Risk Management and Healthcare Policy

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Improving Risk Identification of Adverse Outcomes in Chronic Heart Failure Using SMOTE +ENN and Machine Learning

Ke Wang[1–3]
Jing Tian[4]
Chu Zheng[1,3]
Hong Yang[1,3]
Jia Ren[1]
Chenhao Li[1,3]
Qinghua Han[4]
Yanbo Zhang[1,3]

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

Correspondence: Yanbo Zhang
Department of Health Statistics, School of Public Health, Shanxi Medical University, Yingze District 56 New South Road, Taiyuan, People's Republic of China
Tel +86-0351-3985051
Email sxmuzyb@126.com

Qinghua Han
Department of Cardiology, The First Affiliated Hospital of Shanxi Medical University, Yingze District 85 South Jiefang Road, Taiyuan, People's Republic of China
Tel +86 3100113031
Fax +86 351 4867146
Email syhqh@sohu.com

Purpose: This study sought to develop models with good identification for adverse outcomes in patients with heart failure (HF) and find strong factors that affect prognosis.

Patients and Methods: A total of 5004 qualifying cases were selected, among which 498 cases had adverse outcomes and 4506 cases were discharged after improvement. The study subjects were hospitalized patients diagnosed with HF from a regional cardiovascular hospital and the cardiology department of a medical university hospital in Shanxi Province of China between January 2014 and June 2019. Synthesizing minority oversampling technology combined with edited nearest neighbors (SMOTE+ENN) was used to pre-process unbalanced data. Traditional logistic regression (LR), k-nearest neighbor (KNN), support vector machine (SVM), random forest (RF), and extreme gradient boosting (XGBoost) were used to build risk identification models, and each model was repeated 100 times. Model discrimination and calibration were estimated using F1-score, the area under the receiver-operating characteristic curve (AUROC), and Brier score. The best performing of the five models was used to identify the risk of adverse outcomes and evaluate the influencing factors.

Results: The SME-XGBoost was the best performing model with means of F1-score (0.3673, 95% confidence interval [CI]: 0.3633–0.3712), AUC (0.8010, CI: 0.7974–0.8046), and Brier score (0.1769, CI: 0.1748–0.1789). Age, N-terminal pronatriuretic peptide, pulmonary disease, etc. were the most significant factors of adverse outcomes in patients with HF.

Conclusion: The combination of SMOTE+ENN and advanced machine learning methods effectively improved the discrimination efficacy of adverse outcomes in HF patients, accurately stratified patients at risk of adverse outcomes, and found the top factors of adverse outcomes. These models and factors emphasize the importance of health status data in determining adverse outcomes in patients with HF.

Keywords: heart failure, machine learning, SMOTE+ENN, XGBoost, SHAP

Introduction

Heart failure (HF) is the leading cause of death in most countries in the world.[1] According to reports, one in every eight deaths in the United States is due to HF.[2] Recent data show that the prevalence of HF increases as the population ages, the cardiovascular risk profile of the population deteriorates, and survival rates for patients with acute cardiovascular disease improve.[3,4] HF puts a heavy burden on society through the extensive use of healthcare resources. Without doubt, accurately identifying the risk of adverse outcomes in HF is of vital importance to patients, the

Received: 9 March 2021
Accepted: 24 May 2021
Published: 8 June 2021

medical system, and society as a whole. Thanks to the digitization of medical information, particularly the introduction of electronic medical records (EMR) and the phenomenon of big data,[5] researchers have been provided with massive amounts of available data. Moreover, the rise of machine learning (ML) algorithms[6–8] offers researchers with new powerful tools. In fact, many researchers are currently focusing on risk identification using ML; however, it has not yet achieved high accuracy for the identification of HF related events.[9] The reasons can be summarized as follows: first, medical data often show severe category imbalances, but many studies have ignored this problem, leading to predictions biased to most categories; second, the variable screening methods of many studies are laggard, and the influence of variables is not considered comprehensively; third, some studies have not improved model selection and parameter optimization despite of the presence of advanced ML models and parameter optimization methods.

Accordingly, our aim was to use ML methods to address the limitations of the previously proposed models, especially for the unbalanced data processing, and eventually establish an ML model that can well identify the risk of adverse outcomes in HF patients and find strong influencing factors, so as to provide the basis for patients, doctors, and clinical researchers to initiate subsequent treatment and intervention measures.

Patients and Methods

Study Population

The patients for this study were enrolled according to inclusion and exclusion criteria from two medical centers in Shanxi Province of China between January 2014 and June 2019. The data were obtained according to the case report form of chronic heart failure (CHF-CRF) developed by our research group according to the case record content and HF guidelines.[10] CHF-CRF included the patient's demographics, medical history, physicals tatus and vitals, currently applied medical therapy, electrocardiogram, echocardiographic, and laboratory parameters.

The inclusion criteria were 1) aged ≥18 years; 2) diagnosed with HF, according to the guideline for the diagnosis and treatment of HF in China (2018)[11]; 3) fall under the New York Heart Association (NYHA) II–IV Classification; and 4) received HF treatment while in the hospital. Patients who had an acute cardiovascular event within 2 months prior to admission or were unable or refused to participate in the project for some reason were excluded.

Data Preprocessing and Feature Selection

Some variables (also called features in ML) in this study were missing in different ratios. Referring to relevant studies on missing value processing,[12–14] the variables with a missing percentage of no more than 30% were retained and filled with the missForest method.[15,16] The quantitative data were normalized, and the multi-categorical variables were processed by One-Hot.[17] After initial screening by single-factor method, recursive feature elimination (RFE) based on random forest (RF) with five-fold cross-validation (CV) was used to screen the overall features. The main idea of RFE is to repeatedly build the model and then select the best feature, pick out the selected feature, and then repeat this process on the remaining features until all features have been traversed.

Model Development

In addition to several commonly used supervised learning algorithms such as logistic regression (LR), k-nearest neighbor (KNN), support vector machine (SVM), random forest (RF),[18] we introduced extreme gradient boosting (XGBoost) algorithm, which has attracted a lot of attention in recent years due to its computational speed, generalization ability and high predictive performance.[19,20] According to whether adverse outcomes occurred, 5003 patients were divided into training set, verification set, and test set in a 3:1:1 ratio by stratified random sampling. The training validation set (training set+verification set) and verification set were pretreated using the synthesizing minority oversampling technology combined with edited nearest neighbors (SMOTE +ENN). We used a Grid Search method with fivefold CV to optimize the hyperparameters of the ML models in the original verification set and the pretreated verification set, respectively, and then used the ML models with the optimal hyperparameters to train the original training verification set and the pretreated training verification set (details in Supplementary Table 1). Finally, the performance of each model was evaluated and compared in the test set. To obtain a more robust performance estimate, avoid reporting biased results and limit overfitting, we repeat the holdout method 100 times with different random seeds and compute the average performance over these 100 repetitions[21] (Figure 1).

SMOTE+ENN is a comprehensive sampling method proposed by Batista et al in 2004,[22] which combines the SMOTE and the Wilson's Edited Nearest Neighbor Rule

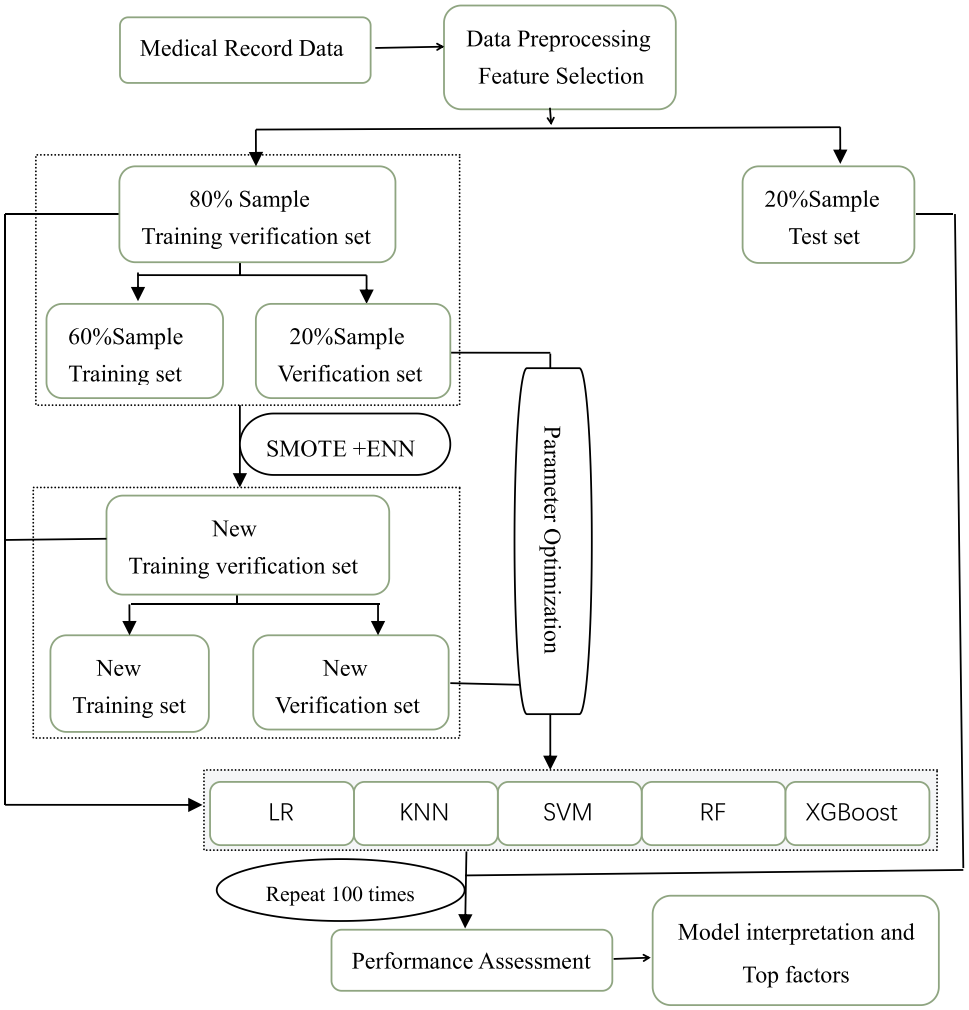

Figure 1 Architecture of the system.

(ENN).[23] SMOTE is an over-sampling method, and its main idea is to form new minority class examples by interpolating between several minority class examples that lie together. Although it can effectively improve the classification accuracy of the model, it can also generate noise samples and boundary samples. To create better defined class clusters, ENN is used as a data cleaning method that can remove any example whose class label differs from the class of at least two of its three nearest neighbors. Since some majority class examples might in vade the minority class space and vice versa, SMOTE+ENN reduces the possibility of overfitting introduced by synthetic examples.[22]

The KNN method is a popular classification method in data mining and statistics because of its simple implementation and significant classification performance.[24] The idea is that if the majority of the k most similar samples (ie, the nearest neighbors in the feature space) of a sample belong to a certain category, the sample also belongs to this category, where K is usually not greater than 20. In the KNN algorithm, the selected neighbors are all objects that have been correctly classified. This method only determines the category to which the sample to be classified belongs based on the category of the nearest sample or samples.

SVM is one of the most important methods in ML, which is broadly applied to image recognition and image processing.[25] It is used to classify data through approximate inter-class distance in high dimensional space, and can satisfactorily solve the problems of small sample size, nonlinearity, and high dimensional data recognition and classification. The SVM looks for an optimal plane that can divide the sample observed in multi-dimensional space into two optimal planes. This optimal plane enables the two categories to be separated with the greatest possible distance from the nearest point. On the spacing boundary, the point that determines the spacing is the support vector, and the segmented hyperplane is in the middle of the spacing.

An RF algorithm is a scheme that was proposed in the 2000s by Breiman for building a predictor ensemble with a set of decision trees that grow in randomly selected subspaces of data.[26] Integration is not just a simple bagging integration,[27] it combines the idea of bagging integration and feature selection. The RF classifier consists of a combination of tree classifiers, where each classifier is generated using a random vector that is independent of the input vector samples, and each tree votes for the most classes to classify the input vector. Numerous studies conducted worldwide have shown that RF algorithms perform very well in classification and prediction in various fields.[28]

Tree boosting[29] is a highly effective and widely used ML method. XGBoost is an ensemble learning algorithm based on gradient boosting theory, it is a scalable end-to-end tree enhancement system proposed by Chen and Guestrin in 2016.[30] Owing to its good scalability and high efficiency in the face of large data sets, it has been widely used by data scientists and has obtained the most advanced results in many ML challenges in recent years. Compared with the traditional gradient boosting decision tree, XGBoost has further improved the loss function, regularization, and parallelization,[31] and has achieved good results in many application scenarios for classification problems and regression problems.

Performance Evaluation

Multiple evaluation indexes such as F1-score, the area under the receiver-operating characteristic curve (AUROC), and Brier score[32] were used to comprehensively evaluate the discrimination and calibration of ML models (details in Supplementary materials).

Model Interpretation and Feature Importance

We used the best-performing of the five ML models to assess the importance of each variable. Moreover, we implemented SHapley Additive exPlanations (SHAP), which is a recent approach to explain the output of a ML model, to illustrate the individual feature-level impacts. In brief, SHAP is an additive feature attribution method that provides an explanation of the tree ensemble's overall impact in the form of particular feature contributions and is relatively consistent with human intuition.[33]

Software Packages

All operations were implemented in Python 3.6.5, and various Python modules were used to conduct the analysis. The GridSearchCV from sklearn.model_selection was used for grid search with 5-fold cross-validation. The SMOTEENN from imblearn.combine was used for SMOTE+ENN. The LogisticRegression from sklearn.linear_model was used for Logistic regression. The KNeighborsClassifier from sklearn.neighbors was used for KNN. The SVC from sklearn.svm was used for SVM. The RandomForestClassifier from sklearn.ensemble was used for RF. The XGBClassifier from xgboost.sklearn was used for XGBoost.

Results

Patient Characteristics

A total of 5004 inpatients were included in this study, including 3292 males (65.79%), with an average age of 65.73 ± 11.58 years old and 1712 females (34.21%), with an average age of 70.80 ± 10.32 years old. Among these patients, 498 patients had adverse outcomes (deterioration or death), 4506 patients improved and were discharged, and the ratio of the two types of patients was 1:9.05, which represents an imbalanced data set.

Variables Selected

After feature selection by single factor and the RFE-RF with fivefold CV, the final optimal number of features was 44 (Figure 2, Table 1) (details in Supplementary Table 2).

Outcomes of the ML Models

Among the evaluated ML models, SME-XGBoost yielded the highest F1-score and AUROC. The Brier score was also relatively low (Table 2). Therefore, SME-XGBoost was used as the optimal model for further study.

Categorization of Prediction Score and Risk Distributions

The best performing SME-XGBoost model was used to identity the risk of adverse outcomes in the test set. The Brier score of the model was 0.1769, indicating that the final model was well calibrated and could accurately identify patients with adverse outcomes. The patients were separated into two groups, low and high prediction scores, using the maximal Youden's index as an optimal cut-off value (0.3739) (Figure 3A). At this cut-off, the prediction scores was associated with a sensitivity and specificity of 0.798 and 0.690, respectively. The distribution plots of the patient risk sequence identified by the model showed a certain aggregation of patients who had adverse

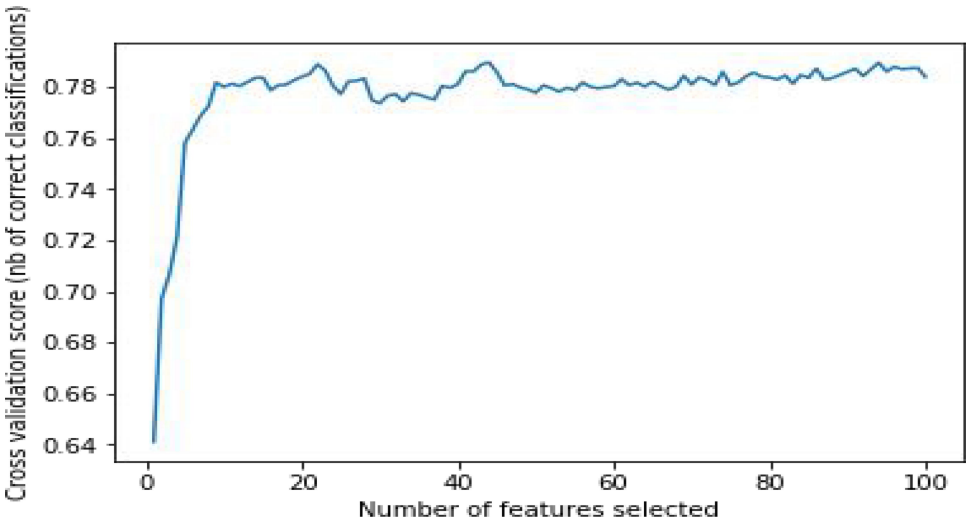

Figure 2 Results of feature screening by RFE-RF with fivefold CV.

outcomes (Figure 3B), indicating that the model accurately stratified patients at low or high risk.

Model Interpretation and Feature Importance

SHAP plot can give physicians an intuitive understanding of key features in the model and it visually displays the top 20 risk factors (Figure 4). Older age, higher value of N-terminal pronatriuretic peptide (NT-proBNP), direct bilirubin (DBIL), QRS wave, creatinine (CR), heart rate, glucose (GLU), red blood cell volume distribution width (RDW), anteroposterior diameter of right atrium (RA), diastolic pressure (DP), and lower value of albumin (ALB), urine-specific gravity (SG), systolic pressure, red blood cells (RBC), chloride ion concentration (CL) were associated with higher risk probability of adverse outcomes in patients with HF. In addition, pulmonary disease (PUMONARY), high level of New York Heart Association (NYHA) clinical classifications, and pulmonary aortic valve regurgitation (PVSIAI-1) were also higher risk factors for adverse outcomes.

Discussion

HF damages the quality of life more than almost any other chronic diseases.[4] Accurate identification of prognostic risks is fundamental to patient-centered care, both in selecting treatment strategies and in informing patients as a foundation for shared decision making.[32] Although published reports are abundant with different models identifying the risk of either mortality or hospitalizations in patients with HF,[34] the present study extends this knowledge in several important ways. First, most standard algorithms assume or expect balanced class distributions or equal misclassification costs. When presented with imbalanced data sets these algorithms fail to properly represent the distributive characteristics of the data, and thus providing unfavorable accuracies across the classes of the data.[35] Unfortunately, in the field of biomedicine, unbalanced data are ubiquitous, as the number of healthy people for whom medical data has been collected is often much larger than that of unhealthy ones. This provides us with new challenges in exploring disease risk identification models. If the problem of category imbalance was ignored, the risk identification model built with imbalanced data sets tends to envisage a higher accuracy rate for the majority class and ignore the minority class. The detailed performance is that the F1-score of the models is very close to or even equal to 0. It indicates that the ability of the model to identify true positive outcomes is very poor, which can be confirmed in our study (Table 2). Studies have shown that for several base classifiers, a balanced data set provides improved over all classification performance compared to an imbalanced data set.[36,37] Thus, it is essential to use an effective preprocessing method to deal with imbalances before modeling so as to improve the accuracy of the model.[38] In some reports, SMOTE is a typical oversampling technique which can effectively deal with the imbalanced data. However, it brings noise and other problems, affecting the classification accuracy.[39] Our study extends this knowledge in an effective way. We used SMOTE +ENN to preprocess the data. In addition to the data

Table 1 Risk Factors Selected for Adverse Outcomes in Patients with HF

Variable	Adverse Outcomes		P value	Variable	Adverse Outcomes		P value
	No	Yes			No	Yes	
Age (years)	67.0(59.0–76.0)	76.0(68.0–81.0)	<0.001	HDLC (μmol/L)	1.0(0.8–1.1)	1.0(0.9–1.2)	0.004
DP (mmHg)	130(120–140)	130(118–150)	0.029	LDLC (μmol/L)	2.4(1.9–2.9)	2.3(1.8–2.9)	0.008
SP (mmHg)	80(70–85)	76(70–84)	<0.001	BUN (mmol/L)	6.0(4.9–7.6)	7.0(5.4–9.41)	<0.001
Height (cm)	167.0(160.0–171.0)	165.0(160.0–170.0)	0.013	CR (mmol/L)	78.0(66.0–92.9)	91.2(74.9–115.6)	<0.001
Weight (kg)	69.0(60.0–75.0)	65.0(55.0–71.0)	<0.001	UA (μmol/L)	365.0(297.0–443.0)	403.0(324.0–502.1)	<0.001
BMI (kg/m)	24.9(22.5–27.2)	23.4(21.1–25.9)	<0.001	K.1 (mmol/L)	4.1(3.8–4.3)	4.1(3.8–4.4)	0.007
WBC (10^9/L)	6.6(5.5–7.9)	6.9(5.7–8.4)	0.003	NA (mmol/L)	140.0(138.0–142.0)	139.3(137.0–141.2)	<0.001
RBC (10^{12}/L)	4.4(4.0–4.8)	4.2(3.8–4.6)	<0.001	CL (mmol/L)	104.0(101.8–107.0)	102.2(99.4–105.0)	<0.001
RDW (%)	13.8(13.3–14.5)	14.4(13.7–15.3)	<0.001	CYSC (mg/L)	1.1(0.9–1.3)	1.27(1.04–1.6)	<0.001
HGB (g/L)	137.0(125.0–149.0)	130.0(117.0–143.0)	<0.001	NTPROBNP	869.8(324.8–2427.7)	3072.1(1324.3–6324.1)	<0.001
NEU (10^{10}/L)	4.2(3.3–5.3)	4.7(3.6–5.9)	<0.001	SG	1.0(1.0–1.0)	1.0(1.0–1.0)	0.007
N (%)	63.5(57.1–70.0)	68.5(62.3–75.1)	<0.001	Heartrate	70(62–82)	78.5(67–92)	<0.001
ALT (U/L)	19.0(13.4–29.0)	17.0(11.8–28.0)	<0.001	QRS (ms)	96(88–108)	102(90–122)	<0.001
ALB (g/L)	43.6(40–46.9)	40.8(37.0–43.8)	<0.001	QTC (ms)	431(406–462)	447(420–478)	<0.001
TBIL (μmol/L)	14.5(11.0–19.6)	15.3(11.3–21.7)	0.006	LA (mm)	38.4(36.0–42.0)	41.0(38.0–46.0)	<0.001
DBIL (μmol/L)	3.5(2.4–5.2)	4.8(3.1–6.6)	<0.001	RA (mm)	35.0(31.0–40.0)	37.8(33.0–45.0)	<0.001
x.GT (U/L)	27.0(18.1–43.7)	33.0(20.0–56.0)	<0.001	RA1 (mm)	43.0(39.0–47.0)	45.0(40.0–50.0)	<0.001
GLU (μmol/L)	5.1(4.5–6.2)	5.3(4.6–6.8)	<0.001	LVDD (mm)	52.0(47.0–58.0)	55.0(49.0–61.0)	<0.001
TG (mmol/L)	1.4(1.0–1.9)	1.2(0.9–1.6)	<0.001	EF (%)	53.0(41.0–62.0)	45.0(35.0–56.3)	<0.001
Healthcare			<0.001	NYHA			<0.001
Urban employee	2270(50.4%)	263(52.8%)			18(0.4%) 0.4%	0(0.0%)	
Urban residents	559(13.30%)	56(11.2%)			2025(44.9%)	96(19.3%)	
Rural cooperative	1160(25.7%)	103(20.7%)			1696(37.6%)	193(38.8%)	
Poverty relief	6(0.1%)	0(0.0%)		IV	767(17.0%)	209(42.0%)	

Full public	24(0.5%)	11(2.2%)		Pumonary		<0.001
Self-paying	142(3.2%)	31(6.2%)		No	3968(88.1%)	327(65.7%)
Other	305(6.8%)	34(6.8%)		Yes	538(11.9%)	171(34.3%)
Lung Rales			<0.001	PVSI AI		<0.001
No	3648(81.0%)	285(57.2%)		No	2507(55.6%)	179(35.9%)
Moist rales	830(18.4%)	205(41.2%)		Little	1718(38.1%)	246(49.4%)
Dry rales	28(0.6%)	8(1.6%)		Moderate	246(5.5%)	59(11.8%)
Infection			<0.001	Massive	35(0.8%)	14(2.8%)
No	4129(91.6%)	376(75.5%)				
Yes	377(8.4%)	122(24.5%)				

Note: Values are median (interquartile range) or n (%).

imbalance issue, this method also solved the problem that the SMOTE algorithm is prone to overlapping data and noise. The performance of each model constructed on the data processed by SMOTE+ENN improved significantly in the study, particularly for F1-score as indicator that reflect the detection rate of positive events. The above results show that SMOTE+ENN can effectively solve the problem of classification deviation caused by unbalanced data and provide a reference for future classification prediction research of imbalanced data. Second, most of the previous models were developed using traditional statistical approaches. However, the new alternatives, such as ML–based models, have remained not under used.[40] Advanced statistical tools and ML methods can improve the risk identification ability of traditional statistical techniques in various ways.[41] In our study, in addition to the advanced ML model, other ML knowledge that has been shown to effectively improve the performance of risk identification models was also used, such as the missing value filling based on missForest, feature selection based on RFECV, and hyperparameter optimization based on GridSearchCV. Among the evaluated models, SME-XGBoost demonstrated the best performance, and this algorithm was used to evaluate the impact factors. XGBoost combining SMOTE+ENN forms the foundation for future testing of the clinical utility with more accurate risk stratification of patients' care and outcomes. Third, this study found that models constructed from data collected by CHF-CRF can accurately identity the risk of adverse outcomes. If combined with rigorous clinical trials, better risk identification results can be obtained, which is the next step in our research. Fourth, although many ML models can provide the importance of variables, they have difficulty explaining whether variables increase or decrease the occurrence of outcomes. Meanwhile, the lack of intuitive understanding of ML models among clinicians is one of the major obstacles to the implementation of ML in the medical field.[42] In our study, we employed ML methods to account for feature importance in specific domains, apply a visual interpretation of the importance of each feature, and compared the accuracy of different ML models using risk identification for adverse outcomes in patients with HF.

The study ultimately included 44 variables. Majority of them are routinely assessed during the management of HF; therefore, they are readily available from EMR. In our study, we found that age, systolic pressure, creatinine, NYHA, and NT-proBNP were important factors of adverse

Table 2 Results of ML Models for the Unbalanced Data and the Data After Pretreatment with SMOTE+ENN(SME) [Mean (95% CI)]

Models	F1-Score	AUC	Brier Score
LR	0.0000(0.0000,0.0000)	0.7583(0.7542,0.7624)	0.7583(0.7542,0.7624)
KNN	0.0375 (0.0322,0.0429)	0.6721 (0.6675,0.6768)	0.0904 (0.0898,0.0909)
SVM	0.0000 (0.0000,0.0000)	0.7218 (0.7117,0.7318)	0.0869 (0.0865,0.0873)
RF	0.0000 (0.0000,0.0000)	0.7993 (0.7957,0.8030)	0.0796 (0.0793,0.0798)
XGBoost	0.3515 (0.3458,0.3572)	0.7918 (0.7879,0.7957)	0.1733 (0.1728,0.1737)
SME-LR	0.2914(0.2891,0.2936)	0.7819(0.7784,0.7853)	0.2801(0.2782,0.2820)
SME-KNN	0.2667 (0.2631,0.2703)	0.6481 (0.6437,0.6525)	0.3256 (0.3230,0.3283)
SME-SVM	0.1976 (0.1922,0.2030)	0.6963 (0.6925,0.7001)	0.1632 (0.1615,0.1650)
SME-RF	0.3606 (0.3567,0.3645)	0.7983 (0.7947,0.8019)	0.1577 (0.1565,0.1588)
SME-XGBoost[b]	0.3673 (0.3633,0.3712)	0.8010 (0.7974,0.8046)	0.1769 (0.1748,0.1789)
P value[a]	<0.001	<0.001	<0.001

Notes: [a]P value is the result of one-way analysis of variance for the three indicators of models. [b]After multiple comparisons of least-significant difference (LSD), SME-XGBoost is significantly different from other models.

outcomes, which is consistent with the results of a recent systematic review of 117 HF predictive models.[43] Meanwhile, the importance of these factors has also been confirmed in other studies.[32,44,45] However, several highly important factors of adverse outcomes from the present study such as pulmonary disease, albumin, DBIL, QRS, SG and CL were not reported in previous studies to the best of our knowledge. It suggests that these factors should be paid more attention in the future and it also provides a new basis for the future study of the prognosis of HF. In addition, some investigators found that sex, sodium, diabetes, blood urea nitrogen, hemoglobin, ejection fraction, angiotensin-converting enzyme inhibitor treatment and left ventricular systolic dysfunction had significant impact for adverse outcomes in patients with HF,[40,42,45] but these factors did not show strong influence in this study.

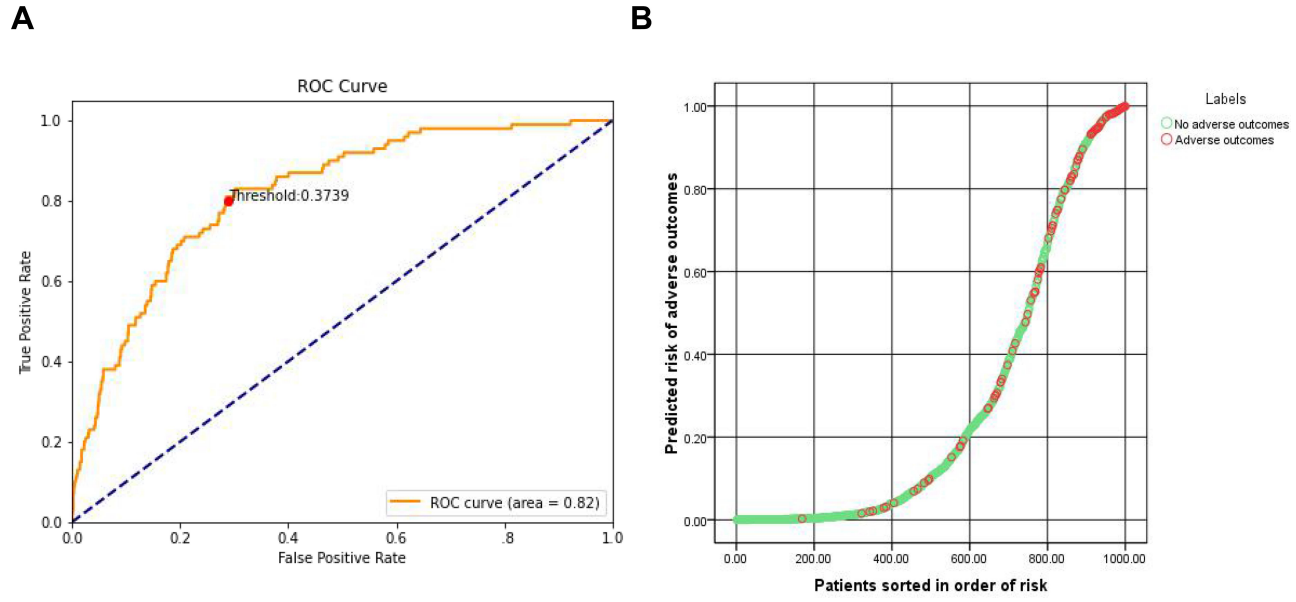

Figure 3 Categorization threshold of prediction score (**A**) and prediction distributions of adverse outcomes in patients with HF (**B**).

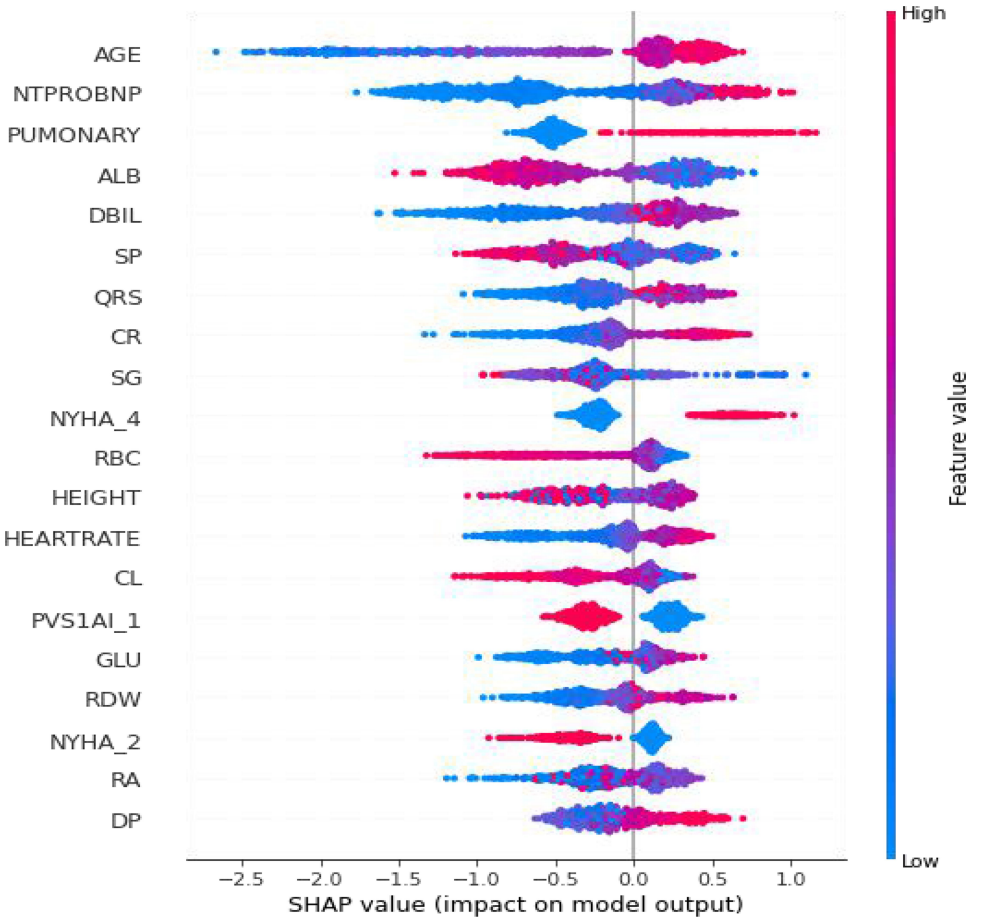

Figure 4 SHAP summary plots for the risk of adverse outcomes in patients with HF. The importance ranking of the top 20 risk factors with stability and interpretation using SME-XGBoost model. The SHAP value (x-axis) is a unified index responding to the impact of a feature in the model. In each feature importance row, all patients' attribution to outcome were plotted using different color dots, in which the red dot represented high risk value and the blue dot represented low risk value.

Limitations and Development

First, this study used a retrospective study—without follow-up of patients—and all patient information was collected in Shanxi Province, meaning it could be stored with a certain bias. In further, we will expand the scope of data collection, make full use of the advantages of EMR information, and carry out patient follow-up, combined with a time factor. Meanwhile, we will collect more data from different hospitals and regions, and use data from different regions as external validation of this model. Second, the information collected in this study was structured data, further research is needed to unearth unstructured information, and add imaging information, biomarkers, environmental factors, and lifestyle habits, as well as other factors to improve prediction. Third, this research solves the problem of data imbalance from the data level. The next step is to combine this with the algorithm level. Fourth, although this study has achieved good results, there is still the possibility of further improvement. With the rapid development of artificial intelligence, deep learning has been applied to the construction of medical models. Future research will introduce deep learning to predict the prognosis of HF, and combine more extensive data and information to conduct research on different levels.

Conclusions

Combining SMOTE+ENN and advanced ML methods effectively improved the risk identification of adverse outcomes in patients with HF, and accurately stratified patients at risk of adverse outcomes. This method can be used to solve the problem of class imbalance in medical data modeling in the future. Moreover, ML model and SHAP plot can provide intuitive explanations of what led to a patients' predicted risk, thus helping clinicians better understand the decision-making process for disease severity assessment. The features can provide a reference for

intervention and the models can be used by clinicians as an important tool for identifying the high-risk patients.

Data Sharing Statement

The datasets during and/or analysed during the current study available from the corresponding author on reasonable request.

Ethical Approval

The study complies with the Declaration of Helsinki and has been approved by the Medical Ethics Committee of Shanxi Medical University. All patients were informed about the purpose of the study and provided written informed consent.

Consent for Publication

Not applicable.

Acknowledgment

We thank Sarah Dodds, PhD, from LiwenBianji, Edanz Editing China (www.liwenbianji.cn/ac), for editing the English text of a draft of this manuscript. We thank Shanxi Cardiovascular Hospital and the First Affiliated Hospital of Shanxi Medical University for their help in the data collection process.

Author Contributions

All authors made substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data; took part in drafting the article or revising it critically for important intellectual content; agreed to submit to the current journal; gave final approval of the version to be published; 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: 818 727 14]; Shanxi Provincial Key Laboratory of Major Diseases Risk Assessment under Grant [number 201805D111006];Youth Science and Technology Research Foundation of Shanxi Province under Grant [number 201801D221423] and Shanxi Provincial Key Laboratory of Major Diseases Risk Assessment under Grant [number 201604D132042].

Disclosure

The authors declare that they have no competing interests.

References

1. Dokainish H, Teo K, Zhu J, et al. Global mortality variations in patients with heart failure: results from the International Congestive Heart Failure (INTER-CHF) prospective cohort study. *Lancet Global Health*. 2017;5(7):e665–e672. doi:10.1016/S2214-109X(17)30196-1
2. Benjamin EJ, Virani SS, Callaway CW, et al. Heart disease and stroke statistics—2018 update: a report from the American Heart Association. *Circulation*. 2018;137(12):e67–e492.
3. Ponikowski P, Anker SD, AlHabib KF, et al. Heart failure: preventing disease and death worldwide. *ESC Heart Fail*. 2014;1:4–25. doi:10.1002/ehf2.12005
4. Mcmurray JJV, Stewart S. The burden of heart failure. *Eur Heart J Suppl*. 2002;(suppl_D):3–13.
5. Gandomi A, Haider M. Beyond the hype: big data concepts, methods, and analytics. *Int J Inf Manage*. 2015;35(2):137–144. doi:10.1016/j.ijinfomgt.2014.10.007
6. Kavakiotis I, Tsave O, Salifoglou A, et al. ML and data mining methods in diabetes research. *Comput Struct Biotechnol J*. 2017;15:104–116. doi:10.1016/j.csbj.2016.12.005
7. Brisimi TS, Xu T, Wang T, Dai W, Paschalidis IC. Predicting diabetes-related hospitalizations based on electronic health records. *Stat Methods Med Res*. 2018;962280218810911. doi:10.1177/0962280218810911
8. Zou Q, Qu K, Luo Y. et al. Predicting diabetes mellitus with machine learning techniques. *Front Genet*;2018. 9. doi:10.3389/fgene.2018.00515
9. Buchan TA, Ross HJ, Mcdonald M, et al. Physician prediction versus model predicted prognosis in ambulatory patients with heart failure. *J Heart Lung Transpl*. 2019;38(4):S381. doi:10.1016/j.healun.2019.01.971
10. Yancy CW, Jessup M, Bozkurt B, et al. 2017 ACC/AHA/HFSA Focused Update of the 2013 ACCF/AHA Guideline for the Management of Heart Failure: a Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Failure Society of America. *J Am Coll Cardiol*. 2016;68(13):1476–1488. doi:10.1016/j.jacc.2016.05.011
11. Heart Failure Group of Chinese Society of Cardiology of Chinese Medical Association; Chinese Heart Failure Association of Chinese. Medical Doctor Association; Editorial Board of Chinese Journal of Cardiology. Chinese guidelines for the diagnosis and treatment of heart failure 2018. *Chin J Cardiol*. 2018;46(10):760.
12. Schmitt P, Mandel J, Guedj M. A Comparison of Six Methods for Missing Data Imputation. *Biomet & Biostats*. 2015;6(1).
13. Lodder P, Rotteveel M, van Elk M. To Impute or not Impute: that's the Question. *Front Psychol*. 2014;5. doi:10.3389/fpsyg.2014.00967
14. Jakobsen JC, Gluud C, Wetterslev J, et al. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. *BMC Med Res Methodol*. 2017;17(1):162. doi:10.1186/s12874-017-0442-1
15. Stekhoven DJ, Buhlmann P. Miss Forest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28(1):112-118.
16. Thio Q, Karhade AV, Bindels B, et al. Development and internal validation of machine learning algorithms for preoperative survival prediction of extremity metastatic disease. *Clin Orthop Relat Res*. 2019;478(2):1.
17. Okada S, Ohzeki M, Taguchi S. Efficient partition of integer optimization problems with one-hot encoding. *Sci Rep*. 2019;9(1). doi:10.1038/s41598-019-49539-6
18. Singh A, Thakur N, Sharma A. A review of supervised machine learning algorithms. 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). IEEE. 2016;1310–1315.

19. Azeez A, Ogunleye W XGBoost model for chronic kidney disease diagnosis. *IEEE/ACM Transac Computat Biol Bioinformat*. 2019.
20. Li M, Fu X, Li D. Diabetes prediction based on XGBoost algorithm. *MS&E*. 2020;768(7).
21. Takeda A, Kanamori T. A robust approach based on conditional value-at-risk measure to statistical learning problems. Elsevier: *European Journal of Operational Research*. 2009, 198(1):287-296.
22. Batista GEAPA, Prati RC, Monard MC. A study of the behavior of several methods for balancing ML training data . *ACM SIGKDD Explor Newsletter*. 2004;6(1):20. doi:10.1145/1007730.1007735
23. Wilson DL. Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems. *Man, Commun*. 1972;2 (3):408–421.
24. Zhang S Efficient kNN classification with different numbers of nearest neighbors. *IEEE Transac Neural Networks Learning Systems*. 2018;5(29).
25. Han J, Jiang W, Dai C, et al. The design of diabetic retinopathy classifier based on parameter optimization SVM[C]// International Conference on Intelligent Informatics & Biomedical Sciences. *IEEE Computer Society*. 2018.
26. Biau G. Analysis of a Random Forests Model. *J ML Res*. 2010;13 (2):1063–1095.
27. Altman N, Krzywinski M. Points of Significance: ensemble methods: bagging and random forests. *Nat Methods*. 2017;14(10):933–934. doi:10.1038/nmeth.4438
28. Kennedy W. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. *Ecol Indic*. 2015;52:394–403. doi:10.1016/j.ecolind.2014.12.028
29. Friedman J. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29(5):1189–1232. doi:10.1214/aos/1013203451
30. Chen T, Guestrin C, Xgboost: a scalable tree boosting system. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. ACM (Association for Computing Machinery) Digital Library. 2016:785–794.
31. Hongshan ZHAO, Xihui YAN, Guilan WANG, et al. Fault diagnosis of wind turbine generator based on deep autoencoder network and XGBoost. *Autom Electr Power Syst*. 2019;43(1):81–90.
32. Angraal S, Mortazavi BJ, Gupta A, et al. ML Prediction of mortality and hospitalization in heart failure with preserved ejection fraction. *JACC Heart Fail*. 2020;8(1):12–21. doi:10.1016/j.jchf.2019.06.013
33. Hu CA, Chen CM, Fang YC, et al. Using a ML approach to predict mortality in critically ill influenza patients: a cross-sectional retrospective multicentre study in Taiwan. *BMJ Open*. 2020;10(2): e033898. doi:10.1136/bmjopen-2019-033898
34. Rahimi K, Bennett D, Conrad N, et al. Risk prediction in patients with heart failure: a systematic review and analysis. *JACC Heart Fail*. 2014;2(5):440–446. doi:10.1016/j.jchf.2014.04.008
35. He H, Garcia EA. Learning from Imbalanced Data. *IEEE Trans Knowl Data Eng*. 2009;21(9):1263–1284. doi:10.1109/TKDE.2008.239
36. Laurikkala J. Improving identification of difficult small classes by balancing class distribution[C]// Proceedings of the 8th Conference on AI in Medicine in Europe: artificial Intelligence Medicine. *Springer Berlin Heidelberg*. 2001.
37. Kanimozhi MA. A Multiple Resampling Method for Learning from Imbalanced Data Sets. *Comput Intell*. 2010;20(1):18–36.
38. Tavares TR, Oliveira ALI, Cabral GG, et al. Preprocessing unbalanced data using weighted support vector machines for prediction of heart disease in children[C]// International Joint Conference on Neural Networks. *IEEE*. 2014.
39. Mi Y. Imbalanced classification based on active learning SMOTE. *Res j Applied Sci, Engineering Technol*. 2013;5(3):944–949.
40. Frizzell JD, Liang L, Schulte PJ, et al. Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of ML and other statistical approaches. *JAMA Cardiol*. 2017;2 (2):204–209. doi:10.1001/jamacardio.2016.3956
41. Mortazavi BJ, Downing NS, Bucholz EM, et al. Analysis of machine learning techniques for heart failure readmissions. *Circ Cardiovasc Qual Outcomes*. 2016;9(6):629–640. doi:10.1161/CIRCOUTCOMES.116.003039
42. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA*. 2017;318(6):517. doi:10.1001/jama.2017.7797
43. Babayan ZV, Mcnamara RL, Nagajothi N, et al. Predictors of cause-specific hospital readmission in patients with heart failure. *Clin Cardiol*. 2010;26(9):411–418. doi:10.1002/clc.4960260906
44. Cunha FM, Pereira J, Ribeiro A, et al. Age affects the prognostic impact of diabetes in chronic heart failure . *Acta Diabetol*. 2018;55 (10):1–8. doi:10.1007/s00592-017-1092-9
45. Chicco D, Jurman G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med Inform Decis Mak*. 2020;20(1):16. doi:10.1186/s12911-020-1023-5