

A Community-Based Machine Learning Model for Predicting Major Adverse Cardiovascular Events in Kazakh Hypertensive Patients

huiyu Shu

Chengdu Medical College

Yining Yuan

Chengdu Medical College

Bo Zhao

Altay City People's Hospital

Jiao Hu

Altay City People's Hospital

Yun Zheng

Chengdu Medical College

Shu Yang

Chengdu University of Traditional Chinese Medicine

xixi Feng

10005800@cmc.edu.cn

Chengdu Medical College

Article

Keywords: Kazakh Ethnicity, Hypertension, MACEs, Incidence Rate, Prediction Model

Posted Date: July 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9308614/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Aims

The current situation in China is that the number of people suffering from hypertension has reached 240 million, with cardiovascular diseases being the leading cause of death among both urban and rural residents, presenting a severe public health challenge.

Methods

A baseline survey was conducted using data from hypertensive patients registered at the "Hypertension Control Center" of Altay City People's Hospital and township health centers in Hongdun, Qiemu'erqieke, and Aweitan from January to December 2020. Prospective follow-up continued until December 31, 2024, with MACEs as the outcome. Patients were randomly divided into Training and Validation set (7:3 ratio) to construct and evaluate the model, respectively. Predictors were selected based on univariate analysis, Lasso regression, and clinician expertise. Four models—Cox regression, random forest, decision tree, and Extreme gradient boosting—were built to predict 3years MACEs risk. Model performance was assessed using positive/negative predictive values (PPV/NPV), accuracy (ACC), F1-score, receiver operating characteristic (ROC) curves, calibration curves, and decision curve analysis (DCA). Shapley additive explanations (SHAP) were applied for interpretability analysis of the optimal model.

Results

Among 1,180 enrolled participants (median follow-up: 1,176 days), 327 (27.7%) developed MACEs. Eight predictors were identified from demographics, physical examinations, laboratory tests, and clinical features. The DT model achieved an AUC of 0.776 for MACEs prediction, with robust performance in both training and Validation sets (ACC = 0.797, PPV = 0.444, NPV = 0.933, F1-score = 0.550), outperforming other models. Calibration and DCA curves confirmed its strong generalizability and clinical utility. SHAP analysis ranked LDL-C, age, office systolic blood pressure, and potassium as the top four risk factors.

Conclusion

The DT-based model effectively predicts 3-year MACEs risk in Kazakh hypertensive patients, and SHAP-derived interpretability supports clinical decision-making for this population.

Lay summary

This study investigated the risk of major adverse cardiovascular events (MACEs) and developed predictive models for Kazakh patients with hypertension in the Altay region of Xinjiang. Analysis of 1,180

patients revealed a MACEs incidence rate of 27.7% during a median follow-up of 1,176 days, with associated risk factors including age, diabetes status, serum potassium, fasting blood glucose, Hcy, TC, LDL-C and blood pressure. The study compared multiple predictive models, with decision tree models demonstrating optimal performance in predictive accuracy, stability, and clinical applicability for assessing MACEs risk in this population. These findings provide crucial evidence for primary prevention strategies targeting hypertension-related cardiovascular events.

Introduction

China is facing a substantial burden of hypertension, with an estimated 240 million affected individuals, against a backdrop where cardiovascular diseases remain the leading cause of death among both urban and rural residents[1]. Hypertension is closely associated with the incidence of major adverse cardiovascular events (MACEs), including unstable angina, non-fatal myocardial infarction, heart failure, and arrhythmia. The Kazakh ethnic group, characterized by distinct lifestyle and dietary habits, has the highest prevalence of hypertension in Xinjiang and also experiences a high frequency of MACEs. Due to the uneven distribution of healthcare resources in Xinjiang, residents in pastoral and rural areas often encounter difficulties in accessing and affording medical services, highlighting the importance of disease prevention. However, existing studies have paid insufficient attention to ethnic minorities in Xinjiang, and current risk prediction models show limitations when applied to the Kazakh population. Therefore, based on a real-world prospective cohort, this study identifies key influencing factors for MACEs, develops and validates a prediction model using follow-up data. To enhance clinical applicability, the Shapley Additive ExPlanations (SHAP) algorithm was applied to interpret the optimal model. Using data from the Hypertension Standard Management Center in Altay City, Xinjiang, this research aims to develop a risk prediction model for MACEs tailored to the local Kazakh hypertensive population, thereby providing a scientific basis for individualized prevention and treatment strategies for this specific group. Furthermore, to facilitate practical application, the prediction model will be deployed as an accessible web-based calculator for ease of use in clinical and community settings.

Material and methods

Study population

The data of this study were derived from the "Hypertension Standardization Center" established by the Altay City People's Hospital as the leading institution, in collaboration with three township health centers in satellite towns. These medical institutions cover the urban area of Altay and its surrounding major farming and pastoral regions, representing Kazakh patients with diverse economic levels, lifestyles, and access to medical resources. A baseline survey was conducted using the records of hypertensive patients archived in the "Hypertension Standardization Center" between January 1, 2020, and December 31, 2020. The study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of Altay City People's Hospital. Written informed consent was obtained from all participants.

Inclusion criteria:

- ☒ Kazakh ethnicity, aged ≥ 18 years;
- ☒ Meeting the diagnostic criteria for hypertension;
- ☒ Complete clinical medical record outcome data;
- ☒ Signed informed consent.

Exclusion criteria:

- ☒ Patients with MACEs;
- ☒ Patients with underlying diseases such as heart, brain, or kidney conditions;
- ☒ Pregnant women;
- ☒ Patients with other severe diseases, such as malignant tumors, immune deficiency, mental disorders, cognitive impairment, etc.

Data Collection and Baseline Assessments

A self-designed questionnaire was administered by uniformly trained investigators through face-to-face interviews. The questionnaire included:

- ☒ General demographic information: age, gender, marital status, smoking status, alcohol consumption, and salt intake score;
- ☒ Personal and medical history: history of diabetes, history of thyroid dysfunction.

Height and weight were measured using a stadiometer and the same model of electronic scale, respectively. Clinic systolic blood pressure (SBP) and diastolic blood pressure (DBP) were measured using an electronic blood pressure monitor (e.g., Omron HBP-1300 series) in a quiet, separate laboratory room after the participant had rested for 5 minutes. Three measurements were taken at 1-minute intervals, and the average of the three readings was recorded. Ambulatory blood pressure monitoring included daytime SBP/DBP, nighttime SBP/DBP, and 24-hour average SBP/DBP. A standardized 24-hour ambulatory blood pressure monitor was used, which automatically measured SBP and DBP at regular intervals (every 15–30 minutes) to continuously track blood pressure fluctuations and collect data on daily and nocturnal blood pressure variations.

Biochemical indicators were measured from fasting venous blood samples using an automatic biochemical analyzer (OLYMPUS AU2007). Parameters included serum potassium, sodium, estimated glomerular filtration rate (eGFR), serum creatinine, uric acid, homocysteine (Hcy), total cholesterol (TC),

low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), triglycerides (TG), and fasting blood glucose.

Left ventricular ejection fraction (LVEF) was assessed using an ultrasound system (Philips EPIQ7C). Dynamic cardiac images were obtained via echocardiography to measure end-diastolic volume (EDV) and end-systolic volume (ESV).

Follow-up and Outcomes

All enrolled patients were followed up four times in 2021, 2022, 2023, and 2024 to record the occurrence of MACEs. Self-reported adverse cardiovascular events during follow-up required supporting clinical diagnosis documentation from a secondary or higher-level hospital. Incident disease was defined as the first occurrence of a disease from the time of enrollment until the end of follow-up. If multiple events occurred in the same participant, the first event was considered the outcome.

The primary endpoint in this study was MACEs, including unstable angina, myocardial infarction, heart failure, ischemic stroke, and death[2]. Diagnoses of MACEs were based on the International Classification of Diseases, Tenth Revision (ICD-10)[3]. ICD-10 codes were as follows: unstable angina (I20.1, I20.8, I20.9); myocardial infarction (I21.0, I21.1, I21.8, I21.9); heart failure (I50.0, I50.1, I50.9); ischemic stroke (I63.0, I63.9); and death (R99).

Operational Definitions of Key Variables

Hypertension: Defined according to the *Clinical practice guideline for the management of hypertension in China*[4] as SBP ≥ 140 mmHg and/or DBP ≥ 90 mmHg, current use of antihypertensive medication, or previous diagnosis of hypertension.

Diabetes: Defined based on the *Standards of medical care for type 2 diabetes in China 2019*[5] as fasting blood glucose ≥ 7.0 mmol/L, current use of glucose-lowering medication, or previous diagnosis of diabetes.

Overweight/Obesity: Based on the *Chinese Guidelines for the Prevention and Control of Overweight and Obesity in Adults*[6], body mass index (BMI) was calculated as weight/height² (kg/m²). Overweight was defined as $24.0 \text{ kg/m}^2 \leq \text{BMI} < 28.0 \text{ kg/m}^2$, and obesity as $\text{BMI} \geq 28.0 \text{ kg/m}^2$.

Smoking: Defined as having smoked more than 100 cigarettes in total or having smoked continuously for at least 6 months.

Alcohol Consumption: Defined as consuming alcoholic beverages (beer, wine, liquor, etc.) at least twice per month.

Sample Size Estimation and Statistical Considerations

To ensure sufficient statistical power for accurately developing a prediction model for MACEs in hypertensive patients, the sample size was calculated according to recent guidelines[7] using the “pmsampsize” package in R. The outcome variable (MACEs) was binary. Based on previous studies, the model was expected to include approximately 10 predictors. According to published literature[1], the incidence of MACEs in hypertensive patients is approximately 24.7%. Therefore, the parameters were set as follows: event_rate = 24.7%, c_statistic = 0.75. The sample size estimation indicated that a minimum of 576 participants, including 143 events (MACEs), were required. The final sample size included in this study was 1,180 as detailed in Table 1.

Table 1
Sample Size Calculation Results

Criteria	Samp_size	Shrinkage	Parameter	CS_Rsq	Max_Rsq	Nag_Rsq	EPP
Criteria 1	576	0.90	10	0.1436	0.673	0.213	14.23
Criteria 2	270	0.81	10	0.1436	0.673	0.213	6.67
Criteria 3	286	0.90	10	0.1436	0.673	0.213	7.06
Final	576	0.90	10	0.1436	0.673	0.213	14.23
Note: CS_Rsq : Composite Squared R							
Max_Rsq : Maximum R Squared							
EPP : Expected Prediction Error							

Derivation and Validation of ML Models

This study employed a three-phase approach to predictor selection, incorporating variables demonstrating statistical significance in univariate analysis, features selected through Least Absolute Shrinkage and Selection Operator (LASSO) regression, and clinically relevant factors identified by cardiovascular specialists. The selected variables have been empirically validated in previous clinical investigations as robust predictors of Major Adverse Cardiovascular Events (MACEs), thereby ensuring both statistical rigor and clinical relevance in model construction.

Hyperparameter optimization was performed using systematic grid search methodology. To evaluate model performance and generalizability, the dataset was partitioned into training (70%) and testing (30%) subsets. Internal validation was conducted through bootstrap resampling, consistent with TRIPOD[8] (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) guidelines. Additionally, 5-fold cross-validation was implemented to further assess model stability through iterative training on four subsets and validation on the remaining subset.

Comprehensive model evaluation was conducted across three critical dimensions:

Discrimination: Quantified using the area under the Receiver Operating Characteristic curve (AUC), with superior performance indicated by higher values.

Calibration: Assessed through calibration plots visualizing agreement between predicted probabilities and observed outcomes. Supplementary metrics including Accuracy (ACC) and F1-score were employed, alongside Decision Curve Analysis (DCA) for clinical utility assessment.

Overall Performance: Evaluated through comprehensive metrics including AUC, Positive Predictive Value (PPV), Negative Predictive Value (NPV), ACC, and F1-score on the independent test set.

To enhance model interpretability, Shapley Additive Explanations (SHAP) were implemented for feature importance visualization. Furthermore, an interactive web-based prediction tool was developed using the R Shiny framework to facilitate clinical implementation and point-of-care risk stratification.

Data availability

The datasets generated and/or analysed during the current study are not publicly available due to patient privacy and institutional regulations but are available from the corresponding author on reasonable request. All data have been de-identified prior to analysis.

Statistical analysis

Data statistical analysis was performed using R (version 4.2.3) and Python 3.12. First, the normality of measurement data was assessed using the Shapiro-Wilk test. Quantitative data conforming to a normal distribution were expressed as mean $\pm$ standard deviation, and comparisons of means between two groups were conducted using the independent samples t-test. For quantitative data that did not follow a normal distribution, values were presented as median (interquartile range), and between-group comparisons were performed using the Mann-Whitney U test. Categorical data were expressed as percentages or rates, and comparisons of rates between groups were carried out using the chi-square test (χ^2 test).

Results

Results of Multicollinearity Analysis among Independent Variables

By examining the color blocks corresponding to each variable in Fig. 1, the Pearson correlation coefficient between *"24h average ambulatory systolic blood pressure"* and *"daytime ambulatory systolic blood pressure"* is 0.892, while that between *"24h average ambulatory diastolic blood pressure"* and *"daytime ambulatory diastolic blood pressure"* is 0.855 (Fig. 1). Since both correlation coefficients exceed

0.8, these variable pairs are considered highly linearly correlated. One variable from each pair was considered for removal.

After removing the redundant variables, a multicollinearity test was performed on the remaining independent variables. The results showed that all variance inflation factor (VIF) values were below 5, indicating no severe multicollinearity. Thus, the dataset is suitable for subsequent regression analysis. Detailed multicollinearity test results are presented in Table 2.

Table 2
Results of Multicollinearity Test

Age	VIF		VIF
	1.53	Fasting blood glucose	1.55
Gender	2.31	TC	1.35
BMI	1.09	TG	1.32
Salt Intake Score	1.15	LDL-C	1.45
Marital status	1.09	HDL-C	1.15
Alcohol Consumption	1.30	LVEF	1.04
Smoking Status	1.25	Office SBP	2.07
Heart rate	1.06	Office DBP	2.09
Diabetes Mellitus	1.55	Daytime DBP (ambulatory)	2.33
Thyroid Dysfunction	1.05	Nighttime DBP (ambulatory)	2.03
Serum potassium	1.08	Nighttime SBP (ambulatory)	1.13
Serum sodium	1.03	Serum creatinine	4.89
eGFR	4.50	Hcy	1.13
Uric acid	1.43	24-hour average SBP (ambulatory)	2.04

Baseline characteristics

From January 2020 to December 2024, a total of 1,180 patients were enrolled in this prospective cohort study at the Hypertension Standardization Center of Altay City People's Hospital. During a median follow-up period of 1,176 days, 327 major adverse cardiovascular events (MACEs) were recorded, resulting in a cumulative incidence rate of 27.7%. The cohort was randomly divided into a training set (n = 826) and a validation set (n = 354). The training set demonstrated a 26.5% incidence of MACEs (219 cases) over a median follow-up of 1,183 days, while the validation set showed a 30.5% incidence (108 cases) over a median follow-up of 1,164 days. No significant difference in MACEs incidence was observed between the two cohorts (Log-rank $P = 0.26$).

Table 3 displayed the baseline clinical characteristics of hypertension patients. The study population had a mean age of 61 ± 11.76 years, including 510 males (43.2%) and 670 females (56.8%). Among them, 1,009 (85.5%) were married, 97 (8.2%) were smokers, and 150 (12.7%) reported alcohol consumption. Body mass index (BMI) distribution was as follows: $<18.5 \text{ kg/m}^2$ ($n = 17$, 1.4%), $18\text{-}23.9 \text{ kg/m}^2$ ($n = 225$, 19.1%), and $> 23.9 \text{ kg/m}^2$ ($n = 938$, 79.5%). The median salt intake score was 20 (IQR: 15.75–30.00). Univariate analysis demonstrated significant differences between patients with and without MACEs. The MACEs group was older (64.72 ± 9.61 vs. 60.55 ± 11.63 years, $P < 0.001$), had a higher proportion of females (61.5% vs. 55.0%, $P = 0.044$), higher smoking rate (13.5% vs. 6.2%, $P < 0.001$), and elevated salt intake score (23.4 vs. 20, $P < 0.001$). No significant differences were observed in marital status, alcohol consumption, or BMI distribution ($P > 0.05$). Among comorbidities, diabetes was significantly more prevalent in the MACEs group (11.3% vs. 6.9%, $P = 0.013$), while thyroid dysfunction showed no significant association.

Table 3
Baseline Characteristics of Study Subjects in This Chinese Kazakhs

Characteristics		Non-MACEs(n = 853)	MACEs(n = 327)	t/Z/ χ^2	P
Age (years)	61.71 ± 11.76	60.55 ± 11.63	64.72 ± 9.61	-5.78	<0.001
Gender				4.052	0.044
Male	510(43.2%)	384(45.0%)	126(38.5%)		
Female	670(56.8%)	469(55.0%)	201(61.5%)		
Marital Status				2.723	0.256
Married	1009(85.5%)	733(85.9%)	276(84.4%)		
Unmarried	117(9.9%)	78(9.1%)	39(11.9%)		
Other	54(4.6%)	42(4.9%)	12(3.9%)		
Smoking Status				16.434	<0.001
Yes	97(8.2%)	53(6.2%)	44(13.5%)		
No	1083(91.8%)	800(93.8%)	283(86.5%)		
Alcohol Consumption				2.106	0.147
Yes	150(12.7%)	101(11.8%)	49(15.0%)		
No	1030(87.3%)	752(88.2%)	278(85.0%)		
BMI(kg/m ²)				1.662	0.436
<18.5	17(1.4%)	10(1.2%)	7(2.1%)		
18.5–23.9	225(19.1%)	165(19.3%)	60(18.3%)		
>23.9	938(79.5%)	678(79.5%)	260(79.5%)		
Salt Intake Score	20 (15.75-30)	20(15–30)	23.4(18–29)	-4.09	<0.001
Diabetes Mellitus				6.118	0.013
Yes	96(8.1%)	59(6.9%)	37(11.3%)		
No	1084(91.9%)	794(93.1%)	290(88.7%)		
Thyroid Dysfunction				0.015	0.902

Characteristics		Non-MACEs(n = 853)	MACEs(n = 327)	t/Z/ χ^2	P
Yes	92(7.8%)	66(7.7%)	26(8.0%)		
No	1088(92.2%)	787(92.3%)	301(92.0)		

Significant differences were observed in several metabolic and cardiovascular parameters. Patients with MACEs exhibited higher fasting blood glucose (5.66 ± 1.56 vs. 5.47 ± 1.26 mmol/L, $P = 0.043$), elevated homocysteine (11.43 ± 6.59 vs. 10.65 ± 4.7 μ mol/L, $P = 0.049$), increased total cholesterol (4.97 ± 1.3 vs. 4.74 ± 1.42 mmol/L, $P = 0.009$), higher LDL-C (3.05 ± 1.29 vs. 2.76 ± 1.54 mmol/L, $P < 0.001$), and higher office systolic blood pressure (154 vs. 150 mmHg, $P < 0.001$). No statistically significant differences were detected in heart rate, serum sodium, creatinine, uric acid, triglycerides, HDL-C, left ventricular ejection fraction (LVEF), office diastolic blood pressure, or any parameters of ambulatory blood pressure monitoring ($P > 0.05$). as detailed in Table 4.

Table 4
Baseline Clinical Characteristics of Study Subjects in This Chinese Kazakhs

Characteristics		Non-MACEs(n = 853)	MACEs(n = 327)	t/Z	P
Heart rate (bpm)	80.03 ± 11.97	80.11 ± 12.36	79.8 ± 10.91	0.42	0.672
Serum potassium(mmol/L)	4.08 ± 0.41	4.07 ± 0.42	4.12 ± 0.36	-1.96	0.051
Serum sodium (mmol/L)	141.06 ± 4.83	141.07 ± 4.89	141.03 ± 4.69	0.12	0.904
Fasting blood glucose (mmol/L)	5.52 ± 1.36	5.47 ± 1.26	5.66 ± 1.56	-2.03	0.043
Hcy (μmol/L)	10.86 ± 5.3	10.65 ± 4.7	11.43 ± 6.59	-1.97	0.049
TC (mmol/L)	4.81 ± 1.39	4.74 ± 1.42	4.97 ± 1.3	-2.63	0.009
TG (mmol/L)	1.44 ± 0.92	1.43 ± 0.91	1.48 ± 0.96	-0.75	0.454
LDL-C (mmol/L)	2.84 ± 1.48	2.76 ± 1.54	3.05 ± 1.29	-3.19	<0.001
HDL-C (mmol/L)	1.4 ± 0.63	1.4 ± 0.69	1.39 ± 0.42	0.21	0.834
LVEF (%)	65(62–69)	65(62–69)	65(63–68)	0.42	0.674
eGFR (mL/min/1.73m ²)	102.3(86.6-120.8)	103.6(86.67-121.27)	99.73(85.92-119.61)	1.18	0.237
Serum creatinine(μmol/L)	62.5(53.3–72.4)	62.9(53.0-73.3)	61.0(54.1–71.0)	0.8	0.424
Uric acid (μmol/L)	287.1(238.5-346.5)	286.19(234.50-345.13)	288.50(241.92-350.06)	1.16	0.248
SBP (mmHg)	150(136–168)	150(135–165)	154(139–173)	-3.13	<0.001
DBP (mmHg)	89(80–100)	89(80–100)	88(80–99)	0.6	0.549
24h average DBP (mmHg)	87(79–95)	87(79–95)	88(79–95)	-1.23	0.218
24h average SBP (mmHg)	142(133–152)	142(132–152)	145(133–154)	-1.2	0.231
Daytime SBP (mmHg)	146(135–157)	145(135–158)	149(135–157)	-0.78	0.436
Daytime DBP (mmHg)	89(80–99)	89(80–99)	90(79–99)	-0.28	0.779
Nighttime DBP (mmHg)	85(78–94)	85(78–94)	86(78–94)	-0.62	0.538
Note: *P < 0.05					

Characteristics		Non-MACEs(n = 853)	MACEs(n = 327)	t/Z	P
Nighttime SBP (mmHg)	140(130–150)	140(130–150)	140(130–150)	-0.39	0.697
Note: *P < 0.05					

Feature Selection for Predictor Screening

Predictor selection was performed based on the union of results from univariate analysis and LASSO regression(Fig. 2), as detailed in Table 5–6. Although serum potassium was not selected in the univariate analysis, it was retained in the LASSO regression and was further supported by clinical expertise regarding its relevance to cardiovascular events. Therefore, serum potassium was included in the final predictor set. The final selected predictors included: age, diabetes status, serum potassium, fasting blood glucose, Hcy, TC, LDL-C, and office systolic blood pressure. These variables were identified through a combination of univariate analysis, LASSO regression screening, and clinical expert evaluation, and were ultimately used as input variables for the prediction model of MACEs in Kazakh hypertensive patients.

Table 5
Univariate Analysis Results of Related Data

	Regression Coefficient	HR(95%CI)	P
Age	0.029	1.029(1.019 ~ 1.039)	<0.001
Gender	0.287	1.332(1.066 ~ 1.665)	0.012
BMI	-0.662	0.516(0.236 ~ 1.129)	0.098
Salt intake score	-0.002	0.998(0.993 ~ 1.003)	0.430
Marital status	0.284	1.328(0.95 ~ 1.859)	0.097
Alcohol consumption	0.131	1.140(0.841 ~ 1.547)	0.398
Smoking status	0.642	1.900 (1.382 ~ 2.611)	<0.001
Heart rate	-0.004	0.996(0.987 ~ 1.005)	0.360
Diabetes mellitus	0.363	1.438 (1.02 ~ 2.028)	0.038
Thyroid dysfunction	-0.105	0.900(0.603 ~ 1.345)	0.608
Serum potassium	-0.193	0.824(0.622 ~ 1.093)	0.179
Serum sodium	-0.003	0.997(0.973 ~ 1.021)	0.797
eGFR	-0.005	0.995(0.991 ~ 0.999)	0.009
Serum creatinine	-0.001	0.999(0.993 ~ 1.006)	0.842
Hcy	0.020	1.020(1.005 ~ 1.036)	0.011
Uric acid	0.000	1.000(0.999 ~ 1.001)	0.917
Fasting blood glucose	0.089	1.093(1.020 ~ 1.172)	0.012
TC	0.078	1.081(1.015 ~ 1.152)	0.016
LDL_C	0.068	1.070(1.022 ~ 1.120)	0.004
HDL_C	0.015	1.015(0.855 ~ 1.204)	0.866
TG	0.022	1.022(0.919 ~ 1.138)	0.688
LVEF	0.000	1.000(0.981 ~ 1.019)	0.965
Office SBP	0.008	1.008(1.003 ~ 1.012)	0.001
Office DBP	-0.003	0.997(0.989 ~ 1.005)	0.479
24h average DBP	0.006	1.006(0.997 ~ 1.016)	0.201
Nighttime SBP	0.000	1.000(0.997 ~ 1.003)	0.999

	Regression Coefficient	HR(95%CI)	P
Age	0.029	1.029(1.019 ~ 1.039)	<0.001
Nighttime DBP	0.003	1.003(0.994 ~ 1.013)	0.489
Daytime DBP	0.001	1.001(0.992 ~ 1.010)	0.864

Table 6
Relevant Feature Variables Identified by Lasso Regression

Intercept	β		β
	-4.494	Diabetes mellitus	0.246
Age	0.029	Serum creatinine	-0.007
Serum potassium	0.056	Fasting blood glucose	0.014
Hcy	0.016	LDL-C	0.063
TC	0.061	Office SBP	0.004

Model Development and Evaluation

The predictive performance of the XGBoost, Random Forest, COX regression, and Decision Tree models was evaluated based on the training dataset. Results (see Table 8 and Fig. 4–15) showed that the XGBoost model achieved the highest AUC value (0.771), though its precision (PPV = 0.149) and F1-score (0.214) were relatively low. The Decision Tree model attained the highest accuracy (ACC = 0.796), negative predictive value (NPV = 0.938), and F1-score (0.536) among all models, along with the second-highest AUC (0.752). The Random Forest model demonstrated a high NPV (0.923) but low precision (PPV = 0.127) and F1-score (0.191). The COX regression model showed the highest precision (PPV = 0.645), yet the lowest accuracy (ACC = 0.580) and AUC (0.635), indicating relatively weak generalizability, potentially due to limitations of linear assumptions in complex data. Overall, the Decision Tree model exhibited the best comprehensive performance on the training set.

Table 8
Performance Evaluation of Four Models in the Training Set

Model	TN	FN	FP	TP	PPV	NPV	ACC	F1score	AUC
XGBoost	542	56	194	34	0.149	0.906	0.697	0.214	0.771
Random Forest	552	46	199	46	0.127	0.923	0.703	0.191	0.610
COX	332	266	81	147	0.645	0.555	0.580	0.459	0.635
Decision tree	561	37	131	97	0.425	0.938	0.796	0.536	0.752

In the validation set (Table 9, Fig. 3), the XGBoost model achieved the highest AUC (0.829) and high accuracy (ACC = 0.788), though its precision remained low (PPV = 0.242), suggesting a high false positive rate in predicting positive cases. Its NPV reached the maximum value of 1.000. The Decision Tree model ranked second in AUC (0.776), while achieving the highest accuracy (ACC = 0.797) and F1-score (0.550), along with relatively high precision (PPV = 0.444), indicating a balanced performance in both positive and negative predictions. It also maintained a high NPV (0.933), reflecting stable overall performance. The Random Forest model performed poorly with the lowest AUC (0.577), accuracy (ACC = 0.701), precision (PPV = 0.121), and F1-score (0.185), indicating limited generalizability, though its NPV (0.925) retained some clinical reference value. The COX model showed the highest precision (PPV = 0.677) but the lowest accuracy (ACC = 0.599), AUC (0.668), and NPV (0.569), reflecting a notably weak ability to predict negative cases.

Table 9
Performance Evaluation of Four Models in the Validation Set

Model	TN	FN	FP	TP	PPV	NPV	ACC	F1score	AUC
XGBoost	255	0	75	24	0.242	1	0.788	0.390	0.829
Random Forest	236	19	87	12	0.121	0.925	0.701	0.185	0.577
COX	145	110	32	67	0.677	0.569	0.599	0.486	0.668
Decision tree	238	17	55	44	0.444	0.933	0.797	0.550	0.776

Across both training and validation sets, the Decision Tree model demonstrated robust and consistent performance, with higher ACC, F1-score, and NPV compared to other models. Although XGBoost showed competitive accuracy and AUC on the training set, its weaker performance on positive-class prediction led to suboptimal validation results. Both COX regression and Random Forest exhibited varying degrees of performance degradation in validation, with Random Forest showing the poorest generalizability.

Decision Curve Analysis (DCA) was used to evaluate the clinical utility of the prediction models by quantifying net benefit across different threshold probabilities. The DCA curves (Fig. 4–17) showed that the Decision Tree model provided stable net benefit in the low to intermediate threshold probability range (0%–50%) and remained clinically useful even at intermediate to high thresholds (50%–75%). Although XGBoost and Random Forest showed acceptable performance within certain threshold ranges, their utility decreased at higher thresholds. The COX regression model had a narrow range of applicability, showing clinical value only at very low thresholds.

The calibration performance of the model was evaluated on the training set using Bootstrap resampling. The mean absolute error (MAE) between the calibration curve and the ideal curve was 0.020 (Fig. 4), indicating good agreement between predicted and observed outcomes. To further assess potential overfitting, the model was validated on the independent validation set. The resulting calibration curve showed an MAE of 0.021 (Fig. 4), demonstrating consistent calibration performance. The Decision Tree model exhibited stable performance across both training and validation sets, with minimal errors in

calibration. Moreover, the model error remained largely unchanged after Bootstrap correction, confirming its reliability and strong generalizability.

In conclusion, the Decision Tree model demonstrated superior stability, predictive performance, and clinical utility, making it more suitable for practical clinical application.

Model Interpretability

The SHAP summary plot (Fig. 5) revealed the following order of feature importance in predicting MACEs: LDL-C, age, office systolic blood pressure, serum potassium, TC, fasting blood glucose, Hcy, and diabetes status. LDL-C was identified as the most influential predictor, indicating its critical role in MACEs occurrence. In contrast, diabetes status had SHAP values clustered near zero, suggesting a minimal impact on model predictions. Elevated values of LDL-C, age, systolic blood pressure, Hcy, and fasting blood glucose, along with decreased serum potassium levels, were associated with increased model output values, reflecting higher predicted risk of MACEs.

Figure 6 Individual prediction explanation for a sample patient. The plot shows how different features, such as elevated Fasting Blood Glucose (7.81 mmol/L) and Total Cholesterol (5.32 mmol/L), push the prediction toward higher risk, while older Age (81 years) and Potassium (4.51 mmol/L) reduce the risk. The final prediction combines all these contributions.

Based on the SHAP importance ranking, the four most influential predictors—LDL-C, age, office systolic blood pressure, and serum potassium—were further analyzed. Boxplots comparing these variables between patients with and without MACEs (Fig. 7–8) confirmed significant differences in their distributions ($P < 0.05$).

SHAP dependence plots (Fig. 9) provided additional insights into feature effects: Lower values (near 0 mmol/L) were associated with negative SHAP values, indicating reduced MACEs risk. Risk increased progressively with higher LDL-C levels. Younger individuals (around 40 years) showed negative SHAP values, suggesting lower risk. Risk increased steadily with advancing age. Values near 100 mmHg corresponded to negative SHAP values, indicating protective effects against MACEs. Risk elevated with increasing blood pressure. Concentrations near 3.0 mmol/L contributed positively to risk prediction, suggesting that hypokalemia may increase the likelihood of arrhythmias and other adverse events. Higher potassium levels were associated with reduced risk.

To increase the convenience and operability of the model, the website "shin yapps" in R software at <http://127.0.0.1:6626> was used. After entering the patient's data in the variable column on the left side of the webpage, the predicted probability of MACEs for the patient was displayed in the prediction column on the right side of the webpage., in order of importance. Figure 10 shows the predicted probability of MACEs in a patient aged 50 years, without diabetes, with a blood potassium concentration of 4 mmol/L, Hcy 10 μ mmol/L, LDL-C 3 mmol/L, TC 5 mmol/L, fasting glucose 5.5 mmol/L, and in-office blood pressure of 120 mm Hg, which is predicted to be a low risk, with a probability of developing MACEs of 25.77%.

Discussion

Medical resources are limited in the Altay region of Xinjiang, and previous studies have found a high risk of cardiovascular morbidity in this population[9]. In the Global Burden of Disease Study 2019 analysis[10], hypertension was listed as one of the major risk factors for adverse cardiovascular events, and the high prevalence and uncontrolled rate of hypertension lead to a large number of cardiovascular deaths and disabilities, especially in the elderly. A serious lack of knowledge, treatment, and control rates among Kazakh herders in mountainous areas was found in previous studies[11], and the survey found that the blood pressure control rate of hypertension among Kazakhs was only 8.6%[12], and nomadic groups tend to lack systematic medical care and health management, which makes it difficult to prevent and treat chronic diseases such as hypertension at an early stage. The relative lack of medical resources in agricultural and pastoral areas, the high cost of medical care for residents, and the low accessibility of health services have resulted in many hypertensive patients failing to receive timely and effective management and intervention. In addition, long-term hypertension is associated with a higher incidence of cardiovascular events. Hypertension being a major risk factor for cardiovascular diseases, patients with chronic hypertension are at increased risk of major cardiovascular events such as heart disease and stroke, hence the higher incidence of cardiovascular events observed in the study. Meanwhile, because Altay region is located in the northernmost part of Xinjiang, medical resources are relatively scarce, and there are fewer opportunities for medical interventions and early diagnosis, which may further exacerbate the occurrence of cardiovascular events.

The higher proportion of women in the MACEs group (61.5% vs. 55.0%, $P = 0.044$) is similar to the results of a study conducted in northern China[13], which may be related to the increased risk of cardiovascular disease in women after menopause. According to a study by Izumi et al[14], the age of menopause in Kazakh women is 43.7 years, which is lower than that of Han Chinese women. Furthermore, related studies have shown[15] that the risk of atherosclerosis and cardiovascular disease increases significantly in women due to the decrease in estrogen levels after menopause, and the prevalence of smoking was higher in the group with MACEs than in the group without MACEs (13.5% vs. 6.2%, $p < 0.001$). Smoking increases the risk of cardiovascular disease through several mechanisms (e.g., promoting atherosclerosis, increasing blood viscosity) [16]. The Framingham study[17] also demonstrated that smoking status was an independent risk factor for MACEs and that smokers had a significantly increased risk of developing MACEs. The salt-scoring questionnaire scores were higher in the MACEs group than in the non-MACEs group (23.4 vs. 20.0, $P < 0.001$), suggesting that a high-salt diet may be associated with the development of MACEs. The Kazakhs traditionally relied on animal husbandry, with meat being the main food source. Salt plays a key role in curing and preserving meat (e.g., air-dried meat, bacon) to prevent spoilage and extend shelf life. Salt is also used in the preparation of fermented dairy products, such as yoghurt and cheese, to help flavour and preserve them. In the Kazakh diet, plain water is consumed less, and milk tea, to which salt is often added, is the main daily drink. This dietary habit makes the salt intake of the Kazakh population generally higher than that of the Han Chinese population. A high-salt diet leads to sodium retention, increasing blood volume and vascular resistance, which in turn elevates blood pressure and increases the risk of cardiovascular disease[18]. It has been shown[19] that

salt intake of more than 6 g per day (equivalent to 24-hour urinary sodium > 100 mmol/d) is associated with the risk of developing MACEs. The proportion of patients with MACEs who developed MACEs in Kazakh hypertension in this study who also had diabetes mellitus was higher than the proportion of patients who did not develop MACEs. A study conducted in northern and northeastern China[20] found that the combination of hypertension and diabetes resulted in more severe cardiovascular outcomes than hypertension and diabetes alone, and the synergistic effect of diabetes combined with hypertension further exacerbated the risk of cardiovascular disease compared with diabetic patients with normotensive blood pressure as demonstrated in the Framingham study[21], in which diabetic patients are often associated with dyslipidemia and obesity, and these metabolic disorders are associated with a higher risk of cardiovascular disease than in patients without MACEs. Diabetic patients are often associated with dyslipidemia and obesity, and these metabolic disorders interact with hypertension to form a vicious circle and increase the incidence of MACEs[21]. In this study, we found that there was a statistical difference between fasting blood glucose, TC, LDL-C, Hcy levels, and office blood pressure systolic blood pressure in the presence or absence of MACEs. Among the investigators who developed MACEs, higher fasting blood glucose levels, higher Hcy, higher TC levels, higher LDL-C levels, and higher in-office systolic blood pressure have been shown to be an independent risk factor for MACEs, and high blood glucose leads to vascular endothelial damage and promotes the development and progression of atherosclerosis, which in turn increases the risk of MACEs[22]. Patients in the MACEs group had higher levels of TC and LDL-C than those in the non-MACEs group, and TC and LDL-C are key risk factors for atherosclerosis. High levels of LDL-C promote the deposition of lipids in the vessel wall and the formation of atherosclerotic plaques, which in turn increases the risk of MACEs[23]. In a study conducted in Xinjiang Production and Construction Corps[24], combined dyslipidemia in hypertensive patients was found to be the most important risk factor for adverse cardiovascular events. In addition, high TC levels also increase blood viscosity, further exacerbating the risk of cardiovascular disease. Elevated systolic blood pressure is an important risk factor for cardiovascular disease. High systolic blood pressure increases cardiac load, leading to myocardial hypertrophy and heart failure, as well as accelerating the process of atherosclerosis[25]. It has been shown that the risk of cardiovascular events approximately doubles for every 20 mmHg increase in systolic blood pressure[26].

High homocysteine levels significantly increase the risk of cardiovascular disease by damaging the vascular endothelium, promoting thrombosis, and increasing oxidative stress[27]. The Framingham study also showed[28] that elevated Hcy levels are an independent risk factor for MACEs. This sclerotic process is further exacerbated with age and hypertension exacerbates the degradation of elastin, an increase in the proportion of collagen, stiffening of the vessel wall, and further expansion of pulse pressure[29]. Smooth muscle cells in the arteries of hypertensive patients of advanced age are more likely to calcify, accelerating the narrowing of the arterial lumen and the rise in resistance to blood flow[30]. Structural damage to the cerebral and cardiovascular vasculature is more pronounced with age in hypertensive patients, and tiny arteries in the brain are prone to rupture or blockage in elderly hypertensive patients, increasing the risk of stroke[31]. Hypertension causes the heart to fight against high afterload for a long time, triggering left ventricular hypertrophy[32]. With age: ventricular compliance

decreases: the left ventricle stiffens and diastolic underfilling further leads to a decrease in cardiac output and induces heart failure. Myocardial fibrosis is higher in elderly hypertensive patients[33], with reduced cardiac compensatory capacity, which predisposes them to adverse cardiovascular events.

The traditional Kazakh diet is based on meat (e.g., beef, lamb, and horse meat) and dairy products (e.g., cream, cheese, and yoghurt), which are rich in saturated fat and cholesterol. Chronic high-fat diets may lead to elevated levels of LDL-C in the blood. High levels of LDL are a key risk factor for the development and progression of atherosclerosis. When LDL-C is elevated in the blood, LDL particles tend to enter the subendothelium of blood vessels where they are modified by oxidation. Oxidised LDL is phagocytosed by macrophages, forming foam cells. The continuous accumulation of these foam cells then forms the lipid core of the atherosclerotic plaque [34, 35]. As the plaque continues to develop, it may lead to narrowing of blood vessels and affect blood supply. When the plaque ruptures, it activates the coagulation system and forms a thrombus[23], which can lead to adverse cardiovascular events such as acute myocardial infarction and stroke. A significant correlation between high LDL-C levels and increased CVD risk and mortality has also been found in related studies[36]. Liang Chen et al[36, 37] found and constructed a CVD risk prediction model around the relationship between LDL and CVD risk and mortality in the Chinese population, in which LDL- C was included as an important risk factor. This is consistent with the inclusion of LDL as a predictor in this study and provides theoretical support for this study. Some Kazakh populations in remote areas have insufficient knowledge of the dangers of hypertension and fail to monitor and control their blood pressure in a timely manner. And medical resources may be relatively scarce, resulting in untimely screening, diagnosis, and treatment of hypertension as well as failure to adhere to regular medication. Systolic blood pressure is the maximum pressure generated by the left ventricle of the heart to pump blood into the aorta and the systemic circulatory system during the contraction phase of the heart. Blood pressure systolic blood pressure is an important indicator used to assess the cardiovascular risk of a patient, and high systolic blood pressure means that the heart needs to generate more pressure during contraction in order to push the flow of blood, which will exacerbate the heart's load, and put more pressure on the walls of blood vessels to accelerate atherosclerosis and damage to the inner wall of the vessels. This increases the load on the heart, puts more pressure on the walls of the blood vessels, accelerates atherosclerosis, damages the lining of the blood vessels, and can lead to cardiovascular and cerebrovascular events if high systolic blood pressure is chronic. In a large cohort study[38] examining the effects of systolic and diastolic blood pressure on cardiovascular outcomes, it was found that elevated systolic blood pressure had a greater impact on adverse cardiovascular events than diastolic blood pressure. The traditional diet of nomadic people is usually high in calories and excessive intake may lead to obesity, which in turn affects cholesterol metabolism. Relevant studies have shown that there is a positive correlation between TC levels and the occurrence of cardiovascular diseases such as atherosclerosis and coronary heart disease[39, 40]. Elevated TC levels are an important risk factor for atherosclerosis [41], which is the main cause of coronary heart disease, heart attacks and strokes[42]. This is due to the fact that when TC is too high, it may be deposited on the walls of blood vessels, which in turn form atherosclerotic plaques. These plaques then lead to narrowing of the arteries and obstruction of blood flow, which can lead to a reduction in blood supply to the heart

and brain[43], thus increasing the risk of cardiovascular events. Atherosclerosis causes the walls of blood vessels to thicken and lose elasticity, and as plaque forms, the inner lumen of the blood vessel gradually narrows, and this narrowing can lead to obstruction of blood flow and insufficient blood supply to the heart, as well as to some organs, for sufficient oxygen and nutrients, especially when the body needs more oxygen [44, 45], such as during exercise or emotional stress, when the heart's need for oxygen increases, and the narrowed blood vessels may not be able to provide enough blood flow, and this is when chest pain, called angina, occurs. If an atherosclerotic plaque ruptures or a blood clot forms at the site of the rupture, this can lead to a complete blockage of the blood vessel, severely interrupting blood flow. Myocardial infarction may occur as the heart muscle is unable to get enough oxygen and the heart tissue begins to die. The prevalence of combined hypokalaemia in our hypertensive patients is more than 8%[2], and recent studies have shown that low blood potassium levels may be associated with an increased cardiovascular risk in hypertensive patients, especially in those who experience adverse cardiovascular events such as heart disease, heart failure or stroke. Fluctuations in blood potassium concentration may affect the regulatory mechanisms of blood pressure[3, 46], and low blood potassium may lead to increased reabsorption of sodium ions in the kidneys, which in turn may lead to an increase in blood pressure. Moreover, the effect of blood potassium on vascular smooth muscle may alter vascular contractility, which in turn affects hemodynamics. Blood potassium is the major intracellular cation involved in the electrical activity of the heart and regulates the action potential of cardiomyocytes. Low blood potassium can lead to rapid depolarization of cardiomyocytes, making the heart more excitable and leading to arrhythmias, particularly ventricular arrhythmias. These arrhythmias can lead to fatal cardiac events such as cardiac arrest or atrial fibrillation[47]. Hcy is a Sulphur-containing amino acid, and primary hypertension with Hcy levels higher than $15 \mu\text{mol/l}$ is known as H-type hypertension[1]. Relevant studies have found that high Hcy is an important independent risk factor for cardiovascular and cerebrovascular diseases[48] and there seems to be a synergistic effect between Hcy and hypertension, especially in hypertensive patients where the combination of high Hcy may increase the incidence of MACEs[49], which is in line with the results of this study. This study was conducted on Kazakh hypertensive patients, and the population included in this study included patients from the township health center of Hong dun Town, Qiemu'erqieke Town, and Aweitan Town in Altay prefecture, where Kazakhs live a nomadic lifestyle, and the diet of Kazakhs in the Xinjiang region consists mainly of pasta as well as beef and lamb, with less intake of vegetables and fruits in the herding lifestyle, and relevant studies have found that the metabolic process of Hcy is closely related to vitamins B6, B12, and folic acid[49]. Related studies have found that Hcy metabolism is closely related to vitamins B6, B12, and folic acid, which are water-soluble vitamins that are easily lost in the body, and that low intake of fruits and vegetables and low folic acid content in the body can lead to a continuous accumulation of Hcy, which ultimately leads to the occurrence of MACEs[50]. In addition to this high Hcy levels are thought to damage vascular endothelial cells through a variety of mechanisms. Endothelial cells are a key component of the blood vessel wall and are responsible for regulating blood flow and blood pressure. High Hcy levels reduce the production of nitric oxide (NO), an important factor in vasodilation, leading to vasoconstriction and reduced blood flow, further increasing blood pressure and exacerbating atherosclerosis[50]. Fasting blood glucose is considered an important independent

predictor of cardiovascular risk, especially in patients with prediabetes or type 2 diabetes. Studies have shown[51–53] that even in the absence of diabetes, mild elevations in fasting blood glucose are associated with the occurrence of cardiovascular events. This is because damage to the vascular endothelium from hyperglycaemia, increased blood viscosity, activation of inflammatory responses and altered lipid metabolism[54] may directly promote atherosclerosis and thrombosis, which may in turn increase the incidence of cardiovascular disease.

In this study, the COX regression model, random forest, decision tree, and XGBoost were trained respectively, and the grid search method was applied for tuning. The selection of models as well as the tuning took into full consideration the characteristics of the study data as well as the advantages and limitations of different models. The COX regression model was chosen based on the fact that the data in this study were follow-up data that needed to be analysed for the relationship between survival time and multiple covariates, and the advantage of the COX regression model was that it did not require strict assumptions about the distribution of survival time, and it was also able to flexibly incorporate multiple influencing factors [55], which ensured that the model could be interpreted while increasing the accuracy of the prediction. The COX regression model in this study was significantly underpowered for positive class prediction, with poor overall performance. This may be due to the fact that COX regression requires a sufficient number of events, and if the positive class sample is small, it is difficult for the model to accurately estimate the risk ratio, or it may be related to the fact that the predictors included in this study did not fully respond to the key factors affecting the survival time, and the predictive performance of the model will be limited. The single learner model was chosen as a decision tree, the decision tree model is a supervised learning algorithm based on tree structure[56, 57], the tree structure is intuitive, which can help to construct the model after the interpretation of the complexity parameter ($cp = 0.006$) after the decision tree in this study, which provides a good balance for the model, both to avoid overfitting and to maintain a better fitting ability, after the adjustment parameter decision tree model performance is very good. In addition, Random Forest and XGBoost, which are commonly used in integrative learning, are also selected in this study. Random Forest can effectively reduce the risk of overfitting of a single decision tree by constructing multiple decision trees to integrate its prediction results[57]. In this study after the tuning parameter $n_estimators = 500$, the number of trees increases, the more the number in the random forest the more it can reduce the risk of overfitting ground. In this study the random forest model performs poorly, the overall performance is not good, the AUC value is 0.60, it may be due to the small sample of positive classes in this study, the random forest tends to predict the majority of the classes, or the parameters may not be optimal. [57]XGBoost is an efficient gradient boosting tree algorithm[58], XGBoost has the highest AUC value in the training set and validation set but its precision and F1 value is lower, AUC value is the overall performance of the model, while accuracy and F1 value depends on a specific threshold, if the threshold is not chosen well the model may not perform well.

This study compared the performance of traditional methods (COX regression) with machine learning models (decision tree, random forest, XGBoost) in the prediction of adverse cardiovascular events in Kazakh hypertensive patients. COX regression model showed poor prediction in this study. In contrast, machine learning models demonstrated greater flexibility: the decision tree model had an intuitive

structure and performed well after parameter tuning; the random forest model theoretically reduced overfitting, but the actual AUC was only 0.60, which may be related to data characteristics or unoptimised parameters; the XGBoost model had the highest AUC value but lower accuracy and F1 values, which may be affected by the category imbalance or threshold selection. The decision tree model performs second only to the XGBoost model in terms of accuracy rate and AUC value in the training set and validation set, and outperforms the random forest and COX regression models. Its accuracy (PPV = 0.444) and F1 value (0.550) are the highest, showing a more balanced performance in positive and negative class prediction. In addition, the decision tree model had a higher negative class prediction value (NPV = 0.933), which can effectively identify low-risk patients and reduce unnecessary medical interventions. Although the XGBoost model had the highest AUC value and the maximum negative class prediction value, it had a lower precision rate, indicating a high false positive rate in predicting positive class samples. The Random Forest and COX regression models were significantly deficient in positive class prediction and had poor overall performance. The balanced performance of the decision tree model in terms of accuracy, AUC value, precision rate, and F1 value makes it highly applicable in clinical practice, which can meet the needs of identifying high-risk patients and screening low-risk patients at the same time. In addition, the performance of the decision tree model is more stable in both the training and validation sets when drawing Calibration curves, and the error of the calibration curves is smaller, and the error of the model after Bootstrap calibration is further reduced, which verifies the reliability and generalisation ability of the model. In the clinical utility evaluation of the decision tree model in this study, it showed stable net gain in the low to medium-high threshold range (0%-75%) and was close to the ideal curve, indicating its strong predictive ability and practical clinical applicability, and the decision tree was able to effectively identify high-risk patients with MACEs by integrating multiple risk factors, which supported personalised interventions, and moreover, the calibration curve of the decision tree in this study was close to the diagonal. curve is also close to the diagonal, indicating good calibration performance, suggesting that the decision tree model in this study can better help clinicians provide reliable risk assessment and support personalised treatment decisions.

The decision tree model is simple and easy to use and does not require complex computational equipment, which makes it suitable for dissemination in areas where medical resources are relatively scarce. It has been shown[59] that decision tree models have high potential for application in resource-limited areas, and their simplicity and interpretability make them more suitable for application in primary organisations. For example, the decision tree-based diabetes screening tool developed by Zhong Xin et al[60] helps physicians to quickly assess the risk of diabetes in residents and effectively identify high-risk groups, which supports the practical application of this study. In this study SHAP summary plot shows the ranking of the contribution of different variables to the prediction model of MACEs in the order of LDL-C, age, office systolic blood pressure, potassium, homocysteine, fasting blood glucose, TC, and whether or not they have diabetes. Age was the most important predictor of this predictive value, and the box-and-line plot showed that hypertensive patients who developed MACEs were older, which is consistent with the results of clinical practice, and based on the SHAP dependence plot, it can be seen that age contributes a negative effect or a smaller positive effect closer to the age of 40 years or smaller

values, and tends to bring a positive effect as age increases, which is consistent with the known theory that increasing age is associated with increased risk of cardiovascular events [28]. According to the analysis of SHAP Dependence Plot, when age is close to 40 years or younger, the Shapley value tends to be negative, indicating that lower age is associated with lower risk of MACEs; as age increases, the Shapley value gradually tends to be positive, indicating that higher age is associated with higher risk of MACEs. This trend further validates the clinical significance of age as an independent risk factor for cardiovascular disease. The next important predictor was LDL-C. When LDL-C values were low, Shapley values tended to be negative, indicating that lower LDL-C values were associated with a lower risk of MACEs; as LDL-C values increased, Shapley values tended to be positive, indicating that higher LDL-C values were associated with a higher risk of MACEs. This result is consistent with the theory that higher LDL-C levels are associated with increased risk of atherosclerosis and cardiovascular events; office systolic blood pressure was the third important predictor. Shapley values tended to be negative when systolic blood pressure was near 120 mmHg or lower, indicating that lower systolic blood pressure was associated with a lower risk of MACEs; as systolic blood pressure increased, Shapley values tended to be positive, indicating that higher systolic blood pressure was associated with a higher risk of MACEs. Elevated systolic blood pressure increases cardiac burden and the risk of vascular injury[61]; the normal range of blood potassium is 3.5-5.0 mmol/L, and the Shapley value tends to be positive in the SHAP dependence plot when the potassium concentration is low, indicating that lower blood potassium concentrations are associated with a higher risk of MACEs. The critical importance of monitoring and regulating blood potassium levels is consistent with clinical experience[3].

Conclusions

MACEs have extremely high rates of disability and mortality and have become an important public health problem worldwide. The Kazakh population in Xinjiang has a high prevalence of hypertension and cardiovascular events. Early identification and control of risk factors are important to prevent the development of MACEs in this population. In this study, based on clinical data from the Hypertension Achievement Centre in Altay City, Xinjiang, independent risk factors for MACEs including age, LDL-C, potassium, clinic blood pressure, systolic blood pressure, TC, fasting blood glucose, Hcy, and whether diabetes mellitus was present were identified and controlled by the clinicians' professional judgment, Hcy, and whether diabetes mellitus, and established the multivariate Cox regression model, random forest, XGBoost and decision tree model to predict the occurrence of MACEs. Through the evaluation and validation of each model, it was concluded that the decision tree had the optimal prediction effect, good prediction performance, and was able to effectively identify high-risk patients.

Abbreviations

AUC

area under the curve

MACEs

Major Adverse Cardiovascular Events

XGBoost
Extreme Gradient Boosting
eGFR
estimated Glomerular Filtration Rate
Hcy
Homocysteine

Declarations

Disclosure

The author(s) report no conflicts of interest in this work.

Ethics approval and consent to participate

This study was approved by the Ethics Committee of Altay City People's Hospital (Project Name: Hypertension Standard Center; Principal Investigator: Bo Zhao. Informed consent was obtained from all participants.

Competing interests

The authors have no conflicts of interest to declare.

Funding

This work was supported by the Hypertension Standardization Center Construction Project, Altay City People's Hospital, [grant numbers 2106-654301-23-05-951274]; and Sichuan Center for Collaborative Innovation in Elderly Care and Health (2020), Study on the Nonlinear Effects of Meteorological Factors on Cardiovascular and Cerebrovascular Diseases, [grant numbers YLZBZ2009]; and Chengdu Medical College Graduate Student Innovation Research Fund, [grant numbers YCX2024-03-09].

Author Contribution

Huiyu Shu is the first author of this study. Xixi Feng and Shu Yang are co-corresponding authors. The study conception was developed by Huiyu Shu, Xixi Feng, and Bo Zhao. Data collection and assembly were carried out by Bo Zhao, Jiao Hu, and Yun Zheng. Statistical analysis was performed by Huiyu Shu and Yun Zheng. Interpretation of the results and drafting of the manuscript were conducted by Huiyu Shu, Xixi Feng, Shu Yang, and Bo Zhao. All authors contributed to critical revision of the article. Xixi Feng

and Shu Yang serve as study guarantors. The corresponding authors confirm that all listed authors meet authorship criteria and that no eligible contributors have been omitted.

Data Availability

The datasets generated and/or analysed during the current study are not publicly available due to patient privacy and institutional regulations but are available from the corresponding author on reasonable request. All data have been de-identified prior to analysis.

References

1. Lloyd-Jones, D. M., Leip, E. P., Larson, M. G., Vasan, R. S. & Levy, D. Novel approach to examining first cardiovascular events after hypertension onset. *HYPERTENSION* **45** (1), 39–45 (2005).
2. Burrello, J. et al. Prevalence of Hypokalemia and Primary Aldosteronism in 5100 Patients Referred to a Tertiary Hypertension Unit. *HYPERTENSION* **75**(4):1025–1033. (2020).
3. Hoppe, L. K. et al. Association of Abnormal Serum Potassium Levels with Arrhythmias and Cardiovascular Mortality: a Systematic Review and Meta-Analysis of Observational Studies. *Cardiovasc. Drugs Ther.* **32** (2), 197–212 (2018).
4. Clinical practice guideline for the. management of hypertension in China. *Chin. Med. J. (Engl)*. **137** (24), 2907–2952 (2024).
5. Jia, W. et al. Standards of medical care for type 2 diabetes in China 2019. *Diabetes Metab. Res. Rev.* **35** (6), e3158 (2019).
6. Chen, C. & Lu, F. C. The guidelines for prevention and control of overweight and obesity in Chinese adults. *BIOMED. ENVIRON. SCI.* **17** (Suppl), 1–36 (2004).
7. Martin, G. P., Riley, R. D., Ensor, J. & Grant, S. W. Statistical primer: sample size considerations for developing and validating clinical prediction models. *Eur J. Cardiothorac. Surg* **67**(5). (2025).
8. Collins, G. S. et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e78378 (2024).
9. Jiang, Y. et al. Cardiovascular Disease Prediction by Machine Learning Algorithms Based on Cytokines in Kazakhs of China. *CLIN. EPIDEMIOL.* **13**, 417–428 (2021).
10. Kreutz, R. et al. 2024 European Society of Hypertension clinical practice guidelines for the management of arterial hypertension. *EUR. J. INTERN. MED.* **126**, 1–15 (2024).
11. Li, N. et al. Ethnic disparities in the clustering of risk factors for cardiovascular disease among the Kazakh, Uygur, Mongolian and Han populations of Xinjiang: a cross-sectional study. *BMC PUBLIC. HEALTH.* **12**, 499 (2012).
12. Tao, J. et al. Prevalence of major cardiovascular risk factors and adverse risk profiles among three ethnic groups in the Xinjiang Uygur Autonomous Region, China. *LIPIDS HEALTH DIS.* **12**, 185 (2013).

13. Qian, X. et al. A Cardiovascular Disease Prediction Model Based on Routine Physical Examination Indicators Using Machine Learning Methods: A Cohort Study. *Front. Cardiovasc. Med.* **9**, 854287 (2022).
14. Izumi, Y. et al. Effect of age at menopause on blood pressure in postmenopausal women. *AM. J. HYPERTENS.* **20** (10), 1045–1050 (2007).
15. Mendelsohn, M. E. & Karas, R. H. The protective effects of estrogen on the cardiovascular system. *N Engl. J. Med.* **340** (23), 1801–1811 (1999).
16. Ambrose, J. A. & Barua, R. S. The pathophysiology of cigarette smoking and cardiovascular disease: an update. *J. AM. COLL. CARDIOL.* **43** (10), 1731–1737 (2004).
17. D'Agostino, R. S. et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *CIRCULATION* **117** (6), 743–753 (2008).
18. He, F. J., Li, J. & Macgregor, G. A. Effect of longer term modest salt reduction on blood pressure: Cochrane systematic review and meta-analysis of randomised trials. *BMJ* **346**, f1325 (2013).
19. Cook, N. R. et al. Long term effects of dietary sodium reduction on cardiovascular disease outcomes: observational follow-up of the trials of hypertension prevention (TOHP). *BMJ* **334** (7599), 885–888 (2007).
20. Wu, J. et al. High prevalence of coexisting prehypertension and prediabetes among healthy adults in northern and northeastern China. *BMC PUBLIC. HEALTH.* **11**, 794 (2011).
21. Jia, G. & Sowers, J. R. Hypertension in Diabetes: An Update of Basic Mechanisms and Clinical Disease. *HYPERTENSION* **78** (5), 1197–1205 (2021).
22. Smulyan, H., Lieber, A. & Safar, M. E. Hypertension, Diabetes Type II, and Their Association: Role of Arterial Stiffness. *AM. J. HYPERTENS.* **29** (1), 5–13 (2016).
23. Kloc, M., Kubiak, J. Z. & Ghobrial, R. M. Macrophage-, Dendritic-, Smooth Muscle-, Endothelium-, and Stem Cells-Derived Foam Cells in Atherosclerosis. *INT J. MOL. SCI* **23**(22). (2022).
24. Pan, Y. et al. Association of ACE2 polymorphisms with susceptibility to essential hypertension and dyslipidemia in Xinjiang, China. *LIPIDS HEALTH DIS.* **17** (1), 241 (2018).
25. Zhang, X., Sessa, W. C. & Fernandez-Hernando, C. Endothelial Transcytosis of Lipoproteins in Atherosclerosis. *Front. Cardiovasc. Med.* **5**, 130 (2018).
26. Lewington, S., Clarke, R., Qizilbash, N., Peto, R. & Collins, R. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *LANCET* **360** (9349), 1903–1913 (2002).
27. Rensma, S. P. et al. Associations of Arterial Stiffness With Cognitive Performance, and the Role of Microvascular Dysfunction: The Maastricht Study. *HYPERTENSION* **75**(6):1607–1614. (2020).
28. Lippi, G., Salvagno, G. L., Targher, G. & Guidi, G. C. Multiple biomarkers for the prediction of first major cardiovascular events and death: considerable costs and limited benefits. *MedGenMed* **9** (1), 34 (2007).

29. Zhang, Y., Lacolley, P., Protogerou, A. D. & Safar, M. E. Arterial Stiffness in Hypertension and Function of Large Arteries. *AM. J. HYPERTENS.* **33** (4), 291–296 (2020).
30. Laurent, S. & Boutouyrie, P. Arterial Stiffness and Hypertension in the Elderly. *Front. Cardiovasc. Med.* **7**, 544302 (2020).
31. Csiszar, A. et al. Hypertension impairs neurovascular coupling and promotes microvascular injury: role in exacerbation of Alzheimer's disease. *GEROSCIENCE* **39** (4), 359–372 (2017).
32. Muller-Brunotte, R. et al. Myocardial fibrosis and diastolic dysfunction in patients with hypertension: results from the Swedish Irbesartan Left Ventricular Hypertrophy Investigation versus Atenolol (SILVHIA). *J. HYPERTENS.* **25** (9), 1958–1966 (2007).
33. Weber, K. T., Brilla, C. G. & Campbell, S. E. Regulatory mechanisms of myocardial hypertrophy and fibrosis: results of in vivo studies. *CARDIOLOGY* **81** (4–5), 266–273 (1992).
34. Zhang, X., Sessa, W. C. & Fernandez-Hernando, C. Endothelial Transcytosis of Lipoproteins in Atherosclerosis. *Front. Cardiovasc. Med.* **5**, 130 (2018).
35. Li, Z., Zhou, Y., Lai, M., Luo, J. & Yan, F. Acoustic Delivery of Plasma Low-Density Lipoprotein into Liver via ApoB100-Targeted Microbubbles Inhibits Atherosclerotic Plaque Growth. *ACS Appl. Mater. Interfaces.* **16** (19), 24206–24220 (2024).
36. Chen, L. et al. Low-Density Lipoprotein Cholesterol, Cardiovascular Disease Risk, and Mortality in China. *JAMA Netw. Open.* **7** (7), e2422558 (2024).
37. Xu, J. J. et al. [The impact of LDL-C/HDL-C ratio on severity of coronary artery disease and 2-year outcome in patients with premature coronary heart disease: results of a prospective, multicenter, observational cohort study]. *Zhonghua Xin Xue Guan Bing Za Zhi.* **51** (7), 702–708 (2023).
38. Flint, A. C. et al. Effect of Systolic and Diastolic Blood Pressure on Cardiovascular Outcomes. *N Engl. J. Med.* **381** (3), 243–251 (2019).
39. Asztalos, B. F., Horvath, K. V. & Schaefer, E. J. High-Density Lipoprotein Particles, Cell-Cholesterol Efflux, and Coronary Heart Disease Risk. *Arterioscler. Thromb. Vasc Biol.* **38** (9), 2007–2015 (2018).
40. Zhang, G. et al. Association between extremely high-density lipoprotein cholesterol and adverse cardiovascular outcomes: a systematic review and meta-analysis. *Front. Cardiovasc. Med.* **10**, 1201107 (2023).
41. Kannel, W. B. Range of serum cholesterol values in the population developing coronary artery disease. *AM. J. CARDIOL.* **76** (9), 69C–77C (1995).
42. Herrington, W., Lacey, B., Sherliker, P., Armitage, J. & Lewington, S. Epidemiology of Atherosclerosis and the Potential to Reduce the Global Burden of Atherothrombotic Disease. *CIRC. RES.* **118** (4), 535–546 (2016).
43. Bentzon, J. F., Otsuka, F., Virmani, R. & Falk, E. Mechanisms of plaque formation and rupture. *CIRC. RES.* **114** (12), 1852–1866 (2014).
44. Tarbell, J., Mahmoud, M., Corti, A., Cardoso, L. & Caro, C. The role of oxygen transport in atherosclerosis and vascular disease. *J. R SOC. INTERFACE.* **17** (165), 20190732 (2020).

45. Yang, X. et al. Oxidative Stress-Mediated Atherosclerosis: Mechanisms and Therapies. *FRONT. PHYSIOL.* **8**, 600 (2017).
46. Wang, X. D. et al. Prognosis of Older Adult Patients Suffering from Atrial Fibrillation and Hypokalemia. *CLIN. INTERV AGING.* **18**, 1363–1371 (2023).
47. Helfant, R. H. Hypokalemia and arrhythmias. *AM. J. MED.* **80** (4A), 13–22 (1986).
48. Ganguly, P. & Alam, S. F. Role of homocysteine in the development of cardiovascular disease. *NUTR. J.* **14**, 6 (2015).
49. Tamura, T. et al. Association between plasma levels of homocysteine, folate, and vitamin B(12), and dietary folate intake and hypertension in a cross-sectional study. *Sci. Rep.* **10** (1), 18499 (2020).
50. Jakubowski, H. & Witucki, L. Homocysteine Metabolites, Endothelial Dysfunction, and Cardiovascular Disease. *INT J. MOL. SCI* **26**(2). (2025).
51. Kaneko, H. et al. Fasting plasma glucose and subsequent cardiovascular disease among young adults: Analysis of a nationwide epidemiological database. *ATHEROSCLEROSIS* **319**, 35–41 (2021).
52. Zuo, Y. et al. Association of Impaired Fasting Glucose With Cardiovascular Disease in the Absence of Risk Factor. *J. Clin. Endocrinol. Metab.* **107** (4), e1710–e1718 (2022).
53. Wang, X. et al. Changes in Impaired Fasting Glucose and Borderline High Low-Density Lipoprotein-Cholesterol Status Alter the Risk of Cardiovascular Disease: A 9-Year Prospective Cohort Study. *Front. Cardiovasc. Med.* **9**, 882984 (2022).
54. Pi, X., Xie, L. & Patterson, C. Emerging Roles of Vascular Endothelium in Metabolic Homeostasis. *CIRC. RES.* **123** (4), 477–494 (2018).
55. Prentice, R. L. & Zhao, S. Regression Models and Multivariate Life Tables. *J. AM. STAT. ASSOC.* **116** (535), 1330–1345 (2021).
56. Chen, C. Y. & Chang, Y. W. Missing data imputation using classification and regression trees. *PeerJ Comput. Sci.* **10**, e2119 (2024).
57. Becker, T., Rousseau, A. J., Geubbelmans, M., Burzykowski, T. & Valkenborg, D. Decision trees and random forests. *Am. J. Orthod. Dentofac. Orthop.* **164** (6), 894–897 (2023).
58. Yi, F. et al. XGBoost-SHAP-based interpretable diagnostic framework for alzheimer's disease. *BMC Med. Inf. Decis. Mak.* **23** (1), 137 (2023).
59. Chen, J. H. & Asch, S. M. Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. *N Engl. J. Med.* **376** (26), 2507–2509 (2017).
60. Xin, Z. et al. Reanalysis and External Validation of a Decision Tree Model for Detecting Unrecognized Diabetes in Rural Chinese Individuals. *INT. J. ENDOCRINOL.* **2017**, 3894870 (2017).
61. Izumi, Y. et al. Effect of age at menopause on blood pressure in postmenopausal women. *AM. J. HYPERTENS.* **20** (10), 1045–1050 (2007).