Evaluating the Risk of Hypertension in Residents in Primary Care in Shanghai, China with Machine Learning Algorithms

Ning Chen
School of Public Health, Shanghai Jiaotong University School of Medicine
Feng Fan
School of Medicine, Tongji University
Jinsong Geng
Medical School of Nantong University
Yan Yang
School of Economics & Management, Tongji University
Ya Gao
School of Public Health, Shanghai Jiaotong University School of Medicine
Hua Jin
Department of General Practice, Yangpu Hospital, Tongji University School of Medicine
Qiao Chu
School of Public Health, Shanghai Jiaotong University School of Medicine
Dehua Yu
Department of General Practice, Yangpu Hospital, Tongji University School of Medicine
Zhaoxin Wang
School of Public Health, Shanghai Jiaotong University School of Medicine
Jianwei Shi (shijianwei_amy@126.com)
School of Public Health, Shanghai Jiaotong University School of Medicine

Article

Keywords:

Posted Date: March 24th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1457304/v1

Abstract

The prevention of hypertension in primary care requires an effective and suitable hypertension risk assessment model. The aim of this study was to develop and compare the performances of three machine learning algorithms in predicting the risk of hypertension for residents in primary care in Shanghai, China. A dataset of 40,261 subjects over the age of 35 years was extracted from Electronic Healthcare Records of 47 community health centres from 2017 to 2019 in the Pudong district of Shanghai. The XGBoost model outperformed the other two models and achieved an AUC of 0.765 in the testing set. Twenty features were selected to construct the model, including age, diabetes status, urinary protein, BMI, elderly health self-assessment, creatinine, systolic blood pressure of the upper right arm, waist circumference, smoking status, low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, frequency of drinking, glucose, urea nitrogen, total cholesterol, diastolic blood pressure of the upper right arm, exercise frequency, time spent engaged in exercise, high salt consumption, and triglycerides. XGBoost outperformed random forest and logistic regression in predicting the risk of hypertension in primary care. Integration of such a risk assessment model into primary care may improve the prevention and management of hypertension in residents.

Introduction

Hypertension is becoming increasingly common in primary care. There are an estimated 245 million adults with hypertension in China[1]. An early warning after accurately evaluating the risk of hypertension in primary care patients can alert individuals in the healthy population or subhealthy population with unhealthy lifestyles to take measures to slow or stop the progression of hypertension. Similar practices have been implemented in foreign countries. For instance, management of risk factors for various chronic diseases has been implemented in primary care in Australia[2]. Risk assessment models are a cost-effective measure for identifying high-risk individuals with chronic diseases[3,4]. Nevertheless, few existing models can be applied to the health management services provided in primary care. The most intractable problem is that most of these models are targeted at patients in a hospital setting[5]; thus, the data input into the models are all extracted from the EHRs of hospitals, which may not be easily available in primary care and suitable for general practitioners to put into use.

Machine learning (ML) is a nuclear branch of artificial intelligence that has been employed everywhere knowingly or unknowingly, not only in industry and the military but also in medicine and healthcare[6]. As a modern data mining, extraction, and analysis technology, ML has the extraordinary ability to automatically train itself and improve its performance without human instruction or elaborate programming[7,8]. With the ability to identify a pattern or make a decision based on the knowledge input, ML algorithms have demonstrated their excellent performance in the area of risk evaluation of diseases. Higher accuracy separates ML algorithms from various other statistical methods. Highly precise risk prediction models for future hypertension were constructed by artificial intelligence techniques in Japan[9]. Health check-up data from 18,258 Japanese individuals were utilized to develop a risk prediction model

for new-onset hypertension by machine learning techniques. The XGBoost ensemble outperformed the logistic regression models (AUC=0.859), with AUCs of 0.877 and 0.881, respectively. A study based on several easy-to-collect risk factors to predict the risk of hypertension also revealed that the random forest (AUC=0.92), CatBoost (AUC=0.87), and MLP neural network (AUC=0.78) models performed better than logistic regression (AUC=0.77)[10]. Although ML is applicable in an extensive range of contexts, the ML algorithm technique alone is insufficient to solve real-world problems[11]. Thus, health and medical data in a primary care setting were utilized to facilitate the practical implementation of the risk assessment model for residents in primary care.

The objective of this study is to develop and compare the performances of three ML algorithms on predicting the risk of hypertension for residents over the age of 35 years in primary care in Shanghai, China.

Materials And Methods

Data source. The dataset was extracted from the electronic healthcare records of 47 community health centres in the Pudong district of Shanghai. Health records, health examinations and other health-related data of community residents over 35 years old from 2017 to 2019 were collected as the original set of data. A total of 40261 subjects were enrolled in the study. The dataset included 20 variables containing information regarding demographic characteristics, diagnosis, biochemical indicators and lifestyles. The characteristics of the participants in primary care are shown in Table 1.

Definition of hypertension. Hypertension was defined as (1) systolic blood pressure (SBP) ≥140 mmHg and/or diastolic blood pressure (DBP) ≥ 90 mmHg, which was measured three times on different days in the clinic without the use of antihypertensive drugs, according to Chinese guidelines for the prevention and treatment of hypertension (2018 revised edition)[12] and/or (2) a diagnosis of hypertension by a physician and/or (3) antihypertension treatment.

Data processing. Outliers were handled by interquartile range (IQR). The IQR is evaluated as IQR=Q3−Q1. Q3 is the upper quartile, and Q1 is the lower quartile. Outliers were defined as records that fell below Q1-(1.5*IQR) or above Q3+ (1.5*IQR).

Missing values, such as data with null rows and columns, were deleted. Different methods, such as the mean values, median values, mode values, feature combinations and null values, were adopted for dealing with the missing values according to the characteristics of different variables.

Discretization was performed by splitting the range of the continuous variables into intervals to save time needed to build the risk assessment model and improve the assessment results[13].

Feature selection. Feature selection, which is one of the essential parts of building a good prediction model, was employed in this study to improve the prediction accuracy by choosing the most important variables. Moreover, it facilitates a reduction in the resources (time and space) needed to construct the

model[14]. The embedded method was applied in this study for feature selection. It integrates the feature selection process with the model training process. This method takes variable interactions into consideration and is less computationally demanding than the wrapper method[15].

A total of twenty features were selected to construct the model: age, diabetes status, urinary protein, BMI, elderly health self-assessment (EHSA), creatinine (Cr), systolic blood pressure of the upper right arm (SBP), waist circumference (WC), smoking status, low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), frequency of drinking, glucose, urea nitrogen, total cholesterol (TC), diastolic blood pressure of the upper right arm (DBP), frequency of exercise, time spent engaged in exercise, high salt consumption, and triglycerides (TGs).

Machine learning algorithms. Extreme Gradient Boosting (XGBoost) is a supervised ML algorithm[16]. It generates an ensemble of classification trees and is used in the gradient boosting framework[17]. Overfitting can be largely decreased in XGBoost by means of parallel calculation and regularization lifting technology[18].

Random forest is a supervised classification algorithm[19]. It works by learning simple decision rules extracted from the data features and overcomes the limitation of overfitting of the decision trees[20].

Logistic regression is an algorithm that classifies values through the application of a logistic function to coefficients calculated by linear regression[21]. It requires that the dependent variable be a second-level score or a second-level evaluation.

Model evaluation and validation. A confusion matrix was employed to evaluate the performance of the models based on ML algorithms for the assessment of hypertension risk. The distinguishing abilities of the risk assessment model were evaluated with the receiver operator characteristic curve (ROC) and the area under the receiver operating characteristic curve (AUC)[22]. The performance of the models was evaluated with the sensitivity (true positive rate, TPR), specificity (true negative rate, TNR), positive predictive value (PPV), negative predictive value (NPV), accuracy (ACC), and F1-score[23.24].

Determination of the cut-off point. The evaluations were kinds of probabilities; thus, a cut-off point was needed to classify the prediction probabilities. The probability of having hypertension was represented by ‘P’ in the model. The cut-off point was utilized to classify the evaluated probabilities belonging to the positive results or negative results. We adopted a cut-off point of 0.5 in this study, which meant that participants were evaluated to be at high risk for hypertension when $P \geq 0.5$; otherwise, they were not.

Statistical analysis. Basic descriptive statistics were used to depict the characteristics of the subjects, including demographic characteristics and health-related factors. All normally distributed measurement data are depicted as the mean ± standard deviation (X±SD), nonnormally distributed measurement data are reported as the median (25th percentile, 75th percentile), and the counting data are expressed as the frequency and proportion. Between groups, normally distributed measurement data were compared by T

test, nonnormally distributed measurement data were compared by rank sum test, and the counting data were analysed by chi-square test. P values less than 0.05 were considered statistically significant. All statistical analyses were performed using IBM SPSS Statistics version 22.0 (IBM Corp., Armonk, NY, USA).

For the assessment models, ML algorithms, XGBoost, random forest and logistic regression were utilized for the evaluation of the risk of hypertension and the effects of the risk factors. Python 3.7.3 was used for the construction of the risk assessment models of hypertension.

Ethics approval. This study was approved by the Ethics Committees of Tongji University (ref: LL-2016-ZRKX-017). The participants provided written informed consent to participate in this study.

Reporting guidelines. Results are presented in accordance to the Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines. STROBE and RECORD guidelines for observational studies and studies using routinely collected health data were also considered. The study was conducted in accordance with relevant institutional guidelines.

Results

Characteristics of the study population. A total of 40,261 subjects were included, with a mean age of 72.429 ± 7.643 years, and the mean age of patients with hypertension was 73.216 ± 7.696 years. The sample prevalence of hypertension was almost 62.19%. The differences between participants with hypertension and normotensive participants were statistically significant in terms of age, diabetes status, urinary protein, BMI, EHSA, Cr, SBP, WC, smoking status, LDL-C, HDL-C, frequency of drinking, glucose, urea nitrogen, TC, DBP, frequency of exercise, high salt consumption, and TG ($P < 0.01$). There were no statistically significant differences ($P>0.05$) in terms of time spent engaged in exercise. The characteristics of the study participants are summarized in Table 1.

Table 1
Characteristics of the participants in primary care

Feature	Hypertension(n = 25,038)	Normal (n = 15,223)	χ^2	P
age*	72.00(68.00–78.00)	70.00(66.00–75.00)	683.51[a]	<0.01
diabetes			2,077.18[b]	<0.01
no	16,512(65.95)	13,177(86.56)		
yes	8,526(34.05)	2,046(13.44)		
urinary protein			32.33[b]	<0.01
negative	8,261(32.99)	8,392(55.13)		
positive	581(2.32)	405(2.66)		
BMI*	24.98(23.01–27.30)	24.16(22.10–26.30)	458.44[a]	<0.01
EHSA			563.15[b]	<0.01
1	6,973(27.85)	5,973(39.24)		
2	12,604(50.34)	6,387(41.96)		
3	358(1.43)	219(1.44)		
4	277(1.11)	149(0.98)		
5	163(0.65)	46(0.30)		
Cr*	69.00(58.00–84.00)	66.00(56.00-77.70)	229.09[a]	<0.01
SBP*	140.00(130.00-153.00)	139.00(126.00-148.00)	326.93[a]	<0.01
WC*	87.00(81.00–93.00)	85.00(79.00–91.00)	157.52[a]	<0.01
smoking status			200.85[b]	<0.01
1	19,171(76.57)	10,238(67.25)		
2	1,159(4.63)	857(5.63)		
3	2,028(8.10)	1,700(11.17)		
LDL-C*	2.89(2.20–3.41)	2.99(2.46–3.63)	402.35[a]	<0.01
HDL-C*	1.35(1.11–1.54)	1.40(1.20–1.66)	586.65[a]	<0.01
frequency of drinking			97.64[b]	<0.01

Feature	Hypertension(n = 25,038)	Normal (n = 15,223)	χ^2	P
1	18,096(72.27)	9,837(64.62)		
2	2,753(11.00)	1,771(11.63)		
3	199(0.79)	151(0.99)		
4	918(3.67)	764(5.02)		
glucose*	5.60(5.13–6.90)	5.50(5.00-6.33)	247.31[a]	<0.01
urea nitrogen*	5.63(4.80–6.83)	5.63(4.80–6.37)	306.45[a]	<0.01
TC*	4.82(4.01–5.52)	4.99(4.35–5.72)	267.34[a]	<0.01
DPB*	78.00(72.00–84.00)	78.00(70.00–82.00)	235.77[a]	<0.01
frequency of exercise			17.48[b]	<0.01
1	14,751(58.91)	8,460(55.57)		
2	815(3.26)	391(2.57)		
3	1,495(5.97)	926(6.08)		
4	5,471(21.85)	3,331(21.88)		
high salt consumption			17.24[b]	<0.01
no	24,938(99.60)	15,199(99.80)		
yes	100(0.40)	24(0.20)		
TG*	1.39(1.12–1.84)	1.39(1.00-1.80)	13.22[a]	<0.01
time spent engaged in exercise*	30.00(30.00–30.00)	30.00(30.00–30.00)	0.41[a]	0.52

* refers to nonnormally distributed measurement data, reported as the median (25th percentile, 75th percentile). [a] refers to results of the rank sum test. [b] refers to the results of the chi-square test.

Construction of the risk assessment models. The training set and validation set were utilized to determine the optimal parameters for XGBoost, random forest and logistic regression. The parameters of each model under optimal performance are exhibited in Table 2. For other unlisted parameters in the three ML algorithms, default values were set.

Table 2
Configuration of parameters in each ML algorithm

ML algorithm	Parameter	Optimal value
XGBoost	learning_rate	0.05
	n_estimators	200
	gamma	5
	subsample	0.4
	colsample_bytree	0.9
	min_child_weight	5
	max_depth	6
	objective	binary:logistic
Random forest	n _estimators	40
	criterion	gini
	max_depth	None
	min_samples_split	200
	min_samples_leaf	1
	max_features	auto
Logistic regression	C	100
	class_weight	None
	max_iter	10
	solver	liblinear

Feature importance. The significant features of the XGBoost model, random forest model and logistic regression model are listed in Figs. 1–3, respectively. Urea nitrogen was the highest ranked feature for hypertension prediction in both the XGBoost model and the random forest model. BMI, SBP, TG, Cr, LDL-C, and glucose were ranked in the top 10 in all three models.

Model performance. We utilized various methods and evaluation metrics to assess the performances of the XGBoost, random forest, and logistic regression models in the training, validation, and testing sets. On the whole, the XGBoost model outperformed the other two models in TPR (0.864), TNR (0.488), PPV (0.735), NPV (0.686), ACC (0.722), F1-score (0.795), and AUC (0.765) in the testing set (Table 3).

Figure 4 summarizes the ROC curve areas obtained from the XGBoost model, random forest model and logistic regression model in the testing set. The areas under the ROC curves were different among the three models. The AUCs in the test set were 0.765 for XGBoost, 0.756 for random forest, and 0.707 for logistic regression (Table 4). The AUC of the XGBoost model was higher than that of the random forest and logistic regression models. Our results demonstrated that the XGBoost model had better predictive performance than the random forest and logistic regression models.

Table 3
The fitting results of the XGBoost, random forest, and logistic regression models for the training, validation, and testing sets.

ML algorithm	Dataset	TPR	TNR	PPV	NPV	ACC	F1-Score	AUC
XGBoost	training	0.886	0.530	0.756	0.739	0.752	0.816	0.818
	validation	0.862	0.480	0.732	0.678	0.717	0.791	0.753
	testing	0.864	0.488	0.735	0.686	0.722	0.795	0.765
Random Forest	training	0.896	0.434	0.723	0.718	0.722	0.800	0.782
	validation	0.871	0.446	0.721	0.678	0.711	0.789	0.745
	testing	0.816	0.548	0.748	0.644	0.714	0.780	0.756
Logistic regression	training	0.827	0.411	0.698	0.591	0.670	0.757	0.705
	validation	0.822	0.418	0.699	0.588	0.669	0.756	0.692
	testing	0.829	0.430	0.705	0.604	0.678	0.762	0.707

Table 4
AUC for the XGBoost, random forest, and logistic regression models for the training, validation, and testing sets

ML algorithm	Dataset	AUC
XGBoost	training	0.818
	validation	0.753
	testing	0.765
Random Forest	training	0.782
	validation	0.745
	testing	0.757
Logistic regression	training	0.705
	validation	0.692
	testing	0.707

Discussion

Among the twenty selected features in this study, BMI, SBP, TG, Cr, LDL-C, and glucose had a strong effect on hypertension prediction and were included among the top 10 in the ranking of the feature importance for all three models. Similar to the results of previous studies, features such as age[25,29,30], BMI[26,29], diabetes status[27], Cr[26], blood pressure[28,30], WC[29], smoking status[29], LDL-C[26,29], HDL-C[26], drinking[29], glucose[30], TC[26,27], exercise[31], salt intake[32], and TG[27] were found to be predictors of hypertension in the risk assessment model of hypertension.

However, to the best of our knowledge, urinary protein, urea nitrogen, and EHSA entered the models as new components that have not been included in risk evaluation models of hypertension in previous studies.

A study collected data from three exams in the Strong Heart Study, explored the risk factors for hypertension by means of generalized linear models and demonstrated that systolic blood pressure was significantly and positively associated with albuminuria, age, and obesity and negatively associated with smoking. Moreover, participants with more severe albuminuria status or older age developed higher SBP, while DBP was not significantly affected by albuminuria status[33]. This study in American Indians revealed that having macro/microalbuminuria is a significant risk factor for hypertension, which can explain why urinary protein was selected as one of the features in our model to some extent. Urinary protein may also affect the development of hypertension in Chinese individuals or facilitate the risk assessment of hypertension in Chinese individuals. Furthermore, Kim et al. reported that subjects with high normal BP had an independently significant association with microalbuminuria by means of

multiple logistic regression analysis, with an odds ratio of 1.692 and a 95% confidence interval from 1.097 to 2.611[34]. These results from a Korean population indicated that compared to individuals with normal BP, those with high normal BP have more risk factors for hypertension and cardiovascular diseases, for instance, albuminuria. Since the incidence of urinary protein was significantly higher in the prehypertensive population than in the normal population, urinary protein should receive attention in future predictive studies and intervention measures.

Although we rarely found urea nitrogen to be included as a predictive factor in the risk prediction models, it was found to be a significant risk factors for hypertension. A case-control study conducted among university staff found that staff with high serum urea levels had a higher risk of hypertension than those with normal urea levels (OR=1.452), which implies that the level of urea is also of great importance as one of the risk factors for hypertension[35]. Not coincidentally, this phenomenon has been found among middle-aged and elderly people. SBP was found to be positively correlated with the concentration of blood urea nitrogen (r=0.16424, P=0.0105) and the concentration of blood uric acid (r=0.16023, P=0.0126) among middle- and older-aged populations in Guangzhou, China, as well as DBP (concentration of blood urea nitrogen: r =0.13506, P =0.0358; blood uric acid: r=0.16562, P=0.0099)[36]. The results of stepwise regression analysis also indicated that there was still a significant positive correlation between SBP, DBP and concentrations of blood urea nitrogen and blood uric acid. The role of urea nitrogen, one of the features entered into our risk assessment model, in the occurrence and development of hypertension still needs to be further investigated.

EHSA was also one of the predictors entered into our model. Kaplan and Camacho have already demonstrated that the association between level of perceived health and mortality persisted in multiple logistic analyses controlling for age, sex, physical health status, health practices, social network participation, income, education, health relative to age peers, anomy, morale, depression, and happiness[37]. The results reminded us that self-assessment of health might serve as a comprehensive reflection of unmeasurable factors and as an indication of some underlying diseases or an early stage of the diseases. Evidence has shown that psychosocial factors have a strong influence on health status measures[38]. Zhang et al. revealed that the proportion of elderly individuals with poor or normal health self-assessments suffering from common chronic diseases was significantly increased[39]. The health self-assessment epitomizes the health concept and self-perception of health status of elderly individuals to some extent, which might have an underlying predictive value on the prediction of the risk of hypertension and should thus be given more attention in future research, as well as the practice in primary care.

Unlike traditional risk assessment methods, our study employed ML algorithms for model construction. XGBoost exhibited the best performance compared to random forest and logistic regression. Logistic regression assumes that every variable should be independent, and the model possesses only a linear partition surface. However, the associations between exposure factors and diseases are often affected by various confounding factors, which leads to the large deviation and low accuracy when fitting the model

through logistic inference. In contrast, XGBoost and random forest are nonparametric algorithms[40] that do not assume that there is a functional relationship between the features and outcomes, as required by logistic regression. A greedy algorithm is executed to determine the optimal splits in the data that reduce the entropy of the outcome to the utmost extent during every split. As a result, once a feature is selected, the significance of any highly related feature will decrease greatly due to the completion of the effective split done by the original feature previously. Consequently, the entropy of the outcome will no longer be reduced effectively by related features. Therefore, XGBoost and random forest are robust to related features. The reason why XGBoost outperforms the other methods may be that it introduces the regularized loss function[41] and combines gradient lifting algorithms and decision trees, which preserves the correlation between features during the modelling process[42].

After the risk assessment of hypertension, subsequent interventions and management to prevent or postpone the occurrence and development of hypertension are crucially important in high-risk populations. Continuous monitoring and management are imperative for high-risk patients. On the one hand, realtimeness and continuity monitoring can detect any problem without delay. On the other hand, early signs of detected symptoms can alert both general practitioners (GPs) and individuals in a timely manner. For high-risk populations, corresponding individual intervention strategies targeting the main risk factors should be prescribed by GPs in primary care. For instance, lifestyle factors such as exercise, eating habits, and drinking habits can be improved under the guidance of GPs after risk assessment. Evidence has revealed that a high concentration of parks or playgrounds in residential areas may reduce the risk of hypertension, mainly attributable to the cultivation and formation of exercise habits, which implies the importance of interventions in communities[43].

However, there were several limitations in our study. One of the limitations of the study was that it had a cross-sectional design, and the results could not indicate causality in this situation. A prospective cohort study is needed to further identify the cause-and-effect relationships. Second, the risk assessment model was designed considering only variables available in the setting of primary care, and variables regarding mental health and hereditary factors were not included. Third, we measured several variables, such as age, urinary protein, BMI, and Cr, on only a single occasion and did not take changes in these variables into consideration.

Conclusion

Early identification and the corresponding preventive strategies in primary care remain insufficient in China. XGBoost outperformed random forest and logistic regression in predicting the risk of hypertension in primary care. Integration of such a risk assessment model into primary care may help general practitioners target populations at high-risk for hypertension, tailor the corresponding preventive measures and treatment strategies to those at high risk, improve the awareness of residents regarding health risks and their adherence towards targeted intervention, and eventually facilitate individuals' health and quality of life while decreasing healthcare costs.

Declarations

Data availability

The datasets generated and/or analysed during the current study are not publicly accessible but are available from the corresponding author upon reasonable request.

Acknowledgements

We thank all the participants involved in this study.

Author contributions

N.C. was involved in design of study, analysis of results and wrote the manuscript. F.F performed data collection, proofread and modified the format. D.Y., Z.W. and J.S. supervised the work and were involved in study design. J.G. and Y.Y. helped with data interpretation and graphing. Y.G., H.J. and Q.C. revised the manuscript. All authors reviewed and approved the final version of the manuscript.

Funding

This study was supported by Shanghai Education Science Research Project (C2021039), National Natural Science Foundation of China (71774116 and 71603182), Shanghai Public Health Outstanding Young Personnel Training Program (GWV-10.2-XD07), soft science project of Shanghai Science and Technology Commission (22692107200), National Key Research and Development Program of China (2018YFC2000700), and Shanghai Pujiang Program (2020PJC080). The funding sources had no role in the design of this study nor any role during its execution, analyses, data interpretation, or decision to submit results.

Competing interests

The authors declare no competing interests.

References

1. Wang, Z. W. et al. Status of Hypertension in China: Results From the China Hypertension Survey, 2012-2015. *Circulation* **137**, 2344-2356. https://doi.org/10.1161/CIRCULATIONAHA.117.032380 (2018).
2. The Royal Australian College of General Practitioners. Guidelines for preventive activities in general practice. 9th edition. https://www.racgp.org.au/getattachment/1ad1a26f-9c8b-4e3c-b45b-3237272b3a04/Guidelines-for-preventive-activities-in-general-practice.aspx (2021).
3. Chen, X. et al. Risk score model of type 2 diabetes prediction for rural Chinese adults: the Rural Deqing Cohort Study. *J. Endocrinol. Invest.* **40**, 1115-1123. https://doi.org/10.1007/s40618-017-0680-4 (2017).

4. Hart, G. R., Roffman, D. A., Decker. R. & Deng, J. A multi-parameterized artificial neural network for lung cancer risk prediction. *PLoS One* **13**, e0205264. https://doi.org/10.1371/journal.pone.0205264 (2018).
5. Andriani, P. & Chamidah, N. Modelling of Hypertension Risk Factors Using Logistic Regression to Prevent Hypertension in Indonesia. *Journal of Physics: Conference Series* **1306**, 012027 (7 pp.). https://doi.org/10.1088/1742-6596/1306/1/012027 (2019).
6. Dash, S. S., Nayak, S. K. & Mishra, D. A Review on Machine Learning Algorithms. In: Mishra D., Buyya R., Mohapatra P., Patnaik S. (eds) Intelligent and Cloud Computing. Smart Innovation, Systems and Technologies, vol 153. Springer, Singapore. https://doi.org/10.1007/978-981-15-6202-0_51 (2021).
7. Alpaydin, E. Introduction to machine learning. (MIT press, 2014).
8. Marsland, S. Machine learning: an algorithmic perspective. (CRC press, 2015).
9. Kanegae, H. et al. Highly precise risk prediction model for new-onset hypertension using artificial intelligence techniques. *J. Clin. Hypertens. (Greenwich).* **22**, 445-450. https://doi.org/10.1111/jch.13759 (2020).
10. Zhao, H. et al. Predicting the Risk of Hypertension Based on Several Easy-to-Collect Risk Factors: A Machine Learning Method. *Front. Public. Health.* **9**, 619429. https://doi.org/10.3389/fpubh.2021.619429 (2021).
11. Benton, W. C. Machine Learning Systems and Intelligent Applications. *IEEE SOFTWARE* **37**, 43-49. DOI:10.1109/MS.2020.2985224 (2020).
12. Writing Group of 2018 Chinese Guidelines for the Management of Hypertension. et al. 2018 Chinese guidelines for the management of hypertension. *Chin. J. Cardiovasc. Med.* **24**, 24-56. https://doi.org/10. 3969 /j. issn. 1007-5410. 2019. 01. 002 (2019).
13. Kurgan, L. & Cios, K. J. Discretization algorithm that uses class-attribute interdependence maximization. *IC-AI'2001: PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, VOLS I-III*, 980-986. https://www.webofscience.com/wos/alldb/full-record/WOS:000173960400153 (2001).
14. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**, 1157-1182. https://doi.org/10.1162/153244303322753616 (2003).
15. Shi, X. et al. An ensemble-based feature selection framework to select risk factors of childhood obesity for policy decision making. *BMC. Med. Inform. Decis. Mak.* **21**, 222. https://doi.org/10.1186/s12911-021-01580-0 (2021).
16. Plagnol, V. et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* **28**, 2747-2754. https://doi.org/10.1093/bioinformatics/bts526 (2012).
17. Kabiraj, S. et al. Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm. *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 4 pp. https://doi.org/10.1109/ICCCNT49239.2020.9225451 (2020).

18. Zhang, Y. et al. Comparison of Prediction Models for Acute Kidney Injury Among Patients with Hepatobiliary Malignancies Based on XGBoost and LASSO-Logistic Algorithms. *Int. J. Gen. Med.* **14**, 1325-1335. https://doi.org/10.2147/IJGM.S302795 (2021).
19. Prasad, A. M., Iverson, L. R. & Liaw, A. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems* **9**, 181-199. https://doi.org/10.1007/s10021-005-0054-1 (2006).
20. Sakr, S. et al. Using machine learning on cardiorespiratory fitness data for predicting hypertension: The Henry Ford Exerclse Testing (FIT) Project. *PLoS One* **13**, e0195344. https://doi.org/10.1371/journal.pone.0195344 (2018).
21. Buya, S., Tongkumchum, P. & Owusu, B. E. Modelling of land-use change in Thailand using binary logistic regression and multinomial logistic regression. *Arab. J. Geosci.* **13**, 12. https://doi.org/10.1007/s12517-020-05451-2 (2020).
22. Moons, K. G., Harrell, F. E. & Steyerberg, E. W. Should scoring rules be based on odds ratios or regression coefficients? *J. Clin. Epidemiol.* **55**, 1054-1055. https://doi.org/10.1016/s0895-4356(02)00453-5 (2002).
23. Cai, Q. C. et al. Derivation and validation of a prediction rule for estimating advanced colorectal neoplasm risk in average-risk Chinese. *Am. J. Epidemiol.* **175**, 584-593. https://doi.org/10.1093/aje/kwr337 (2012).
24. Lavrac, N. Selected techniques for data mining in medicine. *Artif. Intell. Med.* **16**, 3-23. https://doi.org/10.1016/s0933-3657(98)00062-1 (1999).
25. Ren, Z. G, et al. A novel predicted model for hypertension based on a large cross-sectional study. *Sci. Rep.* **10**, 10615. https://doi.org/10.1038/s41598-020-64980-8 (2020).
26. Akdag, B. et al. Determination of risk factors for hypertension through the classification tree method. *Adv. Ther.* **23**, 885-892. https://doi.org/10.1007/BF02850210 (2006).
27. Kshirsagar, A. V. et al. A hypertension risk score for middle-aged and older adults. *J. Clin. Hypertens. (Greenwich).* **12**, 800-808. https://doi.org/10.1111/j.1751-7176.2010.00343.x (2010).
28. Kanegae, H., Oikawa, T., Suzuki, K., Okawara, Y. & Kario, K. Developing and validating a new precise risk-prediction model for new-onset hypertension: The Jichi Genki hypertension prediction model (JG model). *J. Clin. Hypertens. (Greenwich).* **20**, 880-890. https://doi.org/10.1111/jch.13270 (2018).
29. Xu, F. et al. Development and validation of prediction models for hypertension risks in rural Chinese populations. *J. Glob. Health.* **9**, 020601. https://doi.org/10.7189/jogh.09.020601 (2019).
30. Chien, K. L. et al. Prediction models for the risk of new-onset hypertension in ethnic Chinese in Taiwan. *J. Hum. Hypertens.* **25**, 294-303. https://doi.org/10.1038/jhh.2010.63 (2011).
31. Niiranen, T. J., Havulinna, A. S., Langén, V. L., Salomaa, V. & Jula, A. M. Prediction of Blood Pressure and Blood Pressure Change With a Genetic Risk Score. *J. Clin. Hypertens. (Greenwich).* **18**, 181-186. https://doi.org/10.1111/jch.12702 (2016).
32. Xu, Y. et al. Establishment and verification of a nomogram prediction model of hypertension risk in Xinjiang Kazakhs. *Medicine (Baltimore)* **100**, e27600.

https://doi.org/10.1097/MD.0000000000027600 (2021).

33. Wang, W. et al. A longitudinal study of hypertension risk factors and their relation to cardiovascular disease: the Strong Heart Study. *Hypertension* **47**, 403-409. https://doi.org/10.1161/01.HYP.0000200710.29498.80 (2006).
34. Kim, B. J. et al. Comparison of microalbuminuria in 2 blood pressure categories of prehypertensive subjects. *Circ. J.* **71**, 1283-7. https://doi.org/10.1253/circj.71.1283 (2007).
35. Guan, X. P., Xiang, H. & Xia, H. Risk factors of essential hypertension among university staff: a case-conrol study. *Chin. J. Public. Health.* **27**, 501-503 (2011).
36. Xiao, M. et al. Relationship between blood pressure and blood uric acid, urea nitrogen in middle and older-aged population in Guangzhou. *South China Journal of Cardiovascular Diseases* **15**, 457-460. https://doi.org/10.3969/j.issn.1007-9688.2009.06.012 (2009).
37. Kaplan, G. A. & Camacho, T. Perceived health and mortality: a nine-year follow-up of the human population laboratory cohort. *Am. J. Epidemiol.* **117**, 292-304. https://doi.org/10.1093/oxfordjournals.aje.a113541 (1983).
38. Ring, D. et al. Self-reported upper extremity health status correlates with depression. *J. Bone. Joint. Surg. Am.* **88**, 1983-1988. https://doi.org/10.2106/JBJS.E.00932 (2006).
39. Zhang, F. M. & Xu, H. J. Research on the relationship between self-assessment of health and chronic diseases in elderly population. *Chinese Journal of Gerontology* **28**, 2353-2355 (2008).
40. Chen, T. Q., He, T., Benesty M, & Tang, Y. *Understand your dataset with XGBoost*. https://cran.r-project.org/web/packages/xgboost/vignettes/discoverYourData.html (2017).
41. Pan, B. Y. Application of XGBoost algorithm in hourly PM2.5 concentration prediction. 3RD INTERNATIONAL CONFERENCE ON ADVANCES IN ENERGY RESOURCES AND ENVIRONMENT ENGINEERING. *Book Series:IOP Conference Series-Earth and Environmental Science* **113**, 012127. https://doi.org/10.1088/1755-1315/113/1/012127 (2018).
42. Thomas, J., Hepp, T., Mayr, A. & Bischl, B. Probing for Sparse and Fast Variable Selection with Model-Based Boosting. *Comput. Math. Methods. Med.* **2017**, 1421409. https://doi.org/10.1155/2017/1421409 (2017).
43. Ye, C. et al. Prediction of Incident Hypertension Within the Next Year: Prospective Study Using Statewide Electronic Health Records and Machine Learning. *J. Med. Internet. Res.* **20**, e22. https://doi.org/10.2196/jmir.9268 (2018).

Figures

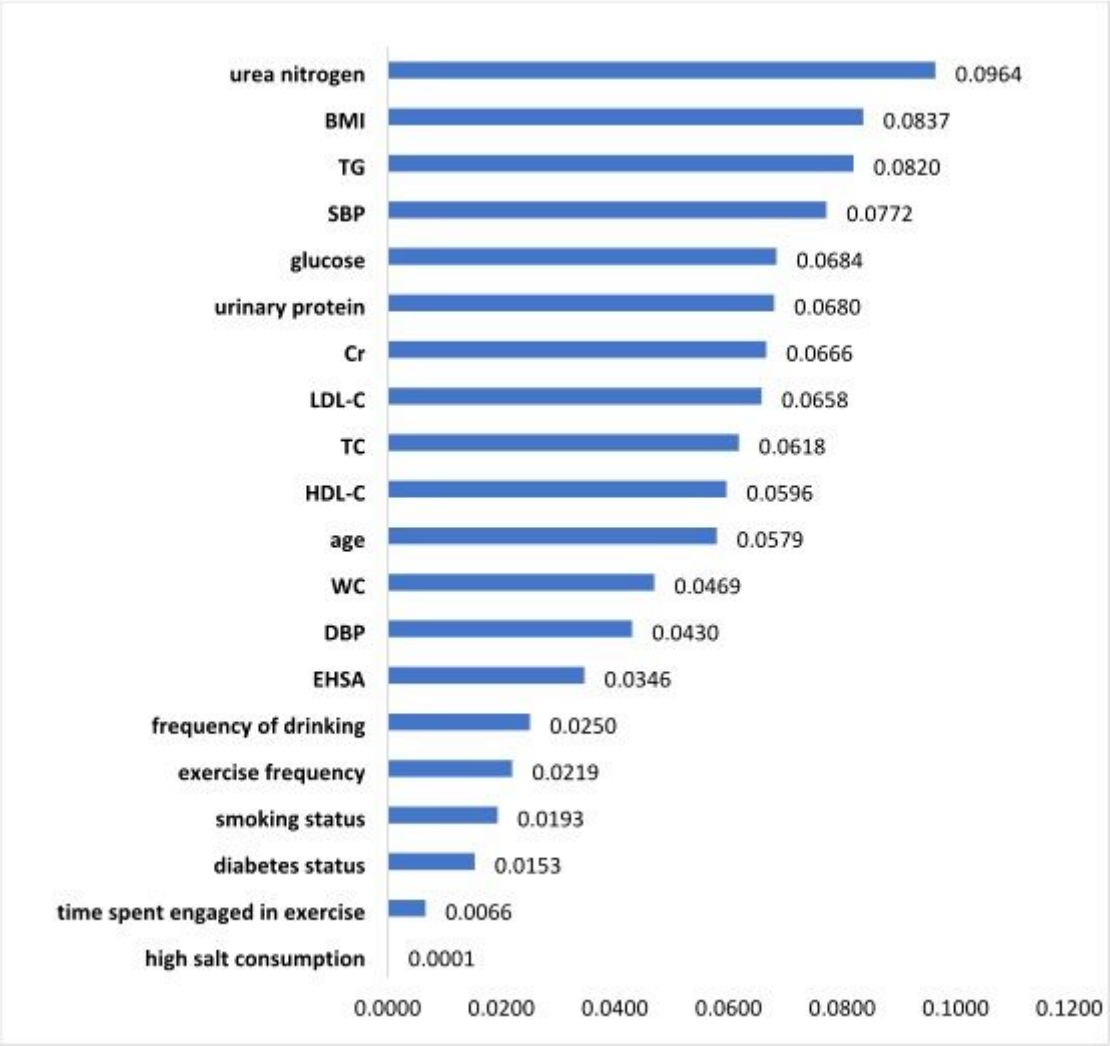

Figure 1

Feature importance of the XGBoost model

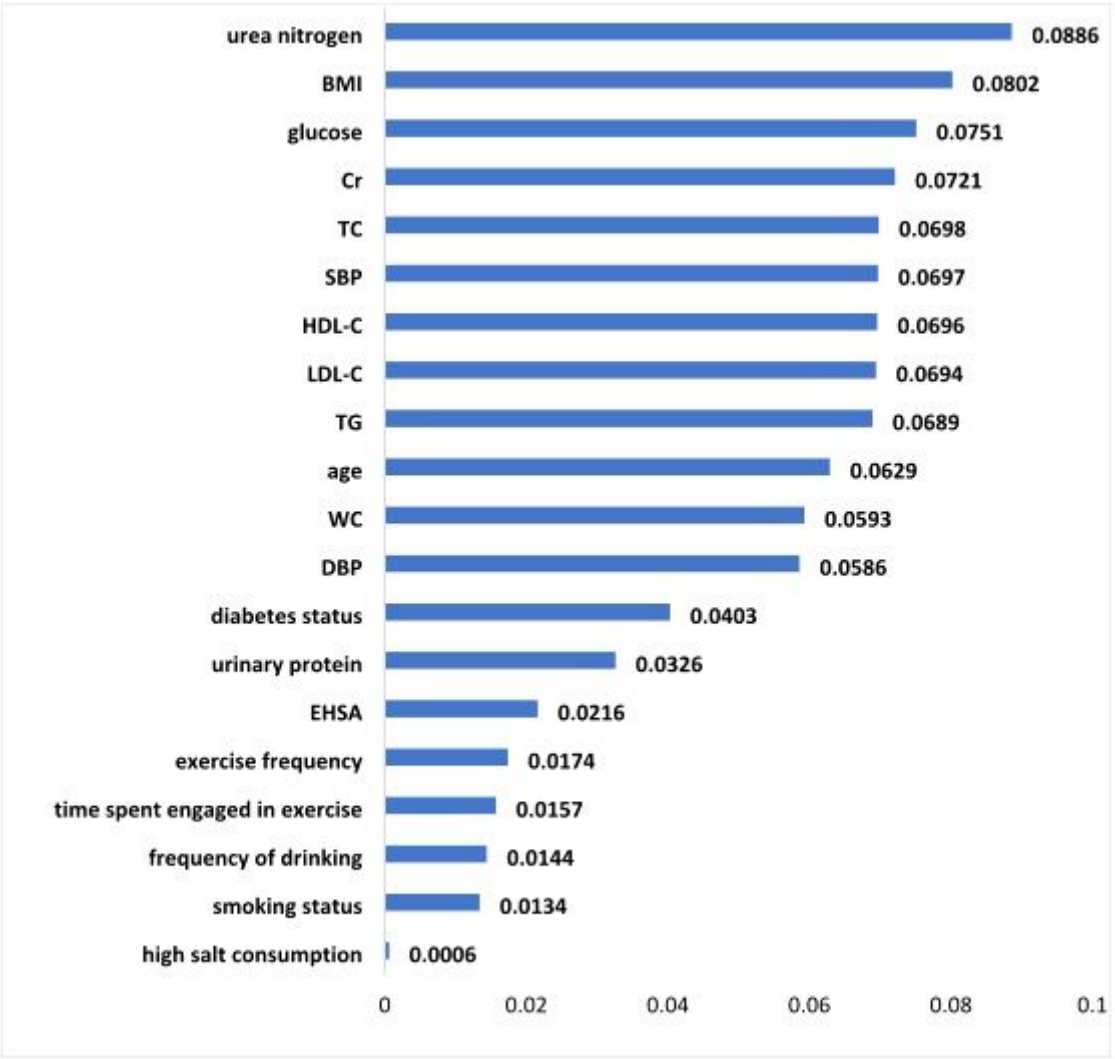

Figure 2

Feature importance of the random forest model

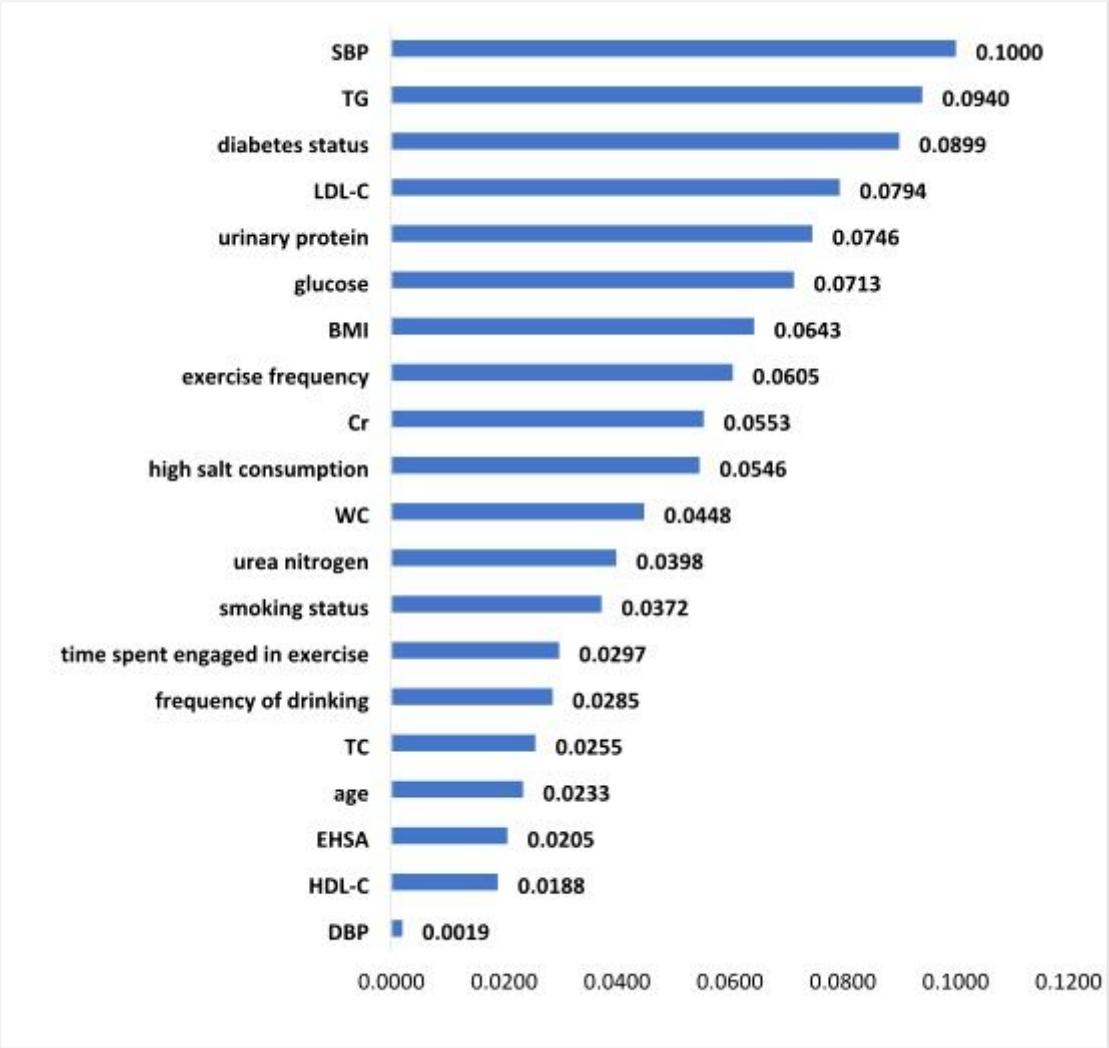

Figure 3

Feature importance of the logistic regression model

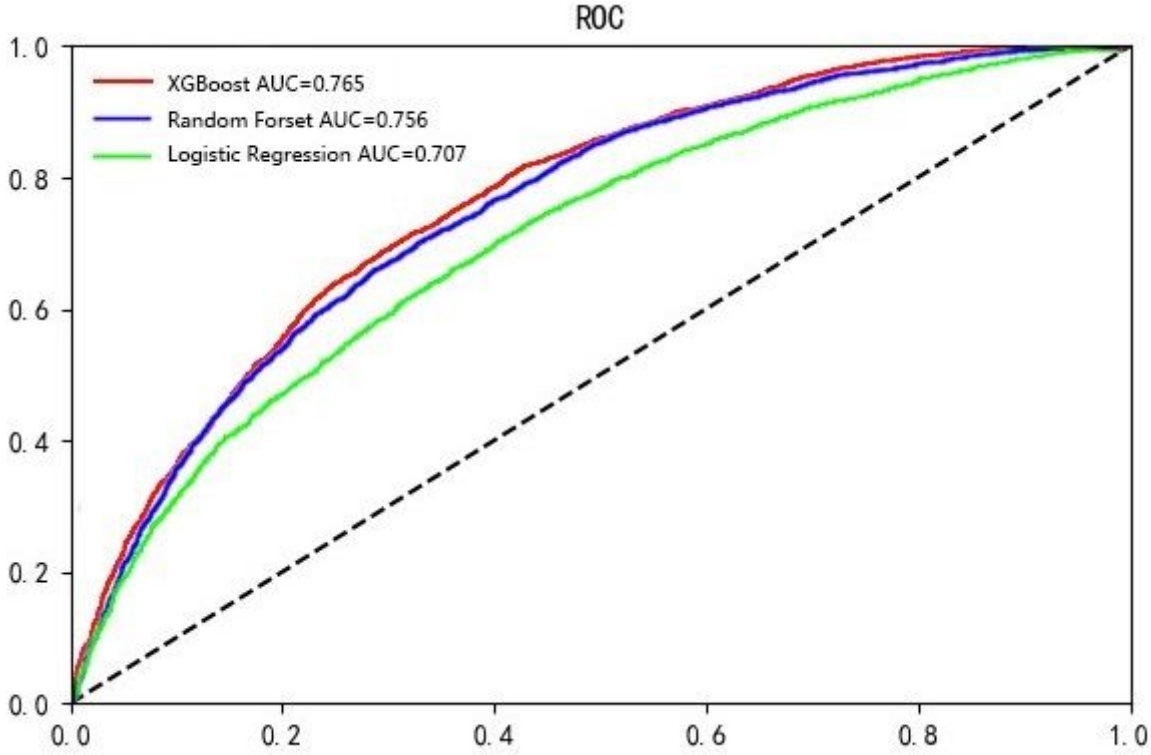

Figure 4

The ROC curves obtained from the XGBoost model, random forest model and logistic regression model. X axis: 1-specificity, Y axis: sensitivity. The reference line is shown as a dashed line (the black line).