IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 30, 2017, accepted December 31, 2017, date of publication January 5, 2018, date of current version March 9, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2789898

Predictive Modeling of Hospital Mortality for Patients With Heart Failure by Using an Improved Random Survival Forest

FEN MIAO[1,2], YUN-PENG CAI[1,2], YU-XIAO ZHANG[3], XIAO-MAO FAN[1,2], AND YE LI[1,2]

[1]Key Laboratory for Health Informatics, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
[2]Shenzhen Engineering Laboratory of Health Big Data, Shenzhen 518055, China
[3]Department of Senile Cardiovascular Medicine, The General Hospital of the People's Liberation Army, Beijing 100853, China

Corresponding author: Ye Li (ye.li@siat.ac.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61502472 and Grant 61771465, in part by the National 863 Project of China under Grant SS2015AA020109, in part by the Basic Research and Discipline Layout Project of Shenzhen under Grant JCY20170413161515911, in part by the Basic Research Program of Shenzhen under Grant JCYJ20150630114942316, in part by the Science and Technology Planning Project of Guangdong Province under Grant 2015B010129012, and in part by the Guangdong Province Special Support Program under Grant 2014TX01X060.

ABSTRACT Identification of different risk factors and early prediction of mortality for patients with heart failure are crucial for guiding clinical decision-making in Intensive care unit cohorts. In this paper, we developed a comprehensive risk model for predicting heart failure mortality with a high level of accuracy using an improved random survival forest (iRSF). Utilizing a novel split rule and stopping criterion, the proposed iRSF was able to identify more accurate predictors to separate survivors and nonsurvivors and thus improve discrimination ability. Based on the public MIMIC II clinical database with 8 059 patients, 32 risk factors, including demographics, clinical, laboratory information, and medications, were analyzed and used to develop the risk model for patients with heart failure. Compared with previous studies, more critical laboratory predictors were identified that could reveal difficult-to-manage comorbidities, including aspartate aminotransferase, alanine aminotransferase, total bilirubin, serum creatine, blood urea nitrogen, and their inherent effects on events; these were determined to be critical indicators for predicting heart failure mortality with the proposed iRSF. The experimental results showed that the developed risk model was superior to those used in previous studies and the conventional random survival forest-based model with an out-of-bag C-statistic value of 0.821. Therefore, the developed iRSF-based risk model could serve as a valuable tool for clinicians in heart failure mortality prediction.

INDEX TERMS Heart failure, survival analysis, risk prediction, random survival forest, predictor.

LIST OF ABBREVIATIONS

ACE-I	Angiotensin-converting enzyme inhibitor.
ARB	Angiotensin-receptor blockers.
CCA	Calcium channel antagonists.
K	Potassium [mEq/L] in Blood.
NA	Sodium [mEq/L] in Blood.
WBC	Leukocytes [K/uL] in Blood.
RBC	Erythrocytes [m/uL] in Blood.
SCR	Creatinine [mg/dL] in Serum.
BUN	Urea nitrogen [mg/dL] in Serum.
CKPK	Creatine kinase.total [IU/L] in Serum.
CKMB	Creatine kinase.MB in Serum.
AST	Aspartate aminotransferase [IU/L] in Serum.
ALT	Alanine aminotransferase [IU/L] in Serum.
BR	Total bilirubin [mg/dL] in Serum.
PT	Prothrombin time (seconds) in Blood by Coagulation assay.
APTT	Activated partial thromboplastin time (seconds) in Blood by Coagulation assay.
INR	International normalized ratio in Blood by Coagulation assay.
Glucose	Glucose [mg/dL] in Blood.
TRIG	Triglyceride [mg/dL] in Blood.
HGB	Hemoglobin (%) in Blood.
BMI	Body mass index (kg/m^2).
RSF	Random survival forest.

I. INTRODUCTION

Heart failure, a major cause of death, occurs when the heart is unable to provide sufficient pumping action to maintain blood flow to meet the body's requirements [1]. In developed countries, approximately 2% of adults suffer from heart failure, increasing to 6%–10% for elderly people aged over 65 years [2]. Evaluating a patient's mortality based on a reliable predictive model is an effective method of identifying critical factors related to poor outcomes, and can thus assist clinicians in identifying those in need of intensive monitoring, therapy, or hospice care in ICU cohorts.

Several risk models exist for predicting heart failure mortality [3]–[9]. Most are based on traditional clinical risk factors, such as hypertension, diabetes, and prior cardiovascular diseases. An ecological study was also conducted [10] to identify areas of Brazil where residents were at risk of mortality related to cardiovascular disease. Laboratory information such as blood urea nitrogen (BUN), serum creatine (SCR), leukocytes (WBC), aspartate aminotransferase (AST), alanine aminotransferase (ALT), and total bilirubin indicates comorbid conditions that are difficult for clinicians to manage and thus lead to high mortality rates. These conditions have not been studied comprehensively. Moreover, because most laboratory results are real-valued, their normal ranges are difficult to define to a certain degree. An effective survival model to handle real variables is essential for accurate risk analysis that can identify critical real variables. Traditional proportional hazards used in existing risk models (i.e., the Cox proportional hazard model) are weak at identifying this type of predictors through manual transformation [11].

Machine learning algorithms can automatically reconstruct relationships between variables and response values from big data and thus provide an efficient method of improving the performance of traditional proportional hazard models in identifying critical predictors [12]. Machine learning approaches such as a least squares support vector machine, decision tree, Bayesian network, and association rule have medical applications (e.g., in cardiovascular disease prediction and coronary heart disease detection) [13]–[17]. Survival trees are flexible nonparametric alternatives to parametric or semiparametric models because they can automatically identify certain types of interactions rather than specify them beforehand for time-to-event data. Survival forests combine survival trees and the ensemble method to form a powerful predictive tool. Therefore, a random survival forest (RSF) was proposed as an extension of a random forest [18] for nonparametric survival analysis through the automatic assessment of nonlinear effects and complex interactions among multiple variables [19]. RSF has been used in risk models for various types of diseases, such as heart failure [7] and breast cancer [20]; however, improvements have been limited, as demonstrated in the experimental results. Furthermore, our previous study found that RSF was weak at identifying predictors in relatively small populations [21], [22]. For example, RSF has a limited ability to identify the newer class III of antiarrhythmic agents (which has not been widely used in large-scale population) [21], as its stopping criterion is based on minimum unique deaths. However, we cannot conclude that it is not predictive, because it was demonstrated to be an ideal agent for reducing mortality in patients with heart disease [23].

Therefore, in this paper, we present an improved RSF with a novel split rule and stopping criterion that can identify the discriminative variables separating survivors from non-survivors in a small population. We used the proposed improved RSF to develop a high-accuracy risk model fitted with more laboratory data for predicting hospital mortality and identifying critical risk factors among ICU patients with heart failure.

This paper is organized as follows: Section II describes the proposed iRSF and statistical analysis procedure; Section III presents the experimental results of the proposed approach; Section IV presents a discussion of the results; and Section V presents the conclusions.

II. METHODOLOGY

A. PROPOSED IMPROVED RANDOM SURVIVAL FOREST

RSF was proposed in 2008 for survival data analysis based on random forest. A split rule, such as the log-rank test, which maximizes the survival difference between daughter nodes, was used for each survival tree in the forest [19]. In traditional RSF, each survival tree is grown to full size under the constraint that a terminal node should have no fewer than $d_0 > 0$ unique deaths. However, the log-rank test was shown to be asymptotically optimal under the proportional hazards alternative because of an equal censoring pattern hypothesis in the two groups. In addition, the stopping criterion is arbitrary and demonstrates bias toward predictors with a larger population. This is because it is difficult for predictors with a smaller population to satisfy the criterion, especially when d_0 is large.

Therefore, we proposed an improved RSF (iRSF) with a novel split rule and stopping criterion for identifying more accurate predictors that can separate survivors and non-survivors and thus improve discrimination ability. First, a weighted log-rank test was used to split the node, which was proposed by Yang and Prentice [24] and can be applied to non-proportional hazard situations to improve the test for a range of alternative hypotheses.

We let $d(t)$ be the number of deaths, $Y(t)$ be the individuals at risk, and t be time. The hazard function estimate $H(t)$ at a time t with the Nelson–Aalen estimator can be expressed as:

$$H(t) = \frac{d(t)}{Y(t)}, \quad t <= \tau_0 \tag{1}$$

where $\tau_0 = 365$ in our study. The model proposed in [25] can be used in our study and expressed as:

$$H_R(t) = \frac{\theta_1\theta_2}{\theta_1 + (\theta_2 - \theta_1 Y_L(t))} H_L(t) \tag{2}$$

where $H_R(t)$ and $H_L(t)$ are the respective hazard functions of the right branch and left branch of a grown tree,

$Y_R(t)$ and $Y_L(t)$ are the respective survival functions of the right branch and left branch, $\theta_1 = \lim_{t\downarrow 0} H_R(t)/H_L(t)$, and $\theta_2 = \lim_{t\uparrow \tau_0} H_R(t)/H_L(t)$. A χ^2 test using the two estimating functions of the right and left branches was used by Yang and Prentice to test the hypothesis of significant difference [25].

We let $t_{1,h} < t_{2,h} < \ldots < t_{N(h),h}$ be the N(h) distinct event times in a node of the grown tree. The cumulative hazard function can be expressed as $CH(t) = \sum_{t_{l,h}<=t} H(t)$. $CH_R(t)$ and $CH_L(t)$ are cumulative hazard functions of the right branch and left branch. The relative cumulative hazard function at each distinct event time can be computed using:

$$rH(t) = abs(\log \frac{CH_R(t)}{CH_L(t)}), \quad t_{1,h} <= t <= t_{N(h)} \tag{3}$$

To identify the predictors that can discriminate low risk (with a small cumulative hazard function) and high risk (with a large cumulative hazard function), combined with the fact that high-risk populations have high mortality rates in the short term, the split function was defined as follows:

$$Splitfun = \sum_{t_{1,h}\le t\le t_{N,h}} \frac{rH(t)}{t} t_{N,h} \tag{4}$$

The stopping criterion is defined as the split function decreasing.

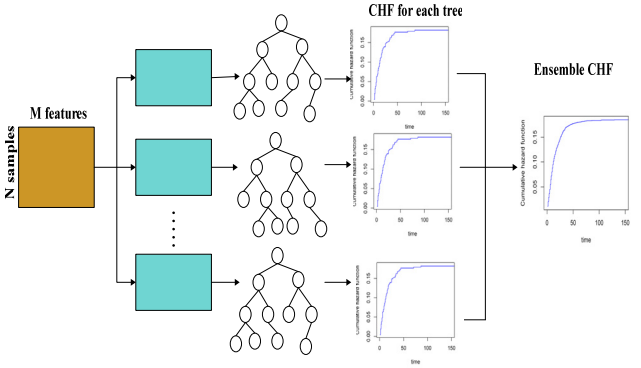


FIGURE 1. Block diagram for developing a risk model based on the proposed iRSF. CHF: cumulative hazard function.

B. STATISTICAL ANALYSIS

As presented in Fig. 1, we developed the risk prediction model based on the proposed iRSF through the following methods:

1. One thousand bootstrap samples were randomly selected from the entire database (N samples, M features). For each bootstrap sample, 37% of the data were excluded for verification (i.e., the out-of-bag (OOB) data).
2. Using each bootstrap sample, a survival tree was grown based on all variables. Then all survival trees were ensembled to develop the risk model. The aforementioned split rule was used to split each node in the tree.
3. Each survival tree was grown to full size under the constraint that the split function presented in formula (4) decreased. The cumulative hazard function was then calculated for each terminal node of the grown tree using Nelson–Aalen estimator.
4. All of the cumulative hazard functions from each tree in the forest were averaged to acquire the ensemble cumulative hazard function. Subsequently, Harrell C-statistics presented in [26] was used to compute the prediction error, whereas the b_{th} value represented the error rate evaluated using the first b trees in the forest. Generally, variables that split branches closed to the tree trunks most frequently were identified as the most crucial variables. Therefore, predictive variables were selected based on the minimal depth of a maximum subtree; specifically, the shortest distance from the tree trunk to the branch level of the maximal tree. Smaller minimal depth meant greater predictive power. Fig. 2 presents an illustration of minimal depth with a single tree grown from the experimental dataset (described in detail in section IIC) as an example. In Fig. 2, due to splitting the tree trunk, BUN has a minimal depth of 0, whereas age has a minimal depth of 1 and AST has a minimal depth of 2. By averaging over the forest, variables that with a minimum depth smaller than mean minimum depth were identified as predictive variables.

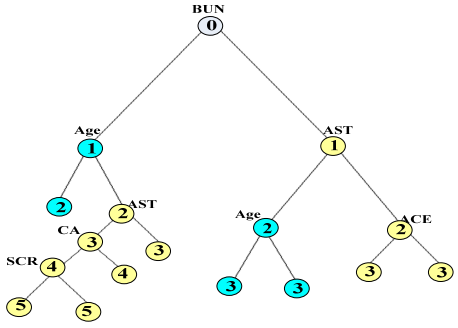


FIGURE 2. Illustration of minimal depth. Blue and yellow points are maximal subtrees for age and AST, respectively. The minimal depth for age is 1 and AST is 2.

To validate the performance of the proposed iRSF-based model, a traditional 1000-tree based RSF with a stopping criterion that unique deaths at the terminal node should be no less than 3 was used for comparison [19]. The discrimination performance of the iRSF-based model was compared with RSF in terms of OOB C-statistics [27].

We conducted our analyses based on R version 3.4.1 (www.R-project.org).

C. EXPERIMENTAL DATA

We used the public MIMIC II database [28], [29] in our study. Beginning in 2001 and lasting over a 7-year period, the data were collected from a variety of ICUs in Boston's Beth Israel Deaconess Medical Center. Clinical records including laboratory results, medications, and the Ninth Revision of the International Classification of Diseases (ICD-9) diagnoses

for 32,536 patients were provided in the database. Patients with heart failure were defined as those with ICD-9 code 428, ∗∗ which encompassed conditions from acute to chronic heart failure. The start point of each patient was defined as the time of ICU admission, whereas the end point was defined as the time of death in hospital or 365-day after the start point. Through these means, 8,059 patients with heart failure were extracted in our study to develop the risk model, including 1,346 patients who died in hospital over a 1-year follow-up period.

Thirty-two variables were assessed for prognostic value: demographics such as age, sex, and body mass index (BMI); clinical variables such as cardiac arrest, hypertension, diabetes, myocardial infarction, cardiac murmur, and atrial fibrillation; laboratory variables such as glucose, blood sodium, blood potassium, SCR, BUN, erythrocytes (RBC), WBC, prothrombin time (PT), activated partial thromboplastin time (APTT), international normalized ratio in blood by coagulation assay (INR), total bilirubin, AST, ALT, hemoglobin, creatine kinase.MB (CKMB), creatine kinase.total (CKPK) and triglyceride; and medications such as beta-blockers, ACE-I, ARB, diuretics, calcium channel antagonists (CCA), and digoxin. The laboratory results were defined as the continuous values measured at the time of ICU admission. If a patient was prescribed and took a type of medication during the ICU stay, the medication was defined as 1, otherwise as 0. Dummy variables were used for categorical variables and log-transformed variables for continuous measurements.

III. EXPERIMENTAL RESULTS

A. BASELINE CHARACTERISTICS OF THE STUDY COHORT

Baseline characteristics during 1-year follow-up of non-survivors and survivors who presented with heart failure are shown in Table 1. During a median follow-up of 1 year, 1,346 individuals died in hospital and 6,713 were censored. We used a two-sample t-test to verify whether the variables differed in non-survivors and survivors, with $p < 0.05$ indicating statistical significance. Table 1 demonstrates that demographics and clinical risk factors such as age, BMI, cardiac arrest, and cardiac murmur were significantly different between live and dead patients, as most of the laboratory results and all of the medications.

TABLE 1. Baseline Characteristics of patients who died in hospital and those who survived during 1-year follow-up. Mean ± SD or counts (%) for real-valued and categorical variables, respectively.

		Non-survivor	Survivor	P-Value
n		1346	6713	
Demographics	Age, years	74.7±13.8	72.8±23.3	<0.001
	Gender, male	727(54%)	3625(54%)	0.1
	BMI, kg/m^2	29.7±10.2	32.2±25.9	<0.001
Clinical variables	Cardiac arrest	175(13%)	201(3%)	<0.001
	Atrial fibrillation	646(48%)	2752(41%)	<0.001
	Myocardial infraction	323(24%)	1544(23%)	0.529
	Cardiac murmur	363(27%)	2484(37%)	<0.001
	Hypertension	310(23%)	1745(26%)	0.006
	Diabetes	269(20%)	1745(26%)	0.001
Laboratory variables	Potassium, mEq/L	4.84±0.95	4.83±0.83	0.759
	Sodium, mEq/L	138.8±4.6	138.5±3.1	0.012
	WBC, K /uL	22.9±31.3	16.5±11.1	<0.001
	RBC, K /uL	4.05±0.62	4.13±0.58	<0.001
	ALT, IU/L	215.7±663	93.9±445	<0.001
	AST, IU/L	418.3±1391	143.2±789	<0.001
	CKMB, IU/L	26.4±77.1	25.4±77.7	0.663
	Triglyceride, mg/dL	126.2±77	126.3±72	0.939
	HGB, mg/dL	12.12±1.67	12.3±1.66	<0.001
	CKPK, IU/L	671±5088	502±2304	0.232
	SCR, mg/dL	2.85±2.03	2.32±2.37	<0.001
	BUN, mg/dL	67.7±37.1	46.9±29.2	<0.001
	Glucose, mg/dL	198.3±88	189.7±72	<0.001
	PT, seconds	23.7±16.9	19.9±11.9	<0.001
	INR	3.52±6.40	2.38±2.80	<0.001
	APTT, seconds	80.78±48.1	68.9±44.6	<0.001
	Total bilirubin, mg/dL	3.49±7.61	1.16±2.18	<0.001
Medications	CCA	188(14%)	1208(18%)	0.001
	Beta-blocker	942(70%)	5169(77%)	<0.001
	ACE-I	363(27%)	3356(50%)	<0.001
	ARB	13(1%)	335(5%)	<0.001
	Diuretic	1090(81%)	5706(85%)	<0.001
	Digoxin	215(16%)	940(14%)	0.04

B. PREDICTOR COMPARISON BETWEEN THE PROPOSED IRSF AND TRADITIONAL RSF

Six randomly selected trees from the 1000-tree forest were presented in Fig. 3. We computed the minimal depth for each variable in each tree and then averaged across the forest to obtain the most predictive variables; specifically, those whose minimal depths were smaller than the mean minimal depth.

The identified predictors based on iRSF and RSF are presented in Fig. 4(a) and 4(b), respectively. Eleven variables were selected as predictive for heart failure mortality with iRSF, including age, AST, ARB, ACE-I, SCR, BUN, ALT, cardiac arrest, WBC, total bilirubin, and triglyceride (the detailed minimal depths of all variables are presented in Fig. 4(a), in which the 11 predictive variables are identified as those below the horizontal line). Another 11 variables were selected as predictive for heart failure mortality with RSF, including cardiac arrest, ACE-I, BUN, total bilirubin, WBC, AST, triglyceride, ALT, age, INR, and SCR (the detailed minimal depths of all variables can be seen in Fig. 4(b)). The selected predictors and their relative predictive powers are slightly different in the two models. ARB, a medication with a high predictive power but fewer individuals from the baseline characteristics description, was not identified as an independent predictor with RSF, as detailed in the Discussion section of this paper. Moreover, compared with previous risk models

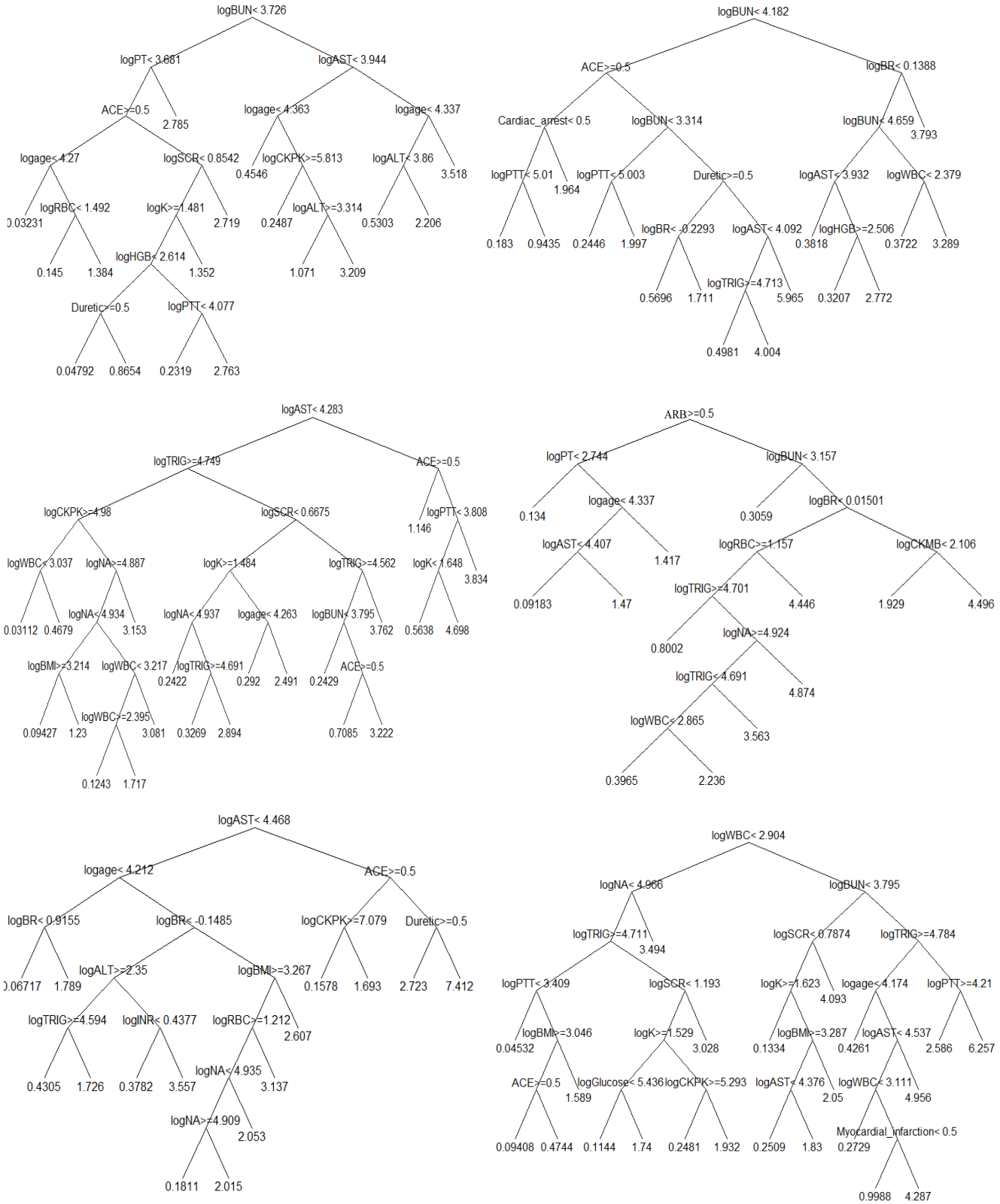


FIGURE 3. Six random trees selected from 1000-tree forest. log* indicates log of index.

for heart failure mortality, more real-valued laboratory variables (including liver indicators that involved AST, ALT, and total bilirubin; renal indicators that involved SCR and BUN; and the inflammation indicator WBC) were demonstrated to have high predictive power for heart failure mortality.

TABLE 2. OOB C-statistics comparison for the proposed iRSF- and RSF-based models.

	iRSF	RSF	P-Value
Comprehensive model	0.821	0.804	<0.01

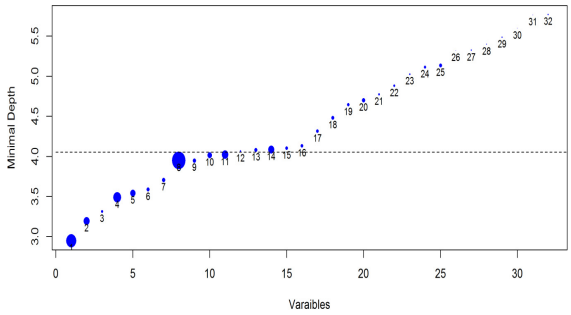
C. DISCRIMINATION PERFORMANCE COMPARISON BETWEEN IRSF- AND RSF-BASED MODELS

The discrimination performance comparison for the two models in our study (iRSF- vs RSF-based) is presented in Table 2. The table reveals that the iRSF-based model improved the discrimination ability to a certain degree with a favorable OOB C-statistics value of 0.821, compared with 0.804 for RSF. Fig. 5 provides the estimated error rates of the iRSF- and RSF-based risk models, using 32 variables for differently grown trees. As shown in Fig. 5, the iRSF-based model has a minimum error rate of 0.179. The error rate decreased as the number of trees increased and became stable when the number of trees exceeded 900. The RSF-based model had a minimum error rate of 0.196. Despite the iRSF-based model exhibiting relatively larger error rates than the RSF-based model when the grown trees were fewer than 50, it decreased sharply as the number of grown trees increased. The one-sided t-test revealed that the iRSF-based error rate was significantly smaller than the RSF-based error rate, with a p value of less than 0.01.

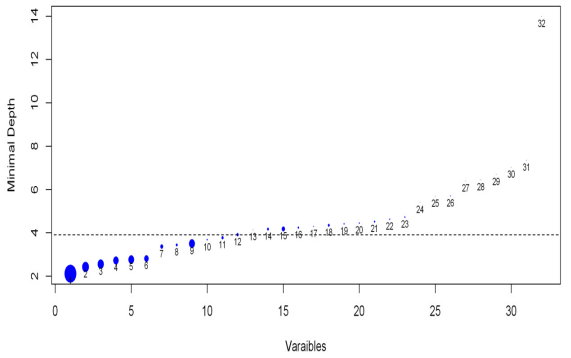
Fig. 6 provides the correlations among the ensemble survival function estimated using iRSF, RSF, and the Kaplan–Meier estimator. Compared with RSF, the estimated survival function using iRSF is closer to the curve of the Kaplan–Meier estimator. We evaluated their proximity in terms of the Euclidean distance between the two curves. The Euclidean distance between the iRSF-estimated survival function and the Kaplan–Meier estimator is 0.00789, compared with 0.0687 for RSF, which demonstrates a superior estimation of survival function.

D. DISCRIMINATION PERFORMANCE COMPARISON BETWEEN THE PROPOSED MODEL AND PREVIOUS MODELS

In our study, a comprehensive risk model for predicting hospital mortality was developed using a 1000-tree iRSF. To demonstrate the effectiveness of the proposed risk model, we compared its performance with previous models based on survival analysis [3]–[7], as presented in Table 3. From the table, we can observe that the proposed risk model is superior to the previous models presented in [3]–[7] in terms of discrimination performance. Although more variables, including echocardiographic and radiographic variables, are used in the



(a) iRSF-based model. .
1. Log of age; 2. Log of AST; 3. ARB; 4.ACE-I; 5. Log of SCR; 6. Log of BUN; 7. Log of ALT; 8. Cardiac arrest; 9. Log of WBC; 10. Log of total bilirubin; 11. Log of triglyceride; 12. Log of PT; 13. Log of BMI; 14. Log of hemoglobin; 15. Log of CKMB; 16. Log of INR; 17. Log of RBC; 18. Log of glucose; 19. Log of CKPK; 20. Log of potassium; 21. Log of sodium; 22. CCA; 23. Log of PTT; 24. Beta-blocker; 25. Diuretic; 26. Cardiac murmur; 27. Diabetes; 28. Digoxin; 29. Myocardial infarction; 30. Hypertension; 31. Atrial fibrillation; 32. Sex.



(b) RSF analysis based model.
1. Cardiac arrest; 2. ACE-I; 3. Log of BUN; 4. Log of total bilirubin; 5. Log of WBC; 6. Log of AST; 7. Log of triglyceride; 8. Log of ALT; 9. Log of Age; 10. Log of INR; 11. Log of SCR; 12. Beta-blocker; 13. Log of PT; 14. Log of BMI; 15. Diuretic; 16. Log of sodium; 17. Log of PTT; 18. Log of potassium; 19. Log of CKMB; 20. Log of Glucose; 21. Log of RBC; 22. Log of hemoglobin; 23. Log of CKPK; 24. Atrial fibrillation; 25. Sex; 26. Cardiac murmur; 27. Myocardial infarction; 28. Diabetes; 29. Hypertension; 30. Digoxin; 31. CCA; 32. ARB.

FIGURE 4. Predictors identified from iRSF and RSF-based models. Y-axis presents the minimal depth of the maximum subtree for specific variables. The horizontal line is the threshold separating predictive variables (below the line) from non-predictive ones. The diameter of each circle in the plot is proportional to the average number of maximal subtrees in the forest for that variable. (a) iRSF-based model. 1. Log of age; 2. Log of AST; 3. ARB; 4.ACE-I; 5. Log of SCR; 6. Log of BUN; 7. Log of ALT; 8. Cardiac arrest; 9. Log of WBC; 10. Log of total bilirubin; 11. Log of triglyceride; 12. Log of PT; 13. Log of BMI; 14. Log of hemoglobin; 15. Log of CKMB; 16. Log of INR; 17. Log of RBC; 18. Log of glucose; 19. Log of CKPK; 20. Log of potassium; 21. Log of sodium; 22. CCA; 23. Log of PTT; 24. Beta-blocker; 25. Diuretic; 26. Cardiac murmur; 27. Diabetes; 28. Digoxin; 29. Myocardial infarction; 30. Hypertension; 31. Atrial fibrillation; 32. Sex. (b) RSF analysis based model. 1. Cardiac arrest; 2. ACE-I; 3. Log of BUN; 4. Log of total bilirubin; 5. Log of WBC; 6. Log of AST; 7. Log of triglyceride; 8. Log of ALT; 9. Log of Age; 10. Log of INR; 11. Log of SCR; 12. Beta-blocker; 13. Log of PT; 14. Log of BMI; 15. Diuretic; 16. Log of sodium; 17. Log of PTT; 18. Log of potassium; 19. Log of CKMB; 20. Log of Glucose; 21. Log of RBC; 22. Log of hemoglobin; 23. Log of CKPK; 24. Atrial fibrillation; 25. Sex; 26. Cardiac murmur; 27. Myocardial infarction; 28. Diabetes; 29. Hypertension; 30. Digoxin; 31. CCA; 32. ARB.

TABLE 3. Performance comparison between the proposed model and previous models.

	Outcome	Variables	Statistical method	Performance evaluation method	Performance
The proposed model	**All-cause mortality**	**Demographics and clinical variables, laboratory variables, medications**	**iRSF**	**OOB C-statistics**	**0.821**
MAGGIC study [3]	All-cause mortality	Demographics and clinical variables, laboratory variables, medications	Multivariable Poisson regression methods	Goodness-of-fit	
MUSIC risk score [4]	All-cause mortality	Demographic and clinical variables, Radiographic variables, Laboratory variables, Echocardiographic variables, 12-Lead ECG and 24-h Holter monitoring variables, medications	Multivariable Cox model	OOB C-index	0.76
SENIORS study [5]	All-cause mortality	Demographics and clinical variables, laboratory variables, medications	Multivariable Cox model	C-statistics	0.72
Seattle heart failure model [6]	All-cause mortality	Demographics and clinical variables, laboratory variables, medications	Multivariable Cox model	AUC	0.729
RSF-based heart failure model [7]	All-cause mortality	Demographics and clinical variables, laboratory variables, medications	RSF	OOB C-statistics	0.705

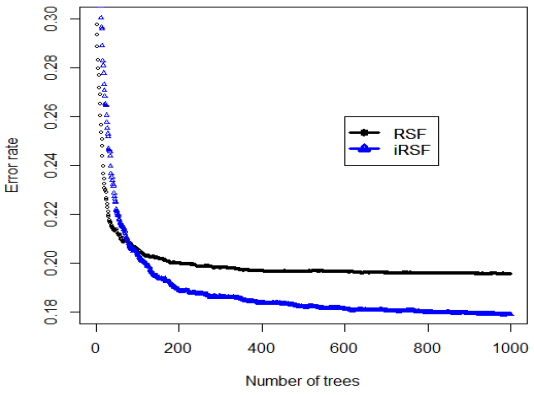


FIGURE 5. Estimated error rate comparison for iRSF- and RSF-based models with differently grown trees.

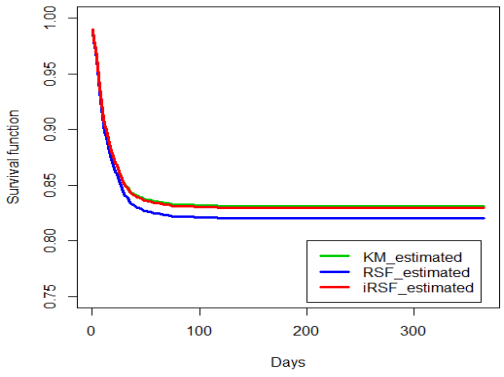


FIGURE 6. Ensemble survival function comparison of iRSF, RSF, and Kaplan–Meier estimator.

MUSIC risk score, the performance is as limited as the Cox hazard model and a small amount of laboratory variables are used. Similar to the Meta-Analysis Global Group in Chronic Heart Failure (MAGGIC) study, in which many laboratory variables such as AST, ALT, and BUN were not merged. Overall, our proposed model exhibited superior performance compared with previous risk models because of its inclusion of a greater number of laboratory variables to reflect comorbidities and its use of the new iRSF-based survival model.

IV. DISCUSSION

A. REAL-VALUED PREDICTORS IDENTIFIED FOR HEART FAILURE MORTALITY

In our study, 17 laboratory risk factors, including BUN, SCR, AST, ALT, and WBC, were studied to predict heart failure mortality. An iRSF-based survival model was proposed to objectively identify the critical real-valued predictors and their nonlinear effects on mortality. Age, AST, SCR, BUN, ALT, WBC, total bilirubin, and triglyceride were identified as the eight crucial real-valued predictors with iRSF. Compared with previous studies, liver indicators including AST, ALT, and total bilirubin, renal indicators including SCR and BUN, and the inflammation indicator WBC were demonstrated to have high predictive power for heart failure mortality. Their estimated effects on the ensemble survival function are presented in Fig. 7 with red lines. Their Kaplan–Meier estimated survival functions are also provided in Fig. 7 with yellow lines for comparison; the estimated survival functions closely conform to the Kaplan–Meier curves. In other words, the proposed iRSF is an effective method to reproduce the inherent relationship between the predictors and the events.

Fig. 7 suggests that different predictors exert different effects on survival rate. For example, BUN, an indicator for renal diseases, which was demonstrated to be predictive for heart failure mortality [30], was found to decrease survival sharply with its elevation, independent of the normal range. However, for SCR, another renal function indicator, survival

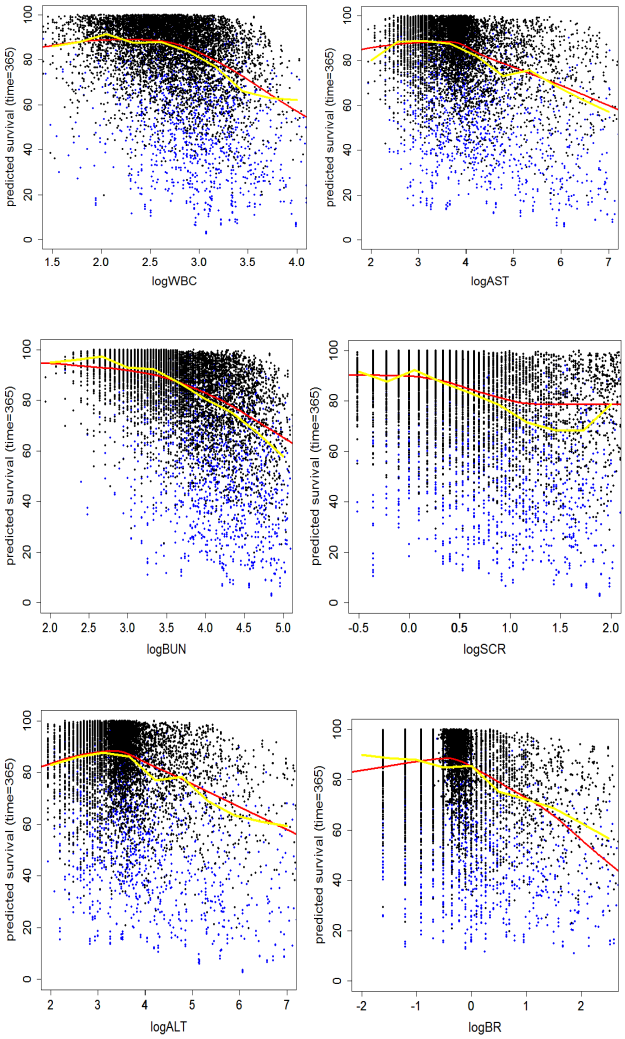


FIGURE 7. Ensemble survivals and Kaplan–Meier estimates plotted against the six newly identified critical continuous predictors: WBC, AST, BUN, SCR, ALT, and total bilirubin. Red line represents predicted survival based on the proposed iRSF, while yellow line represents Kaplan–Meier estimated survival. Each blue point corresponds to each event and black point corresponds to censored observation. log* = log of index. BR = total bilirubin. The Y-axis presents the predicted survival function based on the proposed iRSF. Log*: log of ∗.

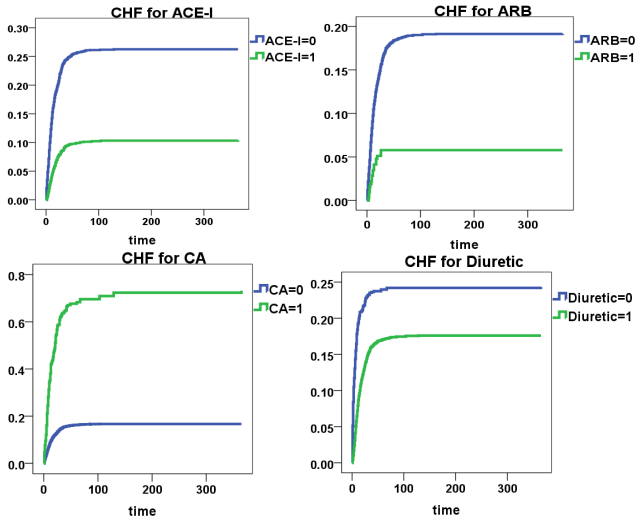


FIGURE 8. Cumulative hazard function computed from the Nelson–Aalen estimator for four typical categorical variables: ACE-I, ARB, cardiac arrest, and diuretics. The Y-axis presents the cumulative hazard function (CHF). Blue line represents the CHF for the subjects without the corresponding variable, while green line presents CHF for the subjects with the corresponding variable. CA = cardiac arrest.

was stable in the normal range and then began to decrease beyond it. WBC, a direct indicator of inflammation, was demonstrated to be an independent predictor of heart failure mortality in the Seattle heart failure model [6]. In our study results, WBC demonstrated its nonlinear effect on predicting heart failure mortality; specifically, values that were too far outside the normal range led to low survival rates. In addition, AST and ALT, which are commonly used as indicators of liver disease, were identified as valuable predictors for heart failure mortality in our study. Fig. 7 suggests that the survival rate decreases sharply when AST and ALT exceed 40 IU/L. Several mechanisms may exist through which AST and ALT are associated with an increased risk of death. First, serum levels of AST and ALT have been demonstrated to be associated with future mortality in community residents [31]. Second, AST and ALT might also be markers of cardiovascular diseases and stroke [32], potentially aggravating the risk of death. Third, AST and ALT elevations may reflect another serious comorbid condition that increases the risk of death for patients with heart failure. Bilirubin is a metabolic byproduct of the breakdown of hemoglobin degradation, which itself must be metabolized for appropriate excretion. Although in previous studies total bilirubin has been shown to correlate inversely with cardiovascular disease in high-risk patients [33], [34], serum bilirubin appears to be a sensitive indicator of liver function and thus has been verified to represent a significant risk of death or renal failure in patients with spontaneous bacterial peritonitis [35], which was significantly greater in patients who died owing to postoperative complications than in the patients who survived [36]. As for patients with heart failure, a high level of bilirubin implied the comorbidities of heart failure and renal failure and thus was a valuable predictor of short-term mortality.

This analysis indicates that the proposed iRSF-based model can automatically identify the most crucial continuous predictors and their inherent effects on events. We recommend monitoring the following laboratory indicators to accurately predict which patients are at high risk: AST, BUN, SCR, WBC, ALT, total bilirubin, and triglyceride.

B. CATEGORICAL PREDICTORS IDENTIFIED FOR HEART FAILURE MORTALITY

The comprehensive model with iRSF identified ARB, ACE-I, and cardiac arrest as three predictive categorical factors of survival in the cohort of 8,059 ICU patients with heart failure, whereas ARB was not identified using the

RSF-based model. Fig. 8 provides the cumulative hazard function for the three variables and one nonpredictive variable (diuretics) for comparison. We can observe that ACE-I, ARB, and cardiac arrest can separate the survival functions for subjects with and without corresponding variables (blue line and green line in Fig. 8) to a greater extent, whereas the predictive ability of diuretics is limited, especially in the short term. The results verified the findings of the iRSF-based model.

ACE-Is are the first line of therapy for patients with heart failure, improving survival rates and quality of life. They have been shown to reduce mortality in patients with left ventricular dysfunction in numerous randomized trials [37]. ARBs, which are often used when patients are intolerant of the adverse effects produced by ACE-Is, may be useful because they act to prevent the action of angiotensin II at the AT1 receptor, leaving the AT2 receptor unblocked. Therefore, they were found to be a useful medication in reducing heart failure mortality in our study. In addition, the effectiveness of beta-blockers and diuretics was determined to be limited in improving short-term survival, even though a trial showed that beta-blockers reduced the absolute risk of death by 4.5% over a 13-month period [38]. Although diuretics are widely used, evidence on their efficacy and safety is limited [39].

V. CONCLUSIONS AND FUTURE WORK

In this study, we developed a prognostic model with favorable discrimination performance based on 32 risk factors, including demographics, clinical information, laboratory information, and medications, for predicting the hospital mortality of ICU patients with heart failure using an iRSF. The iRSF aimed to identify the predictors that could separate high-risk and low-risk individuals, especially for real-valued variables and variables in small populations. This study identified 11 independent predictors of mortality in heart failure. Compared with existing models, more laboratory factors that could reflect comorbidities were merged into the risk model and presented to have great predictive power, including the liver indicators AST, ALT, and total bilirubin, and the renal indicators SCR and BUN. ARB, which is effective only in small populations, was found to be an independent predictor with the proposed iRSF and thus demonstrated its performance in identifying critical predictors. As a result, the model separated the 1-year survivors and non-survivors with much greater accuracy than previous heart failure models, returning an OOB C-statistics value of 0.821. Moreover, the proposed iRSF demonstrated a more accurate estimation of the overall survival function compared with traditional RSF in our study.

In the future, we will merge physiological signals such as electrocardiograph, photoplethysmograph, and blood pressure variation to update and improve the performance of the risk model. In addition, we will study a simplified risk model with few variables for predicting heart failure mortality, to be widely applied in different scenarios, such as ICU cohorts or monitoring settings in the home.

ACKNOWLEDGMENT

The authors would like to thank Prof. Constantinos S. Pattichis from University of Cyprus for kind suggestions on improving this manuscript.

REFERENCES

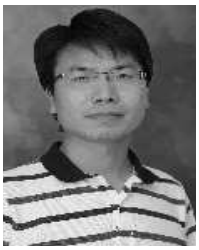
[1] H. Denolin, H. Kuhn, H. Krayenbuehl, F. Loogen, and A. Reale, "The defintion of heart failure," *Eur. Heart J.*, vol. 4, no. 7, pp. 445–448, 1983.
[2] J. J. McMurray and M. A. Pfeffer, "Heart failure," *Lancet*, vol. 365, no. 9474, pp. 1877–1889, 2005.
[3] S. J. Pocock *et al.*, "Predicting survival in heart failure: A risk score based on 39 372 patients from 30 studies," *Eur. Heart J.*, vol. 34, no. 9, pp. 1404–1413, 2013.
[4] R. Vazquez *et al.*, "The MUSIC Risk score: A simple method for predicting mortality in ambulatory patients with chronic heart failure," *Eur. Heart J.*, vol. 30, no. 9, pp. 1088–1096, 2009.
[5] L. Manzano *et al.*, "Predictors of clinical outcomes in elderly patients with heart failure," *Eur. J. Heart Failure*, vol. 13, no. 5, pp. 528–536, 2011.
[6] W. C. Levy *et al.*, "The Seattle heart failure model: Prediction of survival in heart failure," *Circulation*, vol. 113, no. 11, pp. 1424–1433, 2006.
[7] E. Hsich *et al.*, "Identifying important risk factors for survival in patient with systolic heart failure using random survival forests," *Circulat., Cardiovascular Quality Outcomes*, vol. 4, no. 1, pp. 39–45, 2011.
[8] M. Senni *et al.*, "Predicting heart failure outcome from cardiac and comorbid conditions: The 3C-HF score," *Int. J. Cardiol.*, vol. 163, no. 2, pp. 206–211, 2013.
[9] P. C. Austin, J. V. Tu, and D. S. Lee, "Logistic regression had superior performance compared with regression trees for predicting in-hospital mortality in patients hospitalized with heart failure," *J. Clin. Epidemiol.*, vol. 63, no. 10, pp. 1145–1155, 2010.
[10] P. C. O. Rodrigues, E. S. Santos, E. Ignotti, and S. D. Hacon, "Space-time analysis to identify areas at risk of mortality from cardiovascular disease," *BioMed Res. Int.*, vol. 2015, no. 4, Aug. 2015, Art. no. 841645.
[11] L. Breiman, "Heuristics of instability and stabilization in model selection," *Ann Stat.*, vol. 24, no. 6, pp. 2350–2383, 1996.
[12] T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 2003.
[13] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. 5, pp. 8869–8879, 2017.
[14] B. J. Lee and J. Y. Kim, "Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning," *IEEE J. Biomed. Health Inform.*, vol. 20, no. 1, pp. 39–46, Jan. 2016.
[15] S. Bandyopadhyay *et al.*, "Data mining for censored time-to-event data: A Bayesian network model for predicting cardiovascular risk from electronic health record data," *Data Mining Knowl. Discovery.*, vol. 29, no. 4, pp. 1033–1069, 2015.
[16] U. R. Acharya *et al.*, "An integrated index for detection of Sudden Cardiac Death using Discrete Wavelet Transform and nonlinear features," *Knowl.-Based Syst.*, vol. 83, pp. 149–158, Jul. 2015.
[17] G. J. Simon, P. J. Caraballo, T. M. Therneau, S. S. Cha, M. R. Castro, and P. W. Li, "Extending association rule summarization techniques to assess risk of diabetes mellitus," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 1, pp. 130–141, Jan. 2015.
[18] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[19] H. Ishwaran *et al.*, "Random survival forests," *Ann. Appl. Stat.*, vol. 2, no. 3, pp. 841–860, 2008.
[20] I. K. Omurlu, M. Ture, and F. Tokatli, "The comparisons of random survival forests and Cox regression analysis with simulation and an application related to breast cancer," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 8582–8588, 2009.
[21] F. Miao, Y. P. Cai, Y. T. Zhang, and C. Y. Li, "Is random survival forest an alternative to Cox proportional model on predicting cardiovascular disease?" in *Proc. 6th Eur. Conf. Int. Fed. Med. Biol. Eng.*, 2015, pp. 740–743.
[22] F. Miao, Y.-P. Cai, Y.-X. Zhang, Y. Li, and Y.-T. Zhang, "Risk prediction of one-year mortality in patients with cardiac arrhythmias using random survival forest," *Comput. Math. Methods Med.*, vol. 2015, Jul. 2015, Art. no. 303250, doi: 10.1155/2015/303250.

[23] J. W. Mason, "A comparison of seven antiarrhythmic drugs in patients with ventricular tachyarrhythmias," *New England J. Med.*, vol. 329, no. 7, pp. 452–458, 1993.
[24] S. Yang and R. Prentice, "Semiparametric analysis of short-term and long-term hazard ratios with two-sample survival data," *Biometrika*, vol. 92, pp. 1–17, Mar. 2005.
[25] S. Yang and R. Prentice, "Improved logrank-type tests for survival data using adaptive weights," *Biometrics*, vol. 66, no. 1, pp. 30–38, 2010.
[26] E. W. Steyerberg *et al.*, "Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis," *J. Clin. Epidemiol.*, vol. 54, no. 8, pp. 774–781, 2001.
[27] H. Uno *et al.*, "On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data," *Stat. Med.*, vol. 30, no. 10, pp. 1105–1117, 2011.
[28] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. E215–E220, 2000.
[29] M. Saeed *et al.*, "Multiparameter intelligent monitoring in intensive care II (MIMIC-II): A public-access intensive care unit database," *Critical Care Med.*, vol. 39, no. 5, pp. 952–960, 2011.
[30] P. N. Peterson *et al.*, "A validated risk score for in-hospital mortality in patients with heart failure from the American Heart Association get with the guidelines program," *Circulat., Cardiovascular Quality Outcomes*, vol. 3, no. 1, pp. 25–32, 2010.
[31] T. H. Lee, W. Kim, J. T. Benson, T. M. Therneau, and L. J. Melton, "Serum aminotransferase activity and mortality risk in a United States community," *Hepatology*, vol. 47, no. 3, pp. 880–887, 2008.
[32] H. C. Kim *et al.*, "Elevated serum aminotransferase level as a predictor of intracerebral hemorrhage: Korea medical insurance corporation study," *Stroke*, vol. 36, no. 8, pp. 1642–1647, 2005.
[33] J. P. Lin *et al.*, "Association between the UGT1A1*28 allele, bilirubin levels, and coronary heart disease in the Framingham Heart Study," *Circulation*, vol. 114, no. 14, pp. 1476–1481, 2006.
[34] A. Nguyen *et al.*, "Total bilirubin is an independent risk factor for the prevalence of coronary artery disease in men," *Circulation*, vol. 130, no. Suppl 2, p. A16164, 2014.
[35] R. Terg *et al.*, "Serum creatinine and bilirubin predict renal failure and mortality in patients with spontaneous bacterial peritonitis: A retrospective study," *Liver Int.*, vol. 29, no. 3, pp. 415–419, 2009.
[36] Y. Yokoyama *et al.*, "Predictive power of prothrombin time and serum total bilirubin for postoperative mortality after major hepatectomy with extrahepatic bile duct resection," *Surgery*, vol. 155, no. 3, pp. 504–511, 2014.
[37] T. S. Investigators, "Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure," *New England J. Med.*, vol. 325, no. 5, pp. 293–302, 1991.
[38] A. M. Pritchett and M. M. Redfield, "β-blockers: New standard therapy for heart failure," *Mayo Clinic Proceedings*, vol. 77, no. 8, pp. 839–846, 2002.
[39] T. G. Lueder, D. Atar, and H. Krum, "Diuretic use in heart failure and outcomes," *Clin. Pharmacol. Therapeutics*, vol. 94, no. 4, pp. 490–498, 2013.

FEN MIAO received the Ph.D. degree in computer science from the University of Chinese Academy of Sciences in 2015. She is currently an Assistant Researcher with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. She is also presiding the National Natural Science Foundation of China and the Basic Research Foundation of Shenzhen. She has published nearly 20 SCI/EI indexed papers in recent five years. Her research interests include health data analysis, cardiovascular disease modeling, and cuff-less blood pressure estimation.

YUN-PENG CAI is currently an Associate Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. He is also the Executive Director of Guandong Engineering Center and Shenzhen Engineering Laboratory for Health Big Data Analysis. He has published over 60 peer-reviewed papers, including multiple publications on top journals, such as *Nucleic Acids Research*, *Briefings in Bioinformatics*, *The ISME Journal*, and *PLOS Computational Biology*. He is the PI or Co-PI of various national or regional funds, including the NSFC fund, the National Key Technology R&D Program, the National High-Tech R&D Program, the CAS Science and Technology Service Program, and the Guangdong Applied Technology R&D Program. His research interests include health big data, health informatics, bioinformatics, machine learning, and evolution.

YU-XIAO ZHANG received the Doctorate of Medicine degree from the Medical College of People's Liberation Army. She has been involved in clinical medicine for 20 years. She is currently the Chief Physician with the Department of Senile Cardiovascular Medicine, The General Hospital of the People's Liberation Army. She is an Expert in cardiovascular diseases treatment, such as hypertension, heart failure, and coronary heart disease. She has published over 40 peer-reviewed papers.



XIAO-MAO FAN received the bachelor's degree in chemistry from Nanchang University, Nanchang, China, in 2003, and the master's degree in system engineering from the Shanghai University of Science and Technology, Shanghai, China, in 2011. He is currently pursuing the Ph.D. degree in computer application technology from the University of Chinese Academy of Sciences. His current research interests are in machine learning and high performance computing.

YE LI received the Ph.D. degree from Arizona State University in 2002. He is currently a Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. He has applied for more than 90 related patents and software copyrights, more than 40 of which have been granted and many successfully achieved industrialization. He has published over 120 SCI/EI indexed journal or international conference papers in the fields of mobile medical and health management and biomedical information technology in the last five years, over 30 papers were indexed by SCI, and some were cited by journals of important influence, such as *Nature Review*. He has served as an Organizing Member/Workshop Chairman for many times in the important IEEE biomedical engineering related international conference for many times.