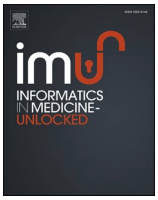
Informatics in Medicine Unlocked 26 (2021) 100772

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu



Survival prediction of heart failure patients using machine learning techniques

Asif Newaz *, Nadim Ahmed, Farhan Shahriyar Haq

Department of Electrical and Electronic Engineering, Islamic University of Technology, Gazipur, Bangladesh

ARTICLE INFO

Keywords:
Machine learning
Random forest
Recursive feature elimination
Heart failure
Ejection fraction
Imbalanced classification

ABSTRACT

The goal of this research is to develop a reliable decision-support system for the survival prediction of heart failure patients by utilizing their clinical records and laboratory test results. Forecasting heart-failure related events in clinical practice tend to be quite inaccurate and highly variable. Identifying the key drivers of heart failure is also clinically very important. In this regard, we develop a model to accurately identify the patients who are at risk utilizing machine learning techniques. This can help clinicians make informed decisions regarding the intensity of treatment required for a patient. For this study, we have utilized a heart failure dataset originally collected from the Faisalabad Institute of Cardiology and the Allied Hospital in Faisalabad, Pakistan. Sampling strategy is incorporated into the ensemble learning framework to develop a more robust Random Forest Classifier that can effectively deal with the imbalanced nature of the data and provide a more generalizable result with higher accuracy. Two different feature selection techniques - Chi-square test and Recursive Feature Elimination are utilized to identify the features that are most significant in terms of survival prediction of heart failure patients. Using our proposed approach, a maximum G-mean score of 76.83% with a sensitivity score of 80.21% was achieved, which is significantly higher than what has been reported by other researchers. Thus, our proposed framework has the potential to be an effective tool to identify the patients who are at risk and guide clinicians accordingly to take pertinent measures.

1. Introduction

Heart failure (HF) is a condition that occurs when the heart is unable to pump enough blood to the body, and it is usually caused by chronic conditions such as coronary heart disease, high blood pressure, or other heart conditions or diseases [1]. The number of patients with HF worldwide has increased drastically, moving from 33.5 million in 1990 to a staggering 64.3 million in 2017 [2]. Approximately 6 million people over the age of 20 in the United States have HF, with approximately 1 million new cases diagnosed annually, a number that continues to rise [3], and the total cost of HF treatment exceeds $30 billion in the United States annually [4].

Fast diagnosis and risk assessment are crucial to providing cost-effective and timely care for HF patients [5]. Understanding expected risks and communicating anticipated future disease trajectories to patients and their families constitute important aspects of patient-physician interactions in heart failure [6,7]. Knowledge of future risks can help clinicians make informed decisions regarding the intensity of treatment or providing end-of-life care to patients [8]. On the other hand, identifying low-risk patients could also reduce patient anxiety and the additional cost of treatment. However, how to best estimate the risk in patients is less clear. A number of tools are available to assess the risk of HF in patients such as biomarkers [9], risk scores [10], and their combination [11]. However, conventional risk prediction strategies for HF are only able to provide modest predictive power [12]. Given the importance of a vital organ like the heart, predicting heart failure has become a priority for clinicians, however to date forecasting heart failure-related events in clinical practice usually has failed to reach high accuracy [13,14]. The complex nature of HF produces a significant amount of information that is too difficult for clinicians to process as it requires simultaneous consideration of multiple factors and their interactions. Artificial intelligence and machine learning techniques can be utilized in this scenario to develop a reliable decision support system to assist clinicians in properly interpreting the patients' records to make informed decisions.

The purpose of this study is not only to develop an accurate survival prediction model but also to discover essential factors for the survival prediction of heart failure patients. In that regard, we utilized a dataset

* Corresponding author. 1704, Bangladesh.
E-mail addresses: asifnewaz@iut-dhaka.edu (A. Newaz), nadimahmed@iut-dhaka.edu (N. Ahmed), farhanshahriyar@iut-dhaka.edu (F. Shahriyar Haq).

https://doi.org/10.1016/j.imu.2021.100772
Received 18 August 2021; Received in revised form 7 October 2021; Accepted 22 October 2021
Available online 23 October 2021

from the UCI repository that contained the medical records of 299 heart failure patients. The data was originally collected from the Faisalabad Institute of Cardiology, Pakistan, and the Allied Hospital in Faisalabad, Pakistan in 2015 [6]. The dataset contains a total of 11 covariates and the target variable - whether the patient survived or not. Out of the 299 patients, a total number of 203 patients survived the study period while 96 patients died during the study period. So, one class is relatively underrepresented compared to the other one. This imbalance present in the data is quite common for medical datasets, however, it creates some complications in the prediction task [15]. Standard classification algorithms have an accuracy-oriented design. They are built based on the assumption of an equal number of instances in each class. The classifiers are trained to minimize the overall number of wrong predictions, irrespective of the class. As a result, the prediction gets biased towards the majority class. The minority class on the other hand suffers from a high misclassification rate. Even if a classifier is misclassifying most of the minority class samples, it'll assume that it is performing well as long as the number of wrong predictions is minimum. The classifier is inherently unable to distinguish between the misclassifications of majority and minority class instances. Therefore, some appropriate measures need to be adopted to mitigate the bias introduced by the imbalanced nature of the data.

Ensemble learning is a machine learning paradigm where multiple models called 'weak learners' are combined to enhance the performance of any single one of them individually. Weak learners generally do not perform very well by themselves. They often suffer from high bias or high variance. When a group of weak learners is properly combined, a more robust model can be obtained. The aggregated model tends to provide better predictive performance with reduced bias and variance.

Decision Trees are one of the most popular classification algorithms due to their interpretability and easy implementation [16]. However, they are very prone to overfitting and easily biased if the data is imbalanced. To overcome the drawbacks of the decision trees, an ensemble of trees can be formed. Random Forest (RF) is one such ensemble technique that combines the simplicity of decision trees with the flexibility and power of an ensemble model [17]. In the RF classifier, a forest of N decision trees is built where each tree is trained using a random sample from the dataset. This process is called 'Bootstrapping'. Training each tree on a different set of samples reduces the overfitting problem. Moreover, a random subset of features is evaluated at each node during splitting. This generates a set of un-correlated trees. Finally, the decisions made by each tree are aggregated to get a prediction. This way, RF combines the process of bagging or bootstrap sampling and random feature subspace selection to create a more robust prediction model. Thus, it is usually able to provide a more generalizable and accurate performance over other classification algorithms. However, the model still remains susceptible to class imbalance. Hence, some modifications need to be brought into the classifier design to alleviate the class imbalance problem.

Data sampling is a standard technique in imbalanced learning [18]. This refers to rebalancing the dataset using certain techniques to produce a more balanced data distribution. This can be done by undersampling the majority class instances or by oversampling the minority class instances. The classifier is trained on the balanced resampled dataset allowing it to make predictions in a similar manner to standard classification algorithms. The sampling is generally performed in the data preprocessing step. Then the resampled data is used in synergy with other machine learning algorithms. Many different techniques have been proposed by researchers to perform sampling [19–21]. SMOTE, which stands for Synthetic Minority Oversampling Technique, is one of the most popular oversampling techniques [22]. It generates synthetic samples using interpolation to balance the dataset. CNN (Condensed Nearest Neighbors), ENN (Edited Nearest Neighbors) are some popular undersampling methods [21]. However, these techniques perform oversampling using the entire minority class or remove examples from the entire majority class to produce a balanced distribution in the data. The generation of a large number of synthetic examples results in overfitting of the minority class. On the other hand, eliminating examples from the majority class this way results in loss of information as a large part of the majority class is not used in the training process.

In this study, we propose to merge the undersampling process into the model-building step of the RF classifier [23]. In a standard RF classifier, a bootstrap sample is drawn from the dataset to train each tree in the forest. Some modifications are introduced in the process to address the class imbalance scenario. First, Stratified K-Fold cross-validation technique was utilized to preserve the percentage of samples for each class on each training and testing fold. Next, a bootstrap sample i.e. a random subset of samples with replacement is drawn from the training dataset. Then undersampling was performed on the majority class to keep the same number of instances from both classes in the data. This constitutes the training set for a particular tree. The process is randomized and repeated for all the trees in the forest. This way the tress can be trained on balanced bootstrapped data. This averts the problem with class imbalance while also mitigating the effect of information loss due to undersampling. Thus, our proposed methodology is able to provide a balanced predictive performance with high accuracy on both the majority and minority classes.

Another important aspect of this study is to identify the key features for the survival prediction of heart failure patients. Identifying the features that are most significant in the prediction of survival can better guide clinicians in the decision-making process. Moreover, it also reduces the number of laboratory tests required for the prediction task. In this regard, two different feature selection techniques were utilized to properly assess the importance of the attributes. Chi-square test [24], which is a popular method used in statistics to test the independence of two events, is utilized in this study to identify the features which are highly dependent on the response variable. Recursive Feature Elimination (RFE), which is a popular wrapper-based approach, was utilized to identify a subset of features that can optimize the prediction performance for the proposed balanced RF classifier. By using only 3 selected features, the classifier was able to make more accurate predictions compared to using all the features in the original dataset.

2. Literature review

Several researchers have analyzed ML techniques as a tool to improve survival prediction in HF patients. The majority of these researches are aimed at identifying the primary risk factors that influence the likelihood of mortality in heart failure patients. Panahiazar et al. [25] compared several ML models with the Seattle Heart Failure Model (SHFM) [26] for survival prediction in patients with Heart Failure with reduced Ejection Fraction (HFrEF). Using EHR data, their model was able to improve AUC by 11% over the SHFM for predicting 1, 2, and 5-year survival. Ahmad et al. [6] applied traditional biostatistics time-dependent models such as Cox regression and Kaplan-Meier survival plots to predict mortality in 299 patients with HF condition who were hospitalized at the Institute of Cardiology and Allied Hospital in Faisalabad, Pakistan. They made their dataset publicly available online [27]. Their study was particularly focused on estimating the death rates of heart failure patients and the major factors responsible for increased risk of mortality among heart failure patients. They identified growing age, renal dysfunction, blood pressure, anemia, and ejection fraction as the prominent risk factors. Following this, Zahid et al. [28] used the same dataset to propose two different survival prediction models based on gender. They claimed in their study that the survival prediction models and top risk factors for male and female patients with heart failure varied significantly. However, it needs to be assessed on a larger population to corroborate the generalizability of their findings. Chicco et al. [13] applied univariate statistical analysis to rank the most relevant factors contributing to a higher risk of death in heart failure patients. They provided empirical evidence that only two features, serum creatinine and ejection fraction, are somewhat sufficient to predict the

survival of heart failure patients. In their study, using all features, the best results were obtained using RF classifier with an MCC score of 38.4% and a sensitivity score of only 49.1%. When the same classification technique was performed using only two features (serum creatinine and ejection fraction), the MCC score improved to 41.8% but the accuracy was only 58.5%. They obtained a maximum sensitivity score of 54.1% with a specificity score of 85.5%. There is a clear disparity present in the performance of their proposed approach. This is due to the imbalanced nature of the dataset which was not considered in their study. Moreover, the sensitivity score is only 50%, meaning that almost 50% of the time, their model is failing to predict the risk in patients. However, predicting the patients at risk is very important to initiate the treatment in time and provide the necessary care. Therefore, having a low sensitivity is unenviable for a reliable decision support system. To address the class imbalance issue, Kim et al. [29] proposed the use of 3 different oversampling approaches - SMOTE, Borderline-SMOTE, and ADASYN. Utilizing oversampling techniques improved the sensitivity score and the highest sensitivity of 71.23% was achieved with the SMOTE algorithm. In another literature, Hasan et al. [30] utilized two different feature selection techniques - Minimum Redundancy Maximum Relevance (MRMR), and Recursive Feature Elimination (RFE) to identify the most relevant features. They reported the highest g-mean score of 69.52% using only two features with the Decision Tree classifier.

In this study, we try to overcome the shortcomings of the above-mentioned studies and develop a more reliable and robust decision-support system for the survival prediction of heart failure patients. With our proposed approach, maximum sensitivity of 80.21% with a specificity of 74.45% was achieved using only 3 selected features out of 11 from the original dataset utilizing feature selection techniques. The classification accuracy obtained was 76.25%. Thus, our proposed method far outperforms the model presented by Chicco et al. [13] and others. Our model has the ability to provide a more balanced prediction performance and help identify the patients at risk more accurately.

3. Materials and methods

3.1. Dataset description

The dataset was originally collected from the Faisalabad Institute of Cardiology, Pakistan, and the Allied Hospital in Faisalabad, Pakistan back in 2015 [6]. It is available in the UCI Machine Learning Repository [27]. It contains the medical records of 299 heart failure patients. No missing entries were found. All patients had left ventricular systolic dysfunction and had suffered from previous instances of heart failure that placed them in classes III or IV of the New York Heart Association (NYHA) classification of the stages of heart failure [31]. The original dataset contains a total of 11 features, a time variable (follow-up period, 130 days on average), and the response variable (death event). The time variable was excluded as the focus of the study is to identify the patients who are at risk based on their clinical features. The target variable is whether a patient will survive or not. It has been reported by the donors that all legal factors were considered when collecting the data. A description of the dataset is provided in Table 1.

Table 1
Description of the dataset.

Feature	Description	Statistics
Age	Age of Patient	Range: 40–95 Mean = 60.834 years
Anemia	Absence or presence of Anemia	0 = absence (170 patients) 1 = presence (129 patients)
Creatine Phosphokinase (CPK)	Level of CPK enzyme in blood in mcg/L	Range: 23–7861 Mean = 581.839
Diabetes	Absence/presence of diabetes	0 = absence (174 patients) 1 = presence (125 patients)
Ejection Fraction	Percentage of blood leaving the heart at each contraction	Range: 14–80 Mean = 38.084
High Blood Pressure	Absence/presence of hypertension	0 = absence (194 patients) 1 = presence (105 patients)
Platelets	Platelets in the blood in kiloplatelets/mL	Range: 25.01–850.00 Mean = 263.358
Serum Creatinine	Level of creatinine in blood in mg/dL	Range: 0.50–9.40 Mean = 1.394
Serum Sodium	Level of sodium in blood in mEq/L	Range: 114–148 Mean = 136.625
Sex	Sex of patient	0 = female (105 patients) 1 = male (194 patients)
Smoking	whether the patient has a smoking habit or not	0 = false (203 patients) 1 = true (96 patients)
Target	whether the patient survived or died during follow up period	0 = survived (203 patients) 1 = deceased (96 patients)

3.2. Proposed methodology

The dataset was first split into training and test set using stratified 5-fold cross-validation scheme. This splitting strategy randomly splits the data into 5 folds by preserving the percentage of samples for each class on each training and testing folds. This ensures that each fold has a similar distribution of data as was in the original dataset. Next, our proposed Balanced Random Forest Classifier (BRF) was trained on the training set. The process is then integrated with feature selection techniques to further enhance the performance of the classifier.

Chi-square test was performed to identify the features that are highly related to the target variable. This can help discern the variables that are the most likely cause of fatality in heart failure patients. However, filter methods like the chi-square test ignore the notion that a feature can be less informative on its own but when combined with other features, it can provide valuable insight into the data [32]. Therefore, a popular wrapper approach - Recursive Feature Selection (RFE) is utilized to select a suitable subset of features that optimizes the prediction performance of our BRF classifier. Wrapper methods heuristically search for sub-optimal subsets of features that can provide the best predictive performance. Thus, it generally provides superior results over filter techniques. 6 different performance measures – Accuracy, Sensitivity, Specificity, G-mean Score, MCC, and ROC-AUC were utilized to properly assess the performance of our proposed method. The outline of our proposed framework is illustrated in Fig. 1.

3.2.1. Balanced random forest classifier (BRF)

Data-level modification or Sampling is a standard approach to deal with the imbalanced classification problem. Standard RF classifier fails to tackle the class imbalance issue, getting biased towards the majority class. In order to prevent that, the sampling strategy is incorporated in the model construction stage of the RF classifier. The architecture of our proposed approach is described below:

Step-1: The number of trees (N) to be used to build the forest is decided. N was taken as 100 in this study. A large number of trees ensures better generalizability and reduced information loss due to undersampling.

Step-2: N number of bootstrap samples are generated from the training set. Bootstrapping is a type of resampling where large numbers

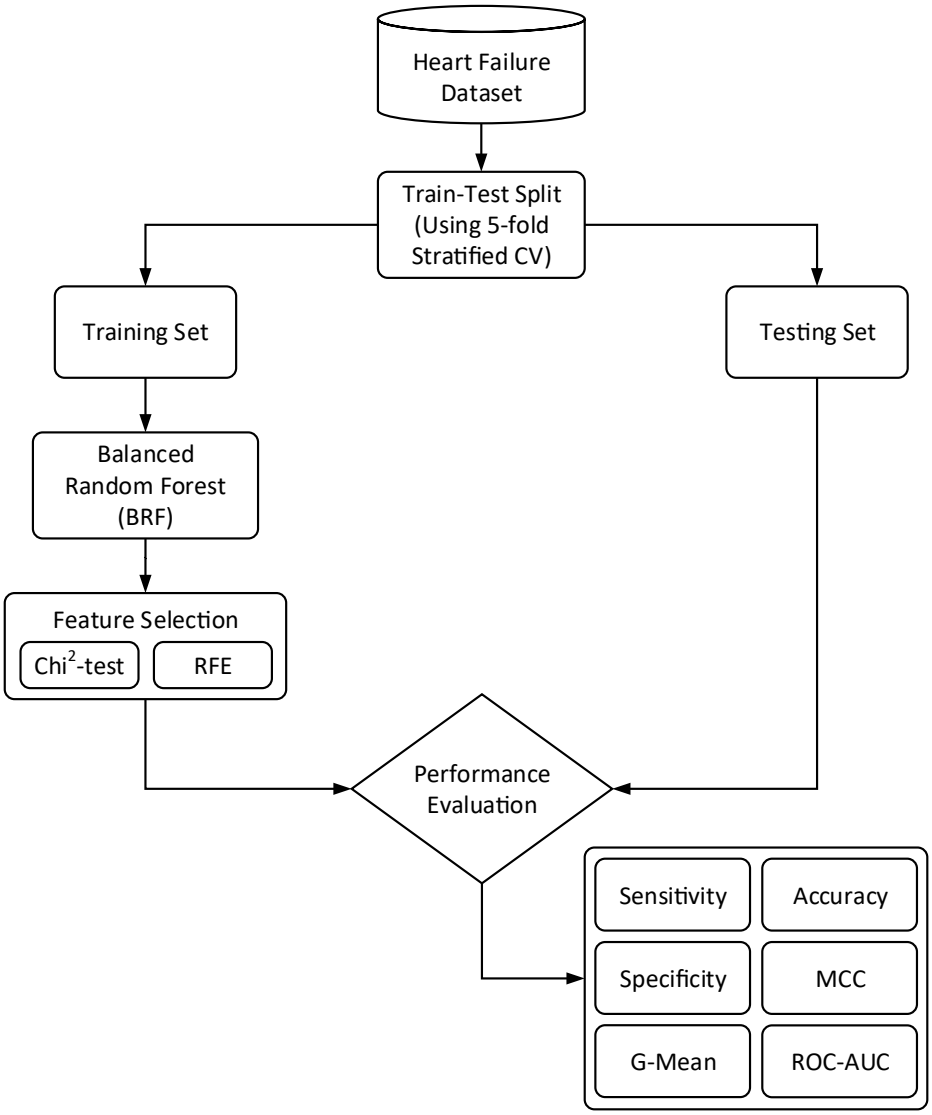


Fig. 1. Outline of the proposed framework.

of smaller samples of the same size are repeatedly drawn, with replacement from a single original sample [33].

Step-3: Random undersampling is performed on the majority class on each bootstrap sample to balance the class distribution in the data.

Step-4: Each tree is then trained on the balanced resampled data. This mitigates the problem of imbalance classification. Moreover, as the samples are drawn with replacement (samples are returned to the population after it has been used to form a sample set before the next unit is drawn), the problem of information loss due to undersampling is alleviated.

Step-5: At each node of each decision tree, Gini Index is calculated to pick the feature to go in that node that optimizes the metric. However, randomness is inserted here to grow un-correlated trees for better generalization. A random subset of features is chosen at each node for evaluation. This way, including features that have high predictive power in every tree can be avoided. The process is carried on until the whole tree is formed.

Step-6: Steps 3 to 5 are repeated for all the trees in the forest. This way a massive forest with a wide variety of trees is formed. Each tree is trained on a random balanced subset of the data. By generating a diverse set of trees like this, variance can be greatly reduced and a better performing model can be achieved.

Step-7: Finally, to make a prediction, each tree provides a decision for each sample in the test set. The final prediction is the most frequent prediction made by the trees in the forest.

The construction scheme of our proposed BRF classifier is illustrated in Fig. 2.

3.2.2. *Feature selection*

Using irrelevant features during the training process of the classifiers results in poor generalization while using redundant features only increases the complexity [34]. Therefore, identifying the most representative subset of features for the prediction task can not only reduce the overfitting and complexity of the model but also improve the prediction performance and reduce computation time [32] In this regard, we employed two different feature selection techniques: Chi-square test and Recursive Feature Elimination (RFE). Chi-square test is a filter approach that uses statistical measures to calculate variable dependency while RFE is a wrapper approach that is wrapped around a classification algorithm (BRF in our case) to identify a subset of features that optimizes its prediction performance.

3.2.2.1. Chi-square test. Chi-square test is a popular statistical method to test the independence of two events [24]. Given the data of two variables, it measures how expected count (E) and observed count (O) deviate from each other. This can be utilized to determine the

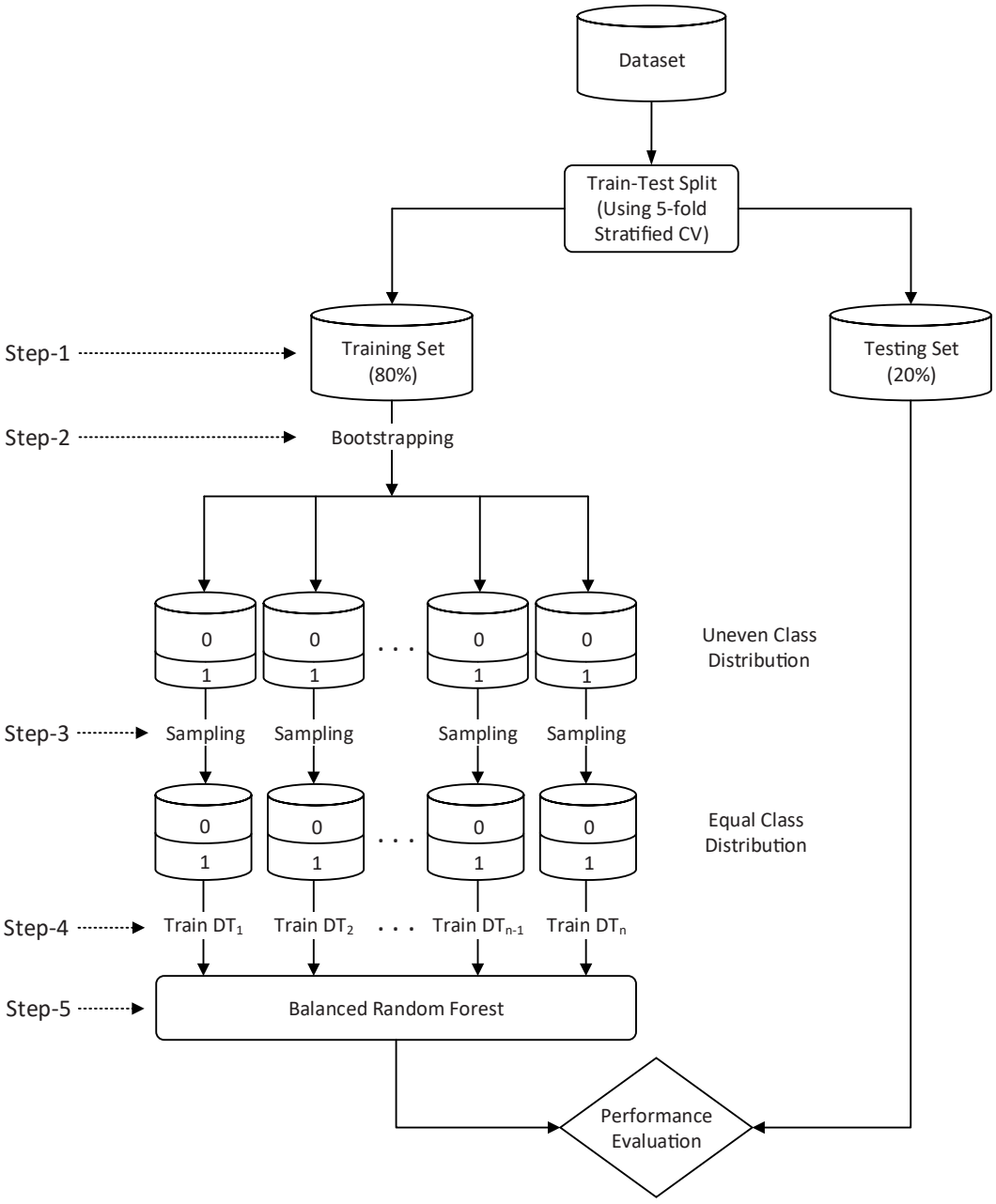


Fig. 2. Construction of proposed Balanced Random Forest classifier.

relationship between the predictor variables and the response variable. The covariates that are highly dependent on the response variable can be considered as the main causes of fatality in heart failure patients.

The formula for the chi-square test is:

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

When two variables are independent of each other, the observed count is close to the expected count, resulting in a smaller chi-Square score. Contrarily, the chi-square score would be higher if there lies a significant dependency between the two variables.

3.2.2.2. Recursive Feature Elimination (RFE). Recursive Feature Elimination (RFE) is one of the most popular feature selection algorithms due to its flexibility and ease of use [35]. It is a wrapper-based backward feature elimination technique. It starts with the entire set of features to train the model. The basic idea behind this algorithm is to measure the change in performance caused by removing a feature. Based on the performance criteria, it computes the ranking of all features this way. The feature with the smallest ranking is removed from the feature set. Then the model is trained again and importance scores of the features are calculated. The least important feature is removed and the process is repeated. The subset of features that optimizes the performance of the classifier is used to build the final model. The G-mean score was utilized as the performance measure. One important thing to notify here is that the features that are top-ranked by the RFE are not necessarily the ones that are individually most relevant [35]. Rather those features only taken together are able to optimize the prediction performance.

3.2.3. Performance metrics

Evaluation metrics quantify the performance of a predictive model. Accuracy is the most commonly used metric in classification tasks. However, in the case of imbalanced datasets, the accuracy metric can be quite misleading as it gets biased towards the majority class. Hence,

class-specific performance metrics are more useful in imbalance classification tasks. These can be better illustrated with the help of a confusion matrix (see Table 2). In the case of a binary classification task, there are two possible outcomes: True (1) and False (0).

TP and TN refer to the correct prediction of the positive (1) and negative (0) class instances respectively. FP refers to predicting a negative case as positive, while FN refers to predicting a positive case as negative. For our scenario, the deceased patients are marked as 1 and they represent the minority class. The patients who survived the study period are marked as 0 and they constitute the majority class. Based on these 4 basic terms, the metrics are defined.

3.2.3.1. Accuracy. Classification accuracy is defined as the ratio of the total number of correct predictions to the total number of instances.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

The majority class in the dataset dominates the classification accuracy of a model. The classifier might attain high accuracy when data is imbalanced. However, this high accuracy is over-optimistic and doesn't really represent the performance of the classifiers properly. Therefore, in order to get a better reflection on the performance of a classifier, different performance metrics need to be considered that take into account possible class imbalance scenarios.

3.2.3.2. Sensitivity or recall. Sensitivity, also known as Recall or True Positive Rate (TPR), represents the performance of the classifier on the positive (minority) class. A higher value of sensitivity reflects that the classifier is good at predicting the minority class instances.

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

3.2.3.3. Specificity. Specificity, also known as True Negative Rate (TNR), is a measure that determines how well the negative class was predicted. It is commonly used in tandem with sensitivity. Since negative class generally renders the majority of instances, classifiers normally manifest a higher specificity but lower sensitivity.

$$Specificity = \frac{TN}{FP + TN} \tag{4}$$

3.2.3.4. Geometric Mean (G-mean). Geometric Mean (G-mean) combines the two metrics (sensitivity and specificity) and provides a more balanced performance measure. A low G-Mean score is an indication of poor performance in the classification if either of the two classes has a higher misclassification rate.

$$G - mean = \sqrt{Sensitivity \times Specificity} \tag{5}$$

3.2.3.5. MCC. It stands for Matthews Correlation Coefficient. It is a reliable statistical metric that produces a high score only if the prediction obtained good results in all of the 4 confusion matrix categories. However, it doesn't differentiate between majority and minority class performances.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6}$$

The value ranges from −1 to +1. A value of +1 represents a perfect prediction, 0 is no better than a random prediction, and −1 is the worst possible prediction.

Table 2
Confusion matrix.

	Predicted False (0)	Predicted True (1)
Actual False (0)	TN	FP
Actual True (1)	FN	TP

3.2.3.6. ROC-AUC. ROC stands for Receiver Operating Characteristics and it is a widely used performance measure in classification tasks. AUC is the Area Under the Curve of True Positive Rate (TPR) vs. False Positive Rate (FPR). A higher TPR is desirable while the lower the FPR value, the better. AUC value can range from 0.5 to 1. A higher score is better.

4. Results and discussion

4.1. Performance comparison of models developed using standard RF classifier, SVM, KNN, LR, AdaBoost

Class Imbalance is frequently observed in many real-world applications [36,37] and it has brought a lot of attention from researchers due to the fact that standard classification algorithms are not designed to deal with such imbalance scenarios [38]. Because of the disparity, the prediction gets biased towards the majority class, resulting in a high misclassification rate for the minority class. This can be observed from the huge disproportion between sensitivity and specificity measures obtained when standard RF classifier was used for classification. The sensitivity, specificity, accuracy, G-mean, MCC, and ROC-AUC measures obtained using RF classifier are outlined in Table 3. The sensitivity or TPR is only 49.05% while specificity.

or TNR is 85.73%. Therefore, the classifier is clearly biased and the number of False Negative (FN) predictions are way too high to be considered as a reliable tool in survival prediction task.

For comparison, we also used four other popular classification algorithms: Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LR), and Adaptive Boosting (AdaBoost). The results are presented in Table 3. It can be observed that SVM and KNN classifier performed very poorly on the minority class prediction, providing a sensitivity score of only 9.42% and 14.58% respectively. The MCC score is very close to 0, which indicates that those classifiers are not performing any better than random prediction. LR performed comparatively better than SVM and KNN classifier, but lower than RF classifier. It is evident that the ensemble approach provides the best results, however still biased towards the majority class. It can be observed that AdaBoost, which is another ensemble technique employing the boosting methodology, provided a slightly higher sensitivity score than the RF classifier. However, the RF classifier performed better in terms of G-mean score, MCC, and ROC-AUC. Therefore, the RF classifier outperforms the other classification algorithms but still remains susceptible to class imbalance.

4.2. Performance comparison of models developed using standard RF classifier, SVM, KNN, LR, AdaBoost after applying SMOTE

Some special strategies need to be adopted to deal with such imbalance scenarios. SMOTE is one of the most popular sampling techniques for imbalance classification. It oversamples the minority class instances to balance the dataset. The classifier is then trained on the balanced data, reducing the bias caused by the majority class. To avoid data leakage, the data was first split into training and validation sets using a 5-fold CV. Sampling was performed only on the training

Table 3
Performance measures obtained (in percentage) using standard classification algorithms.

	SVM	KNN	LR	AdaBoost	RF
Sensitivity	9.42	14.58	34.32	51	49.05
Specificity	97.05	91.65	93.1	82.24	85.73
G-mean	26.44	35.72	56.12	64.25	64.31
Accuracy	68.89	66.89	74.23	72.23	73.92
MCC	11.57	9.43	35.16	34.59	37.84
ROC-AUC	53.23	53.11	63.71	66.62	67.39

folds. The classifiers were trained on the training sets.

and their performance was evaluated on the validation sets. The mean of the performance measures on 5 testing folds was taken. The performance measures obtained when SMOTE was first utilized to resample the dataset are presented in Table 4.

Improvement in performance is noticeable when SMOTE was applied. Especially for the SVM classifier, the sensitivity score improved significantly. In general, all the classifiers performed better when SMOTE was used to resample the data. However, in terms of MCC, G-mean score, accuracy, or ROC-AUC, the RF classifier again outperformed all other classifiers. The maximum MCC score obtained was 38.68% with the RF classifier. However, the sensitivity score achieved with the RF classifier was 59.53%, which is still relatively low. This can be due to the fact that generating a large number of synthetic samples from the existing instances can result in loss of generalization and hence the classifiers were not performing up to the mark on the testing sets.

4.3. Performance results of the model developed using BRF classifier

To overcome the shortcomings of standard classification algorithms as well as SMOTE, we propose a model where sampling technique is incorporated into the construction of an RF classifier. In this study, we under-sampled each bootstrapped dataset to attain a balanced class distribution. Thus, each tree in the forest is trained on balanced bootstrapped data. This approach not only reduces the bias produced by imbalanced data but also ensures better generalizable performance with minimal loss of information. Valuable information can be lost if undersampling is performed beforehand in the data preprocessing step. Moreover, a significant portion of the data remains unused in the model development stage that can lead to biased prediction. The problem can be alleviated by exploiting the bootstrapping methodology embedded in the construction of the RF classifier. Once a bootstrap sample is created, the instances are returned to the original population. This way no majority class instances are removed from the original population. By building a large forest in this manner, all the samples are likely to be utilized in model development, thus reducing the loss of information. The performance measures obtained using our proposed methodology are presented in Table 5.

As can be observed from Table 5, there is a significant improvement in performance when our proposed BRF classifier is employed to predict mortality in HF patients. The sensitivity score obtained is 71.9%, which is much higher compared to other classification algorithms, even 22.85% higher than the standard RF classifier which provided the best performance so far. The score is 12.37% higher than when SMOTE was applied beforehand to resample the dataset. The highest G-mean score of 72.67% is obtained using the BRF classifier, which is 8.36% higher than the conventional RF classifier. In terms of all metrics (except specificity), the BRF classifier outperformed all the other classifiers by a significant margin. Standard classification algorithms provided higher specificity due to the fact that they were biased towards the majority class. As the bias is alleviated using the BRF classifier, the specificity score slightly dropped. However, if we look at more robust performance measures like G-mean score or MCC, it is evident that the BRF classifier performs much better than others as well as popular techniques like SMOTE.

Table 4
Performance measures obtained (in percentage) after applying SMOTE.

	SVM	KNN	LR	AdaBoost	RF
Sensitivity	56.42	33.47	65.63	58.42	59.53
Specificity	69.5	73.46	71.98	73.93	79.34
G-mean	61.98	48.24	68.61	65.32	68.41
Accuracy	65.21	60.54	69.92	68.91	72.93
MCC	25	7.25	35.97	32.01	38.68
ROC-AUC	62.96	53.47	68.8	66.17	69.43

Table 5
Performance measures obtained (in percentage) using our proposed BRF classifier.

	Balanced Random Forest (BRF) Classifier
Sensitivity	71.9
Specificity	73.45
G-mean	72.67
Accuracy	72.93
MCC	43.39
ROC-AUC	72.67

4.4. Performance results of the model developed using Random Undersampling (RUS) with RF classifier

To compare with the conventional undersampling approach, we first randomly under-sample the training dataset. This resampled dataset was then used to train the RF classifier and the performance was measured on the test sets. The performance measures obtained are presented in Table 6. As it can be observed, the performance achieved using this approach is quite lower than our proposed methodology. The MCC score obtained using the BRF classifier is 7.45% higher than this. Therefore, the BRF classifier can be considered the most accurate method for risk prediction in patients with heart failure.

4.5. Performance comparison of models developed using BRF with features selected by RFE and chi-square

As it has been established that the BRF classifier outperforms the other conventional approaches, we also want to take a look into the features that are most significant in terms of predicting heart failure. In that regard, we first employed the chi-square test to evaluate the dependency of the features on the target variable. The results are depicted in Fig. 3. Ejection Fraction and Serum Creatinine have the highest scores meaning that they are highly related to the survival of heart failure patients. Other variables like age and serum sodium also scored comparatively higher. However, some variables like diabetes, gender, or smoking habit have almost 0 scores indicating that they have no relation with the survival of heart failure patients.

One thing to consider here is that the chi-square test is a univariate feature ranking technique. It does not take into consideration interdependency among features. A feature can be less informative on its own but when combined with other features, it can provide valuable insight into the data. Therefore, to attain a more concrete subset of features that can optimize the prediction performance, we employed the Recursive Feature Elimination technique. The process is wrapped around our proposed BRF classifier and the G-mean score was considered as the optimization parameter. A total number of 5 features were selected by RFE. Their importance scores calculated by RFE are presented in Fig. 4. It can be observed that Serum Creatinine and Ejection Fraction are the two highest scored features from RFE as well. This establishes the fact that these two features are indeed the key driver in the survival of heart failure patients. Patient's age is found as the 3rd important attribute which is also ranked 3rd in chi-square test score. Hence, a patient's age is also a critical factor in mortality in heart failure

Table 6
Performance measures obtained (in percentage) using Random Undersampling with RF classifier.

	Random Undersampling (RUS) + RF
Sensitivity	66.79
Specificity	70.96
G-mean	68.57
Accuracy	69.57
MCC	35.94
ROC-AUC	68.88

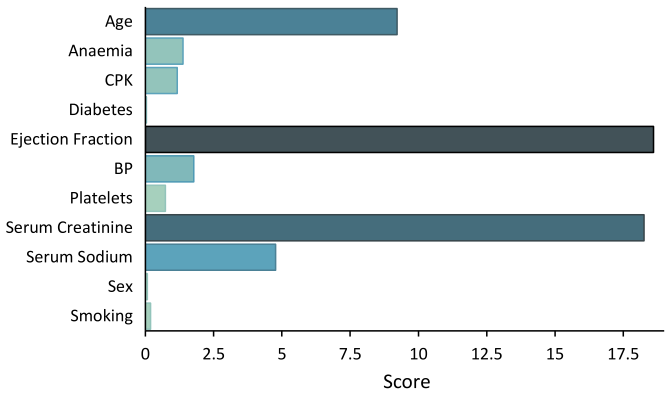


Fig. 3. Chi-square test score.

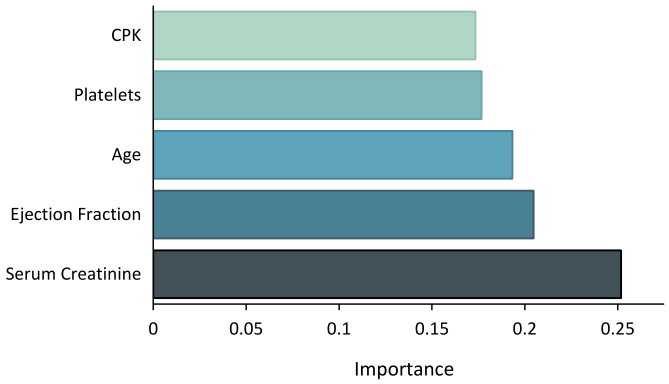


Fig. 4. Features selected by RFE and their importance score.

patients. The other two features selected by RFE are Platelets count and Creatine Phosphokinase (CPK). Although these two features' chi-square test score is comparatively low, when taken with the other 3 features, they constitute a feature subset that can improve the performance of the classifier. Utilizing only the 5 selected features attained from RFE, our proposed BRF was classifier was trained again. Improvement was noticeable in the prediction performance of the classifier when feature selection was incorporated into the framework. The results obtained by incorporating feature selection techniques are presented in Table 7.

As for the selected features, Ejection Fraction (EF) and Serum Creatinine are the two most important predictors for survival prediction. Serum Creatinine is a key biomarker in renal dysfunction and EF is an important measurement used to classify heart failure. They are well-known in the literature as a major driver of heart failure [8]. Age is identified as the 3rd most important factor which is expectable. The last two features - Platelets count and CPK showed small dependence on the target variable. However, CPK and platelets count is related to other major issues like renal dysfunction or heart injury, which are major causes of mortality. Thence they may not be directly related to the target variable, but have significant relation with other variables. When combined with other features, they provide better predictive performance.

From Table 7, it can be observed that the sensitivity score improved by 6.31% when only the 5 selected features from RFE were utilized. The improvement was 8.31% when only the top 3 features from the chi^2 -test were utilized. This is important since a high sensitivity score represents a high accuracy in the prediction of mortality in heart failure patients. By identifying the patients at risk accurately, treatment can be timely started and clinicians can make informed decisions regarding the intensity of treatment required which can prove highly beneficial for the patients. The highest G-mean score of 76.83% and MCC of 52.53% were achieved this way. Thus, by incorporating feature selection methodology along with our proposed BRF classifier, the best performance was obtained. This final model produces a more balanced predictive performance using only 3 features.

Fig. 5 illustrates the improvement in the G-mean score obtained using the selected feature set in conjunction with the BRF classifier, as compared to the G-mean score obtained using the standard RF classifier, RF with oversampled data (SMOTE), RF with under-sampled data (RUS) and BRF classifier without feature selection.

Table 7
Performance measures obtained (in percentage) using selected features by applying RFE.

	BRF + RFE	BRF + Chi^2
Sensitivity	78.21	80.21
Specificity	70.51	74.45
G-mean	74.26	76.83
Accuracy	72.93	76.25
MCC	46.33	52.53
ROC	74.36	77.33

4.6. *Performance comparison of our proposed approach with previous works*

Performance comparison of our proposed approach with previous works has been reported in Table 8. Some measures were not reported by the authors. Therefore, they are kept empty. It can be observed from the table that in terms of G-mean score or ROC-AUC measure, our approach clearly outperforms the previous works. Among other works, the maximum sensitivity of 71.23% was reported by Kim et al. [29] using oversampling techniques. Sensitivity score achieved using our proposed methodology is 80.21% which is significantly higher than the other works. Hasan et al. [30] reported an accuracy of 80% which is a bit higher than the accuracy obtained from our method. However, accuracy is not the ideal metric for imbalanced data which is apparent from the low sensitivity score (51.72%) obtained with the approach proposed by Hasan et al. [30].

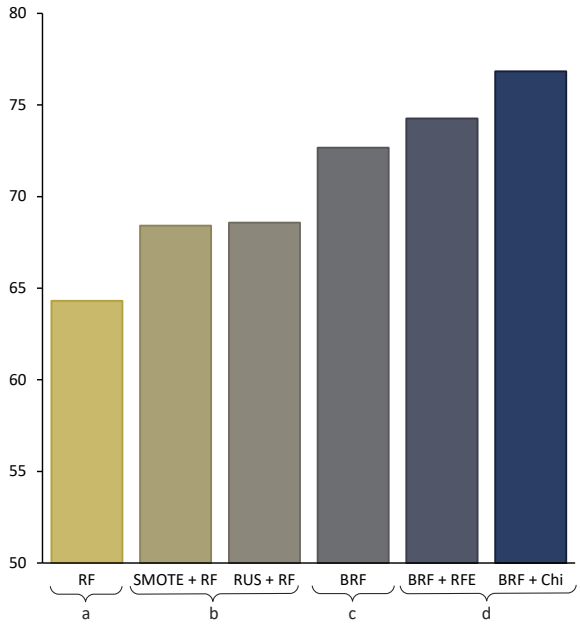


Fig. 5. Performance comparison (in terms of G-mean score) of our proposed approach with other techniques – (a) Standard RF classifier (b) Sampling techniques to handle imbalanced classification problem: Random Undersampling (RUS), SMOTE (c) Balanced RF classifier (d) BRF with Feature Selection methods: RFE and Chi-square test.

Table 8
Performance comparison of our proposed approach with previous works.

Study	Methodology	Accuracy	G-Mean	Sensitivity	Specificity	AUC	MCC
Chicco [13]	RF with univariate feature selection strategy	58.5%	68.01%	54.1%	85.5%	69.8%	41.8%
Kim [29]	RF + SMOTE	–	73.14%	71.23%	75.11%	–	–
Hasan [30]	DT + MRMR + RFE	**80%**	69.52%	51.72%	**93.44%**	72.58%	–
This Study	BRF + Chi2	76.25%	**76.83%**	**80.21%**	74.45%	**77.33%**	**52.53%**

5. Conclusion

In this study, we have developed a model to build a reliable decision-support system for the survival prediction of heart failure patients. The performance achieved on the publicly available HF dataset as well as the versatility of the proposed approach indicates that it has the potential to be a reliable tool that can be utilized in clinical practice to assist clinicians and practitioners in decision-making. Serum Creatinine and Ejection Fraction have been identified as the key factors in predicting risk. However, the patient's age is also a critical factor. Using only these 3 factors, the heart failure patients who are currently at risk can be accurately identified with our proposed method.

A limitation of the current study is that the model is developed on a comparatively smaller dataset owing to the lack of a publicly available dataset for this purpose. A large dataset from a different geographical region would certainly enhance the robustness of the model as well as provide a better understanding of the features that are the most likely cause of mortality in heart failure patients.

In our model, we have incorporated undersampling approach into the construction of bootstrapped samples to build our balanced random forest classifier. However, other sampling techniques can also be utilized and their performance can be compared with the model presented in this study. This is something we plan to look into in future developments.

Conflicts of interest disclosure

We declare that there is no conflict of interest.

Sources of funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study utilized heart failure patient's data collected from the Faisalabad Institute of Cardiology, Pakistan, and the Allied Hospital in Faisalabad, Pakistan. We thank Assia Munir for the data curation.

References

[1] National Heart Lung and Blood Institute (NHLBI). Heart failure. https://www.nhlbi.nih.gov/health-topics/heart-failure/. [Accessed 6 October 2021]. accessed.
[2] Bragazzi NL, Zhong W, Shu J, Abu Much A, Lotan D, Grupper A, Younis A, Dai H. Burden of heart failure and underlying causes in 195 countries and territories from 1990 to 2017. Feb 12 Eur J Prevent Cardiol 2021. https://doi.org/10.1093/eurjpc/zwaa147.
[3] Virani SS, Alonso A, Aparicio HJ, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Cheng S, Delling FN, Elkind MS. Heart disease and stroke statistics—2021 update: a report from the American Heart Association. Feb 23 Circulation 2021;143(8):e254–743. https://doi.org/10.1161/CIR.0000000000000950.
[4] Heidenreich PA, Albert NM, Allen LA, Bluemke DA, Butler J, Fonarow GC, Ikonomidis JS, Khavjou O, Konstam MA, Maddox TM, Nichol G. Forecasting the impact of heart failure in the United States: a policy statement from the American Heart Association. May Circulation: Heart Fail 2013;6(3):606–19. https://doi.org/10.1161/HHF.0b013e318291329a.
[5] Balogh EP, Miller BT, Ball JR. Improving diagnosis in health care. https://doi.org/10.17226/21794.
[6] Ahmad T, Munir A, Bhatti SH, Aftab M, Raza MA. Survival analysis of heart failure patients: a case study. Jul 20 PLoS One 2017;12(7):e0181001. https://doi.org/10.1371/journal.pone.0181001.
[7] Dickstein K. The task force for the diagnosis and treatment of acute and chronic heart failure 2008 of the European Society of Cardiology: developed in collaboration with the heart failure association of the ESC (HFA) and endorsed by the European Society of intensive care medicine (ESICM). Eur Heart J 2008;29:2388–442.
[8] Rahimi K, Bennett D, Conrad N, Williams TM, Basu J, Dwight J, Woodward M, Patel A, McMurray J, MacMahon S. Risk prediction in patients with heart failure: a systematic review and analysis. Oct JACC (J Am Coll Cardiol): Heart Fail 2014;2(5):440–6. https://doi.org/10.1016/j.jchf.2014.04.008.
[9] McKie PM, Cataliotti A, Lahr BD, Martin FL, Redfield MM, Bailey KR, Rodeheffer RJ, Burnett JC. The prognostic value of N-terminal pro–B-type natriuretic peptide for death and cardiovascular events in healthy normal and stage A/B heart failure subjects. May 11 J Am Coll Cardiol 2010;55(19):2140–7. https://doi.org/10.1016/j.jacc.2010.01.031.
[10] Sartipy U, Dahlström U, Edner M, Lund LH. Predicting survival in heart failure: validation of the MAGGIC heart failure risk score in 51 043 patients from the Swedish Heart Failure Registry. Feb Eur J Heart Fail 2014;16(2):173–9. https://doi.org/10.1111/ejhf.32.
[11] Sawano M, Shiraishi Y, Kohsaka S, Nagai T, Goda A, Mizuno A, Sujino Y, Nagatomo Y, Kohno T, Anzai T, Fukuda K. Performance of the MAGGIC heart failure risk score and its modification with the addition of discharge natriuretic peptides. Aug ESC Heart Fail 2018;5(4):610–9. https://doi.org/10.1002/ehf2.12278.
[12] Pocock SJ, Ariti CA, McMurray JJ, Maggioni A, Køber L, Squire IB, Swedberg K, Dobson J, Poppe KK, Whalley GA, Doughty RN. Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. May 14 Eur Heart J 2013;34(19):1404–13. https://doi.org/10.1093/eurheartj/ehs337.
[13] Chicco D, Jurman G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. Dec BMC Med Inf Decis Making 2020;20(1):1–6. https://doi.org/10.1186/s12911-020-1023-5.
[14] Buchan TA, Ross HJ, McDonald M, Billia F, Delgado D, Posada JD, Luk A, Guyatt GH, Alba AC. Physician prediction versus model predicted prognosis in ambulatory patients with heart failure. Apr 1 J Heart Lung Transplant 2019;38(4):S381. https://doi.org/10.1016/j.healun.2019.01.971.
[15] Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Jan 1 Intell Data Anal 2002;6(5):429–49. https://doi.org/10.3233/IDA-2002-6504.
[16] Safavian R. S., David Landgrebe A survey of decision tree classifier methodology. IEEE Trans Syst, Man Cybernet.;213. https://doi.org/10.1109/21.97458.
[17] Breiman L. Bagging predictors. Aug Mach Learn 1996;24(2):123–40. https://doi.org/10.1007/BF00058655.
[18] Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. Learning from imbalanced data sets. Oct 22. Berlin: Springer; 2018. 10.1007%2F978-3-319-98074-4.
[19] Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. Safe-level-smote: safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. Apr 27. In Pacific-Asia conference on knowledge discovery and data mining. Berlin, Heidelberg: Springer; 2009. p. 475–82. https://doi.org/10.1007/978-3-642-01307-2_43.
[20] He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. Jun 1. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). IEEE; 2008. p. 1322–8. https://doi.org/10.1109/IJCNN.2008.4633969.
[21] Wilson DL. Asymptotic properties of nearest neighbor rules using edited data. Jul IEEE Trans Syst, Man Cybernet 1972;(3):408–21. https://doi.org/10.1109/TSMC.1972.4309137.
[22] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Jun 1 J Artif Intell Res 2002;16:321–57. https://doi.org/10.1613/jair.953.
[23] Chen C, Liaw A, Breiman L. Jul. Using random forest to learn imbalanced data, vol. 110. Berkeley: University of California; 2004. p. 24. 1-12.
[24] McHugh ML. The chi-square test of independence. Jun 15 Biochem Med 2013;23(2):143–9. https://doi.org/10.11613/BM.2013.018.
[25] Panahiazar M, Taslimitehrani V, Pereira N, Pathak J. Using EHRs and machine learning for heart failure survival analysis. Stud Health Technol Inf 2015;216:40. https://doi.org/10.3233/978-1-61499-564-7-40.
[26] Levy WC, Mozaffarian D, Linker DT, Sutradhar SC, Anker SD, Cropp AB, Anand I, Maggioni A, Burton P, Sullivan MD, Pitt B. The Seattle Heart Failure Model:

prediction of survival in heart failure. Mar 21 Circulation 2006;113(11):1424–33. https://doi.org/10.1161/circulationaha.105.584102.
[27] UCI Machine Learning Repository. Uci.edu. https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records. accessed: 06-Oct-2021.
[28] Zahid FM, Ramzan S, Faisal S, Hussain I. Gender based survival prediction models for heart failure patients: a case study in Pakistan. Feb 19 PLoS One 2019;14(2): e0210602. https://doi.org/10.1371/journal.pone.0210602.
[29] Kim YT, Kim DK, Kim H, Kim DJ. A comparison of oversampling methods for constructing a prognostic model in the patient with heart failure. In2020. Oct 21 Int Conf Informat Commun Technol Converg (ICTC) 2020:379–83. https://doi.org/10.1109/ICTC49870.2020.9289522. IEEE.
[30] Al Mehedi Hasan M, Shin J, Das U, Yakin Srizon A. Identifying prognostic features for predicting heart failure by using machine learning algorithm. Mar 17 In2021 11th Int Conf Biomed Eng Technol 2021:40–6. https://doi.org/10.1145/3460238.3460245.
[31] Dolgin M, Association NYH. Nomenclature and criteria for diagnosis of diseases of the heart and great vessels. Boston: Little, Brown; 1994.
[32] Chandrashekar G, Sahin F. A survey on feature selection methods. Jan 1 Comput Electr Eng 2014;40(1):16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024.
[33] Wright DB, London K, Field AP. Using bootstrap estimation and the plug-in principle for clinical psychology data. May J Exper Psychopathol 2011;2(2): 252–70. https://doi.org/10.5127/jep.013611.
[34] John GH, Kohavi R, Pfleger K. Irrelevant features and the subset selection problem. 1994 Jan 1 InMachine Learn Proc 1994:121–9. https://doi.org/10.1016/B978-1-55860-335-6.50023-4. Morgan Kaufmann.
[35] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Jan Mach Learn 2002;46(1):389–422. https://doi.org/10.1023/A:1012487302797.
[36] Devarriya D, Gulati C, Mansharamani V, Sakalle A, Bhardwaj A. Unbalanced breast cancer data classification using novel fitness functions in genetic programming. Feb 1 Expert Syst Appl 2020;140:112866. https://doi.org/10.1016/j.eswa.2019.112866.
[37] Tavallaee M, Stakhanova N, Ghorbani AA. Toward credible evaluation of anomaly-based intrusion-detection methods. May 17 IEEE Trans Syst, Man Cybernet, C (Appl Rev) 2010;40(5):516–24. https://doi.org/10.1109/TSMCC.2010.2048428.
[38] Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Jan 1 Intell Data Anal 2002;6(5):429–49. https://doi.org/10.3233/IDA-2002-6504.