IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 16, 2021, accepted February 23, 2021, date of publication February 24, 2021, date of current version March 4, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3062033*

# Machine Learning Method for Continuous Noninvasive Blood Pressure Detection Based on Random Forest

**XIAOHUI CHEN, SHUYANG YU, YONGFANG ZHANG, FANGFANG CHU, AND BIN SUN**
School of Automation, School of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

Corresponding author: Bin Sun (ffgz366@163.com)

This work was supported in part by the National Science Foundation of China under Grant 61801239.

**ABSTRACT** In order to reduce the influence of differences in human characteristics on the blood pressure prediction model and further improve the accuracy of blood pressure prediction, this paper establishes support vector machine regression model and random forest regression model for accurate blood pressure measurement. First, the photoelectric method is used to obtain the photoelectric plethysmography signal (PPG) and ECG signals from people of different ages, and the blood pressure value is roughly estimated based on the high-quality physiological signals and the vascular elastic cavity model; then the human body characteristics are used as the input parameters of the blood pressure prediction model, and the model parameters are used to find the best parameter combination to improve the prediction performance of the model; finally, through a lot of training and learning, the best blood pressure prediction model is selected to achieve accurate measurement of blood pressure values. It has been verified by experiments that the average absolute error of diastolic and systolic blood pressure based on the random forest optimization model meets the standard of less than 5mmHg formulated by AAMI (American Medical Instrument Promotion Association), which is better consistent with the method of mercury sphygmomanometer, and has more excellent performance than support vector machine regression model under the same conditions.

**INDEX TERMS** Blood pressure detection, random forest, support vector regression, human body characteristics.

## I. INTRODUCTION

Blood pressure is one of the important indicators to measure human health, and its abnormal fluctuation will bring serious harm to human body [1], [2]. In recent years, with the acceleration of people's life rhythm, unreasonable diet structure, irregular work and rest and other adverse phenomena are becoming increasingly serious, which makes cardiovascular disease, hypertension and other blood pressure related diseases more and more common. And in the early stage of these diseases, there are no obvious symptoms outside the body, but the blood pressure has deviated from the normal value at this time, so it is extremely important to strengthen the continuous measurement of blood pressure [3].

At present, wearable physiological monitoring equipment is booming, and its advantages such as simple operation, convenience and comfort, are very suitable for use in the family. As an important indicator of the human body, blood pressure has become its main research direction. As an important index of human body, blood pressure has become the main research direction. Through the photoelectric sensor and electrode on the wearable device, the human body's PPG and ECG signals are obtained, and the relationship between the model and the elastic model of the vascular lumen is established to realize the blood pressure measurement [4]. The method is simple to operate and easy to measure repeatedly, and the tester does not have discomfort during the test [5]. Due to the difference of human body characteristics, different mathematical models should be established for the population, which will inevitably increase the workload and complexity of blood pressure measurement. In recent years, machine learning methods have achieved ideal results in many fields, and some scholars have applied machine learning methods to the medical field [6]–[8]. Wu *et al.* [9] used BP neural network method and radial basis function method to build blood pressure prediction model, which is easy to implement, but has low time complexity and weak generalization ability; Wu [10] established a prediction

The associate editor coordinating the review of this manuscript and approving it for publication was Hiu Yung Wong.

model by using deep neural network and combining human physiological characteristics, and the prediction result of this method is obviously better than that of BP neural network method, but its running speed is slow and it is not suitable for real-time monitoring.

In response to these problems, this paper establishes a support vector machine and a random forest regression model to predict blood pressure to improve the accuracy of blood pressure measurement and the generality of the model. Firstly, the blood pressure value is preliminarily estimated through high-quality physiological signals, and then the human body characteristics are used as the input parameters of the prediction model, and the best parameter combination of the model is selected through genetic algorithm and grid search method. Finally, through a lot of training, learning and experimental comparison, Optimal blood pressure prediction model to minimize the impact of individual characteristics on model accuracy.

## II. SUPPORT VECTOR MACHINE AND RANDOM FOREST REGRESSION MODEL PRINCIPLE

### A. REGRESSION PRINCIPLE OF SUPPORT VECTOR MACHINE

Support vector machine regression model transforms the nonlinear functional relationship in low-dimensional space into high-dimensional space through mapping, establishes the optimal hyperplane, and transforms the nonlinear function relationship into linear function relationship for processing [11]. Suppose the training data set is:

$$T_{(n)} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\} \tag{1}$$

where $x_i$ is the characteristic quantity, $y_i$ is the true value of blood pressure corresponding to the characteristic quantity, $i = (1, 2, \cdots, n)$, $n$ is the number of training sets.

Then the objective function is:

$$f(x) = \omega\varphi(x) + b' \tag{2}$$

where $\varphi(x)$ is a mapping function, $\omega$ is the coefficient of regression function and $b'$ is Error value.

The optimal hyperplane is established by $\omega$ and $b'$. Due to the existence of fitting error, penalty parameters $C$ and insensitive parameters $\varepsilon$ need to be introduced into the model. Then, by using the relationship between Lagrange function and dual function, the following function relations are established.

$$\begin{cases} \max[-\frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{L}(a_i - a_i^*)K(x_i, x_j) \\ -\sum_{i=1}^{L}(a - a_i^*)\varepsilon + \sum_{i=1}^{L}(a_i + a_i^*)y_i] \\ s.t. \begin{cases} \sum_{i=1}^{L}(a - a_i^*) = 0 \\ 0 \le a_i \le C \\ 0 \le a_i^* \le C \end{cases} \end{cases} \tag{3}$$

where $a_i$, $a_i^*$ is Lagrange operator.

The training samples of $a^{(*)} = (a_1, a_1^*, a_2, a_2^*, \cdots, a_L, a_L^*)$, $(a_i - a_i^*) \neq 0$ are support vectors.

The decision regression function is obtained as follows:

$$f(x) = \sum_{i=1}^{L}(a - a_i^*)K(x_i, x_j) + b^* \tag{4}$$

### B. PRINCIPLE OF RANDOM FOREST REGRESSION

Random forest algorithm is an integrated algorithm based on cart decision tree [12]. Usually, the training samples are obtained by self-help sampling method, and the decision tree is formed by splitting the sample eigenvalues. The splitting value can be expressed by entropy, Gini index and variance. The calculation method is as follows:

$$g(D, B) = H(D) - H(D|B) \tag{5}$$

$$Gini(D) = 1 - \sum_{i=1}^{k}\left(\frac{|Q_i|}{|D|}\right) \tag{6}$$

$$V = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2 \tag{7}$$

where $B$ represents a feature in the sample data set, $H(D)$ is the empirical entropy of the set $D$, $Q_i$ is the sample subset of the ith class in the training data set, $y_i$ represents the label of a sample instance, and $\mu$ represents the mean value of the sample instance.

In the random forest regression model, each regression tree corresponds to a partition in the feature space and the output value on the partition unit. Assuming that each feature has $s_i(i \in (1, M))$ values, according to the principle of minimum loss function, when the value of feature makes the loss function minimum, it is a kind of division. The output value of each region is $c_i$, and its expression is

$$\min_{j,s}\left[\min_{c1} Loss(y_i, c_1) + \min_{c_2} Loss(y_i, c_2)\right] \tag{8}$$

Suppose the space is divided into f elements by partition $R_1, R_2, \cdots, R_F$, Then the expression of the regression tree is:

$$f(x) = \sum_{f=1}^{F} c_f I(x \in R_f) \tag{9}$$

Finally, multiple regression trees are formed. Assuming that the set of regression trees is $\{T_1, T_2, \cdots, T_s\}$, when new data is input, each tree will have a prediction value, and the average value of the prediction results of each tree is the final prediction result. The expression is:

$$f(x) = \frac{1}{s}\sum_{i=1}^{s} T_i(x) \tag{10}$$

To sum up, the random forest regression model integrates the prediction of each tree, thus reducing the variance of the model. In addition, the model has strong generalization ability, can tolerate a small number of outliers and missing values, and the training speed is fast.

## III. ESTABLISH SVM AND RF BLOOD PRESSURE PREDICTION MODEL

### A. PRELIMINARY ESTIMATION OF BLOOD PRESSURE

By extracting the optimal optical capacitance pulse wave signal and ECG signal at the same time [13]–[15], combined with auscultation method, the blood pressure prediction equation was constructed by using the pulse wave conduction time, and the blood pressure value was preliminarily estimated.

The calculation formula of systolic $P_s$ and diastolic $P_d$ blood pressure is as follows:

$$\begin{aligned} p_s &= z \cdot PTT + v \\ P_d &= P_s \cdot e^{\frac{-T_d}{m_1 KT + m_2}} \end{aligned} \tag{11}$$

where, $PTT$ is the pulse wave conduction time, $z, v$ is constant, $K$ is the characteristic value of pulse wave, $T$ is the pulse cycle, $T_\mathrm{d}$ is the time of the diastolic period of the descending pulse wave, $m_1$ and $m_2$ are the constant.

### B. DATA NORMALIZATION

Since the individual's gender, age, height, weight, body fat rate and real-time measured heart rate all affect blood pressure, the preliminary rough blood pressure estimated value obtained from the high-quality physiological signal in the previous step is used as the characteristic input parameter of the blood pressure prediction model. The blood pressure measured by the corresponding mercury meter is used as the output parameter of the learning model. However, the order of magnitude and physical significance of these blood pressure influencing factors are different, so normalization is needed.

The normalization formula is as follows:

$$x'_{ij} = \frac{x_{ij} - x_{ij}(Min)}{x_{ij}(Max) - x_{ij}(Min)} \tag{12}$$

where $x_{ij}$ represents a blood pressure influencing factor in a sample.

### C. BP PREDICTION MODEL BASED ON SUPPORT VECTOR MACHINE REGRESSION

Firstly, the blood pressure was estimated by processing the optimal physiological signals at the same time. Then, the individual features are imported and normalized. In order to obtain the best prediction model, 10-fold cross-validation method is used to divide the training set and the validation set. In order to improve the processing speed of the prediction model, the principal component analysis method is introduced to reduce the dimension of input parameters. The data after dimension reduction is used as input parameters, which are imported into support vector machine regression model. The actual value measured by mercury sphygmomanometer at the same time is taken as the real value to construct the mapping relationship between input parameters and real values. In order to further improve the prediction accuracy of the model and avoid the model falling into the state of over fitting or under fitting, this paper uses genetic algorithm to find the optimal model parameters. The flow chart of blood pressure prediction model based on support vector regression is shown in Figure 1.

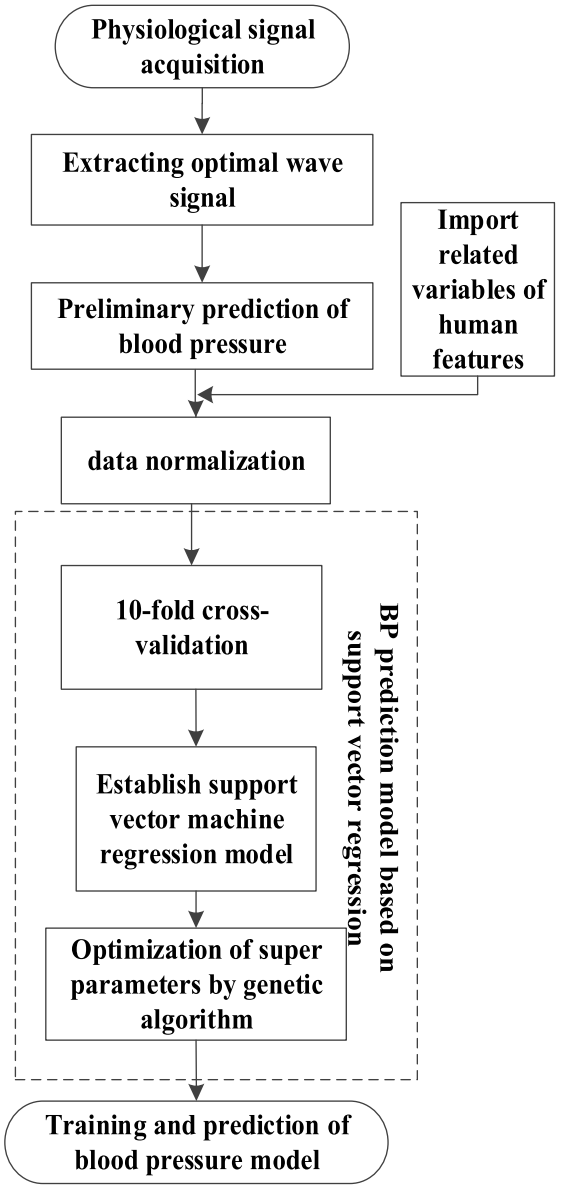



**FIGURE 1.** Flow chart of blood pressure prediction model based on support vector regression.

### D. BLOOD PRESSURE PREDICTION MODEL BASED ON RANDOM FOREST REGRESSION

Taking the normalized blood pressure influence factor as the input parameter, the training data set and the out-of-package data set are divided by the self-sampling method, and the out-of-package data set is used to verify the training effect of the model. According to the analysis and selection of the importance of blood pressure influencing factors, the random forest regression model continuously extracts and divides, and selects the optimal variables according to the values of

entropy, Gini index, and variance, splits the nodes, and forms multiple regression trees. The regression tree averages the predicted blood pressure, and the result is the final predicted blood pressure. The specific process is shown in Figure 2.

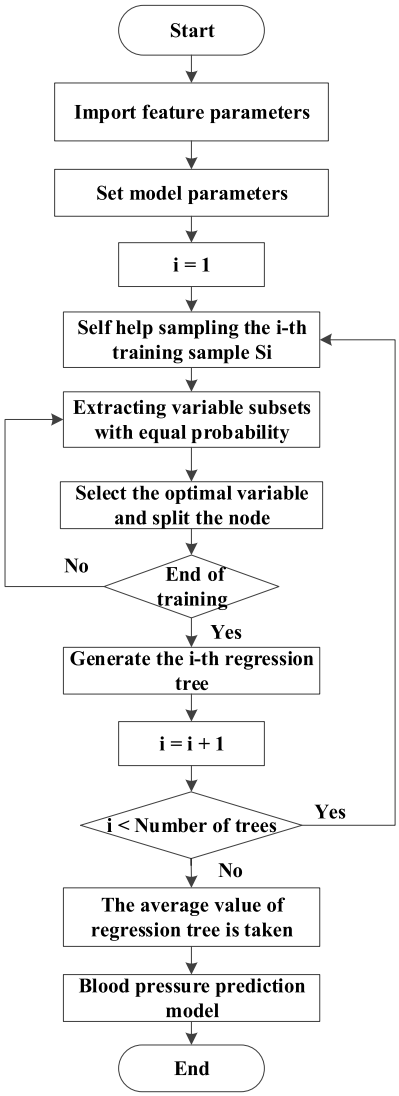



**FIGURE 2.** Flow chart of blood pressure prediction by random forest regression model.

Finally, the choice of model parameters plays a key role in model performance [16]–[18]. In the random forest model, the major influences are the number of decision trees and the number of variables randomly selected when the decision tree splits nodes. This paper uses grid search to achieve model parameter tuning. The main method is to compare model errors, and the parameter combination with the smallest error is the best combination;

The mse calculation formula is:

$$MSE = \frac{1}{N}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{13}$$

Among them, $N$ represents the number of blood pressure record samples in the training set, $y_i$ represents the actual blood pressure measurement value of the i-th sample, and $\hat{y}_i$ represents the blood pressure prediction value obtained by the prediction model for the i-th sample.

## IV. EXPERIMENT ANALYSIS

### A. DATA SET INTRODUCTION

The data of this experiment comes from the database (http://dataset.kangdollar.com/) collected and established by the team. The experimental data was obtained by randomly selecting 1060 groups. Take 1000 groups as the training data set for establishing the blood pressure model, and the remaining 60 groups as the test set for the final blood pressure model. And these two datasets both contain data for people aged 16 to 80, and the data distribution is shown in Figure 3. In addition, there are 512 male samples and 488 females in the training dataset; 32 male samples and 28 female samples in the test set. And the distribution range of blood pressure values is systolic blood pressure: 90-175mmHg, diastolic blood pressure: 57-108mmHg. Basically cover most people, so that the prediction model has strong generalization ability.

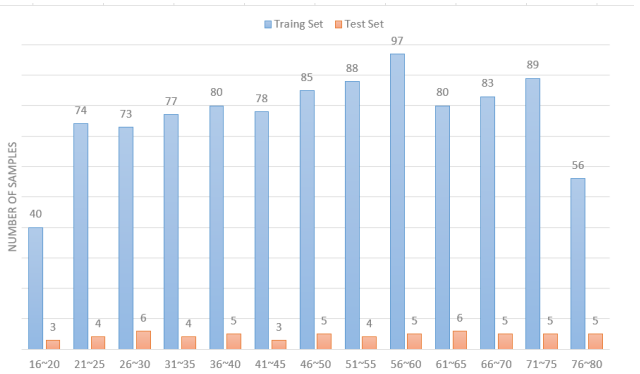
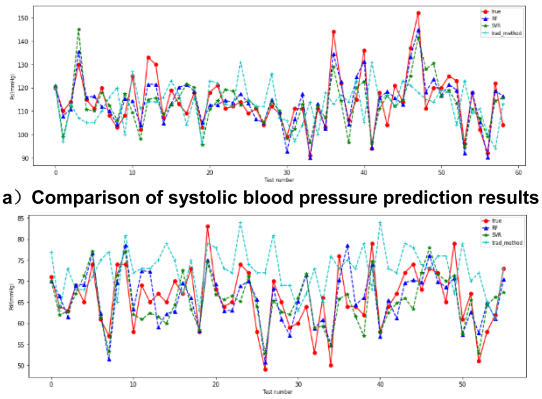



**FIGURE 3.** Data set age distribution map.

### B. ANALYSIS OF EXPERIMENTAL RESULTS

Import 60 groups of test set data into the trained model to test the blood pressure model. The measured value of the mercury sphygmomanometer at the same time is marked as true, the blood pressure predicted value of the support vector machine regression model is marked as SVR, the blood pressure predicted value of the random forest regression model is marked as RF, and the preliminary measurement value of the traditional method is marked as trad_method. The prediction results of the random forest regression model and the support vector machine regression model are shown below.

It can be clearly seen from Figure 4 that the prediction performance of the machine learning method is greatly improved compared with the traditional method. Compared with the support vector machine model, the random forest



(a) Comparison of systolic blood pressure prediction results

(b) Comparison of diastolic blood pressure prediction results

**FIGURE 4.** Comparison of prediction results between random forest regression model and support vector machine regression model.

regression model has a better fitting effect, and the prediction results of systolic and diastolic blood pressure are closer to the real measured values, especially the prediction of systolic blood pressure.

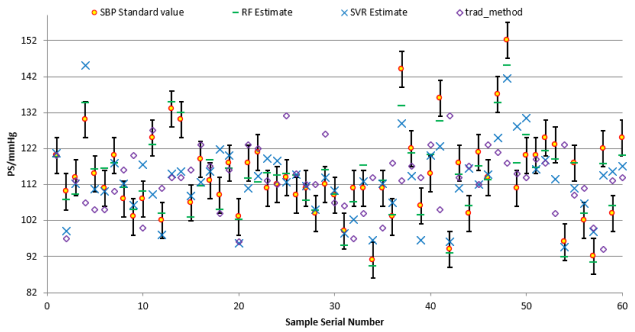



**FIGURE 5.** Systolic pressure error analysis chart.

In order to further analyze the experimental results, the error analysis diagrams shown in Figures 5 and 6 are given, in which the size of the error band is determined by AAMI as ±5mmHg.

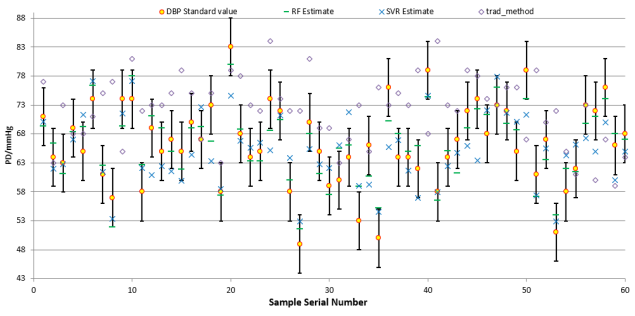



**FIGURE 6.** Diastolic pressure error analysis chart.

In order to evaluate the performance of the two models, systolic blood pressure ($P_s$) and diastolic blood pressure ($P_d$) were compared from the mean absolute error $MAE$, determination coefficient $R^2$ and accuracy rate $AR$. The calculation formula is as follows:

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|h_i - \hat{h}_i\right| \tag{14}$$

$$R^2 = \frac{\sum_{i=1}^{m}\left(\hat{h}_i - \bar{h}\right)^2}{\sum_{i=1}^{m}\left(h_i - \bar{h}\right)^2} \tag{15}$$

$$AR = \frac{\text{Number of qualified samples}}{\text{Total samples}} \times 100\% \tag{16}$$

Among them, $h_i$ is the true blood pressure value of the i-th sample, $\hat{h}_i$ is the predicted blood pressure value of the i-th sample, and $\bar{h}$ is the average value of the true blood pressure of the m samples. The number of samples that meet the error tolerance conditions in AR is obtained from Figures 5 and 6. The smaller the value of $MAE$, the higher the accuracy of the model. $R^2$ is used to weigh the fit of the model. The closer its value is to 1, the higher its reference value. The larger the $AR$ value, the more accurate the prediction of the model.

Through the analysis of the experimental results, the data information of the relevant evaluation indicators as shown in Table 1 is obtained.

**TABLE 1.** Comparison of model evaluation index results.

| BP | Model | MAE | $R^2$ | AR |
|---|---|---|---|---|
| $P_s$ | RF | 4.45 | 0.92 | 96% |
| | SVR | 6.63 | 0.83 | 80% |
| | trad_method | 10.81 | 0.69 | 28% |
| $P_d$ | RF | 3.95 | 0.93 | 93% |
| | SVR | 4.19 | 0.85 | 83% |
| | trad_method | 8.54 | 0.65 | 41% |

It can be seen from Table 1 that the values of the random forest regression model are 0.77 and 0.57, which are closer to 1. $MAE$ meets the requirement of less than 5mmHg, and is significantly smaller than the $MAE$ of the support vector machine regression model, indicating the function constructed by the random forest regression model The relationship is more reliable, the prediction accuracy is higher, and the stability is better. Compared with the traditional model based on $PTT$ blood pressure prediction, the prediction accuracy of the random forest regression model for systolic blood pressure is increased by 68%, and the diastolic blood pressure is increased by 52%.

To test the feasibility of using random forest to predict blood pressure. Using the Bland-Altman method, the consistency between the method proposed in this paper and the blood pressure measured by the mercury meter was analyzed and tested. The error analysis of the above 60 groups of predicted systolic and diastolic blood pressure values is given, and the analysis diagrams shown in Figures 7 and 8 are given. Where is the mean of the blood pressure difference obtained by this method, and is the standard deviation of the mean of the blood pressure difference obtained by the two methods.

It can be seen from figures 7 and 8 that 95% of the difference between the blood pressure value obtained by random forest measurement method and the blood pressure

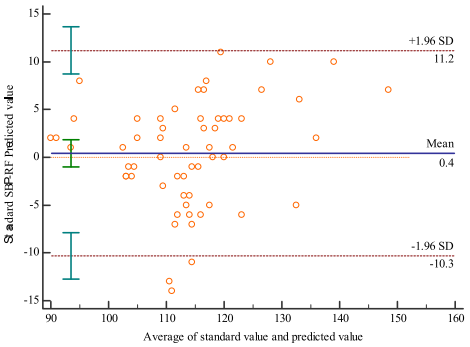



**FIGURE 7.** RF prediction systolic blood pressure Bland-Altman analysis chart.

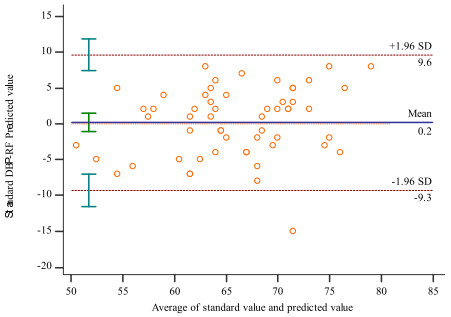



**FIGURE 8.** RF prediction diastolic blood pressure Bland-Altman analysis chart.

value measured by mercury meter is between the $Mean \pm 1.96 \times SD$ consistency limit, and The mean values are 0.4 and 0.2 respectively, which are close to 0, so the random forest method and the traditional mercury sphygmomanometer measurement method have higher consistency, that is to say, the random forest noninvasive blood pressure measurement method proposed in this paper is feasible.

## V. CONCLUSION

In order to reduce the impact of individual characteristic differences on the blood pressure prediction model, enhance the versatility of the model, and improve the accuracy of the model, this article will use the blood pressure estimate derived from the relationship between the physiological signal and the vascular elastic cavity model to construct the blood pressure influence factor in combination with the human body parameters, As the input parameters of the prediction model, and then establish the support vector machine regression model and the random forest regression model to deeply explore the relationship between the influence factors to realize the blood pressure prediction, and finally use the genetic algorithm and the grid search method to optimize the model. Choose the best combination of parameters. After a lot of model training and experimental comparison, it is concluded that the average absolute error of the random forest regression model meets the requirement of less than 5mmHg, and has higher consistency with the mercury sphygmomanometer measurement method, which is significantly better than the support vector machine regression under the same conditions model.

## REFERENCES

[1] J. A. Dias, S. Murali, D. Atienza, and F. Rincon, "Estimation of blood pressure and pulse transit time using your smartphone," in *Proc. IEEE Digit. Syst. Design*, Aug. 2015, pp. 173–180, doi: 10.1109/DSD.2015.90.
[2] S. Yusuf, J. Bosch, and G. Dagenais, "Cholesterol lowering in intermediate-risk persons without cardiovascular disease," *J. Vascular Surg.*, vol. 64, no. 3, p. 827, Sep. 2016.
[3] G. L. L. M. Coelho, "Worldwide trends in blood pressure from 1975 to 2015: A pooled analysis of 1479 population-based measurement studies with 19· 1 million participants," *Lancet.*, vol. 389, no. 10064, pp. 37–55, Jan. 2017.
[4] S. S. Thomas, V. Nathan, C. Zong, K. Soundarapandian, X. Shi, and R. Jafari, "BioWatch: A noninvasive wrist-based blood pressure monitor that incorporates training techniques for posture and subject variability," *IEEE J. Biomed. Health Inform.*, vol. 20, no. 5, pp. 1291–1300, Sep. 2016.
[5] J. Liu, Y. Li, X.-R. Ding, W.-X. Dai, and Y.-T. Zhang, "Effects of cuff inflation and deflation on pulse transit time measured from ECG and multi-wavelength PPG," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 5973–5976.
[6] X. Y. Zheng, "Research on cardiovascular disease prediction system based on machine learning," M.S. thesis, Dept. Inf. Sys. Eng., Beijing JiaoTong Univ., Beijing, China, 2018.
[7] H. W. Wen, Q. Q. Lu, and H. G. He, "Application of machine learning in diagnosis and predcition of neuropsychiatric diseases," *Med. J. Peking Union Med. College Hospital*, vol. 9, no. 1, pp. 19–24, Feb. 2018.
[8] C. Sideris, H. Kalantarian, E. Nemati, and M. Sarrafzadeh, "Building continuous arterial blood pressure prediction models using recurrent networks," in *Proc. IEEE Int. Conf. Smart Comput. (SMARTCOMP)*, May 2016, pp. 1–5.
[9] T. H. Wu, G. K.-H. Pang, and E. W.-Y. Kwong, "Predicting systolic blood pressure using machine learning," in *Proc. 7th Int. Conf. Inf. Autom. Sustainability*, Dec. 2014, pp. 1–6.
[10] D. Wu, "Research on continuous non-invasive blood pressure detection and its application based on deep neural network," Ph.D. dissertation, Dept. Adv. Technol., Univ. Chin. Acad. Sci., Shenzhen, China, 2017.
[11] Z. Yongfang, C. Xiaohui, and Z. Yongsheng, "Non-invasive real-time blood pressure prediction method based on machine learning," in *Proc. 5th IEEE Int. Conf. Cloud Comput. Intell. Syst. (CCIS)*, Nov. 2018, pp. 23–25.
[12] S. Janitza, G. Tutz, and A.-L. Boulesteix, "Random forest for ordinal responses: Prediction and variable selection," *Comput. Statist. Data Anal.*, vol. 96, pp. 57–73, Apr. 2016.
[13] B. Sun, C. Wang, J. F. Chen, Y. F. Zhang, and X. H. Chen, "Algorithm for extracting high quality signals from PPG sequence with motion artifact," *Chin. J. Sci. Inst.*, vol. 39, no. 9, pp. 57–73, Sep. 2018.
[14] B. Sun, C. Wang, X. Chen, Y. Zhang, and H. Shao, "PPG signal motion artifacts correction algorithm based on feature estimation," *Optik*, vol. 176, pp. 337–349, Jan. 2019.
[15] S. Hanyu and C. Xiaohui, "Motion artifact detection and reduction in PPG signals based on statistics analysis," in *Proc. 29th Chin. Control Decis. Conf. (CCDC)*, May 2017, pp. 28–30.
[16] L. Ma, "Research on optimization and improvement of random forests algorithm," M.S. thesis, Dept. Appl. Math., Jinan Univ., Guangdong, China, 2016.
[17] X. X. Cheng, "Research on particle swarm optimization weighted random forest algorithm," M.S. thesis, Dept. Elect. Eng., Zhengzhou Univ., Zhengzhou, China, 2017.
[18] B. W. Wen, W. Dong, and W. Xie, "Parameter optimization method for random forest based on improved grid search algorithm," *Comput. Eng. Appl.*, vol. 54, no. 905, pp. 159–162, May 2018.

**XIAOHUI CHEN** received the M.Sc. degree from Harbin Engineering University, in 1991, and the Ph.D. degree from Southeast University, in 2003.

He is currently a Professor with the Nanjing University of Posts and Telecommunications. His main research interests include networked measurement, control systems, embedded systems, intelligent instruments, sensor networks, and information fusion.

**SHUYANG YU** received the B.Sc. degree form Nanjing Tech University Pujiang Institute, in 2018. He is currently pursuing the master's degree with the Nanjing University of Posts and Telecommunications.

His main research interests include networked measurement and control technology.

**FANGFANG CHU** was born in 1996. She is currently pursuing the degree in instrument science and technology with the School of Automation, Nanjing University of Posts and Telecommunications.

Her main research interests include networked measurement and control technology.

**YONGFANG ZHANG** was born in 1992. She received the degree in instrument science and technology from the School of Automation, Nanjing University of Posts and Telecommunications.

Her main research interests include networked measurement and control technology.

**BIN SUN** received the B.Sc. degree from the Jiangsu University of Science and Technology, in 2008, and the Ph.D. degree from the Nanjing University of Science and Technology, in 2015.

He is currently a Lecturer with the Nanjing University of Posts and Telecommunications. His main research interests include photoelectric signal processing and image information fusion.