


Machine learning-based model for worsening heart failure risk in Chinese chronic heart failure patients

Ziyi Sun^{1,2} , Zihan Wang^{2,3}, Zhangjun Yun^{2,4}, Xiaoning Sun¹, Jianguo Lin¹, Xiaoxiao Zhang¹, Qingqing Wang¹, Jinlong Duan¹, Li Huang³, Lin Li^{3*} and Kuiwu Yao^{1,5*}

¹Guang'anmen Hospital, China Academy of Chinese Medical Sciences, Beijing, China; ²Graduate School, Beijing University of Chinese Medicine, Beijing, China; ³China-Japan Friendship Hospital, Beijing, China; ⁴Dongzhimen Hospital, Beijing University of Chinese Medicine, Beijing, China; and ⁵Academic Administration Office, China Academy of Chinese Medical Sciences, Beijing, China

Abstract

Aims This study aims to develop and validate an optimal model for predicting worsening heart failure (WHF). Multiple machine learning (ML) algorithms were compared, and the results were interpreted using SHapley Additive exPlanations (SHAP). A clinical risk calculation tool was subsequently developed based on these findings.

Methods and results This nested case–control study included 200 patients with chronic heart failure (CHF) from the China-Japan Friendship Hospital (September 2019 to December 2022). Sixty-five variables were collected, including basic information, physical and chemical examinations, and quality of life assessments. WHF occurrence within a 3-month follow-up was the outcome event. Variables were screened using LASSO regression, univariate analysis, and comparison of key variables in multiple ML models. Eighty per cent of the data was used for training and 20% for testing. The best models were identified by integrating nine ML algorithms and interpreted using SHAP, and to develop a final risk calculation tool. Among participants, 68 (34.0%) were female, with a mean age (standard deviation, SD) of 68.57 (12.80) years. During the follow-up, 60 participants (30%) developed WHF. N-terminal pro-brain natriuretic peptide (NT-proBNP), creatinine (Cr), uric acid (UA), haemoglobin (Hb), and emotional area score on the Minnesota Heart Failure Quality of Life Questionnaire were critical predictors of WHF occurrence. The random forest (RF) model was the best model to predict WHF with an area under the curve (AUC) (95% confidence interval, CI) of 0.842 (0.675–1.000), accuracy of 0.775, sensitivity of 0.900, specificity of 0.833, negative predictive value of 0.800, and positive predictive value of 0.600 for the test set. SHAP analysis highlighted NT-proBNP, UA, and Cr as significant predictors. An online risk predictor based on the RF model was developed for personalized WHF risk assessment.

Conclusions This study identifies NT-proBNP, Cr, UA, Hb, and emotional area scores as crucial predictors of WHF in CHF patients. Among the nine ML algorithms assessed, the RF model showed the highest predictive accuracy. SHAP analysis further emphasized NT-proBNP, UA, and Cr as the most significant predictors. An online risk prediction tool based on the RF model was subsequently developed to enhance early and personalized WHF risk assessment in clinical settings.

Keywords Chronic heart failure; Machine learning; Prediction; Risk models; Worsening heart failure

Received: 23 February 2024; Revised: 25 May 2024; Accepted: 21 August 2024

*Correspondence to: Lin Li, Department of Integrated Chinese and Western Medicine Cardiology, China-Japan Friendship Hospital, No. 2 Cherry Blossom Garden East Street, Beijing, China. Email: lilinxcy@126.com;

Kuiwu Yao, Academic Administration Office, China Academy of Chinese Medical Sciences, No. 16, Nanxiaojie, Dongzhimennei, Dongcheng District, Beijing, China. Email: yaokuiwu@126.com

Ziyi Sun, Zihan Wang and Zhangjun Yun contributed equally to this work and share first authorship.

Introduction

Heart failure (HF) is the end stage of a wide range of cardiovascular diseases and is characterized by variable lengths of time for symptoms to stabilize. Despite continuous treatment, the condition worsens repeatedly throughout the

disease.¹ Worsening HF (WHF) has been proposed as a specific stage during HF. WHF is defined as a patient with chronic HF (CHF) whose signs and symptoms of HF continue to worsen despite stable background therapy, requiring urgent intensification of therapy, including inpatient, emergency, or outpatient intravenous or oral diuretic therapy.¹ Studies have

shown that patients with WHF are directly associated with higher mortality rates.^{2,3} This suggests that early identification and warning of WHF is necessary. Moreover, according to recent epidemiological findings, there are approximately 300 million patients with stage B (pre-HF) and stage C (symptomatic HF) in China, which further adds to the urgency of addressing this clinical problem.⁴

Machine learning (ML) algorithms are now widely used for early warning prediction of diseases. ML can effectively utilize large multidimensional data. ML algorithms are not constrained by assumptions of distributional normality, non-informative or random census, and linearity of hazard risk and are more advantageous in dealing with clinical multi-dimensional data and outperform traditional modelling approaches compared to previous models.⁵ Most current predictive modelling studies of HF have focused on predicting patients' long-term mortality and readmission risk.^{6,7} Recently, Parikh *et al.*⁸ developed a predictive model based on the gradient boosted decision tree models for predicting the occurrence of WHF in patients with different types of HF with an area under the curve (AUC) of 0.76 and a mean squared error of 0.13. It is clear that prediction for the occurrence of WHF brings the risk forward compared with prediction for the occurrence of cardiovascular endpoint events. This is very beneficial in reducing the incidence of short-term WHF and long-term cardiovascular mortality in patients with HF. More importantly, most current HF prediction models are based on Western populations. Considering the increasing burden of HF and healthcare in China, developing risk prediction models tailored to the Chinese population has become particularly important.

This study aims to develop an early warning model for WHF occurrence based on ML algorithm. The model results were interpreted with the help of the SHapley Additive exPlained (SHAP) method, and ultimately a risk calculation tool was built to suit the clinical application. This helps further improve the stratified treatment strategy for HF patients and provides more reference for clinicians.

Materials and methods

Study design

The study was a nested case-control study that followed the TRIPOD reporting specifications.⁹ From September 2019 to December 2022, 263 inpatients with New York Heart Association (NYHA) class II or higher CHF were screened at the Department of Integrated Chinese and Western Medicine, China-Japan Friendship Hospital. Patients with acute HF, patients with stage D HF (patients receiving left ventricular assist devices or heart transplantation), patients in the acute phase of acute coronary syndrome, patients with

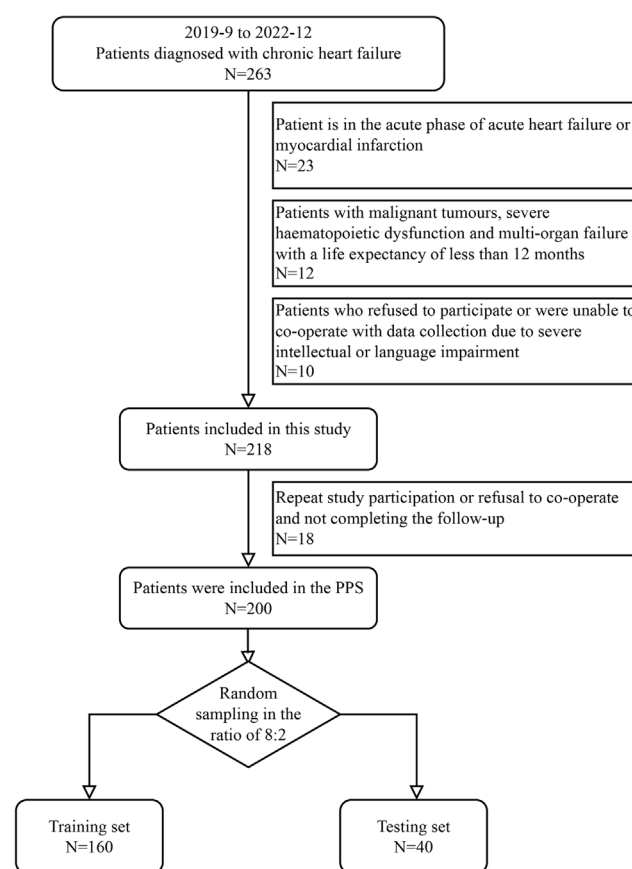
cardiac disorders such as severe arrhythmias with haemodynamic instability, patients with severe haematopoietic dysfunction, patients with malignant neoplasms and patients with multiorgan dysfunction, as well as patients with severe psychological disorders, psychiatric disorders, or speech disorders are excluded. Forty-five patients were excluded from the study, and the patient screening process is shown in *Figure 1*.

The study was conducted following the principles of the Declaration of Helsinki, and all patients participated voluntarily and provided informed consent. The study protocol was approved by the Clinical Research Ethics Committee of China-Japan Friendship Hospital (2019-120-K82) and registered with the China Clinical Trial Registry (ChiCTR1900024482).

Diagnostic criteria

The diagnostic criteria for WHF are the Deterioration of HF signs and symptoms in a patient with CHF despite previous stable background therapy. Requires urgent escalation of therapy, including hospitalization, emergency department

Figure 1 Flow chart of patient inclusion. PPS: per-protocol set.



visit, or outpatient intravenous diuretic therapy, \pm outpatient oral therapy.^{1,10} The criteria for diagnosing and classifying CHF refer to the 2022 AHA/ACC/HFSA Guidelines for the Management of HF.¹¹

Measurement

Basic information

Basic information about the study population was collected, including demographics (gender, age, height, weight, and education), vital signs (heart rate, non-invasive arterial systolic and diastolic blood pressure), co-morbidities [coronary artery disease, hypertension, type 2 diabetes mellitus, chronic kidney disease (CKD), stroke, atrial fibrillation, and hyperlipidaemia], and personal habits (history of tobacco use, alcohol use, and frequency of daily exercise). The frequency of the exercise routine was left to the patient's discretion. Exercise frequency was defined as hardly exercise (aerobic exercise less than four times per month for more than 30 min each time), little exercise (aerobic exercise less than three times per week for more than 30 min each time), and regular exercise (aerobic exercise more than three times per week for more than 30 min each time).¹²

Testing and examination

Fasting venous blood was collected from all participants on the morning of the second day after enrolment in the trial. Routine blood tests [white blood cell count (WBC), red blood cell count (RBC), haemoglobin (Hb), haematocrit (HCT), and platelets (PLT)], liver and kidney function [alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), total bilirubin (TBIL), blood urea nitrogen (BUN), and uric acid (UA)], lipids [total cholesterol (CHO), triglycerides (TG), low-density lipoprotein (LDL), and high-density lipoprotein (HDL)], electrolytes (K^+ , Na^+ , Cl^- , Ca^{2+}), high-sensitivity CRP (Hs-CRP), glycated haemoglobin (HbA1c), and the N-terminal brain natriuretic peptide precursor (NTpro-BNP) were performed in the clinical laboratory of China-Japan Friendship Hospital. Echocardiography was uniformly conducted for all participants in the Ultrasound Department of the China-Japan Friendship Hospital by two experienced sonographers. The assessed indices included left ventricular ejection (LVEF), left ventricular end-diastolic diameter (LVDd), left ventricular end-systolic diameter (LVDs), left ventricular fractional shortening (LVFS), and so on.

Medications

In this study, there was no intervention in the participants' medication, and their supervising physician prescribed the patients' treatment. We recorded the use of angiotensin-converting enzyme inhibitors (ACEIs), angiotensin II receptor blockers (ARBs), angiotensin receptor enkephalinase inhibitors (ARNIs), beta-blockers, loop diuretics, spironolactone, digoxin, aspirin/clopidogrel, statins, nitrates, and traditional

Chinese medicines (TCMs) or proprietary Chinese medicines (PCMs) by the patients in the first 24 h after hospital admission. Similarly, we recorded each participant's type of ACEI, ARB, ARNI, beta-blocker, loop diuretic, and spironolactone. We labelled one if the patient used one of these four classes of drugs.

Quality of life assessment

The Minnesota Heart Failure Quality of Life Questionnaire (MHFLQ) was used to assess participants' quality of life. An independent investigator collected patient information, and physical, emotional, and other scores were calculated based on published criteria.¹³

Follow-up and outcome

Independent investigators conducted follow-up visits, and patients received outpatient or telephone follow-up visits every month after admission to collect relevant information until the end of the 3-month study period. During the follow-up period, patients were considered to have developed WHF if they met any of the following criteria: (1) re-hospitalization for more than 24 h due to HF exacerbation and (2) patients presenting to the emergency department with a precise diagnosis of HF were further evaluated to ensure their visits were indeed related to WHF. During the confirmation process, researchers inquired about specific CHF symptoms at WHF occurrence, such as lower extremity oedema, wheezing, and difficulty lying flat. The primary diagnosis recorded in the emergency department should include HF or acute exacerbation of CHF. Additionally, records were reviewed for elevated NT-proBNP levels and documentation of intravenous or increased oral diuretic therapy. Patients meeting at least two of these criteria were classified as experiencing WHF, (3) receiving intravenous diuretic treatment or an increase in diuretic use or dosage in an outpatient setting due to WHF signs and symptoms.⁸ Participants withdrew informed consent, and patients who could not be contacted or could not cooperate in completing follow-up were considered to be dislodged.

Sample size calculation

A total of 65 variables were included in this study, and we predicted that five to eight variables might be associated with the occurrence of WHF. According to previous literature, the readmission rate of HF patients discharged for 60–90 days is about 30%.¹⁴ Based on the calculation of Events per variable, the minimum sample size is 167 cases.¹⁵ Considering a 20% non-response rate, the final study cohort included 218 patients with CHF.

Statistical analysis

Missing value handling

Patients who were eventually able to complete follow-up were selected for pre-protocol set analysis (PPS). To ensure

data quality, standardized procedures were established before the study commenced. Before statistical analysis, efforts were made to minimize missing data by thoroughly reviewing the original medical records. After these adjustments, variables with more than 30% missing values were excluded. For the remaining variables, missing data patterns were visualized using the VIM package v6.2.2 in R (Figure S1). In addition, we constructed shading matrices to explore correlations between indicator variables and between indicator variables and observable variables (Figures S2 and S3). Based on the above steps and in conjunction with clinical experience, we assumed this study's missing data pattern was at least missing at random (MAR). Therefore, multiple interpolation was filled using mice package v 3.16.0 in R. Predictive mean matching, logistic regression, and polynomial regression were used for numeric, dichotomous, and multicategorical variables, respectively.

Variable selection and optimal modelling

All statistical analyses and visualizations were performed using SPSS (27.0), R (3.6.3), and Python (3.7.0), respectively. Normally distributed data were expressed as mean \pm standard deviation, non-normally distributed data as median and quartiles, and count data as percentages. In comparing the sample means of the two groups, the independent samples t-test was used for normally distributed data, the rank sum (Wilcoxon) test was used for non-normally distributed data, and the chi-square test was used for count data. $P < 0.05$ was considered statistically significant for all analyses.

Using Python (sklearn 0.22.1), the data was randomly divided into a training set and a test set at a ratio of 8:2. The training set is used to screen the variables and construct the categorical multivariate model. The test set is used only for the performance evaluation of the final model. To avoid overfitting and eliminate multicollinearity between variables, the least absolute shrinkage and selection operator (LASSO) regression analysis was performed on all variables in the training set using the R package (glmnet4.1.2). A 10-fold cross-validation of the model was performed to compute the minimum λ -value and to extract variables with non-zero coefficients. The screened variables were also statistically significant in univariate analysis ($P < 0.05$) and will be used for further categorical multimodel construction. Python (sklearn 0.22.1, XGBoost 1.2.1, LightGBM 3.2.1) was employed to construct a multimodel classification framework comprising eXtreme Gradient Boosting (XGBoost), Logistic Regression, Light Gradient Boosting Machine (LightGBM), Random Forest (RF), Adaptive Boosting (AdBoost), Multilayer Perceptron (MLP), Support Vector Machine (SVM), K-Nearest Neighbours (KNN), and Gaussian Naïve Bayes (GNB). The model is validated by a 10-fold cross-validation method, and the receiver operating characteristic curve (ROC) is plotted to compare

the modelling effectiveness of different algorithms and select the optimal model. The determination of the optimal model requires a combination of several metrics. ROC curves were constructed, and AUC was calculated using Python (sklearn 0.22.1) to assess the accuracy of the predictive model.¹⁶ Plot decision curve analysis (DCA) using R software (rmda 1.6) to assess the clinical applicability of the model.¹⁷ Calibration curves were plotted using Python (sklearn 0.22.1) to measure model calibration.¹⁸ Recall (PR) curves and calculated areas under the PR (AP) were plotted as a complement to the ROC curves using Python (sklearn 0.22.1).¹⁹

Once the optimal algorithm has been determined, the construction of the optimal model begins. The training set is used for model training, 10% of the data in the training set is randomly selected to form the validation set, and 10-fold cross-validation is performed for model tuning and evaluation. The test set data is only used to evaluate the final performance of the model. Considering the clinical utility of the model, more variables are challenging to understand and apply, so we intend to streamline the variables in the model further. Due to the sample size limitations, we used a more sensitive ML algorithm to screen the key variables. By comparing the variable importance rankings output by the three algorithms RF, XGBoost, and AdaBoost, we note that these algorithms have a high degree of consistency in evaluating the importance of variables. Therefore, we chose the variable that ranked high in importance among the three algorithms as the critical variable for the study. The key variables obtained were again subjected to categorical multimodel synthesis analysis and optimal modelling to compare the final performance of the complete and simplified versions of the model.

Subgroup analysis

A comprehensive subgroup analysis was conducted to verify the stability of the optimal model's predictive ability across various patient subgroups. These subgroups included age (whether ≥ 65 years), gender, NYHA classification (whether reaching IV), presence of hypertension, CKD, initial admission due to HF, type of HF, and LVEF (whether LVEF $> 35\%$).

Interpretation tools for the model

Since most ML models are black-box models, they are less interpretable, limiting their use in clinical settings. To address this, we interpreted the model using Python (SHAP 0.39.0), plotting significance and SHAP plots of contributions to the model. SHAP is a unified approach that accurately calculates the contribution and impact of each feature to the final prediction. The SHAP values show how each predictor positively or negatively influences the target variable. Each observation in the dataset can also be interpreted in terms of a specific set of SHAP values.²⁰ In addition, based on the

best-performing models, we created an online risk calculator that predicts the risk of developing WHF by inputting information about CHF patients. Finally, the models' corresponding sensitivity, specificity, positive predictive value, negative predictive value, and F1 score were also calculated.

Results

Study population

Through standardized data collection procedures and appropriate handling of missing data, no patients were excluded due to excessive missing information. Ultimately, only five duplicate cases and 13 patients with incomplete follow-up were excluded. A total of 200 patients with complete data were included in the statistical analysis (Figure 1). There are 160 cases in the training set and 40 cases in the test set. During the 3-month follow-up period, 60 patients (30%) developed WHF, and the characteristics of the two groups in the training set are shown in Table 1. The difference in baseline information between the training and test sets was not statistically significant (Table S1).

Filtering variables and multimodel classification construction

The LASSO regression results show that the minimum value of λ is 0.075, which at this point corresponds to 10 variables with non-zero coefficients. Variables included were whether statins were applied, emotional area, Hb, Cr, UA, NT-proBNP, NYHA cardiac function class, history of CKD, history of hypertension, and whether traditional Chinese medicine was applied (Figure 2). Table 1 shows that all 10 variables were statistically different in the univariate analysis. Therefore, we constructed categorical multivariate models for the above 10 variables. The ROC curve results showed that the RF algorithm had the highest model accuracy with an AUC of 0.801 (95% CI: 0.563–0.993) (Figure 3A,B). The RF also showed good accuracy in the calibration curve (Figure 3C). DCA showed that the RF also had good clinical applicability (Figure 3D). Similarly, RF showed the best AP values (Figure 3E,F). The combined analysis concludes that RF is the best model.

The best model building

These 10 variables were RF modelled in the training set with 10-fold cross-validation, and finally, the models were evaluated in the test set. The results showed an average AUC of 1.000 for the training set, 0.800 (95% CI: 0.544–0.997) for the validation set, and 0.822 (95% CI: 0.644–0.999) for the

test set (Figure 4A–C). The performance of the validation set is lower than that of the test set under the AUC metric, so the model fit can be considered successful. Figure 4D shows the changes in AUC during the RF model's training process. It can be seen that the AUC of the training set and the validation set is finally stabilized at around 0.8, and the model has a better prediction effect. The performance metrics of the model are shown in Table 2. These results show that the RF model can be used for dataset classification modelling tasks. To improve the efficiency of model application in clinical settings and to compare the modelling effects of the full model and the simplified model, we simplified the model variables by comparing the results of the variable importance ranking of the three ML algorithms. The results showed that the five variables, NT-proBNP, Cr, UA, Hb, and Emotional area, were significant in all three ML algorithms (Figure 5A–C). The results of the categorical multiple modelling show that the RF algorithm has the highest model accuracy, with a validation set AUC of 0.785 (95% CI: 0.533–0.981) (Figure 5D,E). A simplified version of the RF model containing five variables was again modelled, validated, and tested. The results showed an average AUC of 1.000 for the training set, 0.797 (95% CI: 0.549–0.993) for the validation set, and 0.842 (95% CI: 0.675–1.000) for the test set (Figure 6A–C). Figure 6D shows the changes of AUC during the training process of the RF model, and it can be seen that the AUC of the training set and the validation set is finally stabilized at around 0.8, and the model has a good prediction effect. The performance metrics of the model are shown in Table 2. Considering that the performance of the streamlined version of the model is not inferior to the full version and significantly improves the efficiency of clinical use. Therefore, we considered the streamlined version of the model as the final model and implemented further model interpretation and predictor development.

Subgroup analysis

The results of the subgroup analyses are shown in Table S2. Due to sample size limitations, we cannot assess the performance of the best model in the combined CKD subgroup. Summarizing the model's performance across subgroups, we find that the final model fits well across populations (all AUCs >0.6). Specifically, the model demonstrated higher predictive accuracy for patients without hypertension, with NYHA classification II–III, HFrEF and HFpEF. However, in the subgroups without CKD, first hospitalization due to HF, and HFmrEF, while the model exhibited high sensitivity, its specificity and positive predictive value were low. Additionally, for the subgroup with LVEF $\leq 35\%$, both accuracy and sensitivity were also low. This indicates that the results of this model must be interpreted with caution when predicting the risk of WHF in these populations.

Table 1 Characteristics of participants in the training set grouped according to WHF occurrence

Characteristics	Total (N = 160), N (%)	Non-WHF (N = 110), N (%)	WHF (N = 50), N (%)	P-value
Gender				0.436
Male, n (%)	103 (64.38)	73 (66.36)	30 (60.00)	
Female, n (%)	57 (35.63)	37 (33.64)	20 (40.00)	
Age, years, median (IQR)	72.00 [63.00, 78.00]	72.00 [63.00, 78.00]	70.00 [60.00, 80.00]	0.761
Vital signs				
Heart rate, b.p.m., median (IQR)	76.00 [70.00, 83.00]	74.00 [70.00, 84.00]	77.00 [72.00, 83.00]	0.507
Systolic blood pressure, mmHg, mean (SD)	127.93 ± 22.27	129.90 ± 21.82	123.60 ± 22.65	0.098
Diastolic blood pressure, mmHg, mean (SD)	76.39 ± 13.22	76.89 ± 13.54	75.30 ± 12.44	0.484
BMI, kg/m ² , median (IQR)	24.68 [22.03, 27.74]	24.78 [22.04, 27.89]	24.16 [21.50, 27.68]	0.850
Educational attainment				0.050
Primary and below, n (%)	12 (7.50)	6 (5.45)	6 (12.00)	
Junior middle school, n (%)	44 (27.50)	37 (33.64)	7 (14.00)	
Senior high school, n (%)	65 (40.63)	41 (37.27)	24 (48.00)	
University degree or above, n (%)	39 (24.38)	26 (23.64)	13 (26.00)	
History of smoking, n (%)	63 (39.38)	49 (44.55)	14 (28.00)	0.047
History of drinking, n (%)	34 (21.25)	27 (24.55)	7 (14.00)	0.131
Exercise frequency				0.124
Hardly any exercise, n (%)	67 (41.88)	42 (38.18)	25 (50.00)	
Get some exercise, n (%)	81 (50.63)	57 (51.82)	24 (48.00)	
Regular exercise, n (%)	12 (7.50)	11 (10.00)	1 (2.00)	
First admission for HF				0.101
Yes	77 (48.13)	60 (52.17)	17 (37.78)	
No	83 (51.88)	55 (47.83)	28 (62.22)	
Co-morbidities				
Coronary artery disease, n (%)	68 (42.50)	47 (42.73)	21 (42.00)	0.931
Hypertension, n (%)	58 (36.25)	30 (27.27)	28 (56.00)	<0.001
Diabetes, n (%)	96 (60.00)	61 (55.45)	35 (70.00)	0.082
Chronic kidney disease, n (%)	25 (15.63)	10 (9.09)	15 (30.00)	<0.001
Stroke, n (%)	19 (11.88)	11 (10.00)	8 (16.00)	0.277
Atrial fibrillation, n (%)	42 (26.25)	28 (25.45)	14 (28.00)	0.734
Hyperlipidaemia, n (%)	52 (32.50)	31 (28.18)	21 (42.00)	0.084
Treatment				
Angiotensin-converting enzyme inhibitors/angiotensin-receptor blockers/angiotensin receptor enkephalinase inhibitors, n (%)	72 (45.00)	45 (40.91)	27 (54.00)	0.123
Beta-blocker, n (%)	92 (57.50)	59 (53.64)	33 (66.00)	0.143
Loop diuretics, n (%)	112 (70.00)	76 (69.09)	36 (72.00)	0.710
Spironolactone, n (%)	85 (53.13)	59 (53.64)	26 (52.00)	0.848
Digoxin, n (%)	13 (8.13)	6 (5.45)	7 (14.00)	0.067
Aspirin/clopidogrel, n (%)	43 (26.88)	28 (25.45)	15 (30.00)	0.548
Statin, n (%)	17 (10.63)	7 (6.36)	10 (20.00)	0.009
Nitrate drugs, n (%)	14 (8.75)	11 (10.00)	3 (6.00)	0.407
Traditional Chinese medicine, n (%)	83 (51.88)	65 (59.09)	18 (36.00)	0.007
Drug type				0.022
1, n (%)	33 (20.63)	23 (20.91)	10 (20.00)	
2, n (%)	59 (36.88)	46 (41.82)	13 (26.00)	
3, n (%)	58 (36.25)	38 (34.55)	20 (40.00)	
4, n (%)	10 (6.25)	3 (2.73)	7 (14.00)	
Types of heart failure				0.690
HFrEF, n (%)	62 (38.75)	44 (40.00)	18 (36.00)	
HFmrEF, n (%)	35 (21.88)	22 (20.00)	13 (26.00)	
HFpEF, n (%)	63 (39.38)	44 (40.00)	19 (38.00)	
Functional status by New York Heart Association level				0.010
II, n (%)	23 (14.37)	19 (17.27)	4 (8.00)	
III, n (%)	94 (58.75)	69 (62.73)	25 (50.00)	
IV, n (%)	43 (26.88)	22 (20.00)	21 (42.00)	
Echocardiography				
LVEF, %, median (IQR)	44.00 [36.00, 57.00]	44.00 [37.00, 58.00]	44.00 [29.00, 56.00]	0.471
LVDd, mm, mean (SD)	57.71 ± 8.81	56.90 ± 8.39	59.50 ± 9.42	0.084
LVDs, mm, mean (SD)	43.31 ± 10.33	43.12 ± 10.21	43.74 ± 10.56	0.726
LVFS, %, median (IQR)	24.00 [18.00, 32.00]	23.00 [18.00, 32.00]	27.00 [21.00, 32.00]	0.276

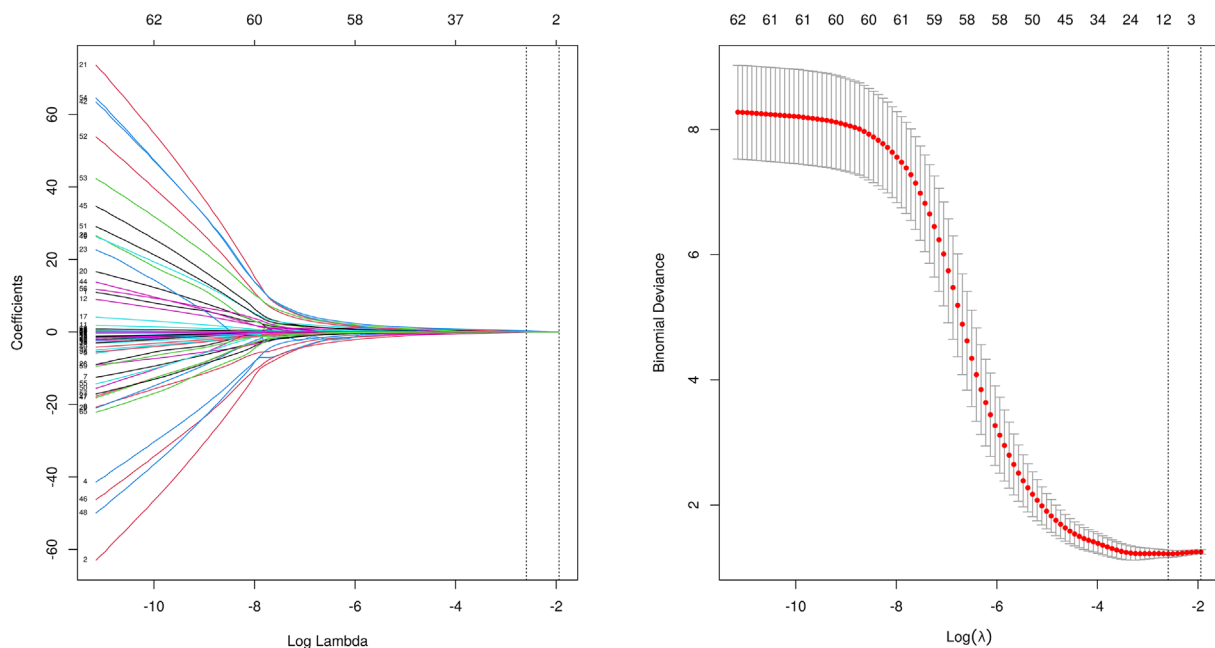
(Continues)

Table 1 (continued)

Characteristics	Total (N = 160), N (%)	Non-WHF (N = 110), N (%)	WHF (N = 50), N (%)	P-value
Laboratory test				
NT-proBNP, pg/mL, median (IQR)	2009.00 [812.00, 4117.00]	1528.00 [639.00, 2780.00]	3612.00 [1876.00, 8121.00]	<0.001
UA, $\mu\text{mol/L}$, median (IQR)	448.00 [365.00, 557.00]	419.00 [354.00, 512.00]	538.00 [425.00, 615.00]	0.001
Cr, $\mu\text{mol/L}$, median (IQR)	88.70 [73.70, 123.70]	83.80 [72.30, 113.50]	114.70 [74.50, 168.00]	0.007
BUN, mmol/L, median (IQR)	7.86 [6.16, 9.97]	7.55 [5.92, 9.27]	9.48 [6.75, 11.26]	0.007
TBIL, $\mu\text{mol/L}$, median (IQR)	15.04 [9.41, 21.66]	13.37 [8.59, 21.26]	16.32 [12.63, 22.24]	0.043
Hb, g/L, mean (SD)	121.56 \pm 24.04	125.86 \pm 23.60	112.10 \pm 22.21	<0.001
HCT, %, median (IQR)	0.36 [0.32, 0.42]	0.37 [0.31, 0.43]	0.36 [0.32, 0.38]	0.046
RBC, $\times 10^9/\text{L}$, median (IQR)	3.95 [3.53, 4.49]	3.95 [3.51, 4.61]	3.92 [3.56, 4.26]	0.310
Quality of life assessment				
Physical area, score, median (IQR)	22.00 [16.00, 28.00]	20.00 [12.00, 28.00]	25.00 [19.00, 29.00]	0.014
Emotional area, score, median (IQR)	5.00 [1.00, 8.00]	4.00 [0.00, 7.00]	6.00 [4.00, 8.00]	0.005
Other areas, score, median (IQR)	22.00 [13.00, 27.00]	21.00 [13.00, 27.00]	22.00 [13.00, 30.00]	0.491
MLHFQ total score, score, mean (SD)	47.13 \pm 19.87	44.87 \pm 19.22	52.10 \pm 20.35	0.033
LVEF, %				
HFrEF, median (IQR)	33.00 [27.00, 38.00], (n = 61)	33.00 [29.00, 38.00], (n = 47)	29.00 [23.00, 38.00], (n = 14)	0.239
HFmrEF, median (IQR)	44.00 [42.00, 46.00], (n = 32)	43.00 [42.00, 46.00], (n = 18)	44.00 [43.00, 48.00], (n = 14)	0.112
HFpEF, mean (SD)	59.35 \pm 9.74, (n = 67)	58.73 \pm 10.03, (n = 52)	61.00 \pm 8.45, (n = 15)	0.464

BMI, body mass index; BUN, blood urea nitrogen; Cr, creatinine; Hb, haemoglobin; HCT, haematocrit; HFmrEF, heart failure with mildly reduced ejection fraction; HFpEF, heart failure with preserved ejection fraction; HFrEF, heart failure with reduced ejection fraction; LVDd, left ventricular end-diastolic internal diameter; LVDs, left ventricular end-systolic internal diameter; LVEF, left ventricular ejection fraction; LVFS, left ventricular fractional shortening; MLHFQ, Minnesota Heart Failure Quality of Life Questionnaire; NTproBNP, N-terminal brain natriuretic peptide precursor; RBC, red blood cell count; TBIL, total bilirubin; UA, uric acid; WHF, worsening heart failure.

Figure 2 LASSO regression analysis was used for preliminary variable screening.



Interpretation of models

To visually explain the selected variables, we use SHAP to illustrate how the variables in these models predict the occur-

rence of WHF. Figure 7A shows the order of importance of the five variables assessed in terms of mean absolute SHAP values, with the SHAP values on the x-axis indicating the significance of the predictive model, and it can be seen that

Figure 3 Model accuracy for categorical multimodel evaluation. (A) ROC curves and AUC of the training group. (B) ROC curves and AUC of the validation group. (C) Calibration curves for the validation set, where the horizontal coordinate is the average predicted probability, the case coordinate is the actual probability of the event, the diagonal dashed line is the reference line, and the other smooth solid lines are the different model fit lines. The closer the fitted line is to the reference line, the smaller the values in parentheses are and the more accurate the model predictions are. (D) Validation set DCA, where the black dashed line represents the hypothesis that all patients have WHF, and the red dashed line and the thin black line represent the hypothesis that no patients have WHF. The remaining solid lines represent different models. (E) PR curves and AP for the training set and (F) PR curves and AP for the validation set. The y-axis represents precision, and the x-axis represents recall. Suppose the precision curve of another model completely covers the precision curve of one model. In that case, the latter can be considered better than the former, and the higher the AP value, the better the model performance. Different colours in the graph represent the corresponding models.

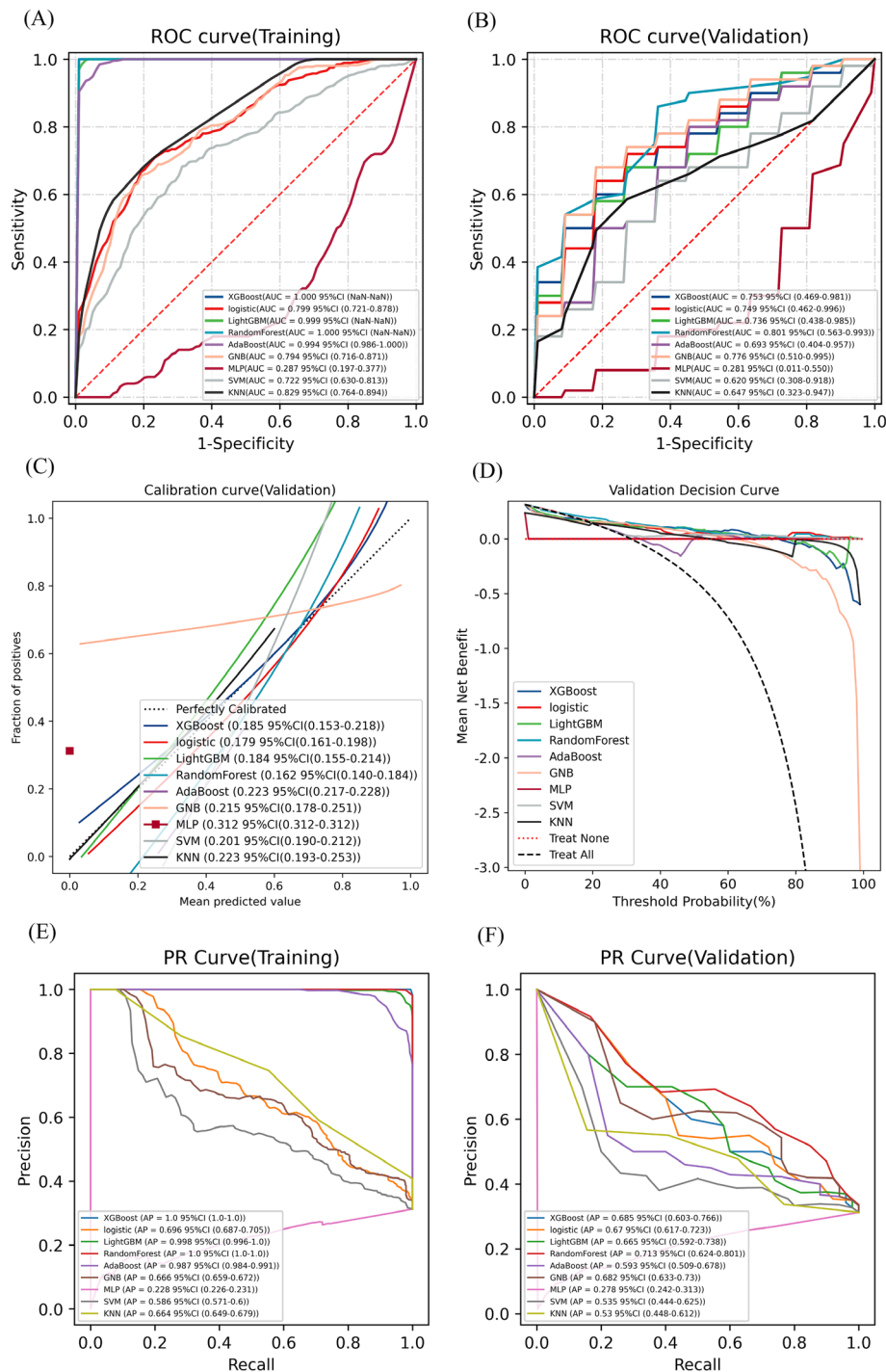
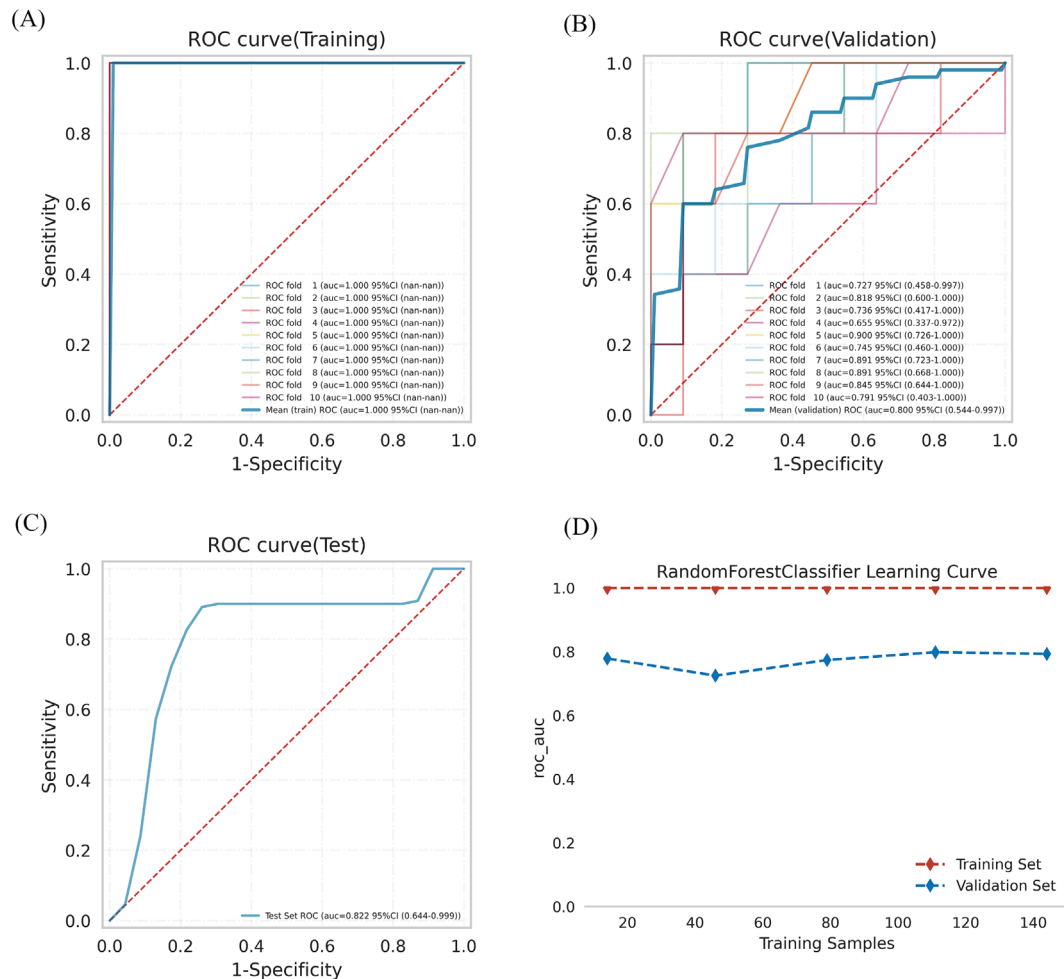


Figure 4 Full version RF model training, validation, and testing. (A) Training set of ROC curves and AUC. (B) Validation set of ROC curves and AUC. (C) Testing set of ROC curves and AUC. (D) Learning curve. The red dashed line represents the training set, and the blue dashed line represents the validation set.



NT-proBNP, UA and Cr were the more important predictors of WHF in patients. In the summary plot of SHAP values in *Figure 7B*, each feature is represented along the y-axis, and the corresponding SHAP values are represented along the x-axis. These SHAP values indicate how much each feature affects the average predicted value of the model, with blue to red representing feature values from low to high. It can be seen that Hb was negatively associated with the development of WHF, whereas NT-proBNP, Cr, emotional area, and UA were positively associated with the development of WHF. We also visualized individual predictions, which are shown in *Figure 7C* for non-WHF (NWHF) patients and in *Figure 7D* for WHF patients. Bold numbers are probabilistic predictions ($f(x)$) and baseline values are predictions not entered into the model. $f(x)$ is the log ratio for each observation. Red features indicate that they make the risk of WHF increased and blue features indicate that the risk of occurrence is reduced. The length of the arrows helps to visualize

the extent to which predictions are affected, with longer arrows having a greater impact on the eventual occurrence of events.²¹

Application of predictive models

Although SHAP values provide a reliable explanation for ML models, it is still challenging to derive insights for practical clinical applications directly from SHAP values. To enable clinicians to translate our findings into practical applications, we have developed a web-based version of an online risk calculator based on a simplified version of the RF model, with detailed instructions and precautions for use (www.xsmartanalysis.com/model/list/predict/model/html?mid=11456&symbol=7170qN6FAcA32640kr67) (*Figure 8*). When applied to the clinic, the model predicts the risk of developing WHF in patients with CHF by inputting five identified

Table 2 Predictive performance of RF model in different datasets

Model	Groups	AUC (95% CI)	Cut-off (95% CI)	Accuracy (95% CI)
Full version of the RF model	Training set	1.000 (Na-Na)	0.609 (0.592–0.626)	0.992 (0.989–0.994)
	Validation set	0.800 (0.544–0.997)	0.609 (0.592–0.626)	0.738 (0.697–0.778)
	Testing set	0.822 (0.644–0.999)	0.61	0.8
Streamlined version of the RF model	Training set	1.000 (Na-Na)	0.605 (0.590–0.620)	0.991 (0.988–0.994)
	Validation set	0.797 (0.549–0.993)	0.605 (0.590–0.620)	0.763 (0.722–0.803)
	Testing set	0.842 (0.675–1.000)	0.59	0.775

AUC, area under the curve; RF, random forest.

Table 2 (continued)

Model	Sensitivity (95% CI)	Specificity (95% CI)	Positive predictive value (95% CI)	Negative predictive value (95% CI)	F1 score (95% CI)
Full version of the RF model	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	0.988 (0.984–0.992)	1.000 (1.000–1.000)
	0.800 (0.683–0.917)	0.800 (0.705–0.895)	Na (Na-Na)	0.742 (0.711–0.773)	Na (Na-Na)
	0.9	0.8	0.75	0.806	0.818
Streamlined version of the RF model	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	0.987 (0.983–0.991)	1.000 (1.000–1.000)
	0.860 (0.758–0.962)	0.755 (0.619–0.890)	Na (Na-Na)	0.768 (0.738–0.797)	Na (Na-Na)
	0.9	0.833	0.6	0.8	0.72

AUC, area under the curve; RF, random forest.

variables. For example, if a patient with CHF has an NTpro-BNP of 21 937 pg/mL, a UA of 452 mmol/L, a Cr of 106.9 mmol/L, a Hb of 72 g/L, and a computed MHFLQ-Emotional area score of 0 when first tested after admission, the model would yield a probability of the patient developing WHF in the next 3 months of 59.00%.

In addition to the web calculator described above, cut-off values, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score were calculated for both models (Table 2). Combining the evaluation metrics of the models, the simplified model has a slightly higher classification ability and prediction strength while maintaining a high sensitivity but is weakened in certain other aspects (accuracy and positive predictive value). This suggests that a simplified version of the model is a more efficient option, especially when there is a need to balance model complexity with practical clinical application scenarios.

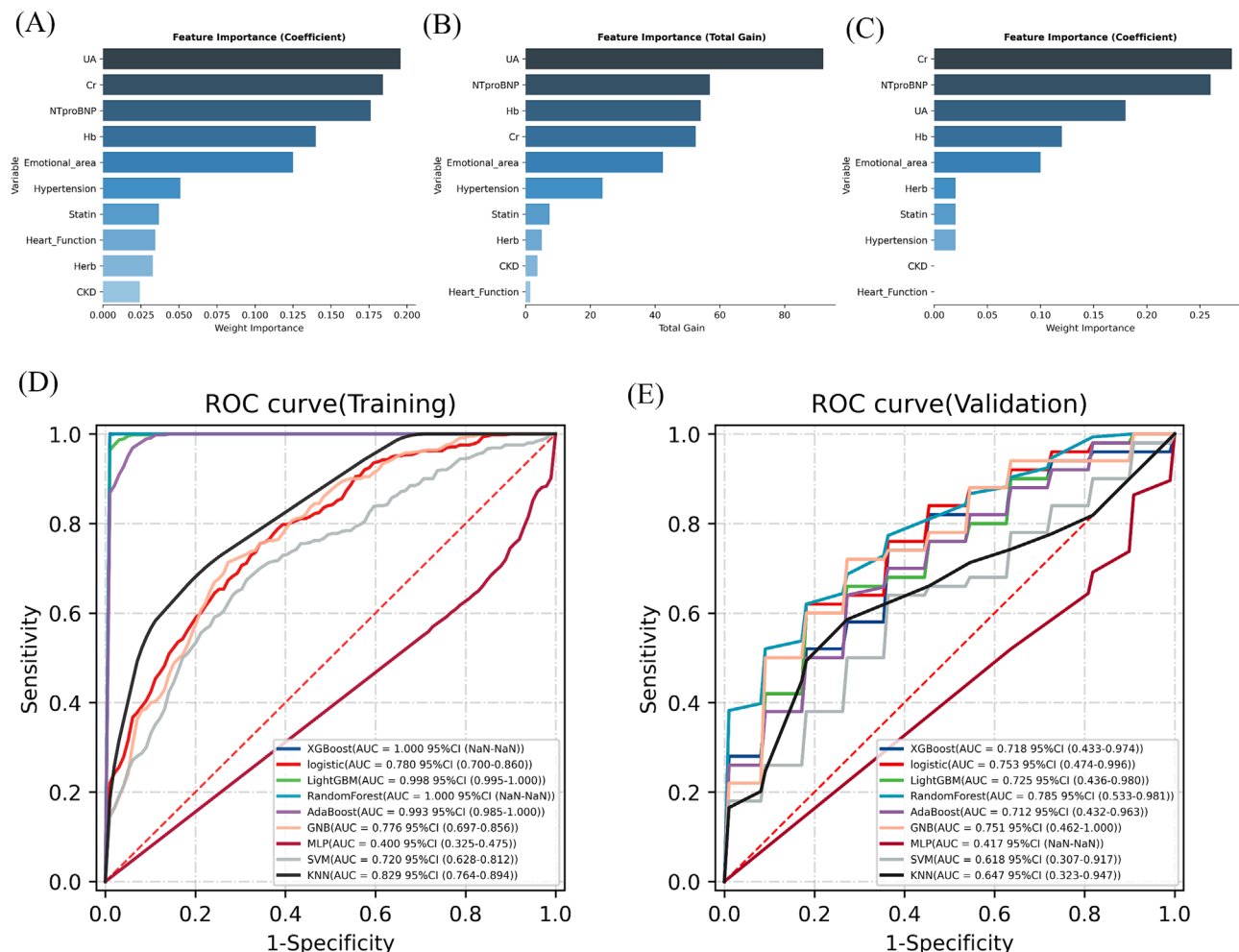
Discussion

WHF is considered a specific stage in the course of HF, and WHF signals a poorer prognosis.^{2,3} Recent guidelines also emphasize the clinical importance of WHF.^{11,22} The study by Butler *et al.*²³ included 11 064 patients with HF with reduced ejection fraction (HFrEF), 17% of whom developed WHF 1.5 years after diagnosis, and the 2-year mortality rate for patients who developed WHF was 22.5%. Of note, the study defined WHF as stable HF for ≥ 90 days, with subsequent deterioration requiring hospitalization. However, studies have

shown that non-admitted patients, who account for 50% of total WHF events, have as poor a prognosis as hospitalized patients.¹⁰ This suggests that the risk of WHF is seriously underestimated. In a recent review, researchers noted the need to recognize WHF as a marker of a new stage in the development of HF.¹ Especially in the context of studies showing that the use of invasive pulmonary artery pressure home monitoring technology can identify changes in the condition of CHF patients before clinical symptoms worsen, potentially leading to a 48% reduction in hospitalization rates, this further confirms the importance of timely detection and intervention of preclinical exacerbations in reducing hospital admissions.²⁴ These findings reinforce the need to develop predictive models of WHF for early identification of high-risk CHF patients and effective risk stratification.

In our study, the prevalence of WHF in CHF patients within 3 months was 30%, which is consistent with the results of epidemiologic surveys.¹⁰ This study utilizes the most recent definition of WHF, which makes our work significantly different from existing studies. Previous studies have preferred to use HF or all-cause readmissions as predictors of outcome (Table 3). As mentioned earlier, the occurrence of WHF events in non-hospitalized patients makes the targeted development of WHF prediction models of practical relevance. In addition, we note that existing studies prefer to select the most accurate model by comparing multiple ML algorithms, which often contain a large number of variables, making it difficult to generalize their results widely to clinical applications. More critically, these studies need more interpretability provided for ML models. Due to the black-box nature of ML,

Figure 5 Filtering variables and multimodel classification construction. (A) Top 10 order of importance of variables in RF models. (B) Top 10 order of importance of variables in XGBoost models. (C) Top 10 order of importance of variables in AdaBoost models. (D) ROC curves and AUC of the training group. (E) ROC curves and AUC of the validation group. NTproBNP: N-terminal brain natriuretic peptide precursor. Cr: creatinine. TBIL: total bilirubin. UA: uric acid. Hb: haemoglobin. CKD: chronic kidney disease.

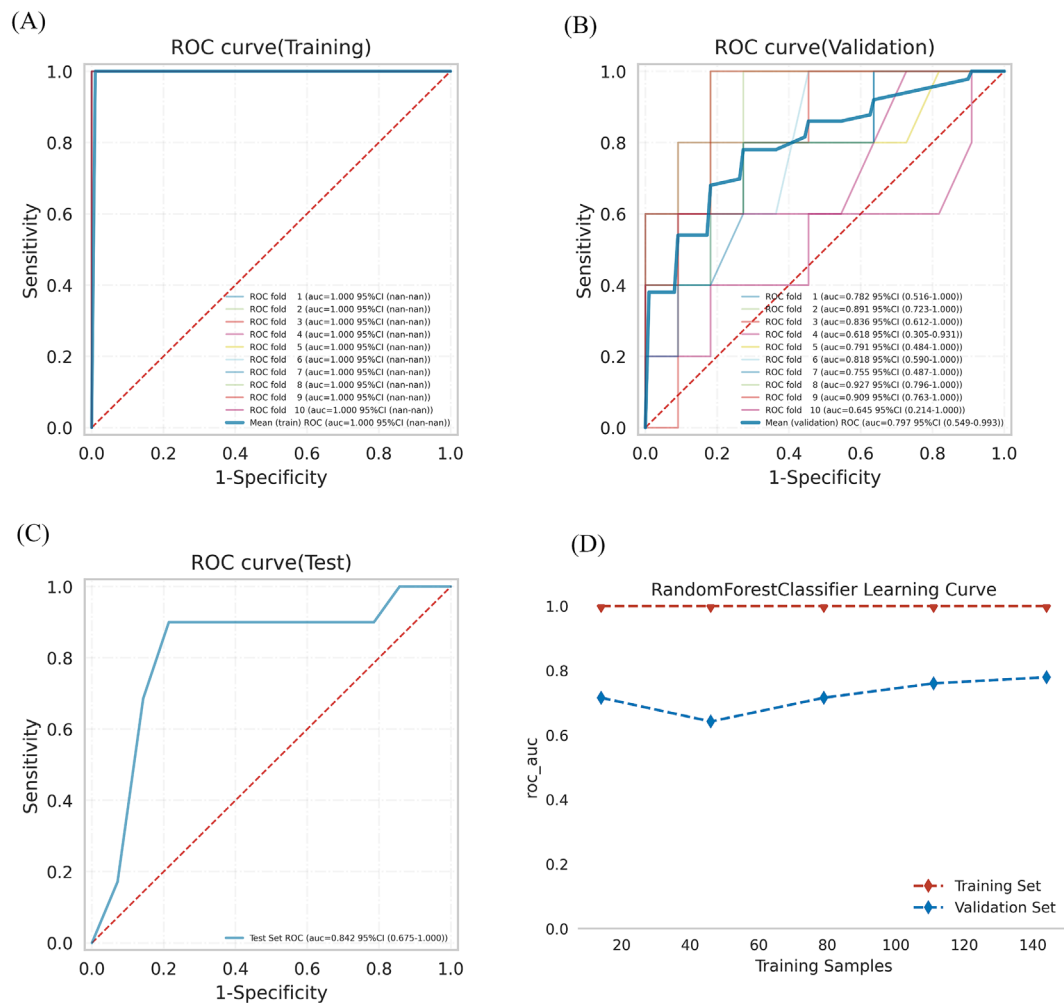


this has been a significant barrier to translating its results into clinical applications. In this case, we provide an interpretable ML model that contains only five clinically common indexes. The generated online risk calculator also provides a convenient way for clinical applications, a significant improvement over the past. Similarly to our study, Parikh et al.⁸ developed a predictive model based on the gradient boosted decision tree model in 338 426 adult HF patients to predict the incidence of WHF in patients with different HF types. They validated the excellent performance of the model. The results showed that brain-type natriuretic peptide (BNP), the number of occurrences of previous lower extremity edema and pulmonary rales, and diuretic use were the most vital predictors associated with the development of WHF. However, the study was based on a retrospective collection of electronic health record data from a medical facility in the Northern California region of the United States. The applicability of

the model results to the Chinese population needs to be verified, and inevitably, some indicators, such as patient-reported outcome metrics (PROMs), will be missed. This study is the first predictive model of WHF in the Chinese population. Prospective data collection ensured the accuracy and validity of the data, and we comprehensively compared the modelling effectiveness of multiple ML algorithms and built a quantifiable risk calculator. This is an improvement over previous studies.

Due to the limited size of the study and to minimize the risk of overfitting, we screened the variables by LASSO regression, multistage ML algorithmic screening, and 10-fold cross-validation. Through LASSO regression and univariate analysis, we initially developed an RF model with 10 variables that performed well. In order to improve the efficiency of the clinical use of this model, we achieved the streamlining of the variables through the ML algorithm, which is because the ML

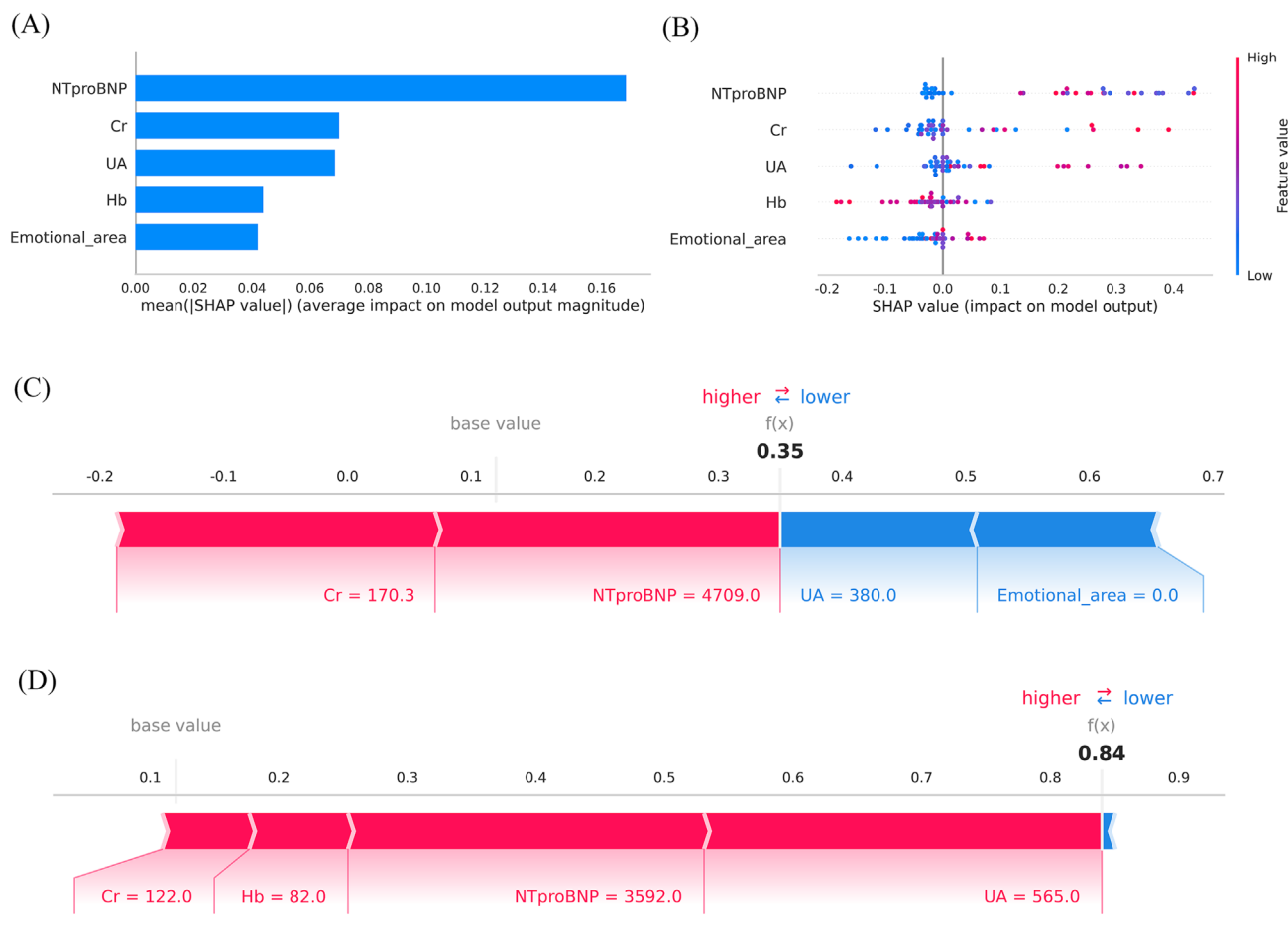
Figure 6 Streamlined RF model training, validation, and testing. (A) Training set of ROC curves and AUC. (B) Validation set of ROC curves and AUC. (C) Testing set of ROC curves and AUC. (D) Learning curve.



algorithm does not have strict requirements on the sample size, thus maximizing the correctness of the results. We finalized five key variables: NT-proBNP, Cr, UA, Hb, and emotional area and used them to construct the RF model with the best performance. Evaluating the model by metrics such as F1 scores, we find that the simplified version performs better and that continuous variables play a more prominent role in the model than other categorical variables. This may be because continuous variables are more conducive to the model's sensitive detection of changes in the condition. In subgroup analyses, the final model demonstrated good generalizability across different populations. Specifically, the model exhibited higher predictive accuracy for patients without hypertension, within NYHA classes II-III, HFmrEF and HFpEF. However, although the model's sensitivity was high in the subgroup of patients without CKD, first hospitalization due to HF and HFmrEF, its specificity and positive predictive value were poorer. This indicates the model may generate more

false positive results when predicting WHF events in these subgroups. This may be related to the physiological mechanisms closely associated with CKD and heart conditions, which are not as pronounced in patients without CKD, affecting the model's predictive performance. For patients hospitalized for the first time due to HF, the potential pathophysiological mechanisms may not have fully manifested due to the short duration of the disease. HFmrEF, as an intermediate state, typically exhibits characteristics of both HFmrEF and HFpEF, and may transition into other categories as the disease progresses.²⁵ This complexity and diversity in pathophysiology increase the difficulty of model prediction and may contribute to the poor predictive performance observed. For patients with LVEF $\leq 35\%$, the presence of more severe structural cardiac changes and haemodynamic abnormalities, combined with the rapid and severe nature of their condition, may render conventional parameters insufficient to timely reflect changes in their condition. In addition, we

Figure 7 Interpretation of the model using SHAP. (A) Importance ranking of features displayed by SHAP. (B) Characterization attributes in SHAP. (C) A patient with NWHF and (D) a patient with WHF.



recognize that these results need to be considered and addressed in subsequent studies due to the lack of external validation and the relatively small study size.

NT-proBNP, UA, and Cr in the model were the three most significant variables in predicting the occurrence of WHF. This is consistent with the findings of Parikh et al.⁸ that BNP, Cr, and Hb are all significant predictors of WHF occurrence in patients with HF, with BNP being the most important predictor. Since NT-proBNP testing is more common in our medical units, this still proves the value of NT-proBNP in predicting the occurrence of WHF. Considering that the current definition of WHF needs to be supplemented with additional biomarkers, it is reasonable to believe that NT-proBNP or BNP levels are expected to be the basis for the diagnosis of WHF.¹ Elevated UA is common in patients with HF due to long-term diuretic use and co-morbid renal dysfunction. The results of the meta-analysis showed that UA levels were positively associated with the occurrence of HF and the risk of a composite endpoint event in patients with HF.²⁶ A recent study by Nishino et al.²⁷ demonstrated that UA is an independent risk factor for readmission in patients with ejection

fraction-preserved HF (HFpEF) and that UA-lowering therapy is associated with a reduced risk of death. Many models have shown that PROMs (MHFLQ, Kansas City Cardiomyopathy Questionnaire) are important predictors of death and readmission in patients with HF.^{28,29} The non-invasive nature of PROMs indicators gives them a unique predictive advantage. Our findings suggest that focusing on the domain of emotion is more important in predicting the future occurrence of WHF than other domains in quality of life indicators. It also means that the long-term management of patients with HF requires multidisciplinary interventions, including psychiatrists. Most previous HF predictive modelling studies have addressed the prediction of readmission and death. A systematic review of 117 predictive models for HF showed that BUN, blood sodium level, and race had the highest predictive value for readmission.³⁰ This is different from our findings. Consider that the definition of WHF used in this study includes worsening of symptoms, signs, and increased outpatient diuretics in addition to readmission for HF. This may be partly responsible for the discrepancy. Study demonstrated that the predictive value of echocardiographic parameters (LVEF, LVDd,

Figure 8 Web-based WHF risk calculator.



LVDs, and LVFS) is not significant, although similar studies have reached comparable conclusions. Additionally, our study indicated that the predictive value of echocardiographic parameters (LVEF, LVDd, LVDs, and LVFS) is not significant, consistent with findings from similar studies.^{8,32,33} It should be noted that during the initial phase of this study, there was limited awareness of HFpEF in China. Consequently, parameters of diastolic function were not routinely reported in our medical unit early in the study. The high rate of missing data led to the exclusion of these variables, which may have limited the evaluation of their predictive value for WHF. The insignificance of systolic function indices such as LVEF and LVDd could also be attributed to the inclusion of a substantial number of HFpEF patients in our study. This underscores the need for future research to focus on the comprehensive collection of echocardiographic parameters in patients with different

HF phenotypes and to investigate the relationship between these parameters and the occurrence of WHF. Interestingly, no significant differences in WHF were observed among the three subtypes, which may be attributed to the limited sample size and short follow-up period. Additionally, due to differences in treatment approaches and disease management, HFrEF patients typically receive more aggressive and comprehensive care. This proactive management may reduce the incidence of WHF in HFrEF patients. In contrast, the lack of treatment options for HFmEF may explain why the incidence of WHF in HFmEF patients is similar to that in HFrEF patients. Finally, although HFmrEF is generally considered a relatively stable phase of the disease, our study showed that HFmrEF patients who were hospitalized still had a high short-term risk of WHF, indicating the need for enhanced management of HFmrEF patients.

Table 3 Comparing of this study with previous studies

Author	Year	Design	No. patients	ML model used	Outcome	Verification	Model contains variables	Model interpretation	Performance metrics reported
Parikh ⁸	2023	Retrospective	338 426	Gradient Boosted Decision Tree Models	WHF	Internal	BNP, number of occurrences of lower limb oedema, number of occurrences of rales, use of diuretics, number of past heart failure events as important predictors	SHAP	AUC: 0.764 MSE: 0.127
Ru ³¹	2023	Retrospective	30 687	XGBoost	WHF	Internal	Demographic and clinical data	NA	AUC: 0.64 AUPR: 0.489 Precision: 0.542 Recall: 0.536 AUC: 0.76 Brier score: 0.19
Angraal ²⁹	2020	Retrospective	1767	RF	HF readmission	Internal	Hb, BUN, previous HF hospitalization duration, and KCCQ scores are the most important predictors	NA	AUC: 0.73 ACC: 0.81 SPE: 0.85 SEN: 0.50 AUC: 0.60 ACC: 0.63 SPE: 0.66 SEN: 0.53
Sabouri ³²	2023	Prospective	737	SVM	30-day readmission	Internal	NYHA functional classification, baseline Cr, UA, inferior vena cava size, worsening of renal function	NA	
Sabouri ³²	2023	Prospective	737	KNN	90-day readmission	Internal	Base Cr, oedema, diabetes, tricuspid regurgitation, ICD/CRT, inotropes, atrial fibrillation, smoking, Hb, wide QRS	NA	
Awan ³³	2019	Retrospective	10 757	MLP	30 days readmission or death	Internal	Demographics, admission characteristics, medical history, co-morbidities, and so on for a total of 47 variables	NA	AUC: 0.63 AUPR: 0.46 ACC: 64.93 SEN: 48.42 SPE: 70.01 AUC: 0.84 ACC: 0.78 SEN: 0.90 SPE: 0.83 NPV: 0.80 PPV: 0.60
This study	-	Retrospective	200	RF	WHF	Internal	NT-proBNP, Cr, UA, Hb, and the emotional area from the MLHFQ	SHAP and online risk predictor	

ACC, accuracy; AUC, area under the curve; AUPR, area under the precision-recall curve; BNP, B-type natriuretic peptide; BUN, blood urea nitrogen; Cr, creatinine; Hb, haemoglobin; HF, heart failure; ICD/CRT, implantable cardioverter-defibrillator/cardiac resynchronization therapy; KCCQ, Kansas City Cardiomyopathy Questionnaire; KNN, K-nearest neighbours; MCC, Matthews correlation coefficient; ML, machine learning; MLHFQ, Minnesota Living with Heart Failure Questionnaire; MLP, multilayer perceptron; MSE, mean squared error; NPV, negative predictive value; NT-proBNP, N-terminal pro B-type natriuretic peptide; NYHA, New York Heart Association; PPV, positive predictive value; RF, random forest; SEN, sensitivity; SHAP, SHapley Additive exPlanations; SPE, specificity; SVM, support vector machine; UA, uric acid; WHF, worsening heart failure; XGBoost, eXtreme Gradient Boosting.

In summary, we conducted a study that culminated in constructing an RF model to predict the occurrence of WHF by screening variables and comparing multiple modelling approaches. The model includes five important variables commonly found in clinical practice: NT-proBNP, Cr, UA, Hb, and MLHFQ-emotional area. Since machine learning models are poorly interpretable, SHAP improves the interpretation of the results. The results suggest that the predictive value of the variables NT-proBNP, UA, and Cr may be higher. The online computational tool developed allows clinicians to easily apply the results of this study to their clinical practice. It is expected to improve the scalability and cost-effectiveness of the healthcare system.

Limitation

This study still has some limitations. First, although the study's target population was Chinese WHF patients, the source of the cases was a single centre, which may cause bias and reduce the extrapolation of the findings. Additionally, the cohort size of 200 patients is relatively small. The limited sample size hindered an in-depth analysis of the rarer HF phenotypes (such as hypertrophic and dilated cardiomyopathy) during the model training process, which has consequently restricted the model's applicability. Future research should aim to increase the sample size and incorporate a broader spectrum of HF phenotypes to enhance the model's accuracy and clinical utility. Finally, in our study, no detailed information was collected on patients' signs of congestion, such as previously presented lower extremity oedema and lung rales. These factors were important in predicting the occurrence of WHF in the study by Parikh *et al.*⁸ and may further improve the model's predictions in this study. Therefore, these important variables should be taken into account in future studies. Based on this study, future research should involve multicentre collaboration with different institutions to ensure data diversity and model generalizability, thereby further validating and optimizing our model. Additionally, extending the follow-up period, optimizing the study design, and employing advanced models such as long short-term memory (LSTM) networks to better capture patients' dynamic changes will enhance the accuracy and applicability of predictions, representing a key direction for future research.

Conclusions

In summary, by comparing the performance of various ML algorithms in predicting the occurrence of WHF, we ultimately developed an RF model. The interpretation of the model

through SHAP highlighted the significance of NT-proBNP, Cr, UA, Hb, and emotional area, culminating in creating an online risk calculation tool. This provides clinicians with early identification of patients at high risk for CHF, improves medical decision-making, and provides personalized treatment.

Acknowledgements

Thanks to <https://www.xsmartanalysis.com/model/index/> for technical support.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

This study was funded by grants from the National Key Research and Development Program of China (2019YFC1708703), Natural Science Foundation of Beijing Municipality (7232324), and Central High-Level Traditional Chinese Medicine Hospital Project of Eye Hospital China Academy of Chinese Medical Science (GSP2-02).

Data Availability Statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding authors.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1. Characteristics of participants in the training and testing sets.

Table S2. Performance of the final model in different subgroups of the test set.

Figure S1. Visualization of missing data patterns. (A) Number of missing variables. (B) Patterns of missing values for variables.

Figure S2. Heat map of correlation between indicator variables.

Figure S3. Heat map of correlation between indicator and observable variables.

References

- Greene SJ, Bauersachs J, Brugs JJ, Ezekowitz JA, Lam CSP, Lund LH, et al. Worsening heart failure: nomenclature, epidemiology, and future directions: JACC review topic of the week. *J Am Coll Cardiol* 2023;**81**:413-424. doi:10.1016/j.jacc.2022.11.023
- Dharmarajan K, Hsieh AF, Lin Z, Bueno H, Ross JS, Horwitz LI, et al. Diagnoses and timing of 30-day readmissions after hospitalization for heart failure, acute myocardial infarction, or pneumonia. *JAMA* 2013;**309**:355-363. doi:10.1001/jama.2012.216476
- Ambrosy AP, Pang PS, Khan S, Konstam MA, Fonarow GC, Traver B, et al. Clinical course and predictive value of congestion during hospitalization in patients admitted for worsening signs and symptoms of heart failure with reduced ejection fraction: findings from the EVEREST trial. *Eur Heart J* 2013;**34**: 835-843. doi:10.1093/eurheartj/ehs444
- Cai A, Zheng C, Qiu J, Fonarow GC, Lip GYH, Feng Y, et al. Prevalence of heart failure stages in the general population and implications for heart failure prevention: reports from the China hypertension survey 2012-15. *Eur J Prev Cardiol* 2023;**30**:1391-1400. doi:10.1093/eurjpc/zwad223
- Segar MW, Jaeger BC, Patel KV, Nambi V, Ndumele CE, Correa A, et al. Development and validation of machine learning-based race-specific models to predict 10-year risk of heart failure: a multi-cohort analysis. *Circulation* 2021;**143**:2370-2383. doi:10.1161/CIRCULATIONAHA.120.053134
- Voors AA, Ouwerkerk W, Zannad F, van Velthuisen DJ, Samani NJ, Ponikowski P, et al. Development and validation of multivariable models to predict mortality and hospitalization in patients with heart failure. *Eur J Heart Fail* 2017;**19**: 627-634. doi:10.1002/ehf.785
- O'Connor C, Fiuzat M, Mulder H, Coles A, Ahmad T, Ezekowitz JA, et al. Clinical factors related to morbidity and mortality in high-risk heart failure patients: the GUIDE-IT predictive model and risk score. *Eur J Heart Fail* 2019;**21**: 770-778. doi:10.1002/ehf.1450
- Parikh RV, Go AS, Bhatt AS, Tan TC, Allen AR, Feng KY, et al. Developing clinical risk prediction models for worsening heart failure events and death by left ventricular ejection fraction. *J Am Heart Assoc* 2023;**12**:e029736. doi:10.1161/JAHA.122.029736
- Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015;**350**: g7594. doi:10.1136/bmj.g7594
- Ambrosy AP, Parikh RV, Sung SH, Tan TC, Narayanan A, Masson R, et al. Analysis of worsening heart failure events in an integrated health care system. *J Am Coll Cardiol* 2022;**80**:111-122. doi:10.1016/j.jacc.2022.04.045
- Heidenreich PA, Bozkurt B, Aguilar D, Allen LA, Byun JJ, Colvin MM, et al. 2022 AHA/ACC/HFSA guideline for the management of heart failure: executive summary: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation* 2022;**145**:e876-e894. doi:10.1161/CIR.0000000000001062
- Pelliccia A, Sharma S, Gati S, Bäck M, Börjesson M, Caselli S, et al. 2020 ESC guidelines on sports cardiology and exercise in patients with cardiovascular disease: the task force on sports cardiology and exercise in patients with cardiovascular disease of the European Society of Cardiology (ESC). *Eur Heart J* 2021;**42**: 17-96. doi:10.1093/eurheartj/ehaa605
- Rector TS. Patients' self-assessment of their congestive heart failure. Part 2: content, reliability and validity of a new measure, the Minnesota living with heart failure questionnaire. *Heart failure* 1987;**3**:
- Greene SJ, Fonarow GC, Vaduganathan M, Khan SS, Butler J, Gheorghide M. The vulnerable phase after hospitalization for heart failure. *Nat Rev Cardiol* 2015;**12**:220-229. doi:10.1038/nrcardio.2015.14
- van Smeden M, Moons KG, de Groot JA, Collins GS, Altman DG, Eijkemans MJ, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. *Stat Methods Med Res* 2019;**28**:2455-2474. doi:10.1177/0962280218784726
- Obuchowski NA, Bullen JA. Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine. *Phys Med Biol* 2018;**63**:07TR01. doi:10.1088/1361-6560/aab4b1
- Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006;**26**:565-574. doi:10.1177/0272989X06295361
- Fenlon C, O'Grady L, Doherty ML, Dunnion J. A discussion of calibration techniques for evaluating binary and categorical predictive models. *Prev Vet Med* 2018;**149**:107-114. doi:10.1016/j.prevetmed.2017.11.018
- Li W, Guo Q. Plotting receiver operating characteristic and precision-recall curves from presence and background data. *Ecol Evol* 2021;**11**:10192-10206. doi:10.1002/ece3.7826
- Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: *In Proceedings of the 31st International Conference on Neural Information Processing Systems NIPS'17*. (Curran Associates Inc.); 2017:4768-4777.
- Li J, Liu S, Hu Y, Zhu L, Mao Y, Liu J. Predicting mortality in intensive care unit patients with heart failure using an interpretable machine learning model: retrospective cohort study. *J Med Internet Res* 2022;**24**:e38082. doi:10.2196/38082
- Authors/Task Force Members McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Böhm M, et al. 2021 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure: developed by the task force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). With the special contribution of the Heart Failure Association (HFA) of the ESC. *Eur J Heart Fail* 2022;**24**:4-131. doi:10.1002/ehf.2333
- Butler J, Yang M, Manzi MA, Hess GP, Patel MJ, Rhodes T, et al. Clinical course of patients with worsening heart failure with reduced ejection fraction. *J Am Coll Cardiol* 2019;**73**:935-944. doi:10.1016/j.jacc.2018.11.049
- Abraham WT, Stevenson LW, Bourge RC, Lindenfeld JA, Bauman JG, Adamson PB, et al. Sustained efficacy of pulmonary artery pressure to guide adjustment of chronic heart failure therapy: complete follow-up results from the CHAMPION randomised trial. *Lancet* 2016;**387**:453-461. doi:10.1016/S0140-6736(15)00723-0
- Savarese G, Stolfo D, Sinagra G, Lund LH. Heart failure with mid-range or mildly reduced ejection fraction. *Nat Rev Cardiol* 2022;**19**:100-116. doi:10.1038/s41569-021-00605-5
- Huang H, Huang B, Li Y, Huang Y, Li J, Yao H, et al. Uric acid and risk of heart failure: a systematic review and meta-analysis. *Eur J Heart Fail* 2014;**16**: 15-24. doi:10.1093/eurjhf/hft132
- Nishino M, Egami Y, Kawanami S, Sugae H, Ukita K, Kawamura A, et al. Lowering uric acid may improve prognosis in patients with hyperuricemia and heart failure with preserved ejection fraction. *J Am Heart Assoc* 2022;**11**:e026301. doi:10.1161/JAHA.122.026301
- Heidenreich PA, Spertus JA, Jones PG, Weintraub WS, Rumsfeld JS, Rathore SS, et al. Health status identifies heart failure outpatients at risk for hospitalization or death. *J Am Coll Cardiol* 2006;**47**:752-756. doi:10.1016/j.jacc.2005.11.021
- Angraal R, Mortazavi BJ, Gupta A, Khara R, Ahmad T, Desai NR, et al. Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction. *JACC Heart Fail* 2020;**8**:12-21. doi:10.1016/j.jchf.2019.06.013
- Ouwerkerk W, Voors AA, Zwinderman AH. Factors influencing the predictive

- power of models for predicting mortality and/or heart failure hospitalization in patients with heart failure. *JACC Heart Fail* 2014;**2**:429-436. doi:[10.1016/j.jchf.2014.04.006](https://doi.org/10.1016/j.jchf.2014.04.006)
31. Ru B, Tan X, Liu Y, Kannapur K, Ramanan D, Kessler G, *et al.* Comparison of machine learning algorithms for predicting hospital readmissions and worsening heart failure events in patients with heart failure with reduced ejection fraction: modeling study. *JMIR Form Res* 2023;**7**:e41775. doi:[10.2196/41775](https://doi.org/10.2196/41775)
 32. Sabouri M, Rajabi AB, Hajianfar G, Gharibi O, Mohebi M, Avval AH, *et al.* Machine learning based readmission and mortality prediction in heart failure patients. *Sci Rep* 2023;**13**:18671. doi:[10.1038/s41598-023-45925-3](https://doi.org/10.1038/s41598-023-45925-3)
 33. Awan SE, Bennamoun M, Sohel F, Sanfilippo FM, Dwivedi G. Machine learning-based prediction of heart failure readmission or death: implications of choosing the right model and the right metrics. *ESC Heart Fail* 2019;**6**:428-435. doi:[10.1002/ehf2.12419](https://doi.org/10.1002/ehf2.12419)