Iran J Public Health, Vol. 51, No.9, Sep 2022, pp.2099-2107

Original Article

Logistic Regression Model Based on Ultrafast Pulse Wave Velocity and Different Feature Selection Methods to Predict the Risk of Hypertension

Xue Bai [1], **Wenjun Liu* [1], *Hui Huang* [2], *Huan You* [1]

1. *School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing 210044, China*
2. *Department of Ultrasound, Affiliated Hospital of Nanjing University of CM, Nanjing 210029, China*

***Corresponding Author:** Email: wjliu@nuist.edu.cn

(Received 24 Apr 2021; accepted 12 Jul 2021)

Abstract

Background: Hypertension is the main reason why the incidence of cardiovascular disease has increased year-by-year and early diagnosis of hypertension is necessary to reducing the incidence of cardiovascular disease. This also puts forward higher requirements for the accuracy of diagnosis. We tried a variety of feature selection methods to improve the accuracy of logistic regression (LR).

Methods: We collected 397 samples from Nanjing, Jiangsu, China between Jan 2016 and Dec 2017, including 178 hypertension samples and 219 control samples. It includes not only clinical and laboratory data, but also imaging data. We focused on the difference of imaging attributes between the control group and the hypertension group, and analyzed the correlation coefficients of all attributes. In order to establish the optimal LR model, this study tried three different feature selection methods, including statistical analysis, random forest (RF) and extreme gradient boosting (XGBoost). The area under the ROC curve (AUC) and accuracy were used as the main criterion for model evaluation.

Results: In the prediction of hypertension, the performance of LR with RF as the feature selection method (accuracy: 0.910; AUC: 0.924) was better than the performance of LR with XGBoost as the feature selection method (accuracy: 0.897; AUC: 0.915) and the performance of LR with statistical analysis as the feature selection method (accuracy: 0.872; AUC: 0.926).

Conclusion: LR with RF as the feature selection method may provide accurate results in predicting hypertension. Carotid intima-media thickness (cIMT) and pulse wave velocity at the end of systole (ESPWV) are two key imaging indicators in the prediction of hypertension.

Keywords: Hypertension; Ultrafast pulse wave velocity; Feature selection; Logistic regression

Introduction

The incidence of cardiovascular disease is increasing year by year, mainly due to arteriosclerosis and hypertension, and cardiovascular disease can be diagnosed and prevented in advance (1, 2). In the early stage of onset, although the patient has not yet felt the symptoms, the elasticity of the blood vessels has changed. Pulse wave velocity (PWV) is a classic indicator reflecting arterial elasticity and it can be used to evaluate arterial wall dilatability and stiffness, and early detection of cardiovascular disease (3).

The increase in PWV value is closely related to the prevalence of cardiovascular patients, not only including patients with hypertension (4), but also including patients with diabetes (5) and chronic kidney disease (6). In assessing the pulse wave propagation time, the traditional PWV measurement method still has great limitations (7, 8). Ultrafast pulse wave velocity (ufPWV) is the latest arterial elasticity non-invasive detection technology, which can directly measure the PWV of local blood vessels and reflect the changes in the stiffness of the blood vessel wall (9-11). Compared with the traditional measurement method, ufPWV greatly improves the sampling efficiency (12).

In recent years, ufPWV has played an important role in assessing atherosclerosis (13, 14), arterial stiffening in healthy people and vascular Ehlers-Danlos syndrome (vEDS) (15), and changes on coronary slow flow (16). However, there are few studies on the role of ufPWV in hypertension research, especially in the early prevention of hypertension. This study established a prediction model for hypertension based on ufPWV.

Logistic regression (LR) analysis is a generalized linear regression analysis model, which is often used to explore the risk factors of disease and predict the probability of disease occurrence based on the risk factors. Besides, LR is a traditional classification model in medical research, but sometimes its classification accuracy is not high, related to the selection of features. In our study, we tried a variety of feature selection methods to improve the accuracy of LR.

Methods

Study population and data preprocessing

We collected the dataset from the Department of Cardiology of Affiliated Hospital of Nanjing University of Chinese Medicine (Nanjing, Jiangsu, China) between Jan 2016 and Dec 2017. The dataset includes two parts, the hypertension group (including 178 samples) and the control group (including 219 samples). The inclusion criteria for the hypertension group were the following: (I) At least 3 outpatient measurements of systolic blood pressure (SBP) ≥140 mmHg and (or) diastolic blood pressure (DBP) ≥90 mmHg; (Ⅱ) Ambulatory blood pressure monitoring mean blood pressure during the day≥135/85 mmHg or 24h average blood pressure ≥130/80 mmHg.

We collected clinical and laboratory findings of each sample, besides imaging findings. The laboratory results are collected by three experienced doctors and then averaged. The Aixplorer cIMT measure system was used to measure cIMT, and a 2 to 10 MHz linear array transducer SL10-2 (Aixplorer; Supersonic Imagine, Aix-en-Provence, France) was used to measure ufPWV.

Data preprocessing mainly includes the processing of missing data and abnormal data. For missing data, the attribute removal standard was the amount of missing data for this attribute was greater than 15%. If the amount of missing data was less than 15%, the attribute was retained and linear interpolation was used to deal with missing values. Outlier analysis is to check whether the data has input errors or whether it contains unreasonable data. The identification of outliers in this paper was based on the principle of Box Plot Diagram. The outliers were converted into missing values, and then linear interpolation was applied to deal with the missing values.

Logistic regression model and feature selection methods

Logistic regression (LR) is an algorithm for classification, suitable for research where the dependent variable is a categorical variable. LR can statistically estimate the magnitude of the numerical influence of each independent variable on the probability of the dependent variable taking a certain value when other independent variables are fixed. For binary dependent variables, $y = 1$ indicates that the event occurred, $y = 0$ indicates that the event does not occur. The relationship between the conditional probability of event occurrence $P(y = 1|X)$ and x_i is non-linear, and it usually monotonous. Logistic is based on sigmoid function. The sigmoid function value range is between (0, 1), and it is a gradual process, which is suitable for describing probability $P(y = 1|X)$. After the probability value is obtained, it is judged according to the threshold value. Let $P(y = 1|X) = p$, then the LR model represented by the connection function is:

$$log\,it(y) = In\left(\frac{p}{1-p}\right) = \beta X \qquad [1]$$

the probability of an event not occurring is $P(y = 0|X)$=1-p, and the ratio of the probability of an event occurring to the probability of not occurring is $\frac{p}{1-p}$. Two-category LR uses cross-entropy loss:

$$J(\beta) = -\frac{1}{n}\sum_{n=1}^{n}\left[y^{(i)}In\hat{p_i} + (1 - y^{(i)})In(1-\hat{p_i})\right] \qquad [2]$$

where $y^{(i)}$ is the true category of the sample, $\hat{p_i} = P(y = 1|X)$ is the probability that the prediction is a positive example. After obtaining the predicted probability, the threshold is usually 0.5 to predict the dependent variable as positive or negative. This is only suitable for balanced classification problems. For the problem of unbalanced classification, taking 0.5 as the threshold is often not the optimal choice. It is necessary to select the optimal threshold according to specific needs.

Figure 1 shows the process of predicting hypertension based on the LR model. First, the important features in the process of identifying hypertension are selected based on the existing data. Then we use these characteristics to build a LR model and this LR model can be used to determine whether a new sample is at risk of hypertension.

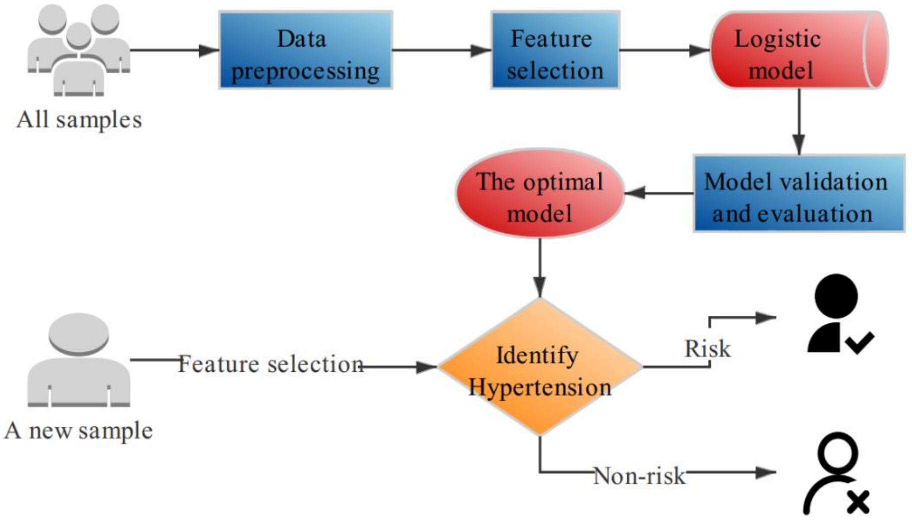

Fig. 1: The flow chart for predicting hypertension based on LR

We chose three methods (statistical analysis, RF and XGBoost) to select important features for modeling. The purpose of our research is to predict whether the sample is at risk of hypertension. Based on this, we can use statistical analysis to select features that have significant differences between the hypertension group and the control group for modeling. Besides, RF is an ensemble learning method based on bagging idea and decision trees, and XGBoost is a representative algorithm based on the boosting idea (17,18). Both RF and XGBoost can be used to deal with regression and classification problems, and can be used to calculate feature importance scores.

Model evaluation

We selected five evaluation indicators (accuracy, precision, sensitivity, specificity and AUC) to compare the prediction effects of each model. Their calculation formula are as follows:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad [3]$$

$$precision = \frac{TP}{TP+FP} \quad [4]$$

$$sensitivity = \frac{TP}{TP+FN} \quad [5]$$

$$specificity = \frac{TN}{TN+FP} \quad [6]$$

where TP represents the number of people with hypertension predicted to hypertension patients, TN represents the number of people without hypertension predicted to non-hypertensive patients, FP represents the number of people without hypertension predicted to hypertension patients and FN represents the number of people with hypertension predicted to non-hypertensive patients. The ROC curve is a curve that compares the ratio of TP to the ratio of FP at different classification thresholds. The closer the ROC curve is to the upper left corner, the better the classification effect. The area under the ROC curve (AUC) can quantify this judgment, that is, the larger the AUC, the better the classification effect.

Statistical analysis

This study used R4.0.3 for statistical analysis. Continuous variables were recorded using mean ± standard deviation, while categorical variables were recorded using count and proportion. Besides, Student's *t*-test and Chi-square test are used to evaluate continuous variables and categorical variables, respectively. Statistical significance was defined as: *P*<0.05.

Ethical approval

The study has been approved by the ethics committee of the Affiliated Hospital of Nanjing University of Chinese Medicine (Protocol No. 2016NL-018-03). Informed consent has been obtained from all the participants. Written informed consent for publication has been obtained from all the patients.

Results

Characteristics of the patients

This study enrolled 397 patients with 794 carotids. Among these, there were 178 patients in the hypertension group and 219 patients in the control group. Our research was from the perspective of the patients and the carotids. At the patient level, we not only discussed the imaging attributes, but also the clinical and laboratory attributes. At the carotid level, we mainly discussed the imaging attributes, which are cIMT and ufPWV. Table 1 showed the results of a single factor analysis of the clinical and laboratory attributes of patients. Sex, smoking and drinking (all *P*>0.05) have no significant effect on the risk of hypertension, while age, weight and BMI are significant influencing factors (all *P*<0.001). In addition, most laboratory attributes showed significant differences between the hypertension group and the control group. Table 2 showed the results of univariate analysis of imaging attributes at the patient level and the carotid artery level, and cIMT and ESPWV are significantly different between the hypertension group and the control group, while BSPWV does not.

Description of the attributes

We explored the correlation between any two variables. The correlation bubble chart cannot only distinguish positive correlation from negative correlation by color, but also display the size of the correlation coefficient through the size of the bubble. Therefore, the method of drawing a bubble chart of the variables was used to explore the correlation between two variables. The bubble chart of the variables was shown in Fig. 2. There is a significant positive correlation between age and cIMT (or ESPWV), that is, the older the age, the greater the value of cIMT and ESPWV. Besides, LDL and TC also had an obvious positive correlation, while the correlation between other variables was not obvious.

Table 1: Summary of clinical and laboratory attributes

Patients	***Total***	***Hypertension group***	***Control group***	***P-value***
No. of patients	397	178	219	-
No. of carotids	794	356	438	-
Sex				
Male (%)	198 (49.87)	95 (53.37)	103 (47.03)	0.248
Female (%)	199 (50.13)	83 (46.63)	116 (52.97)	
Smoking				
Yes (%)	83 (20.91)	38 (21.35)	45 (20.55)	0.943
No (%)	314 (79.09)	140 (78.65)	174 (79.45)	
Drinking				
Yes (%)	65 (16.37)	34 (19.10)	31 (14.16)	0.235
No (%)	332 (83.63)	144 (80.90)	188 (85.84)	
Duration (years)	-	10.18	-	<0.001
Age (years)	55.841 ± 14.702	61.169 ± 12.837	51.511 ± 14.725	<0.001
Weight (kg)	65.146 ± 10.513	68.253 ± 10.792	62.621 ± 9.587	<0.001
BMI (kg/m^2)	23.784 ± 3.144	25.046 ± 3.220	22.758 ± 2.678	<0.001
Laboratory findings				
TG (mmol/L)	1.276 ± 0.632	1.500 ± 0.635	1.093 ± 0.569	<0.001
TC (mmol/L)	4.502 ± 0.927	4.284 ± 1.029	4.679 ± 0.794	<0.001
Glu (mmol/L)	4.925 ± 0.609	4.906 ± 0.682	4.940 ± 0.543	0.596
Cr (μmol/L)	72.824 ± 17.280	77.757 ± 18.553	68.815 ± 15.063	<0.001
Urea (mmol/L)	5.496 ± 1.583	5.949 ± 1.702	5.128 ± 1.376	<0.001
UA (μmol/L)	317.62 ± 92.60	347.53 ± 91.87	293.31 ± 86.02	<0.001
HDL (mmol/l)	1.366 ± 0.364	1.241 ± 0.351	1.467 ± 0.342	<0.001
LDL (mmol/l)	2.454± 0.646	2.382 ± 0.719	2.512 ± 0.576	0.051
Blood pressure (mmHg)				
SBP	129.37 ± 19.38	136.78 ± 18.48	123.35 ± 17.99	<0.001
DBP	74.254 ± 11.013	75.725 ± 12.072	73.059 ± 9.939	0.019

Table 2: Summary of imaging attributes.

Performance level	***Values***	***Total***	***Hypertension group***	***Control group***	***P-value***
Per patient	cIMT	0.057 ± 0.011	0.061 ± 0.010	0.053 ± 0.010	<0.001
	BSPWV	6.186 ± 1.186	6.147 ± 1.197	6.217 ± 1.179	0.563
	ESPWV	8.592 ± 2.133	9.354 ± 1.814	7.973 ± 2.176	<0.001
Per carotid	cIMT	0.056 ± 0.012	0.060 ± 0.011	0.053 ± 0.011	<0.001
	BSPWV	6.172 ± 1.466	6.139 ± 1.524	6.198 ± 1.419	0.575
	ESPWV	8.602 ± 2.445	9.377 ± 2.423	7.973 ± 2.424	<0.001

Figure 3 showed the boxplot of imaging attributes. The cIMT and ESPWV of hypertensive patients were much greater than those of the control group, but the value of BSPWV was not much different between the hypertension group and the control group.

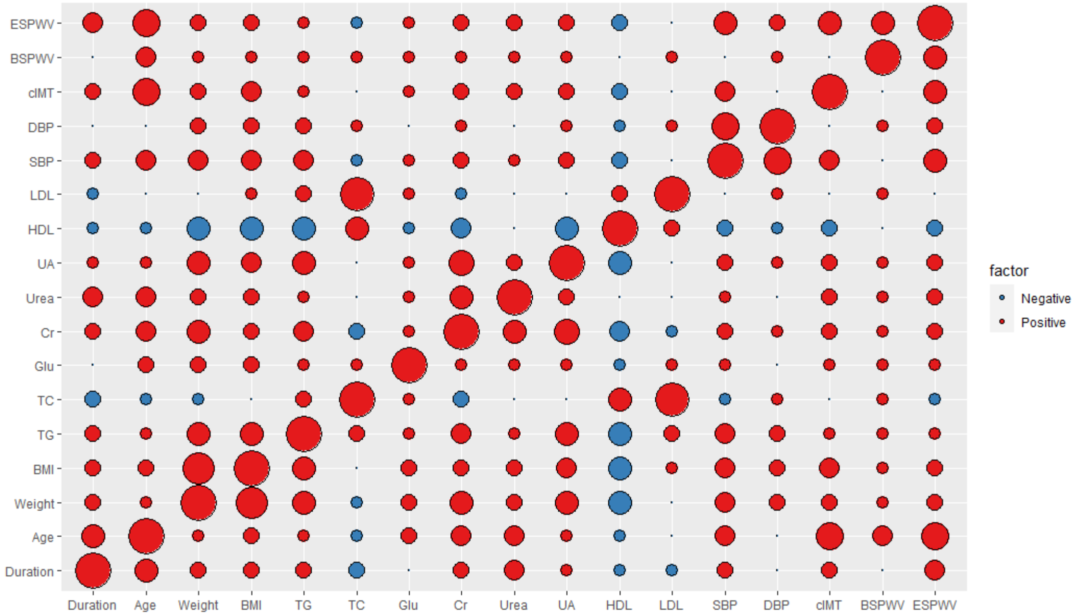

Fig. 2: The correlation bubble chart

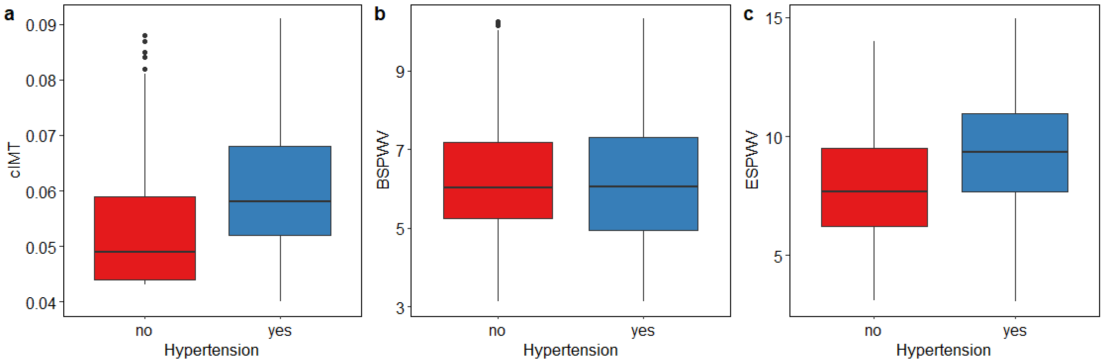

Fig. 3: The boxplot of imaging attributes

Feature selection and model evaluation

We selected three feature selection methods: statistical analysis, RF and XGBoost. Statistical analysis to select features is mainly based on the results in Table 1, that is, the significant differences between the hypertension group and the control group were selected for modeling. The feature selection results of RF and XGBoost were shown in Figs. 4 and 5, and the importance scores of the features were ranked from high to low. Whether it is RF or XGBoost, the top five features are duration, BMI, cIMT, SBP and ES-PWV.

Table 3 showed the model evaluation results based on different feature selection methods. The performance of LR with RF as the feature selection method was better than the performance of LR with XGBoost as the feature selection method and the performance of LR with statistical analysis as the feature selection method.

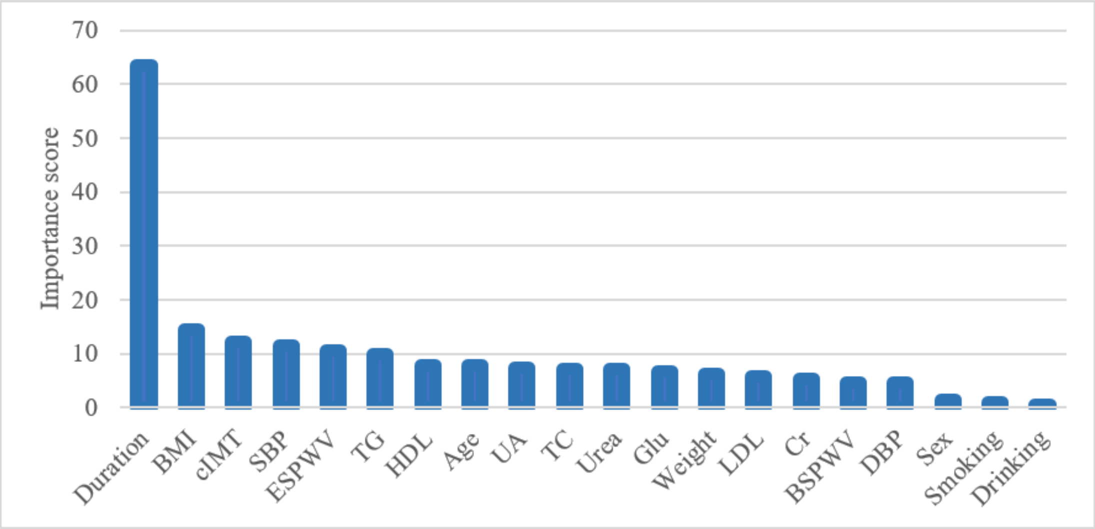

Fig. 4: Importance score ranking based on RF

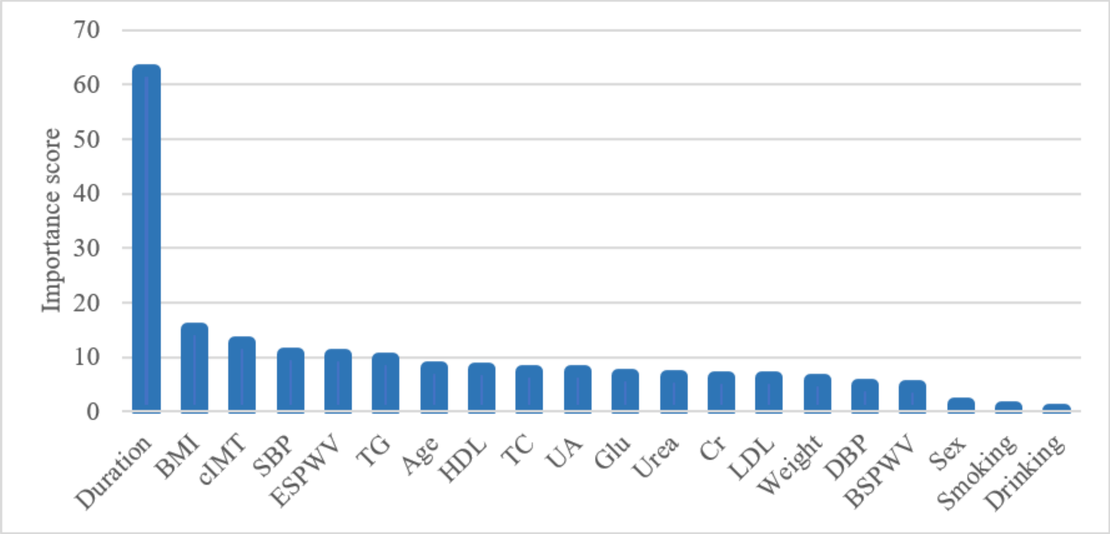

Fig. 5: Importance score ranking based on XGBoost

Table 3: Model evaluation based on different feature selection methods

Feature selection methods	*Accuracy*	*Precision*	*Sensitivity*	*Specificity*	*AUC*
Statistic analyse	0.872	0.851	0.930	0.800	0.926
RF	0.910	0.875	0.977	0.829	0.924
XGBoost	0.897	0.857	0.977	0.800	0.915

Discussion

The purpose of our study was developing a hypertension prediction model based on ufPWV and RF. There were 397 patients with 794 carotids participated in our study, including 178 hypertension patients and 219 non-hypertension patients. Three different feature selection models (statistic analyse, RF, and XGBoost) were used to build LR risk prediction model.

ufPWV is a more accurate PWV measurement method, widely used in cardiovascular disease research, but few studies use ufPWV to assess hypertension. Atherosclerosis and hypertension are two most important components of cardiovascular disease. The difference between ESPWV and BSPWV was quantitatively evaluated in the atherosclerotic population in a recent study (14). In this study, ESPWV is significantly different between the atherosclerosis risk group and the control group, while BSPWV is not significantly different. This may be caused by ESPWV being more accurate than BSPWV in measurement technology (19). In our research, we compared and analyzed ESPWV and BSPWV from three perspectives. Firstly, in the univariate analysis in Table 2, ESPWV was statistically different between the hypertension group and the control group, but BSPWV did not. Secondly, in the boxplot of imaging attributes, the ESPWV of the hypertension group is much greater than that of the control group, but there is no significant change in BSPWV. Finally, in the feature importance scores calculated by RF and XGBoost, the score of ESPWV is much greater than that of BSPWV.

In the hypertension risk prediction model, based on three different feature selection methods, we had obtained three different LR prediction results. Among them, the LR model based on RF had the best prediction effect, followed by XGBoost, and finally by statistical analysis. However, for different data sets, the performance of feature selection methods will be different, so we should evaluate multiple feature selection methods and choose the most appropriate one. In addition, this paper selected five evaluation indicators (accuracy, precision, sensitivity, specificity and AUC) of the two-classification model, which are suitable for the evaluation of most two-classification problems.

Limitations

Generally, our study had two main limitations. Firstly, the size of our samples was small. Secondly, attributes with more than 15% missing data were removed and this may bring some bias. However, this study might provide guiding significances for clinical diagnosis. At first, we compared the impact of different feature selection methods on the LR model, and significantly improved the accuracy of the prediction. The model can be applied to reality. In addition, our feature selection method gave a ranking of the importance of features, which can provide an important reference for doctors' diagnosis.

Conclusion

LR using RF as a feature selection method can provide more accurate results in predicting hypertension compared with statistical analysis or XGBoost as a feature selection method. Besides, Duration, BMI, cIMT, SBP and ESPWV are five important features in hypertension prediction research. Among them, cIMT and ESPWV are two key imaging indicators.

Journalism Ethics considerations

Ethical issues (Including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc.) have been completely observed by the authors.

Acknowledgements

The study was supported by 1) National Natural Science Foundation of China (11771216) and 2) the Key Research and Development Program of

Jiangsu Province (Social Development) (BE2019725).

Conflict of interest

The authors declare that there is no conflict of interests.

References

1. Wang TJ, Gona P, Larson MG, et al (2006). Multiple biomarkers for the prediction of first major cardiovascular events and death. *N Engl J Med,* 355:2631-2639.
2. Laurent S, Boutouyrie P, Asmar R, et al (2001). Aortic stiffness is an independent predictor of all-cause and cardiovascular mortality in hypertensive patients. *Hypertension,* 37:1236-1241.
3. Bérard E, Bongard V, Ruidavets JB, Amar J, Ferrières J (2013). Pulse wave velocity, pulse pressure and number of carotid or femoral plaques improve prediction of cardiovascular death in a population at low risk. *J Hum Hypertens,* 27:529–534.
4. O'Brien E, Atkins N, Stergiou G, et al (2010). European society of hypertension international protocol revision 2010 for the validation of blood pressure measuring devices in adults. *Blood Press Monit,* 15:23-38.
5. Cruickshank K, Riste L, Anderson Simon G, et al (2002). Aortic pulse-wave velocity and its relationship to mortality in diabetes and glucose intolerance: an integrated index of vascular function? *Circulation,* 106:2085-2090.
6. Townsend RR, Wimmer NJ, Chirinos JA, et al (2010). Aortic PWV in chronic kidney disease: a CRIC ancillary study. *Am J Hypertens,* 23:282-289.
7. Lorenz MW, Polak JF, Kavousi M, et al (2012). Carotid intima-media thickness progression to predict cardiovascular events in the general population (the PROG-IMT collaborative project): a meta-analysis of individual participant data. *Lancet,* 379:2053-2062.
8. Williams B, Mancia G, Spiering W, et al (2018). 2018 ESC/ESH Guidelines for the management of arterial hypertension: The Task Force for the management of arterial hypertension of the European Society of Cardiology (ESC) and the European Society of Hypertension (ESH). *Eur Heart J,* 39:3021-3104.
9. Coude M, Pernot M, Messas E, et al (2011). Ultrafast imaging of the arterial pulse wave. *IRBM,* 32(2): 106-108.
10. Robertson CM, Gerry F, Fowkes R, et al (2012). Carotid intima-media thickness and the prediction of vascular events. *Vasc Med,* 17:239-248.
11. Vermeersch SJ, Dynamics B, Society L (2010). Determinants of pulse wave velocity in healthy people and in the presence of cardiovascular risk factors: establishing normal and reference values. *Eur Heart J,* 31:2338-2350.
12. Wilkinson IB, McEniery CM, Schillaci G, et al (2010). ARTERY Society guidelines for validation of non-invasive haemodynamic measurement devices: Part 1, arterial pulse wave velocity. *Artery Res,* 4:34-40.
13. Zhu ZQ, Chen LS, Jiang XZ, et al (2021). Absent atherosclerotic risk factors are associated with carotid stiffening quantified with ultrafast ultrasound imaging. *Eur Radiol,* 31:3195-3206.
14. Zhu ZQ, Chen LS, Wang H, et al (2019). Carotid stiffness and atherosclerotic risk: non-invasive quantification with ultrafast ultrasound pulse wave velocity. *Eur Radiol,* 29:1507-1517.
15. Mirault T, Pernot M, Frank M, et al (2015). Carotid stiffness change over the cardiac cycle by ultrafast ultrasound imaging in healthy volunteers and vascular Ehlers-Danlos syndrome. *J Hypertens,* 33:1890-1896.
16. Yang W, Wang Y, Yu Y, et al (2020). Establishing normal reference value of carotid ultrafast pulse wave velocity and evaluating changes on coronary slow flow. *Int J Cardiovas Imaging,* 36:1931-1939.
17. Breiman L (2001). Random Forests. *Mach Learn,* 45:5-32.
18. Chen TQ, Guestrin C (2016). *XGBoost: A scalable tree boosting system.* ACM, New York, pp785-794.
19. Hermeling E, Reesink KD, Kornmann LM, et al (2009). The dicrotic notch as alternative time-reference point to measure local pulse wave velocity in the carotid artery by means of ultrasonography. *J Hypertens*, 27:2028-2035.