An Empirical study on Predicting Blood Pressure using Classification and Regression Trees

Bing Zhang,Zhiyao Wei,Jiadong Ren,Yongqiang Cheng and Zhangqi Zheng

Abstract—Blood pressure diseases have become one of the major threats to human health. Continuous measurement of blood pressure has proven to be a prerequisite for effective incident prevention. In contrast with the traditional prediction models with low measurement accuracy or long training time, non-invasive blood pressure measurement is a promising use for continuous measurement. Thus in this paper, classification and regression trees (CART) are proposed and applied to tackle the problem. Firstly, according to the characteristics of different information, different CART models are constructed. Secondly, in order to avoid the over-fitting problem of these models, the cross-validation method is used for selecting the optimum parameters so as to achieve the best generalization of these models. Based on the biological data collected from CM400 monitor, this approach has achieved better performance than the common existing models such as linear regression, ridge regression, the support vector machine and neural network in terms of accuracy rate, root mean square error, deviation rate, Theil IC, and the required training time is also comparatively less. With increasing data, the accuracy rate of predicting systolic blood pressure and diastolic blood pressure by CART exceeds 90%, and the training time is less than 0.5s.

Index Terms—classification and regression tree, pruning, blood pressure prediction, health monitoring.

1 Introduction

BLOOD pressure (BP) is one of the most important physiological indicators of human body. Every beat of the heart sends blood to the arteries, and *BP* changes throughout the process. When the heart contracts, *BP* in the vessels reaches its maximum, termed systolic blood pressure (Ps); when the heart rests, the minimum *BP* reaches, termed diastolic blood pressure (Pd). *BP* is closely related to cardiac function and peripheral blood vessels, and it is used as a direct measurement of the cardiovascular functions of human body. Therefore, accurate and effective *BP* measurement plays a vital role in clinical practice. As is well-known, *BP* disease is one of the leading causes of death and heart and kidney diseases. The accurate and timely *BP* diagnosis is thus essential to human health because the prevention of *BP* is a useful medical means for controlling and lowering the physical harm of a *BP* disorder. Although it is necessary to monitor *BP* levels routinely in elderly [1-3], in fact this has not been carried out in large elderly population. Early detection of abnormal *BP* would reduce the risk to further inducing the diseases of the heart, brain and other organs. Furthermore, most hypertensive patients dont show alert signs or symptoms when their *BP* reaches an emergency level. Therefore, routine *BP* measurement is indispensible for preventive healthcare.

The existing *BP* measurement methods can be divided into two categories, invasive and non-invasive measurements [4]. Invasive measurement is more accurate but traumatic, limitting its widespread clinical use; Non-invasive methods, however, are more likely to be accepted by patients because of their less intrusion.

Non-invasive *BP* measurement indirectly measures *BP*, of which the two commonly used approaches are measuring the pulse of the arterial wall and the pulse wave transmission time respectively. Non-invasive measurement methods can also be divided into intermittent *BP* measurement and continuous *BP* measurement according to their continuity and discontinuity.

Intermittent *BP* measurement is typical of the Kochi's auscultation or the oscillatory method. The Kochi's auscultation method is one of the most commonly used methods in non-invasive *BP* measurement based on the detection of Koch's sound. Kochi's auscultation devices are low cost and easy to use, which has been widely used since it was introduced. Unfortunately, the measurement results almost entirely depend on the operators ability of identifying the readings, so the results are subjectively affected, by the operators hearing, visual coordination and sensitivity and might be different. Besides, it also causes discomfort and potential injury risk to patient during obstructing blood flow. Although these disadvantages of Kochis auscultation method has limited its applications, it is still considered as the "golden standard" in non-invasive *BP* measurement methods due to its accuracy that satisfies noninvasive *BP* measurement researchers. The limitation of Kochis auscultation method has, however, encouraged the introduction of other *BP* measurement, such as the oscillation method known as oscillatory method or vibration method adopted by most electronic sphygmomanometers available in the market. Sphygmomanometer is easy to use, not affected by the operator and has low measurement error rate. It is also convenient for home users without professional medical

- *Bing Zhang,Zhiyao Wei, Jiadong Ren and Zhangqi Zheng are with the School of Information Science and Engineering, Yanshan University, Qinhuangdao, Hebei P.R.China. E-mail:bingzhang@ysu.edu.cn;841182743@qq.com;jdren@ysu.edu.cn; 451499304@qq.com*
- *Yongqiang Cheng is with the School of Engineering and Computer Science, University of Hull, Hull, UK, HU6 7RX. E-mail:y.cheng@hull.ac.uk*

Manuscript received XXX.

knowledge.

However, the oscillation method can hardly detect the *BP* changes of high frequency as its main use is to detect the amplitude of the oscillation wave superimposed on *BP* signal. If there occurs a sudden change in the patient's *BP*, the oscillation method will fail to pick up this alteration. In addition, the oscillation method has low tolerance to motion artifacts. As the patient needs to remain calm during the measurement, so it is often needed to judge whether the process has been disturbed by the patients movement so as to ensure that the *BP* measurement results are accurate. Although the Kochi's auscultation and the oscillatory method are the most widely used measurement with the largest market shares in *BP* detection, the two methods are required to inflate and deflate the cuff to obstruct blood flow. In other words, they are only suitable for intermittent *BP* measurement with a single reading record in each measurement, which has hindered its application for cardiovascular and cerebrovascular diseases analysis requiring continuous *BP* readings.

In order to meet the needs of people's long terms health care monitoring, continuous *BP* measurement methods were put forward [5], including volume compensation, tension measurement, optical coherence tomography and pulse wave measurement. Volume compensation is an effective and relatively mature non-invasive continuous *BP* measurement method adopted by the current non-invasive continuous *BP* measurement equipment in the market. But with regards to the continuous measurement for a longer time, venous accumulation will exert greater influence on the accuracy of the measurement results. Meanwhile, the prolonged measurement need to keep patients arm under pressure, which may cause patients discomfort. In this regard, it is not suitable for long time use. The tension measurement method applies external pressure to the arteries to balance the internal *BP* so that the surface becomes flat. By measuring external pressure, it can be the arterial pulse pressure wave can be obtained and then the *BP*. Theoretically, the tension measurement method can continuously monitor the arterial *BP*, but it requires very high position and angle precision in taking readings. Hence the long-time continuous *BP* measurement can easily cause relative movement changes in the position of sensors. At present, due to the fact the research of the volume pulse blood flow signal is still in its infancy with immature theoretical research of optical volume recording, so the accuracy of the detection is relative low.

In view of the above problems, this paper proposed the CART model for analyzing the vital signs collected by CM400 experimental equipment. CM400 is capable of continuously measuring left atrium (AVL), right atrium (AVR), anterior atrium (AVF), photoplethysmography (PPG), heart rate (HR), pulse wave translation time (PTT) and blood oxygen (SPO2). In order to prevent the over fitting or under fitting of the CART model, cross-validation method was adopted in the training process, with these parameters searched and optimized automatically. The CART model were compared with other four traditional machine learning models (supporting vector machine SVM, linear regression, ridge regression, neural network) combining four evaluation indexes, i.e. the accuracy rate [6], root mean square error RMSE [7], the Theil inequality Coefficient *TIC* and the deviation rate [8]. By means of feature correlation analysis, it has been found that the two strongest features related to *BP* prediction are HR and PTT. We also found that the CART model has high time efficiency suitable for real-time prediction.

2 Related work

Traditional *BP* measurement methods and devices are not available for real-time, continuous and non-invasive *BP* prediction. With the development and changes of information perception of information and communication technologies like mobile Internet, cloud computing, smart sensors and robots, machine learning methods have gained their application in the field of medical science, which can be used to build accurate prediction models and discover implicit knowledge in data. The current medical standards are able to ensure rapid and convenient collection of measured characteristic data for *BP* prediction, therefore, the measurement indirectly based on machine learning methods is more worthy of in-depth study. Yi-yen, Hsieh, et al. [9] built a linear regression model using human PTT to predict high pressure and low pressure, whose experimental results are relatively consistent with the real values. For example, the correlation coefficient of high pressure is more than 0.8, and the correlation coefficient of low pressure reaches about 0.95. However, the prediction accuracy of high pressure is not accurate enough and needed to be improved. Costas Sideris [10] using the fingertip oximeter collected data, established a recurrent neural network model. The continuous *BP* values were predicted by use of it and the experimental results were satisfactory, but due to the too slow neural network training time, it lacks real time processing capability. Robabeh Abbasi et al [11] proposed a prediction model of *BP* prediction using multivariate time series based on fuzzy functions to aid doctors choosing a suitable treatment method in clinical diagnosis. However, it takes rather long time to complete a cycle, and the results vary among individuals. Besides, the prediction accuracy decreases with long time usage. Xiaohan Li et al [12] used context layer recursive model to predict *BP* whose data set was obtained from a certain type of wireless home *BP* monitor synchronized with a mobile phone and was further stored in the cloud server. In doing this, the whole historical data in the cloud server are used to train the *BP* prediction model. Although the accuracy of its *BP* prediction is relative high, the model based on the whole time series data requires a significant amount of data accumulation, so the security of the data cannot be guaranteed. Tony, Hao and Wu et al. [13] used neural networks and radial basis function networks to predict *BP* by modeling features such as body mass index, age, movement, alcohol consumption, and smoking, but the training time of this method was too long, leading to not high accuracy. Shrimanti and Ghosh et al. [14] established a linear regression model for human PPG and PTT in order to predict *BP* values. The accuracy was high in ideal situation, but as the linear model was too sensitive to noise, the *BP* measured in a real environment didnt perform well. Robert Munnoch, et al. [15] used Pulse wave velocity (PWV) calculated from PTT based on EIMO devices to predict *BP*

by comparing the usage of three linear regression models as shown in equation (1), (2), (3).

$$P = A(PWV)^2 + B \tag{1}$$

$$P = A(HR) + B \tag{2}$$

$$P = A(PWV)^2 + B(HR) + C \tag{3}$$

For the prediction of *Ps*, the best model is (3), and the average absolute error is 6.15 mmHg. For the prediction of *Pd*, the best model formula is (1), and the average absolute error is 8.36 mmHg. It is shown that these approaches have not achieved overall high accuracy.

3 CART MODEL

3.1 Model overview

The CART (classification and regression tree) model proposed by Breiman et al [16] in 1984 has become a widely used decision tree learning method. The CART method assumes that the decision tree is a binary tree. The input feature space is divided into finite units, based on which the predicted values are determined, i.e. the predicted output values are mapped to the given conditions. The CART model is composed of the following three main steps:

(1) CART initialization: generating a decision tree based on training data set;

(2) CART pruning and optimization: the regression tree is pruned according to some constraints, such as the maximum depth of the tree, the minimum sample number of the leaf node and the node's minimum impurity; and the model has best generalization through the combination of different parameters (the maximum depth of the tree (*max_depth*), the minimum sample number of leaf nodes (*min_samples_leaf*), the minimum impurity of the nodes (*min_impurity_split*)), for each combination of different parameters generated different CART models.

(3) CART prediction: put the test set into the trained model and predict it. Compared with other classical classification and regression models, CART has the following advantages: 1) Less data preparation: no data normalization is required; 2) The ability to handle continuous and discrete data simultaneously; 3) Capable of handling multiple classification problems; 4) The prediction process can be easily explained using Boolean logic as opposed to other learning models such as neural network and SVM.

3.2 CART Construction

CART is a supervised learning model [17] used to recursively divide the feature space into several sections (or nodes) based on the relationship between one output and one or more input factors [18]. In this paper, multiple input characteristics combinations (AVR, AVL, AVF, PPG, SPO2, PTT, HR) are used to establish the optimal CART trees for the predictions of *Ps* and *Pd*. The detailed descriptions of these characteristics are as follows.

- ECG, consisted of AVR, AVL and AVF, is a non-invasive examination technique which reflects the electrophysiological activity of the heart. Medical analysis shows that ECG has some relationship with *BP* classification, and it can be an important method to analyze hypertension and heart disease. Dynamic electrocardiogram and ambulatory *BP* monitoring can simultaneously observe the relationship among *BP* changes, ischemia myocardial and arrhythmia, which is contributed to the reliable clinic diagnosis and treatment.
- The photoplethysmography technology (PPG) is a noninvasive technique for detecting changes in blood volume in living tissues by photoelectric means, which can be used to predict many important health related parameters such as heart rate, hemoglobin, and blood glucose levels [19], and it can also be used to predict the value of continuous *BP* [20]. The low frequency part of PPG contains the breath, *BP* control, body temperature adjustment information, but the high frequency part contains the information related to the heart pulsation [21]. Cardiovascular detection methods based on PPG technology have achieved some success.
- Heart rate (HR) is the rate of heart beating, which has a great significance in human physiological function, health status evaluation, and the analysis of cardiovascular disease. An increase or decrease in HR will correspondingly cause an increase or decrease in *Ps* and *Pd*.
- Pulse transit time (PTT) is the time taken for cardiac ejection and arterial pulse wave propagation from the aortic valve to the peripheral branch vessel [22-25]. In recent years, as one of the characteristics for *BP* prediction, PTT has been widely applied [26-27] in non-invasive and real-time measurement.
- Oxygen saturation (SPO2) is an important physiological parameter to measure the content of the human blood oxygen. The arterial oxygen saturation is always estimated by analyzing the waveform of the light through the capillary tissue bed. The oxygen saturation of hemoglobin has become the standard of health care[28].

In general, the mean square error minimization rule [29] is used during the recursive process for feature selection to generate the binary tree. According to the characteristic of the related problems, the feature space X can be treated as the combination of N m-dimensional vector x. Where $n > 0$ is the number of samples and m=7 represents the seven input features {AVR,AVL,AVF,PPG,SPO2,PTT,HR}, the models output can be expressed as Y=$\{y_1,y_2,y_3,...,y_n\}$. By processing *Ps* and *Pd* separately, y_i is obtained either as *Ps* or *Pd*. For easier description, only *Ps* is discussed and *Pd* should follow the same processing procedure. By appending the output Y to X, Matrix D is obtained and determined by X and Y which is shown in equation (4) below.

$$D^{n\times m} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1(m-1)} & y_{1m} \\ X_{21} & X_{22} & \cdots & X_{2(m-1)} & y_{2m} \\ X_{31} & X_{32} & \cdots & X_{3(m-1)} & y_{3m} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{n(m-1)} & y_{nm} \end{pmatrix} \tag{4}$$

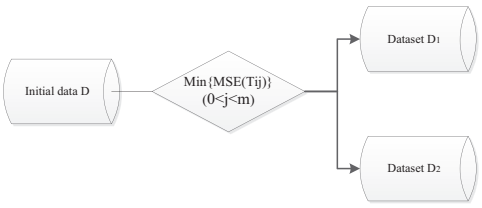

Fig. 1. The first division of tree construction.

The X_{ij} is the j^{th} attribute of the ith data sample. The CART model starts with sorting D in small to large order according to the values of the first column X_{i1}. In the sorted attribute values, the n sequence are represented as a_1, a_2,...,a_n where there will be total n-1 partitions, T_{i1}, divided by the n values as shown in the equation (5).

$$T_{i1} = \frac{a_i + a_{i+1}}{2} \tag{5}$$

For the X_{i1} feature column, the corresponding arbitrary partition point T_{ij} divides the data set D into two subsets D_i and D_{i+1}, The D_i and D_{i+1} in CART tree are represented by the left subtree and the right subtree respectively. The choice of T_{ij} needs to satisfy two conditions: 1) the sum of the mean square error (*MSE*) of D_i and D_{i+1} is minimum; 2) The individual *MSE* of D_i and D_{i+1} are minimum. In general, the choice of T_{ij} is to minimize the value of the equation (6).

$$MSE(T_{ij}) = [\min_{c_1} \sum_{x_i \in D_i} (y_j - c_1)^2 + \min_{c_2} \sum_{x_i \in D_{i+1}} (y_j - c_2)^2] \tag{6}$$

Where c_1 represents the average value of the *Ps* value in the dataset D_i, c_2 represents the average value of the *Ps* in the D_{i+1} dataset. y_j is the X_{i1} feature in the column corresponding to the ith sample of *Ps* value. Similarly, other feature columns X_{i2},X_{i3},..., X_{im} are calculated successively to obtain each partition point T_{ij} (1<j<m). Finally, the partition node T_{ij} (0<j<m) which has the minimum *MSE* among all nodes is selected as the root node. The first division of the binary tree is shown in **Fig 1**.

The process repeats as each split new into datasets D_i and D_{i+1} until the tree reach a steady state.

3.3 Pruning strategy of classification and regression tree

In the process of recursive construction of the CART model, *max_depth*, *min_samples_leaf* and *min_impurity_split* are three parameters used to govern the tree construction process, in which α=*max_depth*, β=*min_samples_leaf*, γ=*min_impurity_split*. α represents the allowed maximum depth of a tree; β represents the minimum number of samples required for a node to become a leaf node; γ denotes the minimum impurity required for node splitting. When the depth of the current tree is less than or equal to α, the current impurity of the node is less than or equal to γ and the number of samples in the current node is greater or equal to β, the node splits; If the current depth of the tree is equal to α, the lowest level, the tree stops splitting. If the number of samples is greater than or equal to β and impurity of the node is greater than γ, the node continues to split. If α of a tree is set to a too small value, it would make the model be under fitting. If it is set to a too large value, it would make the model over fitting; if β or γ is set to a too small value, it would lead to the over fitting of the model and if either is too large, it would lead to its under fitting. Therefore, to prevent the "overfitting" ort the "underfitting" problem in the rendering process, the tree is needed to be pruned according to the three parameters. In this paper, the pre-pruning method is adopted, in three steps:

(1) When the node splitting is determined, the *H(D')* of the node is calculated as equation(7) below.

$$H(D') = \frac{1}{N} \sum_{i \in D} (y_i - c)^2 \tag{7}$$

Where, D' is the dataset to be split, N is the sample number of the dataset D', y_i the *Ps* value of the ith sample in the dataset D, and c the average value of the *Ps* value in the dataset D. Calculate the *H(D')* of each dataset D' to be split, if *H(D')* $< \gamma$, it stops splitting and will turn the dataset to a leaf node, taking the predictive value of c.

(2) If the sample size of the dataset $< \beta$, then it cannot be converted into leaf nodes, and the samples are abandoned.

(3) If the depth of the tree is equal to α, then the split is stopped and the data set is changed into a leaf node with a predicted value of c.

3.4 CART model for *BP* prediction

According to the mentioned above, the *BP* prediction using CART model can be divided into three processes, namely construction, pruning and prediction, which are shown in **Algorithm 1**.

Algorithm 1 DecisionTreeRegressor

Input: D={(X_1, y_1), (X_2, y_2), ...,(X_n, y_n))}, X={X_1, X_2, X_3, ..., X_n} represents a feature property set, Y= {y_1, y_2, y_3, ..., y_n} represents the predicted attribute set
Output: CART model sets

```
for αϵ(1,Q)
    for βϵ(1,Q)
        for γϵ(1,Q)
            for x_i in X
                for T_j in x_i
                    search min(MSE);
                end for
            end for
            Build Trees TreeModel from T_j and x_i;
            if H(D) ≤ γ && the current depth < α && |D| ≥ β
                Divide D to D_1 and D_2
                DecisionTreeRegressor(D_1, α, β, γ);
                DecisionTreeRegressor(D_2, α, β, γ);
            else if |D| < β
                drop D
            else
                D → leaf_Node; // convert to leaf node
                predictive value = avg(D); // average of samples
                break;
            end if
```

21. **end for**
22. **end for**
23. **end for**
24. *TreeModel* set ← *TreeModel*;
25. **return** *TreeModel* set (*TMS*)
26. *BP* Prediction on *TMS*

In **Algorithm 1**, line 1 to 3 traverse all the feature parameter combinations, and each parameter combination is modeled subsequently. Line 4 to 9 mean the least square error minimization criterion used to obtain the feature attributes and partition points of the split selection. When the feature selection is started, build a tree step by step. Line 10 to 11 describe the pruning process of the tree according to three indicators (γ, α, β), once the condition satisfied, the dataset D would be split into two sub trees D_1 and D_2, which is expressed in line 12 to 13. Line 14 to 15 suggest that if the sample number of the nodes is less than β, the node would be dropped. Line 16 to 19 shows that if the node does not satisfy all these conditions, it will convert itself to a leaf node whose output value is the average of the samples on the node. In line 20 to 25, through different combinations of parameters, the different *TreeModel* are obtained and stored in *TreeModel* set. Finally, the process of *BP* prediction is conducted by using the obtained *TreeModel* in Line 26.

3.5 Optimization of CART model

To measure the fitness of the model, there are two commonly used indicators: absolute deviation squared $bias^2$ and *variance*. $bias^2$ describes the gap between the expected and the true values of the predicted value, and the greater $bias^2$, the more the fitted data deviated from the true data. *Variance* describes the range of change in the predicted value, i.e. the degree of dispersion, which *Variance* indicates the stability of the model and the greater the *variance*, the more dispersed the distribution of data. Therefore, the over fitting of the model has high *variance* and low $bias^2$ whist the under fitting features low *variance* and high $bias^2$. The optimization targets of the two indicators are often contradicting with each other. $bias^2$ is calculated as follows.

When the node splitting is determined, the *H(D')* of the node is calculated as equation(8).

$$bias^2 = \frac{\sum_{i=1}^{k}(m_i - n_i)^2}{k} \tag{8}$$

Variance is expressed below.

$$variance = \frac{\sum_{i=1}^{k}(m_i - m)^2}{k} \tag{9}$$

In equation (8) and (9), m_i represents the predictive value of the i^{th} sample, n_i the actual value of the i^{th} sample, K the total number of samples, and m the average of the predicted value. The accuracy of the training model is related to $bias^2$ and *variance* as shown in **Fig 2**.

As illustrated in **Fig 2**, with the decrease of $bias^2$ and the rise of *variance*, the Total Error (model prediction error) becomes smaller until the minimum point is reached. It should be pointed out that this minimum point is the best model for the generalization ability (dotted line). As $bias^2$ continues to decline and *variance* continues to rise, Total Error is beginning to rise again. What expected here is the model parameters corresponding to the dotted lines. Note that it is not the point at which $bias^2$ and *variance* intersect. The cross-validation method is used to select the optimal parameters. First, the training set D is divided into L large collections, then select L-1 as the training set of the model, combined with a set of D_i remaining as the test set and test accuracy of $TreeMode_{l_0}$, **Fig 3** shows the way to divide the training set and test set, L=5.

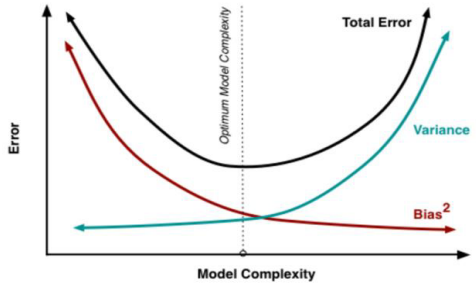

Fig. 2. $Bias^2$ and *variance* changes

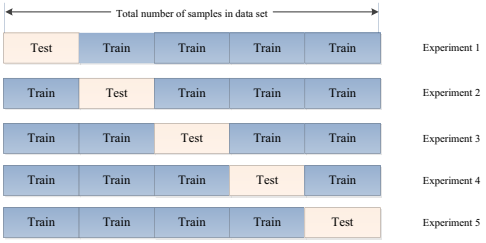

Fig. 3. Partition diagram of training set and test set

Each model evaluation uses the following equation (10) as the model evaluation method [30].

$$score = 1 - \frac{\sum_{i=0}^{k-1}(y_i - m_i)^2}{\sum_{i=0}^{k-1}(y_i - p)^2} = 1 - \frac{bias^2}{variance} \tag{10}$$

Where y_i represents the true value of the i^{th} sample, m_i the predicted value of the i^{th} sample, and P the average value of the sample true value. Cycle execution of the procedure L times to ensure that there is no duplication occurred. Then, for all the score averages, this is the *scores* of the model, and the equation is as follows.

$$scores = \frac{\sum_{i=1}^{L} score_i}{L} \tag{11}$$

By calculating the *scores* of the models, we search the optimal parameters for CART model of *Ps*. According to **Algorithm 1**, $\alpha \in (1, Q), \beta \in (1, Q)$ and $\gamma \in (1, Q)$, there will be Q^3 CART models based on the combination of different parameters, let *TMS*= $\{TreeMode_{l_0}, TreeMode_{l_1}, ..., \text{TreeModel}_{Q^3}\}$. Then according to each parameter combination, the corresponding *scores* of the models are calculated and recorded, and the model with the best *scores* will be chosen as shown in **Algorithm 2**.

Algorithm 2 Finding the best parameters and model of CART

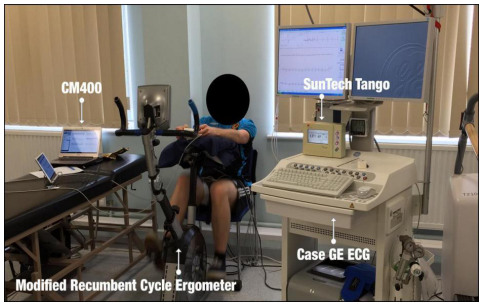

Fig. 4. Data Collection by the equipment *CM400*

Input: TMS={$TreeModel_0$, $TreeModel_1$, ...,$TreeModel_{Q^3}$}
Output: α, β, γ, $Treemodel_{best}$
1. temp=0; score=0;
2. **for each** model in TMS
3. **for** i=1:L
4. Calculate the $bias^2$, $variance$
5. $score+ = 1 - \frac{bias^2}{variance}$
6. **end for**
7. $scores = \frac{score}{L}$
8. **if** model$_i$.scores $>$ temp
9. temp = model$_i$.scores;
10. Treemodel$_{best}$ = model$_i$
11. **end if**
12. **end for**
13. **return** α, β, γ from TreeModel$_{best}$

Line 1 is for initialization of the temporary variables temp and score. Line 2 to Line 7 evaluate the *scores* of each model. Line 8 to line 12 are the searched model with the largest *scores*. The rest of the pseudo codes return these parameters.

4 Experimental Results and Analysis

The characteristic attributes data were collected via CM400, a PC based patient monitor from Contect Medical Systems Co. Ltd. The device provides a USB connection and a PC side program that shows the signal data captured by the device on a computer. The target *BP* measurements were recorded by SunTech Tango automated *BP* measuring apparatus. The data processing was carried out by Anaconda3-4.4.0-Windows-x86_64 software using Python3 programming language running on a 64-bit windows 7 ultimate operation system which has Inter(R) Xeon(R) E3-1231V3 @3.40G Hz CPU, 16.0GB Memory. The experiment setup is shown in **Fig 4**.

The data were collected from 18 healthy young people (including 12 males and 6 females) in simulated resting and exercise environments. The protocol consists of 30 minutes resting measurement, followed by 45 minutes exercising session and 3 minutes cooling down. Characteristic attributes data were recorded continuously and the cuff based *BP* measurements were taken in evenly distributed time slots which allowed a minimum of 2 minutes to relax arms and reduce discomfort. Six *BP* measures were taken during resting period, twelve *BP* measures during excising period and one final *BP* measurement after recovery. The data were imported into the analysis experimental platform, and the data set was randomly divided into training set and test set. Then, the training set was fed into the established CART model for training, and the parameters were optimized during the training to maximize the generalization ability of the model.

4.1 Data collection and analysis

There are a total of 15628501 sets of valid characteristic attributes data of AVR, AVL, AVF, PPG, SPO2, HR and PTT and 384 *BP* readings. Each Cuff based *BP* measurement took around 1 minute to complete, hence one *BP* reading has multiple corresponding characteristic attributes readings. In contrast, the time stamp was relaxed to match *BP* with the attributes data by moving certain number of records forward and backward results of five data sets. The training and testing ratio of the data samples in each group is 4:1, namely 80% of training data and 20% of testing data. The details are listed as follows:

(1) *Group 1*: Extract the exactly matched time stamps for taking *BP* and their characteristic attributes.

(2) *Group 2*: For each *BP* readings, extract 5 characteristic attributes records ahead of the matching time stamps and 5 records behind the matching time stamps.

(3) *Group 3*: For each *BP* readings, extract 10 characteristic attributes records ahead of the matching time stamps and 10 records behind the matching time stamps.

(4) *Group 4*: For each *BP* readings, extract 25 characteristic attributes records ahead of the matching time stamps and 25 records behind the matching time stamps.

(5) *Group 5*: For each *BP* readings, extract 50 characteristic attributes records ahead of the matching time stamps and 50 records behind the matching time stamps.

4.2 Model Parameters Optimization

In order to optimize the generalization of the model, experiments are carried out for different value parameter combinations,i.e. $\alpha \in (1, Q), \beta \in (1, Q)$, and $\gamma \in (1, Q)(Q = 50)$, through **Algorithm 1** and **Algorithm 2**, the optimal parameters of *Ps* and *Pd* prediction models are obtained as shown in **Table 1**.

TABLE 1
Different parameter values corresponding to the optimization parameters of *Ps* and *Pd*

	Group 1			Group 2			Group 3			Group 4			Group 5		
	α	β	γ	α	β	γ	α	β	γ	α	β	γ	α	β	γ
Ps	2	33	1	48	1	7	21	1	1	43	1	1	47	1	1
Pd	4	32	40	19	1	1	47	1	1	35	1	1	29	1	1

Then, CART model is set up according to the optimized parameters, on which the prediction analysis is carried out. **Table 1** shows that as the amount of data increases, the value of β and γ are maintained at 1, but with very different maximum depths of the trees. Take *Group 1* for *Ps* as an example to find the optimal parameters, as shown in **Fig 5** below.

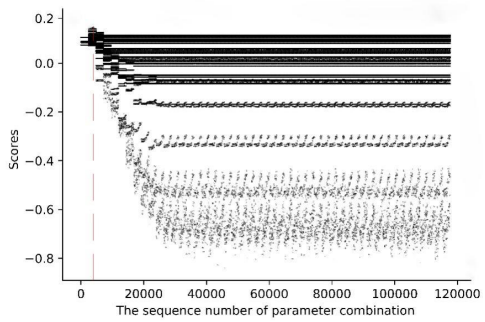

Fig. 5. Search for optimal parameters

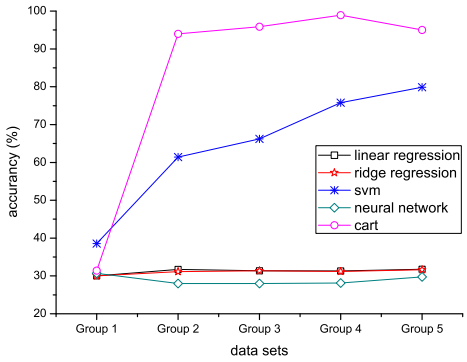

Fig. 6. The *Ps* predication accuracy of each model

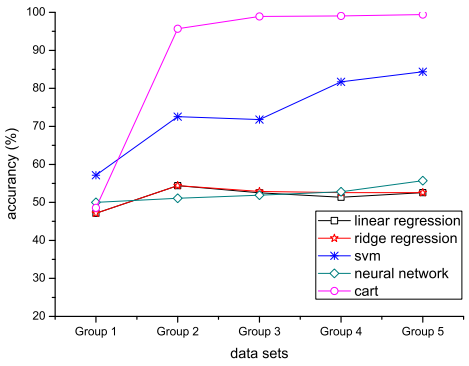

Fig. 7. The *Pd* predication accuracy of each model

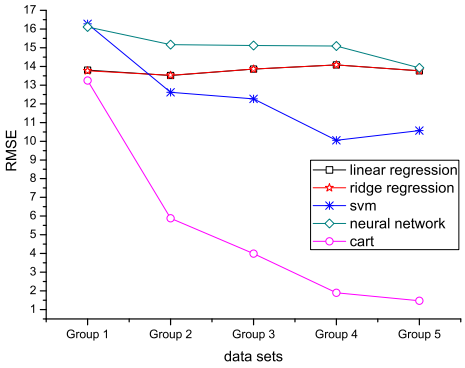

Fig. 8. The model prediction *RMSE* of *Ps*

4.3 Comparative analysis of experimental results

In order to evaluate more accurately and quantitatively the performance of the model, the accuracy rate, *RMSE*, deviation rate and *TIC* are applied accordingly.

1) Accuracy rate

We assess the model in *BP* prediction error range [-5 mmHg, 5 mmHg] according to *ANSI/AAMISP10-1992* standard. If the difference between *BP* prediction and the measured *BP* is less than 5 mmHg, the prediction is deemed to be correct; otherwise incorrect. The accuracy rate ψ is shown below.

$$\psi = X/N \tag{12}$$

where *X* is the count number of *BP* predictions satisfying the error range and *N* is the total *BP* measurement.

Fig 6 and **Fig 7** show the accuracy rate results obtained by CART, SVM, linear regression, neural network and ridge regression models. The *X* axial *Group 1*, *Group 2*, *Group 3*, *Group 4*, *Group 5* denote the results obtained from the five groups of data extracted in different ways mentioned before.

As shown in **Fig 6** and **Fig 7**, when the *Group 2* dataset is more than 5 records before and after marked data, the predictive accuracy of *BP* increased obviously and then reached a stable state. Considered the comparison of these models, the accuracy of ridge regression, linear regression and neural network model remains stable despite of the increment of training data set. The prediction correct rate of *Ps* is around 30%, with *Pd* slightly over 50%. The SVM achieved relatively high accuracy, *Ps* 79.87% and *Pd* 84.37%, when large amount of data was used although its accuracy rate was only slightly better than those of the rest of the three models i.e. 38.57% for *Ps* and 57.14% for *Pd* with using *Group 1* with limited samples. CART model obtained 31.43% for *Ps* and 48.57% for *Pd* using *Group 1* but the prediction accuracy rate reaches as high as 95% for *Ps* and 99.40% for *Pd* in *Group 5*. With appropriate amount of data, CART model proves to have very high *BP* prediction correct rate.

2) *RMSE*

RMSE is the square root of the observation value and the true value of the square of the deviation and the number of observations. The ratio of n. *RMSE* is very sensitive to extreme errors in the data set, e.g. positive infinity errors and negative infinitesimal errors, which can reflect the predictive performance. *RMSE* is expressed below.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{obj,i} - X_{model,i})^2}{n}} \tag{13}$$

where n represents the number of samples, $X_{obj,i}$ the true *BP* for i^{th} sample, and $X_{model,i}$ the i^{th} predicted *BP* by the model. It is shown that the smaller *RMSE* has better performance.

Fig 8 and **Fig 9** illustrate the *BP* prediction in terms of *RMSE* by different models. The figures show that *RMSE* of predicting the *Pd* as a whole by each model is less than the prediction of *Ps*. The reason is that the variance range of human *Ps* is greater than *Pd*. *RMSE* by ridge regression, neural network and linear regression models are all rela-

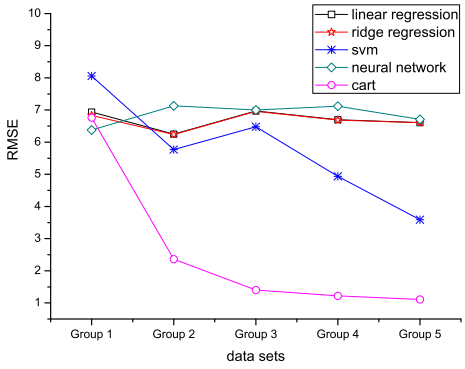

Fig. 9. The model prediction *RMSE* of *Pd*

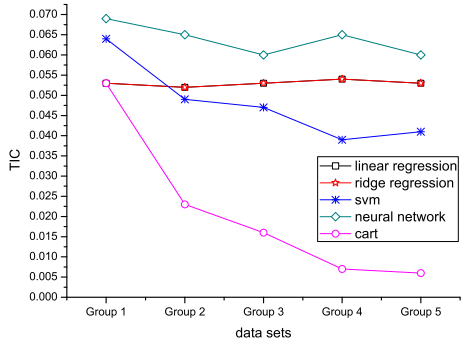

Fig. 10. The *TIC* of *Ps* of prediction models

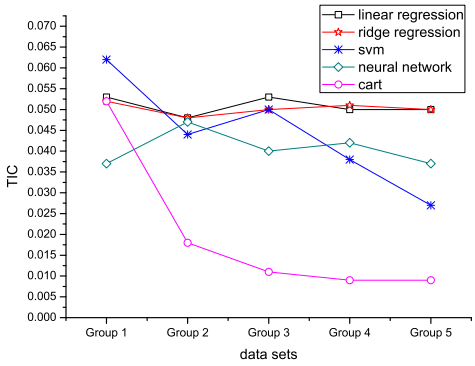

Fig. 11. The *TIC* of *Pd* prediction models

tively large as *RMSE* of the *Ps* is higher than 10, *RMSE* of *Pd* is higher than 6 and it does not decrease significantly with the increase of the training data. Unlike, due to the small amount of data, *RMSE* of CART model is relative large, 13.64 on *Group 1*, but it decreases significantly with the increase of available data. Evaluation of *BP* in *RMSE* verifies that CART model has higher accuracy rate compared to others.

3) Deviation rate ξ reflects the difference between the mean of the predicted values and the mean of the actual values, as shown as follows.

$$\xi = \frac{(K-M)^2}{\frac{1}{n}\sum_{i=T+1}^{T+n}(X_i - Y_i)^2} \tag{14}$$

where *K* represents the mean *BP* predicted by the model, *M* denotes the average of the actual measured *BP*. In the data set, X_i indicates the ith predicted *BP*, Y_i the ith measured *BP* and *n* the number of *BP*. The deviation rates achieved by the models are shown in **Table 2**.

TABLE 2
The deviation rate of prediction models

model	BP	Group1	Group2	Group3	Group4	Group5
ridge model	Ps	0.001	0.002	5.95e-10	4.48e-05	4.37e-05
	Pd	0.031	0.002	0.001	1.69e-05	1.20e-06
neural network	Ps	1.563	4.451	0.428	4.900	4.205
	Pd	34.84	7.963	29.45	28.74	41.35
linear model	Ps	0.003	0.002	6.66e-07	4.40e-05	4.41e-05
	Pd	0.031	0.002	0.001	1.74e-05	1.26e-06
SVM	Ps	0.071	0.0005	0.0007	0.003	0.003
	Pd	0.176	0.061	0.0556	0.044	0.049
CART	Ps	5.59e-05	0.0001	0.006	0.0016	6.09e-05
	Pd	0.049	0.0002	0.0002	0.0002	9.6e-05

As shown in **Table 2**, the deviation rate for *Pd* predicted by neural network model is much higher than the deviation rate calculated by other models. CARTs deviation rate is very small and with the increase of the training data, the deviation rate decreases. The results once again demonstrate that the prediction correct rate of *BP* by CART model outperformed other models.

4) Theil IC (TIC) means the difference between the true value and the predicted value. The numerical results ranged from 0 to 1. The closer *TIC* to 0, the smaller the difference between the predicted value and the true value, and the higher the prediction accuracy. When the result is 0, it means a 100% fitting. *TIC* is expressed below.

$$TIC = \frac{\sqrt{\frac{1}{n}\sum_{i=T+1}^{T+n}(X_i - Y_i)^2}}{\sqrt{\frac{1}{n}\sum_{i=T+1}^{T+n}(X_i)^2} + \sqrt{\frac{1}{n}\sum_{i=T+1}^{T+n}(Y_i)^2}} \tag{15}$$

where, X_i indicates the model predicted *BP* for the ith sample, Y_i the true *BP* of the ith sample and *N* represents the total number of *BP*.

From **Fig 10** and **Fig 11**, it also can be seen that CARTs *TIC* decreases when the amount of data increases. After *Group 2*, *TIC* inequality values are much smaller than those of other models. Regarding CART model in *Group 5*, the *Ps* *TIC* is 0.007, which is very close to 0, and far less than the *TIC* of other models. It is shown that CART model is very suitable for the prediction of *BP*.

4.4 Comparison of model running time

The above analysis mainly focused on the results comparison of different models and validated the prediction accuracy of *BP*. As the real-time prediction is also critical, the analysis of the model execution time starting from the model training and the results were recorded in **Table 3**.

According to results shown in **Table 3**, the training time of neural network and SVM model takes a long time, i.e. 87.39s. *Group 5* took 872.0s of the training time, far greater than CARTs 0.169s. The training time of SVM for *Group 2*

TABLE 3
Compare the running time of each model

Times(s)	Group 1	Group 2	Group 3	Group 4	Group 5
ridge model	0.141	0.265	0.057	0.121	0.248
neural network	9.067	87.39	174.37	436.7	872.0
linear model	0.016	0.032	0.068	0.107	0.200
SVM	0.145	3.243	9.345	41.538	135.7
CART	0.001	0.018	0.037	0.093	0.169

is 3.243s, while the training time for *Group 5* is 135.7s, also much larger than the training time of CART. The respective training time of CART, ridge regression and linear regression model is far less than the neural network and SVM model, meaning the three models have good timeliness, but according to the analysis of the accuracy rate, CART model has an absolute advantage in the timeliness and *BP* prediction accuracy.

4.5 Feature correlation analysis

4.5.1 Correlation feature selection

Given an appropriate amount of data, CART model is able to accurately predict *BP* in real time. However, not all vital signs contribute to the calculation of *BP*. Hence its important to identify the factors with the largest impacts on the prediction so as to improve the efficiency of the models. This paper employed Pearson correlation coefficients for attributes filtering. The Pearson correlation coefficient *r*, also known as product moment correlation (or product moment correlation), is a linear correlation method proposed by British statistician Pearson in 20th century. The calculation method is expressed in equation(16).

$$r = \frac{\sum(X-k)(Y-m)}{\sqrt{\sum(X-k)^2 \sum(Y-m)^2}} \tag{16}$$

where *K* represents the mean expectation of variable *X*, and *m* represents the mean expectation of variable Y. *Pearson* is a value between -1 and 1, when the linear relationship between the two variables is enhanced, the correlation coefficient tends to be 1 or -1; When a variable increases and so does another variable, they are positively correlated, which is denoted by a correlation coefficient greater than 0; if one variable increases whilst another decreases, they are in negative correlation with the correlation coefficient less than 0; and if the correlation coefficient is equal to 0, there is no linear correlation between them. **Fig 12** shows the correlation between the potential correlated attributes and *BP*. As the positive and negative signs only reflect the correlation trends, the paper tries to identify the intensity of the correlation values by taking the calculation process of *Pearson* to the absolute values. By means of Pearson correlation analysis, the correlation coefficients of *Ps* and *Pd* of each characteristic attribute are obtained.

As shown in **Fig 12**, the most correlated coefficient of attributes is the HR with *Ps* 0.3499 and *Pd* 0.2133, followed by the PTT attribute with Pearson coefficient with *Ps* 0.1821 and *Pd* 0.0674, so PTT and HR are selected for further analysis in the remain part of the paper.

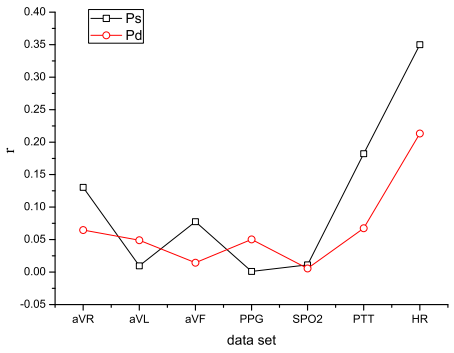

Fig. 12. Pearson correlation coefficients of each eigenvalue

4.5.2 PTT and HR analysis results of CART model

PTT and HR have been used to predict *BP* by many researchers, for example, Rui and He et al. [31] established the random forest model using PTT to predict *BP* and the mean and standard deviation evaluation index to validate it. For *Pd*, the average error (mmHg) is about 4, and the average error of *Ps* is about 9. Wu et al [13], based on attributes (age, gender, body fat and HR), established the neural network model for *BP* prediction and use the accuracy rate index to evaluate the model performance. For male subjects, the *BP* accuracy rate is about 30% ranging (-5 mmHg, +5 mmHg), while the accuracy rate of female *BP* is about 28%. Due to the high correlation of both PTT and HR with *BP*, its not sufficient to use only a single attribute in the prediction of *BP*. The results of experiment of PTT and HR for CART models are shown in **Table 4**.

TABLE 4
CART prediction and evaluation results

Index	Feature	BP	Group1	Group2	Group3	Group4	Group5
ψ (%)	PTT HR	Ps	30.00	94.00	92.79	93.23	94.07
		Pd	48.57	98.57	98.29	98.49	98.23
	all	Ps	31.43	94.00	95.86	98.91	95.00
		Pd	48.57	95.71	98.93	99.06	99.40
τ(s)	PTT HR	-	0.002	0.004	0.007	0.015	0.031
	all	-	0.004	0.018	0.037	0.093	0.169
ω	PTT HR	Ps	0.002	0.004	0.008	0.016	0.033
		Pd	0.002	0.004	0.007	0.015	0.032
	all	Ps	0.004	0.019	0.038	0.094	0.178
		Pd	0.002	0.019	0.037	0.094	0.170

In **Table 4**, the results achieved by using only PTT and HR were compared with those using all the characteristics attributes of CART model in terms of the *BP* prediction accuracy rate ψ and training and prediction time τ. *Group 1* has a small amount of training data, so the accuracy rate is relatively low. For other groups of data, although their training prediction time were significantly shortened with less attributes, the accuracy rate only decreased slightly, still reaching over 94%. By defining the time accuracy ratio evaluation index w, the decrease in time relative to the advantage of accuracy reduction can be well explained. The w can be expresses as follows.

$$\omega = \frac{\tau}{\psi} \tag{17}$$

where τ represents the sum of model training and prediction time measured in seconds. When the value of time is smaller and the accuracy rate is greater, the overall value of w is smaller, i.e. the model is more efficient. According to **Table 4**, the time efficiency is significantly improved by reducing the attributes to *PTT* and *HR* from 0.17 to 0.03. Therefore, the establishment of CART model to predict *BP* from the characteristics of high correlation can greatly improve the time efficiency of the model and achieve real-time and accurate prediction of human *BP*. What's more, the scalability (time complexity) of CART model to predict *BP* is O(N*M*α), where N is the size of sample, M is the number of features, and α is the depth of the tree.

5 Discussion

The CART model proposed in this paper has achieved high accuracy, which has shortened the training time for non-invasive *BP* prediction within the error range (-5 mmHg, +5 mmHg). To avoid the problem of over fitting and improve the generalization of the model, the experiment uses cross-validation to find the optimal parameters of the model. However, it is a time-consuming process. Therefore, it is suggested in this paper to use offline parameter optimizations running on the background when a new pair of attributes and true *BP* measurement is added to the data set. By this background processing, the system takes advantage of new input data, ensuring that the model is optimized. Secondly, the accuracy rate of the *BP* prediction varies over time due to the randomly selected training data set. Based on exhausted experiments, the variance is always below 5% range. Finally, the data were collected within age group of 18 to 22 years old, which may have less accuracy rate for predicting the *BP* for the elderly and people with abnormal *BP* without retraining.

6 Conclusion

This paper has proposed CART model for *BP* prediction based on biological attributes data ECG (AVR, AVL, AVF), PPG, PTT, SPO2 and HR collected from a health monitor *CM400*. To avoid model overfitting, the optimum parameters of the model were calculated by the cross-validation method. To verify the effect of the model, CART model were compared with other classical methods such as linear regression, ridge regression, SVM and neural networks in matrix of accuracy rate, *RMSE*, deviation rate and *TIC*. Pearson correlation coefficient was also applied to selecting the most correlated variables. The experimental results show that the prediction result of CART model outperformed the other four models, with prediction accuracy rate of more than 90% within error range of [-5 mmHg, +5 mmHg]; Finally, by the feature correlation analysis, it is found that PTT and HR are the most related attributes to *BP* prediction. The prediction model using only PTT and HR has significantly reduced the training and predicting time with a relatively identical accuracy rate compared with other models using all the seven available attributes.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. 61772451, No. 61772449, No. 61472341, No. 61572420 and the Natural Science Foundation of Hebei Province P. R. China under Grant No. F2016203330. The authors are grateful to the valuable comments and suggestions of the reviewers.

References

[1] Marik, P. E., and J. Varon. "Hypertensive crises: challenges and management. " Chest 131.6(2007):1949.
[2] Vasan, R. S., et al. "Residual lifetime risk for developing hypertension in middle-aged women and men: The Framingham Heart Study. " Jama the Journal of the American Medical Association 287.8(2002):1003-1010.
[3] Conway, Jlio C. D., et al. "Modelling of the relationship between systolic blood pressure and glucose with the magnesium ion present in the blood plasma: an approach using artificial neural networks." Health1.3(2009):211-219.
[4] Ogedegbe, G, and T. Pickering. "Principles and techniques of blood pressure measurement." Cardiology Clinics 28.4(2010):571-586.
[5] Parati, G, et al. "Non-invasive beat-to-beat blood pressure monitoring: new developments." Blood Pressure Monitoring 8.1(2003):31.
[6] Piper, M. A., et al. "Diagnostic and predictive accuracy of blood pressure screening methods with consideration of rescreening intervals: a systematic review for the U.S. Preventive Services Task Force." Annals of Internal Medicine 162.3(2015):192.
[7] Li, Xiaohan, S. Wu, and L. Wang. "Blood Pressure Prediction via Recurrent Models with Contextual Layer." International Conference on World Wide Web International World Wide Web Conferences Steering Committee, 2017:685-693.
[8] , Mathukumalli Vidyasagar. "An Elementary Derivation of the Large Deviation Rate Function for Finite State Markov Chains." Decision and Control, 2009 Held Jointly with the 2009, Chinese Control Conference. Cdc/ccc 2009. Proceedings of the, IEEE Conference on IEEE, 2014:1599-1606.
[9] Hsieh, Yi Yen, et al. "A linear regression model with dynamic pulse transit time features for noninvasive blood pressure prediction." Biomedical Circuits and Systems Conference IEEE, 2017:604-607.
[10] Sideris, Costas, et al. "Building Continuous Arterial Blood Pressure Prediction Models Using Recurrent Networks." IEEE International Conference on Smart Computing IEEE, 2016:1-5.
[11] Abbasi, Robabeh, M. H. Moradi, and S. F. Molaeezadeh. "Long-term prediction of blood pressure time series using multiple fuzzy functions."Biomedical Engineering IEEE, 2015:124-127.
[12] Li, Xiaohan, S. Wu, and L. Wang. "Blood Pressure Prediction via Recurrent Models with Contextual Layer." International Conference on World Wide Web International World Wide Web Conferences Steering Committee, 2017:685-693.
[13] Wu, Tony Hao, K. H. Pang, and W. Y. Kwong. "Predicting Systolic Blood Pressure Using Machine Learning." International Conference on Information and Automation for Sustainability IEEE, 2015:1-6.
[14] Ghosh, Shrimanti, et al. "Continuous blood pressure prediction from pulse transit time using ECG and PPG signals." Healthcare Innovation Point-Of-Care Technologies Conference IEEE, 2016:188-191.
[15] Munnoch, Robert, and P. Jiang. "A personal medical device for multi-sensor, remote vital signs collection in the elderly." Science and Information Conference IEEE, 2015:1122-1131.
[16] Breiman. Classification and regression trees /. Wadsworth International Group, 1984.
[17] Hastie, T, R. Tibshirani, and J. H. Friedman. "Elements of Statistical Learning." Springer 167.1(2001):192-192.
[18] Golino, Hudson Fernandes, et al. "Predicting Increased Blood Pressure Using Machine Learning." Journal of Obesity 2014.5(2014):637635.
[19] Allen, John. "Photoplethysmography and its application in clinical physiological measurement. " Physiological Measurement 28.3(2007):R1.
[20] Teng, X. F., and Y. T. Zhang. "Continuous and noninvasive estimation of arterial blood pressure using a photoplethysmographic approach."Engineering in Medicine and Biology Society, 2003. Proceedings of the, International Conference of the IEEE IEEE, 2003:3153-3156 Vol.4.

[21] Aoyagi, T, et al. "Multiwavelength pulse oximetry: theory for the future. "Anesthesia & Analgesia 105.6 Suppl(2007):S53.
[22] Poon, C. C. Y., and Y. T. Zhang. "Cuff-less and Noninvasive Measurements of Arterial Blood Pressure by Pulse Transit Time."International Conference of the Engineering in Medicine & Biology Society Conf Proc IEEE Eng Med Biol Soc, 2005:5877.
[23] Jadooei, Ali, O. Zaderykhin, and V. I. Shulgin. "Adaptive algorithm for continuous monitoring of blood pressure using a pulse transit time."Electronics and Nanotechnology IEEE, 2013:297-301.
[24] Ye, S. Y., et al. "Estimation of systolic and diastolic pressure using the pulse transit time." World Academy of Science Engineering & Technology67(2010):726.
[25] Chen, Mr W., et al. "Continuous estimation of systolic blood pressure using the pulse arrival time and intermittent calibration." Medical & Biological Engineering & Computing 38.5(2000):569-574.
[26] Obrist, P. A., et al. "Pulse transit time: relationship to blood pressure and myocardial performance." Behavior Research Methods & Instrumentation16.3(1979):623-626.
[27] Wong, M. Y. M, and Y. T. Zhang. "The Relationship between Pulse Transit Time and Systolic Blood Pressure on Individual Subjects after Exercises." Distributed Diagnosis and Home Healthcare, 2006. D2h2. Transdisciplinary Conference on IEEE, 2006:37-38.
[28] Cheng Y., Jiang P., Peng Y. "Increasing big data front end processing efficiency via locality sensitive Bloom filter for elderly healthcare." Computational Intelligence in Big Data. IEEE, 2015:1-8.
[29] Salman, Mohammad Shukri, O. Kukrer, and A. Hocanin. "Recursive inverse algorithm: Mean-square-error analysis." Digital Signal Processing(2017).
[30] Gujarati, Damodar N. "Basic econometrics." Economics 24.1(1978):20.
[31] He, Rui, et al. "Beat-to-beat ambulatory blood pressure estimation based on random forest." IEEE, International Conference on Wearable and Implantable Body Sensor Networks IEEE, 2016:194-198.