Cardiovascular Diabetology

RESEARCH

Open Access

Construction of machine learning diagnostic models for cardiovascular pan-disease based on blood routine and biochemical detection data

Zhicheng Wang[1,2,3†], Ying Gu[1†], Lindan Huang[1,2†], Shuai Liu[1], Qun Chen[1], Yunyun Yang[2*], Guolin Hong[2*] and Wanshan Ning[1*]

Abstract

Background Cardiovascular disease, also known as circulation system disease, remains the leading cause of morbidity and mortality worldwide. Traditional methods for diagnosing cardiovascular disease are often expensive and time-consuming. So the purpose of this study is to construct machine learning models for the diagnosis of cardiovascular diseases using easily accessible blood routine and biochemical detection data and explore the unique hematologic features of cardiovascular diseases, including some metabolic indicators.

Methods After the data preprocessing, 25,794 healthy people and 32,822 circulation system disease patients with the blood routine and biochemical detection data were utilized for our study. We selected logistic regression, random forest, support vector machine, eXtreme Gradient Boosting (XGBoost), and deep neural network to construct models. Finally, the SHAP algorithm was used to interpret models.

Results The circulation system disease prediction model constructed by XGBoost possessed the best performance (AUC: 0.9921 (0.9911–0.9930); Acc: 0.9618 (0.9588–0.9645); Sn: 0.9690 (0.9655–0.9723); Sp: 0.9526 (0.9477–0.9572); PPV: 0.9631 (0.9592–0.9668); NPV: 0.9600 (0.9556–0.9644); MCC: 0.9224 (0.9165–0.9279); F1 score: 0.9661 (0.9634–0.9686)). Most models of distinguishing various circulation system diseases also had good performance, the model performance of distinguishing dilated cardiomyopathy from other circulation system diseases was the best (AUC: 0.9267 (0.8663–0.9752)). The model interpretation by the SHAP algorithm indicated features from biochemical detection made major contributions to predicting circulation system disease, such as potassium (K), total protein (TP),

†Zhicheng Wang, Ying Gu and Lindan Huang have contributed equally to this work.

*Correspondence:
Yunyun Yang
yyy1213yyy@126.com
Guolin Hong
xmhgl9899@xmu.edu.cn
Wanshan Ning
ningwanshan@xmu.edu.cn

Full list of author information is available at the end of the article

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

albumin (ALB), and indirect bilirubin (NBIL). But for models of distinguishing various circulation system diseases, we found that red blood cell count (RBC), K, direct bilirubin (DBIL), and glucose (GLU) were the top 4 features subdividing various circulation system diseases.

Conclusions The present study constructed multiple models using 50 features from the blood routine and biochemical detection data for the diagnosis of various circulation system diseases. At the same time, the unique hematologic features of various circulation system diseases, including some metabolic-related indicators, were also explored. This cost-effective work will benefit more people and help diagnose and prevent circulation system diseases.

Keywords Metabolic indicator, Blood routine, Biochemical detection, Machine learning, Cardiovascular disease, Circulation system disease

Background

Cardiovascular diseases (CVDs), also known as circulatory system diseases, encompass a range of conditions including coronary heart disease (CHD), cerebrovascular disease, arrhythmias, valvular heart disease, cardiomyopathy, heart failure, and other related disorders [1]. With the widespread adoption of unhealthy lifestyle habits, CVDs continue to be the leading cause of mortality and morbidity worldwide, imposing a significant health burden and economic strain on both patients and society [2, 3]. The impact of CVDs is particularly severe in China. According to the China Health Statistical Yearbook 2021, CVDs rank first in both morbidity and mortality rates among urban and rural residents, surpassing cancer and other diseases [4].

Traditional diagnostic approaches for CVDs, including electrocardiograms (ECG), echocardiography, coronary angiography, stress testing, magnetic resonance imaging, and intracoronary ultrasonography, are often costly and not ideal for early-stage detection [5]. These methods are frequently inaccessible to primary healthcare facilities and economically disadvantaged regions due to the prohibitive costs of the required equipment. Moreover, many CVDs are asymptomatic in their early stages, and their progression can be slow, leading to clinical diagnoses often occurring at an advanced stage of the disease or incidentally during routine check-ups or assessments for other conditions. Therefore, it is crucial to identify more accessible and early screening indicators for CVDs.

Clinical laboratory tests, including hematological and biochemical analyses, provide quantitative measurements in the blood of both xenobiotics (foods, drugs, and their metabolites) and biotics (biomarkers) using validated, robust assays [6, 7]. Biochemical changes induced by disease can significantly impact various aspects of bioanalysis. Specifically, metabolic changes such as hyperglycemia, hypertriglyceridemia, high-density lipoprotein (HDL), cholesterol, hypertension, and a pro-inflammatory state are often present even in the early stages of CVDs [8]. However, doctors often focus on significantly abnormal parameters, potentially overlooking a substantial amount of other test data and the interrelationships between laboratory parameters, which may lead to an underestimation of the diagnostic potential of these tests. Therefore, it is essential to study the reference range and variation characteristics of hematological and biochemical indicators for early identification of preventable risk factors and early-stage CVD diagnosis, especially for indicators related to metabolic health, to assist doctors in early-stage CVD detection.

With the advancement of electronic medical record systems, an increasing amount of clinical laboratory test data has become more accessible and reliable. The use of this data, in combination with artificial intelligence (AI), for disease diagnosis, prediction, monitoring, and prognosis is a rapidly growing field [9, 10]. Machine learning (ML), a subset of AI, has shown great promise in aiding the diagnosis of CVDs [1, 11–13]. Current ML-based studies on CVDs generally overlook clinical laboratory test data, instead focusing on more expensive and/or invasive imaging techniques such as computed tomography angiography (CTA), heart ultrasound, computed tomography (CT), ECG, and echocardiography [14–18]. Additionally, existing research often emphasizes predicting the risk and prognosis of individual diseases [19, 20]. However, there is limited systematic analysis of the distinguishing features and unique hematological characteristics of CVDs.

In summary, this study aims to address several key questions: (1) to develop cost-effective, large-scale screening models based on blood routine and biochemical test data using clinical data from the First Affiliated Hospital of Xiamen University. The models we developed, after undergoing multiple rounds of parameter optimization, have achieved high accuracy. These models can accurately distinguish between cardiovascular disease patients and healthy individuals, as well as differentiate between most types of cardiovascular diseases; (2) to leverage the strengths of machine learning to explore the diagnostic performance of multi-indicator combinations in blood routine and biochemical test data, identifying universal indicators for the diagnosis and classification

of cardiovascular diseases; (3) to systematically compare and evaluate the unique hematological and metabolic characteristics of cardiovascular disease patients, providing clinicians with specialized insights for diagnosis and disease prevention.

Methods

Data collection and processing

All the raw data we collected came from inpatients in the Departments of Neurology and Cardiology and healthy people who had physical examinations in the First Affiliated Hospital of Xiamen University between 2018 and 2023. These data were from the hospital information system. For all patients, we screened the blood routine and biochemical test data from the first test after hospitalization as features for the construction of models, while for healthy people, we selected the blood routine and biochemical test data from the first physical examination every year as features. Because too many missing values may affect the prediction accuracy, we removed the features with a missing value ratio greater than 50% and finally screened out 22 features from the blood routine and 28 features from the biochemical test data (Supplementary Tables 1 and 2). Diagnostic information for all patients was determined according to The International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10). To ensure that the sample size for each circulation system disease was sufficient, we removed circulation system diseases with fewer than 100 samples. At the same time, we also deleted samples with a greater proportion of than 50% missing features. In the end, 25,794 healthy people and 32,822 patients with circulation system disease were used to construct our models (Fig. 1; Table 1). These data were randomly divided into a training set (70%) and a validation set (30%).

Machine learning methods

Logistic regression (LR), also known as logistic regression analysis, is a generalized linear regression analysis model, which is often used in data mining, automatic disease diagnosis, economic forecasting, and other fields. Logistic regression estimates the probability of an event occurring based on a given dataset of independent variables, and since the outcome is a probability, the dependent variable ranges between 0 and 1. Random forest (RF) is a classifier with many decision trees, which can be used to deal with classification and regression problems, as well as for dimensionality reduction problems. It also has a good tolerance for outliers and noise and has

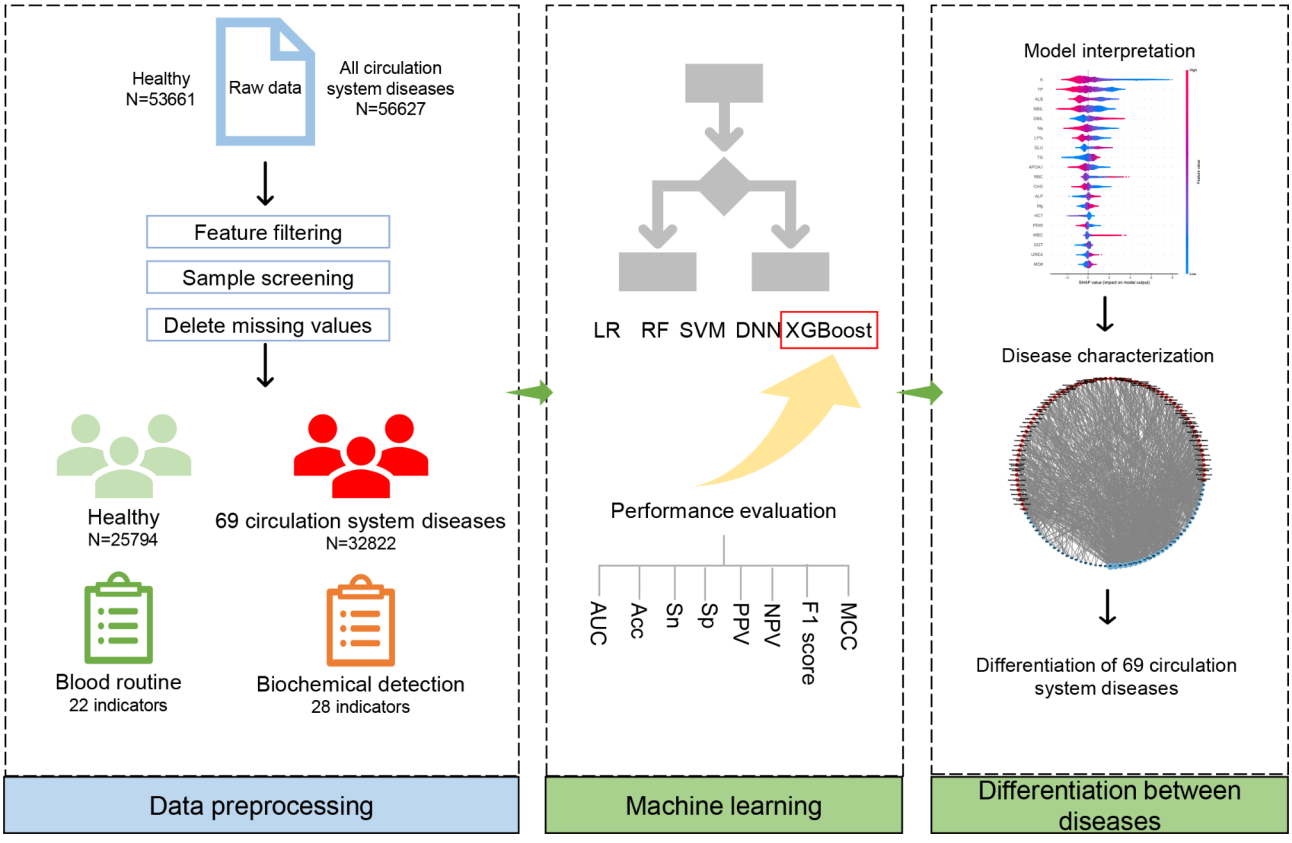

Fig. 1 The flow chart of this study

Table 1 Data distribution of diseases

ICD-10 disease code	Disease	Number
I10.x00	Idiopathic (primary) hypertension	1935
I10.x00×002	Hypertension	3121
I10.x00×028	Hypertension grade 2 (very high risk)	226
I10.x00×032	Hypertension grade 3 (very high risk)	701
I10.x03	Hypertension grade 1	269
I10.x04	Hypertension grade 2	867
I10.x05	Hypertension grade 3	2188
I11.901	Hypertensive heart disease	105
I20.000	Unstable angina	1655
I21.401	Acute non-ST-elevation myocardial infarction	142
I21.900	Acute myocardial infarction	233
I25.102	Coronary atherosclerosis	997
I25.103	Coronary atherosclerotic heart disease	2479
I42.000	Dilated cardiomyopathy	129
I45.102	Complete right bundle branch block	147
I47.101	Atrial tachycardia	101
I47.110	Atrial reentrant tachycardia	152
I48.x01	Atrial fibrillation	933
I48.x02	Paroxysmal atrial fibrillation	251
I49.100	Premature atrial depolarization	383
I49.100×001	Atrial premature contractions (premature atrial contractions)	291
I49.300	Premature ventricular depolarization	593
I49.301	Frequent ventricular extraphase contractions	102
I49.500	Sick sinus syndrome	145
I49.900	Arrhythmia	958
I50.900×002	Cardiac insufficiency	175
I60.900	Subarachnoid hemorrhage	324
I61.004	Basal ganglia hemorrhage	382
I61.101	Lobar hemorrhage	105
I61.500×001	Ventricular hemorrhage	103
I61.802	Hemorrhage in the thalamus	205
I61.900	Intracerebral hemorrhage	229
I62.001	Subdural hematoma	116
I62.003	Chronic subdural hematoma	188
I62.900	Intracranial hemorrhage (non-traumatic)	178
I63.200	Cerebral infarction caused by occlusion or stenosis of the anterior artery into the brain	122
I63.300	Cerebral infarction caused by cerebral artery thrombosis	159
I63.501	Cerebral artery stenosis, cerebral infarction	316
I63.502	Cerebral artery occlusion, cerebral infarction	130
I63.800	Cerebral infarction, others	164
I63.801	Lacunar cerebral infarction	380
I63.900	Cerebral infarction	1643
I63.901	Brainstem infarction	320
I63.902	Massive cerebral infarction	187
I63.904	Cerebellar infarction	122
I63.905	Multiple cerebral infarctions	633
I63.906	Basal ganglia infarction	189
I63.907	Thalamic infarction	131
I65.001	Vertebral artery stenosis	366
I65.002	Vertebral artery occlusion	117
I65.102	Basilar artery stenosis	102
I65.200×001	Carotid artery stenosis	103
I65.201	Internal carotid artery stenosis	379

Table 1 (continued)

ICD-10 disease code	Disease	Number
I65.203	Internal carotid artery occlusion	264
I66.001	Middle cerebral artery stenosis	449
I66.002	Middle cerebral artery occlusion	285
I67.200	Cerebral atherosclerosis	456
I67.200×011	Cerebral atherosclerosis	379
I67.202	Internal carotid atherosclerosis	617
I67.500	Moyamoya disease	217
I69.100	Sequelae of intracerebral hemorrhage	188
I69.300	Sequelae of cerebral infarction	1098
I70.804	Subclavian atherosclerosis	361
I70.806	Carotid arteriosclerosis	828
I70.900	Systemic atherosclerosis	333
I70.900×003	Arteriosclerosis	291
I70.900×004	Atherosclerosis	127
I72.002	Internal carotid aneurysm	128
I80.303	Venous thrombosis of the lower extremities	130

ICD-10: The International Statistical Classification of Diseases and Related Health Problems 10th Revision

better prediction and classification performance than decision trees. Support vector machine (SVM) is a kind of generalized linear classifier that classifies data binarily according to supervised learning, and its decision boundary is the maximum margin hyperplane solved by the learning sample. eXtreme Gradient Boosting (XGBoost) is an algorithm or engineering implementation based on the Gradient Boosting Decision Tree (GBDT). XGBoost is efficient, flexible, and lightweight, and has been widely used in data mining, recommender systems, and other fields. The deep neural network (DNN) is a framework for deep learning, that is a neural network with at least one hidden layer. Similar to shallow neural networks, deep neural networks can also provide modeling for complex nonlinear systems, but the extra layers provide a higher level of abstraction for the model, thus improving the model's capabilities. LR can optimize features through regularization. RF naturally reduces the impact of feature noise by combining multiple decision trees, thereby optimizing feature usage. SVM uses kernel functions and regularization parameters to find an appropriate hyperplane in high-dimensional space, indirectly affecting feature selection and optimization. XGBoost optimizes feature usage in the tree structure through gradient boosting. DNN can automatically learn and optimize features, particularly when dealing with complex data, by progressively extracting and refining features through multiple hidden layers. In summary, each of these algorithms has its strengths in feature optimization. For comparing the performance of different machine learning methods, we selected LR, RF, SVM, XGBoost, and DNN to construct the model [21–25].

To eliminate the impact of different feature scales on the accuracy of the prediction models, we standardized both the training and validation sets. We then performed hyperparameter selection for five machine learning algorithms using a combination of grid search cross-validation (CV) and manual fine-tuning. The parameters adjusted for LR were C, max_iter, penalty, and solver. For RF, the parameters were max_depth, min_samples_leaf, and n_estimators. For SVM, the parameters adjusted were C, gamma, and kernel. For XGBoost, the parameters were colsample_bytree, gamma, learning_rate, max_depth, n_estimators, and subsample. For DNN, the adjusted parameters included activation, number of layers, and number of neurons per layer. All optimal parameters were determined within the training set for the models distinguishing cardiovascular disease patients from healthy individuals. A 5-fold cross-validation was employed, with area under the curve (AUC) serving as the primary performance evaluation metric, to identify the best estimator (Supplementary Data 1).

The LR, RF, and SVM were used through scikit-learn (version 1.3.0), XGBoost was used through the xgboost package (version 2.0.2), and the DNN by tensorflow (version 2.0.2) in python.

Model performance evaluation

All models were trained using the best estimator and then validated on the validation set. Sensitivity (Sn), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV), F1 score, matthews correlation coefficient (MCC), and accuracy (Acc) were utilized for model performance evaluation. Their formulas are shown below [26–28]:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}$$

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}}$$

$$\text{F1 score} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FN} + \mathrm{FP}}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative separately. Meanwhile, we also made use of the AUC of the receiver operating characteristics curve (ROC) to evaluate the model performance comprehensively. Additionally, to further assess the robustness of the models, all performance evaluation metrics were calculated on the validation set using the bootstrapping method to determine their 95% confidence intervals (CI) [29–31].

Model interpretation

Machine learning makes it difficult to explain the contribution of each feature due to its black-box principle, so the SHAP algorithm was introduced in this study. The SHAP algorithm assigns a SHAP value to each feature, which is used to explain the impact of the feature on the predictive model [32]. The SHAP value of each feature was computed by the shap python package (version 0.44.0).

Identification of features for various types of CVDs

To identify the unique hematological and metabolic features of various cardiovascular diseases, we applied the SHAP algorithm to calculate SHAP values for 50 features across the 69 models distinguishing between different diseases. To ensure that the raw SHAP values were accurately represented in the heatmap, we did not normalize the values. We then performed hierarchical clustering on both rows and columns of the heatmap, reordered them according to the clustering results, and finally plotted the heatmap using Python.

To further explore the universal features distinguishing between various diseases, we selected the top ten features from the 69 models and connected these features with the respective diseases in a network graph. The size of each feature's node in the network increases if it appears frequently among the top ten features across the models, indicating its potential as a universal distinguishing feature between the diseases. The network was visualized using Cytoscape (version 3.10.2) [33].

Results

Circulation system disease prediction model construction

To ensure the accuracy of our prediction models, the number of various circulation system diseases was all over 100 (Table 1). The male-to-female ratio between healthy people and circulation system disease patients was similar, all close to 1:1. The number of healthy people for 40–60 years old and circulation system disease patients for 60–80 years old was the most population, 12,828 and 18,868 respectively (Supplementary Fig. 1). Subsequently, we chose five machine learning methods (LR, RF, SVM, XGBoost, and DNN) and utilized 22 features from blood routine and 28 features from biochemical detection to construct the circulation system disease prediction models. The results showed the comprehensive performance of XGBoost was the best (AUC: 0.9921 (0.9911–0.9930); Acc: 0.9618 (0.9588–0.9645); Sn: 0.9690 (0.9655–0.9723); Sp: 0.9526 (0.9477–0.9572); PPV: 0.9631 (0.9592–0.9668); NPV: 0.9600 (0.9556–0.9644); MCC: 0.9224 (0.9165–0.9279); F1 score: 0.9661 (0.9634–0.9686)) (Table 2). Meanwhile, we also attempted to construct the models only using blood routine or biochemical detection data. We found the model performance of the blood routine combined with biochemical detection was the best (Fig. 2B-D and Supplementary Data 2). Considering the imbalance for sample number among 69 circulation system diseases, we also used each circulation system disease to construct 69 models. The AUC of these models were all beyond 0.9, the highest one reached 0.9996 (0.9992–0.9999) (Fig. 2A; Table 2). These models all showed nice performance and robustness (Table 2).

Classification of various circulation system diseases

To further subdivide various circulation system diseases, we constructed 69 models distinguishing a kind of circulation system disease from other circulation system diseases, such as distinguishing venous thrombosis of the lower extremities from other circulation system diseases. The XGBoost was selected to construct models because of its good performance. The results showed the AUC of these models ranged from 0.5256 to 0.9267. Surprisingly, the model performance of distinguishing dilated cardiomyopathy (DCM) from other circulation system diseases was the best. DCM is a type of cardiomyopathy characterized by enlargement of the left or both ventricles of the heart with systolic dysfunction. The diagnosis of DCM primarily depends on ultrasonic cardiogram and cardiac magnetic resonance, not the blood routine

Table 2 Model performance evaluation results (circulation system diseases vs. healthy, XGBoost)

Model	AUC (95%CI)	Acc (95%CI)	Sn (95%CI)	Sp (95%CI)	PPV (95%CI)	NPV (95%CI)	MCC (95%CI)	F1 score (95%CI)
Circulation system disease	0.9921 (0.9911–0.9930)	0.9618 (0.9588–0.9645)	0.9690 (0.9655–0.9723)	0.9526 (0.9477–0.9572)	0.9631 (0.9592–0.9668)	0.9600 (0.9556–0.9644)	0.9224 (0.9165–0.9279)	0.9661 (0.9634–0.9686)
Idiopathic (primary) hypertension	0.9934 (0.9903–0.9960)	0.9859 (0.9833–0.9882)	0.8357 (0.8038–0.8655)	0.9969 (0.9956–0.9981)	0.9517 (0.9332–0.9695)	0.9881 (0.9857–0.9903)	0.8846 (0.8640–0.9029)	0.8899 (0.8698–0.9078)
Hypertension	0.9777 (0.9718–0.9832)	0.9699 (0.9661–0.9734)	0.7956 (0.7710–0.8231)	0.9901 (0.9878–0.9923)	0.9029 (0.8816–0.9241)	0.9767 (0.9733–0.9801)	0.8313 (0.8115–0.8511)	0.8458 (0.8276–0.8640)
Hypertension grade 2 (very high risk)	0.9538 (0.9263–0.9758)	0.9923 (0.9905–0.9942)	0.3506 (0.2424–0.4546)	0.9987 (0.9979–0.9995)	0.7297 (0.5758–0.8724)	0.9936 (0.9919–0.9952)	0.5027 (0.3827–0.6035)	0.4737 (0.3584–0.5794)
Hypertension grade 3 (very high risk)	0.9803 (0.9694–0.9888)	0.9870 (0.9846–0.9894)	0.5817 (0.5102–0.6453)	0.9979 (0.9969–0.9990)	0.8832 (0.8298–0.9359)	0.9889 (0.9865–0.9910)	0.7110 (0.6548–0.7607)	0.7014 (0.6416–0.7537)
Hypertension grade 1	0.9750 (0.9485–0.9925)	0.9927 (0.9907–0.9944)	0.4286 (0.3234–0.5317)	0.9988 (0.9981–0.9995)	0.8000 (0.6785–0.9091)	0.9938 (0.9920–0.9954)	0.5825 (0.4853–0.6725)	0.5581 (0.4528–0.6552)
Hypertension grade 2	0.9818 (0.9719–0.9896)	0.9847 (0.9820–0.9871)	0.6078 (0.5474–0.6653)	0.9972 (0.9959–0.9983)	0.8757 (0.8246–0.9222)	0.9872 (0.9848–0.9895)	0.7225 (0.6758–0.7629)	0.7176 (0.6667–0.7594)
Hypertension grade 3	0.9873 (0.9821–0.9911)	0.9796 (0.9765–0.9824)	0.7973 (0.7665–0.8272)	0.9953 (0.9939–0.9968)	0.9365 (0.9167–0.9561)	0.9828 (0.9797–0.9856)	0.8536 (0.8315–0.8731)	0.8613 (0.8395–0.8801)
Hypertensive heart disease	0.9794 (0.9350–0.9986)	0.9976 (0.9964–0.9985)	0.4839 (0.3124–0.6667)	0.9996 (0.9991-1.0000)	0.8333 (0.6429-1.0000)	0.9979 (0.9969–0.9988)	0.6340 (0.4700-0.7683)	0.6122 (0.4324–0.7556)
Unstable angina	0.9987 (0.9981–0.9992)	0.9933 (0.9915–0.9949)	0.9146 (0.8886–0.9381)	0.9982 (0.9972–0.9991)	0.9691 (0.9514–0.9840)	0.9947 (0.9930–0.9963)	0.9380 (0.9207–0.9533)	0.9411 (0.9247–0.9557)
Acute non-ST-elevation myocardial infarction	0.9972 (0.9937–0.9994)	0.9970 (0.9958–0.9981)	0.5106 (0.3673-0.6400)	1.0000 (1.0000–1.0000)	1.0000 (1.0000–1.0000)	0.9970 (0.9957–0.9981)	0.7135 (0.6049–0.7991)	0.6761 (0.5373–0.7805)
Acute myocardial infarction	0.9993 (0.9986–0.9997)	0.9977 (0.9965–0.9987)	0.7887 (0.6866–0.8788)	0.9996 (0.9991-1.0000)	0.9492 (0.8800-1.0000)	0.9981 (0.9969–0.9990)	0.8641 (0.7972–0.9182)	0.8615 (0.7903–0.9167)
Coronary atherosclerosis	0.9828 (0.9721–0.9907)	0.9882 (0.9857–0.9904)	0.7352 (0.6848–0.7836)	0.9975 (0.9964–0.9986)	0.9174 (0.8804–0.9498)	0.9903 (0.9882–0.9923)	0.8155 (0.7802–0.8486)	0.8162 (0.7795–0.8507)
Coronary atherosclerotic heart disease	0.9957 (0.9935–0.9973)	0.9876 (0.9853–0.9899)	0.8893 (0.8674–0.9107)	0.9970 (0.9957–0.9982)	0.9663 (0.9522–0.9798)	0.9895 (0.9873–0.9918)	0.9204 (0.9057–0.9346)	0.9262 (0.9123–0.9396)
Dilated cardiomyopathy	0.9996 (0.9992–0.9999)	0.9985 (0.9976–0.9994)	0.7381 (0.6052–0.8788)	0.9999 (0.9996-1.0000)	0.9688 (0.8947-1.0000)	0.9986 (0.9977–0.9994)	0.8449 (0.7580–0.9255)	0.8378 (0.7418–0.9231)
Complete right bundle branch block	0.9749 (0.9476–0.9947)	0.9965 (0.9951–0.9978)	0.4444 (0.2922-0.6000)	0.9997 (0.9994-1.0000)	0.9091 (0.7778-1.0000)	0.9968 (0.9955–0.9981)	0.6343 (0.4947–0.7477)	0.5970 (0.4347–0.7251)
Atrial tachycardia	0.9885 (0.9786–0.9964)	0.9970 (0.9958–0.9982)	0.2333 (0.1000-0.4092)	1.0000 (1.0000–1.0000)	1.0000 (1.0000–1.0000)	0.9970 (0.9957–0.9982)	0.4823 (0.3157–0.6391)	0.3784 (0.1818–0.5807)
Atrial reentrant tachycardia	0.9665 (0.9260–0.9940)	0.9955 (0.9938–0.9970)	0.4043 (0.2632–0.5610)	0.9991 (0.9983–0.9997)	0.7308 (0.5500-0.8948)	0.9964 (0.9950–0.9977)	0.5416 (0.3975–0.6672)	0.5205 (0.3714–0.6512)
Atrial fibrillation	0.9937 (0.9895–0.9969)	0.9913 (0.9892–0.9934)	0.8000 (0.7529–0.8493)	0.9979 (0.9969–0.9990)	0.9310 (0.8991–0.9635)	0.9931 (0.9913–0.9947)	0.8587 (0.8259–0.8914)	0.8606 (0.8274–0.8933)

Table 2 (continued)

Model	AUC (95%CI)	Acc (95%CI)	Sn (95%CI)	Sp (95%CI)	PPV (95%CI)	NPV (95%CI)	MCC (95%CI)	F1 score (95%CI)
Paroxysmal atrial fibrillation	0.9850 (0.9692–0.9958)	0.9946 (0.9931–0.9962)	0.4815 (0.3766–0.5890)	1.0000 (1.0000–1.0000)	1.0000 (1.0000–1.0000)	0.9946 (0.9931–0.9961)	0.6920 (0.6118–0.7656)	0.6500 (0.5472–0.7413)
Premature atrial depolarization	0.9827 (0.9726–0.9905)	0.9924 (0.9905–0.9943)	0.5315 (0.4386–0.6286)	0.9990 (0.9982–0.9996)	0.8806 (0.7922–0.9531)	0.9933 (0.9915–0.9951)	0.6809 (0.6026–0.7499)	0.6629 (0.5765–0.7380)
Atrial premature contractions (premature atrial contractions)	0.9669 (0.9461–0.9839)	0.9925 (0.9905–0.9944)	0.4694 (0.3714–0.5686)	0.9991 (0.9983–0.9997)	0.8679 (0.7660–0.9574)	0.9933 (0.9915–0.9951)	0.6352 (0.5466–0.7161)	0.6093 (0.5124–0.6977)
Premature ventricular depolarization	0.9716 (0.9591–0.9823)	0.9867 (0.9842–0.9891)	0.4972 (0.4210–0.5731)	0.9982 (0.9973–0.9991)	0.8654 (0.7966–0.9271)	0.9884 (0.9859–0.9906)	0.6504 (0.5854–0.7085)	0.6316 (0.5611–0.6948)
Frequent ventricular extraphase contractions	0.9275 (0.8592–0.9829)	0.9967 (0.9952–0.9978)	0.2973 (0.1599-0.4500)	1.0000 (1.0000–1.0000)	1.0000 (1.0000–1.0000)	0.9966 (0.9952–0.9978)	0.5443 (0.3993–0.6699)	0.4583 (0.2757–0.6207)
Sick sinus syndrome	0.9982 (0.9964–0.9994)	0.9976 (0.9964–0.9986)	0.6042 (0.4633–0.7436)	1.0000 (1.0000–1.0000)	1.0000 (1.0000–1.0000)	0.9975 (0.9964–0.9986)	0.7763 (0.6797–0.8618)	0.7532 (0.6332–0.8529)
Arrhythmia	0.9812 (0.9744–0.9866)	0.9834 (0.9804–0.9863)	0.6416 (0.5827–0.6993)	0.9964 (0.9950–0.9977)	0.8704 (0.8220–0.9114)	0.9866 (0.9838–0.9891)	0.7394 (0.6969–0.7840)	0.7387 (0.6926–0.7837)
Cardiac insufficiency	0.9968 (0.9936–0.9991)	0.9973 (0.9961–0.9983)	0.7069 (0.5832–0.8149)	0.9995 (0.9990–0.9999)	0.9111 (0.8163–0.9811)	0.9978 (0.9968–0.9987)	0.8013 (0.7113–0.8754)	0.7961 (0.7000-0.8724)
Subarachnoid hemorrhage	0.9975 (0.9946–0.9994)	0.9969 (0.9957–0.9981)	0.8000 (0.7169–0.8785)	0.9996 (0.9991-1.0000)	0.9655 (0.9221-1.0000)	0.9973 (0.9961–0.9985)	0.8774 (0.8258–0.9234)	0.8750 (0.8191–0.9224)
Basal ganglia hemorrhage	0.9927 (0.9860–0.9976)	0.9949 (0.9933–0.9964)	0.7083 (0.6209–0.7911)	0.9994 (0.9987–0.9999)	0.9444 (0.8941–0.9885)	0.9955 (0.9939–0.9969)	0.8156 (0.7563–0.8688)	0.8095 (0.7425–0.8657)
Lobar hemorrhage	0.9824 (0.9546–0.9993)	0.9986 (0.9977–0.9994)	0.6774 (0.5000-0.8401)	0.9999 (0.9996-1.0000)	0.9545 (0.8500-1.0000)	0.9987 (0.9978–0.9995)	0.8035 (0.6786–0.9024)	0.7925 (0.6511–0.8980)
Ventricular hemorrhage	0.9987 (0.9967–0.9999)	0.9988 (0.9979–0.9996)	0.7714 (0.6110–0.9091)	0.9999 (0.9996-1.0000)	0.9643 (0.8710-1.0000)	0.9990 (0.9982–0.9996)	0.8619 (0.7604–0.9426)	0.8571 (0.7441–0.9412)
Hemorrhage in the thalamus	0.9793 (0.9610–0.9930)	0.9956 (0.9941–0.9969)	0.5469 (0.4200-0.6567)	0.9994 (0.9987–0.9999)	0.8750 (0.7659–0.9706)	0.9963 (0.9947–0.9974)	0.6899 (0.5858–0.7771)	0.6731 (0.5566–0.7692)
Intracerebral hemorrhage	0.9886 (0.9750–0.9972)	0.9969 (0.9956–0.9981)	0.7143 (0.6076–0.8193)	0.9995 (0.9990–0.9999)	0.9259 (0.8511–0.9828)	0.9974 (0.9963–0.9985)	0.8118 (0.7366–0.8796)	0.8065 (0.7273–0.8772)
Subdural hematoma	0.9910 (0.9799–0.9987)	0.9976 (0.9964–0.9986)	0.5952 (0.4358–0.7429)	0.9997 (0.9994-1.0000)	0.9259 (0.8077-1.0000)	0.9978 (0.9968–0.9988)	0.7414 (0.6138–0.8382)	0.7246 (0.5806–0.8333)
Chronic subdural hematoma	0.9840 (0.9567–0.9987)	0.9970 (0.9958–0.9982)	0.6364 (0.5091–0.7551)	0.9996 (0.9991-1.0000)	0.9211 (0.8235-1.0000)	0.9974 (0.9963–0.9985)	0.7643 (0.6697–0.8422)	0.7527 (0.6493–0.8364)
Intracranial hemorrhage (non-traumatic)	0.9971 (0.9955–0.9985)	0.9961 (0.9947–0.9976)	0.5714 (0.4339–0.7060)	0.9988 (0.9981–0.9995)	0.7568 (0.6176–0.8919)	0.9973 (0.9961–0.9983)	0.6558 (0.5298–0.7615)	0.6512 (0.5217–0.7595)
Cerebral infarction caused by occlusion or stenosis of the anterior artery into the brain	0.9837 (0.9729–0.9922)	0.9956 (0.9941–0.9970)	0.3095 (0.1739–0.4510)	0.9994 (0.9987–0.9999)	0.7222 (0.5000-0.9375)	0.9963 (0.9948–0.9976)	0.4711 (0.3037–0.6090)	0.4333 (0.2666–0.5833)

Table 2 (continued)

Model	AUC (95%CI)	Acc (95%CI)	Sn (95%CI)	Sp (95%CI)	PPV (95%CI)	NPV (95%CI)	MCC (95%CI)	F1 score (95%CI)
Cerebral infarction caused by cerebral artery thrombosis	0.9710 (0.9327–0.9947)	0.9949 (0.9933–0.9965)	0.3269 (0.2037–0.4615)	0.9994 (0.9987–0.9999)	0.7727 (0.5909–0.9460)	0.9955 (0.9939–0.9970)	0.5007 (0.3676–0.6256)	0.4595 (0.3188–0.5927)
Cerebral artery stenosis, cerebral infarction	0.9842 (0.9761–0.9910)	0.9925 (0.9904–0.9943)	0.4490 (0.3469–0.5532)	0.9994 (0.9988–0.9999)	0.8980 (0.8055–0.9737)	0.9931 (0.9911–0.9949)	0.6320 (0.5426–0.7168)	0.5986 (0.4960–0.6906)
Cerebral artery occlusion, cerebral infarction	0.9822 (0.9580–0.9955)	0.9973 (0.9961–0.9983)	0.5750 (0.4186–0.7180)	0.9995 (0.9990–0.9999)	0.8519 (0.7058–0.9643)	0.9978 (0.9968–0.9987)	0.6987 (0.5614–0.8099)	0.6866 (0.5422–0.8044)
Cerebral infarction, others	0.9821 (0.9673–0.9934)	0.9958 (0.9942–0.9972)	0.3469 (0.2195–0.4889)	0.9999 (0.9996-1.0000)	0.9444 (0.8180-1.0000)	0.9959 (0.9945–0.9973)	0.5711 (0.4433–0.6831)	0.5075 (0.3508–0.6486)
Lacunar cerebral infarction	0.9519 (0.9260–0.9761)	0.9911 (0.9889–0.9931)	0.5410 (0.4531–0.6305)	0.9982 (0.9972–0.9991)	0.8250 (0.7432–0.9079)	0.9928 (0.9909–0.9946)	0.6640 (0.5909–0.7345)	0.6535 (0.5766–0.7289)
Cerebral infarction	0.9891 (0.9853–0.9923)	0.9825 (0.9796–0.9852)	0.7780 (0.7414–0.8146)	0.9955 (0.9938–0.9970)	0.9161 (0.8866–0.9434)	0.9861 (0.9833–0.9886)	0.8353 (0.8090–0.8614)	0.8414 (0.8153–0.8666)
Brainstem infarction	0.9915 (0.9860–0.9953)	0.9922 (0.9902–0.9941)	0.5446 (0.4457–0.6422)	0.9981 (0.9970–0.9990)	0.7857 (0.6901-0.8800)	0.9941 (0.9923–0.9957)	0.6505 (0.5672–0.7291)	0.6433 (0.5548–0.7244)
Massive cerebral infarction	0.9947 (0.9888–0.9992)	0.9976 (0.9964–0.9986)	0.7143 (0.5918–0.8333)	0.9996 (0.9991-1.0000)	0.9302 (0.8511-1.0000)	0.9979 (0.9969–0.9988)	0.8140 (0.7290–0.8918)	0.8081 (0.7158–0.8889)
Cerebellar infarction	0.9735 (0.9538–0.9902)	0.9963 (0.9950–0.9977)	0.4048 (0.2571–0.5642)	0.9995 (0.9990-1.0000)	0.8095 (0.6429-1.0000)	0.9968 (0.9956–0.9981)	0.5709 (0.4277–0.6985)	0.5397 (0.3823–0.6769)
Multiple cerebral infarctions	0.9842 (0.9756–0.9920)	0.9898 (0.9878–0.9919)	0.6522 (0.5829–0.7222)	0.9978 (0.9966–0.9988)	0.8759 (0.8163–0.9265)	0.9918 (0.9898–0.9937)	0.7510 (0.6992–0.8004)	0.7477 (0.6931–0.7988)
Basal ganglia infarction	0.9855 (0.9726–0.9946)	0.9949 (0.9932–0.9964)	0.3455 (0.2222–0.4717)	0.9995 (0.9990–0.9999)	0.8261 (0.6667–0.9615)	0.9954 (0.9938–0.9969)	0.5323 (0.3932–0.6480)	0.4872 (0.3429–0.6154)
Thalamic infarction	0.9571 (0.9290–0.9808)	0.9955 (0.9940–0.9969)	0.3043 (0.1818–0.4444)	0.9996 (0.9992-1.0000)	0.8235 (0.6364-1.0000)	0.9959 (0.9943–0.9973)	0.4991 (0.3544–0.6287)	0.4444 (0.2857–0.5902)
Vertebral artery stenosis	0.9718 (0.9593–0.9830)	0.9920 (0.9901–0.9939)	0.4343 (0.3367–0.5341)	0.9991 (0.9984–0.9997)	0.8600 (0.7500-0.9524)	0.9928 (0.9909–0.9946)	0.6080 (0.5166–0.6918)	0.5772 (0.4733–0.6667)
Vertebral artery occlusion	0.9652 (0.9411–0.9838)	0.9958 (0.9942–0.9972)	0.2955 (0.1666–0.4250)	0.9997 (0.9994-1.0000)	0.8667 (0.6667-1.0000)	0.9960 (0.9946–0.9973)	0.5046 (0.3454–0.6244)	0.4407 (0.2712–0.5807)
Basilar artery stenosis	0.9898 (0.9815–0.9959)	0.9967 (0.9952–0.9979)	0.3243 (0.1749–0.4688)	0.9999 (0.9996-1.0000)	0.9231 (0.7500-1.0000)	0.9968 (0.9955–0.9981)	0.5461 (0.3823–0.6727)	0.4800 (0.2857–0.6342)
Carotid artery stenosis	0.9527 (0.8986–0.9923)	0.9965 (0.9951–0.9977)	0.2286 (0.0937–0.3751)	1.0000 (1.0000–1.0000)	1.0000 (1.0000–1.0000)	0.9965 (0.9951–0.9977)	0.4773 (0.3055–0.6116)	0.3721 (0.1713–0.5455)
Internal carotid artery stenosis	0.9752 (0.9574–0.9875)	0.9920 (0.9899–0.9939)	0.5299 (0.4380–0.6230)	0.9990 (0.9982–0.9996)	0.8857 (0.8088–0.9559)	0.9929 (0.9910–0.9947)	0.6817 (0.6050–0.7485)	0.6631 (0.5802–0.7383)
Internal carotid artery occlusion	0.9911 (0.9824–0.9970)	0.9941 (0.9925–0.9957)	0.6163 (0.5200-0.7177)	0.9983 (0.9973–0.9991)	0.8030 (0.6969–0.8919)	0.9957 (0.9943–0.9970)	0.7007 (0.6153–0.7803)	0.6974 (0.6099–0.7765)
Middle cerebral artery stenosis	0.9710 (0.9530–0.9859)	0.9902 (0.9879–0.9924)	0.5401 (0.4524–0.6228)	0.9982 (0.9972–0.9991)	0.8409 (0.7586–0.9131)	0.9919 (0.9899–0.9938)	0.6696 (0.5940–0.7337)	0.6578 (0.5803–0.7273)

Table 2 (continued)

Model	AUC (95%CI)	Acc (95%CI)	Sn (95%CI)	Sp (95%CI)	PPV (95%CI)	NPV (95%CI)	MCC (95%CI)	F1 score (95%CI)
Middle cerebral artery occlusion	0.9776 (0.9632–0.9899)	0.9931 (0.9913–0.9948)	0.4574 (0.3548–0.5568)	0.9996 (0.9991-1.0000)	0.9348 (0.8519-1.0000)	0.9934 (0.9916–0.9951)	0.6513 (0.5604–0.7276)	0.6143 (0.5116–0.7037)
Cerebral atherosclerosis (I67.200)	0.9826 (0.9748–0.9896)	0.9920 (0.9900-0.9938)	0.6015 (0.5217–0.6801)	0.9987 (0.9979–0.9995)	0.8889 (0.8256–0.9518)	0.9932 (0.9914–0.9949)	0.7276 (0.6609–0.7877)	0.7175 (0.6468–0.7807)
Cerebral atherosclerosis (I67.200×011)	0.9693 (0.9484–0.9847)	0.9897 (0.9874–0.9920)	0.4615 (0.3695–0.5596)	0.9977 (0.9966–0.9987)	0.7500 (0.6528–0.8507)	0.9919 (0.9897–0.9939)	0.5837 (0.5004–0.6723)	0.5714 (0.4810–0.6632)
Internal carotid atherosclerosis	0.9727 (0.9619–0.9821)	0.9850 (0.9823–0.9875)	0.4789 (0.4131–0.5481)	0.9974 (0.9962–0.9986)	0.8198 (0.7431–0.8889)	0.9873 (0.9849–0.9897)	0.6201 (0.5600–0.6800)	0.6047 (0.5407–0.6688)
Moyamoya disease	0.9335 (0.8938–0.9678)	0.9944 (0.9926–0.9959)	0.3279 (0.2187–0.4375)	0.9996 (0.9991-1.0000)	0.8696 (0.7143-1.0000)	0.9947 (0.9931–0.9961)	0.5320 (0.4081–0.6288)	0.4762 (0.3437–0.5870)
Sequelae of intracerebral hemorrhage	0.9819 (0.9607–0.9955)	0.9962 (0.9947–0.9974)	0.5455 (0.4181–0.6786)	0.9994 (0.9987–0.9999)	0.8571 (0.7273–0.9655)	0.9968 (0.9955–0.9979)	0.6821 (0.5687–0.7836)	0.6667 (0.5476–0.7789)
Sequelae of cerebral infarction	0.9900 (0.9837–0.9943)	0.9866 (0.9840–0.9890)	0.7391 (0.6938–0.7841)	0.9969 (0.9956–0.9981)	0.9084 (0.8739–0.9421)	0.9892 (0.9869–0.9914)	0.8128 (0.7817–0.8434)	0.8151 (0.7831–0.8458)
Subclavian atherosclerosis	0.9681 (0.9469–0.9831)	0.9890 (0.9866–0.9913)	0.4034 (0.3181–0.4912)	0.9981 (0.9970–0.9990)	0.7619 (0.6551–0.8616)	0.9909 (0.9886–0.9929)	0.5497 (0.4605–0.6257)	0.5275 (0.4347-0.6100)
Carotid arteriosclerosis	0.9803 (0.9721–0.9874)	0.9870 (0.9842–0.9894)	0.6707 (0.6107–0.7297)	0.9970 (0.9959–0.9982)	0.8777 (0.8342–0.9214)	0.9896 (0.9872–0.9918)	0.7610 (0.7169–0.8027)	0.7604 (0.7133–0.8025)
Systemic atherosclerosis	0.9858 (0.9769–0.9933)	0.9923 (0.9902–0.9944)	0.5377 (0.4392–0.6316)	0.9986 (0.9977–0.9994)	0.8382 (0.7411–0.9231)	0.9937 (0.9919–0.9954)	0.6680 (0.5871–0.7476)	0.6552 (0.5683–0.7352)
Arteriosclerosis (I70.900×003)	0.9739 (0.9579–0.9871)	0.9935 (0.9917–0.9951)	0.5408 (0.4400-0.6374)	0.9992 (0.9986–0.9997)	0.8983 (0.8113–0.9688)	0.9942 (0.9925–0.9957)	0.6943 (0.6158–0.7688)	0.6752 (0.5867–0.7558)
Atherosclerosis (I70.900×004)	0.9677 (0.9227–0.9947)	0.9956 (0.9941–0.9972)	0.2564 (0.1281-0.4000)	0.9994 (0.9987–0.9999)	0.6667 (0.4118-0.9000)	0.9963 (0.9948–0.9976)	0.4118 (0.2394–0.5592)	0.3704 (0.1967–0.5263)
Internal carotid aneurysm	0.9665 (0.9440–0.9857)	0.9964 (0.9950–0.9977)	0.3250 (0.1860–0.4706)	0.9999 (0.9996-1.0000)	0.9286 (0.7643-1.0000)	0.9965 (0.9951–0.9978)	0.5482 (0.3929–0.6732)	0.4815 (0.3110–0.6275)
Venous thrombosis of the lower extremities	0.9930 (0.9862–0.9987)	0.9972 (0.9960–0.9983)	0.6500 (0.5000-0.8001)	0.9990 (0.9982–0.9996)	0.7647 (0.6127–0.8920)	0.9982 (0.9973–0.9991)	0.7036 (0.5859–0.8165)	0.7027 (0.5797–0.8158)

95%CI: 95% confidence intervals (Lower–Upper bound)

and biochemical detection. These results indicated these models could help doctors well distinguish different circulation system diseases (Fig. 3 and Supplementary Data 3).

Analysis of circulation system disease-specific indicators

To help us better understand the contributions of 50 features for the circulation system disease prediction model and find the circulation system disease-specific indicators, we used the SHAP algorithm to compute the contribution degree of each feature. For the constructed model only utilizing the blood routine, the top 10 features were lymphocyte percentage (LY%), red blood cell count (RBC), absolute value of monocyte (MO#), hematocrit (HCT), absolute value of neutrophil (NE#), mean erythrocyte hemoglobin concentration (MCHC), plateletcrit (PCT), white blood cell count (WBC), platelet distribution width (PDW), and mean platelet volume (MPV) (Fig. 4A). For the constructed model only utilizing the biochemical detection data, the top 10 features were potassium (K), albumin (ALB), total protein (TP), indirect bilirubin (NBIL), direct bilirubin (DBIL), sodium (Na), glucose (GLU), triglycerides (TG), cholesterol (CHO), Apolipoprotein A1 (APOA1) (Fig. 4B).

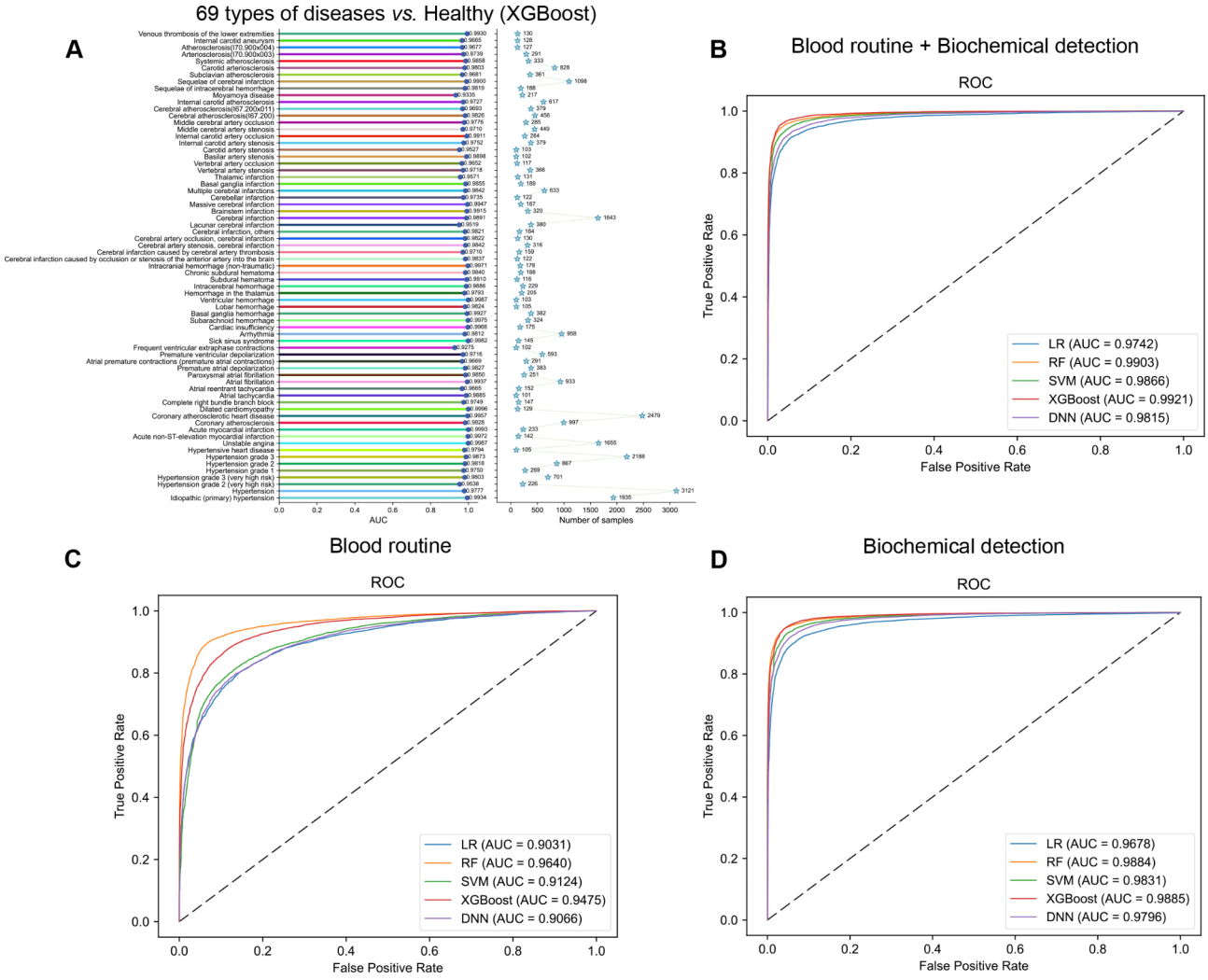

Fig. 2 Construction of circulation system disease prediction model using clinical blood samples. (A) The AUC of 69 circulation system disease prediction models. ROC curves of five machine learning methods using different data. (B) Blood routine combined with biochemical detection. (C) Blood routine. (D) Biochemical detection

Interestingly, for the constructed model utilizing the blood routine combined with biochemical detection, only one feature from the blood routine, LY%, was one of the top 10 features (Fig. 4C). These results indicated features from biochemical detection made major contributions to predicting circulation system disease (Fig. 4D). Additionally, to further validate the importance of these features, we also calculated the top 20 features ranked by the other four machine learning methods. As shown in the results, although there were slight variations in feature rankings across different methods, there was considerable overlap among the top 20 features, indicating that our model interpretation approach demonstrates good stability (Supplementary Data 4).

To verify whether the performance of the XGBoost was affected by redundant features, we constructed the model to distinguish cardiovascular disease patients from healthy individuals using only the top 10 features ranked by the SHAP algorithm (Fig. 4C). The results showed the model built using all 50 features performed better, indicating that the performance of our models was not impacted by redundant features (Supplementary Fig. 2).

Analysis of characteristic indicators of discrimination between various circulation system diseases

After exploring circulation system disease-specific indicators, we also hoped to further explore characteristic indicators of discrimination between various circulation system diseases. Then, we displayed the SHAP value of each feature through a heatmap. Rows and columns were clustered separately, and the more similar the features or diseases, the closer they were. We found that every circulation system disease had distinctive characteristics (Fig. 5A and Supplementary Data 5). At the same time,

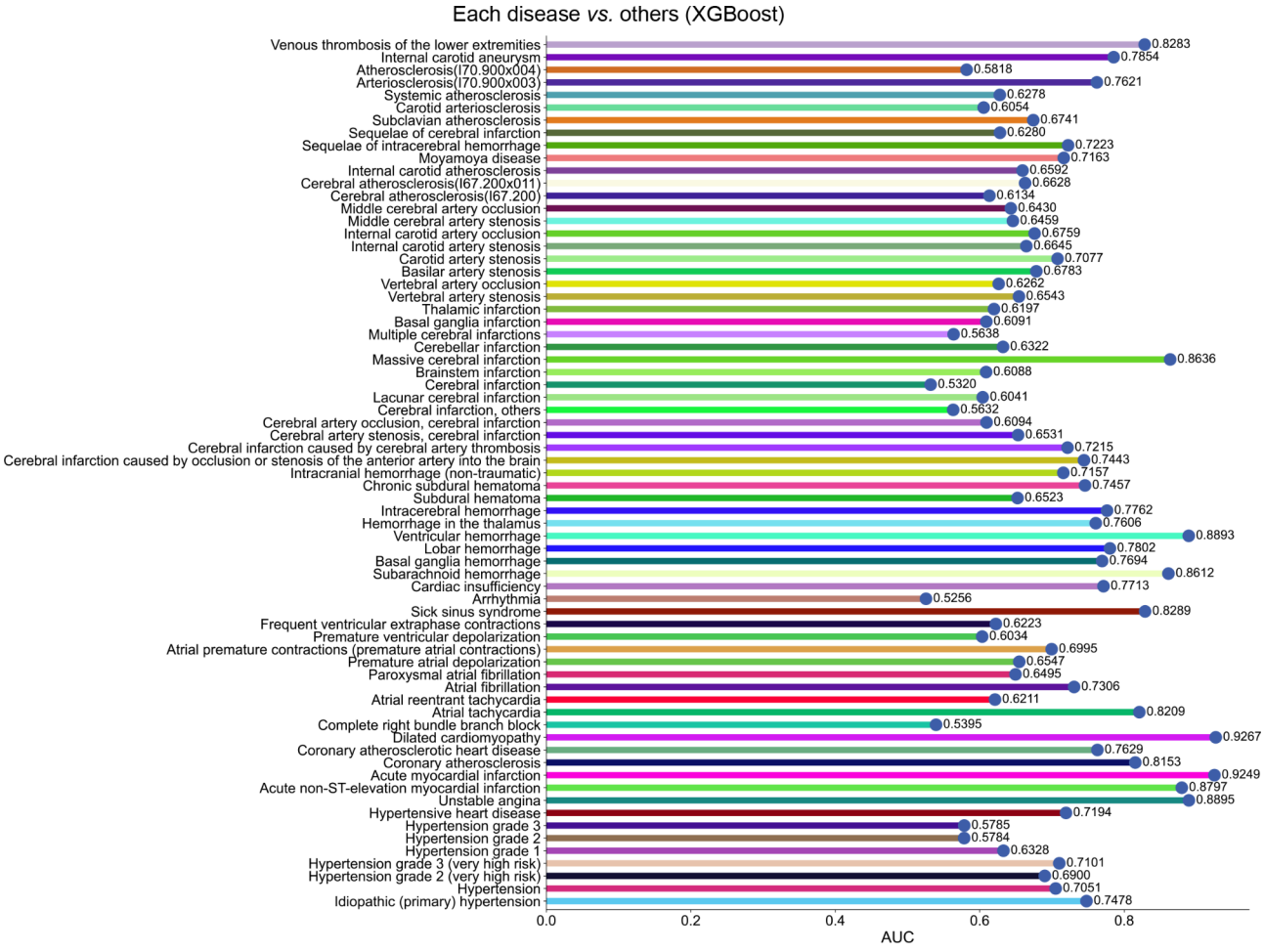

Fig. 3 The AUC of 69 models distinguishes a kind of circulation system disease from others

a network displaying the intersection features among various circulation system diseases showed that RBC, K, DBIL, and GLU were the top 4 features subdividing various circulation system diseases (Fig. 5B). Elevated GLU is often associated with diabetes, but we found that GLU could also be used to distinguish between different circulation system diseases. DBIL, also known as conjugated bilirubin, is produced by the combination of indirect bilirubin into the liver by the action of intrahepatic glucuronosyltransferase and glucuronic acid, and its elevation is usually related to various liver dysfunctions. But as we can see in our results, it also has great potential for predicting various circulation system diseases. The numerical distributions of the top 4 features among various circulation system diseases and healthy people were different (Fig. 5C). The results proved our models were reliable.

Discussion

Cardiovascular disease (CVD) remains the leading cause of death globally [2]. Early-stage detection of CVD is an important way of reducing this toll. An advanced detection of cardiovascular disease is required to improve therapeutic strategies and patient risk stratification. Therefore, an urgent need exists for novel effective, and targeted therapies with more precise risk stratification, which necessitates a deeper understanding of the underlying molecular mechanisms that drive the progression of CVD.

From a 6-year population-based cohort of the First Affiliated Hospital of Xiamen University, this study enrolled 32,822 CVD and 25,794 CVD-free participants. We implemented 5 kinds of ML-based data-driven pipeline (LR, RF, SVM, XGBoost, and DNN) to identify predictors from 50 candidate variables covering 22 features from the blood routine and 28 features from blood biochemical tests and assessed multiple ML classifiers to establish risk prediction models on CVD. Our models obtained satisfied discriminative performance with the

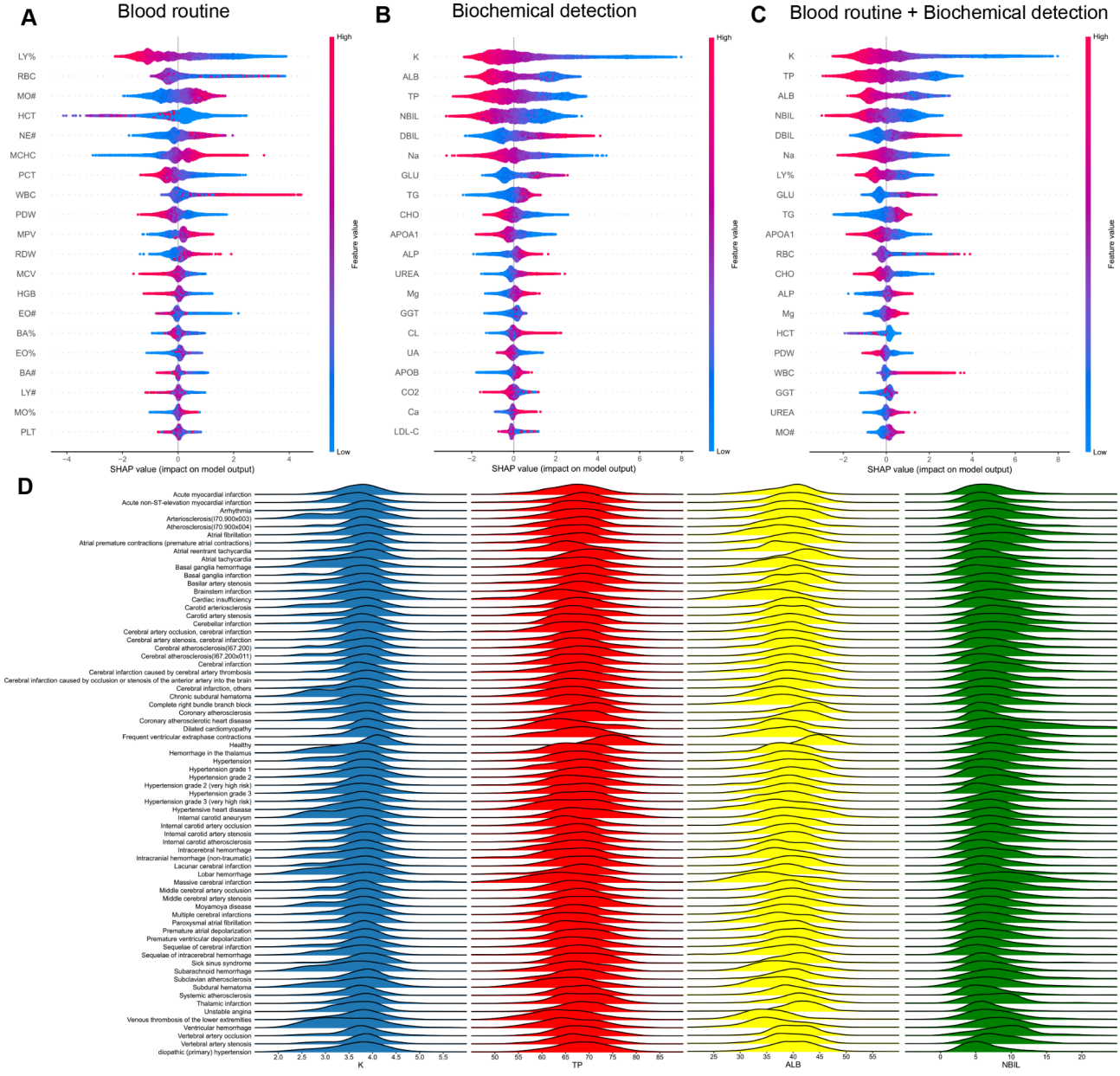

Fig. 4 The top 20 features for the circulation system disease prediction model using different data. (A) Blood routine. (B) Biochemical detection. (C) Blood routine combined with biochemical detection. The red represents a high value, and the blue represents a low value. If the SHAP value is positive, it represents the positive effect of the feature on the model, and vice versa. All features are listed in order of importance from top to bottom. (D) The joyplot of numerical distributions of K, TP, ALB, and NBIL among various circulation system diseases and healthy people

best AUC of 0.9921. Further, we attempted to construct predictive models to distinguish among 69 common CVDs. All these prediction models can discriminate among multiple CVDs, with particularly notable performance in distinguishing DCM (AUC=0.9267) from others.

In this study, we developed predictive models using blood routine and biochemical test data. These models have the potential to be reliable methods for early diagnosis and large-scale screening for CVDs in populations. Recent studies have shown that bilirubin is not just a byproduct of heme degradation but also a crucial endogenous antioxidant [34]. The biochemical processes underlying the relationship between raised DBIL and higher CHD risk remain unclear, although in middle-aged and older adults, DBIL is independently linked to a linear dose-response increased risk for CHD incidence [35]. It has been reported that DBIL is more readily available in an active state because it is soluble in serum and only weakly bound to albumin. In the meantime, it may

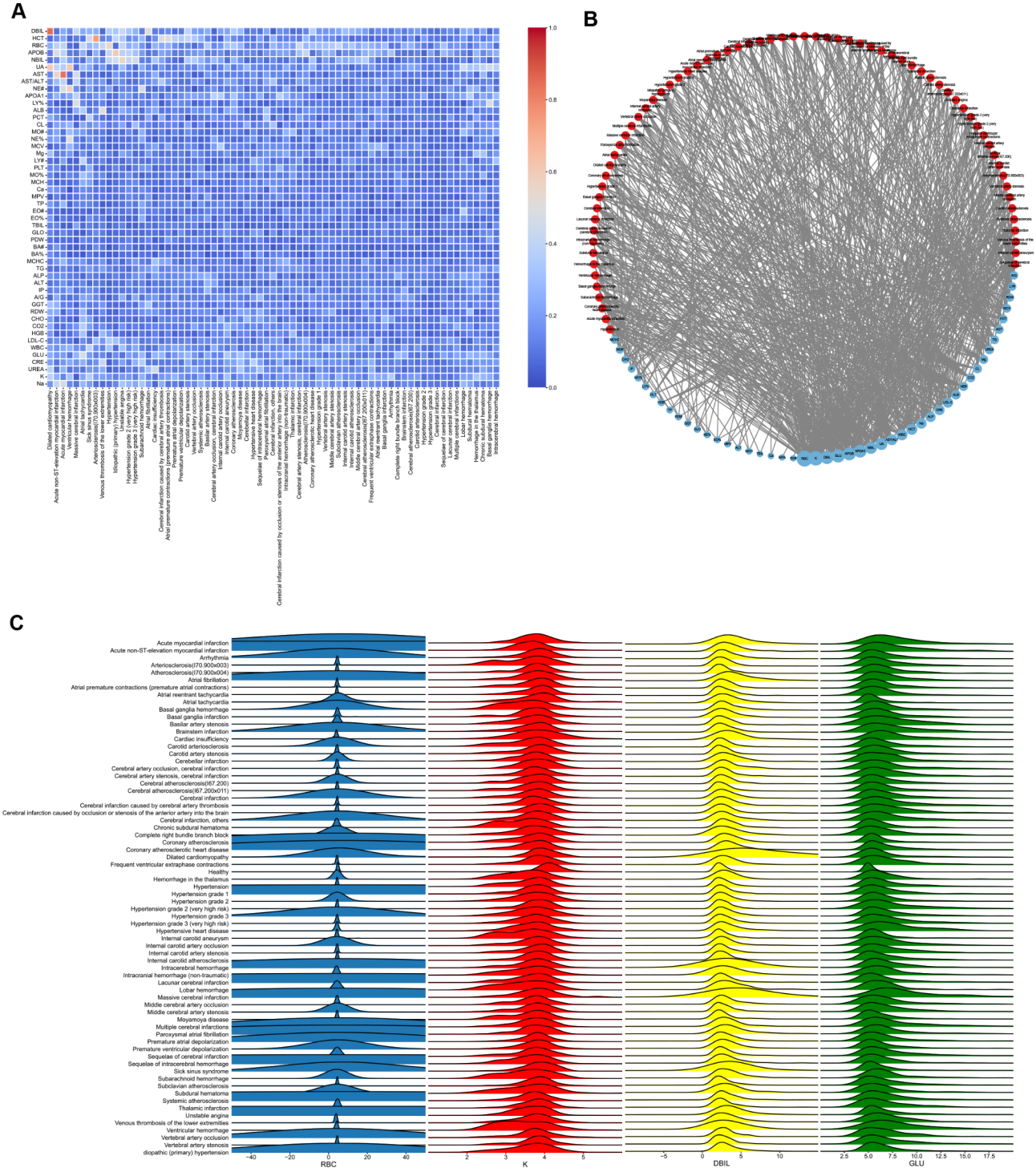

Fig. 5 Analysis of specific indicators for differentiation between different circulation system diseases. (A) The heatmap displays SHAP values of 50 features for each disease differentiation model. The positive SHAP value is added to the absolute value of the negative SHAP value to form the final SHAP value to be displayed. (B) The network shows the intersection top 10 features among different disease differentiation models. The red circles represent various circulation system diseases, and the blue circles represent various features. The larger the blue circle, the more the intersection features. (C) The joyplot of numerical distributions of RBC, K, DBIL, and GLU among various circulation system diseases and healthy people

be hard for water-soluble DBIL to penetrate the vascular intima of the atherosclerotic plaque and function as an antioxidant [36]. This affirmed the significance of circulating bilirubin, including DBIL and NBIL, as predictive features in our models. The diagnostic relevance pertains to the utility of circulating bilirubin concentrations as a novel and reliable marker of cardiovascular disease risk. This biomarker can be readily measured in clinical laboratories and implemented in medical practice.

Diabetes mellitus is associated with a significantly increased risk of cardiovascular diseases. Chronic hyperglycemia is known to induce mitochondrial dysfunction and endoplasmic reticulum stress, promote the accumulation of reactive oxygen species (ROS), and consequently lead to cardiovascular damage [37]. Similarly, hypertriglyceridemia has been implicated in promoting cardiovascular disease through multiple mechanisms, including the upregulation of signaling pathways that mediate inflammation, oxidative stress, thrombosis, endothelial dysfunction, and vascular impairment [38]. Furthermore, serum cholesterol (CHO) levels have been demonstrated to be associated with an increased risk of CVD [39]. Apolipoprotein (APO) A1, the principal apolipoprotein of plasma high-density lipoproteins (HDLs), possesses multiple well-documented cardioprotective functions [40]. Our models confirm alignment with these established metabolic risk factors—GLU, TG, CHO, and APOA1—highlighted as top predictors in this study. This alignment enhances the models' clinical utility, demonstrating their potential to identify individuals at risk of cardiovascular events based on readily accessible parameters.

Furthermore, vascular inflammation and associated chronic pro-inflammatory states are considered key factors in the development of CVD [41]. Previous clinical investigations have demonstrated that peripheral blood lymphocytes are associated with the prognosis of heart failure [42]. Lymphocytopenia in chronic heart failure patients may result from programmed lymphocyte death due to excessive sympathetic activation and increased oxidative stress and pro-inflammatory status [43]. The model constructed in our work demonstrates the role of lymphocyte percentage in the diagnosis of CVD, which is consistent with these previous investigations.

Overall, the predictors derived in our data-driven pipeline have been validated by numerous studies, proving the reliability of our model; however, it is the first time that the ten predictors were combined to establish a CVD risk prediction model. Our models underscore the importance of blood lipid and glucose levels, as well as circulating bilirubin, in the prediction of CVDs.

One notable strength of our study is that all the top 10 predictors for model development can be easily obtained through blood sampling, which provides the general population with the opportunity to perform automated and rapid health screening. It also gives clinicians a tool to help them diagnose heart problems early on. As a result, it will be easier to treat patients effectively and avoid serious repercussions.

While earlier studies have primarily focused on the prediction and diagnosis of specific cardiovascular diseases, such as coronary artery disease [44, 45], atrial fibrillation [46], major adverse cardiovascular events in patients with diabetes [47], and heart failure [48], comprehensive approaches that encompass the entire cardiovascular system remain relatively underexplored. We performed an extensive analysis of 69 prevalent cardiovascular diseases and developed diagnostic models. Additionally, our comprehensive approach in constructing the model included an analysis of distinctions between different CVDs, thereby providing physicians with improved diagnostic differentiation.

While the analysis of clinical data is commonly employed in diagnosis, this practice is less prevalent in CVD diagnostics, where available data is often limited to advanced imaging modalities and invasive hemodynamic assessments [49–51]. The availability of data are essential prerequisite for advancements in the clinical application of machine learning. Our research utilizes hematological data, which is not only more readily accessible but also significantly more cost-effective.

Several caveats should be considered. Given the effect of biological variables such as sex and age on cardiovascular risk [52], it is imperative to integrate datasets from increasing numbers of donors to evaluate the influence of these variables on human cardiovascular disease. Moreover, patients with cardiovascular disease, especially those of advanced age, often have comorbidities such as diabetes mellitus, obesity, and high blood pressure, which will need to be considered in the analysis and interpretation. This study's limitation is its single-center retrospective design, with a sample confined to patients from the First Affiliated Hospital of Xiamen University. Consequently, some results may not be generalizable to other populations. Further validation requires studies involving diverse populations and multiple centers.

Conclusions

In summary, our study developed cost-effective, large-scale screening models based on blood routine and biochemical test data. These models are capable of distinguishing not only cardiovascular disease patients from healthy individuals but also differentiating between various types of cardiovascular diseases (Supplementary Fig. 3). We identified K, TP, ALB, and NBIL as universal indicators for distinguishing cardiovascular disease patients from healthy individuals, while RBC, K, DBIL, and GLU were found to be universal indicators for distinguishing between different types of cardiovascular diseases. Additionally, we

identified unique hematological and metabolic characteristics for each type of cardiovascular disease, which could provide clinicians with specialized insights for early disease prevention and diagnosis.

Abbreviations

CVDs	Cardiovascular diseases
LR	Logistic regression
RF	Random forest
SVM	Support vector machine
XGBoost	Extreme Gradient Boosting
GBDT	Gradient boosting decision tree
DNN	Deep neural network
DCM	Dilated cardiomyopathy
RBC	Red blood cell count
LY%	Lymphocyte percentage
HCT	Hematocrit

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12933-024-02439-0 .

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements
The authors would like to thank all the patients who participated in this trial as well as their families. We thank other team members for assisting with manuscript preparation.

Author contributions
ZW, WN, YG, and SL wrote the manuscript together. YY, ZW, WN, QC, YG, and LH completed the data collection, investigation, and analysis. ZW and LH contributed to the methodology. WN, YY, and GH designed the study and contributed to conceptualization, funding acquisition, reviewing, and editing. All authors read and approved the final version of the manuscript.

Funding
This work was supported by the National Key R & D Program of China [2021ZD0201300 and 2022YFC2704300], the National Natural Science Foundation of China [82371200 and 82171474], and the Natural Science Foundation of Fujian Provincial [2020J05310]. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Availability of data and materials
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
The protocol has been approved by the Ethics Committee of the First Affiliated Hospital of Xiamen University (XMYY-2023KYSB088).

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Institute for Clinical Medical Research, School of Medicine, The First Affiliated Hospital of Xiamen University, Xiamen University, Xiamen 361003, Fujian, China
[2]Department of Laboratory Medicine, Xiamen Key Laboratory of Genetic Testing, School of Medicine, the First Affiliated Hospital of Xiamen University, Xiamen University, Xiamen 361003, Fujian, China
[3]Department of Otolaryngology, School of Medicine, Xiamen University, Xiamen 361003, Fujian, China

Received: 22 July 2024 / Accepted: 11 September 2024
Published online: 28 September 2024

References

1. Cheng X, Manandhar I, Aryal S, et al. Application of artificial intelligence in cardiovascular medicine. Compr Physiol. 2021;11(4):2455–66.
2. Roth GA, Mensah GA, Johnson CO, et al. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. J Am Coll Cardiol. 2020;76(25):2982–3021.
3. Lindstrom M, DeCleene N, Dorsey H, et al. Global burden of cardiovascular diseases and risks collaboration, 1990–2021. J Am Coll Cardiol. 2022;80(25):2372–425.
4. The W. Report on cardiovascular health and diseases in China 2022: an updated summary. Biomed Environ Sci. 2023;36(8):669–701.
5. Leening MJ, Siregar S, Vaartjes I, et al. Heart disease in the Netherlands: a quantitative update. Neth Heart J. 2014;22(1):3–10.
6. Bandesh K, Jha P, Giri AK, et al. Normative range of blood biochemical parameters in urban Indian school-going adolescents. PLoS ONE. 2019;14(3):e0213255.
7. Wolthuis A. Impact of disease on interferences in blood bioanalysis. Bioanalysis. 2011;3(19):2223–31.
8. Menotti A, Lanti M, Zanchetti A, et al. The role of HDL cholesterol in metabolic syndrome predicting cardiovascular events. The Gubbio population study. Nutr Metab Cardiovasc Dis. 2011;21(5):315–22.
9. Rabbani N, Kim G, Suarez CJ, et al. Applications of machine learning in routine laboratory medicine: current state and future directions. Clin Biochem. 2022;103:1–7.
10. Ronzio L, Cabitza F, Barbaro A et al. Has the flood entered the basement? A systematic literature review about machine learning in laboratory medicine. Diagnostics (Basel) 2021;11(2).
11. Mathur P, Srivastava S, Xu X, et al. Artificial intelligence, machine learning, and cardiovascular disease. Clin Med Insights Cardiol. 2020;14:1522409556.
12. Attia ZI, Harmon DM, Behr ER, et al. Application of artificial intelligence to the electrocardiogram. Eur Heart J. 2021;42(46):4717–30.
13. Fernandez-Luque L, Imran M. Humanitarian health computing using artificial intelligence and social media: a narrative literature review. Int J Med Inform. 2018;114:136–42.
14. Panjiyar BK, Davydov G, Nashat H, et al. A systematic review: Do the use of machine learning, deep learning, and artificial intelligence improve patient outcomes in acute myocardial ischemia compared to clinician-only approaches? Cureus. 2023;15(8): e43003.
15. Chen L, Han Z, Wang J, et al. The emerging roles of machine learning in cardiovascular diseases: a narrative review. Ann Transl Med. 2022;10(10):611.
16. Muse ED, Topol EJ. Guiding ultrasound image capture with artificial intelligence. Lancet. 2020;396(10253):749.
17. Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. Nat Med. 2019;25(1):70–4.
18. Shu S, Ren J, Song J. Clinical application of machine learning-based artificial intelligence in the diagnosis, prediction, and classification of cardiovascular diseases. Circ J. 2021;85(9):1416–25.
19. Roh J, Houstis N, Rosenzweig A. Why don't we have proven treatments for HFpEF? Circ Res. 2017;120(8):1243–5.
20. Shah SJ, Katz DH, Selvaraj S, et al. Phenomapping for novel classification of heart failure with preserved ejection fraction. Circulation. 2015;131(3):269–79.
21. Wang H, Liang P, Zheng L, et al. eHSCPr discriminating the cell identity involved in endothelial to hematopoietic transition. Bioinformatics. 2021;37(15):2157–64.
22. Tang H, Zhao YW, Zou P, et al. HBPred: a tool to identify growth hormone-binding proteins. Int J Biol Sci. 2018;14(8):957–64.
23. Kumar A, Loharch S, Kumar S, et al. Corrigendum to "Exploiting cheminformatic and machine learning to navigate the available chemical space of potential small molecule inhibitors of SARS-CoV-2" [Computational and

Structural Biotechnology Journal 19 (2021) 424–438]. Comput Struct Biotechnol J. 2023;21:4408.
24. Zhang D, Xu ZC, Su W, et al. iCarPS: a computational tool for identifying protein carbonylation sites by novel encoded features. Bioinformatics. 2021;37(2):171–7.
25. Eichler J. Protein glycosylation. Curr Biol. 2019;29(7):R229–31.
26. Wu H, Wu Y, Jiang Y, et al. scHiCStackL: a stacking ensemble learning-based method for single-cell Hi-C classification using cell embedding. Brief Bioinform. 2022;23(1).
27. Meng L, Chan WS, Huang L, et al. Mini-review: recent advances in post-translational modification site prediction based on deep learning. Comput Struct Biotechnol J. 2022;20:3522–32.
28. Liu M, Zhou J, Xi Q, et al. A computational framework of routine test data for the cost-effective chronic disease prediction. Brief Bioinform. 2023;24(2).
29. Ning W, Lei S, Yang J, et al. Open resource of clinical data from patients with pneumonia for the prediction of COVID-19 outcomes via deep learning. Nat Biomed Eng. 2020;4(12):1197–207.
30. Altan G. DeepOCT: An explainable deep learning architecture to analyze macular edema on OCT images[J]. Eng Sci Technol Int J-JESTECH, 2022;34.
31. Altan G. Breast cancer diagnosis using deep belief networks on ROI images. Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi. 2022;28(2):286–91.
32. Wang K, Tian J, Zheng C, et al. Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP. Comput Biol Med. 2021;137: 104813.
33. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504.
34. Seppen J, Bosma P. Bilirubin, the gold within. Circulation. 2012;126(22):2547–9.
35. Lai X, Fang Q, Yang L, et al. Direct, indirect and total bilirubin and risk of incident coronary heart disease in the Dongfeng-Tongji cohort. Ann Med. 2018;50(1):16–25.
36. Franchini M, Targher G, Lippi G. Serum bilirubin levels and cardiovascular disease risk: a Janus Bifrons? Adv Clin Chem. 2010;50:47–63.
37. Fiorentino TV, Prioletta A, Zuo P, et al. Hyperglycemia-induced oxidative stress and its role in diabetes mellitus related cardiovascular diseases. Curr Pharm Des. 2013;19(32):5695–703.
38. Reiner. Hypertriglyceridaemia and risk of coronary artery disease. Nat Rev Cardiol. 2017;14(7):401–11.
39. Stamler J, Daviglus ML, Garside DB, et al. Relationship of baseline serum cholesterol levels in 3 large cohorts of younger men to long-term coronary, cardiovascular, and all-cause mortality and to longevity. JAMA. 2000;284(3):311–8.
40. Nacarelli GS, Fasolino T, Davis S. Dietary, macronutrient, micronutrient, and nutrigenetic factors impacting cardiovascular risk markers apolipoprotein B and apolipoprotein A1: a narrative review. Nutr Rev. 2024;82(7):949–62.
41. Silveira RJ, Barbalho SM, Reverete DAR, et al. Metabolic syndrome and cardiovascular diseases: going beyond traditional risk factors. Diabetes Metab Res Rev. 2022;38(3): e3502.
42. Ommen SR, Hodge DO, Rodeheffer RJ, et al. Predictive power of the relative lymphocyte concentration in patients with advanced heart failure. Circulation. 1998;97(1):19–22.
43. Weng TP, Fu TC, Wang CH, et al. Activation of lymphocyte autophagy/apoptosis reflects haemodynamic inefficiency and functional aerobic impairment in patients with heart failure. Clin Sci (Lond). 2014;127(10):589–602.
44. Shapiro D, Lee K, Asmussen J, et al. Evolutionary action-machine learning model identifies candidate genes associated with early-onset coronary artery disease. J Am Heart Assoc. 2023;12(17): e029103.
45. Trigka M, Dritsas E. Long-term coronary artery disease risk prediction with machine learning models. Sensors (Basel), 2023;23(3).
46. Lu Y, Chen Q, Zhang H, et al. Machine learning models of postoperative atrial fibrillation prediction after cardiac surgery. J Cardiothorac Vasc Anesth. 2023;37(3):360–6.
47. Abegaz TM, Baljoon A, Kilanko O, et al. Machine learning algorithms to predict major adverse cardiovascular events in patients with diabetes. Comput Biol Med. 2023;164: 107289.
48. Kyodo A, Kanaoka K, Keshi A, et al. Heart failure with preserved ejection fraction phenogroup classification using machine learning. ESC Heart Fail. 2023;10(3):2019–30.
49. Wang YJ, Yang K, Wen Y, et al. Screening and diagnosis of cardiovascular disease using artificial intelligence-enabled cardiac magnetic resonance imaging. Nat Med. 2024;30(5):1471–80.
50. Sun Z. Multislice computed tomography angiography in the diagnosis of cardiovascular disease: 3D visualizations. Front Med. 2011;5(3):254–70.
51. Givertz MM, Fang JC, Sorajja P, et al. Executive summary of the SCAI/HFSA clinical expert consensus document on the use of invasive hemodynamics for the diagnosis and management of cardiovascular disease. J Card Fail. 2017;23(6):487–91.
52. You J, Guo Y, Kang JJ, et al. Development of machine learning-based models to predict 10-year risk of cardiovascular disease: a prospective cohort study. Stroke Vasc Neurol. 2023;8(6):475–85.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.