PARAMETER
JURNAL MATEMATIKA, STATISTIKA DAN TERAPANNYA

e-ISSN. 2829-6303
Vol. 05 No 1, April 2026
Pages 87-110

Clinical Factor Analysis and Comparison of Heart Failure Patient Prediction Models Using Logistic Regression and XGBoost

Novri Suhermi[1]*, Rahida Rihhadatul Aisy[2], Auriga Wiradhiani[3], Edvina Kresnaningrum[4], Fathina Sahirah[5], Faza Inayatulloh[6], Grahsaro Yosha Teduhati[7], Hollyviar Resnias Putri Zalukhu[8], Muhammad Rafi Insani[9], Regytha Puteri Ayuningtyas[10], Yoel Prawira Simamora[11], Linda Dwi Rahmawati[12], Zelika Anindita Rachman[13], Syahwalia Asacha[14]

[1,2,3,4,5,6,7,8,9,10,11,12,13,14]Department of Statistics, Faculty of Science and Data Analytics,
Institut Teknologi Sepuluh Nopember,
Jl. Raya ITS Sukolilo, Surabaya, 60111, Jawa Timur, Indonesia

E-mail Correspondence Author: *novri@statistika.its.ac.id*

Abstract

Heart failure is a serious chronic condition and a leading cause of death globally. Early detection of mortality risk among heart failure patients remains a challenge due to the complexity of clinical data. This study aims to identify the most influential clinical factors associated with patient mortality and to compare the performance of two classification models, Logistic Regression and Extreme Gradient Boosting, in predicting death risk. The dataset includes clinical and demographic variables of heart failure patients. Key predictors identified include serum creatinine, ejection fraction, time, and age, which are clinically associated with kidney function, cardiac output, and treatment progression. These features were selected based on their relevance and contribution to the model's predictive performance. Model performance was evaluated using accuracy, precision, recall, F1-score, and AUC. Results indicate that XGBoost slightly outperformed Logistic Regression in terms of accuracy (85%) and recall (63%) compared to Logistic Regression (83% and 58%). However, Logistic Regression achieved a higher AUC (0.88) and showed more consistent results between training and testing data. Its interpretability also makes it more appropriate for clinical applications. This study underscores the potential of data-driven approaches in enhancing risk stratification and guiding early interventions in heart failure management.

Keywords: *Heart Failure, Logistic Regression, Machine Learning, XGBoost*

: https://doi.org/10.30598/parameterv5i1pp87-110

1. INTRODUCTION

Heart failure is a serious chronic condition in which the heart is unable to pump blood effectively to meet the body's needs. It is a leading cause of death worldwide, with approximately 78% of cardiovascular-related deaths occurring in low- and middle-income countries, according to the World Health Organization (WHO) and the World Heart Federation (WHF) [1]. In Asia, it is estimated that by 2025, heart disease will become the leading cause of death. According to data from the 2018 Basic Health Research (Riskesdas) by the Ministry of Health of the Republic of Indonesia, the prevalence of heart failure in Indonesia reached 0.5%, up from 0.3% in 2013 [2]. Based on medical diagnoses, the actual rate is estimated to be around 5%, with a higher incidence in men (66%) compared to women (34%). These data show a rising number of heart failure cases in Indonesia, which may place a significant burden on the national healthcare system, especially in regions with limited access to adequate health facilities and medical personnel. Despite advances in medical technology, early detection of the risk of death from heart failure is still a challenge due to the complexity of clinical data, highlighting the need for data-driven approaches to uncover hidden patterns.

As information and computing technology advances, numerous studies have explored the application of statistical and machine learning methods in health data analysis, including in the context of heart failure. For example, a study by Sugiharto [3] analyzed the use of the Logistic Regression model on patients with heart failure at Sosodoro Djatikoesoemo Bojonegoro Hospital. Another study by Farida and Bahri [4] applied the Support Vector Machine (SVM) method for heart failure classification and achieved the highest accuracy of 86.92%. However, some attributes in clinical datasets may provide little or no useful information for predictions, which can reduce the comprehensiveness and performance of the resulting models. Moreover, research gaps still exist, particularly in the use of classification models such as Logistic Regression and Extreme Gradient Boosting (XGBoost) to predict the likelihood of patient death based on clinical features. Although each model has its own strengths, limited studies have directly compared their performance on heart failure clinical datasets. Therefore, this study aims to address this gap by comparing the effectiveness of these two classification methods in order to identify a more accurate and clinically applicable predictive model.

This research has three main objectives. First, to identify the clinical factors that have the most influence on the mortality of heart failure patients. Second, to build a predictive model that can classify patients based on the risk of death using two different approaches, Logistic Regression as an interpretable statistical method [5], and XGBoost, a machine learning ensemble algorithm known for its efficiency and strong predictive performance in structured data classification tasks [6]. Third, to evaluate the performance of the models using classification evaluation metrics to quantitatively compare the effectiveness of both approaches.

Practically, this research is expected to assist medical professionals in more accurately detecting the risk of death in heart failure patients, so that clinical interventions can be carried out early and on target. Early detection supported by predictive modeling can improve the efficiency of medical resource allocation, especially in healthcare facilities with limited personnel and equipment [7]. Academically, this research contributes to the literature on the application of classification models in healthcare, particularly in the context of clinical heart failure data, which remains relatively underexplored in comparative algorithm studies. Furthermore, the theoretical framework of this research includes the basic concepts of Logistic Regression as a statistical model that quantifies the relationship between predictive variables with

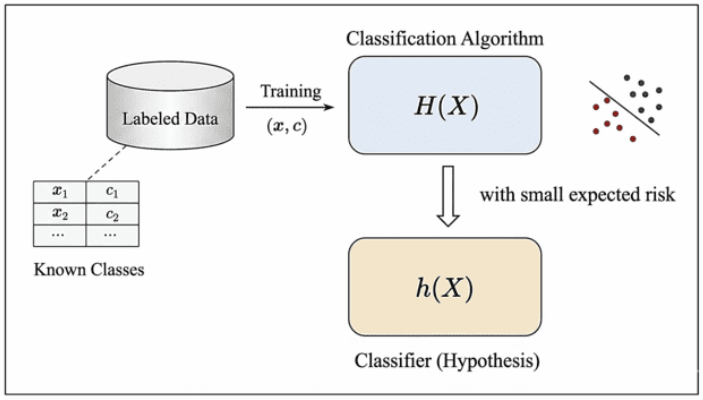
binary event probabilities and XGBoost as a tree-based boosting algorithm known for its effectiveness in structured data classification tasks. Performance evaluations are conducted using various evaluation metrics such as accuracy, precision, recall, and AUC that are commonly used in medical classification studies [8].

2. METHOD

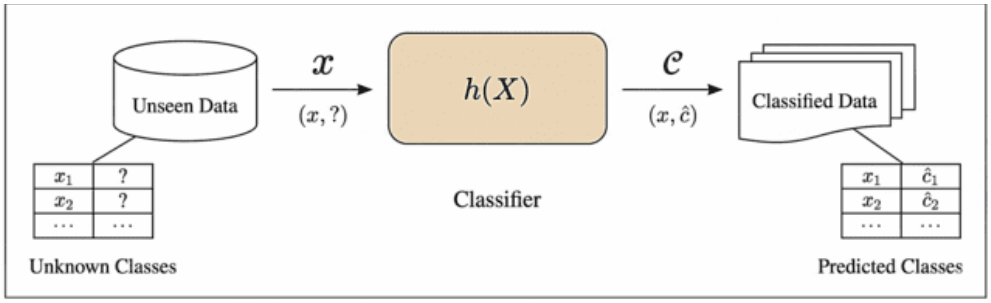
In this study, an analysis was conducted using Logistic Regression and XGBoost algorithms as approaches to classify the condition of heart failure patients. Clinical data were used to build predictive models by considering various variables such as age, blood pressure, ejection fraction, and history of chronic diseases. Both algorithms were analyzed and interpreted based on their respective parameters and used to evaluate predictive performance in a comparative manner.

2.1. Research Methodology

Classification in machine learning is a type of supervised learning [9]. Supervised learning algorithms learn from labeled data, meaning each input is paired with a known output or category. The goal of classification is to build a model that can assign new, unseen data to one of the predefined categories. In simple terms, machine learning classification is the process of identifying the correct label for given data using training data [10].



(a)



(b)

Figure 1. Classification Process from (a) Training and (b) Testing

As shown in Figure 1, the labeled data is used to train a classification algorithm with the aim of minimizing expected risk. The resulting classifier model is then applied to classify unlabeled data, producing predicted class labels. There are two types of classification functions: probabilistic classification and statistical classification. Probabilistic classification produces a probability estimate for each possible label of the

unlabeled data. The final classification is based on selecting the label with the highest probability. The probabilistic classification function can be expressed as follows:

$$h_j(\boldsymbol{x}) = \underset{j}{\operatorname{argmax}}\, P(c = c_j|\boldsymbol{x}) \quad (1)$$

This approach selects the category with the highest probability as the best prediction. Algorithms that utilize probabilistic classification in machine learning include logistic regression, Naive Bayes, and others. Statistical classification, on the other hand, directly assigns a class label to the data without computing the probabilities for all possible labels. The classification function for this classification can be represented as:

$$h_j(\boldsymbol{x}) = \underset{j}{\operatorname{argmax}}\, f(c_j = \boldsymbol{w}_j^T.\boldsymbol{x} + b) \quad (2)$$

Machine learning algorithms that use statistical classification include k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), among others.

2.2. Logistic Regression

Logistic Regression is a widely used statistical method for analyzing datasets in which the dependent variable is binary or dichotomous in nature. It models the relationship between one or more independent variables and the probability of a certain event occurring, such as success/failure, yes/no, or alive/deceased. Unlike linear regression, which assumes a continuous outcome, logistic regression is used to predict the log odds of the outcome variable using the logistic (sigmoid) function **[5]**.

The general form of the logistic regression model is given as:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad (3)$$

where:

p : probability of the event of interest
$X_1, X_2, \ldots, X_n$: predictor variables
$\beta_1, \beta_2, \ldots, \beta_n$: regression coefficients

These coefficients are estimated using Maximum Likelihood Estimation (MLE), which aims to find the parameter values that maximize the likelihood of observing the given data **[10]**. The results of logistic regression are often interpreted in terms of odds ratio, where an odds ratio greater than 1 indicates an increased likelihood of the outcome, while a value less than 1 suggests a reduced likelihood. One of the main advantages of logistic regression is its flexibility in handling various types of predictor variables, including continuous, ordinal, and categorical data. It also supports the inclusion of interaction terms to explore combined effects between variables. Furthermore, logistic regression does not assume normality, linearity, or homoscedasticity of the predictors, making it suitable for diverse practical applications **[11]**. The performance of logistic regression models is typically evaluated using classification metrics such as accuracy, precision, recall, F1-score, and the Area Under the ROC Curve (AUC). These metrics offer a comprehensive understanding of the model's ability to distinguish between outcome classes, particularly in imbalanced datasets. In addition to performance metrics, diagnostic tools such as the Wald test and the Hosmer-Lemeshow goodness-of-fit test are often employed to assess the statistical significance of individual predictors and the overall model fit **[5]**.

2.3. XGBoost

XGBoost (Extreme Gradient Boosting) is an ensemble learning algorithm based on the gradient boosting framework, designed to handle large-scale data efficiently. The algorithm is known for its optimized memory usage, parallel computation capability, and built-in support for handling missing values automatically. In addition, XGBoost employs a second-order Taylor approximation to accelerate optimization and incorporates both L_1 and L_2 regularization techniques to reduce overfitting. These technical advantages distinguish XGBoost from conventional gradient boosting methods. Numerous studies have demonstrated the effectiveness of XGBoost across various application domains. For example, previous studies reported strong predictive performance in medical classification, disaster assessment, forecasting, and malware detection tasks [12], [13], [14], [15], [16]. Due to its efficient algorithmic structure and consistently high empirical performance, XGBoost has become one of the most widely used machine learning algorithms for predictive modeling using structured data.

The main advantage of XGBoost lies in the objective function which includes two components: regularization of model complexity and loss function. The objective function can be written in **equation 4**.

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{4}$$

where $f_t \in \mathcal{F}$ is the new regression tree at iteration t, and $\Omega(f)$ is the complexity regularization function expressed in **equation 5**.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{5}$$

2.4. Feature Engineering

Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning [17]. This technique enables machine learning models to extract data patterns more effectively, especially in structured and relational data [18]. In classification and regression, feature engineering aims to transform / create features in such a way that complex relationships between variables become easier for the model to learn [19].

The first stages include pre-processing such as imputation of missing values, encoding categorical features (e.g., one-hot, label, target encoding), and normalizing or scaling features using z-score or min-max scaling. Such transformations are important for machine learning algorithms that are sensitive to feature scales, including Logistic Regression and distance-based methods such as K-Nearest Neighbors (KNN) and Support Vector Machines (SVM) [20].

One important aspect of feature engineering is feature selection, which is the process of determining the most relevant and informative subset of features in a dataset. The objectives include dimension reduction, training time acceleration, overfitting risk reduction, and model accuracy and interpretability improvement [21]. This process generally consists of data preprocessing, feature selection, method selection, feature subset evaluation, and model validation [22].

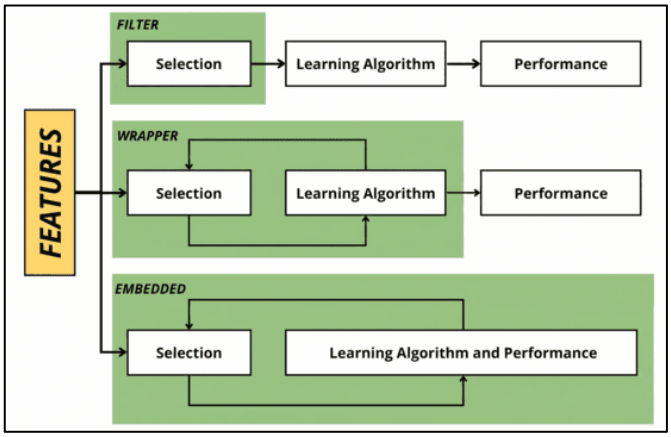


Figure 2. Feature Selection Methods

Another important aspect is feature selection, which is the process of selecting the most relevant subset of features to improve accuracy, reduce overfitting, and speed up training [21], [22]. There are three main approaches: filter (using statistical measures such as chi-square and correlation), wrapper (evaluating feature subsets with models such as SFS and RFE), and embedded (incorporating feature selection into model training, e.g., Lasso and Random Forest) as presented in **Figure 2** [23]. Selecting the right method and performing cross-validation after the selection process is important to ensure that the model remains generalizable. Thus, feature engineering not only optimizes model performance but also simplifies complexity and improves prediction quality in high-dimensional data.

2.5. Model Evaluation

Model evaluation is a crucial step in the development of machine learning models, including in the context of analyzing factors that influence heart failure. Evaluation aims to assess how well a model predicts the correct outcomes, as well as to measure the balance between positive and negative predictions. Four commonly used evaluation metrics are Accuracy, Precision, Recall, and F1-Score. Each metric has a different interpretation and purpose, and the selection of the appropriate evaluation metric depends on the specific task and data characteristic.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$F1 = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \tag{9}$$

where:
TP (True Positive) : number of actual positive instances correctly predicted as positive
TN (True Negative) : number of actual negative instances correctly predicted as negative
FP (False Positive) : number of actual negative instances incorrectly predicted as positive
FN (False Negative) : number of actual positive instances incorrectly predicted as negative

Area Under the Curve (AUC) is a performance metric for classification models, typically measured based on the Receiver Operating Characteristic (ROC) curve. The ROC curve illustrates the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR), while the AUC-ROC is the probability that the model can correctly distinguish between positive and negative classes at random **[24]**.

AUC values range from 0 to 1, where a value closer to 1 indicates better model performance. One advantage of AUC is its relative stability even when class distributions are imbalanced. However, in cases of extreme class imbalance, the Precision-Recall (PR) curve and the corresponding AUC-PR are considered more informative and representative than the ROC curve **[8]**.

2.6. Heart Failure

Heart Failure is one of the complex clinical syndromes with symptoms which resulted from the structural or functional disfigure of ventricular filling or blood ejection [4]. By 2025, heart failure is predicted to become the leading cause of death in Asia. Some common causes of heart failure are ischemic heart disease (coronary heart disease) and myocardial infarction (heart attack), hypertension, and valvular heart disease. Other causes that may not originate from the heart itself such as family genetics or substance abuse such as alcohol, cocaine, or methamphetamine **[6]**. In this case, some of the predicted variables are patient's age, anaemia, the level of creatinine phosphokinase (CPK), diabetes status, percentage of blood when leaving the patient's heart, hypertension status, count of platelets inside, the level of serum creatinine and serum sodium inside, patient's gender, smoking status, and the patient's follow up time after diagnosis.

2.7. Data Source

We analyzed a dataset containing clinical records of heart failure patients, obtained from the UCI Machine Learning Repository. This dataset includes medical information from 299 patients treated at the Faisalabad Institute of Cardiology and Allied Hospital in Faisalabad, Punjab, Pakistan. The dataset was used to support the classification process of patients' clinical conditions in this study.

2.8. Research Variables

Prior to conducting the analysis, it is essential to provide a clear explanation of the twelve variables contained within the dataset. Each variable represents either a clinical measurement, a demographic attribute, or a lifestyle-related factor associated with heart failure. These include both numerical variables-such as ejection fraction, serum sodium, and age-and categorical variables-such as diabetes, smoking status, and sex. A proper understanding of these variables and their respective data types is fundamental to ensuring the accuracy and relevance of the subsequent preprocessing, statistical testing, and predictive modeling. The preview of variables can be seen in

Table 1.

Table 1. Research Data Variable

Variable	Description	Category	Value Explanation
age (X_1)	Determines the age of the person.	Numerical	
anaemia (X_2)	A disorder with insufficient red blood cells; associated with poor health outcomes.	Categorical	0 = No anaemia, 1 = Has anaemia
creatinine_phospho kinase (X_3)	Enzyme level that indicates muscle, heart, or brain tissue damage or stress.	Numerical	
diabetes (X_4)	Indicates presence of diabetes, a metabolic condition with high blood sugar.	Categorical	0 = No diabetes, 1 = Has diabetes
ejection_fraction (X_5)	Measures the percentage of blood leaving the heart each contraction.	Numerical	
high_blood_pressu re (X_6)	Indicates whether the person has high blood pressure.	Categorical	0 = Normal, 1 = High blood pressure
platelets (X_7)	Determines the number of platelets in the blood.	Numerical	
serum_creatinine (X_8)	Indicates creatinine level in blood, used to assess kidney function.	Numerical	
serum_sodium (X_9)	Indicates the level of sodium in the blood.	Numerical	
sex (X_{10})	Gender of the patient (male or female).	Categorical	0 = Female, 1 = Male
smoking (X_{11})	Indicates if the person is a smoker.	Categorical	0 = Non-smoker, 1 = Smoker
time (X_{12})	Follow-up period in days.	Numerical	
death_event(Y)	Indicates whether the patient died during the follow-up period.	Categorical	0 = Survived, 1 = Died

2.9. Data Structure

This study used a dataset consisting of 12 clinical variables collected from heart failure patients, which included demographic, biochemical, and physiological attributes. This dataset includes features such as age, gender, blood test results, and clinical conditions such as anaemia, diabetes, and high blood pressure. These features serve as predictors to analyze and model the likelihood of a patient's death. Here is the data structure. Data structure preview can be seen in **Table 2**.

Table 2. Data Structure

Observation	X_1	X_2	⋯	X_{12}	Y
1	$x_{1.1}$	$x_{2.1}$	⋯	$x_{12.1}$	x_1
2	$x_{1.2}$	$x_{2.2}$	⋯	$x_{12.2}$	x_2
⋮	⋮	⋮	⋱	⋮	⋮

299	$x_{1.299}$	$x_{2.299}$	⋯	$x_{12.299}$	x_{299}

2.10. Analysis Steps

The analytical procedure employed in this study is outlined and presented in flowchart in **Figure 3**.

1. Import and inspect the dataset containing clinical records of 299 heart failure patients from the UCI Machine Learning Repository.
2. Conduct data preprocessing, including checking for missing values, encoding categorical variables, and normalizing numerical features.
3. Perform exploratory data analysis (EDA) to understand variable distributions, detect outliers, and explore basic correlations between clinical features and patient outcomes.
4. Perform feature engineering to enhance model learning, including interaction features and transformation of skewed variables.
5. Conduct feature selection using a hybrid method:

a) Filter method using Pearson correlation and chi-square tests.

b) Embedded method using feature importance scores from Logistic Regression and XGBoost.

6. Split the dataset into training and testing sets using stratified 80:20 ratio to ensure representative class distribution.
7. Develop predictive models using Logistic Regression with Maximum Likelihood Estimation for parameter estimation and XGBoost with default boosting parameters and regularization to avoid overfitting.
8. Evaluate both models using classification metrics (Accuracy, Precision, Recall, F1-score, and AUC).
9. Compare the model performance to identify the most accurate and clinically interpretable classifier in predicting patient mortality risk.
10. Interpret the results using odds ratio from Logistic Regression and feature importance scores from XGBoost.

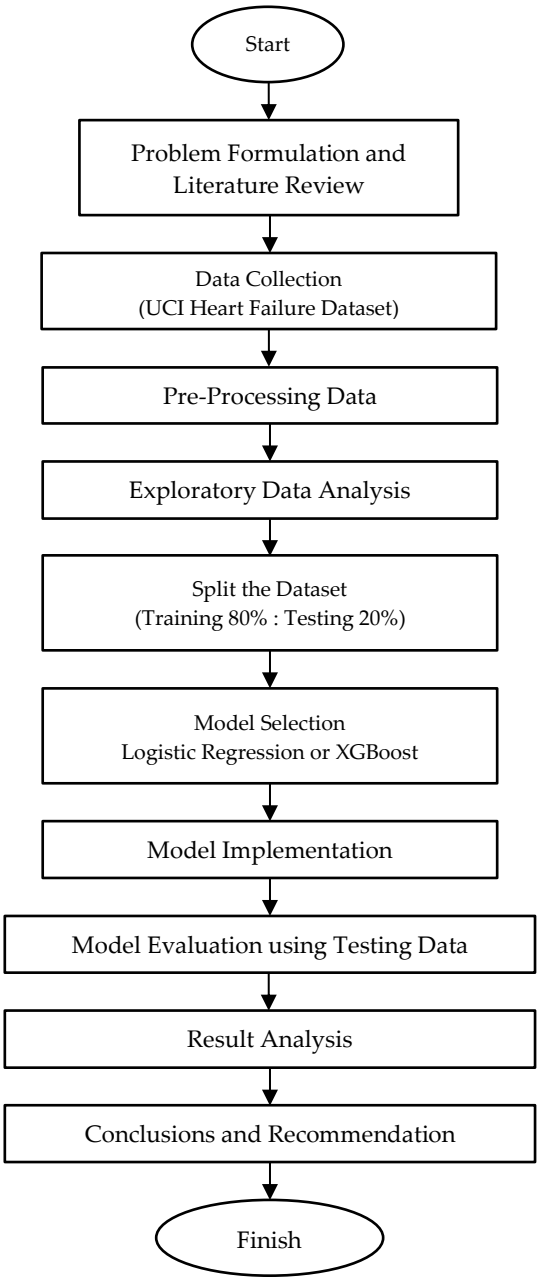


Figure 3. Flowchart analysis

3. RESULTS AND DISCUSSION

3.1. Exploratory Data Analysis for Heart Failure Patient Data

a) Proportion of Patients Survived and Deceased by Diabetes Status

The distribution of survival outcomes among patients is categorized by their diabetes status. **Figure 4** presents the distribution of survival outcomes among heart failure patients based on diabetes status during the follow-up period. Among all patients, the largest proportion consists of individuals without diabetes who survived the follow-up period (39.5%). Conversely, patients with diabetes who died represent the smallest segment (13.4%). It is notable that the combined proportion of deceased patients is higher in the diabetes group compared to the non-diabetes group. Specifically, while 28.4% of patients without diabetes died, a substantial percentage of diabetic patients did not survive. This pattern suggests that the presence of diabetes may contribute to an increased risk of mortality among heart failure patients. The visualization highlights the

importance of considering comorbid conditions such as diabetes when assessing prognosis and planning clinical management strategies.

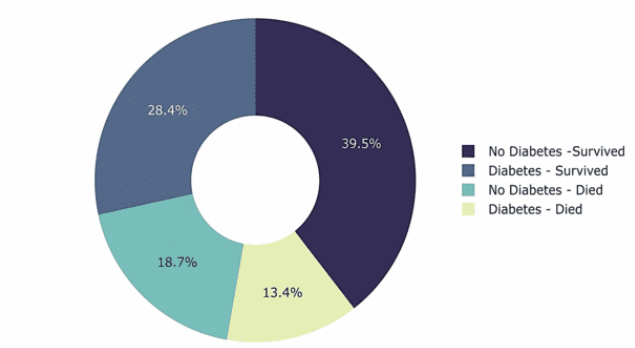
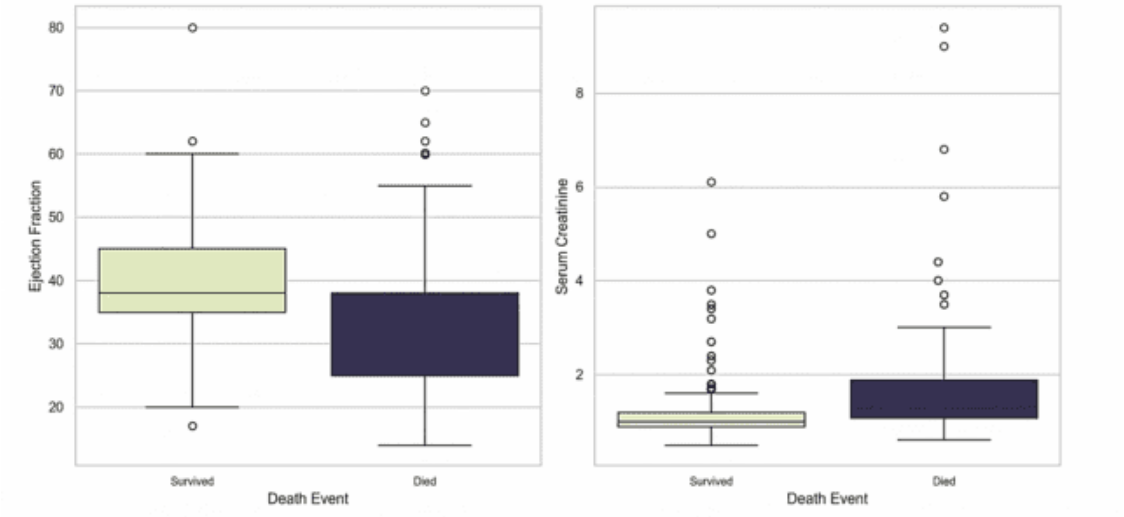


Figure 4. Distribution of Patients Survived and Deceased by Diabetes Status

b) Distribution Ejection Fraction and Serum Creatinine by Death Event

Figure 5 below displays the distribution of ejection fraction and serum creatinine levels between patients who survived and those who died. These boxplots illustrate how these clinical indicators differ according to survival outcomes.



(a) (b)

Figure 5. Boxplot of Ejection Fraction (a) and Serum Creatinine by Death Event (b)

The boxplot in **Figure 5** (a) illustrates the distribution of ejection fraction values among patients categorized by survival outcome. Patients who survived during the follow-up period generally exhibit higher ejection fraction levels, with the median positioned clearly above that of the deceased group. This indicates that, on average, survivors maintained better cardiac function. Additionally, the interquartile range among deceased patients is shifted downward, and the overall spread of values suggests a concentration of low ejection fraction measurements. This pattern highlights that reduced ejection fraction, which reflects impaired pumping capacity of the heart, is strongly associated with higher mortality risk in heart failure patients. The presence of several patients in the deceased group with ejection fraction values near the lower extreme further supports this association.

The boxplot in **Figure 5** (b) presents the serum creatinine levels, an important indicator of kidney function, also stratified by survival status. Patients who died tend to have higher serum creatinine measurements, as evidenced by a higher median and a

wider interquartile range compared to the survivors. The visualization shows a notable number of outliers at the upper end among deceased patients, suggesting that some individuals experienced severe renal dysfunction. This is clinically relevant because impaired kidney function often coexists with heart failure and contributes to poorer prognosis. Overall, the figure implies that elevated serum creatinine levels may be predictive of increased mortality risk, underscoring the importance of monitoring renal biomarkers alongside cardiac parameters in the management of heart failure patients.

c) Comparison of Heart Failure Patients Outcomes by Gender and Smoking Status

From **Figure 6**, the two largest patient groups are female non-smokers and male non-smokers, each forming over 100 individuals. In both groups, most patients survived, as shown by the larger proportion of the bar being the "Survived" category. This may indicate an association between non-smoking status and better survival outcomes, regardless of gender, and may imply that patients who do not smoke have a higher chance of survival when experiencing heart failure.

The group of male smokers also shows a relatively large number of patients. While this group does exhibit a slightly higher proportion of deaths compared to their non-smoking counterparts, the majority of patients still survived. This indicates that although smoking may increase the risk of death, it does not drastically alter survival outcomes for male patients in this dataset. However, the difference may warrant further investigation through statistical modeling to assess the significance and strength of the association.

On the other hand, the group of female smokers is notably small, with only a few patients represented. Despite the limited sample size, most individuals in this category experienced death during the study period. While this might suggest a higher vulnerability for female smokers, the very small number of observations makes it difficult to draw firm conclusions. The result could be influenced by sampling variability or other confounding factors that are not captured in this visualization. The visualization provides an initial insight into how gender and smoking status may relate to patient mortality. It appears that non-smokers, both male and female, generally have better survival outcomes. However, to confirm and quantify these patterns, further analysis using statistical methods such as logistic regression or machine learning models is necessary.

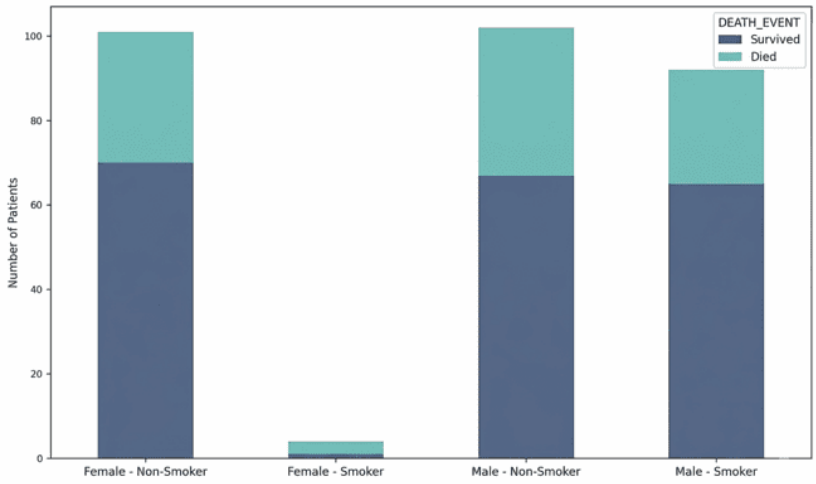


Figure 6. Comparison of Heart Failure Patient Outcomes by Gender and Smoking Status

d) Heatmap Visualization of Variable Correlations in Clinical Data

To explore the relationships among the variables in the dataset, a Pearson correlation matrix was constructed and visualized using a heatmap. The values in the matrix range from -1 to +1, where a value close to +1 indicates a strong positive linear relationship, a value close to -1 indicates a strong negative linear relationship, and a value near 0 indicates little to no linear association between the variables.

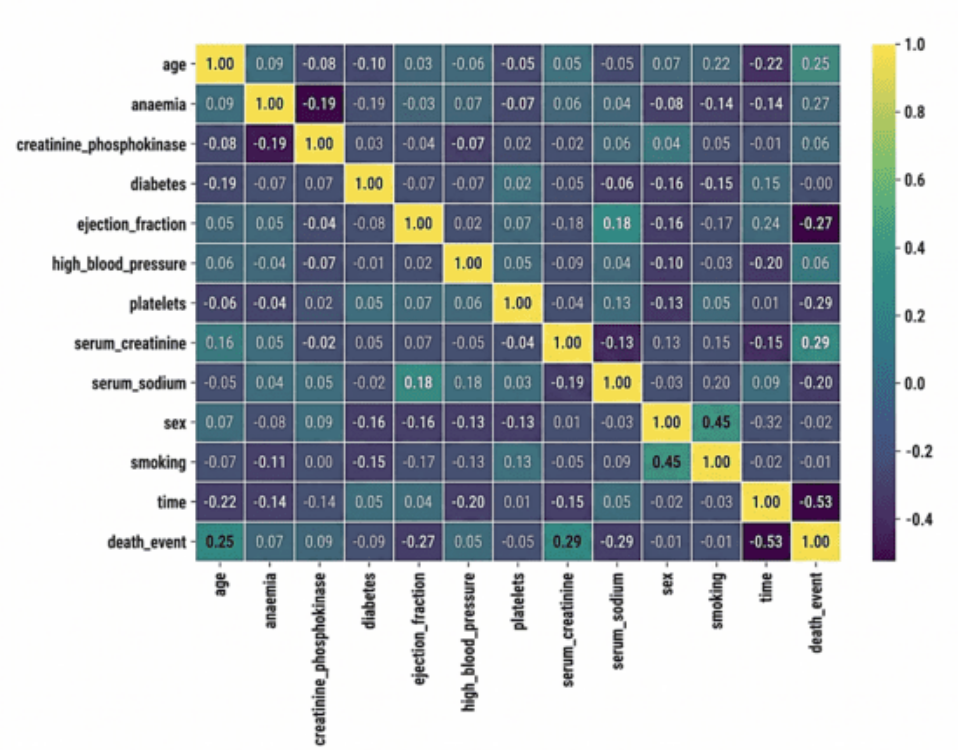


Figure 7. Variable Correlation Visualization

Figure 7 presents the Pearson correlation heatmap among all variables in the dataset. Several variables demonstrated notable linear relationships, particularly with the outcome variable, death_event (Y), which represents patient mortality during the follow-up period. Only variables with an absolute correlation coefficient greater than or equal to 0.25 are emphasized due to their potential statistical relevance in predictive modeling. The strongest negative correlation with death_event was observed for time (X_{12}), with a correlation coefficient of r = -0.53. This result indicates that patients with longer follow-up periods were less likely to experience death during the observation period. The strong inverse relationship suggests that time may serve as an important indicator in survival prediction analysis and should therefore be carefully considered in the modeling process.

In contrast, serum_creatinine (X_8) showed a moderate positive correlation with death_event (r = 0.29). This finding suggests that higher serum creatinine levels, which indicate impaired kidney function, are associated with an increased risk of mortality among heart failure patients. Similarly, age (X_1) demonstrated a positive correlation with death_event (r = 0.25), implying that older patients tend to have a higher probability of death compared to younger patients. Both variables may therefore act as important risk factors in classification and prediction models. Furthermore, ejection_fraction (X_5) exhibited a negative correlation with death_event (r = -0.27). Since ejection fraction measures the percentage of blood pumped out of the heart during each contraction, this

result indicates that patients with higher cardiac pumping efficiency generally have a lower risk of mortality. Consequently, ejection_fraction may serve as a protective clinical indicator in heart failure prognosis.

In addition to correlations with the target variable, a noteworthy inter-variable correlation was identified between sex (X_{10}) and smoking (X_{11}), with a coefficient of r = 0.45. This relationship suggests that smoking behavior differs across gender groups within the dataset, which may indirectly influence heart failure outcomes and should therefore be taken into account during model interpretation and feature analysis. Overall, the variables time, serum_creatinine, age, and ejection_fraction demonstrated meaningful correlations with death_event and may play significant roles in subsequent predictive modeling. Nevertheless, further statistical validation and feature selection techniques are still necessary to confirm their significance and predictive contribution to the proposed machine learning models.

3.2. Descriptive Statistics

Table 3 presents the descriptive statistics of the clinical and demographic variables used in this study. The statistics include the mean, standard deviation, minimum value, quartiles, median, and maximum value for each variable to provide an initial overview of the dataset characteristics.

Table 3. Descriptive Statistics of The Data

Variable	Mean	Std	Min	25%	50%	75%	Max
Age	60.83	11.89	40	51	60	70	95
Anaemia	0.43	0.50	0	0	0	1	1
Creatinine	581.84	970.29	23	116.5	250	582	7861
Diabetes	0.42	0.49	0	0	0	1	1
Ejection Fraction	38.08	11.83	14	30	38	45	80
Blood Pressure	0.35	0.48	0	0	0	1	1
Platelets	263.358	97.804	25.100	212.500	262.000	303.500	850.000
Serum Creatinine	1.39	1.03	0.5	0.9	1.1	1.4	9.4
Serum Sodium	136.63	4.41	113	134	137	140	148
Sex	0.65	0.48	0	0	1	1	1
Smoking	0.32	0.47	0	0	0	1	1
Time	130.26	77.61	4	73	115	203	285

Based on Table 3, several variables exhibit substantial variability, particularly creatinine phosphokinase and platelet counts, as indicated by their large standard deviations and wide value ranges. In addition, binary variables such as anaemia, diabetes, high blood pressure, sex, and smoking demonstrate relatively balanced distributions across patients. Overall, these findings indicate considerable heterogeneity in the demographic and clinical characteristics of heart failure patients included in this study.

The average patient age was 60.83 years, indicating that most participants were older adults, which is consistent with the higher prevalence of heart failure among elderly populations. The age distribution ranged from 40 to 95 years, suggesting substantial variation in patient demographics. Furthermore, the ejection fraction showed a mean value of 38.08% with a standard deviation of 11.83, ranging from 14% to 80%. Since normal ejection fraction typically ranges between 50% and 70%, these results suggest that many patients experienced moderate to severe impairment in cardiac function.

Creatinine phosphokinase levels showed a very high variability, with a mean of 581.84 mcg/L but a standard deviation exceeding 970. This indicates the presence of extreme outliers or cases with markedly elevated enzyme levels, likely reflecting acute or chronic muscle damage, including myocardial injury. Serum creatinine had a mean of 1.39 mg/dL, but the maximum value reached 9.4 mg/dL, indicating that some patients suffered from severe renal dysfunction. Similarly, serum sodium values ranged from 113 to 148 mEq/L, with an average close to 137 mEq/L, reflecting mostly normal to slightly altered electrolyte levels. Regarding categorical variables expressed numerically (0/1), the mean values show that anaemia (43%), diabetes (42%), and high blood pressure (35%) were prevalent comorbid conditions in this sample. The proportion of smokers was lower (32%), while the sex variable had a mean of 0.65, implying a higher proportion of male patients. Finally, the death event variable, with a mean of 0.32, indicates that approximately 32% of patients died during the follow-up period, underscoring the seriousness of heart failure outcomes. Overall, the table shows substantial heterogeneity in patient demographics and clinical characteristics, suggesting that multiple factors, such as impaired heart function, kidney dysfunction, and comorbidities, may contribute to variations in survival.

3.3. Feature Engineering

Based on the data, two important preprocessing techniques used to improve model performance were feature scaling and feature selection using the SelectKBest method. The following subsections describe these preprocessing techniques. Feature scaling is one of the most important preprocessing steps because scaling numerical data to a similar range allows the model to learn features more effectively and improves model robustness. In this study, Min-Max Scaling was applied to transform feature values into a range between 0 and 1. This method is particularly suitable for Logistic Regression because the algorithm is sensitive to differences in feature scales.

Another feature engineering technique used in this study was feature selection, which uses the SelectKBest function to select the appropriate feature (predictor) for each model. In this case, the selected features were based on the highest F-Score values, meaning that the selected features have more effect than the others. Below is the visualization for the F-Score of each predictor.

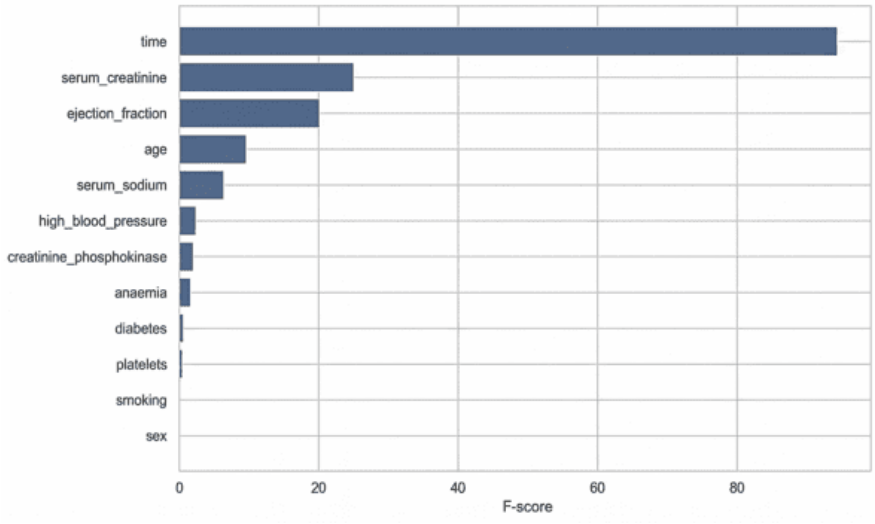


Figure 8. F-Score Results Visualization

Based on the F-Score presented in **Figure 8** and

Table 4, the chosen variables for the Logistic Regression model are Time, Serum Creatinine, Ejection Fraction, Age, and Serum Sodium. Meanwhile the variables used for the XGBoost model include all of the mentioned variables.

Table 4. F-Score Results

Variable	F-Score
Time	94.81
Serum Creatinine	24.88
Ejection Fraction	20.04
Age	9.55
Serum Sodium	6.33
Blood Pressure	2.32
Creatinine	1.97
Anaemia	1.58
Diabetes	0.49
Platelets	0.46
Smoking	0.20
Sex	0.04

3.4. Result of Logistic Regression

Before building the logistic regression model, feature selection was performed using the SelectKBest method with the F-statistical score function (f_classif) to identify the most relevant predictors of the death_event variable. Based on the F-Score results in **Table 4**, the selected features were serum_creatinine, ejection_fraction, age, time, and serum_sodium, as these variables showed the strongest contribution to mortality classification.

Next, hyperparameter tuning was carried out using Grid Search combined with cross-validation techniques. The parameter grid includes the range of values for feature_selection_k (3, 5, 7, 10, 'all'), lrC (0.01, 0.1, 1, 10), lr_penalty ('l2', 'none'), and lr_max_iter (100, 1000, 2500, 5000). The best models are selected based on the average AUC (Area Under the Curve) score during cross-validation. In addition to AUC, additional metrics such as accuracy, precision, recall, and F1-score are also used to evaluate the balance between positive and negative classifications. Based on the Grid Search results, the best parameter combination obtained is {'feature_selection_k': 5, 'lrC': 10, 'lr_max_iter': 100, 'lr_penalty': 'l2'}. This means that the optimal model uses the 5 best features and applies relatively weak L2 regularization (C=10), indicating that the data is less susceptible to overfitting or that stronger regularization may hinder performance.

A logistic regression model with the optimal configuration was evaluated using several classification metrics, including accuracy, precision, recall, and F1-score for each class, as presented in **Table 5**.

Table 5. Results of Logistic Regression Model Evaluation (Testing Set)

Class	Precision	Recall	F1-Score	Support
0 (survived)	0.83	0.95	0.89	41
1 (death)	0.85	0.58	0.69	19
Accuracy			0.83	60
Macro Avg	0.84	0.77	0.79	60
Weighted Avg	0.83	0.83	0.82	60

To measure the model's overall discriminating ability, the AUC (Area Under the Curve) metric is used. The AUC value obtained was 0.88, showing that the model has

good performance in distinguishing positive and negative classes. The ROC curve and AUC (Area Under the Curve) value are shown in Figure 9. ROC-Curve Logistic Regression indicate that the model not only achieved a good accuracy rate of 83% on the test data, but also had adequate discriminative and generalization capabilities. However, it should be noted that the recall in the positive class (death incidence) is lower than in the negative class. This phenomenon reflects a trade-off common in medical classifications, where increased sensitivity to minority classes is often a challenge.

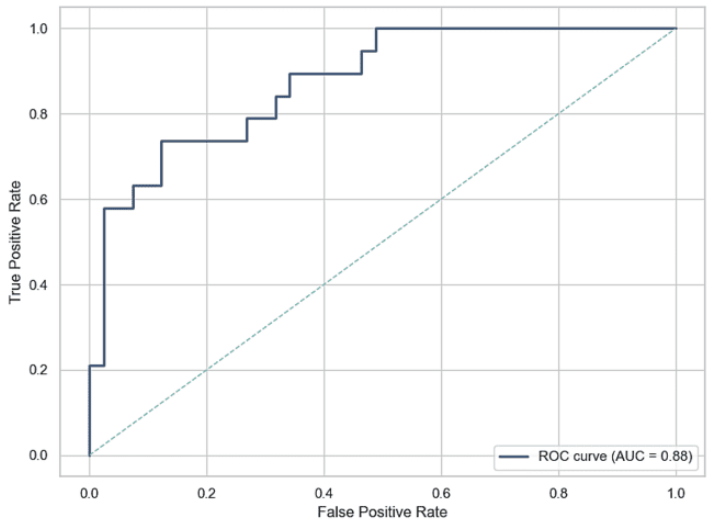


Figure 9. ROC-Curve Logistic Regression

To ensure that the model does not experience overfitting or underfitting, evaluation is carried out on both training data and data testing. Table 6 and Table 7 respectively present a confusion matrix and classification report for the training dataset, providing a detailed overview of the model's performance on the trained data. The model correctly classified 147 of the 162 patients who survived, but erroneously misclassified 22 of the 77 patients who died as non-dead.

Table 6. Confusion Matrix of Logistic Regression (Training Set)

Actual \ Predicted	Pred: 0	Pred: 1
Actual: 0 (Survived)	147	15
Actual: 1 (Death)	22	55

Table 7. Results of Logistic Regression Model Evaluation (Training Set)

Class	Precision	Recall	F1-Score	Support
0 (survived)	0.87	0.91	0.89	162
1 (death)	0.79	0.71	0.75	77
Accuracy			**0.85**	239
Macro Avg	0.83	0.81	0.82	239
Weighted Avg	0.84	0.85	0.84	239

Based on the results of the evaluation on the training data, the model achieved an accuracy of 85%. This performance shows that the model is able to learn from training data without showing signs of extreme overfitting. Performance comparisons on training and test data show that the model is performing stably and does not suffer

significant performance degradation. Although recalls in the minority class (death) tend to be lower, this is a common challenge in unbalanced data classification. Therefore, the logistic regression model in this study can be said to be quite reliable and can be used as a strong baseline to be compared with more complex models.

3.5. Result of XGBoost Model

Based on feature selection using feature importance that has been described in the feature engineering section, all features were used for XGBoost modeling. The selection of these features was further supported by the results of hyperparameter tuning using grid search. The grid search procedure involved 1600 iterations, as it systematically maps all possible combinations of specified hyperparameters into a grid and evaluates each combination exhaustively. The wider the hyperparameter space, the more computationally complex the search process becomes. Therefore, the hyperparameter search space in this study was intentionally limited to reduce computational complexity while still allowing effective parameter optimization. The optimal hyperparameter configuration obtained from the grid search process is presented in **Table 8**.

Table 8. XGBoost Hyperparameters

Hyperparameter	Domain		Optimal Value
	Min	Max	
feature_selection_k	3	all	all
n_estimators	100	200	100
max_depth	3	5	3
learning_rate	0.01	0.1	0.1
subsample	-	0.8	0.8
colsample_bytree	-	0.8	0.8
gamma	0	0.1	0.1
reg_alpha	0	0.1	0
reg_lambda	1	2	1

After obtaining the optimal hyperparameter configuration, the performance of the XGBoost model was evaluated using several classification metrics, including precision, recall, F1-score, and accuracy. The evaluation results on the testing dataset are presented in **Table 9**.

Table 9. Results of XGBoost Model Evaluation (Testing Set)

Class	Precision	Recall	F1-Score	Support
0 (survived)	0.85	0.95	0.90	41
1 (death)	0.86	0.63	0.73	19
Accuracy			**0.85**	60
Macro Avg	0.85	0.79	0.81	60
Weighted Avg	0.85	0.85	0.84	60

Based on the model evaluation presented in **Table 9**, the XGBoost model achieved an overall accuracy of 0.85, indicating that 85% of the predictions were classified correctly. In addition to accuracy, the Receiver Operating Characteristic (ROC) curve was used to evaluate the model's ability to distinguish between positive and negative classes across different classification thresholds. The ROC curve and the corresponding Area Under the Curve (AUC) score are presented in **Figure 10**. The model obtained an AUC value of 0.86, indicating good classification performance. This result suggests that the

model is capable of ranking a randomly selected positive instance higher than a randomly selected negative instance approximately 86% of the time.

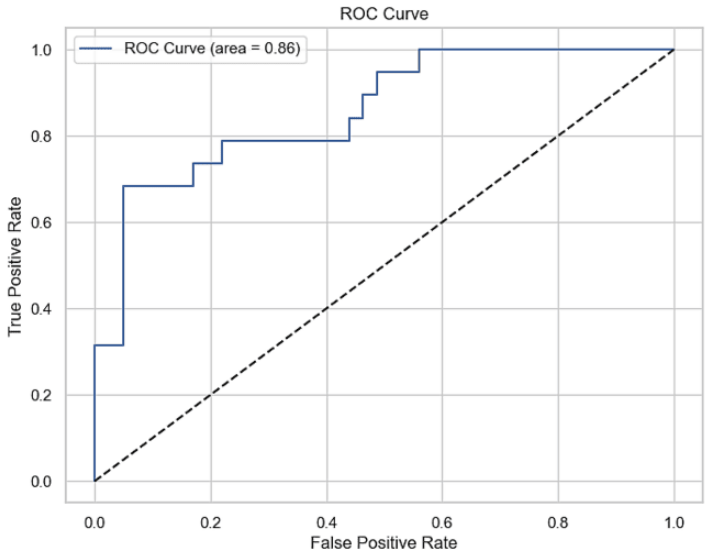


Figure 10. ROC-Curve XGBoost

The best-performing XGBoost model identified the following 12 features as meaningful contributors: age, anaemia, creatinine_phosphokinase, diabetes, ejection_fraction, high_blood_pressure, platelets, serum_creatinine, serum_sodium, sex, smoking, time contributed meaningfully to the model's predictions. By focusing on these features, the model is expected to reduce the influence of irrelevant variables and improve overall predictive performance. After training the final XGBoost model using the optimal hyperparameter configuration, model evaluation was conducted on both the testing and training datasets. To provide a more detailed understanding of the classification results, a confusion matrix for the testing set is presented in **Table 10**.

Table 10. Confusion Matrix of XGBoost (Testing Set)

Actual \ Predicted	Pred: 0	Pred: 1
Actual: 0 (Survived)	39	2
Actual: 1 (Death)	7	12

Based on the confusion matrix in **Table 10**, the model correctly classified 39 out of 41 surviving patients and 12 out of 19 deceased patients. However, several death cases were still misclassified as survived, indicating that the model has some limitations in detecting all positive cases. To further evaluate the consistency of the model, performance evaluation on the training dataset is presented in **Table 11**.

Table 11. Results of XGBoost Model Evaluation (Training Set)

Class	Precision	Recall	F1-Score	Support
0 (survived)	1.00	0.99	0.99	162
1 (death)	0.97	1.00	0.99	77
Accuracy			**0.99**	239
Macro Avg	0.99	0.99	0.99	239
Weighted Avg	0.99	0.99	0.99	239

The model's performance on the training set presented in **Table 11** was significantly higher, with an overall accuracy of 99%. Both classes were classified with near-perfect

precision and recall, resulting in an F1-score of 0.99 for both classes. Meanwhile, on the testing set, the model achieved an accuracy of 85%, with a precision of 0.85 and recall of 0.95 for class 0, and a precision of 0.86 and recall of 0.63 for class 1. These results indicate that the model performs well overall, particularly in identifying class 0 instances. However, the lower recall for class 1 suggests that some positive cases may still be misclassified.

Although the XGBoost model achieved excellent results on the training dataset, the noticeable gap between training and testing accuracy indicates a potential tendency toward overfitting. This suggests that the model may have learned patterns specific to the training data that do not fully generalize to unseen samples. Nevertheless, the model still demonstrated good predictive capability on the testing dataset.

3.6. Model Comparison

Logistic Regression and XGBoost were evaluated to compare their effectiveness in predicting mortality among heart failure patients. The evaluation focused on the performance of both models in classifying the positive class, representing patients who died, as this outcome is clinically critical. Performance was assessed using accuracy, precision, recall, F1-score, and the Area Under the Curve (AUC), measured on training and test datasets to examine predictive capability and generalization. The evaluation results are presented in **Table 12**.

Table 12. Performance Metrics of Logistic Regression and XGBoost

Model	Train Set				Test Set				AUC
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	
Logistic Regression	0.85	0.79	0.71	0.75	0.83	0.85	0.58	0.69	0.88
XGBoost	0.99	0.97	1.00	0.99	0.85	0.86	0.63	0.73	0.86

XGBoost achieved slightly higher accuracy (0.85) compared to Logistic Regression (0.83). In terms of recall, which reflects the model's ability to correctly identify patients who experienced death, XGBoost also outperformed Logistic Regression (0.63 versus 0.58). However, Logistic Regression achieved a slightly higher AUC score of 0.88 on the test set, compared to 0.86 for XGBoost, indicating better discrimination capability across different classification thresholds.

On the training set, XGBoost demonstrated near-perfect performance across all evaluation metrics, with accuracy, precision, recall, and F1-score approaching 1.00. While this indicates the model's strong ability to capture complex patterns in the training data, it also suggests a potential tendency toward overfitting, as reflected by the lower performance on the testing set. In contrast, Logistic Regression showed more consistent results between the training and testing sets, indicating better model stability and generalization capability.

In terms of interpretability, Logistic Regression offers a significant advantage because its coefficients provide direct insights into how each predictor variable influences the probability of mortality. This characteristic facilitates clinical interpretation and decision-making. Conversely, XGBoost, as an ensemble-based model, has a more complex structure and generally requires additional interpretability techniques, such as feature importance analysis or SHAP values, to explain its predictions.

Overall, although XGBoost demonstrated slightly superior predictive performance in several evaluation metrics, Logistic Regression remains valuable due to its higher interpretability and more stable generalization performance. These characteristics are particularly important in clinical applications where transparency and reliability of predictive models are essential.

4. CONCLUSION

Based on the comprehensive analyses conducted, this study demonstrates that machine learning models, particularly Logistic Regression and XGBoost, can effectively predict mortality among heart failure patients using clinical and demographic variables. Key features such as time, serum creatinine, ejection fraction, and age emerged as important predictors, highlighting the significant role of cardiac and renal function in patient survival outcomes. While XGBoost achieved slightly better predictive performance, particularly in identifying deceased patients, its complex structure may reduce model interpretability. In contrast, Logistic Regression provided more stable and interpretable results, making it potentially more suitable for clinical decision-making environments where transparency and explainability are essential.

This study has several limitations. First, the dataset used in this research was relatively small, consisting of only 299 patient records, which may limit the robustness and generalizability of the developed models. Second, the dataset originated from a single source, potentially limiting the applicability of the findings to broader and more diverse populations. In addition, the substantial performance gap between the training and testing results in the XGBoost model suggests a potential tendency toward overfitting. Therefore, future studies should validate these models using larger and multi-center clinical datasets, while also exploring additional ensemble learning approaches and advanced feature selection techniques to improve model generalization and predictive stability.

Overall, the findings of this study highlight the potential of data-driven predictive models to support risk stratification and assist healthcare professionals in identifying high-risk heart failure patients more effectively. With appropriate validation and optimization, such approaches may contribute to improved clinical decision-making and more efficient healthcare resource allocation.

REFERENCES

[1] J. P. Ferreira *et al.*, "World Heart Federation Roadmap for Heart Failure," *Glob. Heart*, vol. 14, no. 3, p. 197, Sep. 2019, doi: 10.1016/j.gheart.2019.07.004.

[2] Badan Penelitian dan Pengembangan Kesehatan, "Laporan Nasional Riskesdas 2018," Jakarta, 2019.

[3] S. A, "PENERAPAN METODE REGRESI LOGISTIK BINER PADA PASIEN PENYAKIT GAGAL JANTUNG DI RSUD SOSODORO DJATIKOESOEMO BOJONEGORO," Universitas Nadhlatul Ulama Sunan Giri, Bojonegoro, 2024.

[4] L. N. Farida and S. Bahri, "Klasifikasi Gagal Jantung menggunakan Metode SVM (Support Vector Machine)," *Komputika : Jurnal Sistem Komputer*, vol. 13, no. 2, pp. 149–156, Oct. 2024, doi: 10.34010/komputika.v13i2.11330.

[5] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*. Wiley, 2013. doi: 10.1002/9781118548387.

[6] L. M. Cendani and A. Wibowo, "Perbandingan Metode Ensemble Learning pada Klasifikasi Penyakit Diabetes," *Jurnal Masyarakat Informatika*, vol. 13, no. 1, pp. 33–44, May 2022, doi: 10.14710/jmasif.13.1.42912.

[7] S. Chae *et al.*, "Developing a clinical decision support framework for integrating predictive models into routine nursing practices in home health care for patients with heart failure," *Journal of Nursing Scholarship*, vol. 57, no. 1, pp. 165–177, Jan. 2025, doi: 10.1111/jnu.13030.

[8] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," *PLoS One*, vol. 10, no. 3, p. e0118432, Mar. 2015, doi: 10.1371/journal.pone.0118432.

[9] H. M and S. M.N, "A Review on Evaluation Metrics for Data Classification Evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, pp. 01–11, Mar. 2015, doi: 10.5121/ijdkp.2015.5201.

[10] S. Menard, *Applied Logistic Regression Analysis*. 2455 Teller Road, Thousand Oaks California 91320 United States of America : SAGE Publications, Inc., 2002. doi: 10.4135/9781412983433.

[11] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An Introduction to Logistic Regression Analysis and Reporting," *J. Educ. Res.*, vol. 96, no. 1, pp. 3–14, Sep. 2002, doi: 10.1080/00220670209598786.

[12] S.-K. Di, Y.-Y. Wang, D. Yang, Y.-H. Liu, J. Zhang, and W.-Z. Zheng, "SMOTE-enhanced XGBoost for rapid seismic damage assessment of bridge portfolios," *Soil Dynamics and Earthquake Engineering*, vol. 199, p. 109712, Dec. 2025, doi: 10.1016/j.soildyn.2025.109712.

[13] N. Qin *et al.*, "Forecasting the mechanical compaction influence on soybean yield using XGBoost-ANN," *Information Processing in Agriculture*, Sep. 2025, doi: 10.1016/j.inpa.2025.09.002.

[14] X. Song, J. Shi, C. Zhu, F. Xian, Z. Dong, and J. Li, "XGBoost machine learning algorithm for predicting unplanned readmission in elderly patients with coronary heart disease," *Geriatr. Nurs. (Minneap).*, vol. 66, p. 103609, Nov. 2025, doi: 10.1016/j.gerinurse.2025.103609.

[15] A. R. Zaidi, T. Abbas, A. Daud, O. Alghushairy, H. Dawood, and N. Sarwar, "Enhancing Android Malware Detection with XGBoost and Convolutional Neural Networks," *Computers, Materials & Continua*, vol. 84, no. 2, pp. 3281–3304, 2025, doi: 10.32604/cmc.2025.063646.

[16] Z. Mustaffa and M. H. Sulaiman, "Advanced forecasting of building energy loads with XGBoost and metaheuristic algorithms integration," *Energy Storage and Saving*, Aug. 2025, doi: 10.1016/j.enss.2025.03.005.

[17] H. Patel, *What is feature engineering—importance, tools and techniques for machine learning*. Towards Data Science, 2021.

[18] S. Reddy, S. B. Dodda, S. Maruthi, and M. Raparthi, "AI-Driven Automated Feature Engineering to Enhance Performance of Predictive Models in Data Science," *International Journal of Control and Automation*, 2020, doi: 10.52783/ijca.v13i4.38349.

[19] F. Nargesian, H. Samulowitz, U. Khurana, E. B. Khalil, and D. Turaga, "Learning Feature Engineering for Classification," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, California: International Joint Conferences on Artificial Intelligence Organization, Aug. 2017, pp. 2529–2535. doi: 10.24963/ijcai.2017/352.

[20] S. Maddula, "Feature Engineering in Machine Learning: A Practical Guide," datacamp.

[21] A. Rahmansyah, O. Dewi, P. Andini, T. H. P. Ningrum, and M. E. Suryana, "Membandingkan Pengaruh Feature Selection Terhadap Algoritma Naïve Bayes dan Support Vector Machine," *Seminar Nasional Aplikasi Teknologi Informasi*, 2018.

[22] U. Umar, I. H. Al Ghozali, and A. R. Handoko, "Analisis Efektifitas Feature Selection dalam Pengkayaan Machine Learning untuk Deteksi Dini Risiko Putus Kuliah Mahasiswa," *Jurnal Edukasi & Penelitian Informatika*, vol. 11, no. 1, 2025.

[23] K. Koirunnisa, A. M. Siregar, and S. Faisal, "Optimized Machine Learning Performance with Feature Selection for Breast Cancer Disease Classification," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 9, no. 4, pp. 1131–1143, Dec. 2023, doi: 10.26555/jiteki.v9i4.27527.

[24] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd international conference on Machine learning - ICML '06*, New York, New York, USA: ACM Press, 2006, pp. 233–240. doi: 10.1145/1143844.1143874.