Hindawi
Scientific Programming
Volume 2022, Article ID 3649406, 17 pages
https://doi.org/10.1155/2022/3649406

Research Article

A Comprehensive Investigation of the Performances of Different Machine Learning Classifiers with SMOTE-ENN Oversampling Technique and Hyperparameter Optimization for Imbalanced Heart Failure Dataset

Mirza Muntasir Nishat,[1] Fahim Faisal ,[1] Ishrak Jahan Ratul,[1] Abdullah Al-Monsur,[1] Abrar Mohammad Ar-Rafi,[1] Sarker Mohammad Nasrullah,[2] Md Taslim Reza,[1] and Md Rezaul Hoque Khan[1]

[1]*Islamic University of Technology, Gazipur, Bangladesh*
[2]*North South University, Dhaka, Bangladesh*

Correspondence should be addressed to Fahim Faisal; faisaleee@iut-dhaka.edu

Received 5 September 2021; Revised 17 December 2021; Accepted 26 January 2022; Published 9 March 2022

Academic Editor: Qianchuan Zhao

Heart failure is a chronic cardiac condition characterized by reduced supply of blood to the body due to impaired contractile properties of the muscles of the heart. Like any other cardiac disorder, heart failure is a serious ailment limiting the activities and curtailing the lifespan of the patient, most often resulting in death sooner or later. Detection of survival of patients with heart failure is the path to effective intervention and good prognosis in terms of both treatment and quality of life of the patient. Machine learning techniques can be critical in this regard since they can be used to predict the survival of patients with heart failure in advance, allowing patients to receive appropriate treatment. Hence, six supervised machine learning algorithms have been studied and applied to analyze a dataset of 299 individuals from the UCI Machine Learning Repository and predict their survivability from heart failure. Three distinct approaches have been followed using Decision Tree Classifier, Logistic Regression, Gaussian Naïve Bayes, Random Forest Classifier, K-Nearest Neighbors, and Support Vector Machine algorithms. Data scaling has been performed as a preprocessing step utilizing the standard and min–max scaling method. However, grid search cross-validation and random search cross-validation techniques have been employed to optimize the hyperparameters. Additionally, the synthetic minority oversampling technique and edited nearest neighbor (SMOTE-ENN) data resampling technique are utilized, and the performances of all the approaches have been compared extensively. The experimental results clearly indicate that Random Forest Classifier (RFC) surpasses all other approaches with a test accuracy of 90% when used in combination with SMOTE-ENN and standard scaling technique. Therefore, this comprehensive investigation portrays a vivid visualization of the applicability and compatibility of different machine learning algorithms in such an imbalanced dataset and presents the role of the SMOTE-ENN algorithm and hyperparameter optimization for enhancing the performances of the machine learning algorithms.

1. Introduction

Heart failure (HF) refers to the condition when the heart cannot pump adequate blood throughout the body. According to the WHO, it has emerged as one of the most lethal and debilitating diseases, claiming approximately 18 million lives each year [1]. Chronic conditions such as weak or damaged heart muscles result in a decreased ejection fraction, which eventually results in heart failure. However, it can also cause severe damage to the body's other vital organs and can strike both children and adults. Age, family history, genetics, lifestyle habits, cardiovascular diseases (CVD), and race or ethnic origin are the major risk factors for heart failure. It is equally prevalent in men and women, but women develop it at a later age [2].

Nevertheless, clinical detection of HF proves to be difficult as patients predominantly present with dyspnea attributed to a wide range of differential diagnoses [3, 4]. The American Heart Association defined heart failure as a progressive dysfunction of the heart where it fails to supply an adequate amount of blood to meet the metabolic demand of the body [5]. In most cases, it is a chronically deteriorating condition known as chronic heart failure (CHF). However, the signs and symptoms may also develop acutely within 24 hours, giving rise to acute heart failure (AHF), which may present with pulmonary edema, cardiogenic shock leading to hypotension, oliguria, and other related features, and decompensating CHF [6]. However, ischemic heart diseases are the most common cause of HF. Cardiomyopathies and valvular heart diseases come next in the line of etiologies [7]. Risk factors include hypertension, diabetes, hypercholesterolemia, obesity, smoking, congenital cardiac diseases, arrhythmias, and family history [8]. Moreover, there is a scarcity of data from the developing nations pertaining to heart diseases [9]. The disease is rare in the young, whereas the incidence of HF increases along with the progression of age after the age of 50 years [10]. Heart failure is among the diseases with high hospitalization rates. Estimates from New Zealand, the USA, Sweden, Scotland, and Netherland revealed that the age-adjusted rate of hospitalization had risen gradually since the 1980s [11]. Furthermore, the disease is responsible for the annual death of around 10%. Mortality due to heart failure is mostly from sudden cardiac death [12]. In spite of the advancement of medical science and associated technologies, the rate of death within 5 years after the diagnosis of HF is still 25% to 50% [13]. Cardiac diseases are the causes of life-long morbidity and medication in the patients. The prognosis of any particular heart disease depends on the early detection and rapid management of the condition, which goes the same for heart failure [14]. Machine learning classification techniques have the potential to significantly benefit the medical field by enabling accurate and rapid disease diagnosis [15–18]. In this alarming situation, recent technological advancements and the computerization of the health sector in Bangladesh may make it easier to implement machine learning models for different disease prediction. Machine learning and data mining have enormous potential for revealing hidden patterns in clinical domain data sets [19–21]. These patterns can be used to aid in medical diagnosis.

In this study, the main contribution would be carrying out a rigorous investigative analysis in applying six supervised machine learning algorithms in a heart failure dataset extracted from the UCI machine learning repository. For the purpose of investigation, three approaches are undertaken, which are as follows:

(i) Approach A: default hyperparameter and no data preprocessing

(ii) Approach B: hyperparameter optimization and data scaling

(iii) Approach C: data sampling by SMOTE-ENN algorithm and hyperparameter optimization

Hence, a comparative analysis has been portrayed with a view to evaluating the performance parameters obtained from the simulation accomplished in Python programming language. In addition, the performances have been compared with other research works. To the best of our knowledge, this dataset has not been investigated before in such a manner that may provide new promising windows in developing an intelligent computer-aided diagnosis system so that timely and proper treatment can be ensured for patients pertaining to heart diseases.

2. Related Works

Machine learning is on the trend in the health sector for a variety of reasons, including disease prediction, medical imaging diagnosis, and personalized medicine [22–25]. Numerous studies have been conducted on the use of data mining techniques to predict heart disease in recent times [26–28]. Several research articles have been studied related to the prediction of heart failure patient's survival using machine learning techniques. Ahmad et al. conducted a research where they performed statistical analysis (Cox regression and Kaplan Meier Plot) to predict the survival probability of heart disease. According to their study, the main dominant features for predicting heart failure are age, ejection fraction (EF), serum creatinine, serum sodium, anemia, and blood pressure are [29]. However, Chicco and Jurman applied several machine learning classifiers to both predict the patient's survival and rank the features corresponding to the most important risk factors. They also used traditional biostatistics methods and carried out a comparative analysis. From both feature rankings, serum creatinine and ejection fraction are the most important attributes for building a prediction model. Considering all features, they achieved 74% accuracy, while with two features (serum creatinine and ejection fraction), they obtained an accuracy of 83.8% [30]. On the other hand, Oladimeji and Olayanju proposed a machine learning-based integrated method for the prediction of survival of heart failure patients. The integrated method deals with the class imbalance in the classification dataset by selecting significant predictive features in order of their ranking. The Random Forest algorithm displayed the highest accuracy of 83.18% [31]. Moreover, Gürfidan and Ersoy implemented different classification algorithms on the heart failure dataset, where the Support Vector Machine (SVM) algorithm showed the highest accuracy of 90% among all the algorithms [32]. Furthermore, Elyassami and Kaddour formed an incremental deep learning model and used stochastic gradient descent to train the model. To increase the performance of the heart disease patient's classification model, they implemented the chi-square test and dropout regularization into the model, and the model achieved a balanced accuracy of 91.43% [33]. However, Rubini et al. presented a comparative analysis of machine learning techniques like Random Forest Classifier (RFC), Logistic Regression (LR), Support Vector Machine (SVM), and Naïve Bayes (NB) in the classification of cardiovascular disease. From their comparative analysis, RFC and LR executed the highest accuracies of 84.81% and 83.82%, respectively [34]. Ishaq et al. employed nine classification models to predict heart failure patients' survival with the synthetic minority oversampling technique (SMOTE) to solve the problem of class imbalance. The experimental results showed that the Extra Tree Classifier (ETC) outperformed the other models and gained an accuracy of 92.62% with SMOTE [35]. On the other hand, Rahayu et al. utilized RFC, DTC, KNN, SVM, NB, and Artificial Neural Network (ANN) with resample and SMOTE techniques where they achieved an accuracy of 94.31% and 85.82%, respectively [36]. Ali et al. developed a feature-driven decision support system consisting of two main stages to improve heart prediction accuracy. In the first stage, the $\chi 2$ statistical model was employed to rank thirteen heart failure features. Using forward best-first search, an optimal subset of features has been formed. In the second stage, Gaussian Naïve Bayes (GNB) classifier was applied as a predictive model, and finally, the proposed method attained a prediction accuracy of 93.33% [37]. While prior research has demonstrated that various machine learning techniques are proved to be quite effective in predicting the survival of patients with heart failure, none of them has achieved an accuracy greater than 95% to the best of our knowledge. This research presents a comprehensive analysis consisting of three approaches with six state-of-the-art machine learning (ML) algorithms to predict the survival of patients with heart failure. In order to enhance the performances of the classifiers, the SMOTE-ENN technique and hyperparameter optimization are incorporated, and different data scaling techniques are employed, which provides a rigorous investigation of this imbalanced dataset.

3. Methodology

3.1. Data Description. This dataset has been extracted from the UCI machine learning repository, which contains medical information for 299 patients, gathered from the Faisalabad Institute of Cardiology and the Allied Hospital in Faisalabad (Punjab, Pakistan) [38]. It consists of information on 105 females and 194 males with Left Ventricular Systolic Dysfunction (LVSD) classified as stage 3 or stage 4 HF by the New York Heart Association (NYHA). The patients' age ranged from 40 to 95 years, and the follow-up time was between 4 and 285 days. The dataset contains 13 attributes that have been assessed during the patients' follow-up at the hospital. Table 1 summarizes the characteristics. However, seven of the thirteen traits are numeric, while the remaining six are Boolean. Therefore, the statistical information of the numerical attributes is tabulated in Table 2. Following that, the dataset was imported into Jupyter Notebook and was subjected to exploratory data analysis to ascertain its general characteristics and validity. Then, a correlation heatmap was developed, as depicted in Figure 1, to determine the degree of correlation among the attributes.

3.2. Feature Scaling. The term "feature scaling" refers to the process of normalizing or standardizing independent features or variables. This is because machine learning algorithms can give more weight to higher values and less weight to lower values regardless of their units. Standardization ensures that the values of specific attributes have a mean of zero and a variance of one [39]. In this work, both min–max and standard scalars are used for investigating the performances of the ML models. The mathematical expressions for the scalars are depicted as follows:

$$\text{Min} - \text{max Scaling}, x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

$$\text{Standardization}, x' = \frac{x - \overline{x}}{\sigma}, \tag{2}$$

where mean, $\overline{x} = (1/N)\sum_{i=1}^{N} x_i$ and standard deviation, $\sigma = \sqrt{(1/N)\sum_{i=1}^{N} (x_i - \overline{x})^2}$.

3.3. Data Sampling. Synthetic Minority Oversampling Technique and Edited Nearest Neighbor (SMOTE-ENN) refers to a sampling technique that combines techniques of over- and undersampling minority classes in an imbalanced dataset. For instance, in this dataset, the number of deaths and survival are 96 and 203, respectively (out of 299 patients). For the purpose of resampling this imbalanced dataset, this algorithm has been utilized to balance the class distributions. It has emerged to be an effective method when there is an imbalance in the distribution of classes, as machine learning algorithms can be biased in favor of the majority class in the presence of an imbalance [40,41]. SMOTE-ENN oversamples the minority class initially using interpolation and then removes redundant samples using the ENN method. Finally, it produces class balanced data that can be used with machine learning algorithms to achieve the desired performance.

3.4. Hyperparameter Optimization. Hyperparameters refer to a collection of parameters that can control the learning procedure of machine learning algorithms. Optimization of hyperparameters has the potential to significantly impact the outcome and performance of machine learning algorithms [42]. This study employs both random search cross validation (RSCV) and grid search cross validation (GSCV) to govern the optimal hyperparameter combination. Grid search is a parameter sweep technique that evaluates all possible combinations of given parameters and returns the optimal result based on previously defined performance metrics. However, it appears to be expensive in terms of consuming time and requires more resources. On the other hand, random search chooses random combinations rather than attempting all possible combinations. It is more time and resource-efficient and is used when parameter influences on outcomes are minimal [43]. The optimal values of hyperparameters are then deployed to enhance the performances of the ML models.

3.5. Workflow. Three different approaches are taken into consideration in order to inspect the performance of six popularly used supervised ML models, namely, Decision Tree Classifier (DTC), Logistic Regression (LR), Gaussian

Table 1: Attribute of the dataset [30].

Sl. no.	Attribute	Data type	Information
1	Age (years)	Numeric	Age of the patient
2	Anemia	Boolean	Decrease of red blood cells or hemoglobin
3	Creatinine_phosphokinase (Mcg/L)	Numeric	Level of creatine phosphokinase enzyme in blood
4	Diabetes	Boolean	If the patient has diabetes
5	Ejection_fraction (percentage)	Numeric	Volume of blood ejected from the left ventricle in each contraction
6	High_blood_pressure	Boolean	If the patient has high blood pressure
7	Platelets (kiloplatelets/mL)	Numeric	Platelets count in the blood
8	Serum_creatinine (Mg/dL)	Numeric	Level of creatinine in the blood
9	Serum_sodium (mEq/L)	Numeric	Level of sodium in the blood
10	Sex	Boolean	Man or woman
11	Smoking	Boolean	If the patient has a smoking habit
12	Time (days)	Numeric	Follow-up period
13	Death_event	Boolean	If the patient died during the follow-up period

Table 2: Statistical information of numeric attributes.

Sl. no.	Numeric attributes	Maximum	Minimum	Mean	Standard deviation
1	Age	95.0	40.0	60.83	11.89
2	Creatinine_phosphokinase	7861.0	23.0	581.84	970.29
3	Ejection_fraction	80.0	14.0	38.084	11.83
4	Platelets	850000.0	25100.0	263358.03	97804.24
5	Serum_creatinine	9.4	0.5	1.40	1.035
6	Serum_sodium	148.0	113.0	136.63	4.41
7	Time	285.0	4.0	130.26	77.61

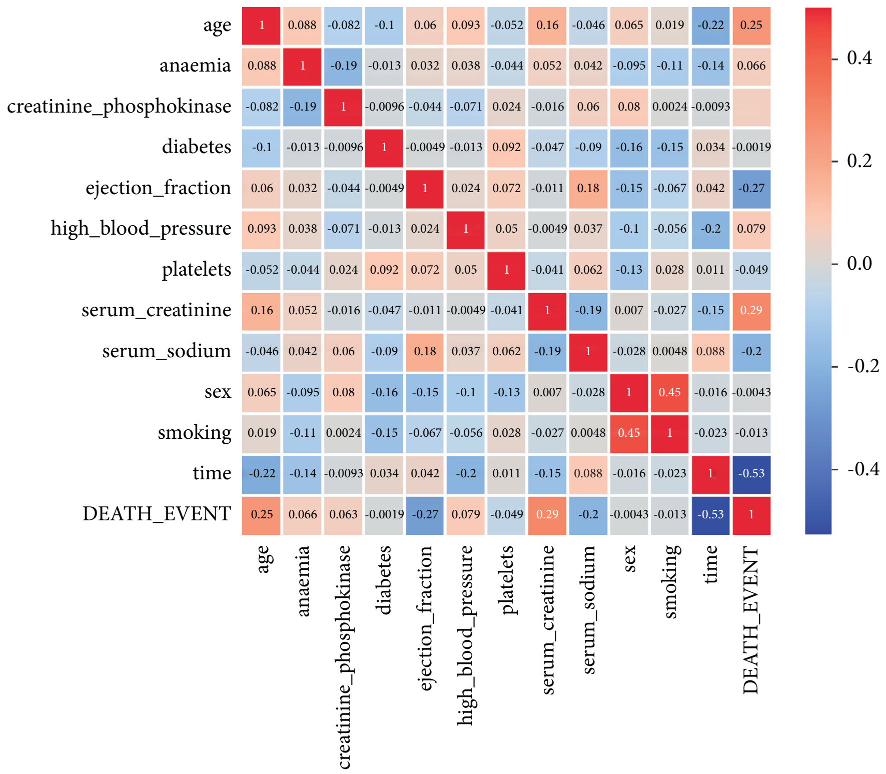


Figure 1: Correlation heatmap.

Naïve Bayes (GNB), Random Forest Classifier (RFC), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). The approaches are highlighted below with their corresponding workflow diagram.

3.5.1. Approach A: Default Hyperparameter and No Data Preprocessing. Firstly, machine learning (ML) models have been constructed, trained, and validated using the default data distribution and no preprocessing. Hence, the

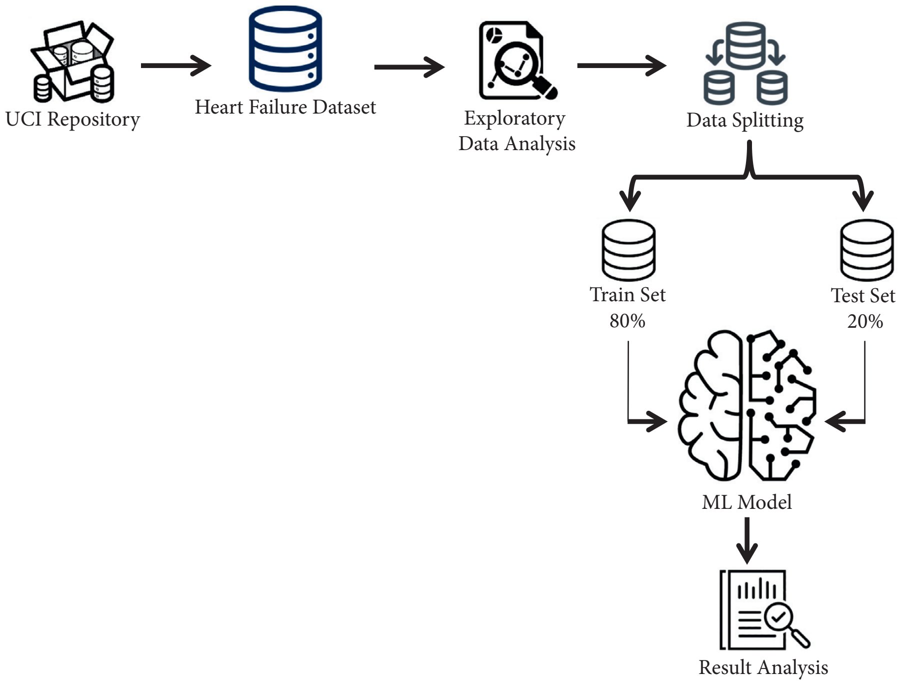


Figure 2: Workflow diagram of Approach A.

performance matrices have been evaluated using a 20% test dataset. However, default hyperparameters are utilized in this method. Figure 2 illustrates the workflow diagram of Approach A.

3.5.2. Approach B: Hyperparameter Optimization and Data Scaling. Secondly, hyperparameter optimization (HPO) has been performed using grid search cross validation (GSCV) and random search cross validation (RSCV). In this approach, data scaling has been accomplished by the use of min–max and standard scalar methods, and the dataset is not class balanced. Then the dataset has been cross-validated by 5-fold and 10-fold, and the optimal hyperparameters have been identified and used to evaluate the ML models. The workflow diagram of Approach B is depicted in Figure 3.

3.5.3. Approach C: Data Sampling (SMOTE-ENN Algorithm) and Hyperparameter Optimization. Finally, in Approach C, the dataset has been resampled by employing SMOTE-ENN to balance the class distributions, which was imbalanced. Then, the dataset has been split into test and train sets and 5-fold and 10-fold cross validations have been performed. After scaling and class balancing the data, RSCV and GSCV have been applied to achieve the optimal combination of hyperparameters to improve the performance of ML models. The workflow diagram of Approach C is presented in Figure 4, and the SMOTE-ENN algorithm is illustrated in Figure 5.

3.6. Experiment Environment. The experiment has been conducted using Jupyter Notebook v6.1.4 (Python 3 version 3.8.5) and Anaconda distribution v4.10.3 on an Intel Core i5-8300H CPU running at 2.30 GHz, 8 GB of RAM, and an NVIDIA GTX 1050 Ti graphics unit with 4 GB of dedicated memory.

4. Experimental Results

4.1. Approach A. The performance metrics such as accuracy, precision, F1, recall, and ROC AUC have been recorded and shown in Table 3. Here, Tables 4 to 9 illustrate the confusion matrices for this approach, and Figure 6 shows the ROC curve.

4.2. Approach B. In this approach, Table 10 summarizes the computational time required for both optimization methods (GSCV and RSCV), where it is seen that GSCV takes more time than RSCV in all algorithms. To prevent algorithms from being biased toward higher values, two scaling methods (standard scaler and min-max scaler) are utilized. Table 11 represents all the eight experiments done with these scaling and hyperparameter optimization methods. The receiver operating characteristic (ROC) curve for this method is shown in Figure 7, and the confusion matrices are included in Tables 12–17. The performance metrics of all mentioned algorithms are presented in Table 18, using the test dataset after scaling and using the hyperparameters obtained from hyperparameter optimization (HPO).

4.3. Approach C. Along with scaling and hyperparameter optimization (HPO), SMOTE-ENN is incorporated into the ML models in this experimental configuration which has enhanced the performances of the classifiers. The comparison of computational time between GSCV and RSCV is tabulated in Table 19. However, the investigation has been

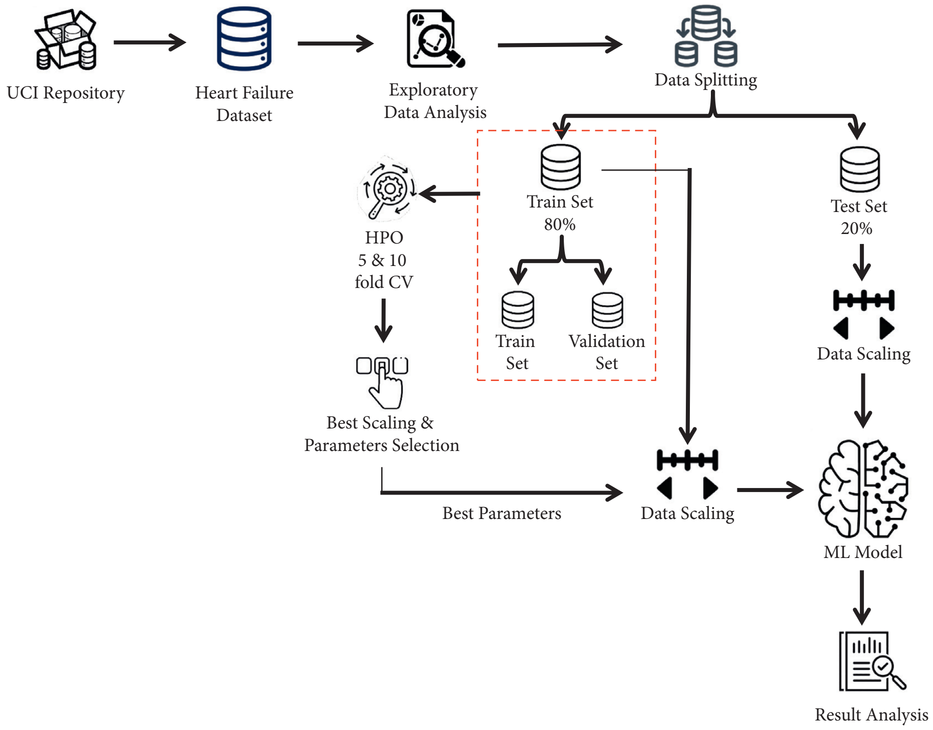


Figure 3: Workflow diagram of Approach B.

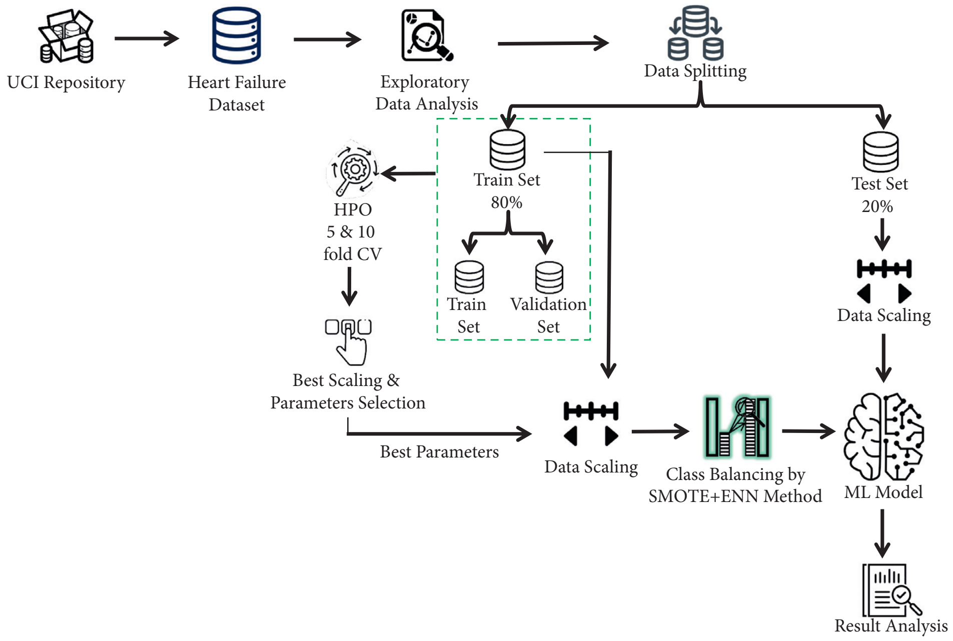


Figure 4: Overall workflow diagram for Approach C.

accomplished by using both 5 and 10-fold cross validation with “standard scaler” and “min–max scaler.” The eight experiments are depicted in Table 20, where it is seen that Support Vector Machine (SVM) with a value of 0.989 has showcased the highest accuracy among all, which has been obtained using Standard Scalar with a GSCV of 10-fold technique. The confusion matrices for the classifiers are presented in Tables 21–26, and the ROC curves for all the classifiers in our investigation are shown in Figure 8. By comparing the True Positive and False Positive rates, the ROC curve can determine the optimal classification model and eliminate suboptimal models [38].

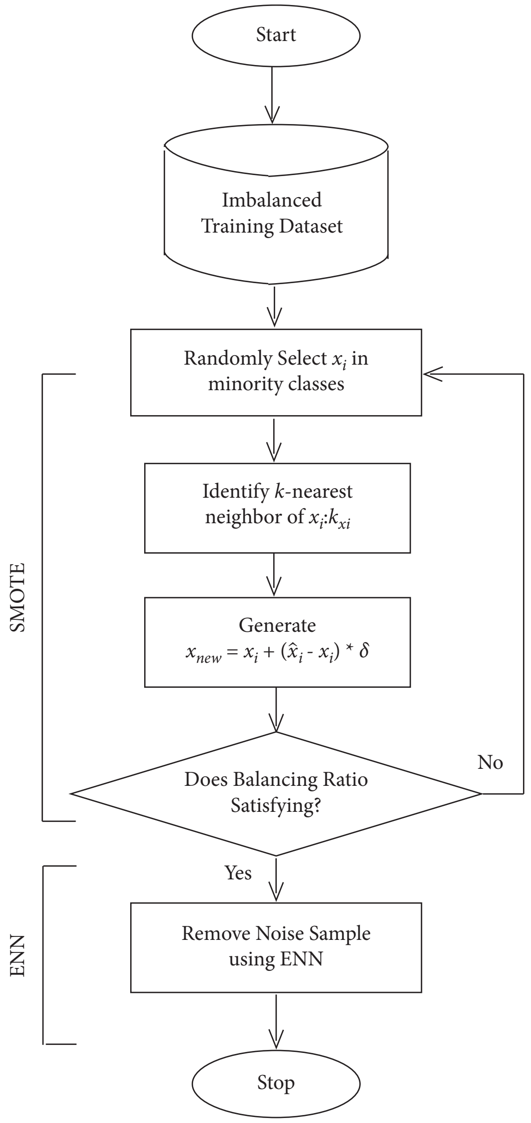

Start
Imbalanced Training Dataset
Randomly Select x_i in minority classes
Identify k-nearest neighbor of x_i:k_{xi}
Generate $x_{new} = x_i + (\hat{x}_i - x_i) * \delta$
Does Balancing Ratio Satisfying?
No
Yes
Remove Noise Sample using ENN
Stop
SMOTE
ENN

Figure 5: SMOTE-ENN algorithm.

Table 3: Performance metrics of ML algorithms by Approach A

Algorithms	Accuracy	Precision	F1	Recall	ROC_AUC
DTC	0.733	0.756	0.795	0.838	0.720
LR	**0.800**	0.878	**0.857**	0.837	0.755
GNB	0.683	**1.000**	0.812	0.683	0.500
RFC	**0.800**	0.854	0.854	**0.854**	**0.769**
KNN	0.667	0.902	0.787	0.698	0.530
SVM	0.683	**1.000**	0.812	0.683	0.500

Table 4: Confusion matrix of Decision Tree Classifier.

Decision Tree Classifier (DTC)		Predicted	
		Positive	Negative
Actual	True	31	13
	False	10	6

Table 5: Confusion matrix of Logistic Regression.

Logistic Regression (LR)		Predicted	
		Positive	Negative
Actual	True	36	12
	False	5	7

Table 6: Confusion matrix of Gaussian Naïve Bayes.

Gaussian Naïve Bayes (GNB)		Predicted	
		Positive	Negative
Actual	True	41	0
	False	0	19

Table 7: Confusion matrix of Random Forest Classifier.

Random Forest Classifier (RFC)		Predicted	
		Positive	Negative
Actual	True	35	13
	False	6	6

Table 8: Confusion matrix of K-Nearest Neighbors.

K-Nearest Neighbors (KNN)		Predicted	
		Positive	Negative
Actual	True	37	3
	False	4	16

Table 9: Confusion matrix of Support Vector Machine.

Support Vector Machine (SVM)		Predicted	
		Positive	Negative
Actual	True	41	0
	False	0	19

5. Discussion

5.1. Approach A. In this approach, from the measured performance metrics, evident in Table 9, it is seen that the Random Forest Classifier (RFC) outperforms in the majority of performance metrics with accuracy, recall, and ROC_AUC values of 0.800,0.854, and 0.769, respectively. However, precision is maximized by the GNB and SVM, and the F1 score is measured highest using the LR algorithm. The Decision Tree Classifier (DTC) has secured the second place in terms of accuracy, scoring 0.733, while LR and RFC have ranked first, scoring 0.800. In Figure 9, bar charts depict the comparison of algorithms in terms of performance metrics. The LR and RFC algorithms are likewise well functioning in this approach, as illustrated in Figure 9 and the ROC curve.

5.2. Approach B. In this approach, data scaling has been performed as [44] shows that the high deviation between the numeric values of the various characteristics can force ML algorithms to bias toward large values. However, hyperparameter optimization can also improve the performance in this type of case, as shown in [45–47]. As seen in Table 10, RSCV takes significantly less time than GSCV since it attempts

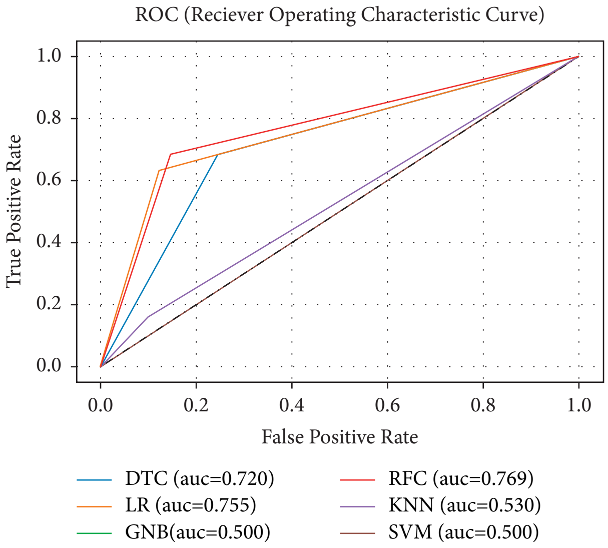


FIGURE 6: ROC curve for all the ML models for Approach A.

TABLE 10: Comparison of computational time.

Algorithms	Computation time Grid search CV (sec)	Computation time Random search CV (sec)
DTC	11.763	0.285
LR	4.207	0.244
GNB	0.140	0.140
RFC	88.745	6.175
KNN	4.160	0.598
SVM	3.955	1.011

TABLE 11: Highest accuracies of classifiers in conducted eight experiments.

Experiment no.	Preprocessing method	K fold CV	Hyperparameter optimization Method	Highest accuracy classifier	Highest accuracy
1	Standard scalar	5-fold	Random search	LR	0.845
2	Standard scalar	10-fold	Random search	RFC	0.854
3	Min–max scalar	5-fold	Random search	RFC	0.850
4	Min–max scalar	10-fold	Random search	RFC	0.858
5	Standard scalar	5-fold	Grid search	RFC	0.866
6	Standard scalar	10-fold	Grid search	RFC	**0.870**
7	Min–max scalar	5-fold	Grid search	RFC	0.862
8	Min–max scalar	10-fold	Grid search	RFC	0.866

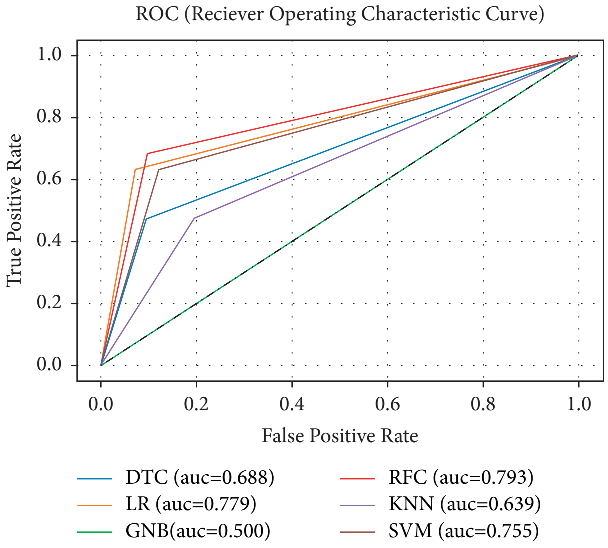


FIGURE 7: ROC curve of all the ML models for Approach B.

Table 12: Confusion matrix of Decision Tree Classifier.

Decision Tree Classifier (DTC)		Predicted	
		Positive	Negative
Actual	True	37	9
	False	4	10

Table 13: Confusion matrix of Logistic Regression.

Logistic Regression (LR)		Predicted	
		Positive	Negative
Actual	True	38	12
	False	3	7

Table 14: Confusion matrix of Gaussian Naïve Bayes.

Gaussian Naïve Bayes (GNB)		Predicted	
		Positive	Negative
Actual	True	41	0
	False	0	19

Table 15: Confusion matrix of Random Forest Classifier.

Random Forest Classifier (RFC)		Predicted	
		Positive	Negative
Actual	True	37	13
	False	4	6

Table 16: Confusion matrix of K-Nearest Neighbors.

K-Nearest Neighbors (KNN)		Predicted	
		Positive	Negative
Actual	True	33	9
	False	8	10

Table 17: Confusion matrix of Support Vector Machine.

Support Vector Machine (SVM)		Predicted	
		Positive	Negative
Actual	True	36	12
	False	5	7

Table 18: Performance metrics of ML algorithms by Approach B.

Algorithms	Accuracy	Precision	F1	Recall	ROC_AUC
DTC	0.767	0.902	0.841	0.787	0.688
LR	**0.833**	0.927	**0.884**	0.844	0.779
GNB	0.683	**1.000**	0.812	0.683	0.500
RFC	**0.833**	0.902	0.881	**0.860**	**0.793**
KNN	0.700	0.805	0.786	0.767	0.639
SVM	0.800	0.878	0.857	0.837	0.755

Table 19: Comparison of computational time (Approach C).

Algorithms	Computation time Grid search CV (sec)	Computation time Random search CV (sec)
DTC	11.428	0.310
LR	4.078	0.315
GNB	0.355	0.148
RFC	65.935	3.558
KNN	4.225	0.578
SVM	2.233	0.880

random combinations of hyperparameters rather than all combinations as GSCV does. However, GSCV performed better in terms of accuracy than RSCV. Figure 10 illustrates the contrast of computational time using a bar plot. Table 11 summarizes the findings from our eight experiments. Eight different combinations of scaling and cross-validation methods have been implemented, and the classifier with the highest accuracy has been identified. It is clear from this table that the GSCV and standard scaling techniques have provided the best performance. Here, RFC produces the best result, with an accuracy of 0.870, as determined by a 10-fold GSCV with standard scaling. As a result, Table 18 evaluates all performance measures using this combination. In this case, RFC exceeds all other algorithms in terms of accuracy, recall, and ROC_AUC. However, LR provides the highest F1 score and also accuracy. The precision is maximized by GNB in this approach. The best accuracy here is 0.833, as determined by RFC and LR, and the second highest is 0.800, as assessed by SVM. Figure 11 depicts a comparison of all algorithms based on performance metrics. This strategy produces a result that is significantly better than Approach A.

5.3. Approach C. This approach adds a class balancing technique called SMOTE-ENN as this dataset was highly imbalanced, with a class ratio of 203 : 96, meaning one class is nearly double that of the other. This kind of imbalance can prevent machine learning algorithms from performing correctly, and there is a tendency to prefer the majority class in the prediction. To correct the imbalance and improve the results, researchers used sampling techniques like SMOTE and SMOTE-ENN to balance class in this type of dataset [48,49]. SMOTE-ENN is used with scaling and hyperparameter optimization (HPO) in this study, yielding more promising outcomes. In Table 19, the computing time for this approach is compared, and it is clear that GSCV takes longer than RSCV, but GSCV provides better accuracy, which is graphically presented in Figure 12.

Following that, eight trials have been performed as Approach B; the SVM provides the highest accuracy with a value of 0.989, which is the best result of all three approaches in terms of accuracy. The best accuracy is currently found with 10-fold GSCV and standard scalar. Table 20 depicts a summary of the experiments. Following that, the performance metrics for this Approach were evaluated using the test dataset and parameters from standard scalar 10-fold GSCV and presented in Table 27. The RFC has the utmost accuracy of 0.900, followed by DTC 0.867. And it also wins in terms of F1 score, recall, and ROC_AUC. However, the DTC has the highest precision value here. The results obtained in this approach are far better than those of the other two approaches. Figure 13 shows a comparison of algorithms using Approach C.

Table 20: Highest accuracies of classifiers in conducted eight experiments (Approach C).

Experiment no.	Preprocessing method	K fold CV	Hyperparameter optimization method	Highest accuracy classifier	Highest accuracy (validation set)
1	Standard scalar	5-fold	Random search	SVM	0.970
2	Standard scalar	10-fold	Random search	SVM	0.970
3	Min–max scalar	5-fold	Random search	RFC	0.987
4	Min–max scalar	10-fold	Random search	RFC	0.980
5	Standard scalar	5-fold	Grid search	KNN	0.986
6	Standard scalar	10-fold	Grid search	SVM	**0.989**
7	Min–max scalar	5-fold	Grid search	RFC	0.981
8	Min–max scalar	10-fold	Grid search	KNN	0.983

Table 21: Confusion matrix of Decision Tree Classifier.

Decision Tree Classifier (DTC)		Predicted	
		Positive	Negative
Actual	True	38	14
	False	3	5

Table 22: Confusion matrix of Logistic Regression.

Logistic Regression (LR)		Predicted	
		Positive	Negative
Actual	True	29	17
	False	12	2

Table 23: Confusion matrix of Gaussian Naïve Bayes.

Gaussian Naïve Bayes (GNB)		Predicted	
		Positive	Negative
Actual	True	33	13
	False	8	6

Table 24: Confusion matrix of Random Forest Classifier.

Random Forest Classifier (RFC)		Predicted	
		Positive	Negative
Actual	True	36	18
	False	5	1

Table 25: Confusion matrix of K-Nearest Neighbors.

K-Nearest Neighbors (KNN)		Predicted	
		Positive	Negative
Actual	True	32	15
	False	9	4

In terms of all performance metrics, Approach C outperforms all other experimental approaches. The best accuracy was found to be 80% in Approach A, 83.3% in Approach B, and 90% in Approach C, indicating the models' successive improvement. The final Approach C performed exceptionally well in predicting the survival of patients with heart failure. Figures 14(a)–14(e) shows the comparison of three approaches based on accuracy, precision, F1 score, recall, and ROC_AUC.

Table 26: Confusion matrix of Support Vector Machine.

Support Vector Machine (SVM)		Predicted	
		Positive	Negative
Actual	True	31	16
	False	10	3

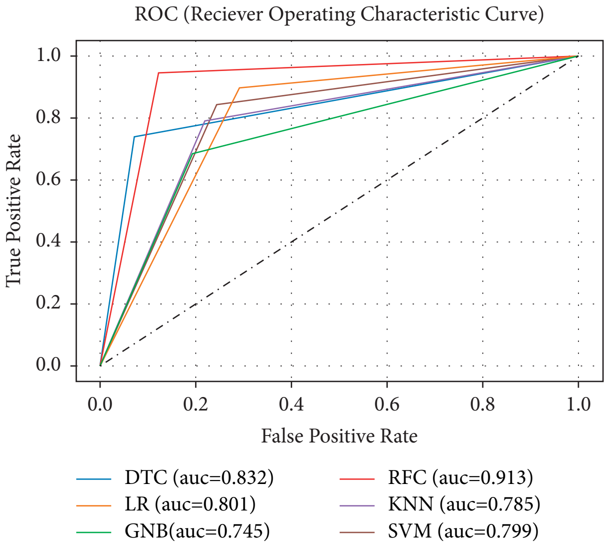


Figure 8: ROC curve of all the ML models for Approach C.

Finally, a detailed comparative analysis has been portrayed in Table 28, where the best accuracies obtained by different researchers have been presented. It is evident that the proposed method (Approach C) depicted the highest

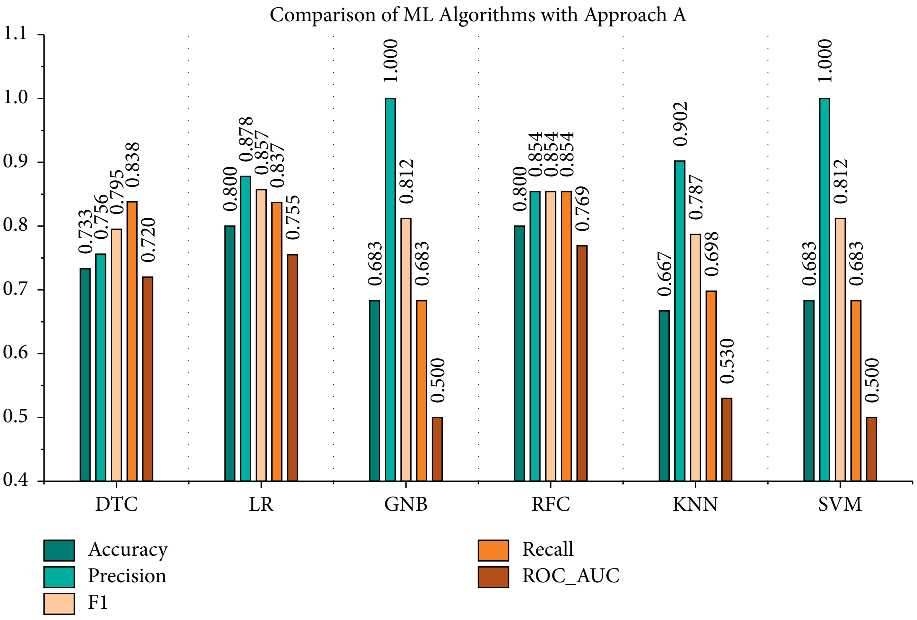


FIGURE 9: Comparative analysis of the performance metrics by Approach A.

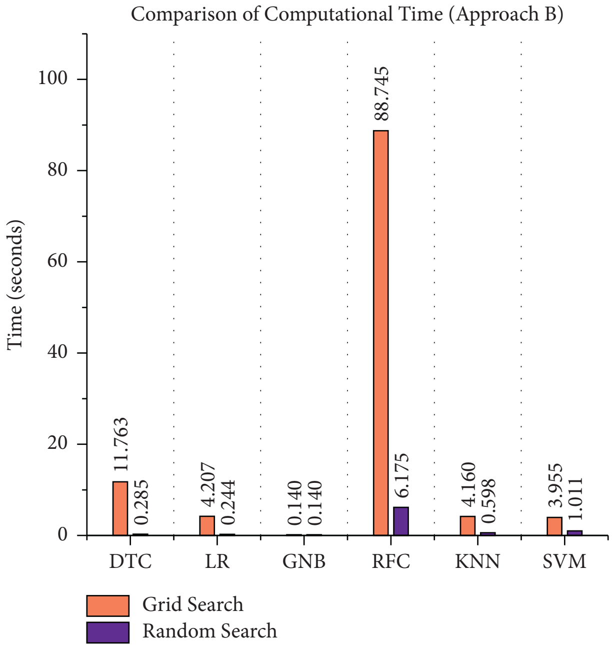


FIGURE 10: Comparison of computational time by Approach B.

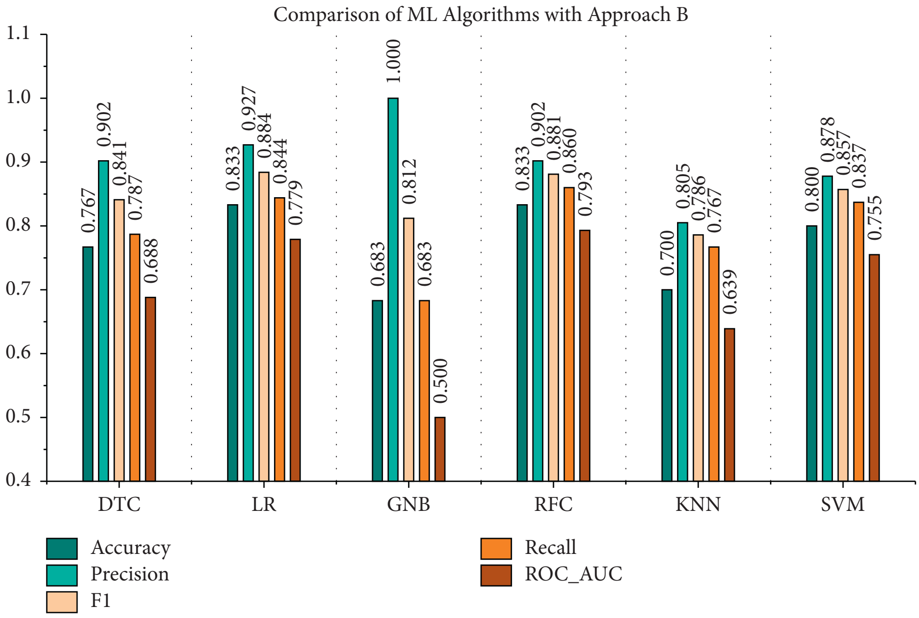


Figure 11: Comparative analysis of the performance metrics by Approach B.

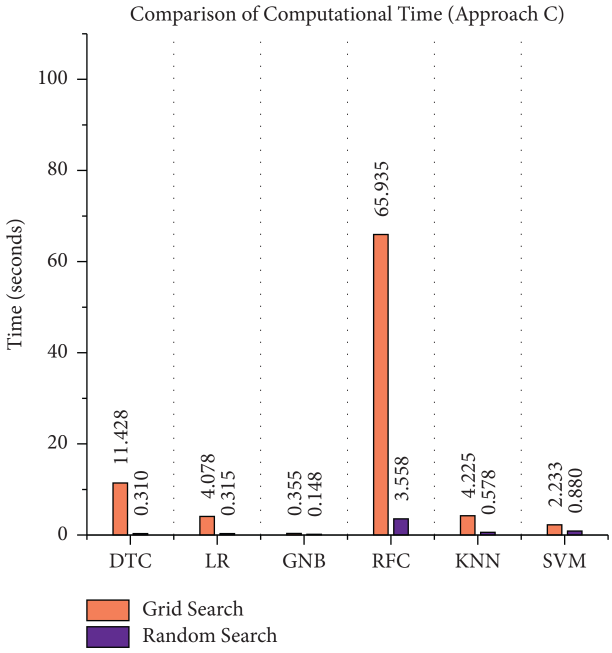


Figure 12: Comparison of computational time by Approach C.

Table 27: Performance metrics of ML algorithms with Approach C.

Algorithms	Accuracy	Precision	F1	Recall	ROC_AUC
DTC	0.867	**0.927**	0.905	0.884	0.832
LR	0.767	0.707	0.806	0.936	0.801
GNB	0.767	0.805	0.825	0.846	0.745
RFC	**0.900**	0.878	**0.923**	**0.973**	**0.913**
KNN	0.783	0.781	0.831	0.889	0.785
SVM	0.783	0.756	0.827	0.912	0.799

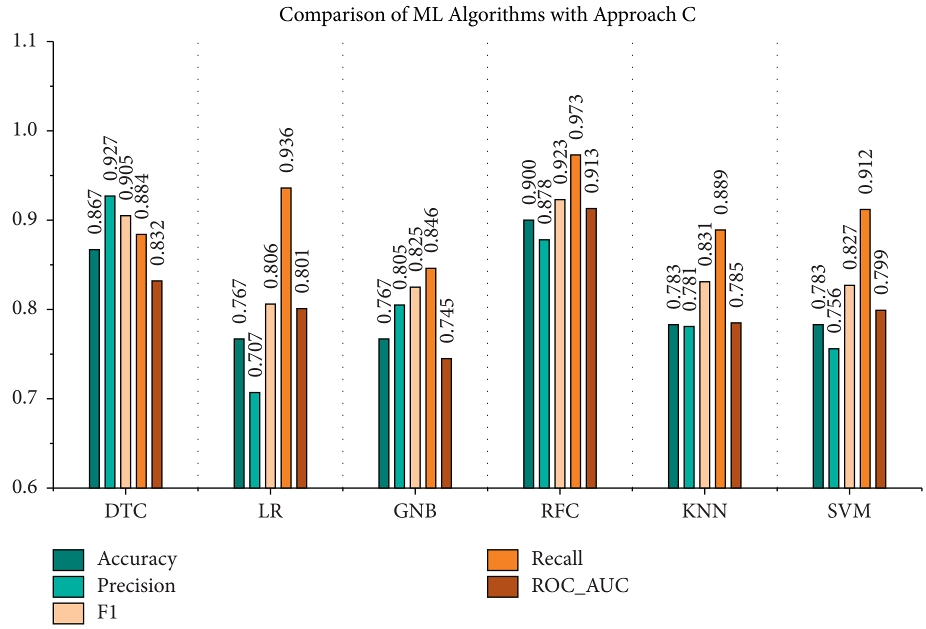
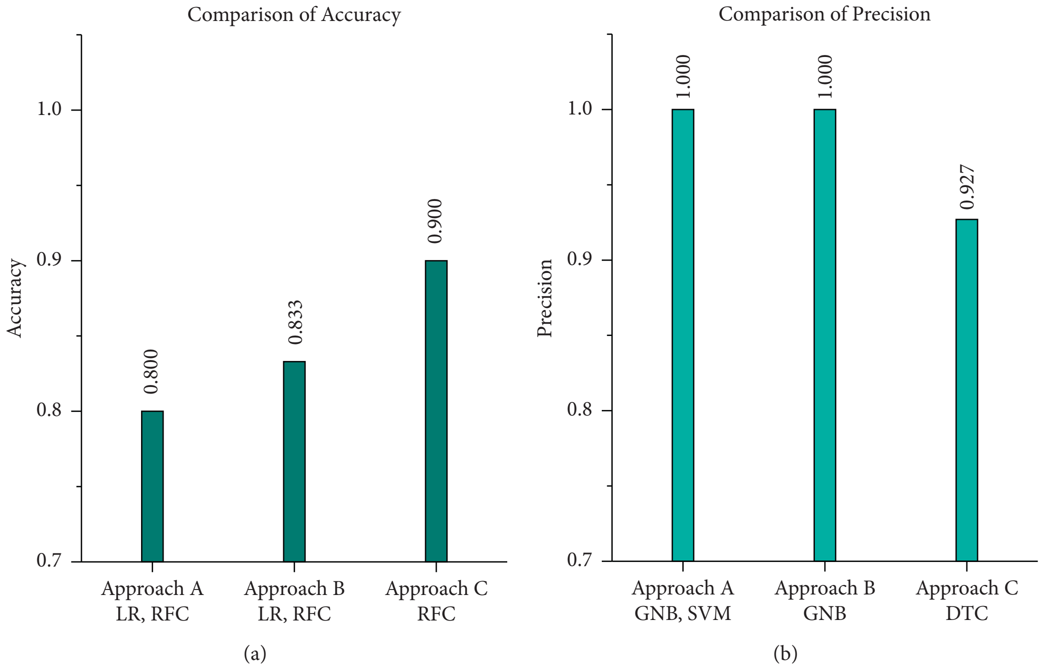


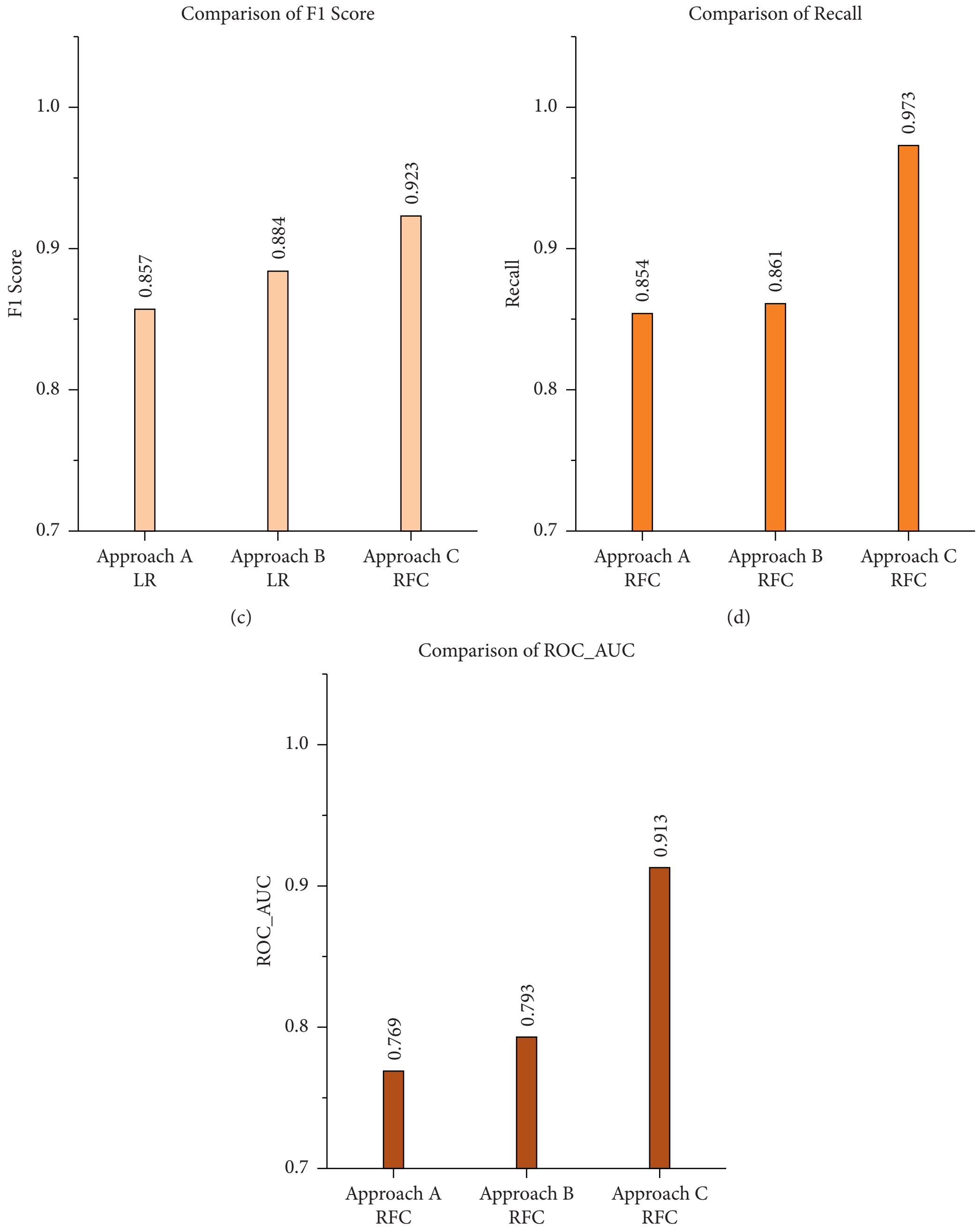
Figure 13: Comparative analysis of the performance metrics by Approach C.



(a)

(b)

Figure 14: Continued.



(c)

(d)

(e)

Figure 14: Comparison of (a) accuracy, (b) precision, (c) F1 score, (d) recall, and (e) ROC_AUC.

Table 28: Comparison of performance with other works.

References	Authors	Algorithms	Best accuracy (validation set)	Best accuracy (test set)
[31]	O. O. Oladimeji and O. Oladimeji	Random Forest Classifier (RFC)	83.1788%	—
[30]	D. Chicco and G. Jurman	Logistic Regression (LR)	83.80%	—
[34]	P. E. Rubini and C. A. Subasini	Random Forest Classifier (RFC)	84.81%	—
[32]	R. Gürfidan and M. Ersoy	Support Vector Machine (SVM)	90.00%	—
[33]	S. Elyassami and A. A. Kaddour	Stochastic gradient descent with chi-square test	91.43%	—
[35]	A. Ishaq et al.	Extra Tree Classifier (ETC) with SMOTE	92.62%	—
[37]	L. Ali et al.	Gaussian Naïve Bayes (GNB) with $\chi 2$ statistical model	93.33%	—
[36]	S. Rahayu and J. Jaya Purnama	Random Forest Classifier (RFC) with SMOTE	94.31%	—
This work		KNN SVM with SMOTE-ENN	98.9% (SVM)	90% (RFC)

validation accuracy of 98.9% and test accuracy of 90%. Therefore, this approach can impose a significant contribution in predicting the survival of patients with heart failure in an efficient way.

6. Conclusion

As heart failure is extremely perilous and prevention is critical, patients must seek the advice of healthcare professionals in a regular fashion. However, healthcare professionals should consider a variety of conditions and parameters when advising or treating patients. On the other hand, the diagnostic instruments and expert medical technologists are insufficient in many cases, and a prompt and accurate diagnosis of the patient's condition is quite challenging. That is why vast amounts of data are collected and analyzed for real-world patient scenarios to assist healthcare professionals. Machine learning and data mining have enormous potential for revealing hidden patterns in large datasets from the clinical domain. These patterns can be used to assist physicians in diagnosing patients. It is a more efficient and advanced technique than statistics for analyzing large amounts of data because it allows for prediction based on prior cases and enables healthcare professionals to make informed decisions. In this study, three different approaches are undertaken to investigate the performances of the ML models in predicting the survival of patients with heart failure. It is observed that Approach C outperforms the other two approaches significantly in terms of accuracy, F1 score, recall, and ROC_AUC. Therefore, it is evident that SMOTE-ENN and hyperparameter optimization have played a significant role in enhancing the performances of the classifiers. Approach C has the best test accuracy of 90%, followed by approaches A and B with 80% and 83.33%. Additionally, Approach C ranks on the top among other approaches in terms of F1 score (0.923), recall (0.973), and ROC AUC (0.913), respectively. On the other hand, Approaches A and B have showcased the values of F1 score of 0.857 and 0.884, recall values of 0.854 and 0.860, and ROC AUC values of 0.769 and 0.793 correspondingly. Therefore, it is evident that RFC (with SMOTE-ENN technique and hyperparameter optimization) triumphs over all other approaches and obtained 90% accuracy with the test dataset. Hence, this study can make a notable contribution in predicting the survival of patients with heart failure and can aid in developing an automated computer-aided diagnosis system for e-healthcare applications.

Data Availability

Heart failure clinical records dataset from UCI Machine Learning Repository was used in order to support this study and is available at "https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records". This prior study and dataset are cited at relevant places within the text as references [30,38].

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] World Health Organization, "Cardiovascular diseases (CVDs)," 2021, https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-cvds.

[2] Nhlbi Nih, "Heart failure," 2021, https://www.nhlbi.nih.gov/health-topics/heart-failure.

[3] K. F. Adams Jr, G. C. Fonarow, C. L. Emerman et al., "Characteristics and outcomes of patients hospitalized for heart failure in the United States: rationale, design, and preliminary observations from the first 100,000 cases in the Acute Decompensated Heart Failure National Registry (ADHERE)," *American Heart Journal*, vol. 149, no. 2, pp. 209–216, 2005.

[4] G. C. Fonarow, W. T. Abraham, N. M. Albert et al., "Organized program to initiate lifesaving treatment in hospitalized patients with heart failure (OPTIMIZE-HF): rationale and design," *American Heart Journal*, vol. 148, no. 1, pp. 43–51, 2004.

[5] American Heart Association, "What is heart failure?," 2020, https://www.heart.org/en/health-topics/heart-failure/what-is-heart-failure.

[6] A. Mosterd and A. W. Hoes, "Clinical epidemiology of heart failure," *Heart*, vol. 93, no. 9, pp. 1137–1146, 2007.

[7] F. Shamsham and J. Mitchell, "Essentials of the diagnosis of heart failure," *American Family Physician*, vol. 61, no. 5, pp. 1319–1328, 2000.

[8] E. Tanai and S. Frantz, "Pathophysiology of heart failure," *Comprehensive Physiology*, vol. 6, no. 1, pp. 187–214, 2016.

[9] G. F. Mendez and M. R. Cowie, "The epidemiological features of heart failure in developing countries: a review of the literature," *International Journal of Cardiology*, vol. 80, no. 2–3, pp. 213–219, 2001.

[10] M. M. Redfield, S. J. Jacobsen, J. C. Burnett, D. W. Mahoney, K. R. Bailey, and R. J. Rodeheffer, "Burden of systolic and diastolic ventricular dysfunction in the Community," *JAMA*, vol. 289, no. 2, pp. 194–202, 2003.

[11] J. J. V. McMurray and M. A. Pfeffer, "Heart failure," *Lancet*, vol. 365, no. 9474, pp. 1877–1889, 2005.

[12] R. E. Lane, M. R. Cowie, and A. W. C. Chow, "Prediction and prevention of sudden cardiac death in heart failure," *Heart*, vol. 91, no. 5, pp. 674–680, 2005.

[13] M. Piepoli, "Editorials Diagnostic and prognostic indicators in chronic heart failure," *European Heart Journal*, vol. 20, no. 19, pp. 1367–1369, 1999.

[14] B. Ziaeian and G. C. Fonarow, "Epidemiology and aetiology of heart failure," *Nature Reviews Cardiology*, vol. 13, no. 6, pp. 368–378, 2016.

[15] A. Onan, "A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer," *Expert Systems with Applications*, vol. 42, no. 20, pp. 6844–6852, 2015.

[16] K. Harimoorthy and M. Thangavelu, "Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 3, pp. 3715–3723, 2021.

[17] A. Onan and S. Korukoğlu, "Exploring performance of instance selection methods in text sentiment classification," *Advances in Intelligent Systems and Computing*, vol. 464, pp. 167–179, 2016.

[18] M. A. A. R. Asif, M. M. Nishat, F. Faisal et al., "Performance evaluation and comparative analysis of different machine learning algorithms in predicting cardiovascular disease," *Engineering Letters*, vol. 29, no. 2, pp. 731–741, 2021.

[19] L. Yang, H. Wu, X. Jin et al., "Study of cardiovascular disease prediction model based on random forest in eastern China," *Scientific reports*, vol. 10, no. 1, pp. 5245–5248, 2020.

[20] A. Onan and S. Korukoğlu, "A feature selection model based on genetic rank aggregation for text sentiment classification," *Journal of Information Science*, vol. 43, no. 1, pp. 25–38, 2017.

[21] M. M. Nishat, F. Faisal, R. R. Dip et al., "Performance investigation of different Boosting algorithms in predicting chronic Kidney disease," in *Proceedings of the 2020 2nd International Conference on Sustainable Technologies for Industry*, pp. 1–5, IEEE, 2020.

[22] A. Onan, "An ensemble scheme based on language function analysis and feature engineering for text genre classification," *Journal of Information Science*, vol. 44, no. 1, pp. 28–47, 2018.

[23] I. Babaoǧlu, O. Findik, and M. Bayrak, "Effects of principle component analysis on assessment of coronary artery diseases using support vector machine," *Expert Systems with Applications*, vol. 37, no. 3, pp. 2182–2185, 2010.

[24] A. Onan, S. Korukoğlu, and H. Bulut, "Ensemble of keyword extraction methods and classifiers in text classification," *Expert Systems with Applications*, vol. 57, pp. 232–247, 2016.

[25] M. A. A. R. Asif, M. M. Nishat, F. Faisal et al., "Computer aided diagnosis of Thyroid disease using machine learning algorithms," in *Proceedings of the 2020 11thInternational Conference on Electrical and Computer Engineering (ICECE)*, pp. 222–225, IEEE, 2020.

[26] A. Onan, S. Korukoğlu, and H. Bulut, "A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification," *Information Processing & Management*, vol. 53, no. 4, pp. 814–833, 2017.

[27] A. Onan and M. A. Tocoglu, "A term weighted Neural language model and stacked Bidirectional LSTM based Framework for sarcasm Identification," *IEEE Access*, vol. 9, pp. 7701–7722, 2021.

[28] M. M. Nishat, F. Faisal, R. R. Dip et al., "A comprehensive analysis on detecting chronic Kidney disease by employing machine learning algorithms," *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 21, no. 29, p. e1, 2021.

[29] T. Ahmad, A. Munir, S. H. Bhatti, M. Aftab, and M. A. Raza, "Survival analysis of heart failure patients: a case study," *PLoS One*, vol. 12, no. 7, p. e0181001, 2017.

[30] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, p. 16, 2020.

[31] O. O. Oladimeji and O. Oladimeji, "Predicting survival of heart failure patients using classification algorithms," *JITCE (Journal of Information Technology and Computer Engineering)*, vol. 4, no. 2, pp. 90–94, 2020.

[32] R. Gürfidan and M. Ersoy, *Classification of Death Related to Heart Failure by Machine Learning Algorithms*, http://www.dergipark.com/aair/, 2021.

[33] S. Elyassami and A. Ait Kaddour, "Implementation of an incremental deep learning model for survival prediction of cardiovascular patients," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 1, pp. 101–109, 2021.

[34] P. E. Rubini, C. A. Subasini, A. Vanitha Katharine, V. Kumaresan, S. Gowdhamkumar, and T. M. Nithya, "A cardiovascular disease prediction using machine learning algorithms," *Annals of the Romanian Society for Cell Biology*, vol. 25, no. 2, pp. 904–912, 2021.

[35] A. Ishaq, S. Sadiq, M. Umer et al., "Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques," *IEEE Access*, vol. 9, pp. 39707–39716, 2021.

[36] S. Rahayu, J. Purnama, A. Pohan, F. Nugraha, S. Nurdiani, and S. Hadianti, "Prediction of survival of heart failure patients using random forest," *Journal of Pilar Nusa Mandiri*, vol. 16, no. 2, pp. 255–260, 2020.

[37] L. Ali, S. U. Khan, N. A. Golilarz et al., "A feature-driven decision support system for heart failure prediction based on statistical model and Gaussian naive bayes," *Computational and Mathematical Methods in Medicine*, vol. 2019, 2019.

[38] UCI Machine Learning Repository, "Heart failure clinical records Data Set," 2020, https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records.

[39] J. Grus, *Data Science from Scratch*, O'Reilly Media, Inc., 2015.

[40] N. Noorhalim, A. Ali, and S. M. Shamsuddin, "Handling imbalanced ratio for class imbalance problem using SMOTE," in *Proceedings of the Third International Conference on Computing, Mathematics and Statistics*, pp. 19–30, 2019.

[41] T. Le, M. T. Vo, B. Vo, M. Y. Lee, and S. W. Baik, "A hybrid approach using oversampling technique and cost-sensitive learning for Bankruptcy prediction," *Complexity*, vol. 2019, Article ID 8460934, 2019.

[42] H. J. P. Weerts, A. C. Mueller, and J. Vanschoren, "Importance of Tuning Hyperparameters of Machine Learning Algorithms," 2020, http://arxiv.org/abs/2007.07588.

[43] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.

[44] S. Mezzatesta, C. Torino, P. D. Meo, G. Fiumara, and A. Vilasi, "A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis," *Computer Methods and Programs in Biomedicine*, vol. 177, pp. 9–15, 2019.

[45] R. S. De Andrades, M. Grellert, and M. B. Fonseca, "Hyperparameter tuning and its effects on cardiac arrhythmia prediction," in *Proceedings of the 2019 8th Brazilian Conference on Intelligent Systems*, pp. 562–567, BRACIS, 2019.

[46] S. Ambesange, A. Vijayalaxmi, S. Sridevi, Venkateswaran, and B. S. Yashoda, "Multiple heart diseases prediction using logistic regression with ensemble and hyper parameter tuning techniques," in *Proceedings of the 2020 Fourth World Conference on Smart Trends in System, Security and Sustainability*, pp. 827–832, WS4, 2020.

[47] M. V Sonth, S. Ambesange, D. Sreekanth, and S. Tulluri, "Optimization of random forest algorithm with ensemble and hyper parameter tuning techniques for Multiple heart diseases," *Solid State Technology*, vol. 63, no. 5, pp. 3961–3972, 2020.

[48] L. Hussain, K. J. Lone, I. A. Awan, A. A. Abbasi, and J.-u.-R. Pirzada, "Detecting congestive heart failure by extracting multimodal features with synthetic minority oversampling technique (SMOTE) for imbalanced data using robust machine learning techniques," *Waves in Random and Complex Media*, vol. 2020, pp. 1–24, 2020.

[49] K. Wang, J. Tian, C. Zheng et al., "Improving risk Identification of Adverse outcomes in chronic heart failure using SMOTE+ENN and machine learning," *Risk Management and Healthcare Policy*, vol. 14, pp. 2453–2463, 2021.