
   Calculate the error of each cell \( j \) of the hidden layer:
err_i = z_i^*(1-z_i) * \sum_j (err_j * w_{ij})

Where i is the neuron in the next layer of j

5: Adjust the weights and thresholds between two layers adjacent:

\[ w_{ij} = w_{ij} + \mu * err_i * z_j \]
\[ b_j = b_j + \mu * err_j \]

Where \( \mu \) is the learning rate

(C) Risk stratification method

Based on the risk value from the output of neural network, we divided the risk of hypertension into high, medium and low, using the accumulation function to find boundaries between high risk and medium risk, medium risk and low risk in the training data [41]. The boundary between high risk and medium risk is chosen when 10% of non-hypertension was assessed as high risk, the boundary between medium risk and low risk is chosen when 10% of hypertension population assessed as low risk. Algorithm 2 presents the pseudo-code of risk stratification method.

**Algorithm 2: risk stratification method**

**Input:** \( m \) risk values \( r_{hi}, r_{h2}, \ldots, r_{hm} \) for hypertension, \( r_{hi} \) is the risk value of hypertension sample \( i \),

\( k \) risk values \( r_{li}, r_{l2}, \ldots, r_{lk} \) for non-hypertension, \( r_{li} \) is the risk value of non-hypertension sample \( i \)

**Output:** boundary between risk levels \( B = \{B_1, B_2\} \), \( B_1 \) is the boundary between medium risk and low risk, \( B_2 \) is the boundary between high risk and medium risk.

1: for \((r = 0; r <= 1; r = r + 0.001)\) do
2: \( h = 0 \)
3: for \((r': r_{hi} - r_{hm})\) do
4: \( \text{if } (r' <= r) \) do
5: \( h = +1 \)
6: end if
7: end for
8: if \((h / m >= 10\%)\) do
9: \( B_1 = r \)
10: break
11: end if
12: end for
13: for \((r = 1; r >= 0; r = r - 0.001)\) do
14: \( l = 0 \)
15: for \((r': r_{li} - r_{lk})\) do
16: \( \text{if } (r' >= r) \) do
17: \( l = +1 \)
18: end if
19: end for
20: if (1 / k >= 10%) do
21: \( B_2 = r \)
22: break
23: end if
24: end for

4. EXPERIMENTS

This section investigate the performance of the proposed approach against previous hypertension risk prediction model in hypertension detection, which is measured in terms of the sensitivity, specificity, AUC and Youden’s index. In addition, the accordance of the training set and testing set is also investigated.

4.1 Datasets and risk factors of hypertension

This subsection illustrates the preparation of the simulation data. The data analyzed in this paper is from a community health service center in Hefei area. There were a total of 4550 samples from April 2018 to July 2019, among which 52 samples were with missing values, and a total of 4498 cases were complete data. There were 1098 cases of hypertension and 3400 cases of non-hypertension in the sample. In addition to general physical examination and laboratory measurements, relevant lifestyle behaviors and habits, including history of chronic disease, lifestyle and physical activity habits were collected through face-to-face questionnaires by a professional worker in the community health service center. According to the diagnostic criteria of the Chinese guidelines for the prevention and treatment of hypertension, hypertension is defined as having been diagnosed as hypertension in hospital, or having an average systolic blood pressure greater than 140 or diastolic blood pressure greater than 90 during physical examination. According to previous researches, the risk factors of hypertension mainly include: gender, age, height, weight, Body Mass Index (BMI), family history of hypertension (“family history” was used for short in the rest of article), smoking, high-salt diets, diabetes, hyperlipidemia, physical activity, etc. [18]. Firstly, all these factors were included in the study. BMI is defined as weight (Kg) divided by square of height (m). Diabetes is defined as being told suffering diabetes by the doctor or fasting blood glucose is greater than 7.0mmol/L or 2 hours postprandial blood glucose is greater than 11.1mmol/L. Hyperlipidemia was defined as being told suffering hyperlipidemia by the doctor or total cholesterol is greater than 572mmol/L or triglycerides is greater than 1.70mmol/l or HDL-cholesterol is less than 9.0mmol/L. Family history is defined as parents or brothers and sisters having hypertension. Smoking is defined as total amount of smoking is more than 100 or smoking now. High-salt diet is defined as taking more than 6g of salt per day at ordinary times. Physical activity is divided into three categories, those with regular exercise habits (at least 3 times a week, 30 minutes each time) are defined as “regular exercise”, those only with housework but without regular exercise were defined as “housework”, and those without any regular physical activity are defined as “no exercise”.
Logistic regression analysis has the capacity to determine significant factors. So the binary logistic regression was used to analyze the sample data, the dependent variable was hypertension or not, 1 stand for hypertension and 0 stand for non-hypertension. The independent variables including continuous variables: age, BMI, and categorical variables: gender, family history, smoking, high-salt diets, diabetes, hyperlipidemia, physical activity. When setting critical value of statistical significance p at 0.05, the correlation of dependent variable and each independent variable showed that age, BMI, family history, smoking, high-salt diets, diabetes, hyperlipidemia and physical activity was significantly correlated to hypertension(p<0.05), which indicate they are predict indicators of hypertension that should be included into the model. While gender was not significantly correlated to hypertension(p>0.05), but clinical studies have shown that gender is one of the influence factors of hypertension, so we would include gender as hypertension risk factors into the model.

4.2 hypertension prediction model based on BP neural network

Hypertension prediction model is established based on BP neural network. Where the input layer is constitute of nine risk factors, they are age, gender, BMI, family history, smoking, high-salt diets, diabetes, hyperlipidemia and physical activity. A node is used to represent the output; the output is a decimal number 0 or 1, which can show the risk of hypertension, and the greater the value means the higher the risk of hypertension. We set one hidden layer, thus can not only reduce the amount of computation, but also prevents over-fitting. As regard to the number of nodes in the hidden layer, which is not only related to the number of nodes in the input and output layer, but also related to the complexity of the problem to be solved, the type of conversion function and the characteristics of sample data. In order to avoid the phenomenon of "over-fitting" and ensure the performance of the network, the most basic principle to determine the number of hidden layer nodes is minimizing the number of hidden layers on the premise of meeting the requirements of precision. The number of hidden layers in this model is determined to be 10, which contributes to the smallest correction error of the model. The Tanh was selected as activation function referring to previous research results [22]. Therefore, the predict model is shown in Fig 4.
4.3 Experimental results and analysis

In order to implement the risk assessment approach of hypertension, sample data were randomly divided into training samples and testing samples according to the ratio 7 (a total of 3148 cases, in which 768 cases hypertension, 2380 cases non-hypertension) : 3 (a total of 1350 cases, in which 330 cases hypertension, 1020 cases non-hypertension). The training samples were used to train the model, while the testing samples were used to verify the validity of the model. Referring to previous studies [42], to generate the binary hypertension status (Yes or No), the threshold is set at 0.5. But due to the unbalanced character of our sample data, which means hypertension population is far less than non-hypertension population, we consider all thresholds. The optimal threshold is the threshold which makes the sum of sensitivity and specificity of the training set reach the maximum value. A person is identified as hypertension when the output of neural network is larger than the threshold value, otherwise, he will be predicted as non-hypertension. We use the testing set to verify the performance of the model.

The accuracy, sensitivity, specificity, area under the curve (AUC) and Youden’s index is used to evaluate the diagnostic performance of our model. Accuracy presents the probability of correctly identifying hypertension and non-hypertension. Sensitivity is defined as the probability of correctly identifying hypertension. Specificity refers to the probability of correctly identifying non-hypertension. AUC is defined as the area bounded by the coordinate axis under the receiver operating characteristic (ROC) curve, as a value, the classifier with larger AUC has better performance. Youden’s index is equal to the sum of specificity and sensitivity minus 1. The sensitivity and specificity at different threshold for both training set and testing set are shown in Fig 5, and the ROC curves of training set and testing set are shown in Fig 6. Under the optimal threshold, for
training set, the accuracy was 77.78% (95% Confidence Interval (CI): 77.73%-77.81%), the sensitivity was 87.73% (95% CI: 87.69%-87.78%), the specificity was 74.61% (95% CI: 74.57%-74.65%), and the AUC was 0.88 (95% CI: 0.87-0.90). The accuracy, sensitivity, specificity and AUC of the testing set were 77.95% (95% CI: 77.91%-77.99%), 87.82% (95% CI: 87.78%-87.86%), 74.75% (95% CI: 74.70%-74.80%), and 0.88 (95% CI: 0.86-0.90) respectively. It is obvious that the training set have good consistency with the testing set in accuracy, sensitivity, specificity and AUC.

We compared the performance of the prediction model in this paper with the previous hypertension prediction models, and the results are shown in table 1. This model outperforms other models in specificity, Youden’s index and AUC, and also performs well in sensitivity.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Youden’ index (%)</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fernando2018[19]</td>
<td>logistic regression</td>
<td>77.00</td>
<td>68.00</td>
<td>45.00</td>
<td>0.73</td>
</tr>
<tr>
<td>Wang 2015[22]</td>
<td>neural network</td>
<td>72.76</td>
<td>67.96</td>
<td>40.72</td>
<td>0.77</td>
</tr>
<tr>
<td>Mevlut Ture 2005[24]</td>
<td>logistic regression</td>
<td>90.48</td>
<td>64.1</td>
<td>54.58</td>
<td>0.79</td>
</tr>
<tr>
<td>This approach</td>
<td>neural network</td>
<td>87.82</td>
<td>74.75</td>
<td>62.57</td>
<td>0.88</td>
</tr>
</tbody>
</table>

The above results show that our model has good performance on predicting hypertension. But in fact, our goal is to predict the risk of hypertension, and finally realize early warning of hypertension. Good performance of hypertension diagnostic, which can ensure the accuracy of risk degree, is the basis of risk classification. So risk stratification of hypertension need to be implemented further. Different from hypertension prediction, we let output of the neural network is not 0 or 1, but a continuous value between 0 and 1, which shows the risk of hypertension. We divided the risk of hypertension into high, medium and low, using the accumulation function to find boundaries between high risk and medium risk, medium risk and low risk in the training
data. The boundary between high risk and medium risk is chosen when 10% of non-hypertension was predicted as high risk, the boundary between medium risk and low risk is chosen when 10% of hypertension population predicted as low risk. The result of risk stratification is shown in Fig 7.

We used the testing set to verify the performance of the risk stratification method. The results were shown in table 2. Among 330 hypertension population, 22 (6.67%) were predicted as low risk, 88 (26.7%) were predicted as medium risk, 220 (66.7%) were predicted as high risk. Among 1020 non-hypertension population, 652 (63.9%) were predicted as low risk, 250 (24.5%) were predicted as medium risk, and 118 (11.6%) were predicted as high risk. which shows that our risk stratification method can effectively identify the risk of hypertension and provide a new idea for the early warning of hypertension.

Table 2. Performance of the risk stratification method in testing set

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>#Low risk</th>
<th>%Low risk</th>
<th>#Medium risk</th>
<th>%Medium risk</th>
<th>#High risk</th>
<th>%High risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>hypertension</td>
<td>330</td>
<td>22</td>
<td>6.67%</td>
<td>88</td>
<td>26.7%</td>
<td>220</td>
<td>66.7%</td>
</tr>
<tr>
<td>Non-hypertension</td>
<td>1020</td>
<td>652</td>
<td>63.9%</td>
<td>250</td>
<td>24.5%</td>
<td>118</td>
<td>11.6%</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS

In the paper, we propose a hypertension risk assessment approach based on mobile crowd sensing. The participants can access the system through smart phones and upload their hypertension-related information anytime and anywhere. In order to encourage participants, the assessment results will be feed back to them in real time. Additionally, optimized communication model is used to reduce the communication cost of non-data-users. Then the collected information can be relayed to the cloud service center through the communication network. The data are processed and analyzed in the cloud.
service center to provide corresponding health care services. In our approach, Binary logistic regression was used to identify risk factors of hypertension, and then BP neural network with risk factors as input was used to calculate the risk value of hypertension. Finally, further risk stratification of hypertension was realized by cumulative distribution function.

4498 sample data in Hefei area were used to implement the simulation experiment. Results show that the average accuracy, sensitivity, specificity and AUC of this model are 77.95%, 87.82%, 74.75%, 0.88 respectively. In addition, the performance of training set is well consistent with the testing set, thus the model outperform the previous hypertension prediction model. Referring to risk level, the error was controlled within 10%, which indicate it can assess the risk of hypertension effectively. Compared with the traditional health care service mode, this approach can collect a large number of hypertension related information effectively and provide real time, dynamic feedback in response to the collected information, which will effectively improve the interaction between individuals and doctors and the benefits of health care. This study provides a new idea for the early warning of hypertension. Although this paper only considers hypertension, this approach is also applicable to risk assessment of other chronic diseases. For future, we will explore effective approaches to alleviate network congestion which is caused by massive real-time data transmission.

Acknowledgment

This work was supported by the fund support of Major Science and Technology Project of Anhui Province (18030801133), the University Natural Science Foundation of Anhui Province (KJ2017B16), Natural Science Foundation of Anhui Province (1608085MF127), and the University Natural Science Foundation of Anhui Province (KJ2017A419).

References

4. S. Lewington, R. Clarke, N. Qizilbash, R. Peto, R. Collins, and C. Prospective Studies, “Age-specific relevance of usual blood pressure to vascular mortality: a


---

**Huan-Huan Zhao** (赵欢欢) received the M.S. degrees in Computing Application Technology from Anhui Normal University, China. She is currently pursuing the Ph.D. degree in the Institute of Intelligent Machines, Chinese Academy of Sciences, China. She is also a lecturer at School of Computer and Information Engineering, ChuZhou University, China. Her research interests include wireless communication networks and health information technology.
Zu-Chang Ma (马祖长) received his Ph.D. degree in precision instrument and machinery from University of Science and Technology of China in 2007. He is currently a professor at Institute of Intelligent Machines, Chinese Academy of Sciences. His research interests include health information technology and public health promotion service theory, technology, method and equipment.

Yi-Ning Sun (孙怡宁) received his Ph.D. degree in electrical engineering from University of Rome in 1993. He is currently a professor at Institute of Intelligent Machines, Chinese Academy of Sciences. His research interests include Technology, system and mode of non-medical health promotion service and Intelligent information processing technology.