2014 IEEE International Conference on Internet of Things (iThings 2014), Green Computing and Communications (GreenCom 2014), and Cyber-Physical-Social Computing (CPSCom 2014)

A Logistic Regression and Artificial Neural Network-based Approach for Chronic Disease Prediction: a Case Study of Hypertension

Aiguo Wang, Ning An, Yu Xia, Lian Li
School of Computer and Information
Hefei University of Technology
Hefei, China
e-mail: wangaiguo2546@163.com, ning.g.an@acm.org

Guilin Chen
School of Computer and Information Engineering
Chuzhou University
Chuzhou, China
e-mail: glchen@chzu.edu.cn

***Abstract*—The global trend of population aging and the continuing maturity of the Internet of Things (IoT) technology drives the rapid development of health care. In the comprehensive applications of IoT technology, developing and constructing a prediction model for chronic diseases is a great improvement to healthcare technology as well as an exploration of IoT technology on the data-analysis and decision-making level. Considering that early detection, diagnosis and screening of hypertension plays a significant role in the prevention and reduction of the onset of cardiovascular diseases as well as the improvement of quality of life, it is of great value to figure out hypertension-related risk factors and further establish a model for the prediction of hypertension with the identified risk factors. Thus, in this paper, we put forward to integrate logistic regression analysis and Artificial Neural Networks (ANNs) model for the selection of risk factors and the prediction of chronic diseases by taking a case study of hypertension. First, binary logistic regression model was applied on experimental dataset collected from Behavior Risk Factor Surveillance System (BRFSS) to select factors statistically significant to hypertension in terms of the pre-defined *p*-value. Then, a Multi-Layer Perception (MLP) neural network model with Back Propagation (BP) algorithm was constructed and trained for the prediction of hypertension with the selected risk factors as inputs to ANNs. Experimental results showed that our proposed approach achieved more than 72% prediction accuracy acceptable in the diagnosis of hypertension and that the Area Under the receiver-operator Curve (AUC) was more than 0.77. The results indicate that integration of logistic regression and artificial neural networks provides us an effective method in the selection of risk factors and the prediction of hypertension, as well as a general approach for the prediction of other chronic diseases.**

Keywords-logistic regression; artifical neural network; hypertension prediction

I. INTRODUCTION

With the continuing maturity and rapid development of Internet of Things (IoT) technology, along with its powerful capacity in sensing information, collecting and communicating data with one another, collaboratively analyzing the context and intelligent monitoring and decision-making, IoT technology has been gradually applied to a variety of spectrums ranging from logistics tracking, transportation to medial and health care fields [1]. Concurrent with the development of IoT technology is the global population aging that strains governments' ability to provide better health care. In addition to increasing costs of health care, chronic disease also directly affects the quality of life of individuals and their family members as well, which drives further researches of the applications of IoT technology in health care. Among these, developing and constructing an effective prediction model for chronic diseases is of great value in healthcare and also a specific application case of IoT technology on the data-analysis and decision-making level.

Chronic disease such as hypertension, diabetes, heart diseases and cancer, is a long-lasting health conditions that can be controlled yet not cured. Data from World Health Organization show that chronic disease is also the major cause of premature death around the world. Although chronic diseases are among the most common and costly health problems, they are also among the most preventable and most can be effectively controlled through reasonable measures. Since many chronic diseases are linked to lifestyle choices that within our own hands to change, therefore, identifying risk factors associated with a certain disease and further constructing a prediction model would be of great importance in the early prevention and effective management of chronic diseases.

Hypertension is a chronic medical condition that affects a wide range of population, particularly the older adults after the age of 55, and even becomes prevalent among adolescents in both developing and developed countries [2]. Besides the fact that prevention and management of hypertension consumes a wealth of medical resources and health care services, resulting in unbalanced medical service distributions and definitely putting on the society considerable financial burdens, hypertension is also a major risk factor for the occurrence and development of cardiovascular diseases such as stroke, heart failure, chronic kidney disease, etc., which are the leading causes of the high morbidity and mortality rates [3,4,5,6]. Consequently, early detection, diagnosis and screening of hypertension is an necessity in the prevention and reduction of the onset of cardiovascular diseases as well as leading to improved quality of life of individuals suffering from hypertension and their families, and potentially saving enormous lives; on the other hand, investigation of the hypertension risk factors are particularly drawing interests from public health and health care researchers with the aim to bring down the onset of

978-1-4799-5967-9/14 $31.00 © 2014 IEEE
DOI 10.1109/iThings.2014.16

hypertension of individuals and improve their health conditions through early warning and prevention.

A number of researchers and medical staff have conducted considerable work in the investigation of hypertension risk factors and in the construction of effective and efficient models for the prediction of hypertension with potential risk factors. There are a variety of factors that are relevant to hypertension prediction mainly including demographics, anthropometry body surface scanning data, clinical test results and even molecular-level data such as genetics, proteins; and a large number of theories and methods, including machine learning and statistical analysis techniques, are employed as powerful tools to facilitate the prediction and diagnosis of hypertension. For example, Ture, et al. conducted a comparative experiment to compare the hypertension prediction accuracy of nine commonly used classifiers on experimental dataset and their experimental dataset consists of demographics, behavior information and clinical laboratory data. On the basis of experimental results, they concluded that Multi-Layer Perceptron (MLP) neural networks and Radial Basis Function (RBF) neural networks performed better than the other three decision trees and four statistical algorithms [2]. The approach proposed by Blinowska, et al. achieved satisfactory prediction accuracy through the application of Bayesian statistical method in the prediction of hypertension by incorporating both prior knowledge and possible costs of wrong decisions, while one of the deficiencies of their study is the difficulty in the collection of sufficient numbers of experimental cases and in ensuring the integrality of each case since Bayesian method is built on statistical theory [7,8]. Besides the use of clinical laboratory data, researchers also turn to other types of available data to improve the prediction accuracy of hypertension. For example, Hsu, et al focused their attentions on the exploration of the relation between hypertension and three-dimensional anthropometric scanning data such as the circumferences of waist, wrist, gluteal, etc. and associated individual medical profiles, and their experimental results demonstrated the effectiveness of anthropometric data in the prediction of hypertension [4]. To investigate the mechanism of hypertension in molecule level, Caulfield, et al conducted a research to identify the genetic factors associated with essential hypertension. Their work presents us novel insights in the pathogenesis mechanisms and prediction of hypertension and the design and discovery of potential therapeutic targets [9]. These researches achieve satisfactory performance in the prediction of hypertension; however, there exist some difficulties and limitations in actual use, especially in hypertension surveillance for a large population. First, the utilization of clinical data, anthropometric body surface scanning data and/or genomic data achieves higher prediction accuracy, but it is not suitable and practical for hypertension prediction in a large population since it involves complex operation processes and costs much, which hinders the collection of sufficient hypertension cases. Second, since being lack of clear clinical effects in the early stage of hypertension and not taking it seriously, individuals easily disregard the occurrence of hypertension, which leads to serious complications potentially [10]. Third, clinical data and/or genomic data are good indicators for the prediction of hypertension, but they present less information about hypertension risk factors, which is of great value in the early prevention and self-management of hypertension. To enable early-stage prevention and management of hypertension in an efficient and economic but effective way and facilitate hypertension surveillance in a large population, developing and constructing an effective hypertension prediction model with easily observed and collected factors are in urgent need.

Since lifestyle behaviors such as drinking, smoking habits and physical activity level contribute to the occurrence and development of hypertension, many researchers have conducted studies in the construction of hypertension diagnosis and prediction model by integrating behavior risk factors and demographics such as age, sex, height and weight with clinical laboratory data [2,4,11,12,13,14]. Compared with clinical laboratory data, anthropometric body surface scanning data and genomic data, behavior information are easily collected and meaningful in the prevention and management of hypertension and more suitable for use in a large population. Lifestyle risk factors could be indicators for hypertension to remind individuals to avoid or circumvent unhealthy behaviors and prediction model could be used in large-scale hypertension surveillance without measuring their blood pressures using instruments. Therefore, selecting significant risk factors and further establishing a prediction model with these factors definitely facilitate the prevention and management of hypertension.

In this paper, we proposed a logistic regression and artificial neural network-based approach for chronic disease prediction by taking hypertension as a study case. Through collecting and cleansing the experimental data publicly available from Behavior Risk Factor Surveillance System (BRFSS) of Centers of Control and Prevention (CDC), we first utilized the binary logistic regression analysis to select risk factors with significant *p*-value. Then, we constructed and trained a Multi-Layer Perceptron (MLP) neural network with Back Propagation (BP) algorithm with the selected factors as inputs to predict whether an individual suffering from hypertension or not. In the construction and training of artificial neural networks, three rule-of-thumbs were employed to narrow down the search space of the parameter values in neural networks towards a balanced tradeoff between speed and accuracy.

Contributions of our research mainly include:

1) integrating logistic regression analysis and Artificial Neural Networks model in the selection of risk factors and the prediction of chronic diseases. Although we just considered hypertension as a study case, it could be used for the prediction of other chronic diseases with corresponding risk factors supplied.

2) presenting detailed discussion of the selection of Artificial Neural Networks architecture and the setting of relevant parameters, which directs researchers in the choice and usage of neural networks towards a balanced tradeoff between speed and accuracy.

3) experimental results showed that the proposed approach achieved satisfactory performance, which

demonstrates the feasibility of the approach in the prediction of not only hypertension, but also other chronic diseases.

This paper is organized as follows: research backgrounds, related work and main contribution statements were presented in introduction section; collection and preparation of hypertension experimental dataset, logistic regression analysis and artificial neural networks model related knowledge were illustrated in materials and methods section; in experimental design and result analysis section, we detailed the construction and the selection of parameters of artificial neural networks, and presented the experimental results of the risk factors with significant *p*-value and the prediction performance achieved by artificial neural network model. The last section concluded the paper with a brief summary.

II. MATERIALS AND METHODS

A. Dataset and Hypertension Risk Factors

Experimental dataset about hypertension was collected from the Behavior Risk Factor Surveillance System (BRFSS) of Centers of Disease Control and Prevention (CDC), data of which are publicly available and downloadable from its website [15]. BRFSS, which has a long history in behavioral and chronic disease surveillance, is the world's largest and on-going telephone health survey system. BRFSS is mainly for tracking and measuring individual health conditions and risk behaviors that contribute to the leading causes of high morbidity and mortality rates in adult population, aged 18 years and older in the United States yearly since 1984. Its survey covers a wide range of health risk factors, preventive health practices and health conditions, including hypertension, diabetes and cancers, and other common chronic diseases. By collecting and recording a variety of health-related information, BRFSS facilitates us to conduct researches to investigate the relations between some specific chronic diseases and behavior information and demographics of an individual.

BRFSS questionnaire consisting of core component, optional modules and state-added questions, is designed by a working group of BRFSS state coordinators and CDC staff. Each item in BRFSS survey system records the reply to each question from an individual. For example, for hypertension: corresponding survey is "Have you been told by a doctor, nurse, or other health professionals that you have high blood pressure?" and its reply is either "YES" or "NO", in which the former means that the individual suffers from hypertension, while the later representing one is not with hypertension; in the same manner, other questions like "About how much do you weight without shoes?" and "About how tall are you without shoes?" The value of other survey items could be obtained in the similar way. BRFSS website provides relevant questionnaire, coding form and detailed illustrations [15].

Through combining the survey items in BRFSS with the potential hypertension risk factors used by previous researchers as discussed in introduction section, we chose 13 survey items as candidate factors and illustrations to each item was presented in Table I. After excluding cases with missing values and transforming the coding of survey item in BRFSS, finally, we got the experimental dataset from the year of 1996 to 2005, which consists of 308,711 cases with one target variable, i.e. hypertension or not, and 13 independent variables relevant to hypertension.

TABLE I. DESCRIPTION OF VARIABLES OF EXPERIMENTAL DATA

No.	Variable	Variable description
1	AGE	'age'
2	SEX	'sex'
3	HEIGHT	'height in inches'
4	WEIGHT	'weight in pounds'
5	MARITAL	'marriage status'
6	EDUC	'education level'
7	INCOME	'income level'
8	EXERANY	'exercises during past month'
9	DIABETES	'ever told having diabetes'
10	TOLDHI	'ever told blood cholesterol high'
11	SMOKE100	'smoke more than 100 in total'
12	SMOKEDAY	'smoke frequency now'
13	ALCDAY	'drink frequency'
14	BPHIGH	'ever told blood pressure high'

B. Logistic Regression Model

Logistic regression model, a type of statistical regression analysis technology, has the capacity to measure the relationship between a categorical dependent variable and one or more independent variables, and is extensively used in numerous disciplines such as medical, bioinformatics and social science fields [30]. According to the number of values of the dependent variable, logistic regression model is categorized into binomial and multinomial regression analysis. In binary logistic regression analysis, dependent variable is usually coded as "0" or "1" to denote an individual suffering from or getting away from a certain disease. In the case of hypertension, the logistic regression model computes the probability of the target disease y (y=1 if the subject suffering from hypertension, otherwise, y=0) as a function of the risk factors. By computing the conditional probability $p(y=1 \mid X)$, where $X = (x_1, x_2 \ldots x_n)$ represents n risk factors associated with the disease, we could calculate the likelihood that an individual suffers from the disease. The logistic regression model takes the following form:

$$\log\left[\frac{p(X)}{1-p(X)}\right] = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \ldots + \beta_n * x_n. \quad (1)$$

, where $X = (x_1, x_2 \ldots x_n)$ stands for the vectors of n risk factors selected by logistic regression model, β_i is the coefficients of corresponding x_i and represents the statistical significance level. By transforming (1), we rewrite the prediction model expressed as (2):

$$p(y \mid X) = 1 / (1 + \exp(-(\beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \ldots + \beta_n * X_n))). \quad (2)$$

By setting 0.5 as the cutoff value, if $p(y=1 \mid X) > 0.5$, we infer that the individual suffers from the disease; otherwise, he/she is free from the disease. Besides this, logistic regression model is endowed with the capacity to select factors that are significant to the disease y on the basis of its statistical significance p-value.

In our work, logistic regression analysis was applied to select risk factors that were significant to hypertension. Further, the selected risk factors were directed to neural networks model as inputs.

C. Multilayer Neural Network

Artificial Neural Networks (ANNs) are computational model inspired by animals' central nervous system. ANNs are data-driven self-adaptive methods since they could adjust themselves to the data without positing any explicit specification of distribution form for the underlying model, which differs from traditional statistical procedures that are established on Bayesian decision theory. Moreover, as a nonlinear mapping model, ANNs are flexible and effective in modeling and reflecting the complex relationships between inputs and outputs in the real word, and the effectiveness and flexibility of neural networks for classification and prediction problems has been tested empirically in a wide variety of classification tasks such as handwriting recognition [16], speech recognition, medical diagnosis [17,18]. The typical processing procedure of an artificial neural network is: a set of input neurons are activated by inputs, then the activations of these neurons are passed on, weighted and transformed by functions given by the network to other neurons, until finally the output neurons are activated and generate results.

Multi-Layer Perceptron (MLP) neural networks are one of the classical and commonly used static neural networks and widely used for classification problems [19]. MLP are feed-forward neural networks trained with the Back Propagation (BP) algorithm, and utilizes supervised learning techniques to transform sets of input data into a desired response. As a modification of the standard linear perceptron, MLP can distinguish data that are not linearly separable. More recently, there has been renewed interest in back propagation networks due to the success of deep learning. In our research, we plan to employ MLP to explore the relationship between hypertension and the selected risk factors and further develop a model for hypertension prediction.

As the core component of MLP, BP training with generalized delta learning rule is an iterative gradient algorithm with the aim to obtain a classification model with high prediction accuracy by minimizing the root mean square error between the actual output of the model and desired output. In general, BP learning algorithm can be divided into two phases: propagation and weight update. The BP algorithm is depicted as shown in algorithm 1.

Algorithm 1: Back Propagation (BP) algorithm

Input: N train samples, with inputs $x(1), x(2), \ldots, x(N)$; corresponding desired output $y(1), y(2), \ldots, y(N)$, where $x(i) = (x_1(i), x_2(i), \ldots, x_k(i))$ is a vector with k features, $1 \le i \le N$

Output: *NN*: a neural network

1: Initializing network weights and biases to small random values.

2: Inputting a study sample $(x(p), y(p))$ $(1 \le p \le N)$.

3: Calculating the actual output of nodes in hidden layer:

$$Y_j^2 = f(\sum_{i=1}^{n_1} W_{ij} * Y_i^1 - b_j) = f(\sum_{i=1}^{n_1} W_{ij} * X_{ip} - b_j), j \in \{1, 2, \ldots, n_2\}. \quad (3)$$

4: Calculating the actual output of nodes in output layer:

$$o_k = f(\sum_{i=1}^{n_2} W_{jk} * Y_j^2 - b_k), k \in \{1, 2, \ldots, m\}. \quad (4)$$

5: Adapting weights W_{ij} and biases b_i using (5) and (6):

$$\Delta w_{i,j}^{(l)} = \mu * x_j * \delta_i^{(l)}. \quad (5)$$

$$\Delta b_i^{(l)} = \mu * \delta_i^{(l)}. \quad (6)$$

, where μ is learning rate, $x_j(n)$ is output of node j at the iteration n.

$$\delta_i^{(l)}(n) = \begin{cases} \varphi'(net_i^{(l)}) * [y_i - o_i], & l=M \\ \varphi'(net_i^{(l)}) * \sum_k w_{ki} * \delta_k^{(l)}, & 1 \le l < M \end{cases}. \quad (7)$$

, in which l is the layer, M is output layer, k is the number of output nodes of *NN*.

6: If left study sample, goto step 2

7: Calculating error function E, if E satisfying, stop; else, goto step 2

D. Evaluation Measures

A confusion matrix, also known as contingency table, contains the desired/actual class and predicted class of a classification model [20], and is applicable to evaluating the performance of a supervised learning algorithm such as MLP with BP algorithm, Support Vector Machine (SVM). Table II presents the confusion matrix case for the prediction of hypertension.

To evaluate the performance of the constructed hypertension prediction model, in this study we used the following measures:

1) Accuracy represents the total accuracy rate of classifying each case correctly.

$$Accuracy = (TP+TN)/(TP+FP+TN+FN). \quad (8)$$

2) Sensitivity stands for the probability of correctly classifying an individual suffering from hypertension.

$$Sensitivity = TP/(TP+FN). \quad (9)$$

3) Specificity represents the probability of correctly determining that an individual is a non-hypertension.

$$Specificity = TN/(FP+TN). \quad (10)$$

4) The Area Under the ROC Curve (AUC) value presents the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [21]. An area of 1 represents a perfect classification, while an area of 0.5 represents a worthless model. The AUC is equivalent to the Mann-Whitney-Wilcoxon sum of ranks statistic and is estimated as follows [22]:

$$AUC = \frac{s - (pos \times (pos + 1) / 2)}{pos \times neg}, \quad (11)$$

, in which s is the sum of ranks of true hypertension cases, pos denotes the number of hypertension cases, and neg denotes the number of non-hypertension cases.

TABLE II. A CONFUSION MATRIX FOR HYPERTENSION PREDICTION

	Real situation	
Class predicted	***Hypertension***	***Non-hypertension***
Hypertension	TP	FP
Non-hypertension	FN	TN

III. EXPERIMENTAL DESIGN AND RESULTS

In this section, we illustrated the experiment design procedure in detail and then presented corresponding experimental results. To select hypertension-associated risk factors and establish a model for the prediction of hypertension with the selected risk factors, we first utilized binary logistic regression analysis on the experimental dataset to select factors that are significant to hypertension according to the statistically significant p-value; then we constructed and trained an ANN model with these risk factors. Through detailed illustration and systematic parameter selection for neural networks, we got the hypertension prediction model, followed by experimental results and analysis.

A. Significant Risk Factors for Hypertension

Logistic regression analysis has the capacity not only function as a prediction model, but also to select significant factors as inputs to the construction of other kind of prediction model and provide disease surveillance researchers an approach to figure out factors that are significant to a certain disease.

In the selection of significant risk factors, multi-factor logistic regression model with partial maximum likelihood estimation and forward-step regression analysis was applied on the experimental dataset. Consequently, 11 hypertension-relevant risk factors (exercise, diabetes, hyperlipemia, age, marriage, education, income, weight, height, sex, smoke, drink) were selected as significant ones and two factors (smoke100, smoke) were filtered out by setting statistical significance p-value less than 0.05 as variable inclusion criteria and p-value greater than 0.1 as variable exclusion criteria (Table III). After investigating the distribution of each variable of the dataset, we found that variable "smoke100" just took one value, thus it was not involved in the calculation of logistic regression analysis. Odd Ratio represented the 95% confidence interval in statistics. All the odd ratio values were located in confidence interval, which proved the effectiveness of the results.

TABLE III. MULTI-FACTOR LOGISTIC REGRESSION ANALYSIS FOR HYPERTENSION

Variable	***p*-value**	**Odd Ratio (95% CI)**
Exercise	<0.001	0.878(0.861~0.895)
Diabetes	<0.001	1.420(1.401-1.439)
Hyperlipemia	<0.001	2.112(2.077~2.148)
Age	<0.001	0.955(0.955~0.956)
Marriage	<0.001	0.987(0.981~0.993)
Education	<0.001	1.047(1.038~1.056)
Income	<0.001	1.079(1.073~1.084)
Weight	<0.001	0.989(0.988~0.989)
Height	<0.001	1.003(1.003~1.003)
Sex	<0.001	0.944(0.925~0.963)
Smoke	0.270	1.006(0.996~1.016)
Drink	<0.001	0.967(0.950~0.984)

B. Neural Network Prediction Model

From its graphical representation, ANNs consists of one input layer, zero or more hidden layers and one output layer, and a collection of neurons with connectivity between two or more network layers. In general, the architecture of ANNs is determined by the number of inputs n and outputs m, the number of hidden layers and neurons in each hidden layer. In this study, binary logistic regression selected 11 risk factors

with significant *p*-value; hence, there were 11 inputs in our constructed neural network. Since our aim is to predict whether an individual suffers from hypertension or not, we set 2 outputs.

On the basis of Kolmogorov theorem, theoretical analysis proves that feed-forward neural networks with single hidden layer have the capacity to approximately denote any continuous function and achieve arbitrary nonlinear mapping [23,24]. Considering that the training time increases with the number of hidden layers increasing, for achieving the tradeoff between speed and accuracy, artificial neural networks with single hidden layer were adopted in our research. As for the choice of the number of neurons *h* in the hidden layer, three rule-of-thumbs were used in the choice of the interval of its possible values rather than in a grid-based or exhaustive way to search for the best-fitting value of *h*.

1) Boger and Guterman pointed out that the number of neurons in hidden layer should be more than two thirds of the number of inputs [25], and it was expressed as the following form in (12):

$$h \geq \frac{2}{3} * n. \tag{12}$$

2) Berry and Linoff suggested that the number of neurons in hidden layer should be less than twice the number of inputs for circumventing high amount of computation during training [26], written as:

$$h \leq 2 * n. \tag{13}$$

3) Blum suggested that the number of neurons in hidden layer should be limited between the number of inputs and outputs [27], presented as:

$$m \leq h \leq n. \tag{14}$$

By taking into consideration constraint conditions (12), (13) and (14) as a whole, we derived that, in the case of our study, the possible value of the number of neurons *h* was constrained between 8 and 11.

In the choice of activation function of hidden layer and output layer, Karlik and Olgac compared five conventional activation functions, including Bi-polar sigmoid, Uni-polar sigmoid, Hyperbolic Tangent (Tanh), Conic Section, and Radial Bases Function (RBF) to evaluate the performance of MLP neural network architecture; and they concluded that activation function Tanh outperformed other activation functions in the vast majority of MLP classification and prediction applications [28,29]. Directed by their experimental results, we chose Tanh as the activation function in both the hidden layer and output layer.

To achieve faster convergence with minimum oscillation, BP algorithm with learning rate μ and momentum *mc* was adopted as an improvement to the basic BP algorithm. Empirical values were assigned to the two parameters. According to the discussion above, parameters of the neural network prediction model and their values were summarized in Table IV.

TABLE IV. A SUMMARY OF THE PARAMETERS OF ANNS

Parameter	Symbol	Value
Number of inputs	*n*	11
Number of outputs	*m*	2
Number of neurons in hidden layer	*h*	[8,9,10,11]
Activation function of hidden layer	*hid_func*	Tanh
Activation function of output layer	*out_func*	Tanh
Learning rate	μ	0.4
Momentum	*mc*	0.9

On the basis of the detailed discussion of the architecture setting and parameter selection of the neural network model, the constructed hypertension prediction model with one hidden layer and two outputs was presented in Fig. 1. In the input layer, there were 11 variables obtained from the binary logistic regression analysis; the number of neurons the hidden layers ranged from 8 to 11. The outputs with 2 neurons predicted whether an individual suffered from hypertension or not.

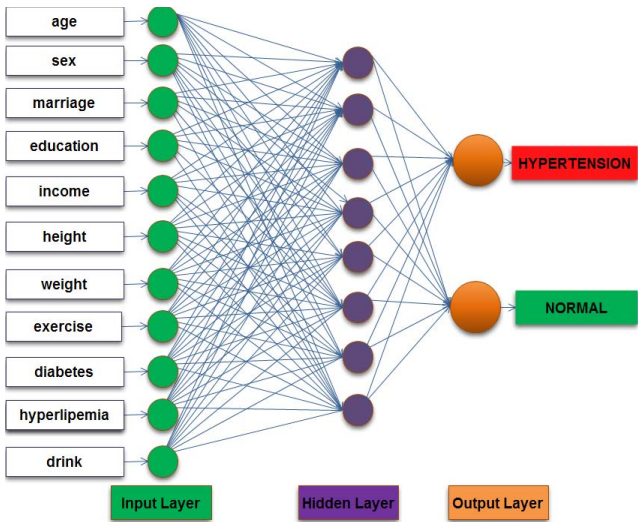

Figure 1. Hypertension prediction model

C. Experimental Results and Analysis

To evaluate the performance of the constructed hypertension prediction model, we randomly partitioned the experimental dataset into train set and test set in the ratio of 7:3. Train set was used for the training of the model to select

the optimized parameters and test set was used for the evaluation of the model. Specially, initial parameters of the model were listed in Table IV and training period varied from 100,000 to 200,000 iterations. We ran each experiment 10 times and presented the averaged results and the standard deviations (SD) in Table V with different number of neurons in the hidden layer.

From Table V, we can see that the average accuracy ranges from 71.91% to 72.12% with good stability and that the average Area Under operator-receiver Curve (AUC) is 0.77 when h varied from 8 to 11. And the best prediction accuracy was found to be close to 72%. Senior physicians suggest that 30% is an acceptable error rate in the diagnosis of hypertension in practice [8]; therefore, prediction accuracy above 70% is acceptable and useful, and this indicates the effectiveness of our approach.

TABLE V. PREDICTION RESULTS WITH DIFFERENT NEURONS IN HIDDEN LAYER

Neurons in Hidden Layer	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**	AUC
$h = 8$	49.20 ± 3.53	84.37 ± 1.73	72.04 ± 0.22	0.77 ± 0.002
$h = 9$	48.69 ± 1.59	84.69 ± 0.80	72.06 ± 0.07	0.77 ± 0.002
$h = 10$	46.85 ± 5.59	85.42 ± 2.25	71.91 ± 0.43	0.77 ± 0.003
$h = 11$	48.91 ± 1.22	84.62 ± 0.68	72.12 ± 0.04	0.77 ± 0.001

As a comparative study, we also conducted experiments by constructing a logistic regression-based prediction model with the selected risk factors being independent variables. Similar to that of neural network-based model, we randomly partitioned the experimental dataset into train set and test set in the ratio of 7:3. Training set was used to get the coefficients of each variable in logistic regression equation represented as (2), and the test set was to evaluate the performance. By setting 0.5 as the cutoff value, we inferred that the individual suffered from hypertension if predicted value was greater than 0.5, otherwise, the individual free from hypertension. After repeating the procedure 10 times, we presented the experimental results and standard deviations (SD) shown in Table VI.

TABLE VI. PREDICTION RESULTS FROM LOGISTIC REGRESSION

	Sensitivity (%)	**Specificity (%)**	**Accuracy (%)**	**AUC**
Logistic regression	44.68 ± 5.17	86.42 ± 2.66	71. 96 ± 0.21	0.74 ± 0.001

Through the comparison of the experimental results in Table V and Table VI, it was observed that logistic regression-based prediction model achieved 71.96% prediction accuracy and an AUC of 0.74. Though its performance were very close to that of neural network-based prediction model, it showed poor stability in prediction with much bigger standard deviations in sensitivity and specificity, while neural network-based model achieved good performance and comparatively small standard deviations when h was equal to 11. This indicated that neural networks were more powerful in adjusting themselves to new environments and more suitable to be employed in our study.

IV. CONCLUSION

With the continuing maturity of the Internet of Things (IoT) technology and its wide applications in various fields, constructing a prediction model for chronic diseases is an exploration of IoT technology on the data-analysis and decision-making level as well as a meaningful research in health care practice. Following this, in this paper, we proposed a logistic regression and artificial neural network-based approach for chronic disease prediction. As a case study of hypertension, binary logistic regression analysis was first applied on experimental dataset to select significant hypertension risk factors by setting p-value. These selected variables were not only the inputs of neural networks, but also risk indicators to warn individuals to avoid or strengthen some certain factors. Then, on the basis of detailed discussion of the design of artificial neural network architecture and the choice of parameters, we constructed a Multi-Layer Perception (MLP) neural network with the selected risk factors as inputs. Finally, we conducted experiments using both artificial neural network model and the logistic regression model as a comparison. Experimental results showed that integration of logistic regression and neural networks was an effective tool in the prediction of hypertension.

In the future work, we plan to explore deeper in the prediction of hypertension by providing logistic regression analysis model more potential risk factors with the aim to find more novel and typical risk factors about hypertension and improve prediction accuracy, and further apply the approach in the prediction of other chronic diseases such as diabetes, asthma as a validation.

ACKNOWLEDGMENT

This work was supported in part by the major project of natural science foundation for Anhui Province higher education under award KJ2011ZD06, the “111 Project” of Ministry of Education and State Administration of Foreign Experts Affairs under Grant No. B14025, the “University Featured Project” of Ministry of Education under Grant No. TS2013HFGY031, the Chinese National Key Technology R&D Program under Grant No. 2013BAH19F01, and the Natural Science Foundation of China under award 61305064. The authors are very grateful to the anonymous reviewers for their constructive comments and suggestions to the improvement of this research.

REFERENCES

[1] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey.*Computer networks*," 2010, *54*(15), pp.2787-2805.

[2] M. Ture, I. Kurt, A. Turhan Kurum, and K. Ozdamar, "Comparing classification techniques for predicting essential hypertension," *Expert Systems with Applications*, 2005, *29*(3).

[3] N. Wong, G. Thakral, S. Franklin, G. L'Italien, M. Jacobs, J. Whyte, and P. Lapuerta, "Prevention and rehabilitation: preventing heart disease by controlling hypertension: impact of hypertensive subtype, stage, age, and sex," *American Heart Journal*, 2003, *145*(5), pp.888-895.

[4] K. Hsu, C. Chiu, N. Chiu, P. Lee, W. Chiu, T. Liu, and C. Hwang, "A case-based classifier for hypertension detection.*Knowledge-Based Systems*," 2011, *24*(1), pp.33-39.

[5] R. Vasan, M. Larson, E. Leip, J. Evans, C. O'Donnell, W. Kannel, and D. Levy, "Impact of high-normal blood pressure on the risk of cardiovascular disease," *New England Journal of Medicine*, 2001, *345*(18), pp.1291-1297.

[6] J. Jeppesen, H. Hein, P. Suadicani, and F. Gyntelberg, "High triglycerides and low HDL cholesterol and blood pressure and risk of ischemic heart disease," *Hypertension*, 2000, *36*(2), pp.226-232.

[7] A. Blinowska, G. Chattellier, A. Wojtasik, and J. Bernier, "Diagnostica-a Bayesian decision-aid system-applied to hypertension diagnosis," *Biomedical Engineering, IEEE Transactions on*, 1993, *40*(3), pp.230-236.

[8] A. Blinowska, G. Chatellier, J. Bernier, and M. Lavril, "Bayesian statistics as applied to hypertension diagnosis," *Biomedical Engineering, IEEE Transactions on*, 1991, *38*(7), pp.699-706.

[9] M. Caulfield, P. Munroe, J. Pembroke, N. Samani, A. Dominiczak, M. Brown, and J. Connell, "Genome-wide mapping of human loci for essential hypertension," *The Lancet*, 2003, *361*(9375), pp.2118-2123.

[10] R. Vasan, M. Larson, E. Leip, J. Evans, C. O'Donnell, W. Kannel, and D. Levy, "Impact of high-normal blood pressure on the risk of cardiovascular disease," *New England Journal of Medicine*, 2001, *345*(18), pp.1291-1297.

[11] C. Chang, C. Wang, and B. Jiang, "Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors," *Expert systems with applications*,2011, *38*(5), pp.5507-5513.

[12] B. Sumathi, D. Santhakumaran, " Pre-diagnosis of hypertension using artificial neural network," *Global Journal of Computer Science and Technology*, 2011, *11*(2).

[13] B. Krawczyk, and M. Wozniak, "Hypertension diagnosis using compound pattern recognition methods," *Journal of Medical Informatics & Technologies*, 2011,*18*, pp.41-50.

[14] M. Wozniak, "Two-stage classifier for diagnosis of hypertension type," In *Biological and Medical Data Analysis* ,2006, pp.433-440. Springer Berlin Heidelberg.

[15] Behavior Risk Factor Surveillance System. http://www.cdc.gov/brfss/. Center of Disease Control and Prevention.

[16] S. Knerr, L. Personnaz, and G. Dreyfus, "Handwritten digit recognition by neural networks with single-layer training," *Neural Networks, IEEE Transactions on*, 1992, *3*(6), pp.962-968.

[17] Q. Al-Shayea, "Artificial Neural Networks in Medical Diagnosis," 2013, *International Journal of Computer Science Issues (IJCSI)*, *8*(2).

[18] K. Chan, S. Ling, T. Dillon, and H. Nguyen, "Diagnosis of hypoglycemic episodes using a neural network based rule discovery system," *Expert Systems with Applications*, 2011, *38*(8), pp.9799-9808.

[19] A. Vellido, P. Lisboa, and J. Vaughan, "Neural networks in business: a survey of applications (1992–1998)," *Expert Systems with Applications*, 1999, *17*(1), pp.51-70.

[20] F. Provost, and R. Kohavi, "Guest editors' introduction: On applied research in machine learning," *Machine learning*, 1998, *30*(2), pp.127-132.

[21] A. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern recognition*, 1997, *30*(7), pp.1145-1159.

[22] J. Vila-Francés, J. Sanchís, E. Soria-Olivas, A. Serrano, M. Martínez-Sober, C. Bonanad, and S. Ventura, "Expert system for predicting unstable angina based on Bayesian networks," *Expert Systems with Applications*, 2013, *40*(12), pp.5004-5010.

[23] A. Kolmogorov, "The representation of continuous functions of many variables by superposition of continuous functions of one variable and addition," *Doklady Akademii Nauk SSSR*, 1957, *114*(5), pp.953-956.

[24] T. Chen, H. Chen, and R. Liu, "Approximation capability in C (R n) by multilayer feedforward networks and related problems," *Neural Networks, IEEE Transactions on*, 1995, *6*(1), pp.25-30.

[25] Z. Boger, and H. Guterman, "Knowledge extraction from artificial neural network models. In *Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference on* (Vol. 4, pp. 3030-3035). IEEE.

[26] M. Berry, and G. Linoff, "Data mining techniques: for marketing, sales, and customer support," John Wiley & Sons, Inc. 1997.

[27] A. Blum, "Neural networks in C++: an object-oriented framework for building connectionist systems," John Wiley & Sons, Inc. 1992.

[28] B. Karlik, and V. Olgac, "Performance analysis of various activation functions in generalized MLP architectures of neural networks. *International Journal of Artificial Intelligence and Expert Systems*, 2010, *1*(4), pp.111-122.

[29] T. Tan, J. Teo, and P. Anthony, "A comparative investigation of non-linear activation functions in neural controllers for search-based game AI engineering," *Artificial Intelligence Review*, 2011, 1-25.

[30] D. Hosmer, and S. Lemeshow, "*Applied logistic regression*," 2004, John Wiley & Sons.