Expert Systems with Applications 42 (2015) 7601–7609

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Predicting hypertension without measurement: A non-invasive, questionnaire-based approach

Aiguo Wang[a], Ning An[a,*], Guilin Chen[b], Lian Li[a], Gil Alterovitz[c,d,e]

[a] *Hefei University of Technology, School of Computer and Information, Hefei, China*
[b] *Chuzhu University, School of Computer and Information Engineering, Chuzhou, China*
[c] *Harvard University, Harvard Medical School, Center for Biomedical Informatics, Boston, USA*
[d] *Children's Hospital Informatics Program, Boston Children's Hospital, Harvard Medical School, Boston, USA*
[e] *Massachussetts Institute of Technology, Department of Electrical Engineering and Computer Science, Cambridge, USA*

ARTICLE INFO

Article history:
Available online 11 June 2015

Keywords:
Hypertension prediction
Non-invasive
Questionnaire
Behavior Risk Factor Surveillance System
Artificial neural networks

ABSTRACT

Early detection of hypertension contributes to the prevention and reduction of the onset of cardiovascular diseases. Since lifestyle choices are linked to the occurrence and development of hypertension, determining hypertension risk factors and further establishing a predictive model with these factors will facilitate the early prevention and effective management of hypertension and improve individual health conditions. This study attempts to construct a prediction model based on the hybrid use of logistic regression and artificial neural networks (ANNs) for hypertension detection in a non-invasive, questionnaire-based way. First, the binary logistic regression model was used to select risk factors significant to hypertension. Second, after detailing the selection of ANNs architecture and the setting of relevant parameters, we constructed a multi-layer perception neural network model with back propagation learning algorithms to predict hypertension. Then, to mitigate the biased prediction results caused by a potentially unbalanced training set, we proposed an effective under-sampling technique and adopted it to balance the dataset prior to the training of the predictive model. To evaluate the performance of the proposed approach, we conducted extensive experiments on the questionnaires collected from Behavior Risk Factor Surveillance System. Experimental results show that ANN-based prediction model obtains over 72.0% accuracy and an area under the receiver-operator curve of 0.77 and achieves good stability in comparison with the logistic regression-based model. Further, the proposed approach obtains balanced prediction performance with the under-sampling technique. The results demonstrate the practicability of hypertension prediction with simple demographic data rather than with clinical tests and genomic data and of developing a hypertension surveillance system for a large scale of population in a non-invasive and economical way. Also, we actually provide a general framework for the simultaneous identification of risk factors and prediction of other chronic diseases.

1. Introduction

Hypertension is a long-lasting chronic health condition and affects a wide range of the population, particularly for adults over the age of fifty-five. Even worse, it is becoming prevalent among adolescents in both developing and developed countries (Ture, Kurt, Turhan Kurum, & Ozdamar, 2005; Midha et al., 2014). Hypertension is also a major risk factor for the occurrence and development of many cardiovascular diseases, such as stroke, heart failure and chronic kidney disease, and the poor management and treatment of hypertension leads to the increase in morbidity and mortality rates (Hsu et al., 2011; Jeppesen, Hein, Suadicani, & Gyntelberg, 2000; Vasan et al., 2001; Wong et al., 2003). Besides the fact that prevention and management of hypertension consumes a wealth of medical resources and healthcare services, it deteriorates the imbalanced distribution of medical resources and definitely puts on the society a considerable financial burden.

The main difficulty associated with hypertension prevention and management is the lack of clear clinical effects in the early stage of hypertension. As a consequence, individuals may easily disregard the occurrence of hypertension and develop potential serious complications (Vasan et al., 2001). Though hypertension

[*] Corresponding author at: Hefei Tunxi Road 193, Hefei University of Technology, Hefei 230009, China. Tel.: +86 180 1995 6086; fax: +86 551 6290 4642.

E-mail addresses: wangaiguo2546@163.com (A. Wang), ning.g.an@acm.org (N. An), glchen@chzu.edu.cn (G. Chen), llian@hfut.edu.cn (L. Li), gil_alterovitz@hms.harvard.edu (G. Alterovitz).

http://dx.doi.org/10.1016/j.eswa.2015.06.012

is among the most common and costly health problems, it is also among the most preventable and can be effectively controlled through reasonable measures due to the fact that lifestyle choices are linked to the occurrence and development of hypertension (Chang, Wang, & Jiang, 2011; Hsu et al., 2011; Krawczyk & Wozniak, 2011; Sumathi & Santhakumaran, 2011; Wozniak, 2006). Therefore, investigating risk factors and identifying hypertension plays a crucial role in the effective prevention and reduction of the onset of cardiovascular diseases as well as better management and intervention of individual health conditions (Hsu et al., 2011). On the other hand, the investigation of hypertension risk factors is a crucial issue for preventive medicine and particularly drawing interests from public health researchers with the aim to bring down the onset of hypertension through early warning and prevention. In comparison with clinical test data, genomic data and anthropometric body surface scanning data (Chiu et al., 2007), lifestyle behavior information provides an alternative way for hypertension prediction, and they are easily collected and more meaningful in the prevention and management of hypertension. Furthermore, lifestyle risk factors could be indicators to remind or warn individuals to avoid or circumvent unhealthy behaviors in order to effectively prevent and better manage hypertension, and the prediction model can be used for locating those individuals who may be at high risk of hypertension and for the large-scale hypertension surveillance without the complex and expensive measurements.

Traditional approaches usually statistical techniques to determine the relationship between hypertension and the risk factors. Among these, artificial neural networks (ANNs) are data-driven methods and have the ability to adjust themselves to the data without positing any explicit specification of distribution form for the underlying model. This differs from traditional statistical procedures that are established on Bayesian statistical theory. As a nonlinear mapping model, ANNs are flexible and effective in modeling complex relationships between inputs and outputs and widely used for the medical diagnostics (Ziada et al., 2001). However, one of the main difficulties in constructing neural networks is the model selection problem. More precisely, one needs to select a suitable ANNs architecture and set its corresponding parameters due to the fact that ANNs are quite sensitive to these factors and inappropriate model selection can degrade their generalization ability.

With the aim to enable early identification of hypertension and risk factors and develop a practical screening tool, in this study, we proposed a questionnaire-based hypertension prediction approach that integrated logistic regression analysis and artificial neural networks with the aim of determining risk factors and predicting hypertension. After collecting and cleaning a publicly available dataset from Behavior Risk Factor Surveillance System (BRFSS) of Centers of Control and Prevention (CDC), we first applied the binary logistic regression model to select risk factors significantly relevant to hypertension and constructed the logistic regression-based prediction model. Then, we trained a multi-layer perception (MLP) neural network with back propagation algorithms using the selected risk factors as inputs to predict whether an individual suffers from hypertension. In the construction and training of ANNs, we detailed the selection of ANNs architecture and proposed to employ three rule-of-thumbs to narrow down the search space of feasible solutions towards a tradeoff between efficiency and accuracy. Additionally, considering that class imbalance problems are common in medical datasets and that the skewed class distribution makes many classification methods less effective and jeopardizes the accuracy of the minority class (Wang & Yao, 2012), we proposed an effective under-sampling technique to adjust the size of training sets prior to the training of ANNs.

The remainder of this paper is organized as follows. Section 2 reviews previous related research work and techniques. Experimental dataset, logistic regression analysis and artificial neural network models are illustrated in Section 3. In the experimental design and result analysis section, we detail the selection of neural network architectures and the setting of corresponding parameters, and describe an experiment to demonstrate their effectiveness for hypertension prediction in comparison with that of logistic regression based prediction model. The last section concludes our work with a brief summary and presents possible directions for the future studies.

2. Related work

A large number of researchers and medical experts have conducted considerable work in investigating hypertension risk factors and indicators and constructing effective prediction models with these factors. There are a variety of factors that can be used to predict hypertension, mainly including demographics, anthropometry body surface scanning data, clinical test data and even molecular-level data (e.g. genomic and proteomic data). To figure out the risk factors, Lee and Entzminger (2006) conducted a cross-sectional study in a Thai population of 1398 patients (382 men and 1016 women), and performed multiple linear regression to determine the relevance of several risk factors for hypertension. They found that old age, body mass index and low education attainment are significant risk factors. Akdag et al. (2006) applied the classification tree method to determine risk factors for hypertension among 1761 adults at the outpatient clinic in western Turkey between January 2002 and July 2004. They studied the effects of fourteen risk factors on hypertension, and their results revealed that body mass index, waist-to-hip ratio, sex, serum triglycerides, serum total cholesterol, hypertension in first-degree relatives, and saturated fat consumption are main risk factors. Accordingly, various machine learning and statistical analysis techniques with different metrics are utilized to find a mapping function between the factors and hypertension. Ture et al. (2005) compared a comparative study to evaluate the performance of nine commonly used classification methods for hypertension prediction among 694 subsets (452 hypertension patients and 242 controls). Their experimental dataset consisted of demographics, lifestyle information and clinical test results. Experimental results revealed that multi-layer perception (MLP) neural network and Radial Basis Function (RBF) neural network outperformed the other three decision tree and four statistical algorithms. Blinowska, Chatellier, Bernier, and Lavril (1991) proposed to apply Bayesian statistical methods that incorporated both prior knowledge and possible costs of wrong decisions for hypertension prediction using demographics and clinical test data, and the proposed method achieved satisfactory accuracy. However, since Bayesian method is built on statistical theory, difficulties in collecting a sufficient number of experimental cases and ensuring the integrity of each case hinder its wide applications in actual use (Blinowska, Chattellier, Wojtasik, & Bernier, 1993; Blinowska et al., 1991). Chang et al. (2011) proposed to use several data mining classifier techniques to determine the risk factors of hypertension in a vote-based scheme, and then build a predictive model using multivariate adaptive regression splines. Besides using clinical test data, researchers also explore the possibility of hypertension prediction using other types of data. For example, Hsu et al. (2011) focused on determining the relationship between hypertension and three-dimensional anthropometric scanning data (e.g. the circumferences of waist, wrist and gluteal), and they proposed to hybridize case-based reasoning and genetic algorithms for hypertension detection. Experimental results revealed the relationship between anthropometric data and hypertension and demonstrated the effectiveness of case matching techniques. In addition, to investigate the mechanism of hypertension at the molecular level, Caulfield et al. (2003) conducted a study to identify genetic factors associated with hypertension. Kesselmeier et al. (2014) used data from the Genetic Analysis Workshop 18 to evaluate the performance of the standard logistic regression methods, and found their strong dependence on a few observations that deviated from the majority of the data. Huang, Xu, and Yang (2014) developed a two-stage hypertension prediction approach using the genotype information. They first detected significant single-nucleotide polymorphisms (SNPs) and then developed a permanental classifier for prediction purposes. Sanada et al. (2015) conducted a study using statistical tests in a Japanese population with 588 hypertensive individuals and 486 normotensive controls, and their study showed that non-synonymous GRK4 variants are associated with essential hypertension. Their work provides us novel insights into the study of pathogenesis mechanisms, the prediction of hypertension, as well as the discovery of potential therapeutic targets.

Though great progresses have been made in hypertension prediction, however, there exist several difficulties and limitations of current methods to be widely applied for hypertension prevention and management in actual use. First, predicting hypertension with clinical test data, anthropometric body surface scanning data and genomic data obtains high prediction performance, but it is not suitable and practical for hypertension prediction in a large population due to the fact that it involves complex operation processes and costs much and that few individuals are willing to take clinical tests. Second, clinical test data and genomic data are good indicators for hypertension prediction, but they present less information about hypertension risk factors that can be of great value for the early prevention and better management of hypertension. On the contrary, lifestyle behaviors are easily collected and provide meaningful insights into hypertension prevention and management. Third, due to lack of clear clinical effects in the early stage of hypertension, individuals may easily disregard the occurrence of hypertension and develop potential serious complications (Vasan et al., 2001). Fourth, predicting hypertension with risk factors is a challenging task, since inappropriate model selection and unbalanced class distribution may deteriorate the accuracy. Studies showed that ANNs can obtain satisfactory prediction performance and outperform many other classification methods, while few presented detailed discussions of the selection of ANNs architecture and the setting of corresponding parameters towards a tradeoff between efficiency and accuracy. Besides the model selection issue, class imbalance problem often emerges in medical datasets and usually leads to a learning bias of the constructed classifier to the majority class. Unfortunately, hypertension prediction is such a case because the number of hypertension cases (called minority class or positive class) is usually much smaller than the number of controls (called majority class or negative class) (Wang & Yao, 2012). For example, suppose there is a data set and the ratio between the number of majority class samples and the number of minority class samples is 100:1. An accuracy-driven classifier that aims to maximize the final classification accuracy may obtain an accuracy of 99% by ignoring the minority class samples and predicting all instances as majority class. Obviously, it is not acceptable in actual use. Based on the discussions above, to enable early-stage effective prevention and later-stage better management of hypertension in an efficient and economical way, we propose to predict hypertension with the collected questionnaires and further explore the model selection problem.

3. Materials and methods

3.1. Dataset and hypertension risk factors

The dataset used in our study was collected from the Behavior Risk Factor Surveillance System (BRFSS) of Centers for Disease Control and Prevention (CDC) and is publicly available and downloadable from the BRFSS website. BRFSS is the world's largest and continuously conducted telephone-based health survey regarding behavioral risk factors, chronic health conditions and use of preventive services. Established in 1984 with 15 states participating in the survey, it has a long history in behavioral and chronic disease surveillance. The primary aim of BRFSS is to track and measure individual health conditions and risk behaviors that contribute to the leading cause of high morbidity and mortality rates in the adult population who are aged 18 years and the elderly in United States. The survey covers a wide range of health risk factors, preventive health practices and health conditions, including hypertension, diabetes and carcinoma related items. By collecting a variety of information and sharing them to the public, BRFSS enables researchers to investigate the relationships between chronic diseases and their risk factors (Mokdad et al., 2003; Oswald & Hu, 2010). Additionally, U.S. government can rely on BRFSS data to compare states to allocate funding and focus interventions, and the states can use the survey results to focus interventions for the public and make better policies. A working group of BRFSS coordinators and CDC staff is in charge of the design of BRFSS survey. Currently, BRFSS questionnaires have three parts: the core components, optional modules and state-added questions. The core component consists of a core set of questions on certain topics like hypertension, exercise or tobacco use and must be asked without modification, while the modules are optimal and state-added questions are designed by each state and may differ among states. The BRFSS system records the reply of every investigated individual to each question. In our study, the diagnosis of hypertension is made when an individual answers yes to the question that "have you ever been told by a doctor, nurse or other health professional that you have high blood pressure?". Similarly, researchers can check other survey items. BRFSS official website provides relevant questionnaires and coding forms to help researchers better understand and exploit the data.

According to the survey items in BRFSS and the potential hypertension risk factors used by many researchers as discussed in the introduction section, we picked out 13 survey items as the candidate risk factors. Table 1 presents the variable names coded in BRFSS, their corresponding meanings, and their coding names used in our study. Independent variables include age, sex, height, weight, marriage, income, Hyperlipemia, diabetes, exercise, education, smoke100, smoke and drink. Particularly, *Marriage* refers to the one's marriage status and there are six possible values for choosing; *Education* is defined as the highest grade or year of school one completed, and six choices are provided; *Income* means one's annual household income level with eight options provided; *Smoke* represents the smoke frequency (every day, some days or not at all) and the remaining risk factors are Boolean variables except *age*, *height* and *weight* with real numbers. After excluding cases with missing values, we collected 308,711 samples from

Table 1
Description of variables of experimental data.

No.	Variable	Variable description	Coding in our work
1	AGE	'Age'	Age
2	SEX	'Sex'	Sex
3	HEIGHT	'Height in inches'	Height
4	WEIGHT	'Weight in pounds'	Weight
5	MARITAL	'Marriage status'	Marriage
6	EDUC	'Education level'	Education
7	INCOME	'Income level'	Income
8	EXERANY	'Exercises during the past month'	Exercise
9	DIABETES	'Ever told having diabetes'	Diabetes
10	TOLDHI	'Ever told blood cholesterol high'	Hyperlipemia
11	SMOKE100	'Smoke more than 100 in total'	Smoke100
12	SMOKEDAY	'Smoke frequency now'	Smoke
13	ALCDAY	'Drink frequency'	Drink
14	BPHIGH	'Ever told blood pressure high'	Hypertension

Table 2
Main characteristics of experimental data.

Variable	Hypertension (n = 108,260)	Control (n = 200,511)
Age (years)	59.8 ± 13.4	49.3 ± 14.8
Sex (female/male)	1.0	1.3
Height (inches)	520.4 ± 37.9	520.4 ± 37.0
Weight (pounds)	186.3 ± 44.7	169.7 ± 39.3
Exercise (%)	70.0	80.0
Diabetes (%)	97.7	85.9
Hyperlipemia (%)	60.0	30.0
Smoke (everyday: someday: no)	3.5: 1.0: 9.0	3.1: 1.0: 5.2
Drink (%)	50.0	60.0

the year of 1996–2005. Each sample consists of with 13 independent variables and the target variable. The main characteristics of the experimental data are shown in Table 2. We can observe that there are 108,260 hypertension samples and 200,511 controls, which indicates that the class imbalance problem occurs.

3.2. Logistic regression analysis

Logistic regression is a type of statistical regression analysis model, and has the capacity to measure the relations between a categorical dependent variable and one or more independent variables, thus, it has been extensively used in numerous disciplines such as the medical and social science fields (Hosmer, Lemeshow, & Sturdivant, 2013). According to the number of possible values of the dependent variable, we can categorize logistic regression models into binary or multinomial models. In the binary logistic regression model, the values of the dependent variable are usually coded as zero and one to denote the two different outputs. In the case of hypertension, logistic regression models compute the probability of hypertension y ($y = 1$ if an individual suffering from hypertension; otherwise, $y = 0$) as a function of the risk factors. Specifically, through computing two conditional probabilities $p(y = 1|X)$ and $p(y = 0|X)$, where $X = (x_1, x_2, \ldots, x_n)$ represents n risk factors that are associated with the hypertension, we can obtain the likelihood that one is at risk of hypertension. The binary logistic regression model usually takes the following form.

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

where $X = (x_1, x_2, \ldots, x_n)$ represents the vectors of n risk factors determined by logistic regression analysis. β_i represents the coefficient of corresponding x_i ($1 \leqslant i \leqslant n$). Then, we can rewrite it as

$$p(y|X) = 1/(1 + \exp(-(\beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \ldots + \beta_n * x_n))). \tag{2}$$

After determining a suitable threshold *delta*, we infer that one is at risk of hypertension if $p(y = 1|X) > delta$. Besides these, logistic regression analysis is endowed with the capacity to select variables that are statistically significant to hypertension. In our study, binary logistic regression analysis is applied not only to select risk factors that are significant to hypertension, but also to build up a prediction model. Further, the selected risk factors are directed to a well-constructed neural network to improve the performance of logistic regression-based prediction model.

3.3. Multilayer perceptron neural network

As a nonlinear mapping model, ANNs are flexible and effective in modeling complex relationships between inputs and outputs. The effectiveness and flexibility of neural networks to solve classification and regression problems has been empirically validated in handwriting recognition (Knerr, Personnaz, & Dreyfus, 1992), speech recognition (Lippmann, 1989), and medical diagnosis (Amato et al., 2013; Chan, Ling, Dillon, & Nguyen, 2011). The typical processing procedure of an artificial neural network takes the following scheme. First, a set of input neurons are activated by the inputs, the activations of these neurons are then weighted, transformed and passed on to other neurons until the output neurons are activated and output the final results.

Multi-layer perception (MLP) neural networks are one of the most commonly used static neural networks (Vellido, Lisboa, & Vaughan, 1999). MLP are feed-forward neural networks that are trained with the back propagation (BP) algorithm, and utilize supervised learning techniques to transform input data into a desired response. Adopting the iterative gradient optimization algorithm, BP is trained with a generalized delta learning rule to obtain a model with high accuracy by minimizing the root mean square error between the actual outputs and desired outputs. The BP algorithm can be divided into two phases: propagation and weight update. In MLP, each layer is fully connected to the previous layer and there is no connection within the same layer. After completing the training procedure, we can obtain the weights on each edge and use them to test unseen samples. Algorithm 1 presents the pseudo-code of MLP. In our study, we employed MLP to explore the relation between the occurrence of hypertension and the risk factors and further optimized the hypertension prediction model.

```
Algorithm 1: Neural network learning with the back propagation algorithm
Input:   $N$ train samples, with inputs $x(1), x(2), \ldots, x(N)$ and
         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, $1 \leqslant i \leqslant N$
Output:  $NN$: a neural network
1:       Initializing network weights and biases to small random values
2:       Inputting a study sample $(x(p), y(p))$, $(1 \leqslant i \leqslant N)$
3:       Calculating the actual output of nodes in the hidden layer:
         $Y_j^2 = f\left(\sum_{i=1}^{n_1} W_{ij} * Y_i^1 - b_j\right) = f\left(\sum_{i=1}^{n_1} W_{ij} * X_{ip} - b_j\right), \quad j \in \{1, 2, \ldots, n_2\}$   (3)
4:       Calculating the actual output of nodes in the output layer:
         $o_k = f\left(\sum_{i=1}^{n_2} W_{jk} * Y_j^2 - b_k\right), \quad k \in \{1, 2, \ldots, m\}$   (4)
5:       Adapting weights $W_{ij}$ and biases $b_i$, using Eqs. (5) and (6):
         $\Delta w_{ij}^{(l)} = \mu * x_j * \delta_i^{(l)},$   (5)
         $\Delta b_i^{(l)} = \mu * \delta_i^{(l)},$   (6)
         where $\mu$ is learning rate, $x_j(n)$ is the output of node $j$ at the iteration $n$.
         $\delta_i^{(l)}(n) = \left\{ \begin{array}{ll} \varphi'\left(net_i^{(l)}\right) * (y_i - o_i), & l = M \\ \varphi'\left(net_i^{(l)}\right) * \sum_k w_{ki} * \delta_k^{(l)}, & 1 \leqslant l < M \end{array} \right\},$   (7)
         where $l$ is the layer, $M$ is output layer, $k$ is the number of output nodes.
6:       If left study sample, goto step 2.
7:       Calculating error function $E$, if $E$ satisfying, stop; else, goto step 2.
```

3.4. K-means algorithm

The aim of a clustering algorithm is to group a set of objects into several clusters so that objects in the same cluster are more similar to each other than to objects in other clusters. Among the various metrics, Euclidean distance is commonly used to measure the similarity between two instances. Given two variables X and Y with N numeric attributes, the similarity between X and Y is defined as

$$d(X,Y)=\sqrt{\sum_{i=1}^{N}(X_i-Y_i)^2}. \tag{8}$$

The K-means cluster is a simple but powerful algorithm and has fast convergence in actual use (MacQueen, 1967). K-means algorithm works in the following way: (1) it first randomly designates K objects as the initial cluster center, and calculates the similarity between each object and the K cluster centers; (2) it puts each object into the closest cluster, re-calculates the centroid of the newly formed clusters and substitutes previous clusters with them; (3) repeat the second step until no change occurs to any cluster. In our study, to circumvent the biased results caused by the class imbalance problem, we proposed an under-sampling technique built on K-means algorithm to balance the experimental data.

3.5. Evaluation measures

In the evaluation of a classifier, a confusion matrix contains the actual outputs and predicted outputs of a classifier, and is applicable to evaluate the classification performance (Provost & Kohavi, 1998). Table 3 presents the confusion matrix for hypertension prediction. To evaluate the performance of the proposed prediction model, we used the following four measures. The higher accuracy, sensitivity, specificity and *AUC* are, the better the proposed prediction model.

(1) *Accuracy* is defined as the total accuracy rate of classifying each case correctly. Accuracy is an index that can present the power of a model in correctly predicting an individual's health condition.

$$Accuracy=(TP+TN)/(TP+FP+TN+FN). \tag{9}$$

(2) *Sensitivity* refers to the probability of correctly predicting an individual at risk of hypertension. A higher sensitivity indicates that the model can easily detect hypertension.

$$Sensitivity=TP/(TP+FN). \tag{10}$$

(3) *Specificity* represents the probability of correctly determining that an individual has no hypertension.

$$Specificity=TN/(TN+FP). \tag{11}$$

(4) The area under the ROC curve (*AUC*) presents the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Bradley, 1997). An area of one 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 estimated using the following formula (Vila-Francés et al., 2013).

$$AUC=\frac{s-(pos\times(pos+1)/2)}{pos\times neg}, \tag{12}$$

where s is the sum of ranks of true hypertension cases, pos denotes the number of hypertension cases, and neg denotes the number of controls.

Table 3
Confusion matrix for hypertension prediction.

Class predicted	Real situation	
	Hypertension	Normal
Hypertension	TP	FP
Normal	FN	TN

4. Experimental design and results

In this section, we detailed the experimental design and presented corresponding results. To determine hypertension-associated risk factors and establish a prediction model with these factors, we first utilized binary logistic regression analysis on the collected dataset to select variables that were significant to hypertension according to the statistically significant p-value. We then built up an ANN-based prediction model with these factors as inputs. As a comparison to the prediction performance obtained by neural networks, we presented the experimental results of logistic regression-based prediction model as well. Furthermore, to mitigate the class imbalance problem, we applied the proposed under-sampling technique to balance the experimental data prior to the training of the prediction model, and experimentally validated it.

4.1. Significant risk factors for hypertension

Logistic regression analysis has the capacity to function as a prediction model and to determine significant factors. In order to select the significant risk factors, a multi-factor logistic regression model with maximum likelihood estimation and forward-step regression analysis was applied. Consequently, eleven 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 when 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 4 presents corresponding results, where B denotes the coefficients of each variable in the logistic regression-based prediction model. After investigating the distribution of each variable of the dataset, we found that variable "smoke100" only had one value, so it was not involved in the regression analysis.

4.2. Artificial neural network-based prediction model

ANNs typically consist of one input layer, one output layer, zero or more hidden layers and a collection of neurons with connectivity between layers. Generally, the architecture of ANNs is determined by the number of inputs n and outputs m, the number of hidden layers and the number of neurons h in each hidden layer. In our study, we set n equal to eleven since there are eleven risk factors, and use two nodes to represent the 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 (Chen, Chen, & Liu, 1995; Kolmogorov, 1957). Due to the fact that the training time of a neural network model increases with the number of hidden layers, in our study, ANNs with single hidden layer are adopted towards a tradeoff between accuracy and time performance. In determining the number of neurons h in the hidden layer, three rules-of-thumb were used to reduce the search space rather than using a grid-based or exhaustive search scheme to search for the best-fitting value of h.

Table 4
Multi-factor logistic regression analysis for hypertension.

Variable	B	Wald	*P*-value	Odd ratio (95% CI)
Exercise	−0.130	177.438	$<10^{-3}$	0.878 (0.861–0.895)
Diabetes	0.350	2616.923	$<10^{-3}$	1.420 (1.401–1.439)
Hyperlipemia	0.748	7650.724	$<10^{-3}$	2.112 (2.077–2.148)
Age	−0.046	18513.797	$<10^{-3}$	0.955 (0.955–0.956)
Marriage	−0.013	15.459	$<10^{-3}$	0.987 (0.981–0.993)
Education	0.046	103.814	$<10^{-3}$	1.047 (1.038–1.056)
Income	0.076	925.898	$<10^{-3}$	1.079 (1.073–1.084)
Weight	−0.011	8795.350	$<10^{-3}$	0.989 (0.988–0.989)
Height	0.003	453.368	$<10^{-3}$	1.003 (1.003–1.003)
Sex	−0.058	32.369	$<10^{-3}$	0.944 (0.925–0.963)
Smoke	0.006	1.219	0.270	1.006 (0.996–1.016)
Drink	−0.034	13.974	$<10^{-3}$	0.967 (0.950–0.984)

(1) Blum suggested that the number of neurons in the hidden layer should be limited between the number of inputs and outputs (Blum, 1992).

$$m \leqslant h \leqslant n. \tag{13}$$

(2) Boger and Guterman pointed out that the number of neurons in the hidden layer should be more than two thirds of the number of inputs (Boger & Guterman, 1997).

$$h \geqslant \frac{2}{3} * n. \tag{14}$$

(3) Berry and Linoff suggested that the number of neurons in the hidden layer should be less than twice the number of inputs for circumventing high computation (Berry & Linoff, 1997).

$$h \leqslant 2 * n. \tag{15}$$

Considering the three constraint conditions simultaneously, we derived that the number of neurons h should be chosen between eight and eleven in our study.

In choosing activation functions, Karlik and Olgac (2011) conducted a comparative study and evaluated the performance of five MLP neural networks with different conventional activation functions, including Bi-polar sigmoid, Uni-polar sigmoid, Hyperbolic Tangent (*Tanh*), Conic Section, and Radial Bases Function (RBF). They concluded that the activation function *Tanh* performed better in the vast majority of MLP applications (Karlik & Olgac, 2011; Tan, Teo, & Anthony, 2011). Directed by their work, we chose *tanh* as the activation function in the hidden layer and output layer.

In addition, we adopted the back propagation algorithm with learning rate μ and momentum *mc* to achieve faster convergence with minimum oscillation, and assigned empirical values to the two parameters. According to the discussions above, parameters of the proposed prediction model and corresponding values were summarized and presented in Table 5. Furthermore, on the basis of the discussion of the neural network architecture, Fig.1 presents the constructed hypertension prediction model. In the input layer, there are eleven variables that are obtained using the binary logistic regression analysis; in the output layer, there are two output nodes to denote hypertension and normal; the number of neurons in the hidden layer ranges from eight to eleven.

4.3. Experimental results and analysis

To evaluate the performance of the proposed approach and compare it with the logistic regression-based prediction model, we randomly partitioned the experimental dataset into a training set and a test set in the ration of 7:3. The training set was used to optimize model parameters and construct the prediction model, while the test set was used to evaluate the model. In our study, the initial parameter values for the neural network prediction model were listed in Table 5, and the value of training period varied from 100,000 to 2,000,000. As a comparison, the logistic regression-based prediction model was built up using Eq. (2), and its coefficients were obtained using the logistic regression analysis and shown in Table 4. In determining hypertension, 0.5 was taken as the threshold. We inferred that the subject suffered from hypertension if the predicted value was greater than 0.5; otherwise, we predicted that the individual did not have high blood pressure. We ran each experiments ten times and recorded the averaged results and standard deviations with the varying number of neurons in the hidden layer. Table 6 presented the experimental results of the proposed approach and its contrast.

From the experimental results in Table 6, we can observe that the artificial neural network-based approach obtained an average prediction accuracy ranging from 71.91% to 72.12% and an average *AUC* of 0.77 with h vary from 8 to 11. And the best prediction accuracy was found up to 72.12% when the number of neurons was equal to 11. Senior physicians suggest that 30.0% is an acceptable error rate for the diagnosis of hypertension (Blinowska et al., 1991). Namely, prediction accuracy over 70.0% is useful, which indicates the effectiveness of our proposed neural network-based prediction model. We can also observe that logistic regression-based prediction model obtained 71.96% accuracy. Though its results were very close to that of neural network-based model, it had larger standard deviations. In contrast, neural network-based model achieved better accuracy and comparatively small standard deviations when h was equal to 11. This indicates that the neural network-based method have better stability and robustness than logistic regression-based model, and that neural network-based method is more powerful in adjusting itself to new environments and more suitable to model the complex relations between variables in real world applications. In terms of *AUC*, logistic regression-based approach obtained an *AUC* of 0.74, which was less than 0.77 achieved by neural network-based approach. This further demonstrated the superiority of neural networks in hypertension prediction.

Notably, we can observe from Table 6 that there is a great difference between the value of sensitivity and specificity. Specifically, the probability of correctly determining that the subject does not have hypertension is twice the probability of correctly classifying that one suffers from hypertension. This is mainly caused by the class imbalance problem, that is, the number of instances belonging to one class is much larger than the ones of other classes. Consequently, constructing a classifier with all the data is generally biased towards the majority class in order to obtain higher accuracy (Fernández, López, Galar, José del Jesus, & Herrera, 2013; Tahir, Kittler, & Yan, 2012). From Table 2, we can see that the number of controls is twice the number of hypertension cases, and this explains why specificity is much larger than sensitivity.

To circumvent the imbalanced sample issue, we proposed an effective under sampling technique built on *K*-means to balance the dataset. The proposed algorithm mainly consists of three steps

Table 5
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

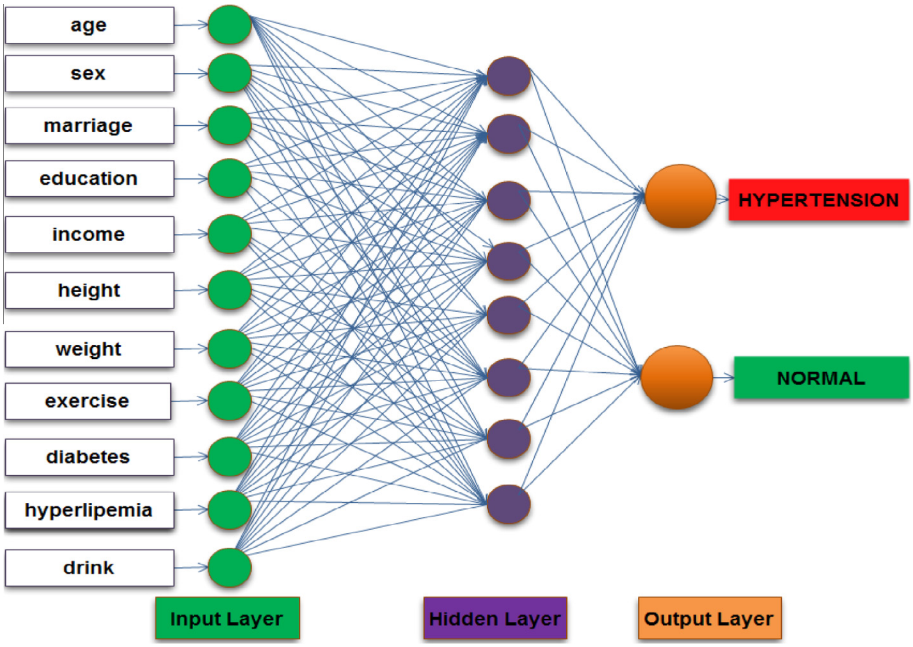

Fig. 1. An artificial neural network-based hypertension prediction model.

Table 6
Prediction results of neural network and logistic regression.

Model		Sensitivity (%)	Specificity (%)	Accuracy (%)	*AUC*
Artificial neural network	*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
Logistic regression		44.68 ± 5.17	86.42 ± 2.66	71.96 ± 0.21	0.74 ± 0.001

(see Algorithm 2). First, *K*-means algorithm is applied to group the majority class samples into *K* clusters. Then, for each cluster, calculate the Euclidean distance between each sample in the cluster and the minor class samples and sort these samples in ascending order. Finally, select majority class samples from each cluster by proportion of the size of each cluster, and combine all these selected majority class samples with the minor class samples to form the final balanced dataset.

```
Algorithm 2. Under-sampling technique using K-means
Input:   D: imbalanced dataset, K: number of clusters
Output:  D': balanced dataset
1:       Extracting positive class samples P and negative
         class samples N. |N|: number of negative class
         samples, |P|: number of positive class samples.
2:       Applying K-means algorithm to group N into K
         clusters: $C_1, C_2, \ldots, C_k$. $|C_j|$: size of each cluster,
         $1 \leqslant j \leqslant K$.
3:       For $j = 1$ to K
             For each sample $X_h$ in $C_j$ ($1 \leqslant h \leqslant |C_j|$), calculating its
             Euclidean distance to P: $dist(X_h, P)$.
             Sorting $dist(X_h, P)$ in ascending order.
             Selecting $C'_j$ samples from $C_j$ by proportion. The size
             of $C'_j$ is $|C'_j| = (|C_j|/|N|) * |P|$, and selecting $|C'_j|$ with
             smaller $dist(X_h, P)$.
4:       Merging all $C'_j$ into a set N'.
5:       Combining P and N' to form a new dataset D'.
```

Table 7
Prediction results of neural networks on balanced dataset.

Neurons in hidden layer	Sensitivity (%)	Specificity (%)	Accuracy (%)	*AUC*
h = 8	72.90 ± 1.12	67.87 ± 1.05	70.37 ± 0.07	0.77 ± 0.001
h = 9	71.39 ± 2.60	69.03 ± 1.84	70.19 ± 0.39	0.77 ± 0.004
h = 10	72.37 ± 1.38	67.96 ± 0.79	70.16 ± 0.49	0.77 ± 0.006
h = 11	72.76 ± 0.69	67.96 ± 0.61	70.34 ± 0.07	0.77 ± 0.001

In our study, we set the number of clusters to be fifteen, and then applied the proposed under-sampling technique to obtain a balanced experimental dataset. Similar to the above experimental settings, we partitioned the balanced dataset into a training set and a test set in the ratio of 7:3 with the training period varying from 100,00 to 2,000,000. We also conducted the experiments ten times and reported the average results and their standard deviations of sensitivity, specificity, accuracy and *AUC*. Table 7 presented the experimental results. We can observe that it can greatly improve the sensitivity to 72.0% from 48.0%, and that the difference between sensitivity and specificity was reduced, which demonstrated the effectiveness of the proposed under-sampling technique. In medical diagnosis, sensitivity is associated with Type I Error and specificity is associated with Type II Error. Type II Error β represents the probability of classifying healthy subjects into hypertension group, and β is equal to 1-specificity. In hypertension prediction, this error is acceptable because wrongly classifying healthy individuals into hypertension group will draw their attentions to hypertension and potential risk factors. Moreover, we can still obtain an accuracy over 70.0% and an *AUC* of 0.77.

5. Conclusion

Predicting hypertension with clinical test data, anthropometric data or genetic data generally obtains better accuracy, but these types of data are indicators for hypertension prediction and provide us less information about risk factors. Due to the fact that lifestyle choices are linked to the occurrence and development of hypertension, therefore, determining hypertension risk factors and further establishing a predictive model will contribute to the early prevention and effective management of hypertension. In this study, we proposed to predict hypertension in a non-invasive way using the simple demographics recorded in questionnaires rather than using clinical or genetic data. The proposed approach mainly consists of three parts. First, the binary logistic regression was used to determine risk factors that are significantly relevant to hypertension. Second, a well-defined neural network was constructed and optimized for hypertension prediction. Last, to improve the biased classification performance, we proposed an effective under-sampling technique prior to the training of prediction model. To show the effectiveness of the proposed approach, we included logistic regression-based prediction model as a comparison, and conducted experimental comparisons on the publicly available BRFSS datasets in terms of sensitivity, specificity, accuracy and *AUC*. Experimental results show that the proposed approach can obtain more than 72.0% accuracy and 0.77 *AUC*, and is applicable to hypertension prediction.

In particular, the main contributions of our study mainly include the following four aspects. (1) We propose to predict hypertension only using the questionnaires other than clinical test data, anthropometric data or genetic data. Its effectiveness demonstrates the practicability of developing a hypertension surveillance system for a large scale of population in a non-invasive and economical way. And the results from this study may be used to guide the development of programs geared towards preventing and mitigating specific hypertension risk factors. (2) We propose to integrate logistic regression analysis and artificial neural networks for simultaneous risk factor selection and hypertension prediction. Though we only consider hypertension as a study case in this paper, the proposed approach is essentially a general framework that can facilitate researchers to analyze other chronic diseases and other types of data. (3) We detail the selection of artificial neural network architecture and the setting of relevant parameters, which is a difficult and challenging task in model learning. This can potentially relieve researchers of the complex model selection issue and enable them to focus on the problems under investigation. (4) To deal with the class imbalance problems, we propose an effective under-sampling technique. Built on a cluster algorithm and selecting the representative samples from each cluster in the proportion of the cluster size, the proposed method can select the most discriminative samples from the majority class while causing us to lose the least amount of information.

For the future work, we plan to work in the following lines. First, although we tested the effectiveness of the proposed approach merely on questionnaires, it is a general framework that can applied to other situations. Thus, one of the future works involves applying the proposed approach to predict other chronic diseases (e.g. diabetes and asthma) as well as to analyze other types of data such as clinical data and anthropometric data for hypertension prediction. Second, model selection greatly influences the finally obtained prediction performance. We then plan to explore other strategies to construct the architecture of artificial neural networks and compare it with the proposed one in this study. Third, to obtain a balanced dataset with the discriminative majority class samples, we used *K*-means clustering algorithm, which requires us to designate the number of clusters to be obtained. So we plan to explore other self-determined cluster algorithms such as DBSCAN and CURE to automate the proposed method. Finally, working on the imbalanced dataset, most classification methods tend to bias towards the majority class, as is the case of hypertension prediction in this study. Therefore, how to effectively deal with the class imbalance problem remains another topic for future research.

Acknowledgements

This work was partially supported by the “111 Project” of Ministry of Education and State Administration of Foreign Experts Affairs (No. B14025), the International S&T Cooperation Program of China (No. 2014DFA11310), the Natural Science Foundation of China (No. 61305064), the “University Featured Project” of Ministry of Education (No. TS2013HFGY031), and the Chinese National Key Technology R&D Program (No. 2013BAH19F01).

References

Akdag, B., Fenkci, S., Degirmencioglu, S., Rota, S., Sermez, Y., & Camdeviren, H. (2006). Determination of risk factors for hypertension through the classification tree method. *Advances in Therapy, 23*, 885–892.

Amato, F., López, A., Peña-Méndez, E. M., Vaňhara, P., Hampl, A., & Havel, J. (2013). Artificial neural networks in medical diagnosis. *Journal of Applied Biomedicine, 11*(2), 47–58.

Berry, M. J., & Linoff, G. (1997). *Data mining techniques. For marketing, sales, and customer support*. John Wiley & Sons, Inc.

Blinowska, A., Chatellier, G., Bernier, J., & Lavril, M. (1991). Bayesian statistics as applied to hypertension diagnosis. *IEEE Transactions on Biomedical Engineering, 38*(7), 699–706.

Blinowska, A., Chattellier, G., Wojtasik, A., & Bernier, J. (1993). Diagnostica–A Bayesian decision-aid system-applied to hypertension diagnosis. *IEEE Transactions on Biomedical Engineering, 40*(3), 230–236.

Blum, A. (1992). *Neural networks in C++: An object-oriented framework for building connectionist systems*. John Wiley & Sons, Inc.

Boger, Z., & Guterman, H. (1997). Knowledge extraction from artificial neural network models. In *IEEE international conference on systems, man, and cybernetics, 1997. Computational cybernetics and simulation* (vol. 4, pp. 3030–3035).

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145–1159.

Caulfield, M., Munroe, P., Pembroke, J., Samani, N., Dominiczak, A., Brown, M., et al. (2003). Genome-wide mapping of human loci for essential hypertension. *The Lancet, 361*(9375), 2118–2123.

Chan, K. Y., Ling, S. H., Dillon, T. S., & Nguyen, H. T. (2011). Diagnosis of hypoglycemic episodes using a neural network based rule discovery system. *Expert Systems with Applications, 38*(8), 9799–9808.

Chang, C. D., Wang, C. C., & Jiang, B. C. (2011). Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors. *Expert Systems with Applications, 38*(5), 5507–5513.

Chen, T., Chen, H., & Liu, R. W. (1995). Approximation capability in C (R¯n) by multilayer feedforward networks and related problems. *IEEE Transactions on Neural Networks, 6*(1), 25–30.

Chiu, C., Hsu, K. H., Hsu, P. L., Hsu, C. I., Lee, P. C., Chiou, W. K., et al. (2007). Mining three-dimensional anthropometric body surface scanning data for hypertension detection. *IEEE Transactions on Information Technology in Biomedicine, 11*(3), 264–273.

Fernández, A., López, V., Galar, M., José del Jesus, M., & Herrera, F. (2013). Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-Based Systems, 42*, 97–110.

Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. Wiley.com.

Hsu, K. H., Chiu, C., Chiu, N. H., Lee, P. C., Chiu, W. K., Liu, T. H., et al. (2011). A case-based classifier for hypertension detection. *Knowledge-Based Systems, 24*(1), 33–39.

Huang, H. H., Xu, T., & Yang, J. (2014). Comparing logistic regression, support vector machines, and permanental classification methods in predicting hypertension. *BMC proceedings* (Vol. 8 (Suppl. 1), pp. S96). BioMed Central Ltd.

Jeppesen, J., Hein, H. O., Suadicani, P., & Gyntelberg, F. (2000). High triglycerides and low HDL cholesterol and blood pressure and risk of ischemic heart disease. *Hypertension, 36*(2), 226–232.

Karlik, B., & Olgac, A. V. (2011). Performance analysis of various activation functions in generalized MLP architectures of neural networks. *International Journal of Artificial Intelligence and Expert Systems, 1*(4), 111–122.

Kesselmeier, M., Legrand, C., Peil, B., Kabisch, M., Fischer, C., Hamann, U., et al. (2014). Practical investigation of the performance of robust logistic regression to predict the genetic risk of hypertension. *BMC proceedings* (Vol. 8 (Suppl. 1), pp. S65). BioMed Central Ltd.

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

Kolmogorov, A. N. (1957). On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. *Dokl. Akad. Nauk SSSR, 114*(5), 953–956.

Krawczyk, B., & Wozniak, M. (2011). Hypertension diagnosis using compound pattern recognition methods. *Journal of Medical Informatics & Technologies, 18*, 41–50.

Lee, M., & Entzminger, L. (2006). Risk factors of hypertension and correlates of blood pressure and mean arterial pressure among patients receiving health exams at Thailand. *Journal of the Medical Association of Thailand, 89*(8), 1213–1221.

Lippmann, R. P. (1989). Review of neural networks for speech recognition. *Neural Computation, 1*(1), 1–38.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 281–297, p. 14).

Midha, T., Krishna, V., Nath, B., Kumari, R., Rao, Y. K., Pandey, U., et al. (2014). Cut-off of body mass index and waist circumference to predict hypertension in Indian adults. *World Journal of Clinical Cases: WJCC, 2*(7), 272.

Mokdad, A. H., Ford, E. S., Bowman, B. A., Dietz, W. H., Vinicor, F., Bales, V. S., et al. (2003). Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001. *Jama, 289*(1), 76–79.

Oswald, A. J., & Hu, S. (2010). Objective confirmation of subjective measures of human wellbeing: Evidence from the USA. *Science, 327*, 576–579.

Provost, F., & Kohavi, R. (1998). Guest editors' introduction: On applied research in machine learning. *Machine Learning, 30*(2), 127–132.

Sanada, H., Yoneda, M., Yatabe, J., Williams, S. M., Bartlett, J., White, M. J., et al. (2015). Common variants of the G protein-coupled receptor type 4 are associated with human essential hypertension and predict the blood pressure response to angiotensin receptor blockade. *The Pharmacogenomics Journal*.

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

Tahir, M. A., Kittler, J., & Yan, F. (2012). Inverse random under sampling for class imbalance problem and its application to multi-label classification. *Pattern Recognition, 45*(10), 3738–3750.

Tan, T. G., Teo, J., & Anthony, P. (2011). A comparative investigation of non-linear activation functions in neural controllers for search-based game AI engineering. *Artificial Intelligence Review*, 1–25.

Ture, M., Kurt, I., Turhan Kurum, A., & Ozdamar, K. (2005). Comparing classification techniques for predicting essential hypertension. *Expert Systems with Applications, 29*(3), 583–588.

Vasan, R. S., Larson, M. G., Leip, E. P., Evans, J. C., O'Donnell, C. J., Kannel, W. B., et al. (2001). Impact of high-normal blood pressure on the risk of cardiovascular disease. *New England Journal of Medicine, 345*(18), 1291–1297.

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

Vila-Francés, J., Sanchís, J., Soria-Olivas, E., Serrano, A. J., Martínez-Sober, M., Bonanad, C., et al. (2013). Expert system for predicting unstable angina based on Bayesian networks. *Expert Systems with Applications, 40*, 5004–5010.

Wang, S., & Yao, X. (2012). Multiclass imbalance problems: analysis and potential solutions. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, 42*(2), 1119–1130.

Wong, N. D., Thakral, G., Franklin, S. S., Gil, J. L., Jacobs, M. J., Whyte, J. L., et al. (2003). Prevention and rehabilitation: preventing heart disease by controlling hypertension: Impact of hypertensive subtype, stage, age, and sex. *American Heart Journal, 145*(5), 888–895.

Wozniak, M. (2006). Two-stage classifier for diagnosis of hypertension type. *Biological and Medical Data Analysis*, 433–440.

Ziada, A. M., Lisle, T. C., Snow, P. B., Levine, R. F., Miller, G., & Crawford, E. D. (2001). Impact of different variables on the outcome of patients with clinically confined prostate carcinoma: Prediction of pathologic stage and biochemical failure using an artificial neural network. *Cancer, 91*(2001), 1653–1660.