Journal of Geriatric Cardiology | JGC
RESEARCH ARTICLE
J Geriatr Cardiol 2022; 19(6): 445–455

Establishment of a diagnostic model of coronary heart disease in elderly patients with diabetes mellitus based on machine learning algorithms

Hu XU[1,2,3,*], Wen-Zhe CAO[1,4,*], Yong-Yi BAI[2], Jing DONG[5], He-Bin CHE[5], Po BAI[6], Jian-Dong WANG[1,3,✉], Feng CAO[1,2,✉], Li FAN[1,2,✉]

1. Chinese PLA Medical School, Chinese PLA General Hospital, Beijing, China; 2. Department of Cardiology, the Second Medical Center, National Clinical Research Center for Geriatric Diseases, Chinese PLA General Hospital, Beijing, China; 3. Department of General Surgery, the First Medical Center, Chinese PLA General Hospital, Beijing, China; 4. Institute of Geriatrics, the Second Medical Center, Chinese PLA General Hospital, Beijing, China; 5. Medical Big Data Research Center & National Engineering Laboratory for Medical Big Data Application Technology, Chinese PLA General Hospital, Beijing, China; 6. Department of Respiratory Diseases, Chinese PLA Rocket Force Characteristic Medical Center, Beijing, China

[*]The authors contributed equally to this manuscript

✉ Correspondence to: vicky1968@163.com; fengcao8828@163.com; fl6698@163.com

https://doi.org/10.11909/j.issn.1671-5411.2022.06.006

ABSTRACT

OBJECTIVE To establish a prediction model of coronary heart disease (CHD) in elderly patients with diabetes mellitus (DM) based on machine learning (ML) algorithms.

METHODS Based on the Medical Big Data Research Centre of Chinese PLA General Hospital in Beijing, China, we identified a cohort of elderly inpatients (≥ 60 years), including 10,533 patients with DM complicated with CHD and 12,634 patients with DM without CHD, from January 2008 to December 2017. We collected demographic characteristics and clinical data. After selecting the important features, we established five ML models, including extreme gradient boosting (XGBoost), random forest (RF), decision tree (DT), adaptive boosting (Adaboost) and logistic regression (LR). We compared the receiver operating characteristic curves, area under the curve (AUC) and other relevant parameters of different models and determined the optimal classification model. The model was then applied to 7447 elderly patients with DM admitted from January 2018 to December 2019 to further validate the performance of the model.

RESULTS Fifteen features were selected and included in the ML model. The classification precision in the test set of the XGBoost, RF, DT, Adaboost and LR models was 0.778, 0.789, 0.753, 0.750 and 0.689, respectively; and the AUCs of the subjects were 0.851, 0.845, 0.823, 0.833 and 0.731, respectively. Applying the XGBoost model with optimal performance to a newly recruited dataset for validation, the diagnostic sensitivity, specificity, precision, and AUC were 0.792, 0.808, 0.748 and 0.880, respectively.

CONCLUSIONS The XGBoost model established in the present study had certain predictive value for elderly patients with DM complicated with CHD.

With the population in China ageing at an increasing rate, the phenomenon of comorbidity among the elderly has become increasingly prominent. Comorbidity is the coexistence of two or more chronic diseases or conditions.[1] The prevalence of comorbidity in individuals aged 65 years or older is 76.6%.[2] Comorbidity has brought a heavy burden to individuals, families and society. In the elderly, comorbidity with coronary heart disease (CHD) and diabetes mellitus (DM) is one of the most common occurrences. There are 114 million patients with DM in China, which is more than any other country in the world.[3] Cardiovascular disease, which is represented by CHD, is the leading cause of death among urban and rural residents. It is estimated that there are 11.39 million

patients with CHD in China.[4] In patients with DM, the risk of CHD increased significantly.[5,6] DM is an independent risk factor for CHD.[7] CHD is one of the most common complications of DM.[8] The life expectancy of patients with DM complicated with cardiovascular disease is significantly decreased.[9]

In elderly patients with comorbid CHD and DM, the symptoms of myocardial ischaemia are relatively mild or absent, which is not easy to determine and diagnose.[10] Irreversible pathological damage is often present when the diagnosis of CHD is confirmed, which seriously affects the quality of life of patients.[11] Coronary angiography is the gold standard for diagnosing of CHD. However, this operation is complex, invasive and expensive, and is not included in routine examinations. Therefore, early screening of CHD in elderly patients with DM is of great significance.

In recent years, many experts and researchers have begun to explore new models of DM diagnosis, making full use of medical big data and machine learning (ML) algorithms, and have achieved good results in prediction and diagnosis of CHD.[12–15] For different data sets, researchers use different algorithms, such as sequential minimal optimization,[16] artificial neural network[17] and neural network models,[18–20] to build CHD diagnosis models, including a variety of features, like electrocardiogram (ECG),[21] heart rate variability,[22] and even facial photos.[23] Currently, there is no specific model to predict the risk of CHD in elderly patients with DM. This study aimed to establish a prediction model of CHD in elderly patients with DM using the ML algorithm, and to provide a new auxiliary screening method from the perspective of medicine and big data science.

METHODS

Study Population

Based on Medical Big Data Research Center of Chinese PLA General Hospital in Beijing, China, we identified a total of 28,059 elderly inpatients (≥ 60 years) who were diagnosed with DM from January 2008 to December 2017. The inclusion criteria were inpatients aged 60 years or older and the discharge diagnosis included DM according to the corresponding diagnostic criteria. The exclusion criteria were malignant tumor, severe hepatic and renal dysfunction, respiratory failure, sepsis, shock, severe autoimmune disease, severe hematological disease, acute rheumatic fever, chronic rheumatic valvular disease, myocarditis, and myocardiopathy. The diagnostic criteria for CHD are based on the European Society of Cardiology guidelines for the diagnosis and management of chronic coronary syndromes.[24] This study was approved by the Ethics Committee of Chinese PLA General Hospital (No.S2018-269-02).

According to the inclusion and exclusion criteria, 4892 patients were excluded. Finally, a total of 23,167 elderly patients with DM were included in the present study. Among them, 10,533 patients complicated with CHD were included in the research group and 12,634 patients complicated without CHD were included in the control group.

To further validate the performance of the models, we collected information on 7447 elderly patients with DM admitted from January 2018 to December 2019. Of these patients, there were 3116 patients with CHD and 4331 patients without CHD.

Candidate Variables

All demographic characteristics, laboratory examinations, history of complications and other clinical data were used to establish prediction models. For missing data, the nonparametric filling missForest algorithm was used. Any variable with > 30% of missing data was removed. Recursive feature elimination (RFE) is a method for feature selection that combines with various ML models to eliminate redundant information, thus identifying the most influential features for each model. In our study, RFE was used to rank the importance of all 67 feature variables, and the top 15 feature variables were selected to construct models. The importance of each variable in each model was also evaluated.

Model Development

To develop ML models, random split validation was used. We selected 80% of the research group samples and 80% of the control group samples as the training dataset. The remaining subset (20%) was reserved as the testing dataset.

Five commonly used ML algorithms were used to train the models and diagnose CHD in elderly patients with DM. Extreme gradient boosting (XGBoost) is an ensemble ML algorithm based on de-

cision tree (DT). Random forest (RF) is a combined classifier composed of multiple DTs, and it is an excellent ensemble learning method. DT is a ML algorithm that continuously introduces features to reduce the uncertainty of original random variables. Adaptive boosting (Adaboost) is an algorithm that integrates the weak classifiers to form a strong classifier with excellent performance. Logistic regression (LR) is a probabilistic nonlinear regression model, which is a multivariate analysis method to study binary output classification.[25,26] Various Python packages were used to conduct this analysis.

Model Performance

The performance of the model was evaluated using the standard format, namely confusion matrix.[27] The evaluation indicators included sensitivity, specificity, accuracy, precision, F1-score, receiver operating characteristic curves, and area under the curve (AUC). By comparing the relevant parameters of various models, the optimal classification model was found out.

Statistical Analysis

Continuous variables were presented as mean ± SD or median (interquartile range) and categorical variables were presented as counts (percentages). The Mann-Whitney *U* test and the Pearson's chi-squared test were used to test the difference of variables between the research group and the control group. Different Python packages were used to establish the five ML models. DeLong method was used to compare AUC between different models, which was realized by MedCalc 11.4.2.0 (MedCalc Software, Ostend, Belgium). Two-sided *P*-value < 0.05 were considered statistically significant. Statistical analysis was performed with SPSS 24.0 (SPSS Inc., IBM, Armonk, NY, USA) and Python 3.7.0 (https://www.python.org/downloads/release/python-370/).

RESULTS

Clinical Characteristics

A cohort of 23,167 elderly patients with DM was enrolled in our study, including 10,533 patients in the research group and 12,634 patients in the control group. We collected demographic characteristics, laboratory examinations, and history of complications. The clinical characteristics of all patients are summarized in Table 1.

Predictor Variables

RFE was used to rank the importance of all 67 feature variables, and the top 15 feature variables were selected to construct models. The top 15 feature variables were pro-B-type natriuretic peptide, hemoglobin A1c, troponin T, high-density lipoprotein cholesterol, total bile acid, D-dimer, glycated serum pr-

Table 1 Clinical characteristics of all patients.

Characteristics	Research group (*n* = 10,533)	Control group (*n* = 12,634)	*P*-value
Demographics			
Male	5537 (52.6%)	6168 (48.8%)	< 0.001
Age, yrs	69 (64–74)*	68 (63–73)*	< 0.001
Body mass index, kg/m^2	25.7 (23.6–28.0)*	25.3 (23.1–27.6)*	< 0.001
Systolic blood pressure, mmHg	138 (126–150)*	138 (126–150)*	0.606
Diastolic blood pressure, mmHg	75 (68–82)*	80 (70–85)*	< 0.001
Blood electrolyte			
Ca^{2+}, mmol/L	2.25 (2.16–2.32)*	2.22 (2.11–2.31)*	< 0.001
P^{5+}, mmol/L	1.09 (0.93–1.24)*	1.10 (0.87–1.25)*	0.002
Mg^{2+}, mmol/L	0.86 (0.80–0.92)*	0.85 (0.76–0.91)*	< 0.001
K^+, mmol/L	3.91 (3.63–4.17)*	3.90 (3.54–4.18)*	< 0.001
Na^+, mmol/L	141.5 (139.1–143.5)*	141.7 (138.6–143.7)*	0.032
Cl^-, mmol/L	103.8 (101.2–106.1)*	103.2 (99.8–105.7)*	< 0.001
CO_2, mmol/L	25.9 (23.6–27.9)*	25.7 (22.9–27.8)*	< 0.001

Continued

Characteristics	Research group (n = 10,533)	Control group (n = 12,634)	P-value
Glycemic control			
Hemoglobin A1c, %	6.30 (3.77–7.60)*	3.78 (3.75–6.80)*	< 0.001
Glycated serum protein, μmol/L	171.0 (120.1–230.3)*	136.2 (120.1–218.1)*	< 0.001
Myocardial enzyme			
Troponin T, ng/mL	0.006 (0.003–0.016)*	0.003 (0.003–0.003)*	< 0.001
Creatine kinase, U/L	66.3 (45.2–101.3)*	57.8 (31.6–89.7)*	< 0.001
Creatine kinase-MB, U/L	7.2 (0.6–13.9)*	9.9 (0.5–14.9)*	< 0.001
Lactate dehydrogenase, U/L	156.4 (135.4–183.7)*	150.0 (123.4–175.1)*	< 0.001
Pro-B-type natriuretic peptide, pg/mL	59.5 (5.0–315.9)*	5.0 (5.0–5.1)*	< 0.001
Blood coagulation			
Thrombin time, s	16.2 (15.4–17.0)*	16.0 (12.7–16.9)*	< 0.001
Activated partial thromboplastin time, s	22.5 (22.3–35.7)*	22.4 (22.2–34.5)*	< 0.001
Prothrombin time, s	13.3 (12.8–14.1)*	13.3 (12.7–14.3)*	0.988
Prothrombin activity, %	96 (86–105)*	96 (82–106)*	< 0.001
International normalized ratio	1.03 (0.97–1.07)*	1.03 (0.97–1.07)*	0.041
Fibrinogen, g/L	3.33 (2.81–3.98)*	3.19 (2.53–3.88)*	< 0.001
D-dimer, μg/mL	0.30 (0.01–0.52)*	0.01 (0.01–0.37)*	< 0.001
Routine blood + C-reactive protein			
Hemoglobin, g/dL	132 (120–143)*	131 (116–142)*	< 0.001
Red blood cell count, × 10^{12}/L	4.35 (3.97–4.70)*	4.30 (3.84–4.67)*	< 0.001
White blood cell count, × 10^{9}/L	6.42 (5.27–7.73)*	6.03 (4.77–7.46)*	< 0.001
Neutrophil, %	0.61 (0.54–0.68)*	0.58 (0.50–0.66)*	< 0.001
Lymphocyte, %	0.28 (0.21–0.35)*	0.29 (0.19–0.36)*	0.225
Monocyte, %	0.06 (0.05–0.07)*	0.06 (0.04–0.07)*	< 0.001
Eosinophil, %	0.02 (0.01–0.03)*	0.02 (0.01–0.03)*	< 0.001
Basophil, %	0.003 (0.002–0.005)*	0.003 (0.002–0.005)*	< 0.001
Hematocrit, %	0.40 (0.37–0.43)*	0.40 (0.37–0.44)*	< 0.001
Mean corpuscular volume, fl	90.0 (87.1–92.9)*	89.4 (85.8–92.4)*	< 0.001
Mean corpuscular hemoglobin, pg	30.4 (29.3–31.4)*	30.2 (28.9–31.3)*	< 0.001
Mean corpuscular hemoglobin concentration, g/L	337 (329–344)*	336 (326–344)*	< 0.001
Red cell distribution width, %	12.8 (12.3–13.4)*	12.8 (12.2–13.3)*	< 0.001
Platelet count, × 10^{9}/L	198 (162–240)*	193 (150–237)*	< 0.001
Mean platelet volume, fl	10.6 (9.9–11.2)*	10.4 (9.7–11.2)*	< 0.001
C-reactive protein, mg/dL	0.1 (0–0.348)*	0.0 (0–0.328)*	< 0.001
Routine urine			
Specific gravity	1.014 (1.010–1.019)*	1.013 (1.005–1.019)*	< 0.001
pH	5.5 (5.0–6.5)*	5.5 (5.0–6.5)*	< 0.001
Blood biochemical			
Alanine aminotransferase, U/L	17.2 (12.3–25.6)*	15.1 (10.1–22.0)*	< 0.001
Aspartate aminotransferase, U/L	16.6 (13.4–21.8)*	15.3 (11.9–19.7)*	< 0.001

Continued

Characteristics	Research group (n = 10,533)	Control group (n = 12,634)	P-value
Total protein, g/L	67.7 (63.5–71.9)*	65.7 (60.4–70.2)*	< 0.001
Albumin, g/L	40.5 (37.7–42.9)*	39.7 (35.3–42.4)*	< 0.001
Total bilirubin, μmol/L	10.1 (7.4–13.5)*	9.4 (6.4–13.0)*	< 0.001
Direct bilirubin, μmol/L	3.1 (2.2–4.3)*	2.8 (1.8–4.1)*	< 0.001
Total bile acid, μmol/L	1.4 (0.002–3.7)*	2.4 (0.003–4.6)*	< 0.001
Homocysteine, μmol/L	0.82 (0.76–13.20)*	1.10 (0.77–13.10)*	< 0.001
Alkaline phosphatase, U/L	65.6 (52.9–79.9)*	64.7 (48.9–80.7)*	< 0.001
Glutamyl transpeptidase, U/L	22.6 (15.8–34.6)*	19.9 (12.5–32.1)*	< 0.001
Glucose, mmol/L	7.02 (5.57–9.46)*	6.37 (4.97–8.32)*	< 0.001
Blood urea nitrogen, mmol/L	5.56 (4.46–6.91)*	5.29 (4.04–6.59)*	< 0.001
Creatinine, μmol/L	70.7 (58.5–84.5)*	64.1 (51.4–78.0)*	< 0.001
Uric acid, μmol/L	296.3 (239.2–359.0)*	269.0 (196.1–333.7)*	< 0.001
Total cholesterol, mmol/L	3.80 (2.99–4.65)*	3.92 (1.24–4.85)*	0.290
Triglyceride, mmol/L	1.27 (0.85–1.81)*	1.11 (0.23–1.69)*	< 0.001
Apolipoprotein A1, g/L	0.95 (0.23–1.23)*	0.77 (0.23–1.22)*	< 0.001
Apolipoprotein B, g/L	0.61 (0.19–0.88)*	0.50 (0.18–0.91)*	< 0.001
High-density lipoprotein cholesterol, mmol/L	0.99 (0.79–1.19)*	0.94 (0.08–1.20)*	< 0.001
Low-density lipoprotein cholesterol, mmol/L	2.18 (1.54–2.90)*	2.26 (0.09–3.04)*	< 0.001
Lipoprotein(a), mg/dL	0.00 (0–16.82)*	0.00 (0–15.51)*	0.543
Apolipoprotein A1/Apolipoprotein B	1.28 (1.19–1.51)*	1.28 (1.18–1.39)*	< 0.001
Estimated glomerular filtration rate, mL/min per 1.73 m^2	98.2 (80.3–117.9)*	108.1 (88.2–135.5)*	< 0.001

Data are presented as n (%). *Presented as median (interquartile range).

otein, activated partial thromboplastin time, triglyceride, low-density lipoprotein cholesterol, total cholesterol, total protein, creatine kinase, creatine kinase-MB and lymphocyte.

Performance of the Predictive Models on Test set

The classification precision in test set of XGBoost, RF, DT, Adaboost and LR models was 0.778, 0.789, 0.753, 0.750 and 0.689, respectively; and the AUC of the subjects was 0.851, 0.845, 0.823, 0.833 and 0.731, respectively (Table 2, Figure 1). Among the five models, XGBoost performed the best with the highest AUC. The AUC is 0.851 (95% CI: 0.841–0.861). The importance scores of the 15 features of XGBoost are shown in Figure 2.

Performance of the XGBoost Model on a Newly Recruited Independent set

The clinical characteristics of newly recruited subjects are shown in Table 3. Applying the XGBoost model with optimal performance to a newly recru-

Table 2 Performance of the five models on the testing sets.

Test set	AUC	95% CI	Precision	Accuracy	F1-score	Sensitivity	Specificity
XGBoost	0.851	0.841–0.861	0.778	0.773	0.732	0.690	0.841
RF	0.845	0.834–0.855	0.789	0.778	0.735	0.688	0.851
DT	0.823	0.812–0.834	0.753	0.757	0.714	0.679	0.820
Adaboost	0.833	0.822–0.844	0.750	0.757	0.715	0.683	0.816
LR	0.731	0.718–0.744	0.689	0.686	0.607	0.542	0.803

Adaboost: adaptive boosting; AUC: area under the curve; DT: decision tree; LR: logistic regression; RF: random forest; XGBoost: extreme gradient boosting.

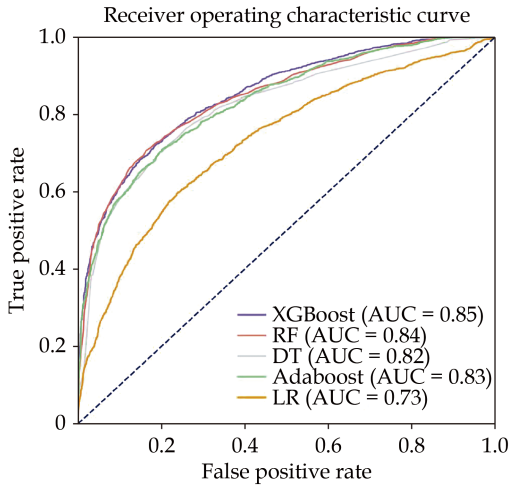

Figure 1 The receiver operating characteristic and the AUC of the five prediction models on the testing dataset. Adaboost: adaptive boosting; AUC: area under the curve; DT: decision tree; LR: logistic regression; RF: random forest; XGBoost: extreme gradient boosting.

ited independent dataset for validation, the diagnostic sensitivity, specificity, precision, and AUC were 0.792, 0.808, 0.748 and 0.880, respectively (Table 4, Figure 3). The XGBoost model performed well in diagnosing CHD in elderly patients with DM.

DISCUSSION

Based on data from the Medical Big Data Research Centre of Chinese PLA General Hospital in Beijing, China, the present study collected hospitalization information for tens of thousands of elderly patients with DM in the past ten years. Five ML algorithms were used to build a diagnostic model of CHD in elderly patients with DM. Ultimately, it was found that the XGBoost model best distinguished patients with CHD and had good performance.

Based on ML algorithms, researchers explored the diagnosis of CHD.[28] A review published in 2019 studied 149 research articles related to ML-based CHD detection.[29] The 67 relevant datasets came from 18 countries and regions of three continents. The sample size ranged from 20 to 240,000, and the median sample size was approximately 350. Similar to our study, Fan, *et al.*[30] trained an artificial intelligence model using RF to predict CHD risk among patients with DM, there were 1273 patients enrolled in the study. Researchers selected the top eight features (age, heart rate, diastolic blood pressure, blood platelet, low-density lipoprotein cholesterol, total cholesterol level, course of hypertension and course of DM) from the total 50 features to develop the model. The model achieved an AUC of 0.77 on the training dataset and 0.80 on the testing dataset. In our study, a dataset of 28,059 elderly patients with DM was constructed. Our research sample size is large, and the performance of the model is better. The results in different studies cannot be generalized due to the differences in the analyzed datasets, sample sizes, features, data collection areas, performance metrics, and applied ML algorithms.

Feature selection (selecting subsets of relevant features for model development) has a significant impact on the model performance.[25] Many feature selection methods have been used for model creation, including information gain, correlation, principal component analysis, Gini index and the genetic algo-

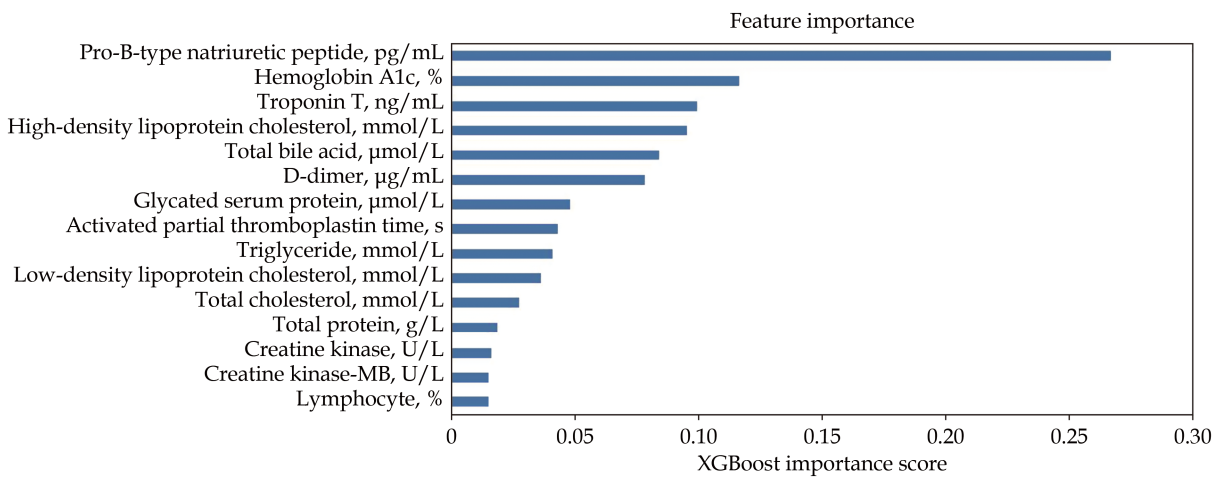

Figure 2 Ranking of the relative importance of XGBoost model features. XGBoost: extreme gradient boosting.

Table 3 Clinical characteristics of newly recruited subjects.

Characteristics	Research group (n = 3116)	Control group (n = 4331)	P-value
Demographics			
Male	1816 (58.3%)	2321 (53.6%)	< 0.001
Age, yrs	68 (64–74)*	67 (64–72)*	< 0.001
Body mass index, kg/m^2	25.6 (23.5–27.7)*	25.3 (23.2–27.4)*	< 0.001
Systolic blood pressure, mmHg	138 (124–150)*	138 (126–150)*	0.082
Diastolic blood pressure, mmHg	75 (67–82)*	78 (70–85)*	< 0.001
Blood electrolyte			
Ca^{2+}, mmol/L	2.25 (2.17–2.33)*	2.22 (2.11–2.3)*	< 0.001
P^{5+}, mmol/L	1.08 (0.93–1.21)*	1.10 (0.90–1.24)*	0.209
Mg^{2+}, mmol/L	0.85 (0.79–0.91)*	0.83 (0.75–0.89)*	< 0.001
K^+, mmol/L	3.84 (3.58–4.11)*	3.88 (3.48–4.19)*	0.326
Na^+, mmol/L	141.1 (139.1–142.7)*	141.1 (138.4–143.0)*	0.238
Cl^-, mmol/L	102.8 (100.4–104.8)*	102.0 (98.7–104.5)*	< 0.001
CO_2, mmol/L	26.6 (24.6–28.6)*	26 (23.3–27.9)*	< 0.001
Glycemic control			
Hemoglobin A1c, %	6.8 (3.8–7.9)*	3.8 (3.8–6.5)*	< 0.001
Glycated serum protein, μmol/L	180.4 (129.9–221.2)*	172.8 (139.9–201.5)*	< 0.001
Myocardial enzyme			
Troponin T, ng/mL	0.009 (0.004–0.019)*	0.003 (0.003–0.008)*	< 0.001
Creatine kinase, U/L	67.1 (45.9–99.6)*	61.1 (35.5–93.8)*	< 0.001
Creatine kinase-MB, U/L	6.0 (0.5–11.7)*	10.2 (0.5–15.8)*	< 0.001
Lactate dehydrogenase, U/L	159.7 (137.7–187.0)*	152.9 (126.4–177.9)*	< 0.001
Pro-B-type natriuretic peptide, pg/mL	103.9 (26.0–355.7)*	5.0 (5.0–76.0)*	< 0.001
Blood coagulation			
Thrombin time, s	15.8 (15.1–16.5)*	15.7 (14.7–16.5)*	< 0.001
Activated partial thromboplastin time, s	22.28 (22.16–22.40)*	22.22 (22.06–22.34)*	< 0.001
Prothrombin time, s	13.3 (12.8–13.9)*	13.3 (12.7–14.4)*	0.048
Prothrombin activity, %	97 (88–105)*	96 (82–107)*	0.006
International normalized ratio	1.02 (0.98–1.07)*	1.03 (0.97–1.07)*	0.854
Fibrinogen, g/L	3.26 (2.79–3.85)*	3.00 (2.34–3.61)*	< 0.001
D-dimer, μg/mL	0.33 (0.23–0.55)*	0.26 (0.01–0.47)*	< 0.001
Routine blood + C-reactive protein			
Hemoglobin, g/dL	132 (121–143)*	129 (114–141)*	< 0.001
Red blood cell count, × 10^{12}/L	4.34 (3.97–4.68)*	4.25 (3.75–4.61)*	< 0.001
White blood cell count, × 10^9/L	6.40 (5.18–7.65)*	5.92 (4.61–7.37)*	< 0.001
Neutrophil, %	0.61 (0.54–0.67)*	0.58 (0.49–0.66)*	< 0.001
Lymphocyte, %	0.28 (0.21–0.34)*	0.28 (0.17–0.35)*	0.008
Monocyte, %	0.06 (0.05–0.08)*	0.06 (0.05–0.07)*	< 0.001
Eosinophil, %	0.02 (0.01–0.03)*	0.02 (0.01–0.03)*	< 0.001
Basophil, %	0.004 (0.003–0.006)*	0.004 (0.002–0.006)*	< 0.001

Continued

Characteristics	Research group (n = 3116)	Control group (n = 4331)	P-value
Hematocrit, %	0.39 (0.36–0.42)*	0.40 (0.36–0.43)*	< 0.001
Mean corpuscular volume, fl	88.0 (85.4–90.9)*	87.7 (84.1–90.8)*	< 0.001
Mean corpuscular hemoglobin, pg	30.4 (29.3–31.5)*	30.3 (28.8–31.4)*	< 0.001
Mean corpuscular hemoglobin concentration, g/L	344 (336–351)*	342 (332–351)*	< 0.001
Red cell distribution width, %	12.6 (12.1–13.1)*	12.4 (11.9–13.0)*	< 0.001
Platelet count, × 10^9/L	202 (164–245)*	197 (149–239)*	< 0.001
Mean platelet volume, fl	10.5 (9.9–11.2)*	10.3 (9.5–11.1)*	< 0.001
C-reactive protein, mg/dL	0.097 (0.050–0.183)*	0.05 (0–0.117)*	< 0.001
Routine urine			
Specific gravity	1.014 (1.010–1.019)*	1.012 (1.000–1.019)*	< 0.001
pH	6.0 (5.5–6.5)*	5.5 (5.0–6.0)*	< 0.001
Blood biochemical			
Alanine aminotransferase, U/L	16.3 (11.7–23.7)*	14.3 (9.5–20.7)*	< 0.001
Aspartate aminotransferase, U/L	15.9 (13.0–20.6)*	14.8 (11.3–18.9)*	< 0.001
Total protein, g/L	67.8 (64.0–71.8)*	65.4 (60.1–69.5)*	< 0.001
Albumin, g/L	40.7 (38.0–43.1)*	40.1 (35.4–42.8)*	< 0.001
Total bilirubin, μmol/L	10.4 (7.5–13.9)*	9.8 (6.6–13.5)*	< 0.001
Direct bilirubin, μmol/L	3.0 (2.0–4.2)*	2.8 (1.7–4.0)*	< 0.001
Total bile acid, μmol/L	1.2 (0.001–1.7)*	1.8 (0.002–4.1)*	< 0.001
Homocysteine, μmol/L	0.79 (0.75–11.70)*	0.83 (0.76–12.40)*	< 0.001
Alkaline phosphatase, U/L	62.9 (50.6–76.8)*	61.0 (46.3–76.1)*	< 0.001
Glutamyl transpeptidase, U/L	22.0 (15.3–32.5)*	18.9 (11.8–29.6)*	< 0.001
Glucose, mmol/L	7.07 (5.66–9.57)*	6.37 (4.85–8.09)*	< 0.001
Blood urea nitrogen, mmol/L	5.44 (4.45–6.79)*	5.07 (3.78–6.37)*	< 0.001
Creatinine, μmol/L	72.6 (60.3–86.4)*	65.7 (51.8–79.6)*	< 0.001
Uric acid, μmol/L	303.3 (245.4–367.7)*	275.6 (203.1–341.6)*	< 0.001
Total cholesterol, mmol/L	3.45 (2.69–4.25)*	3.28 (1.22–4.34)*	< 0.001
Triglyceride, mmol/L	1.25 (0.85–1.77)*	0.95 (0.22–1.56)*	< 0.001
Apolipoprotein A1, g/L	0.89 (0.23–1.21)*	0.24 (0.23–1.22)*	0.066
Apolipoprotein B, g/L	0.45 (0.19–0.79)*	0.21 (0.18–0.84)*	0.120
High-density lipoprotein cholesterol, mmol/L	0.98 (0.77–1.18)*	0.87 (0.07–1.15)*	< 0.001
Low-density lipoprotein cholesterol, mmol/L	2.02 (1.38–2.76)*	1.83 (0.08–2.83)*	< 0.001
Lipoprotein(a), mg/dL	0.00 (0–0.65)*	0.00 (0–8.92)*	< 0.001
Apolipoprotein A1/Apolipoprotein B	1.28 (1.19–1.57)*	1.28 (1.18–1.43)*	< 0.001
Estimated glomerular filtration rate, mL/min per 1.73 m^2	95.9 (78.9–115.1)*	105.4 (87.1–134.3)*	< 0.001

Data are presented as n (%). *Presented as median (interquartile range).

rithm.[29] In our study, we applied RFE as a filter for muting irrelevant features in the process of feature selection. RFE was a feature selection method that trained a model and removed the weakest feature (or features) until a specified number of features were reached. The features were sorted by feature_importances_attributes of the model. In contrast to other methods, RFE took the final desired number of features to use as input, and then recursively reduced the number of features to use that attempts

Table 4 Performance of the five models on the validation sets.

Model	AUC	95% CI	Precision	Accuracy	F1-score	Sensitivity	Specificity
XGBoost	0.880	0.872–0.887	0.748	0.801	0.769	0.792	0.808
RF	0.877	0.869–0.884	0.771	0.811	0.776	0.781	0.833
DT	0.858	0.849–0.865	0.770	0.802	0.760	0.750	0.839
Adaboost	0.869	0.862–0.877	0.692	0.772	0.751	0.820	0.737
LR	0.797	0.788–0.807	0.682	0.741	0.698	0.715	0.761

Adaboost: adaptive boosting; AUC: area under the curve; DT: decision tree; LR: logistic regression; RF: random forest; XGBoost: extreme gradient boosting.

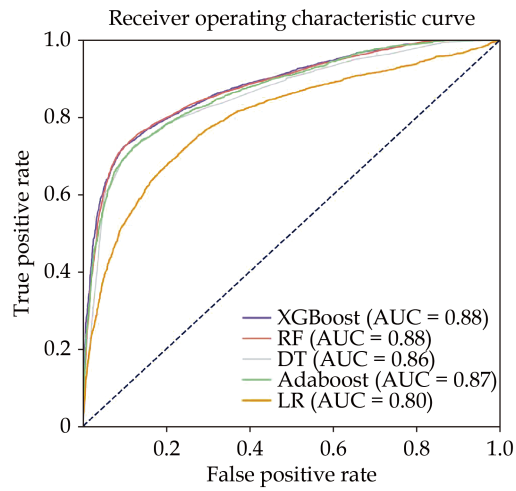

Figure 3 The receiver operating characteristic and the AUC of the five prediction models on the newly recruited dataset. Adaboost: adaptive boosting; AUC: area under the curve; DT: decision tree; LR: logistic regression; RF: random forest; XGBoost: extreme gradient boosting.

to eliminate dependencies and collinearity that may exist in the model.

A variety of feature categories were used to detect CHD. Some datasets only contain ECG features,[21,31] while others contain demographic, laboratory, symptom and examination, fluoroscopy and echo features.[32] In most CHD datasets, demographic features were widely used, such as age and sex. Other features, such as genetic features, were seldom used. Most datasets only show whether the patients have CHD. Few of them contain information on the stenosis severities of the three main arteries of the heart, which is the main limitation of these datasets. In our study, the variables included in the screening were mostly demographic and laboratory examinations. Compared to other categories of features, it is much easier to collect data that belong to these categories. Unfortunately, with the ML method, we have not found a new specific index that can well identify CHD in elderly patients with DM. However, we still believe that the auxiliary diagnosis system based on such research will effectively serve the clinic in the future and play a positive role in the screening, early warning and early diagnosis of CHD in elderly patients with DM.

Among the five models, RF, Adaboost and XGBoost showed better prediction performance. The RF algorithm performed random sampling, and the trained model had small variance and strong generalization ability, but on a sample set with relatively large noise, the RF model was prone to overfitting. Adaboost had high accuracy and fully considered the weight of each classifier, however, when the data was unbalanced, it would lead to a decrease in the classification accuracy of the model. XGBoost added a regular term to the objective function to control the complexity of the model and made the learned model simpler. XGBoost drew on the practice of RF and supported column sampling, which could not only reduce overfitting, but also reduced the workload of calculation. The above three ML algorithms used in this study minimized error and improved prediction accuracy, and identified other latent variables that were not easily observed, but the "black box" characteristics of these models were more difficult to explain, that is, they could not explain the inherent complexity of how risk factor variables interact and their independent effects on outcomes. In the testing sets, the sensitivity of the XGBoost model was not ideal (0.690). However, in the validation sets, the sensitivity was 0.792, which reflected that the XGBoost model had good transportability and generalizability. In the future, we will incorporate richer features and try more algorithms to further improve the performance of the model.

STRENGTHS AND LIMITATIONS

The strengths of this study are in the specific study population, large sample size, and multiple ML models. We collected the hospitalization information of tens of thousands of elderly patients with DM from the past ten years. We used five ML algorithms to establish models. Their performance was assessed with external validation by thousands of patients. We also obtained a prediction model for CHD in elderly patients with DM. However, there were some limitations that must be noted. Firstly, the variables included in the screening were mostly laboratory examinations. The inclusion of symptoms, new biomarkers, environmental factors, ECGs, cardiac ultrasounds and other data may further improve the prediction efficiency of the model. Secondly, doctors may miss diagnosis when writing medical records, which is also the limitation of this study based on medical records. Last but not least, the datasets for training and validation were from the same hospital. If the research is conducted in multiple centres, models would be more substantial and robust. Using larger datasets, more features and ML approaches may achieve better results.

CONCLUSIONS

In summary, this study established a ML model to predict CHD in elderly patients with DM, which may provide a reference for the early detection and intervention of CHD in elderly patients with DM.

ACKNOWLEDGMENTS

This study was supported by the Key Project of Chinese Military Health Care Projects (No.18BJZ32), the Projects of International Cooperation and Exchanges NSFC (No.81820108019), the Technical Fund for the Foundation Strengthening Program of China (2021-JCJG-JJ-1079), the Chinese Military Innovation Project (CX19028), and the Project of National Clinical Research Center for Geriatric Disease (NCRCG-PLAGH-2019024). All authors had no conflicts of interest to disclose.

REFERENCES

[1] Tinetti ME, Fried TR, Boyd CM. Designing health care for the most common chronic condition-multimorbidity. *JAMA* 2012; 307: 2493–2494.
[2] Bähler C, Huber CA, Brüngger B, *et al*. Multimorbidity, health care utilization and costs in an elderly community-dwelling population: a claims data based observational study. *BMC Health Serv Res* 2015; 15: 23.
[3] Zheng Y, Ley SH, Hu FB. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. *Nat Rev Endocrinol* 2018; 14: 88–98.
[4] The Writing Committee of the Report on Cardiovascular Health and Diseases in China. [Report on cardiovascular health and diseases burden in China: an updated summary of 2020]. *Chin Circ J* 2021; 36: 521–545. [In Chinese].
[5] Bragg F, Holmes MV, Iona A, *et al*. Association between diabetes and cause-specific mortality in rural and urban areas of China. *JAMA* 2017; 317: 280–289.
[6] Sarwar N, Gao P, Seshasai SR, *et al*. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *Lancet* 2010; 375: 2215–2222.
[7] Juutilainen A, Lehto S, Rönnemaa T, *et al*. Type 2 diabetes as a "coronary heart disease equivalent": an 18-year prospective population-based study in Finnish subjects. *Diabetes Care* 2005; 28: 2901–2907.
[8] American Diabetes Association. Cardiovascular disease and risk management: standards of medical care in diabetes-2020. *Diabetes Care* 2020; 43: S111–S134.
[9] Di Angelantonio E, Kaptoge S, Wormser D, *et al*. Association of cardiometabolic multimorbidity with mortality. *JAMA* 2015; 314: 52–60.
[10] Chinese Diabetes Society. [Guideline for the prevention and treatment of type 2 diabetes mellitus in China (2020 edition)]. *Chin J Diabetes Mellitus* 2021; 13: 315–409. [In Chinese].
[11] Bartnik M, Malmberg K, Hamsten A, *et al*. Abnormal glucose tolerance-a common risk factor in patients with acute myocardial infarction in comparison with population-based controls. *J Intern Med* 2004; 256: 288–297.
[12] Dinh A, Miertschin S, Young A, *et al*. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Med Inform Decis Mak* 2019; 19: 211.
[13] Segar MW, Vaduganathan M, Patel KV, *et al*. Machine learning to predict the risk of incident heart failure hospitalization among patients with diabetes: the WATCH-DM risk score. *Diabetes Care* 2019; 42: 2298–2306.
[14] Gautam D, Ahmed M, Meena YK, *et al*. Machine learning-based diagnosis of melanoma using macro images. *Int J Numer Method Biomed Eng* 2018; 34: e2953.
[15] Battineni G, Sagaro GG, Chinatalapudi N, *et al*. Applications of machine learning predictive models in the chronic disease diagnosis. *J Pers Med* 2020; 10: 21.
[16] Alizadehsani R, Habibi J, Hosseini MJ, *et al*. A data mining approach for diagnosis of coronary artery disease. *Comput Methods Programs Biomed* 2013; 111: 52–61.
[17] Ayatollahi H, Gholamhosseini L, Salehi M. Predicting coronary artery disease: a comparison between two data mining algorithms. *BMC Public Health* 2019; 19: 448.
[18] Arabasadi Z, Alizadehsani R, Roshanzamir M, *et al*. Computer aided decision making for heart disease detec-

tion using hybrid neural network-genetic algorithm. *Comput Methods Programs Biomed* 2017; 141: 19–26.
[19] Singh P, Singh S, Pandi-Jain GS. Effective heart disease prediction system using data mining techniques. *Int J Nanomedicine* 2018; 13: 121–124.
[20] Lih OS, Jahmunah V, San TR, *et al*. Comprehensive electrocardiographic diagnosis based on deep learning. *Artif Intell Med* 2020; 103: 101789.
[21] Tan JH, Hagiwara Y, Pang W, *et al*. Application of stacked convolutional and long short-term memory network for accurate identification of CAD ECG signals. *Comput Biol Med* 2018; 94: 19–26.
[22] Davari Dolatabadi A, Khadem SEZ, Asl BM. Automated diagnosis of coronary artery disease (CAD) patients using optimized SVM. *Comput Methods Programs Biomed* 2017; 138: 117–126.
[23] Lin S, Li Z, Fu B, *et al*. Feasibility of using deep learning to detect coronary artery disease based on facial photo. *Eur Heart J* 2020; 41: 4400–4411.
[24] Knuuti J, Wijns W, Saraste A, *et al*. 2019 ESC guidelines for the diagnosis and management of chronic coronary syndromes. *Eur Heart J* 2020; 41: 407–477.
[25] Handelman GS, Kok HK, Chandra RV, *et al*. eDoctor: machine learning and the future of medicine. *J Intern Med* 2018; 284: 603–619.
[26] Huynh T, Gao Y, Kang J, *et al*. Estimating CT image from MRI data using structured random forest and auto-context model. *IEEE Trans Med Imaging* 2016; 35: 174–183.
[27] Wu MT. Confusion matrix and minimum cross-entropy metrics based motion recognition system in the classroom. *Sci Rep* 2022; 12: 3095.
[28] Alizadehsani R, Khosravi A, Roshanzamir M, *et al*. Coronary artery disease detection using artificial intelligence techniques: a survey of trends, geographical differences and diagnostic features 1991–2020. *Comput Biol Med* 2021; 128: 104095.
[29] Alizadehsani R, Abdar M, Roshanzamir M, *et al*. Machine learning-based coronary artery disease diagnosis: a comprehensive review. *Comput Biol Med* 2019; 111: 103346.
[30] Fan R, Zhang N, Yang L, *et al*. AI-based prediction for the risk of coronary heart disease among patients with type 2 diabetes mellitus. *Sci Rep* 2020; 10: 14457.
[31] Kampouraki A, Manis G, Nikou C. Heartbeat time series classification with support vector machines. *IEEE Trans Inf Technol Biomed* 2009; 13: 512–518.
[32] Abdar M, Książek W, Acharya UR, *et al*. A new machine learning technique for an accurate diagnosis of coronary artery disease. *Comput Methods Programs Biomed* 2019; 179: 104992.

Please cite this article as: XU H, CAO WZ, BAI YY, DONG J, CHE HB, BAI P, WANG JD, CAO F, FAN L. Establishment of a diagnostic model of coronary heart disease in elderly patients with diabetes mellitus based on machine learning algorithms. *J Geriatr Cardiol* 2022; 19(6): 445–455. DOI: 10.11909/j.issn.1671-5411.2022.06.006