Hindawi
Journal of Sensors
Volume 2022, Article ID 3730303, 21 pages
https://doi.org/10.1155/2022/3730303

Research Article

Prediction of Cardiovascular Disease on Self-Augmented Datasets of Heart Patients Using Multiple Machine Learning Models

Sumaira Ahmed,[1] **Salahuddin Shaikh**,[1] **Farwa Ikram**,[2] **Muhammad Fayaz**,[3] **Hathal Salamah Alwageed**,[4] **Faheem Khan**,[5] **and Fawwad Hassan Jaskani**[6]

[1]*Centre of Computing Research, Department of Computer Science and Software Engineering, Jinnah University for Women, Karachi 74600, Pakistan*
[2]*Department of Computer Engineering, University of Lahore, Pakistan*
[3]*Department of Computer Science, University of Central Asia, Naryn, Kyrgyzstan*
[4]*College of Computer and Information Science, Jouf University, Saudi Arabia*
[5]*Gachon University, Department of Computer Engineering, Republic of Korea*
[6]*Department of Computer Systems Engineering, Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan*

Correspondence should be addressed to Muhammad Fayaz; muhammad.fayaz@ucentralasia.org

Received 24 June 2022; Revised 14 October 2022; Accepted 27 October 2022; Published 23 December 2022

Academic Editor: Rajesh Kaluri

About 26 million people worldwide experience its effects each year. Both cardiologists and surgeons have a tough time determining when heart failure will occur. Classification and prediction models applied to medical data allow for enhanced insight. Improved heart failure projection is a major goal of the research team using the heart disease dataset. The probability of heart failure is predicted using data mined from a medical database and processed by machine learning methods. It has been shown, through the use of this study and a comparative analysis, that heart disease may be predicted with high precision. In this study, researchers developed a machine learning model to improve the accuracy with which diseases like heart failure (HF) may be predicted. To rank the accuracy of linear models, we find that logistic regression (82.76 percent), SVM (67.24 percent), KNN (60.34 percent), GNB (79.31 percent), and MNB (72.41) perform best. These models are all examples of ensemble learning, with the most accurate being ET (70.31%), RF (87.03%), and GBC (86.21%). DT (ensemble learning models) achieves the highest degree of precision. CatBoost outperforms LGBM, HGBC, and XGB, all of which achieve 84.48% accuracy or better, while XGB achieves 84.48% accuracy using a gradient-based gradient method (GBG). LGBM has the highest accuracy rate (86.21 percent) (hypertuned ensemble learning models). A statistical analysis of all available algorithms found that CatBoost, random forests, and gradient boosting provided the most reliable results for predicting future heart attacks.

1. Introduction

Patients often undergo a battery of tests, putting them under unnecessary mental, emotional, and financial strain. Tobacco use, excessive body fat, and cardiovascular disease have all been linked in studies [1]. Pain in the arms and chest is the most common indicator. Cardiac surgeons can benefit from a thorough examination of such a dataset for both diagnostic and operational purposes [2]. It has been attempted in the past [2] to enhance the HF diagnostic process through the use of learning machines and heart disease categories. This project aims at exploring different machine learning techniques and making better use of healthcare data. It is anticipated that classifier efficiency would rise. Heart failure (HF) and other health risks are affected by an individual's unique set of circumstances. Standard HF risk prediction models consider each variable as a covariate, but this approach ignores important characteristics like cardiac

biomarkers. Machine learning (ML) may be more effective than existing modelling approaches for predicting high blood pressure in particular races and explaining key aspects in the development of high blood pressure in diverse races [3]. Most cases of heart failure may be traced back to issues with the anatomy or physiology of the heart. This causes a rise in intracardiac pressure and/or a decrease in cardiovascular output, depending on whether you are at rest or under stress. Because of this, HF has been linked to a lower quality of life and less effort put into physical and mental activities. It is estimated that 1-2% of the general population and 10% of the elderly population in developed countries suffer with HF. The prevalence of heart failure is expected to increase as our population ages (HF). Patients with heart failure (HF) had a 56.6% readmission rate after being discharged from the hospital. When it comes to high frequency (HF), ignoring it now will cause serious issues in the future. One of the most pressing needs right now is to cut down on readmissions.

Many times, people will substitute heart disorders with cardiovascular diseases. Diseases of this type typically involve constricted blood arteries, which can lead to a stroke, chest pain or angina, or a heart attack. Diseases of the heart can also damage the heart's rhythm, valves, or muscles. However, detecting cases of heart disease requires the use of machine learning. Either way, if these can be anticipated in advance, it will be much easier for doctors to learn vital information that is needed for treating and diagnosing patients. It is important to note that heart illness is primarily an erroneous sign of coronary artery disease. Heart disease is distinct from cardiovascular disease, which encompasses a wide range of issues affecting the circulatory system. Python is an object-oriented programming language with a wide range of dynamic building tools and short iteration cycles. Based on the findings of [1] research, it is widely considered to be one of the most secure programming languages with many potential medical applications. In addition, it is widely considered as a popular and accepted programming language, spanning the development of software based on artificial intelligence and many others. Python's convenient foundation makes it simple to develop a programme for the desktop or the web, as recommended in [2]. According to the illustration [2], using Python programming in the healthcare sectors, and particularly for detecting cardiac illnesses, will allow physicians and institutions to give better and improved outcomes for the patients through scalable and dynamic applications. However, pandas, Matplotlib, IPython, NumPy, Python, SciPy, and a plethora of additional coding packages and libraries are being used on this project. Many hospitals and individuals are contributing to a growing data pool in the healthcare industry. Doctors may readily foresee better techniques of therapy and improve the entire healthcare delivery system [3] by making the best use of this data. The Python framework has several key applications; one of the most notable being its ability to assist in making sense of and stimulate computational facilities in extracting meaningful insights from data throughout the healthcare sectors. In addition, Python is widely recognised as one of the best programming languages in the world. It was voted as the most secure language for creating healthcare applications by 32% of UK residents.

Patients experiencing heart failure are often hospitalised for an extended amount of time after the initial episode. Routine blood samples are taken from patients to record a variety of health statistics [4]. Nonhaematological information can be gathered, such as age, gender, and smoking history. Once the data has been analysed, it will be difficult to determine if the patient's health is stable or worsening. The goal of ML algorithms is to learn about the environment and predict future events based on the data provided by users. People with this trait are adaptable, and they can use their past experiences to inform their present-day judgments. Multiple sclerosis is just one of many diseases for which these algorithm models are used in the diagnostic process. The major purpose of this study is to develop a method for estimating the likelihood of a patient dying from heart failure. Multiple machine learning (ML) models, such as logistic regression, support vector machine (SVM), random forest, and naïve Bayes, are used in today's data analysis [5, 6].

Atherosclerosis is the leading cause of cardiac arrest. It goes by a number of names, including hypertension and cardiovascular illness. Nearly 26 million people around the world suffer from cardiovascular disease [7]. This asymmetry makes it hard for the analyst to make accurate predictions about the outcome and the likelihood of survival [6]. Survival rates are also affected by a variety of sociodemographic characteristics, including gender, smoking status, the presence of chronic health conditions, and high blood pressure. Because survival rates for people with heart disease are so variable, it is hard to make accurate prognostic predictions for them [8].

If researchers do not find solutions to the problems we have outlined, we should expect to see this ratio climb in the next years. A healthy lifestyle and a strict diet are not the only things that need to be done to save lives. Patients with heart failure who are diagnosed using their medical records have had their prognoses improved with the application of machine learning techniques.

The goals of this study are as follows:

(a) To collect patient-specific medical and demographic data, including but not limited to age, gender, smoking status, blood pressure, body mass index (BMI), and anaemia

(b) Outliers on the datasets have been addressed as part of the data cleaning and preparation procedure

(c) Linear, ensemble, and boosting-based machine learning algorithms are utilised to diagnose cardiac illness in its earliest stages

(d) To compare the proposed approach to other machine learning methods like linear modes and ensemble model learning

2. Related Work

Heart failure is an area where machine learning models and analysis [9, 10] can be found. Heart failure occurs when the

heart becomes too weak to adequately pump blood throughout the body. For many people, this is their ultimate fate. About 2% of people in affluent countries suffer from heart failure, and that number rises to 6%-10% among those aged 65 and up [6].

Patients in intensive care unit cohorts who could benefit from closer observation, more aggressive therapy, or hospice care could be identified with the help of a death prediction model that better reflects reality. Predictions of mortality due to heart failure can be made using a variety of risk models [11].

Standard clinical risk factors, such as hypertension and diabetes, are used for most applications. High-risk areas for deaths due to cardiovascular disease have also been the subject of research in Brazil. Moreover, half of the people analysed in this study had multiple difficult-to-treat diseases that increased their risk of dying. Insufficient studies have been conducted on this topic [12]. When lab results are evaluated more realistically, it can be more challenging to establish a baseline for normalcy. Using reliable survival models is crucial for estimating risk [13]. Such predictors are ignored by models using conventional proportional risks, such as the Cox proportional hazard model.

Using computational methods to automatically establish relationships between components and big data response values, conventional proportional risk models can improve their capacity to uncover meaningful predictors [1]. Many machine learning techniques, such as the Bayesian network, decision trees, and association rules [14], have found use in the medical field. Nonparametric survival woodlands are a viable alternative to parametric and semiparametric models due to their independence from a time axis in characterising interactions [15]. When combined, survival trees and the ensemble approach produce better predictions.

By automatically evaluating nonlinear effects and intricate interactions between many elements, RSF was designed as a nonparametric enlargement of the random forest [16]. Experiential data shows that RSF-based risk models for some diseases, such as heart failure and breast cancer, have shown modest improvement [2, 17]. Recent research has shown that RSF is unable to identify statistically significant predictors in sample sizes that are too small. New class III antiarrhythmic drug discovery is hindered by RSF's interruption criterion, which requires a certain threshold of human deaths before stopping the search. Because of its usefulness in reducing mortality in cardiovascular patients [6], this cannot be discounted as a predictor. Researchers have employed machine learning algorithms to forecast both the likelihood of a heart attack and the likelihood of a patient's survival after one. Table 1 shows the comparative analysis based on datasets.

KNN and SVM are just two examples of the many supervised learning algorithms that have been put to use in the prediction of heart failure [28, 29]. In [30], the authors elaborate extensively on the supervised machine learning models they employ. This research takes a look at cancer and heart disease statistics using five ML algorithms. The approach has been shown to accurately predict breast cancer and other diseases by the authors. Fundamental reasons of these illnesses are also being investigated. When evaluating breast cancer data, principal component regression (PCR) and random forest (RF) are the best approaches. The authors argue that heart illness prediction can be accomplished with the use of machine learning [31]. Using WEKA, a number of decision tree categorization strategies were compared and contrasted. Many different algorithms, including random forest and logistic model tree (J48), have been examined. Researchers at UCI use the Cleveland Heartland Registry to screen for and validate heart disease in patients. This dataset contains the following types of information. The optimum algorithm for large-scale classification will then be suggested. Data mining can be used to find correlations between patient data and heart disease risk factors in order to get more accurate diagnoses for patients.

The authors' study concludes that machine learning can be used to anticipate cardiac emergencies. The subfield of artificial intelligence known as "machine learning" focuses on the process of teaching a computer to learn new things on its own. They are under constant pressure because of the prevalence of heart attacks among their patients. It is crucial to find ways to reduce the number of deaths caused by heart attacks. Machine learning plays an important role in this study. Scientists have developed a way to reliably predict cardiovascular disease in patients [32]. In this study, we used logistic regression, random forest, and artificial neural network with the ReLU-activated neural network (NNR), *K*-nearest neighbors (KNN), and GNB voting methods to predict the probability of cardiac sickness. The model was developed in Jupyter Notebook with Flask and Python using the Kaggle dataset. To evaluate the model's effectiveness in each of these contexts, we run tests on a wide range of parameters. Another investigation revealed that the test was accurate 90% of the time, precise 91% of the time, and accurate and precise 91% of the time. Since ensemble modelling is more precise than utilising individual models, these findings demonstrate that it contributes to saving lives. It is explained in [1] how a more accurate model for predicting survival in people with heart failure can be developed. Two-hundred-nine patients with heart failure are used to evaluate a survival prediction model. To determine the best ensemble tree method and feature selection approach, a data pipeline is utilised. Accuracy on five of the twelve variables increased from 79.5 to 85.1 using the extra tree classifier thanks to the use of new data and cross-validation (follow-up time, serum creatinine, ejection fraction, age, and diabetes) [33, 34].

So say the authors [35], machine learning allows for the foresight of cardiac arrests. An individual's risk of developing heart disease can be estimated using machine learning. The likelihood that an individual would develop heart disease over the course of their lifetime might be predicted by a machine learning algorithm given access to a large enough dataset. The likelihood of developing cardiovascular disease depends on the individual's present way of life and food. Framingham Heart Study information was used to inform the development of the model. It is the authors' contention that machine learning can be used to forecast who will develop heart failure—type 2 diabetic individuals and an innovative machine learning method for foreseeing the onset of heart failure (T2DM). The authors developed and validated a machine learning-based risk score using publicly available clinical, laboratory, and ECG data.

Table 1: Comparative analysis based on datasets.

Reference	Dataset	Techniques	Accuracy
Moreno-Sanchez [7]	They used a dataset that comprises 299 patients who suffered heart failure.	Receiver operating characteristic (ROC) curve	87.5%
Vijayashree and Iyengar [18]	The UCI machine learning repository provided the HF Indian heart attack dataset.	Post hoc techniques	—
Marimuthu et al. [19]	They used heart disease dataset.	*K*-means and fuzzy *C*-means clustering and random forest, XGBoost, and decision tree	87%
Golande and Kumar [20]	The data comes from a database of heart disease patients' medical records.	SVM (support vector machine)	—
Ayers et al. [21]	Training and validation datasets are used.	*K*-nearest neighbors (KNN), naïve Bayes, and support vector machine (SVM)	84.81%
Haq et al. [22]	It is necessary to make use of a vast number of disparate electronic datasets.	Cardiac magnetic resonance (CMR)	—
Mortazavi et al. [23]		Logistic regression (LR)	83.2%.
Marbaniang et al. [24]	UCI provided a dataset on heart illness with 14 different features.	Random forest (RF), logistic regression, and support vector machine (SVM)	—
Kathare and Gaikwad [13]	Heart study dataset was used.	Support vector machine (SVM)	88.7%
Segar et al. [3]	They gathered data from the website https://cran.r-project.org/web/packages/MASS/index.html.	Support vector machine (SVM) and *K*-nearest neighbors (KNN)	83.9%
Chicco and Jurman [6]	Data from 299 people with heart failure is analysed by the researchers.	Stochastic gradient classifier	—
Nashif et al. [25]	The dataset from the UCI machine learning repository was used.	Data mining modelling techniques	85.1% and 79.5%
Jindal et al. [26]	Using the UCI repository, a dataset with a patient's medical history and attributes is picked.	Logistic regression, KNN, and random forest classifier	—
Solanki and Sharma [27]	To get the data, they employed a variety of methods.	Artificial neural network (ANN)	56.76%

Table 2: Prediction of cardiovascular attacks.

Reference	Gender based	Dataset	Feature selection techniques	Classification models	Accuracy	Optimization
Khan et al. [37]	No	A synthetic dataset across minority data space	No technique	Support machine vector	83.65% for RBF and 84.56% for linear	No optimization technique
Chen et al. [14]	No	17 left ventricle defective patients	No technique	Machine learning basic models	79% on an average	No optimization technique
Mehta et al. [16]	No	Self-created data	No technique	Machine learning basic models	79% on an average	No optimization technique
Ahmad et al. [4]	No	Self-created data	No technique	Machine learning basic models	79% on an average	No optimization technique
Zahid et al. [8]	No	Self-created data	No technique	Support machine vector	72%	No optimization technique
Chicco and Jurman [6]	No	Self-created data	No technique	Support machine vector	76%	No optimization technique

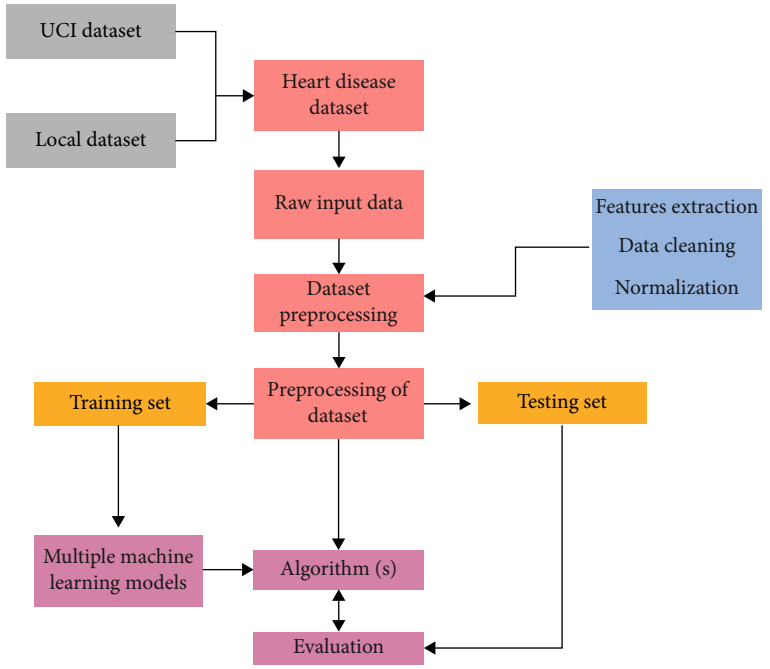


Figure 1: Block diagram.

Table 3: Data attributes.

Feature	Values	Description
Anaemia	0=no, 1=yes	Having blood deficiency
Diabetes	0=no, 1=yes	Having sugar problem
High_blood_pressure	0=no, 1=yes	Having BP problem
Sex	(1=male, 0=female)	Gender of patient
Smoking	0=no, 1=yes	Smoker or not
Time	Time of sample	Follow-up period
Death_event	0=no, 1=yes	Dead or alive

Table 4: Data attributes dataset (local).

Feature	Values	Description
Anaemia	0=no, 1=yes	Having blood deficiency
Diabetes	0=no, 1=yes	Having sugar problem
High_blood_pressure	0=no, 1=yes	Having BP problem
Sex	(1=male, 0=female)	Gender of patient
Smoking	0=no, 1=yes	Smoker or not
Time	Time of sample	Follow-up period
Death_event	0=no, 1=yes	Dead or alive

Machine learning seems to be able to accurately forecast survival times for heart failure patients. This study analyses data from a 2015 database of individuals with heart failure. According to patient charts, blood creatinine and ejection fraction levels are excellent predictors of patients' survival times. Survival rates in patients with heart failure could be predicted with a new method [36].

This article [20] discusses the use of machine learning for cardiac disease prognosis. This study used data analytics to examine heart disease. The authors conducted a study to determine the accuracy and dependability of three distinct data analytic methods (neural networks, SVM, and KNN). The results from neural networks are superior and can be constructed more quickly (accuracy of 93 percent).

In [18], the authors demonstrate how machine learning can be applied to the problem of predicting cardiac disease. According to a novel algorithm developed by Stanford University researchers, a patient's medical history can be used to forecast their risk of developing heart disease. Methods from the field of machine learning, such as logistic regression and

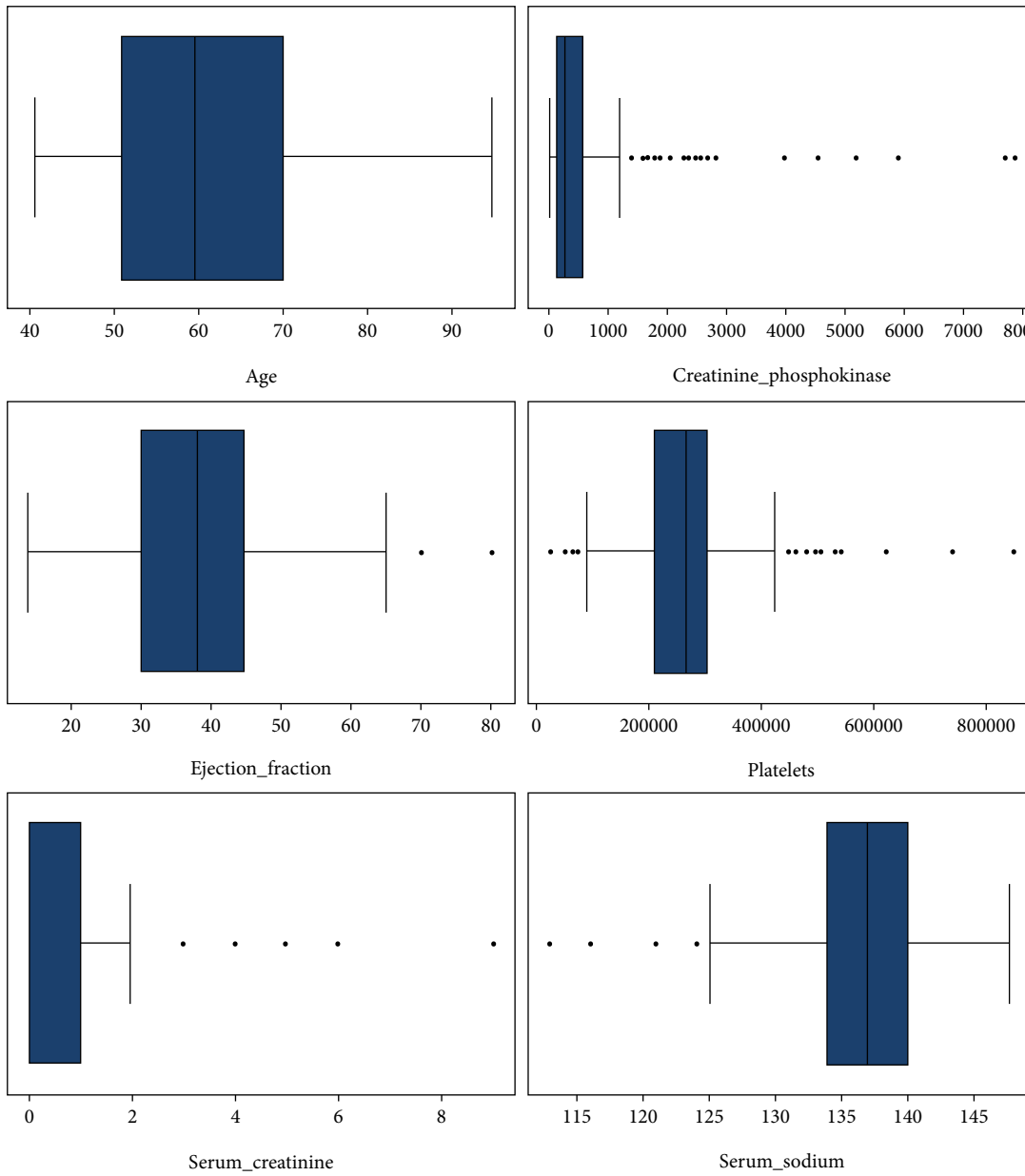
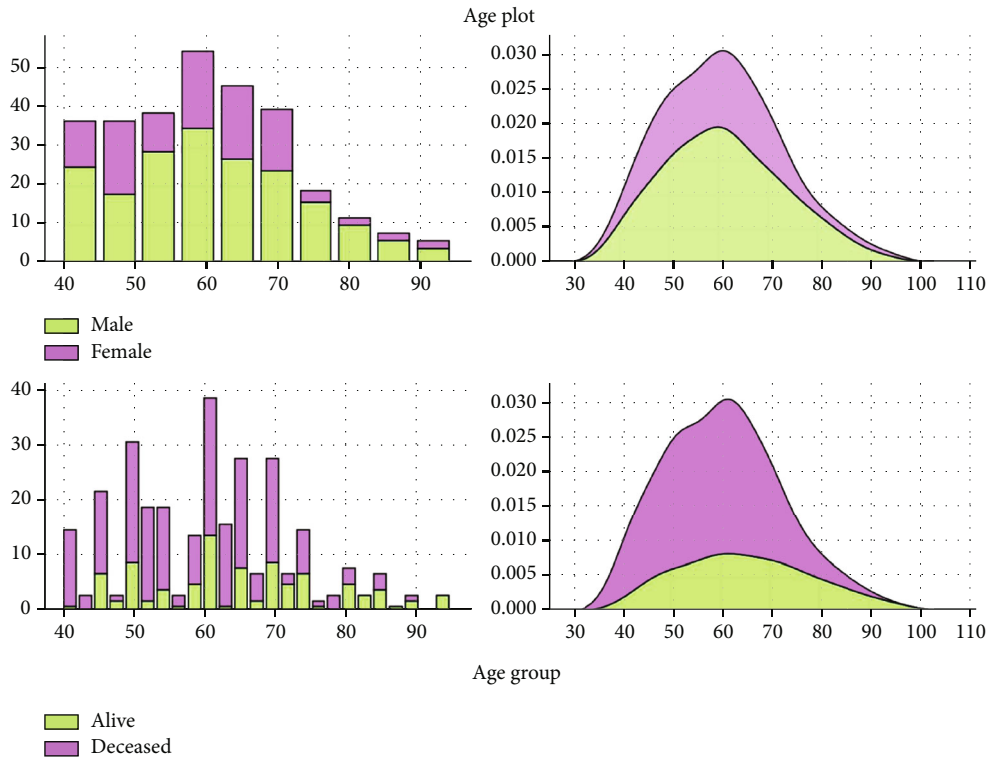


Figure 2: Boxplot of each attribute.

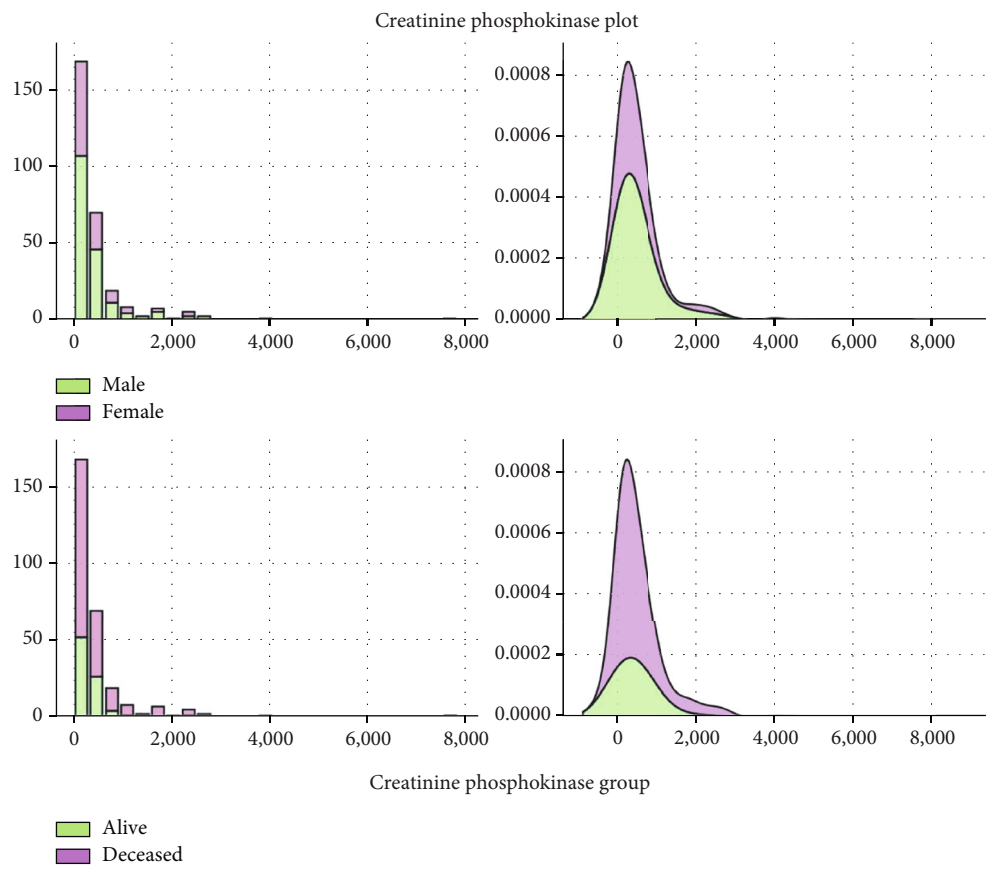
KNN, were used to categorise and predict outcomes for patients with heart disease. By taking these steps, a more reliable model was developed for predicting cardiac events in a broad population. KNN and logistic regression were superior at predicting the presence or absence of heart disease in a single person when compared to other classifiers like naïve Bayes. Saving time and effort in testing whether a classifier can correctly identify cardiac disease is a major benefit of this technology. Use of the offered heart disease prediction method will help you save both time and money. The ipynb file format used for this study's data makes it possible to make inferences about the future health of persons with heart disease. This study evaluates the efficacy of a narrative approach to cardiovascular disease prediction, wherein machine learning is used to find crucial components. Assorted features and classification methods are used to construct a prediction model. The authors claim that a 92% accuracy rate can be attained by combining a random forest with a linear model. In order to better identify heart attacks and forecast cardiovascular illness, the following Table 2 analyses the results of past research utilising demographic and medical variables.

3. Materials and Methods

We explain how we conducted our research here. Feature engineering, model creation, and performance evaluation are all engaged in the gathering, describing, and analysing of datasets. Figure 1 is a flowchart that shows the overall development of the study.

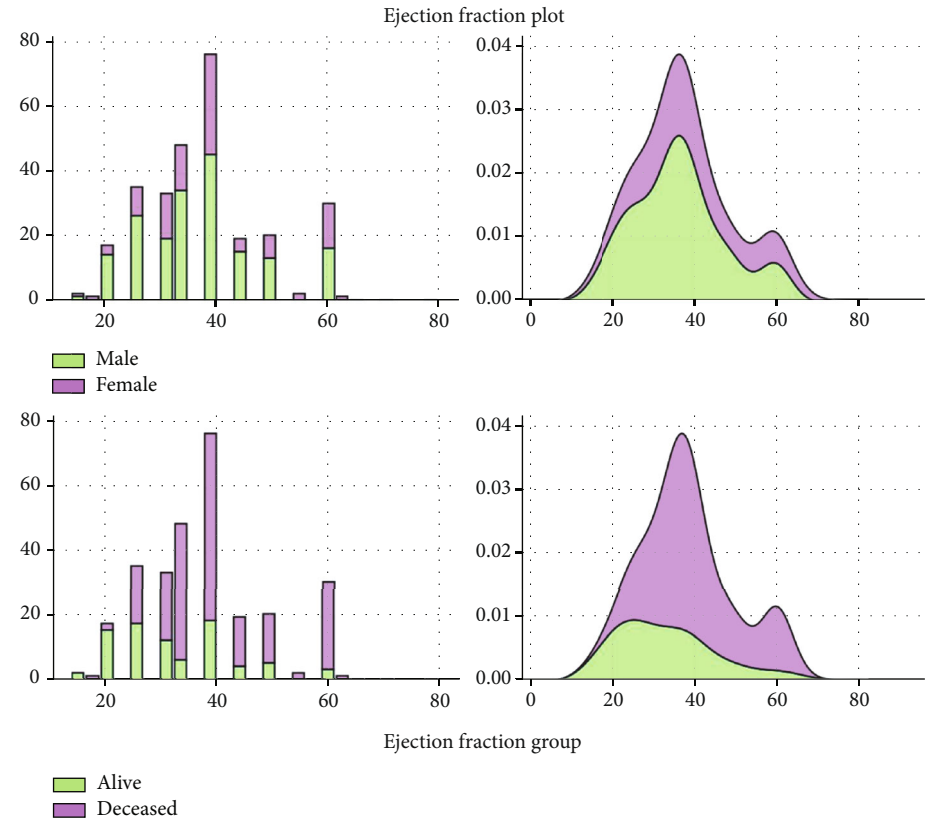


(a) Age plot

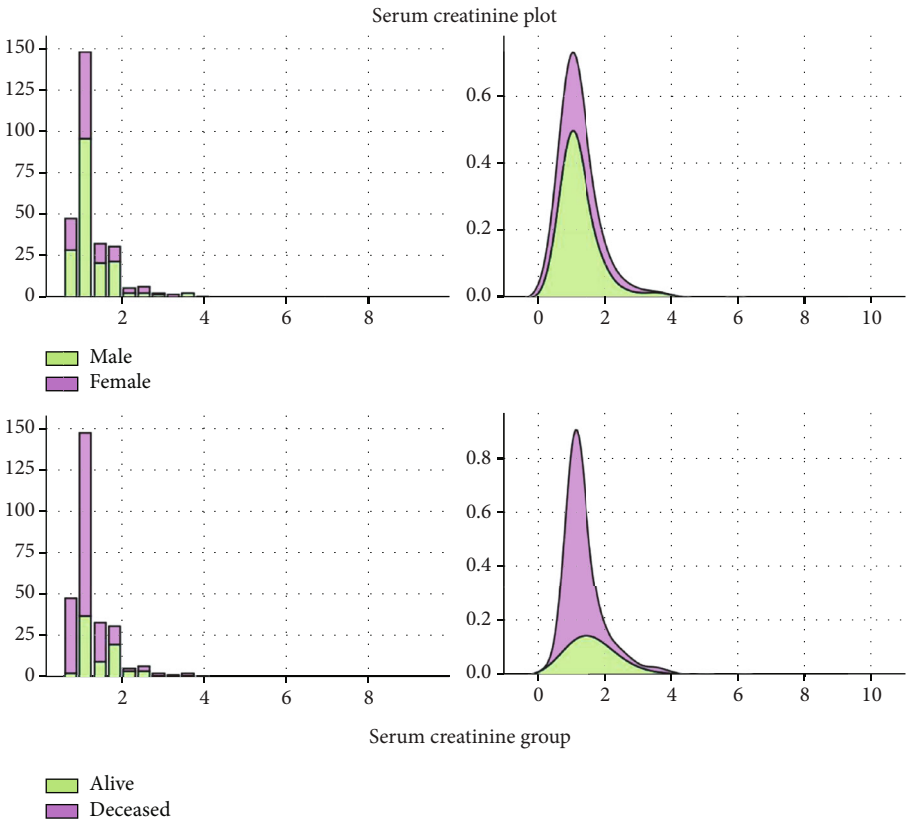


(b) Creatinine phosphokinase

Figure 3: Continued.

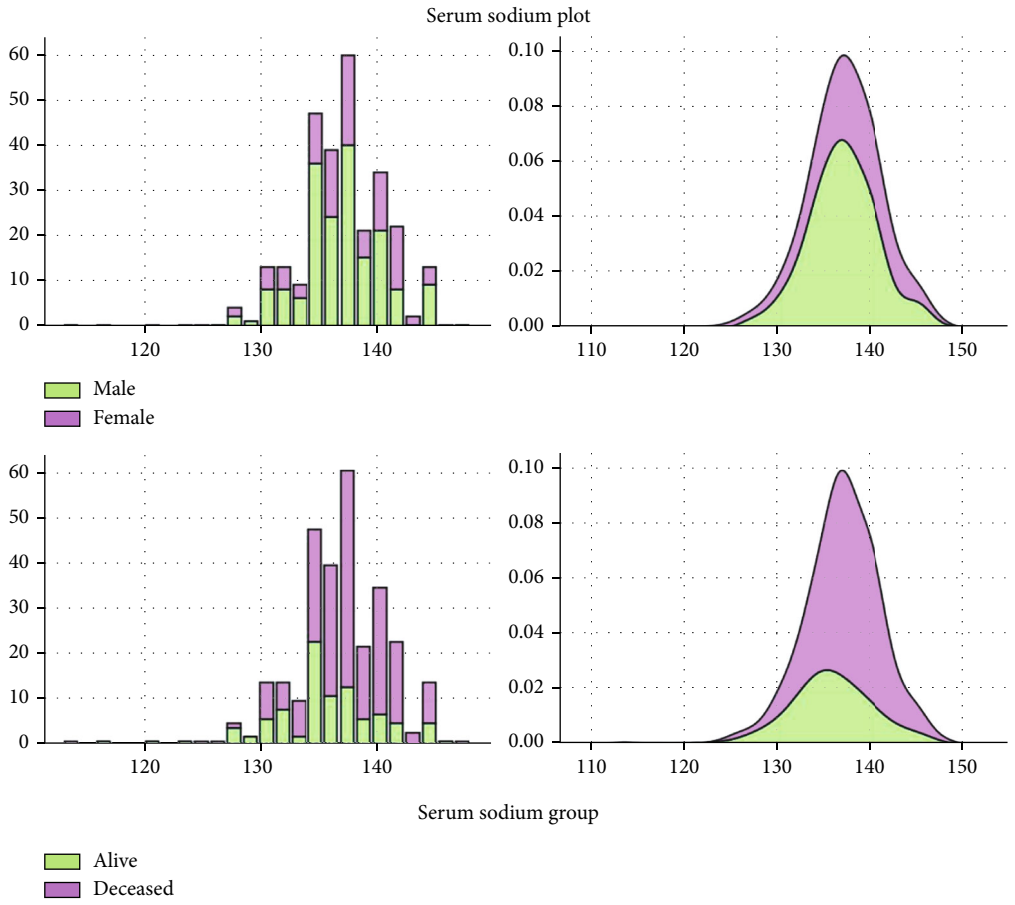


(c) Ejection fraction group



(d) Serum creatinine group

Figure 3: Continued.



(e) Serum sodium group

Figure 3: Data about death from heart failure.

Information regarding the block diagram can be found as follows:

(i) A comma-separated values (.csv) file was generated using information from the heart failure dataset

(ii) Outliers have been removed, and data has been normalised during preprocessing. Cross-validation has been used, and the results have been confirmed. Models based on machine learning have been used

(iii) The most effective classification models have been chosen, and then, Ensemble Learning strategies have been put into place

The prediction dataset is freely available on Kaggle. The American Heart Failure Institute has collected this information. There is one dependent variable and a vast number of unrelated independent variables that make up the dataset.

About 300 individuals with left ventricular systolic dysfunction are represented in the UCI dataset. Specific patient characteristics are described in the 12 columns that follow heart failure. The average duration of a follow-up is 130 days. Find a list of all accessible datasets down below in Table 3. Among the items in the dataset are

It is estimated that roughly 500 patients with left ventricular systolic dysfunction are represented in the local dataset. Specific patient characteristics are described in the 12 columns that follow heart failure. The average duration of follow-up is 130 days. Table 4 shows the data attributes dataset (local).

The boxplots for each of the following attributes can be seen in the following images. Figure 2 shows the boxplot of each attribute.

Data about death from heart failure is shown in the following Figure 3.

Death from heart failure is based on demographic characteristics in Figure 4. The distribution of patients who are dead or alive based on their medical status is depicted in the graph below. Figure 5 shows the outcome as death or alive, while Figure 6 shows the data distribution of each attribute.

Figure 7 shows the frequency of anaemia and Figure 8 shows the frequency distribution of diabetes attribute, while Figure 9 shows the frequency distribution of blood pressure. Figure 10 shows the gender distribution (demographic features). Figure 11 shows the smoking attribute (demographic feature) frequency.

Figure 12 below shows the histograms illustrating the data distribution.

In order to test and evaluate the effectiveness of the suggested method, the HF dataset must be used. Diseases from

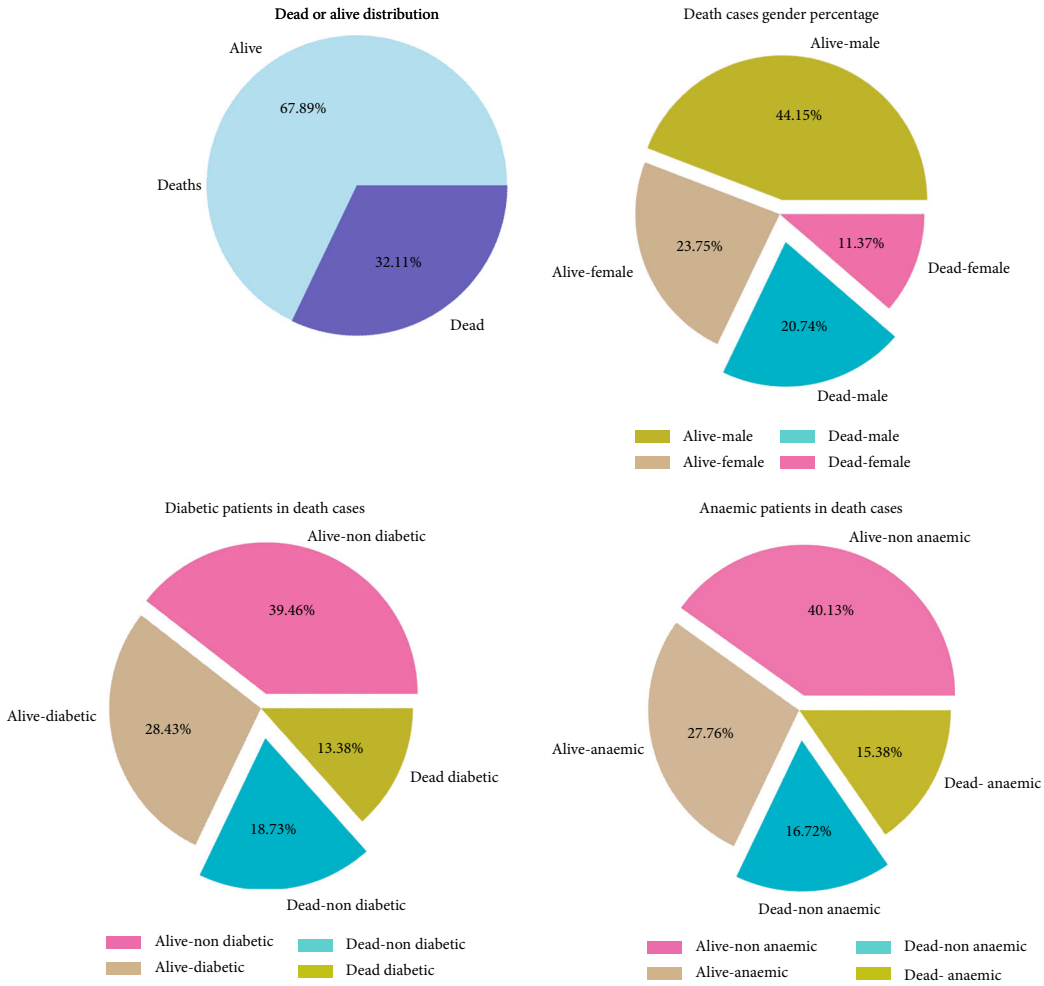


Figure 4: Dead and alive distribution of patients according to their medical deficiencies.

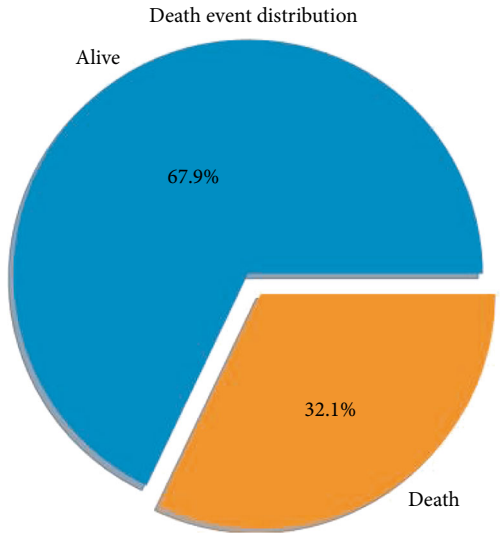


Figure 5: Outcome as death or alive.

many different real-world categories are represented in the HF dataset. We employ the comma-separated values (CSV) file format to do preliminary processing and feature extraction on raw data [35, 36, 38–41].

3.1. Data Cleaning. Through Kaggle, we were able to acquire access to raw data. Multiple methods were employed to get rid of duplicates, null values, and other useless data. Wearables such as electrocardiogram (ECG) monitors, pulse oximeters, thermometers, and blood pressure monitors were utilised to gather this clinical information. When attached to a person, these sensors gathered electrocardiogram (ECG) information, as well as blood pressure and temperature readings. Using the Internet of Things, we were able to collect and store the data in the cloud.

3.2. Data Preprocessing. Data mining relies on this method for transforming unstructured data into a more digestible form. Information gleaned from the real world is sometimes missing, incorrect, or otherwise unusable. Some methods of preprocessing include the ones listed above. Imprecise classification is a barrier to developing accurate predictive models. With the majority of machine learning techniques used for

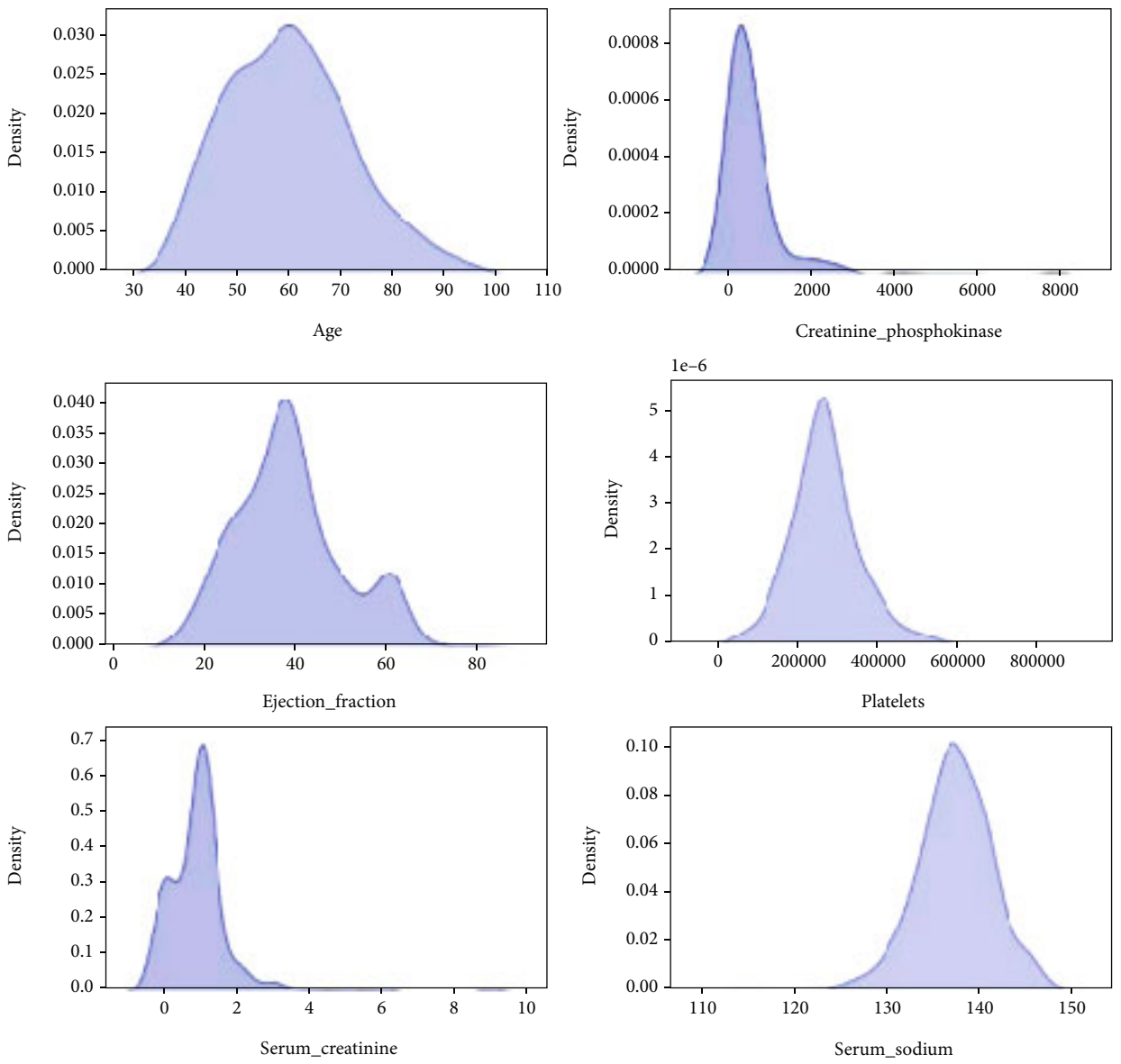


Figure 6: Data distribution of each attribute.

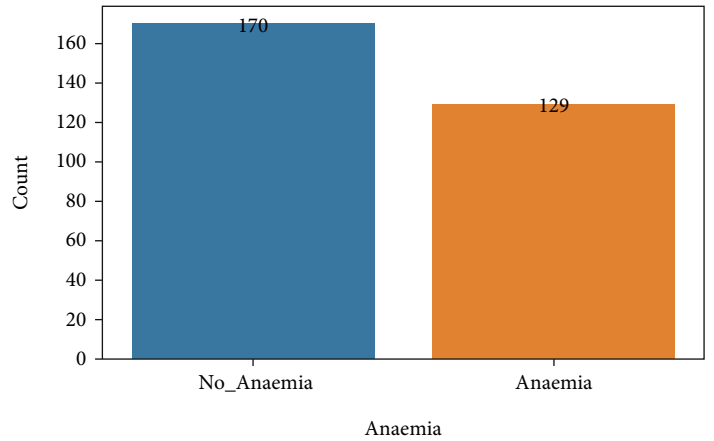


Figure 7: Frequency of anaemia.

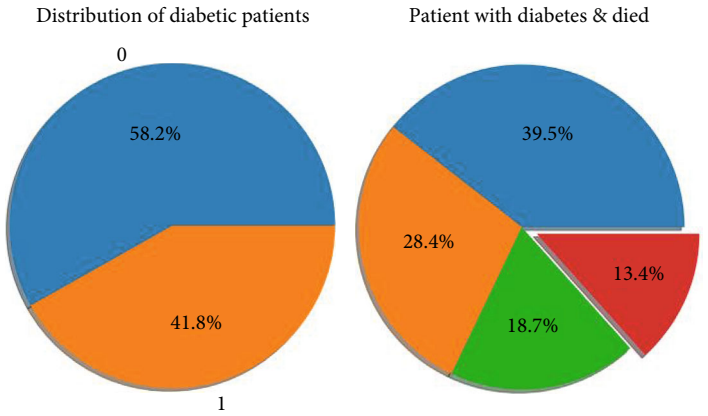


Figure 8: Frequency distribution of diabetes attribute.

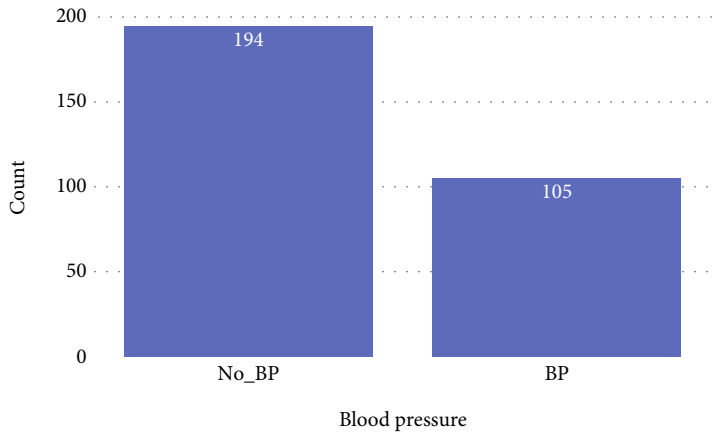


Figure 9: Frequency distribution of blood pressure.

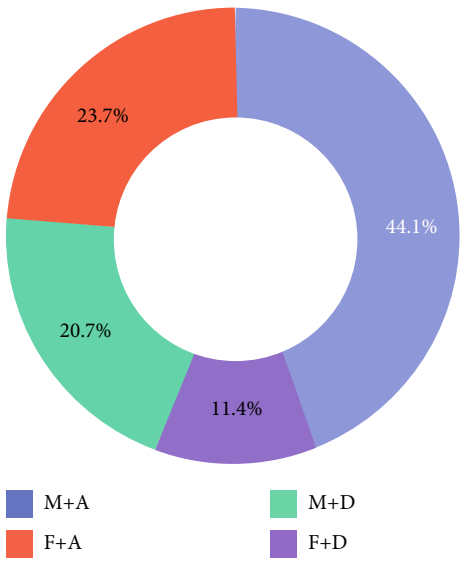


Figure 10: Gender distribution (demographic features).

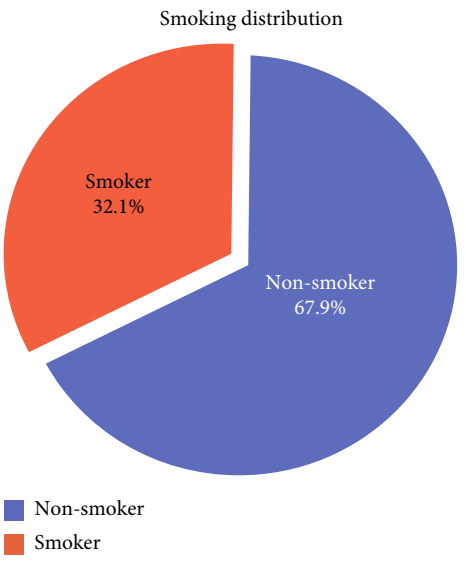


Figure 11: Smoking attribute (demographic feature) frequency.

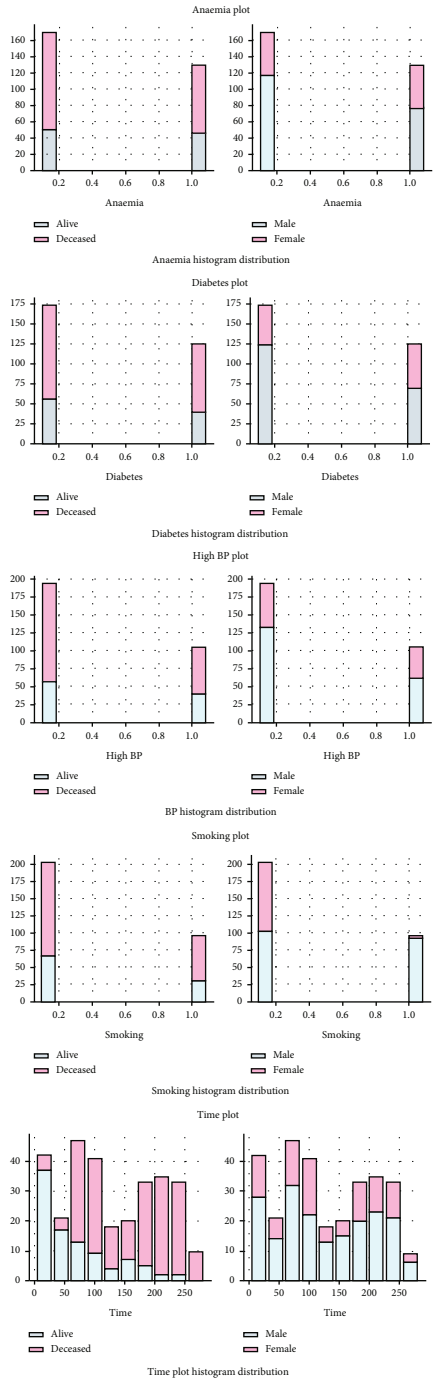


Figure 12: Data distribution visualization in histograms.

categorization, the number of examples in each class is around the same. Because of this, many people of colour end up being portrayed incorrectly. It is concerning since this tendency is exacerbated by the fact that minorities, being numerically underrepresented, are disproportionately affected by mistakes in data analysis and statistical classification. This allowed us to remove the anomalous data from our sample and normalise the overall dataset. As a direct outcome of this study, numerous improved resampling techniques have been presented. We can accomplish undersampling by deleting entries from each cluster, for example, while yet preserving information by collecting the records from the majority class. Oversampling allows us to make slight alterations to our copies of data from underrepresented groups, producing more representative mock samples than would be possible with a 1 : 1 replication of the original data.

3.3. Feature Engineering. Information from a certain domain is used to create functions that can be used by learning algorithms. Extracting and processing raw data is the first step in building a machine learning representation. It is utilised to find out how closely related things are in this investigation.

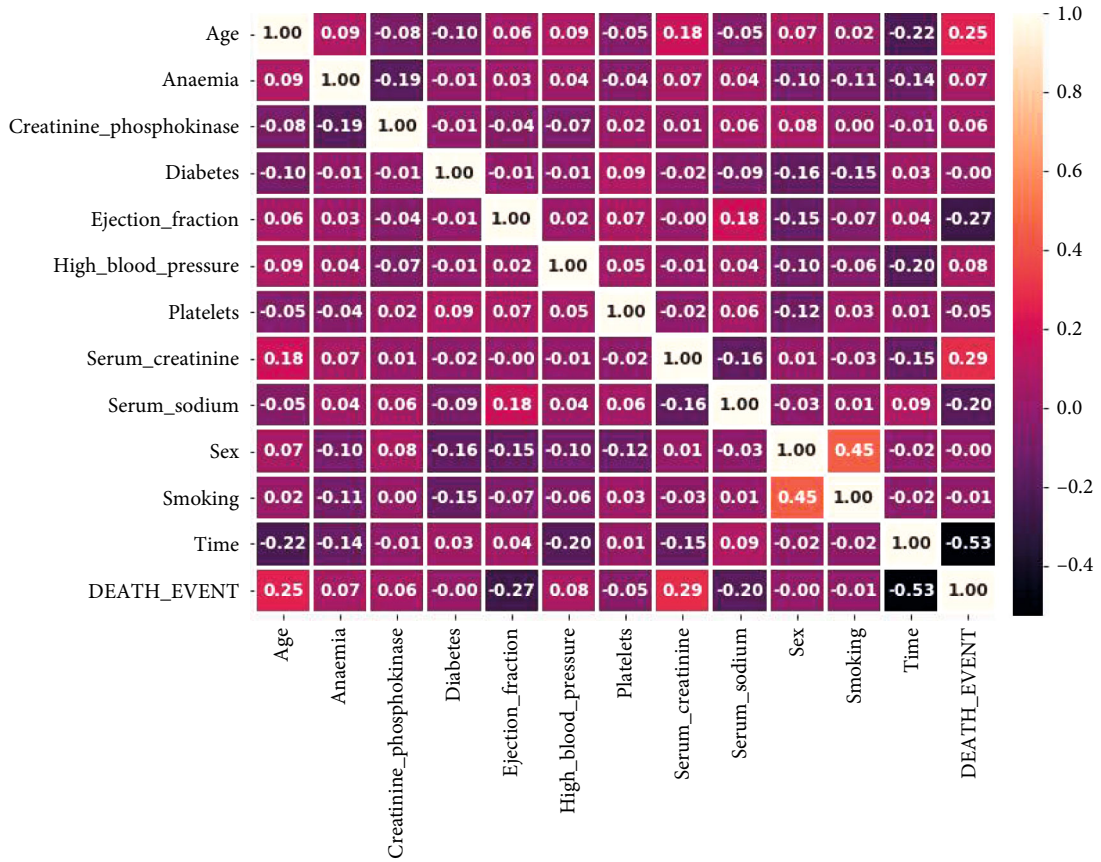
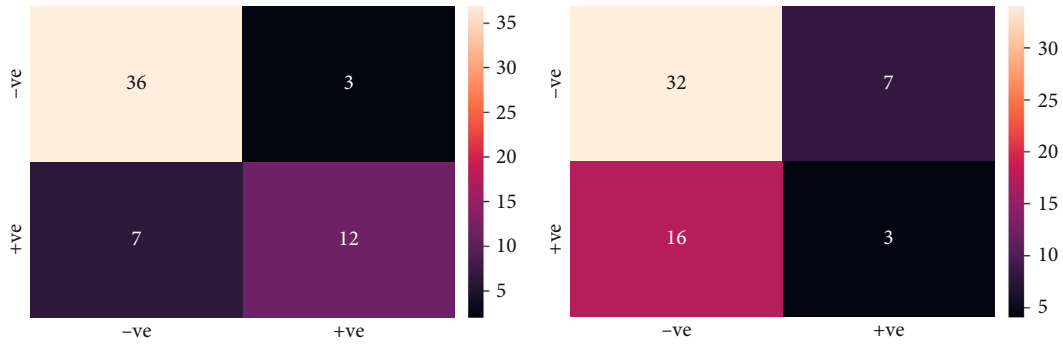


Figure 13: Correlation matrices HF dataset.

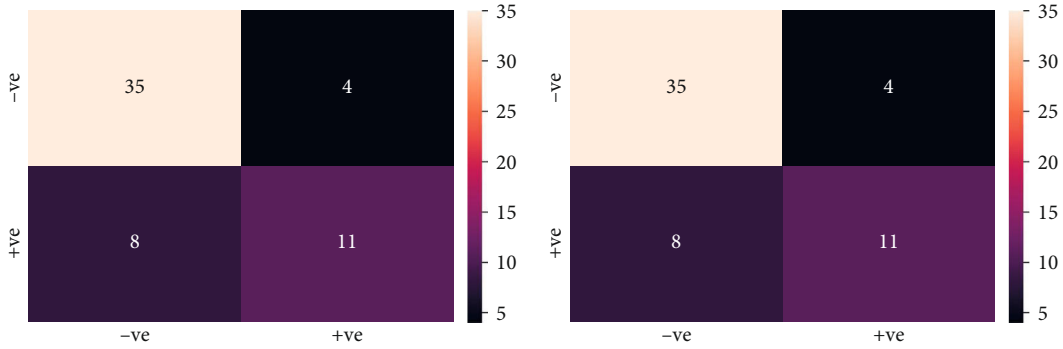
Table 5: Description of Metrics.

Metric	Description
Accuracy	$\text{Accuracy} = \frac{\text{TP}}{(\text{TP} + \text{TN}) * 100}$
Precision	$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP}) * 100}$
Recall	$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN}) * 100}$
F1 score	$\text{F1 score} = \frac{\text{precision.recall}}{\text{precision} + \text{recall} * 100}$
Confusion matrix	Real label: Positive / Negative; Predicted label: Positive / Negative; True Positive (TP); False Positive (FP); False Negative (FN); True Negative (TN); $\text{Precision} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FP}}$; $\text{Recall} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FP}}$; $\text{Accuracy} = \frac{\sum \text{TP} + \text{TN}}{\sum \text{TP} + \text{FP} + \text{FN} + \text{TN}}$



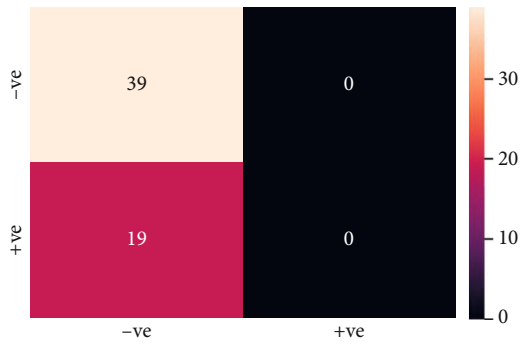
(a) Logistic regression

(b) K-nearest neighbors



(c) Gaussian naïve Bayes

(d) Multinomial naïve Bayes



(e) Support vector machine

Figure 14: Confusion matrix of each linear model.

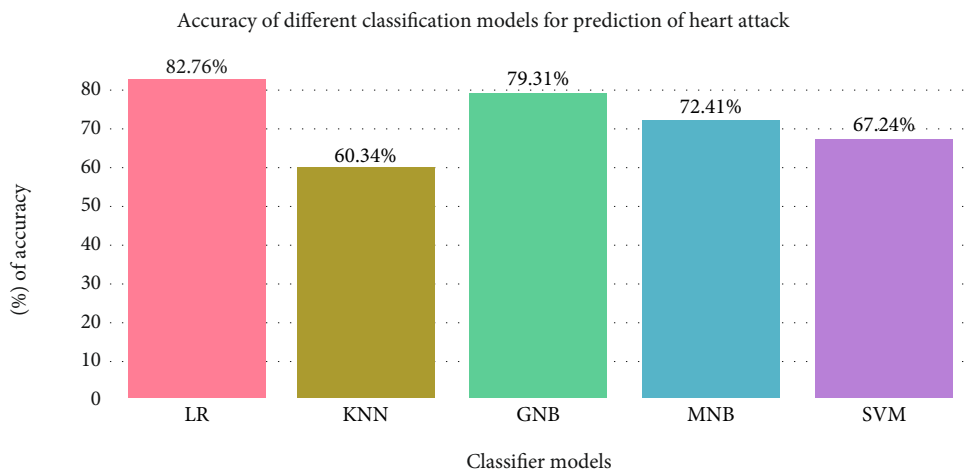


Figure 15: Linear machine learning classification model performance.

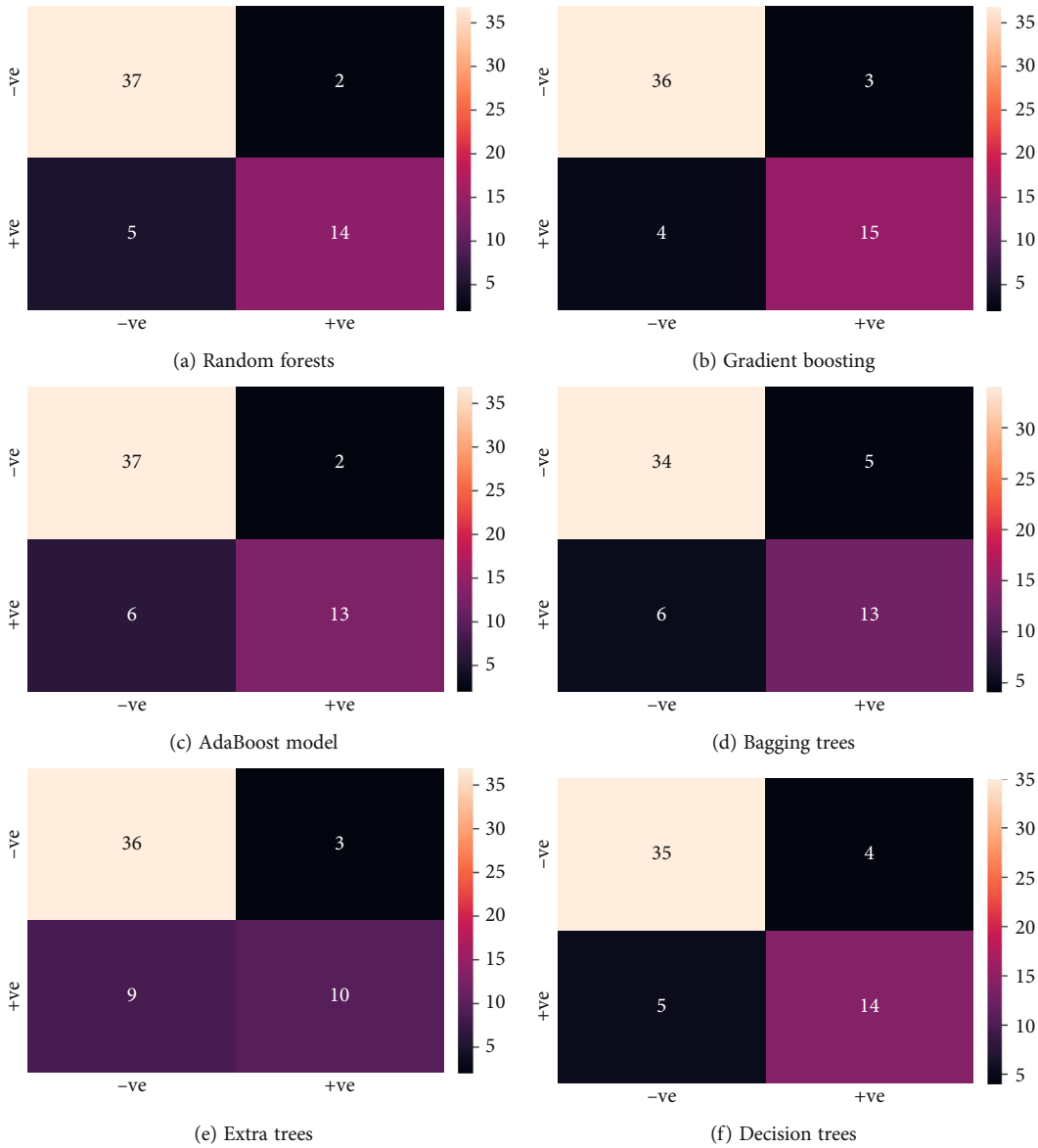


Figure 16: Confusion matrix of each ensemble learning model.

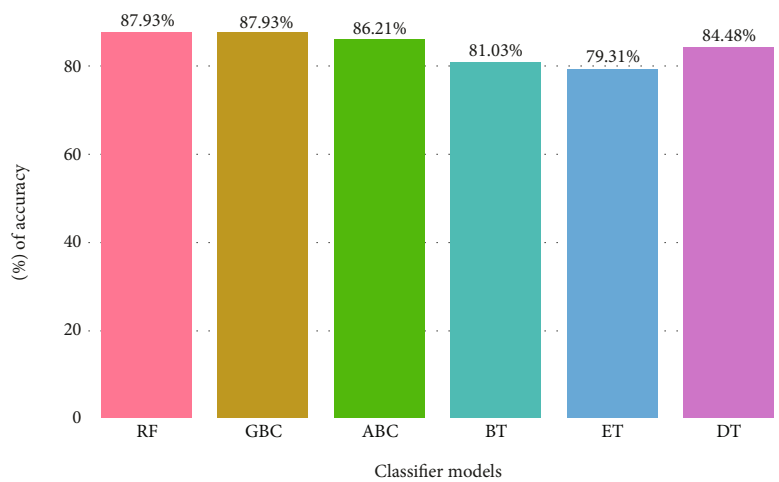
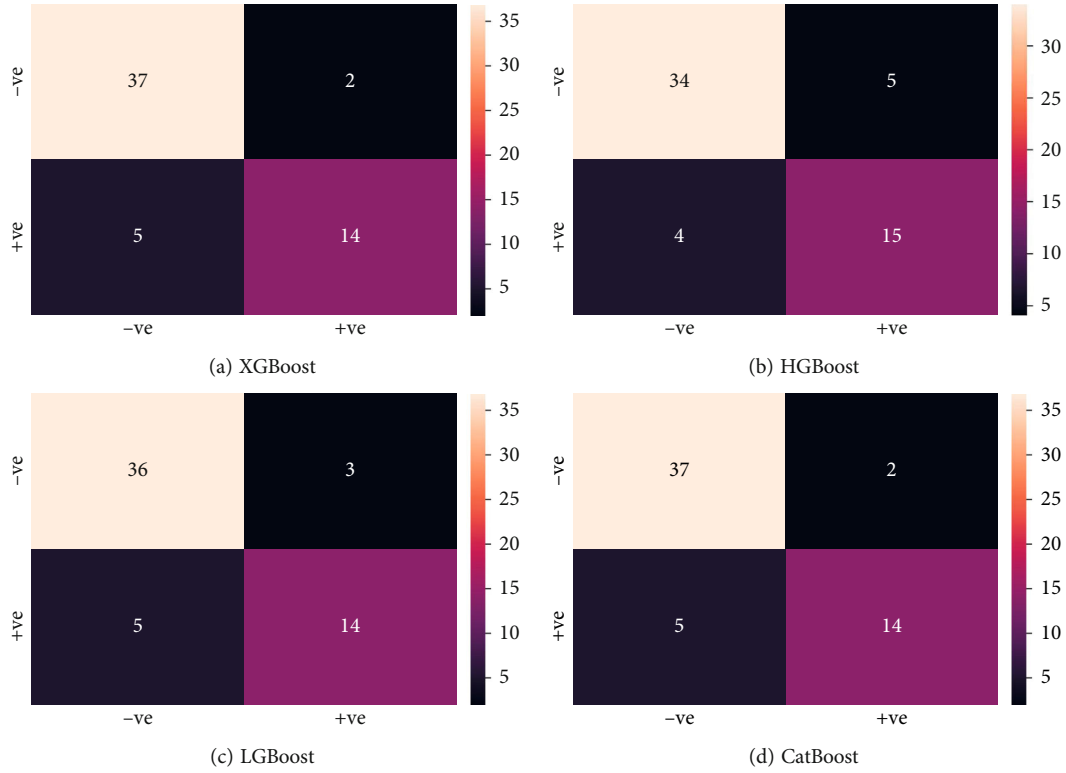


Figure 17: Ensemble machine learning classification model performance.



(a) XGBoost (b) HGBoost

(c) LGBoost (d) CatBoost

Figure 18: Confusion matrix of each boost learning model.

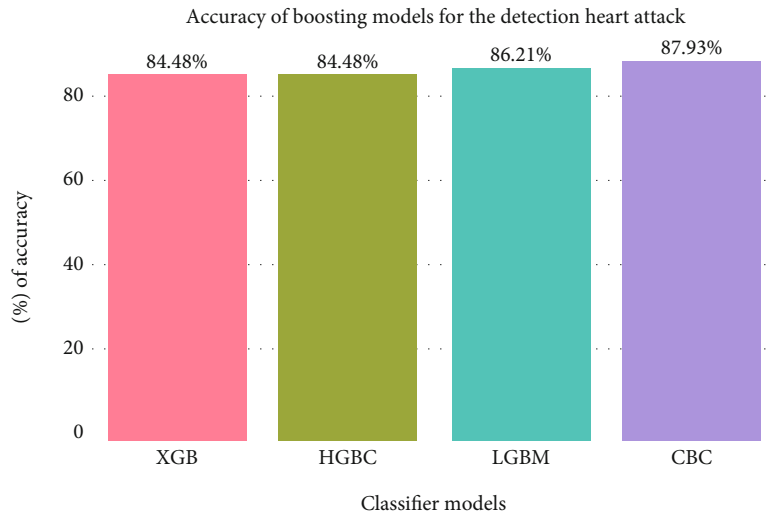


Figure 19: Boosting machine learning classification model performance.

Or, to put it another way, a correlation matrix is really a fancy name for a covariance matrix. The strength of a linear relationship can be summarised by the correlation, which provides a numerical value. To "correlate" two numbers is to draw a straight line between them. Only the numbers 1 and 1+1 are valid. As can be seen in the graph below, skin thickness, insulin, pregnancies, and age are all completely unrelated to one another. Figure 13 shows correlation matrices HF dataset.

The purpose of cross-validation in machine learning is to reevaluate models with a reduced dataset. With just K, you can categorise a certain dataset with pinpoint accuracy. This technique goes by a few different names, including K-fold cross-validation. K-folds should be chosen at random between 5 and 10 times (depending on the amount of data). You may make your model correct by running it through the folds $K - 1$(Kminus 1). Multiple algorithms, including linear machine learning models, boosted machine learning models, and ensemble learning models, are being used to make these determinations. An F1 score was created to evaluate the efficacy of different methods. According to the confusion

matrix, both the classified and misclassified clauses are incorrect. The following metrics were used in this analysis. Table 5 shows the description of metrics.

4. Results and Discussion

4.1. Linear Machine Learning Models. Logistics regression beats SVM, MNB, GNB, and KNN models when attempting to ascertain whether or not a certain individual is still alive. Below is a figure depicting the confusion matrix calculated from the true positive and true negative values of SVM, LR, MNB, GNB, and KNN.

Figure 14 shows the confusion matrix of each linear model. Logistic regression, compared to other linear models for predicting heart failure from a dataset, has the highest number of true positive classified values, as seen in the preceding graphic. Figure 15 shows the linear machine learning classification model performance.

This graph compares the model accuracy. Logistic regression (82.76%) produced the greatest results, followed by SVM (67.24%), KNN (60.34%), GNB (79.51%), and MNB (79.51%).

4.2. Ensemble Learning Models. Random forests and gradient boosting classifiers are the best ensemble learning models for forecasting patient death. This figure illustrates the true positive, false positive, and false negative values of the DT, ET, BT, GBC, ABC, and random forests:

Figure 16 shows the confusion matrix of each ensemble learning model. Since they have the truest positive categorization values, RF and GBC are the best ensemble learning models for predicting heart failure.

Figure 17 shows the ensemble machine learning classification model performance graph showing model accuracy. ABC (86.21%), BT (81.03%), ET (79.31%), and DT (79.31%).

4.3. Boosting Classifiers. CatBoost classifier is the most accurate booster. The following graph shows true positive and false negative values for CatBoost, XBG, LGBM, and Hist GBC.

Figure 18 shows the confusion matrix of each boost learning model. CatBoost has the truest positive categorised values in the dataset, as seen in the graphic above.

Figure 19 shows the boosting machine learning classification model performance. The reliability of several models is shown here. The highest levels of accuracy are achieved by CatBoost (87.93%), LGBM (86.21%), HGBC (84.48%), and XGB (84.48%).

SVM (67.24%), KNN (60.34%), GNB (79.31%), and MNB (72.41%) are the most accurate linear models for logistic regression (82.76 percent). After them, come the 86.21 percent, 81.33 percent, 70.33 percent, and 76.31 percent of the DT and ET, respectively, 84% to be exact. The highest accuracy is achieved by CatBoost (86.21%), followed by HGBC (86.21%) and LGBM (86.21%) (87-93 percent). CatBoost, random forests, and gradient boosting can all be used to forecast cardiac arrests.

This study's proposed machine learning method could be used to improve the prognosis of heart failure (HF) and other diseases by analysing real-time patient data. Based on logistic regression's 82.76 percent accuracy, SVM's 67.24 percent, KNN's 60.34 percent, GNB's 70.31 percent, and MNB's 72.41 percent accuracy, RF and GBC (87.93 percent), ABC (86.11 percent), BBT (81.03 percent), ET (79 percent), and DT (84.58 percent) are the most accurate models in an ensemble learning model. Compared to LGBM (86.21% accurate) and LGBM (HGBC and XGB) (86.21% accurate), CatBoost achieves higher accuracy (84.48 percent). CatBoost, random forests, and gradient boosting, in combination with other predictive algorithms, are optimal for predicting the occurrence of heart attacks.

5. Conclusions

Heart failure (HF) is a common condition that can be fatal in the modern era. Every year, somewhere around 26 million people are infected globally. In cardiology and surgery, it is challenging to predict when a patient may develop heart failure. Classification and prediction models are useful to the medical business because they demonstrate potential applications for medical data. The accuracy of HF projections will be improved with the use of data on cardiovascular disease, by predicting heart failure occurrences in a medical database using machine learning techniques. According on the current results and comparison analyses, it is now possible to more accurately predict heart disease. In this study, we present a machine learning approach that can be used to improve disease prediction, not just for HF but for any condition. Each of the four most accurate linear models—SVM, KNN, GNB, and MNB—has an accuracy rating of 67.24 percent or higher. The accuracy of ensemble learning models such as GBC and ABC (87.93%) and RF and GBC is drastically different from one another (87.3 percent). With an accuracy of 87.93%, CatBoost outperforms LGBM (86.21%), HGBC (84.48%), and XGB (83.78%), 84% to be exact. Methods like CatBoost, random forests, and gradient boosting can accurately foresee almost eight out of ten cardiac arrests. In future studies, this research can be upgraded to predict the survival of patients by using HF dataset.

Data Availability

All the data is available in the manuscript.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] A. F. Grubb, C. A. Pumill, S. J. Greene, A. Wu, K. Chiswell, and R. J. Mentz, "Tobacco smoking in patients with heart failure and coronary artery disease: a 20-year experience at Duke University Medical Center," *American Heart Journal*, vol. 230, pp. 25–34, 2020.

[2] 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.

[3] M. W. Segar, M. Vaduganathan, K. V. Patel et al., "Machine learning to predict the risk of incident heart failure hospitalization among patients with diabetes: the WATCH-DM risk score," *Diabetes Care*, vol. 42, no. 12, pp. 2298–2306, 2019.

[4] 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, article e0181001, 2017.

[5] B. K. Turkmenoglu and O. Yildiz, "Predicting the survival of heart failure patients in unbalanced data sets," in *2021 29th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, Istanbul, Turkey, 2021.

[6] 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, 2020.

[7] P. A. Moreno-Sanchez, "Improvement of a prediction model for heart failure survival through explainable artificial intelligence," 2021, https://arxiv.org/abs/2108.10717.

[8] F. M. Zahid, S. Ramzan, S. Faisal, and I. Hussain, "Gender based survival prediction models for heart failure patients: a case study in Pakistan," *PLoS One*, vol. 14, no. 2, 2019.

[9] E. J. Benjamin, P. Muntner, A. Alonso et al., "Heart disease and stroke statistics-2019 update: a report from the American Heart Association," *Circulation*, vol. 139, no. 10, pp. e56–e528, 2019.

[10] A. Jayakrishnan, R. Visakh, and K. T. Ratheesh, "Computational approach for heart disease prediction using machine learning," in *2021 International Conference on Communication, Control and Information Sciences (ICCISc)*, pp. 1–5, Idukki, India, 2021.

[11] S. B. Wang, P. Mitchell, G. Liew et al., "A spectrum of retinal vasculature measures and coronary artery disease," *Atherosclerosis*, vol. 268, pp. 215–224, 2018.

[12] T. Y. Wong, R. Klein, B. E. K. Klein, J. M. Tielsch, L. Hubbard, and F. J. Nieto, "Retinal microvascular abnormalities and their relationship with hypertension, cardiovascular disease, and mortality," *Survey of Ophthalmology*, vol. 46, no. 1, pp. 59–80, 2001.

[13] S. Kathare and S. Gaikwad, "Practicability of heart attack prediction using machine learning," *International Journal of Research Publication and Reviews*, vol. 2, no. 7, pp. 1473–1477, 2021.

[14] X. J. Chen, E. T. LaPorte, C. Olsen et al., "Heart sound analysis in individuals supported with left ventricular assist devices," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 10, pp. 3009–3018, 2021.

[15] G. J. Chowdary, "Prediction of cardiovascular disease using machine learning algorithms," *International Journal of Engineering and Advanced Technology*, vol. 9, no. 3, pp. 2404–2414, 2020.

[16] D. Mehta, A. Naik, R. Kaul, P. Mehta, and P. J. Bide, "Death by heart failure prediction using ML algorithms," in *2021 4th Biennial International Conference on Nascent Technologies in Engineering (ICNTE)*, pp. 1–5, NaviMumbai, India, 2021.

[17] M. U. Ghani, T. M. Alam, and F. H. Jaskani, "Comparison of classification models for early prediction of breast cancer," in *2019 International Conference on Innovative Computing (ICIC)*, pp. 1–6, Lahore, Pakistan, 2019.

[18] J. Vijayashree and N. C. S. N. Iyengar, "Heart disease prediction system using data mining and hybrid intelligent techniques: a review," *International Journal of Bio-Science and Bio-Technology*, vol. 8, no. 4, pp. 139–148, 2016.

[19] M. Marimuthu, M. Abinaya, K. S. Hariesh, K. Madhankumar, and V. Pavithra, "A review on heart disease prediction using machine learning and data analytics approach," *International Journal of Computers and Applications*, vol. 181, no. 18, pp. 20–25, 2018.

[20] A. Golande and T. Pavan Kumar, "Heart disease prediction using effective machine learning techniques," *International Journal of Recent Technology and Engineering*, vol. 8, no. 1, pp. 944–950, 2019.

[21] B. Ayers, T. Sandholm, I. Gosev, S. Prasad, and A. Kilic, "Using machine learning to improve survival prediction after heart transplantation," *Journal of Cardiac Surgery*, vol. 36, no. 11, pp. 4113–4120, 2021.

[22] I. U. Haq, I. Haq, and B. Xu, "Artificial intelligence in personalized cardiovascular medicine and cardiovascular imaging," *Cardiovascular Diagnosis and Therapy*, vol. 11, no. 3, pp. 911–923, 2021.

[23] B. J. Mortazavi, N. S. Downing, E. M. Bucholz et al., "Analysis of machine learning techniques for heart failure readmissions," *Circulation. Cardiovascular Quality and Outcomes*, vol. 9, no. 6, pp. 629–640, 2016.

[24] I. A. Marbaniang, N. A. Choudhury, and S. Moulik, "Cardiovascular disease (CVD) prediction using machine learning algorithms," in *2020 IEEE 17th India Council International Conference (INDICON)*, pp. 491–495, New Delhi, India, 2020.

[25] S. Nashif, M. R. Raihan, M. R. Islam, and M. H. Imam, "Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system," *World Journal of Engineering and Technology*, vol. 6, no. 4, pp. 854–873, 2018.

[26] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," *IOP Conference Series: Materials Science and Engineering, Materials Science and Engineering*, vol. 1022, no. 1, 2021.

[27] Y. Solanki and S. Sharma, "A survey on risk assessments of heart attack using data mining approaches," *International Journal of Information Engineering and Electronic Business*, vol. 11, no. 4, pp. 43–51, 2019.

[28] G. Kaur and A. Chhabra, "Improved J48 classification algorithm for the prediction of diabetes," *International Journal of Computers and Applications*, vol. 98, no. 22, pp. 13–17, 2014.

[29] A. J. Santos, X. E. Asuncion, C. Rivero-Co et al., "Modeling differential rates of aging using routine laboratory data; implications for morbidity and health care expenditure," 2021, https://arxiv.org/abs/2103.09574.

[30] A. Baccouche, B. Garcia-Zapirain, C. C. Olea, and A. Elmaghraby, "Ensemble deep learning models for heart disease classification: a case study from Mexico," *Information*, vol. 11, no. 4, 2020.

[31] G. Manogaran, R. Varatharajan, and M. K. Priyan, "Retracted article: hybrid recommendation system for heart disease diagnosis based on multiple kernel learning with adaptive neuro-fuzzy inference system," *Multimedia Tools and Applications*, vol. 77, no. 4, pp. 4379–4399, 2018.

[32] X. Liu, X. Wang, Q. Su et al., "A hybrid classification system for heart disease diagnosis based on the RFRS method," *Computational and Mathematical Methods in Medicine*, vol. 2017, Article ID 8272091, 11 pages, 2017.

[33] N. P. Selvaraj, S. Paulraj, P. Ramadass et al., "Exposure of botnets in cloud environment by expending trust model with CANFES classification approach," *Electronics*, vol. 11, no. 15, p. 2350, 2022.

[34] V. Kumar, G. S. Lalotra, P. Sasikala et al., "Addressing binary classification over class imbalanced clinical datasets using computationally intelligent techniques," *Healthcare*, vol. 10, no. 7, 2022.

[35] P. Kashyap and P. Kashyap, "Industrial applications of machine learning," in *Machine Learning for Decision Makers*, pp. 189–233, Apress, 2017.

[36] F. Cabitza, A. Locoro, and G. Banfi, "Machine learning in orthopedics: a literature review," *Frontiers in Bioengineering and Biotechnology*, vol. 6, 2018.

[37] M. F. Khan, R. K. Gazara, M. M. Nofal et al., "Reinforcing synthetic data for meticulous survival prediction of patients suffering from left ventricular systolic dysfunction," *IEEE Access*, vol. 9, pp. 72661–72669, 2021.

[38] C. Kruse, P. Eiken, P. Vestergaard, C. Kruse, P. Eiken, and P. Vestergaard, "Machine learning principles can improve hip fracture prediction," *Calcified Tissue International*, vol. 100, no. 4, pp. 348–360, 2017.

[39] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, pp. 242–264, IGI Global, 2010.

[40] P. Andreasson, J. Johansson, S. Liljestrand, and M. Granath, "Quantum error correction for the toric code using deep reinforcement learning," *Quantum*, vol. 3, p. 183, 2019.

[41] F. Ma, T. Sun, L. Liu, and H. Jing, "Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network," *Future Generation Computer Systems*, vol. 111, pp. 17–26, 2020.