Risk Management and Healthcare Policy

Dovepress

open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Application of the Unbalanced Ensemble Algorithm for Prognostic Prediction Outcomes of All-Cause Mortality in Coronary Heart Disease Patients Comorbid with Hypertension

Jiaxin Zan[1,2], Xiaojing Dong[1,2], Hong Yang[1,2], Jingjing Yan[1,2], Zixuan He[3], Jing Tian[3], Yanbo Zhang[1,2,4]

[1]Department of Health Statistics, School of Public Health, Shanxi Medical University, Taiyuan, People's Republic of China; [2]Shanxi Provincial Key Laboratory of Major Diseases Risk Assessment, Taiyuan, People's Republic of China; [3]Department of Cardiology, The First Hospital of Shanxi Medical University, Taiyuan, People's Republic of China; [4]School of Health Services and Management, Shanxi University of Chinese Medicine, Taiyuan, People's Republic of China

Correspondence: Jing Tian; Yanbo Zhang, School of Public health, Shanxi Medical University, 56 Xinjian Road, Taiyuan, Shanxi Province, People's Republic of China, Tel/Fax +86 15535406059, Email 1105551933@qq.com; sxmuzyb@126.com

Purpose: This study sought to develop an unbalanced-ensemble model that could accurately predict death outcomes of patients with comorbid coronary heart disease (CHD) and hypertension and evaluate the factors contributing to death.

Patients and Methods: Medical records of 1058 patients with coronary heart disease combined with hypertension and excluding those acute coronary syndrome were collected. Patients were followed-up at the first, third, sixth, and twelfth months after discharge to record death events. Follow-up ended two years after discharge. Patients were divided into survival and nonsurvival groups. According to medical records, gender, smoking, drinking, COPD, cerebral stroke, diabetes, hyperhomocysteinemia, heart failure and renal insufficiency of the two groups were sorted and compared and other influencing factors of the two groups, feature selection was carried out to construct models. Owing to data unbalance, we developed four unbalanced-ensemble prediction models based on Balanced Random Forest (BRF), EasyEnsemble, RUSBoost, SMOTEBoost and the two base classification algorithms based on AdaBoost and Logistic. Each model was optimised using hyperparameters based on GridSearchCV and evaluated using area under the curve (AUC), sensitivity, recall, Brier score, and geometric mean (G-mean). Additionally, to understand the influence of variables on model performance, we constructed a SHapley Additive explanation (SHAP) model based on the optimal model.

Results: There were significant differences in age, heart rate, COPD, cerebral stroke, heart failure and renal insufficiency in the nonsurvival group compared with the survival group. Among all models, BRF yielded the highest AUC (0.810; 95% CI, 0.778–0.839), sensitivity (0.990; 95% CI, 0.981–1.000), recall (0.990; 95% CI, 0.981–1.000), and G-mean (0.806; 95% CI, 0.778–0.827), and the lowest Brier score (0.181; 95% CI, 0.178–0.185). Therefore, we identified BRF as the optimal model. Furthermore, red blood cell count (RBC), body mass index (BMI), and lactate dehydrogenase were found to be important mortality-associated risk factors.

Conclusion: BRF combined with advanced machine learning methods and SHAP is highly effective and accurately predicts mortality in patients with CHD comorbid with hypertension. This model has the potential to assist clinicians in modifying treatment strategies to improve patient outcomes.

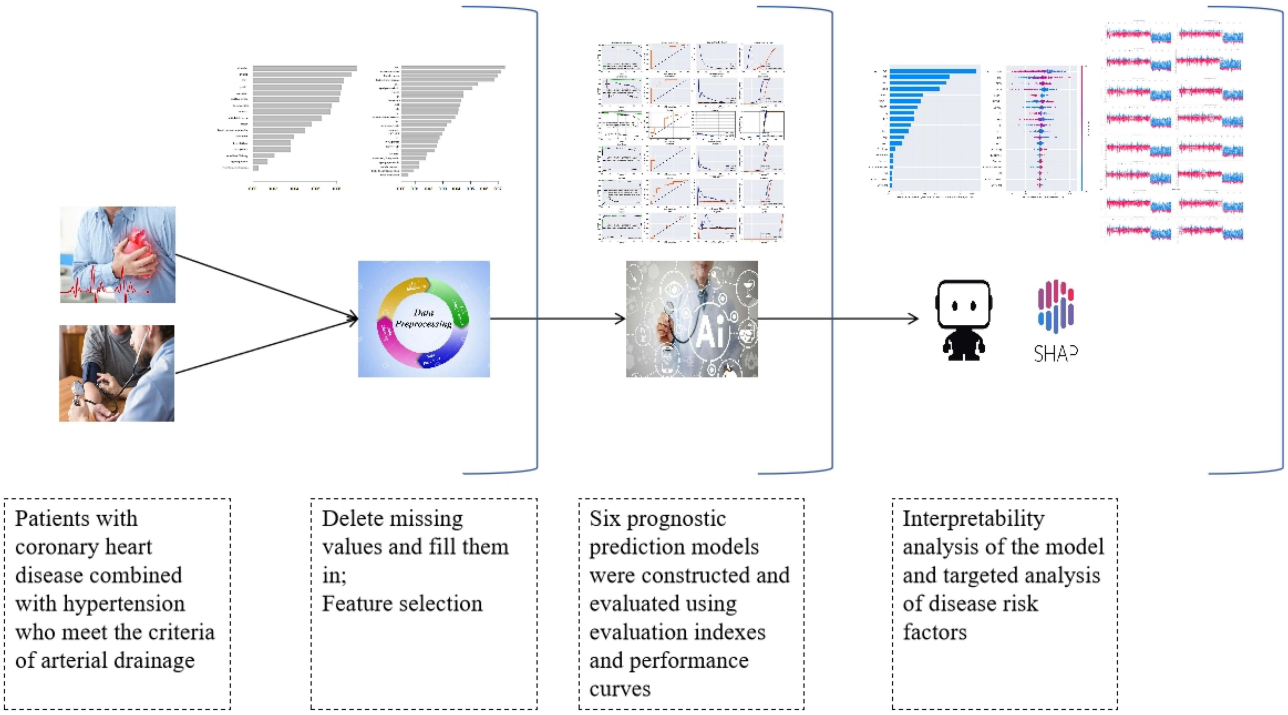
Keywords: coronary heart disease comorbid with hypertension, ensemble learning, balanced random forest, SHAP, Prognosis

Introduction

The incidence of coronary heart disease (CHD) is increasing due to the ageing population and gradual adoption of unhealthy lifestyles.[1] CHD is a significant threat to public health,[2] with approximately 11.39 million individuals currently diagnosed.[3] Hypertension and CHD are closely related[4,5] in that hypertension is the leading preventable risk factor for CHD.[6] According to the consensus of Chinese experts on blood pressure management in patients with hypertension

Received: 5 April 2024
Accepted: 24 July 2024
Published: 6 August 2024

Graphical Abstract



combined with coronary heart disease, patients with hypertension are often exhibit by left ventricular hypertrophy, leading to increased myocardial oxygen consumption and coronary microcirculation disorders. When combined with coronary heart disease, coronary blood supply is reduced, making myocardial ischemia more likely to occur, which poses greater health risks and a more serious disease burden. Epidemiological data shows that the prevalence rate of CHD patients with hypertension in Chinese population is as high as 60%, and about 71.8% of hospitalized CHD patients in China have hypertension. Therefore, predicting the mortality of CHD patients with hypertension is of greater practical and clinical significance, aligning more closely with China's national conditions.[7,8] Accurately predicting the prognosis of patients with CHD and comorbid hypertension has significant implications. It can guide tertiary prevention, inform treatment strategies for healthcare providers, allocate public health resources, and inform government decision-making.

In recent decades, machine learning has emerged as a valuable tool for predicting disease prognosis.[9] Different machine learning algorithms have demonstrated high accuracy and speed in analysing various types of medical data.[10] However, class unbalance, an uneven distribution of positive and negative results, can lead traditional classification methods to produce biased predictions favouring the majority class. This problem poses significant challenges for machine learning and data mining.[11,12] Considerable progress has been made in addressing the challenges of unbalanced datasets. The progress includes the development of various techniques, such as resampling, random undersampling, random oversampling, CUSBoost, and SMOTE, as well as ensemble-based approaches, including Bagging and Boosting.[13–16] Owing to the unique characteristics of various biomedical datasets, different algorithms may be required, and ensemble learning techniques can significantly enhance the performance of most unbalanced methods.[17] In recent years, unbalanced-ensemble algorithms have gained popularity in medical research due to their effectiveness in handling unbalanced datasets. Among these algorithms, SMOTEBoost, RUSBoost, and other machine learning methods have emerged as the most preferred. These unbalanced-ensemble algorithms combine ensemble learning with various weak classifiers and sampling techniques to improve prediction performance and enhance model robustness.[18]

The aim of this study was to construct an optimal ensemble model to address the issue of unbalanced datasets in predicting death outcomes in patients with CHD comorbid with hypertension. The model was developed to accurately predict patient outcomes and evaluate the risk factors associated with mortality. To achieve this, advanced machine learning and SHAP models were used to explain the variables influencing model performance. This study could help clinicians adjust treatment strategies in real time.

Patients and Methods

Study Population

This was a prospective cohort study to predict all-cause mortality outcomes in CHD patients comorbid with hypertension. Patients admitted between November 2017 and June 2020 in the cardiology department of a medical university hospital in Shanxi Province, China, were enrolled in this study in accordance with the inclusion and exclusion criteria.

The inclusion criteria were 1) aged ≥18 years; 2) diagnosed with CHD according to the guidelines for the diagnosis and treatment of coronary heart disease;[19] and 3) diagnosed with hypertension according to the Hypertension Prevention and Treatment Guidelines.[20] Patients with a concurrent mental illness, concurrently suffering from other diseases with an expected survival time of less than 1 year, or refusing to participate in the program were excluded.

Data Collection

This study retrieved patient information from electronic medical records stored in the medical cloud. The information included patient demographics, medical history, physical examination results and vitals, currently prescribed medical therapy, and laboratory parameters. After discharge from the hospital, all patients were followed up by a trained specialist over the phone at the first, third, sixth, and twelfth months. Based on the inclusion and exclusion criteria, 3263 patients with CHD comorbid with hypertension were included.

The variables from the medical records were collected including the following:

General information: demographic characteristics, smoking, drinking, allergy history, family history, surgical history;

Comorbidities: atrial fibrillation, COPD, valvular disease, hyperlipidemia, diabetes, embolism, ventricular aneurysm, renal insufficiency, obesity, hyperkalemia/hypokalemia;

Physical signs: body temperature, blood pressure, height, weight, respiratory rate, heart rate;

Laboratory tests: blood routine, liver function, lipids, kidney function, electrolytes, thyroid function, coagulation function, blood gas analysis, blood glucose, urine routine;

Drug treatment: anticoagulants, such as aspirin, clopidogrel, warfarin, ordinary heparin, enoxaparin, etc.; Statins, such as simvastatin, atorvastatin, rosuvastatin, pravastatin, etc. Calcium antagonists, such as amlodipine, levamamlodipine, benidipine; Beta blockers, such as metoprolol, metoprolol sustained-release tablets, bisoprolol, etc.; Diuretics, such as tolasemide, Furosemide, tolvaptan, hydrochlorothiazide, etc.; Aldosterone antagonists, such as spironolactone, eplerenone, etc.; Cardiotonic, such as digoxin, sildilan, dobutamine, etc.

The cohort used in this study was the Prospective Cohort and Prognostic Study of Coronary Atherosclerotic Heart Disease registered with the Chinese Clinical Trial Registry, registration number ChiCTR2100043434.

Study Outcomes

The primary endpoint of the study was all-cause mortality during the follow-up period. All-cause mortality was defined as death from any cause.

Data Pre-Processing and Feature Selection

Our dataset initially included hundreds of clinical variables. Based on relevant studies on missing value processing,[21] variables with a missing ratio greater than 30% and samples with a missing ratio greater than 20% were excluded from our dataset. Variables with a less than 30% missing ratio were retained and imputed using the MissForest algorithm.[22] Because there are missing values in the data, the missing values are analyzed. Based on relevant literature, or possible influencing factors of patients with coronary heart disease and hypertension, the form is divided into original data

missing value analysis and data missing value analysis with a deletion ratio ≥ 20%. The results are shown in Supplementary Tables 1 and 2.

In this study, feature screening was performed using XGBoost. Studies have indicated that utilising XGBoost for feature selection improves efficiency compared to traditional methods and leads to a superior predictive performance of the model.[23] Given the potential impact of data dimensionality on the effectiveness of feature selection, we ranked categorical and continuous variables based on their importance in predicting disease prognosis. We retained the top 20 variables that significantly impacted disease prognosis.[24]

Statistical Analysis

The data were processed by SPSS24.0, and measurement data conforming to normal distribution were represented by $\bar{x} \pm s$. The comparison between groups was performed using a *t*-test of independent samples, and the intra-group comparison using a *t*-test of paired samples. Measurement data that did not conform to a normal distribution were represented by Md (P_{25},P_{75}). They were made by Wilcoxon's rank sum test, which was made by comparing two independent samples, and the enumeration data was made by n (%) and by chi-square test or exact probability test of Fisher. $P<0.05$ was statistically significant. (See Supplementary Table 3 for normality tests).

Research Methodology

Balanced Random Forest

BRF was employed in this study to generate a system tree from the undersampling data. The BRF algorithm included the following steps: adding random undersampling to each round of self-service sampling, randomly selecting several samples from a few classes, and then randomly selecting the same number of samples from many classes to form a balanced dataset. The algorithm used the balanced dataset as input for each iteration. The largest decision tree was derived from the data without pruning it. The decision tree was derived from the CART algorithm and modified accordingly. Instead of searching for all attributes to obtain the best-split variables on each node, one attribute was randomly selected as a split variable. The steps were repeated to generate multiple decision trees to form a random forest, which eventually led to classification by simple voting.[25]

Easyensemble

The EasyEnsemble algorithm is a data augmentation technique that is particularly effective in dealing with highly unbalanced data. The algorithm randomly divides the majority class samples into N subsets, each with the same number of samples as the minority class. The resulting subsets of the majority and minority samples were combined to train an AdaBoost-based classification model. Finally, the outputs of the individual base classifiers were combined to obtain the final classification model.[16]

RUSBoost

The RUSBoost algorithm is a combined algorithm that integrates undersampling with boosting to improve classification performance. Boosting is a machine learning technique that converts weak classifiers into strong ones. One of the most commonly used boosting algorithms is the AdaBoost algorithm proposed by Freund et al. The RUSBoost algorithm used in this study was based on the AdaBoost algorithm. The RUSBoost algorithm initially trains a base classifier using the initial training set and then adjusts the distribution weight of each training sample based on the training error of the current base classifier. This method strengthens the focus on misclassified samples in subsequent training iterations by increasing their weight. The adjusted sample was then used to train the next base classifier, and this process was iterated N times until N base classifiers were generated. Finally, the test samples were predicted by weighted voting based on the results of the N base classifiers.[26]

SMOTEBoost

SMOTEBoost is a boosting algorithm that integrates the Synthetic Minority Over-sampling Technique (SMOTE) with the traditional boosting algorithm. Boosting combines several weak classifiers to form a strong classifier. However, because boosting assigns the same weight to all examples and samples of the wrong classification, it mainly focuses on

the data pool composed of the majority classes. To reduce the inherent bias caused by class unbalance and increase the sampling weight of minority classes, the SMOTE algorithm is introduced into each class. This introduces synthetic samples into the minority class, thereby increasing the number of learning samples for the minority classes.[15]

AdaBoost

AdaBoost, short for Adaptive Boosting, is an adaptive algorithm that increases the weight of previously misclassified base classifiers while reducing the weight of correctly classified ones. In each iteration, a new weak classifier is added until a predetermined maximum number of iterations or a sufficiently small error rate is reached.[27]

Logistic

Logistic regression is a well-established statistical classification method. It is a linear model that creates regression equations to determine classification boundary lines based on available data. The term "regression" in this context refers to finding the best-fit parameters for the model.

Model Development

In this study, the training and test sets were randomly partitioned in an 8:2 ratio. On the training set, GridSearchCV is used to build a 5-fold cross-validation. One part is taken as the test set without repetition each time, and the other four parts are used as the training set to train the model. The AUC is used as the scoring standard to conduct a grid search to determine the best configuration of the model. To achieve a more reliable performance estimate, mitigate reporting of biased results, and prevent over-fitting, we repeated the persistence method 100 times using different random seeds and calculated the average performance across these 100 repetitions. This study built the model in the imbalanced ensemble and sklearn ensemble libraries in Python 3.7. The model parameters are provided in Supplementary Table 4.

Model Evaluation Metrics

In this study, performance metrics such as AUC, sensitivity, recall, Brier score[28] and G-mean[29] were employed to assess the performance of the machine learning model. Higher values of AUC, sensitivity, recall, and G-mean indicate better performance of the machine learning model. In contrast, a lower Brier score indicates better model stability and can effectively prevent overfitting.

Model performance was evaluated using visualisations such as the receiver operating characteristic (ROC) curve, lift curve,[30] Kolmogorov–Smirnov(KS) curve,[31] and precision-recall curve. A higher area under the ROC curve indicates better model accuracy, whereas a larger lift exponent signifies better model performance. The Kolmogorov–Smirnov curve measures the model's ability to distinguish between classes, with a greater distance between the curves indicating better performance. The precision-recall curve indicates the trade-off between precision and recall, with a curve closer to the upper right indicating better performance (see Figure 1 for model development and evaluation).

Model Interpretability

The most effective machine learning model among the six was selected to evaluate the importance of each variable. Additionally, we utilised SHAP[32] as an updated approach to interpret the machine learning models. This approach helps illustrate the impact of individual features by explaining the overall impact of tree sets in the form of specific feature contributions.

The best way to explain a simple machine learning model is through the model itself, which is easy to understand. However, for complex models, such as ensemble models, the underlying principles are complex and difficult to explain using the original model. Therefore, the SHAP model has emerged as a way to simplify the explanation of complex models.

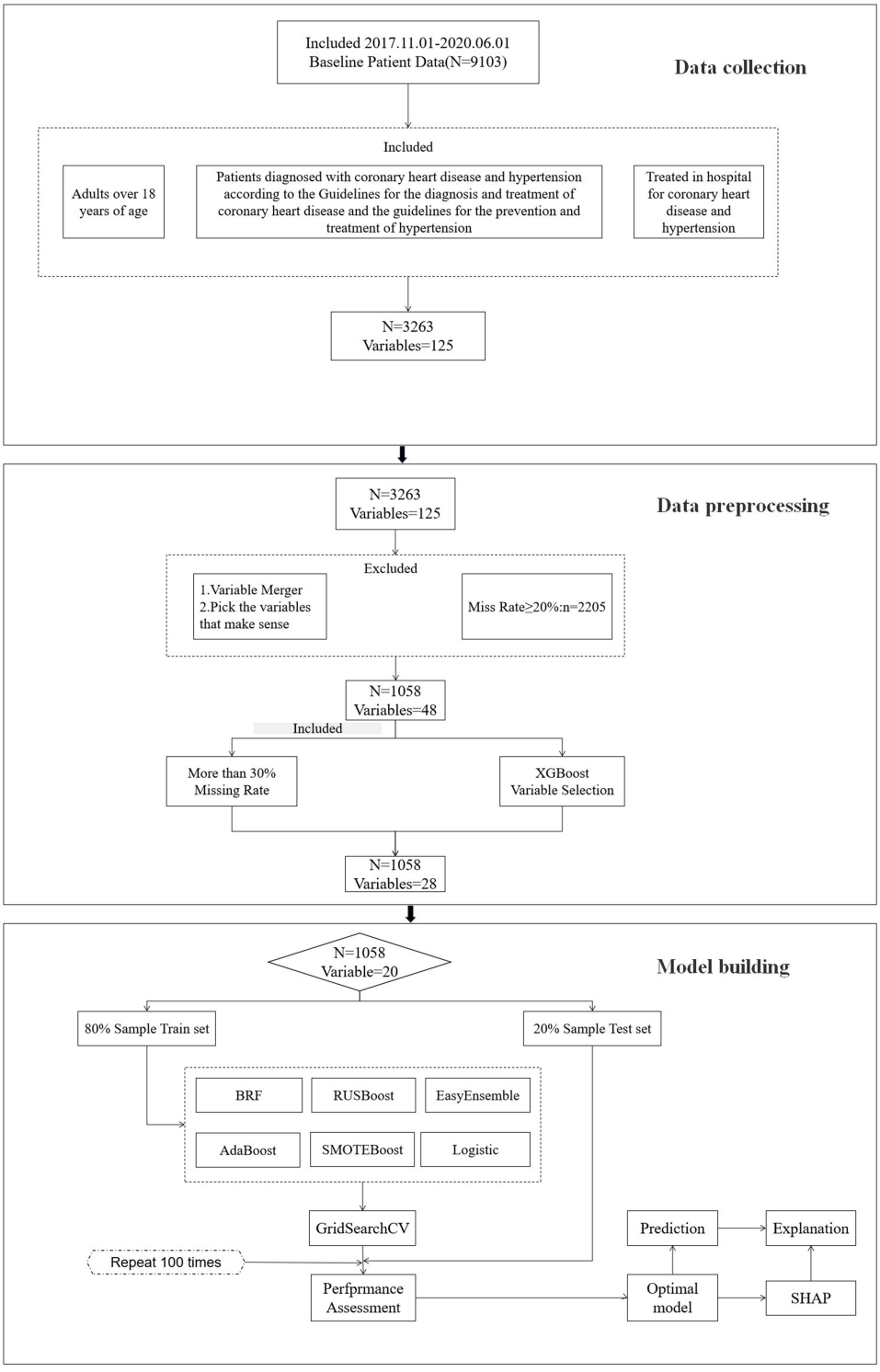


Figure 1 Analytical processes for model development and evaluation.

Results

Basic Information

A total of 1058 hospitalised patients were enrolled, with 751 (71.0%) males and 307 (29.0%) females. Among them, 67 (6.3%) patients experienced mortality, 991 (93.7%) patients were still alive at the end of follow-up. The ratio of the two groups was 1:14.79, indicating that the dataset was unbalanced. The baseline data are shown in Table 1.

Table 1 Single Factor Analysis of Baseline Data

Variable	Survivors (n=991)	Nonsurvivors (n=67)	*p*
Drinking (n%)			0.665
Yes	306.0 (30.9)	19.0 (28.4)	
No	685.0 (69.1)	48.0 (71.6)	
PCI (n%)			0.886
Yes	317.0 (31.9)	22.0 (32.8)	
No	674.0 (68.1)	45.0 (67.2)	
Gender (n%)			0.877
Male	704.0 (71.0)	47.0 (70.1)	
Female	287.0 (29.0)	20.0 (29.9)	
Smoking (n%)			0.664
Yes	535.0 (54.0)	38.0 (56.7)	
No	456.0 (46.0)	29.0 (43.3)	
Age (years)	69.0 (62.0,76.0)	79.0 (70.0,84.0)	0.001
Heart rate (heart rate)	78.0 (70.0,84.0)	82.0 (77.0,92.0)	0.001
Comorbidity			
COPD (n%)			<0.001
Yes	617.0 (62.3)	22.0 (2.8)	
No	374.0 (37.7)	45.0 (67.2)	
Cerebral stroke (n%)			<0.001
Yes	116.0 (11.7)	20.0 (29.9)	
No	875.0 (88.3)	47.0 (70.1)	
Diabetes (n%)			0.452
Yes	759.0 (76.6)	54.0 (80.6)	
No	232.0 (23.4)	13.0 (19.4)	
Hyperhomocysteinemia (n%)			0.670
Yes	40.0 (4.0)	2.0 (3.0)	
No	951.0 (96.0)	65.0 (97.0)	
Hyperlipidaemia (n%)			0.105
Yes	75.0 (7.6)	1.0 (1.5)	
No	916.0 (92.4)	66.0 (98.5)	
Arrhythmia (n%)			0.768
Yes	870.0 (87.8)	58.0 (86.6)	
No	121.0 (12.2)	9.0 (13.4)	
Heart failure (n%)			0.015
Yes	20.0 (2.0)	5.0 (7.5)	
No	971.0 (98.0)	62.0 (92.5)	
Renal insufficiency (n%)			0.001
Yes	983.0 (99.2)	64.0 (94.0)	
No	8.0 (0.8)	4.0 (6.0)	
Laboratory index			
Red blood cell count (10^{12}/L)	4.4 (3.9,4.8)	3.8 (3.2,4.5)	0.001
Blood glucose (mmol/L)	5.3 (4.8,6.6)	5.4 (4.8,7.1)	0.717
Lactate dehydrogenase (U/L)	192.0 (166.0,230.0)	242.0 (174.0,303.0)	0.001
BMI (kg/m^2)	23.6 (22.1,24.9)	22.2 (21.0,23.1)	0.001
Apolipoprotein a1 (g/L)	1.2 (1.0,1.3)	1.1 (0.8,1.2)	0.001
Creatine Kinase Isoenzyme (U/L)	9.2 (3.6,13.2)	7.7 (4.0,12.6)	0.509
Serum potassium (mmol/L)	4.1 (3.8,4.3)	4.2 (4.0,4.4)	0.007
Glutamic oxalacetic transaminase (U/L)	20.9 (16.8,28.3)	24.0 (17.3,32.8)	0.174
Blood magnesium (mmol/L)	0.9 (0.8,1.0)	0.9 (0.8,0.9)	0.403
Blood calcium (mmol/L)	2.3 (2.2,2.3)	2.2 (2.1,2.3)	0.001

(*Continued*)

Table 1 (Continued).

Variable	Survivors (n=991)	Nonsurvivors (n=67)	*p*
Serum sodium (mmol/L)	140.0 (138.0,142.0)	138.0 (134.0,141.0)	0.001
Systolic pressure (mmHg)	123.0 (116.0,131.0)	124.0 (117.0,131.0)	0.766
Diastolic pressure (mmHg)	73.0 (68.0,78.0)	71.0 (67.0,75.0)	0.065
AST/ALT	1.1 (0.9,1.5)	1.5 (1.1,1.9)	0.001
Total cholesterol (mmol/L)	3.6 (3.0,4.5)	3.3 (2.6,4.2)	0.011
Triglyceride (mmol/L)	1.2 (0.9,1.7)	1.0 (0.7,1.4)	0.001
High density lipoprotein cholesterol (mmol/L)	1.8 (1.1,29.7)	19.4 (1.2,31.3)	0.223
Low density lipoprotein cholesterol (mmol/L)	2.1 (1.6,2.6)	1.8 (1.4,2.4)	0.008
Apolipoprotein B (g/L)	0.7 (0.5,0.8)	0.6 (0.5,0.8)	0.060
Lipoprotein (a) (g/L)	24.2 (12.4,37.6)	32.6 (14.4,42.1)	0.042
Uric Acid (mmol/L)	307.0 (255.0,381.0)	305.0 (230.0,357.0)	0.329
Blood chlorine (mmol/L)	106.0 (103.0,108.0)	104.0 (99.0,107.0)	0.002
Serum phosphorus (mmol/L)	1.1 (1.0,1.2)	1.0 (0.9,1.2)	0.105
Cystatin C (mg/L)	0.8 (0.6,0.9)	0.9 (0.7,1.3)	0.001
Myohaemoglobin (g/L)	38.0 (29.1,56.2)	51.4 (36.2,82.8)	0.001
Drug usage			
Antiplatelet drug (n%)			0.001
Yes	462.0 (46.6)	17.0 (25.4)	
No	529.0 (53.4)	50.0 (74.6)	
Heparins (n%)			0.469
Yes	987.0 (99.6)	67.0 (100.0)	
No	4.0 (0.4)	0.0 (0.0)	
Statins (n%)			0.015
Yes	621.0 (62.7)	32.0 (47.8)	
No	370.0 (37.3)	35.0 (52.2)	
Vasodilator drugs (n%)			0.034
Yes	247.0 (24.9)	9.0 (13.4)	
No	744.0 (75.1)	58.0 (86.6)	
Hypotensive drugs (n%)			0.51
Yes	94.0 (9.5)	8.0 (11.9)	
No	897.0 (90.5)	59.0 (88.1)	
Beta-blockers (n%)			0.353
Yes	387.0 (39.1)	30.0 (44.8)	
No	604.0 (60.9)	37.0 (55.2)	

Feature Selection

Feature selection was performed using the XGBoost model. Figure 2 displays the top 10 variables selected for their importance among the counting variables and continuous variables, which include characteristics such as drinking, chronic obstructive pulmonary disease (COPD), percutaneous transluminal coronary intervention (PCI), gender, smoking, cerebral stroke, beta-blockers, diabetes, age, RBC, blood glucose, lactate dehydrogenase, BMI, apolipoprotein A1 (APOA1), creatine kinase isoenzyme, and others. Ultimately, a set of 20 optimal features was identified and used to construct the model.

Model Prediction Performance Comparison

Table 2 presents the performance metrics of the evaluated machine learning models, showing that the BRF model outperformed the other five models with an AUC of 0.810 (95% CI:0.778, 0.839), G-mean of 0.806 (95% CI:0.778, 0.827), sensitivity of 0.990 (95% CI:0.981, 1.000), and recall of 0.990 (95% CI:0.981, 1.000). Furthermore, the Brier score of the BRF model was the lowest at 0.181 (95% CI:0.178, 0.185) compared to the other three unbalanced ensemble

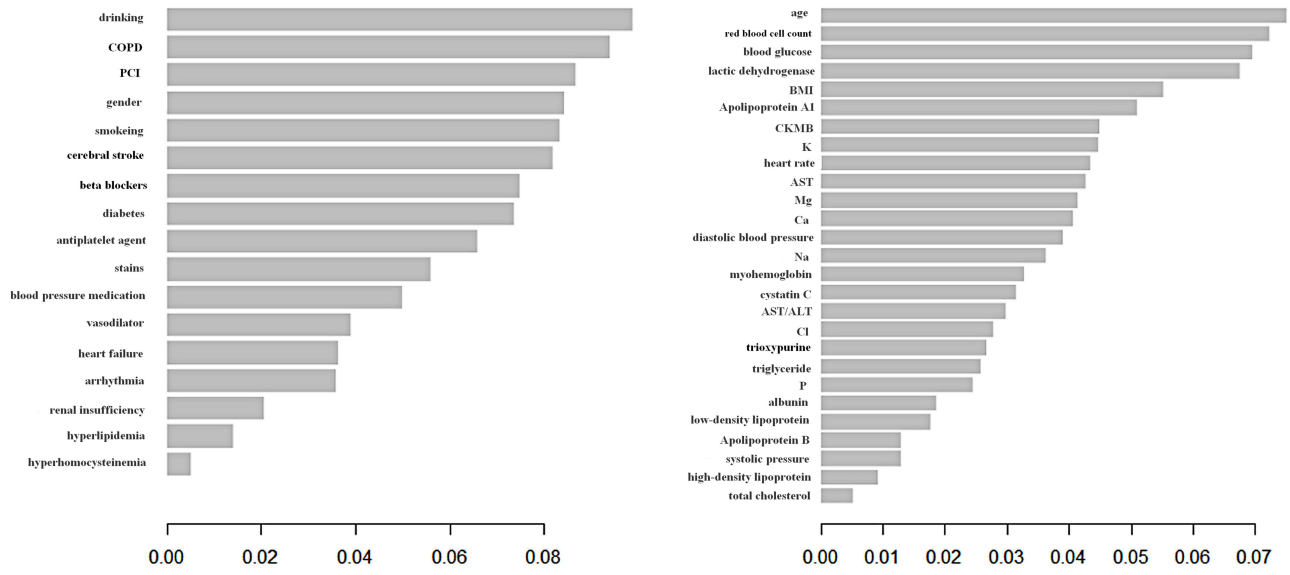


Figure 2 The importance of feature screening.
Notes: BMI, Body Mass Index; CKMB, Creatine Kinase Isoenzyme; K, potassium; AST, glutamic oxaloacetic transaminase; Mg, magnesium; Ca, calcium; Na, sodium; AST/ALT, glutamic-oxaloacetic transaminase/glutamic-pyruvic transaminase; Cl, chlorine; P, phosphorus; COPD, chronic obstructive pulmonary disease; PCI, Percutaneous Transluminal Coronary Intervention).

models, and the difference between the metrics was statistically significant. Therefore, we selected the BRF model as the optimal model for further analysis.

Figure 3 illustrates that the BRF model has the largest area under the ROC curve, and the lift curves indicate lift values greater than 1 and 20, suggesting that the model is both stable and efficient. The KS statistics of the model was 0.995, the highest among all six models, indicating that the model had the best discrimination ability. Moreover, the precision-recall curves of the positive and negative BRF curves had the highest areas of 1.000 and 0.989, respectively. Thus, we conclude that the BRF model is optimal for this study.

Interpretation of Predictive Features

The SHAP values provide information about the contribution of each feature to the final prediction and help explain the model predictions for individual patients. Figure 4A illustrates the influential features of the model, including age, BMI, RBC, lactate dehydrogenase, COPD, and APOA1. In the SHAP interpretation model, each point represents a sample and is stacked lengthwise according to the sample size, with the colour indicating the eigenvalue (red for high values and blue for low values). Figure 4B shows that low age, high RBC, high BMI, low lactate dehydrogenase, high apolipoprotein A1, low blood potassium, low heart rate, and high blood calcium positively affect the model prediction. Specifically, as

Table 2 Indicators of Six Models

Models/indicators	AUC	Sensitive	Recall	Brier	G-mean
BRF	**0.810 (0.778,0.839)**	**0.990 (0.981,1.000)**	**0.990 (0.981,1.000)**	**0.181 (0.178,0.185)**	**0.806 (0.778,0.827)**
EasyEnsemble	0.774 (0.770,0.778)*	0.983 (0.979,0.987)*	0.983 (0.979,0.987)*	0.215 (0.213,0.217)*	0.773 (0.770,0.777)*
RUSBoost	0.713 (0.677,0.745)*	0.963 (0.959,0.978)*	0.963 (0.959,0.978)*	0.219 (0.215,0.221)*	0.696 (0.654,0.737)*
SMOTEBoost	0.624 (0.611,0.662)*	0.955 (0.948,0.969)*	0.955 (0.948,0.969)*	0.209 (0.208,0.210)*	0.562 (0.536,0.609)*
AdaBoost	0.531 (0.500,0.571)*	0.938 (0.928,0.949)*	0.938 (0.928,0.949)*	0.159 (0.156,0.162)*	0.352 (0.235,0.444)*
Logistic	0.497 (0.494,0.500)*	0.934 (0.931,0.938)*	0.934 (0.931,0.938)*	0.064 (0.060,0.069)*	0.246 (0.238,0.252)*

Notes: Predictive models for BRF, EasyEnsemble, RUSBoost, SMOTEBoost, AdaBoost, and Logistic CHD comorbid with hypertension were constructed from the same training set and applied to the same testing set. A comparison of the prediction results of various models used the median statistical description. Median (Q1-Q3) means Median and Interquartile Range. Nonparametric Friedman test and Nemenyi post hoc test were used to make a comparison with the BRF, " *" $P < 0.05$, the bolded font in the table indicates the best index value of the same class).

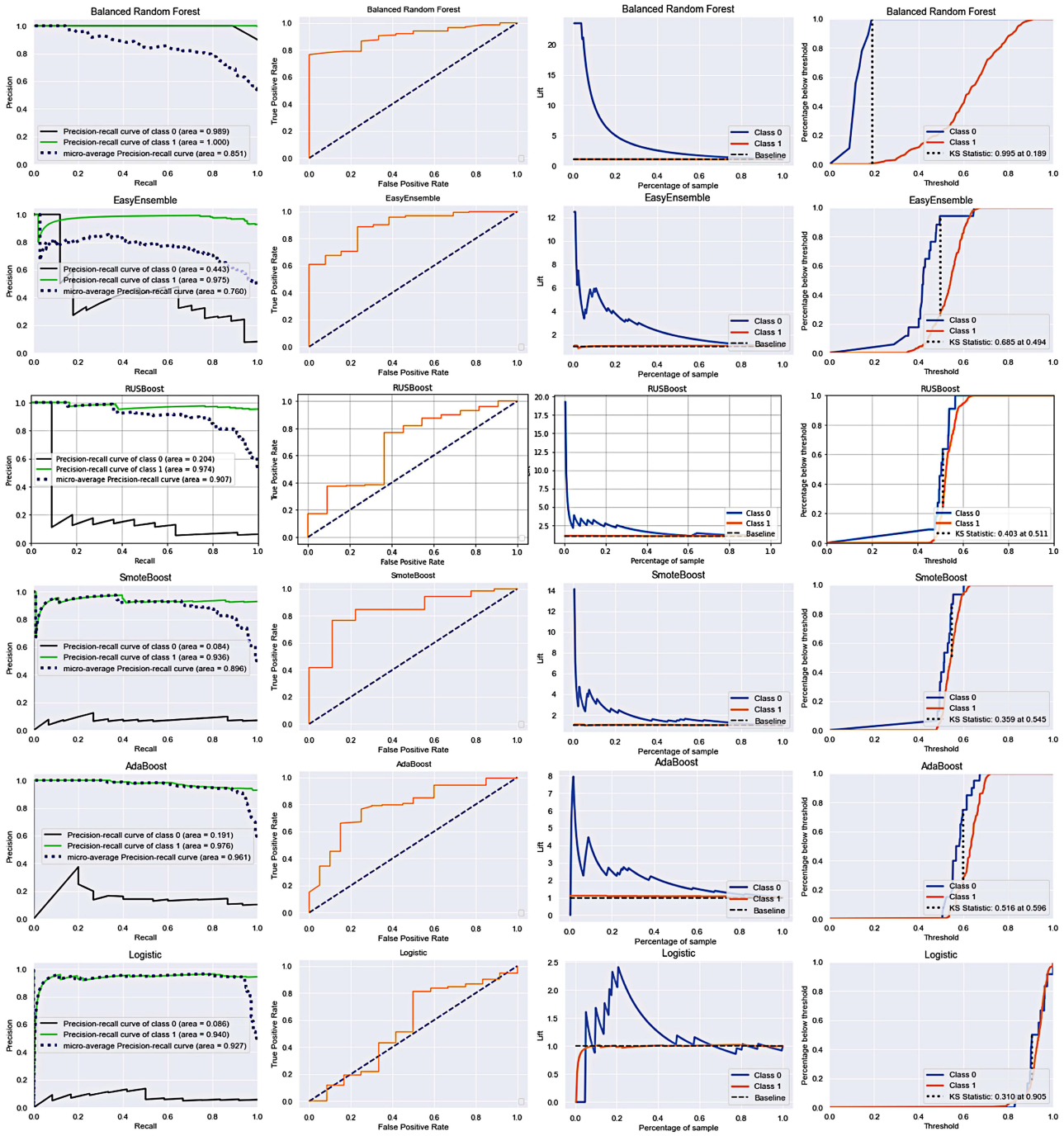


Figure 3 Precise-recall curve, ROC, lift curve, and Kolmogorov Smirnov curve for six models.

shown in Figure 5, these factors play a positive role in predicting improved outcomes and a negative role in predicting death outcomes.

Discussion

Hypertension is considered a crucial risk factor for CHD. The mortality rate of patients with CHD comorbid with hypertension is higher than that of patients with CHD.[33,34] In this study, a combination of an unbalanced ensemble algorithm and advanced machine learning methods was utilised to accurately predict the mortality outcomes of patients with CHD comorbid with hypertension and assess the risk factors associated with mortality.

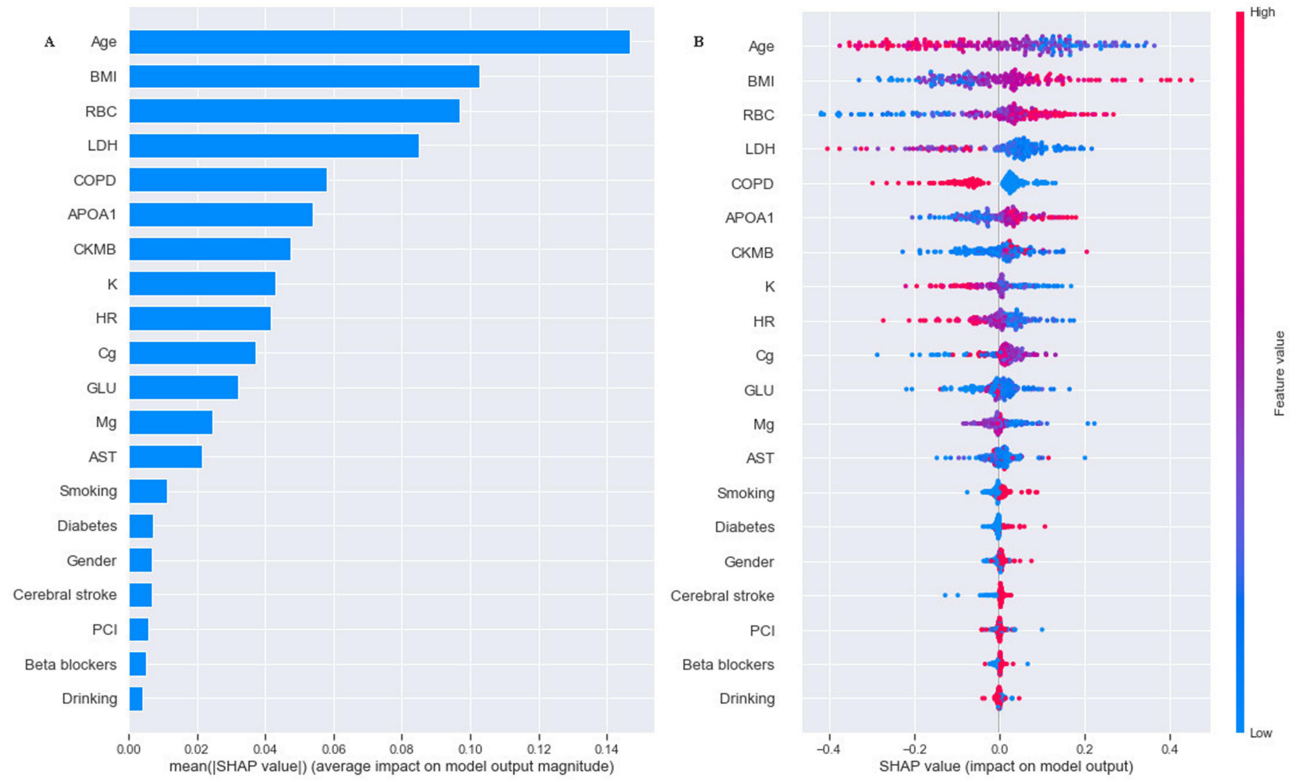


Figure 4 Feature importance (**A**) and model interpretability (**B**).
Notes: RBC, red blood cell count; LDH, lactate dehydrogenase; APOA1, Apolipoprotein A1; HR, heart rate; GLU, blood glucose; BMI, Body Mass Index; CKMB, Creatine Kinase Isoenzyme; K, potassium; AST, glutamic oxalacetic transaminase; Mg, magnesium; Ca, calcium; Na, sodium; AST/ALT, glutamic-oxalacetic transaminase/ glutamic-pyruvic transaminase; Cl, chlorine; P, phosphorus; COPD, chronic obstructive pulmonary disease; PCI, Percutaneous Transluminal Coronary Intervention).

Based on the experimental results, we have demonstrated that the unbalanced ensemble algorithm outperforms traditional machine learning algorithms in predicting the prognosis of CHD patients comorbid with hypertension. Additionally, we found that undersampling may be a better or at least comparable approach to oversampling.[35] BRF, EasyEnsemble, and RUSBoost, proposed in this study, demonstrated clear advantages over SMOTEBoost in predicting the mortality of patients with CHD comorbid with hypertension. In the ensemble model based on undersampling, BRF introduces undersampling in each round of self-service sampling, significantly improving its performance and exceeding that of the other two models.[36] The model's superiority is evident in various metrics, such as AUC, G-mean, and lift curves. In addition, the Brier score of the BRF is also better than that of the other unbalanced ensemble algorithms, which shows that the model has good calibration and stable performance and can avoid the phenomenon of overfitting.[37] The results obtained in this study demonstrate that BRF can effectively address the issue of classification bias arising from unbalanced data. This finding could serve as a useful reference for predicting unbalanced data in the biomedical field.

Most models currently being developed utilise traditional statistical methods. In model construction and utilisation, advanced statistical methods and machine learning techniques have not been fully exploited to enhance the predictive capability of the model.[38] In this study, we utilised not only advanced unbalanced integrated models but also incorporated other effective machine learning methods to improve model performance. For instance, missing value filling was implemented based on missing forests,[39] model optimisation was conducted using GridSearchCV,[40] and feature filtering was performed using XGBoost.[41] In summary, the BRF model demonstrated significant advantages in handling unbalanced data in the CHD comorbid with hypertension dataset. The use of BRF in this study has established a strong foundation for predicting the clinical prognosis of patients with CHD comorbid with hypertension. Accurate predictions of adverse outcomes can facilitate doctors to promptly adjust their treatment plans for better prognosis and treatment.

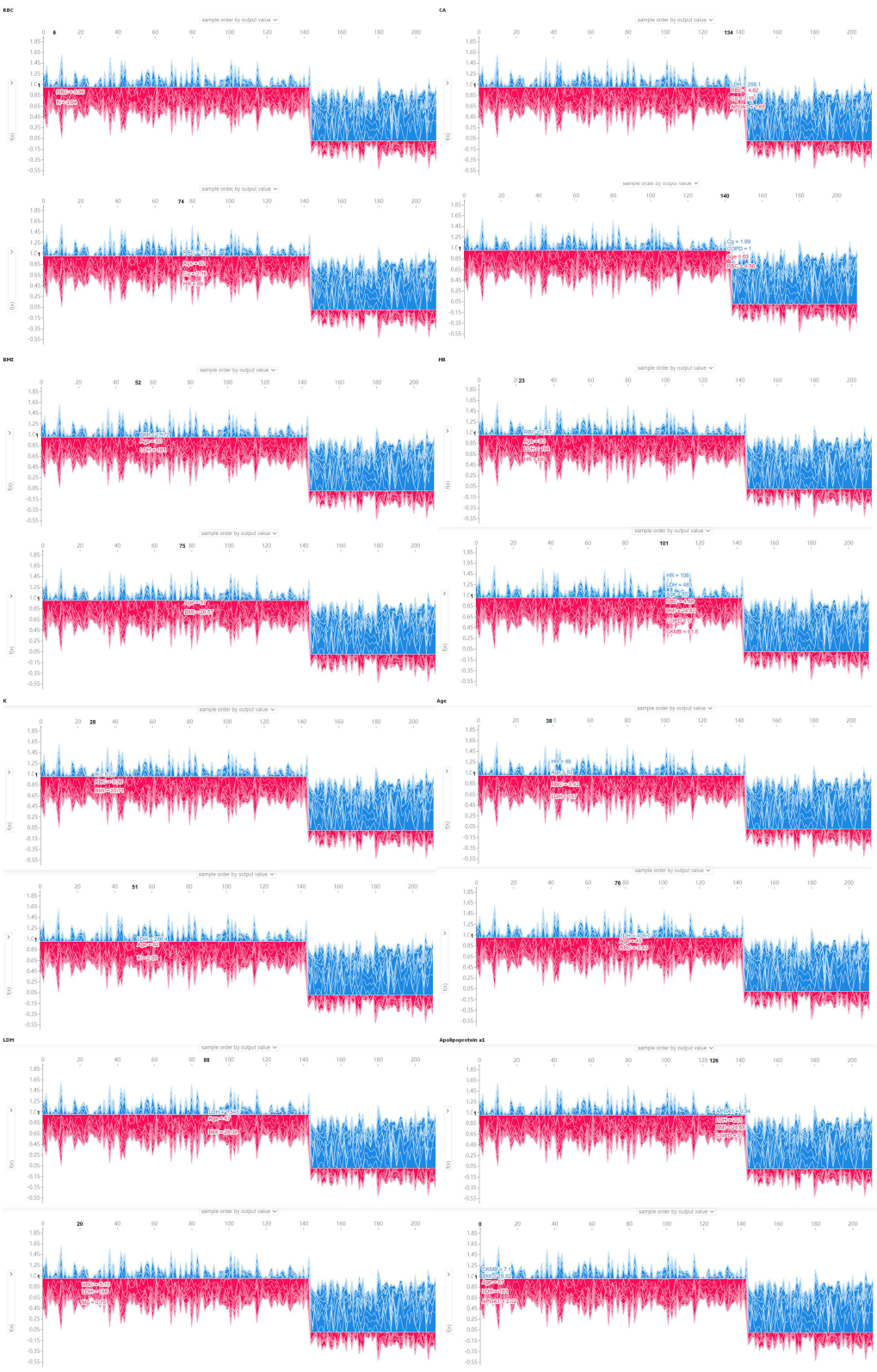


Figure 5 Illustration of the model results with eight sets of samples.
Note:1= Survivors, 0= Nonsurvivors.

Although many machine learning models can display variable importance, it can be challenging to explain the impact of each variable on the model. In addition, the lack of interpretability of machine learning models for clinicians is a major obstacle in biomedical machine learning. In this study, we utilised machine learning techniques to explain the importance of features in a particular model. Specifically, SHAP was employed to elucidate the significance of each feature. Age, BMI, RBC, lactate dehydrogenase, COPD, apolipoprotein A1, creatine kinase isoenzyme, blood potassium, blood calcium, and heart rate were identified as significant predictors. The importance of these factors has been confirmed in other studies. Studies have indicated that low levels of apolipoprotein A1, as well as high or low BMI and high levels of blood potassium, may exacerbate coronary atherosclerosis, which is associated with an increased risk of CHD and hypertension.[42–44] In patients with COPD, the increase in intrathoracic pressure due to dyspnoea-induced hyperinflation of the lungs can result in excessive activation of sympathetic nerves and decreased sensitivity to pressure-sensitive receptors. These factors contribute to elevated blood pressure, which can worsen the progression of primary diseases such as CHD;[45] both lactate dehydrogenase and creatine kinase isoenzyme are independent risk factors for CHD.[46,47] Low blood calcium levels have been associated with a higher incidence of hypertension, and blood calcium plays a role in promoting sodium excretion, which can lower blood pressure. Additionally, blood calcium can affect the synthesis of norepinephrine, inhibiting the development of atherosclerosis.[48] Rapid heart rate may contribute to an increased risk of hypertension, while elevated blood pressure can damage heart function. Increased heart rate in patients with CHD comorbid with hypertension may lead to rupture of atherosclerotic plaques and form block cardiovascular and cerebrovascular thrombosis.[49] Studies have confirmed that age is the most significant factor, and the mortality rate of elderly patients with CHD comorbid with hypertension is higher.[50] Additionally, red blood cell count was identified as one of the top 20 important variables in this study, which has been infrequently reported in previous CHD studies. These results suggest that the RBC count is an effective independent risk factor for mortality in patients with CHD comorbid with hypertension. The prognostic significance of these factors in patients with CHD and comorbid hypertension merits attention.

The next step of this research plan is to introduce an online risk prediction platform based on the prediction model, and integrate the unbalanced ensemble prediction model and SHAP into the online platform through the interactive module by Streamlit, This will enable ordinary patients and medical staff can judge the risk of disease according to the clinical indicators of patients in real time. This provides theoretical support for clinicians to adjust diagnosis and treatment methods and patients to change their lifestyle promptly. The establishment of this system is expected to improve the prediction accuracy of patients with coronary heart disease complicated with hypertension and make medical decision-making more personalized and accurate.

Limitations and Development

All patient information in this study was collected within the Shanxi Province, resulting in a potential geographical bias. Therefore, we plan to collect data from hospitals nationwide and use information from various regions and hospitals as an external validation set for the model. Moreover, because the information collected in this study is limited to structured data, further research is required to extract and utilise unstructured data and incorporate additional features such as imaging data, biomarkers, environmental factors, lifestyle habits, and psychological changes to improve the model's predictive performance. Another limitation of this study is that it only discusses the prognosis of patients with CHD comorbid with hypertension. Although this study achieved conclusive results, further improvement is still possible. With the rapid development of artificial intelligence and big data, future research will combine more extensive data for different levels of research.In the future, we plan to conduct subgroup analysis among different populations to analyze whether there are differences in the efficacy of the prediction model among different populations.

Conclusion

In this study, the unbalanced ensemble algorithm effectively improved the prognosis of death outcomes of patients with CHD comorbid with hypertension and classified them accurately. This method can solve the problem of unbalanced datasets in future medical research. Simultaneously, the advanced machine learning model and SHAP can visually explain the variables affecting the performance of the model. Our study might help clinicians better understand the severity of the disease and promptly adjust the treatment plan.

Data Sharing Statement

Data and any appendix material related to this article can be obtained from the corresponding author on request.

Ethical Approval

This study complied with the Helsinki Declaration and was approved by the Medical Ethics Committee of Shanxi Medical University. All patients were informed of the study objectives and provided written informed consent.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This work was supported by the National Natural Science Foundation of China under Grant [number 82173631]; Shanxi Science and Technology innovation talent team project [number 202204051001026].

Disclosure

The authors report no conflicts of interest in this work.

References

1. Virani SS, Alonso A, Benjamin EJ, et al. Heart disease and stroke statistics-2020 update: A report from the American heart association. *Circulation*. 2020;141(9):e139–e596. doi:10.1161/cir.0000000000000757
2. Bauersachs R, Zeymer U, Brière JB, Marre C, Bowrin K, Huelsebeck M. Burden of coronary artery disease and peripheral artery disease: a literature review. *Cardiovasc Ther*. 2019;2019(8295054). doi:10.1155/2019/8295054
3. Yang X, Li J, Hu D, et al. Predicting the 10-year risks of atherosclerotic cardiovascular disease in Chinese population: the china-par project (prediction for ASCVD risk in china). *Circulation*. 2016;134(19):1430–1440. doi:10.1161/circulationaha.116.022367
4. Beaglehole R, Magnus P. The search for new risk factors for coronary heart disease: occupational therapy for epidemiologists? *Int J Epidemiol*. 2002;31(6):1117–1122. doi:10.1093/ije/31.6.1117
5. Lefèvre G, Puymirat E. Hypertension and coronary artery disease: new concept?. *Annales de cardiologie et d'angeiologie*. 2017;66(1):42–47. doi:10.1016/j.ancard.2016.10.011
6. Scrutinio D, Bellotto F, Lagioia R, Passantino A. Physical activity for coronary heart disease: cardioprotective mechanisms and effects on prognosis. *Monaldi Archives for Chest Dis*. 2005;64(2):77–87. doi:10.4081/monaldi.2005.591
7. Lewis CE, Fine LJ, Beddhu S, et al. Final report of a trial of intensive versus standard blood-pressure control. *New Engl J Med*. 2021:384 (20):1921–1930. doi:10.1056/NEJMoa1901281
8. Gao BF, Shen ZC, Bian WS, Wu SX, Kang ZX, Gao Y. Correlation of hypertension and F2RL3 gene methylation with Prognosis of coronary heart disease. *J Biol Regul Homeost Agent*. 2018;32(6):1539–1544.
9. Feeny AK, Chung MK, Madabhushi A, et al. Artificial intelligence and machine learning in arrhythmias and cardiac electrophysiology. *Circ Arrhythm Electrophysiol*. 2020;13(8):e007952. doi:10.1161/circep.119.007952
10. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC Med Res Meth*. 2019;19(1):64. doi:10.1186/s12874-019-0681-4
11. Kumar A, Sinha N, Bhardwaj A. A novel fitness function in genetic programming for medical data classification. *J biomed informat*. 2020;112:103623. doi:10.1016/j.jbi.2020.103623
12. Du G, Zhang J, Jiang M, et al.: Graph-based class-imbalance learning with label enhancement. *IEEE transactions on neural networks and learning systems* 2021.
13. Wang S, Dai Y, Shen J, Xuan J. Research on expansion and classification of imbalanced data based on SMOTE algorithm. *Sci Rep*. 2021;11 (1):24039. doi:10.1038/s41598-021-03430-5
14. Wu Y, Fang Y. Stroke Prediction with Machine Learning Methods among Older Chinese. *Int J Environ Res Public Health*. 2020;17(6):1828.
15. Wang K, Xue Q, Xing Y, Li C. Improve Aggressive Driver Recognition Using Collision Surrogate Measurement and Imbalanced Class Boosting. *Int J Environ Res Public Health*. 2020;17(7):2375.
16. Liu L, Zhang C, Zhang G, et al. A study of aortic dissection screening method based on multiple machine learning models. *J Thoracic Dis*. 2020;12 (3):605–614. doi:10.21037/jtd.2019.12.119
17. Vong CM, Du J. Accurate and efficient sequential ensemble learning for highly imbalanced multi-class data. *Neural Net*. 2020;128:268–278. doi:10.1016/j.neunet.2020.05.010
18. Tesche C, Bauer MJ, Baquet M, et al. Improved long-term prognostic value of coronary CT angiography-derived plaque measures and clinical parameters on adverse cardiac outcome using machine learning. *Eur Radiol*. 2021;31(1):486–493. doi:10.1007/s00330-020-07083-2

19. Gulati M, Levy PD, Mukherjee D, et al. 2021 AHA/ACC/ASE/CHEST/SAEM/SCCT/SCMR Guideline for the Evaluation and Diagnosis of Chest Pain: a Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation*. 2021;144(22):e368–e454. doi:10.1161/cir.0000000000001029
20. Jones NR, McCormack T, Constanti M, McManus RJ. Diagnosis and management of hypertension in adults: NICE guideline update 2019. *British j Gene Practi*. 2020;70(691):90–91. doi:10.3399/bjgp20X708053
21. Schmitt P, El J, Guedj mjjob, biostatistics: A comparison of six methods for missing data imputation. 2015, 6.
22. Stekhoven DJ, Bühlmann P. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28(1):112–118. doi:10.1093/bioinformatics/btr597
23. Yang M, Fan H, Zhao K. PM(2.5) Prediction with a Novel Multi-Step-Ahead Forecasting Model Based on Dynamic Wind Field Distance. *Int J Environ Res Public Health*. 2019;16(22):4482. doi:10.3390/ijerph16224482
24. Chung H, Ko H, Kang WS, et al. Prediction and feature importance analysis for severity of covid-19 in south korea using artificial intelligence: model development and validation. *J Med Int Res*. 2021;23(4):e27060. doi:10.2196/27060
25. Degórski L, Kobylinski L, AJIMoCS P, Technology I: definition extraction: Improving Balanced Random Forests. 2008:353–357.
26. Kang Q, Chen X, Li S, Zhou M. A noise-filtered under-sampling scheme for imbalanced classification. *IEEE Transactions on Cybernetics*. 2017;47 (12):4263–4274. doi:10.1109/tcyb.2016.2606104
27. Blanchard M, Feuilloy M, Gervès-Pinquié C, et al. Cardiovascular risk and mortality prediction in patients suspected of sleep apnea: a model based on an artificial intelligence system. *Physiol Meas*. 2021;42(10):105010. doi:10.1088/1361-6579/ac2a8f
28. Assel M, Sjoberg DD, Vickers AJ. The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models. *Diagnos Prognostic Res*. 2017;1:19. doi:10.1186/s41512-017-0020-3
29. Du J, Vong CM, Pun CM, Wong PK, Ip WF. Post-boosting of classification boundary for imbalanced data using geometric mean. *Neural Net*. 2017;96:101–114. doi:10.1016/j.neunet.2017.09.004
30. Shahinfar S, Guenther JN, Page CD, et al. Optimization of reproductive management programs using lift chart analysis and cost-sensitive evaluation of classification errors. *J dairy Sci*. 2015;98(6):3717–3728. doi:10.3168/jds.2014-8255
31. Wee S, Choi C, Jeong J. Blind Interleaver Parameters Estimation Using Kolmogorov-Smirnov Test. *Sensors*. 2021;21(10):3458. doi:10.3390/s21103458
32. Wen X, Xie Y, Wu L, Jiang L. Quantifying and comparing the effects of key risk factors on various types of roadway segment crashes with LightGBM and SHAP. *Acc Analysis and Preven*. 2021;159:106261. doi:10.1016/j.aap.2021.106261
33. Turin TC, Okamura T, Afzal AR, et al. Impact of hypertension on the lifetime risk of coronary heart disease. *Hypertension Rese*. 2016;39 (7):548–551. doi:10.1038/hr.2016.23
34. Malakar AK, Choudhury D, Halder B, Paul P, Uddin A, Chakraborty S. A review on coronary artery disease, its risk factors, and therapeutics. *J Cell Physiol*. 2019;234(10):16812–16823. doi:10.1002/jcp.28350
35. Jeong DH, Kim SE, Choi WH, Ahn SH. A comparative study on the influence of undersampling and oversampling techniques for the classification of physical activities using an imbalanced accelerometer dataset. *Healthcare*. 2022;10(7):1255. doi:10.3390/healthcare10071255
36. Yagci AM, Aytekin T, Gurgen FS, IEEE: balanced random forest for imbalanced data streams. In: *24th Signal Processing and Communication Application Conference (SIU)*: *Zonguldak, TURKEY*. 2016: 1065–1068.
37. Nistal-Nuno B. Machine learning applied to a cardiac surgery recovery unit and to a coronary care unit for mortality prediction. *J Clin Mon Compu*. 2022;36(3):751–763. doi:10.1007/s10877-021-00703-2
38. 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA). *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)* 2015:555 pp.
39. Yang H, Li X, Cao H, et al. Using machine learning methods to predict hepatic encephalopathy in cirrhotic patients with unbalanced data. *Comput Meth Progr Biomed*. 2021;211:106420. doi:10.1016/j.cmpb.2021.106420
40. Adnan M, Alarood AAS, Uddin MI, Ur Rehman I. Utilizing grid search cross-validation with adaptive boosting for augmenting performance of machine learning models. *PeerJ Comput Sci*. 2022;8:e803. doi:10.7717/peerj-cs.803
41. Li Z, Liu Z. Feature selection algorithm based on XGBoost. *J Commun*. 2019;40(10).
42. Held C, Hadziosmanovic N, Aylward PE, et al. Body Mass Index and Association With Cardiovascular Outcomes in Patients With Stable Coronary Heart Disease - A STABILITY Substudy. *J American Heart Associa*. 2022;11(3):e023667. doi:10.1161/JAHA.121.023667
43. Thompson A, Danesh J. Associations between apolipoprotein B, apolipoprotein AI, the apolipoprotein B/AI ratio and coronary heart disease: a literature-based meta-analysis of prospective studies. *J Internal Med*. 2006;259(5):481–492. doi:10.1111/j.1365-2796.2006.01644.x
44. Xu Z, Zhang Y, Zhang C, Xiong F, Zhang J, Xiong J. Clinical Features and Outcomes of COVID-19 Patients with Acute Kidney Injury and Acute Kidney Injury on Chronic Kidney Disease. *Aging and Disease*. 2022;13(3):884–898. doi:10.14336/AD.2021.1125
45. Herych PR, Iatsyshyn RI. [Treatment and prevention of combined cardio-respiratory pathology in exacerbation of chronic obstructive-pulmonary disease (new approach)]. *Likars'ka Sprava*. 2014;(7–8):38–46.
46. Liu ZF, Hu WW, Li R, Gao Y, Yan LL, Su N. Expression of lncRNA-ANRIL in patients with coronary heart disease before and after treatment and its short-term prognosis predictive value. *Eur Rev Med Pharmacol Sci*. 2020;24(1):376–384. doi:10.26355/eurrev_202001_19936
47. Zhu W, Ma Y, Guo W, et al. Serum Level of Lactate Dehydrogenase is Associated with Cardiovascular Disease Risk as Determined by the Framingham Risk Score and Arterial Stiffness in a Health-Examined Population in China. *Int J Gene Med*. 2022;15:11–17. doi:10.2147/ijgm.S337517
48. Ha AW, Kim WK, Kim SH. Cow's Milk Intake and Risk of Coronary Heart Disease in Korean Postmenopausal Women. *Nutrients*. 2022;14 (5):1092. doi:10.3390/nu14051092
49. Dalal J, Dasbiswas A, Sathyamurthy I, et al. Heart Rate in Hypertension: review and Expert Opinion. *Int j Hyper*. 2019;2019(2087064):1–6. doi:10.1155/2019/2087064
50. Kikuchi N, Ogawa H, Kawada-Watanabe E, et al. Impact of age on clinical outcomes of antihypertensive therapy in patients with hypertension and coronary artery disease: a sub-analysis of the Heart Institute of Japan Candesartan Randomized Trial for Evaluation in Coronary Artery Disease. *J clin hyperten*. 2020;22(6):1070–1079. doi:10.1111/jch.13891