Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2026.1989

Research Article

Predicting survival rates of patients with cardiovascular diseases using ensemble techniques

Bharati KARARE[1,*], Anushree Anand PANDE[2], Pratibha WAGHALE[1], Amruta Tapas PAUL[1], Chanchla TRIPATHI[1], Lalit DAMAHE[1]

[1]*Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, 441110, India*
[2]*P R Pote Patil College of Engineering and Management Amravati, Maharashtra, 444604, India*

ARTICLE INFO

Article history
Received: 07 August 2024
Revised: 31 October 2024
Accepted: 30 December 2024

Keywords:
Cardiovascular Disease; Kaplan-Meier Estimator; Machine Learning; Survival Rate

ABSTRACT

This study focuses on predicting cardiovascular diseases and forecasting the survival rate of patients using machine learning techniques. In this paper publicly available dataset is sued that contain the several risk factors for cardiovascular disease such as age, anemia, hypertension, diabetes, smoking habits, gender, blood pressure, glucose levels, and alcohol consumption. The dataset is preprocessed to extract relevant features for predicting survival rates based on risk parameters. The ensemble model is designed based on combining the several machine learning algorithms such as Logistic Regression, Random Forest, XGBoost, and Naive Bayes to classify the cardiovascular disease. The study also applies the Kaplan-Meier estimator to predict the survival rate of patient based on continuous variables. The final results indicate that age, serum creatinine, and ejection fraction significantly correlate with the death event. At the same time, smoking, sex, platelets, and diabetes show uncertain statistical significance. The study provides insights into the impact of various factors on cardiovascular disease survival rates. The novelty of this work lies in its integration of survival analysis with ensemble learning for robust prediction and interpretability in clinical applications.

Cite this article as: Karare B, Pande AA, Waghale P, Paul AT, Tripathi C, Damahe L. Predicting survival rates of patients with cardiovascular diseases using ensemble techniques. Sigma J Eng Nat Sci 2026;44(1):404–422.

INTRODUCTION

Heart and blood vessel problems like heart attacks, strokes, and heart failure are known as cardiovascular diseases (CVDs). Around 17 million people die from these diseases every year worldwide [1]. In India, the number of deaths due to CVDs has been increasing for the first time in 30 years. Heart failure occurs when cardiovascular system cannot circulate blood throughout the human body. Excessive hypertension, diabetic complications, or other cardiac problems frequently bring it on. Doctors classify heart failure into two types based on how much blood the heart pumps out with each beat [2]. One type, heart failure with reduced ejection fraction (HFrEF) [3], happens when

*Corresponding author.
*E-mail address: kararebharati@gmail.com

This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkilic

the heart pumps less than 45% of the blood. The other type, heart failure with preserved ejection fraction (HFpEF), occurs when the heart contracts well but doesn't relax properly to fill with blood. Most heart attacks can be prevented by using population-wide strategies to address lifestyle risk factors such as smoking, unhealthy diets and obesity, and alcohol consumption [4]. The machine learning is an effective strategy to early prediction and proper medication of patient with heart failure or who are at risk due to the combination of several risk factors such as BP, diabetics, and other medical history [5]. This potential brings optimism for the future of cardiovascular disease management.

Machine learning (ML) might surpass current modeling methods to accurately predict high blood pressure within specific racial groups and elucidate critical factors contributing to high blood pressure development across diverse races [6]. The majority of heart failure cases can be attributed to structural or physiological issues in the heart, leading to increased intracardiac pressure or reduced cardiovascular output based on the individual's state of rest or stress [7]. Consequently, heart failure is associated with a diminished quality of life and decreased engagement in physical and mental activities. Approximately 1-2% of the general population and 10% of older people in developed nations are affected by heart failure, with its prevalence expected to rise alongside an aging population. In hospital discharge, patients with heart failure (HF) experience a high 56.6% readmission rate. Addressing high frequency promptly is crucial to prevent future severe complications, with a current urgent focus on minimizing readmissions. Cardiovascular diseases like coronary artery disease (CAD), atrial fibrillation (AF), and vascular conditions remain the primary global cause of death [8]. The rising incidence of cardiovascular diseases is a pressing issue that requires immediate attention and innovative solutions. If lifestyle conditions rise and the amount of stress increases, the incidence of cardiovascular diseases is alarmingly increasing. To address this problem, an ensemble approach and survival rate prediction of patients have developed as a possible treatment.

Motivation

Based on the current studies [9]-[10] cardiovascular disease (CVD) is projected to cause the deaths of approximately 23 million individuals by 2030. There are many causes such as heart disease, irregular heartbeat, and heart attack, which are three different forms of cardiovascular disease [11]-[12]. Age, gender, BMI, height, waist circumference, and results from blood tests that check cholesterol levels, liver health, and kidney function are some of the factors used to analysed cardiovascular disease [13]–[14].

Several health issues could arise from the complex interactions across the risk factors. Traditional statistically effective methods are unable to investigate the complex relationship across risk-associated factors due to the large number of components present [15]–[16].

Over the last few decades, several researchers utilized the artificial intelligence (AI) techniques to investigate the new clinical data that help the physicians to analyze the signs and effects of many diseases and the prediction of survival of patients. The continuous efforts to collect all health examination records and consistent clinical data [17]–[18] focus to standardizing of clinical data before investigating previously unknown risk factors. A number of possible risk factors shows the associations regarding the development of diseases that indicate the basic causes of the disorders. Furthermore, a significant amount of medical data must be analyzed in order to create accurate prediction models for disease occurrences [19]–[20]. The risk assessment of CVD frameworks increasingly utilizes AI and large amounts of clinical data aggregation.

Problem Statement

Cardiovascular diseases (CVD) are the leading cause of death globally, with an estimated 17.9 million deaths each year. Early prediction of survival rates for patients with CVDs is crucial for providing timely and effective interventions to improve patient outcomes. Machine learning (ML) algorithms have shown promise in predicting survival rates based on patient data, including demographics, medical history, and clinical measurements. This study designs the ensemble models to detect and classify CVD and also predict the survival rates of patients with cardiovascular diseases. The models trained on a CVD dataset contain the risk factors of CVD. To measure the performance of the models using several evaluation parameters and also predict the survival rate of patients over a specified period. The proposed model helps to improved patient care that enable the healthcare experts to identify high-risk patients and early diagnosis to prevent adverse effect.

Limitation of Existing System

- The existing model worked on limited set of features, such as age, gender, and a few clinical elements.
- The quality and quantity of dataset is used to train the models could vary performance significantly that shows the biased or inaccurate predictions.
- Some existing models might have needed to be more complex, making them difficult to interpret and implement in clinical settings. This need for interpretability is a pressing issue in healthcare machine learning. We must address this complexity to avoid overfitting, where the model performed well on training data but failed to generalize to new data. Machine learning models worked effectively on present clinical procedure in order to be useful in the field of medicine. The medical professionals' resistance and accessibility issues result that many existing methods were not designed with the combination in perspective.

Contribution

Following is a list of this investigation's primary contribution as well as uniqueness:

- To extract seven unique features from the CVD dataset.
- Following the feature extraction, we meticulously normalized the data and divided the CVD dataset into training and testing sets using a 70:30 split, a crucial step in our thorough methodology. This aids in creating an ensemble model that includes a base ML classifier and a DT as a meta-classifier.
- Utilizes the Kaplan-Meier estimator to predict the survival rate of cardiovascular diseases for continuous variables in the dataset.
- Ultimately, we obtained results through various performance parameter analyses and predicted the survival rate of patients.

Paper organization

Section 2 presents the previous work in cardiovascular disease prediction based on ML and DL techniques. Section 3 presents the proposed methodology; Section 4 discusses the result analysis of the ensemble model. Section 4 addressed the conclusion and future scope of the study.

Related Work

Weng et al. suggested four different ML methods using health information from over 300,000 patients from the UK [21]. They found that the Neural Network (NN) method gave the most accurate predictions for CVD when analyzing a large amount of data. Dimopoulos A. C. et al. evaluated three traditional machine learning methods using ATTICA data with 2020 instances for a smaller dataset. Among these, KNN, RF, and DT were tested. Comparatively, RF showed the best outcome when using the Hellenic SCORE tool, calibrating the ESC Score [22].

Given the increasing utilization of ML in the medical domain, Mohan S. et al. suggested a hybrid HRFLM strategy to enhance the prediction accuracy score of the proposed approach, considering the growing use of ML algorithms [23].

In [24], they examined specific areas using different prediction models. They used LR to analyze 32 characteristics related to cardiovascular disease in over 210,000 high-risk patients in China.

Yang et al. [25] suggested the stacking ensemble framework for CVD prediction. They used information about air pollution and weather to understand how the average daily hospitalized percentage for cardiovascular diseases changes with the stacking ensemble model. They started with a basic level of five classifiers to build the stacking model.

Several investigations focus on creating new categories and approaches to enhance the current ones. For instance, studies by [26] suggested employing NLP to develop and evaluate an anxiety forecasting network. A neural network model with a predicted accuracy of 88.3% has been suggested, as reported [27], for predicting the development of diabetes. They indicated that the neural network's effectiveness was approximately 91% in training and 86% in evaluation compared to its previous effectiveness of 89% and 81%.

A technique for increasing the accuracy of cardiovascular disease prediction has been developed by Mohan et al. [28], employing machine learning algorithms for recognizing critical features. Regarding heart disease prediction, the suggested hybrid RF and linear model obtained the 87.9% accuracy. Au et al. [29] suggested the hybrid model to detect the CVD based on the logistic regression and obtained the accuracy score of 88.00%. Investigators have suggested a hybrid approach for forecasting cardiovascular disease [30]. The framework was employed by three ML techniques: DT, RF, and a combination. At 87.8%, the hybrid technique had the most excellent accuracy score.

As reported [31], author suggested the several ML models used to detect CVD and also suggested the ML-based ensemble model to predict and improve the efficiency of model.

We use information on coronary artery disease for our investigation. So, this might be our final investigation, with about 710,000 people and ten features in the data set. In addition, we employ multiple deep learning and machine learning techniques to determine which is most effective in identifying coronary artery disease.

Last two decade there is an advancements and large use of ML based models to predict the CVD and there are still several gaps that are unaddressed. The existing models depends on limited feature sets that fail to capture the complex features among risk factors such as lifestyle, biochemical markers, and physiological features. The usefulness of ensemble approaches in real-world scenarios is limited by the lack of thorough evaluations in several publications. The majority of approaches do not offer apparent findings for medical decision-making, making and ML models difficult to interpret in healthcare settings. Additionally, the datasets utilized in earlier studies are either limited or region-specific, and also limits their applicability to larger populations. Practical interactions between ML and survival analysis techniques, including Kaplan-Meier estimators, are not fully investigated. Addressing these gaps could lead to more robust, interpretable, and clinically actionable models for predicting survival rates in cardiovascular patients.

Proposed Methodology

This section presents the methodology of the various suggested ML models and the survival probability of patients. Figure 1 illustrates the complete architecture of the proposed model, which aims to predict cardiovascular diseases and forecast the survival rate of patients. The study utilizes a CVD dataset for model training, which initially requires preprocessing due to missing values. Strategies from feature extraction are used to extract essential features from data to predict the survival rates based on various parameters. This paper proposed ensemble model based on using several ML models as LR [32], RF [33], XGB [34], and NB [35] to predict the CVD. The study employs the Kaplan-Meier estimator to predict the survival rate of cardiovascular diseases for continuous variables present in the dataset.

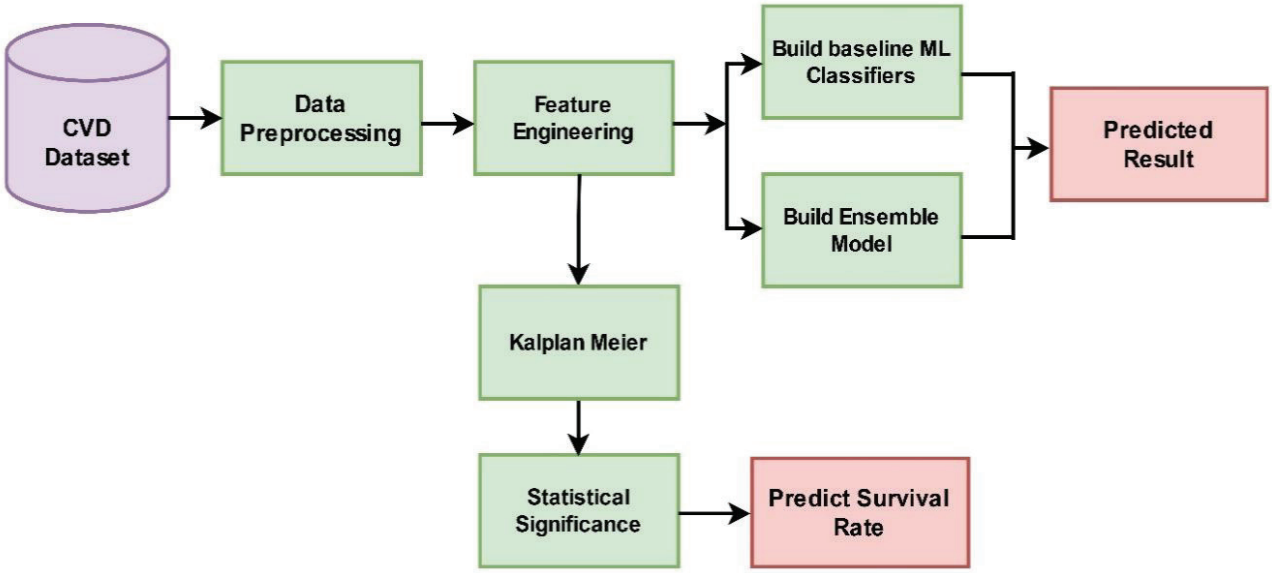


Figure 1. Architecture of proposed model.

Dataset description

The dataset contains comprehensive data on possible risk elements associated with CVD. It contains details on Age, anemia, Hypertension, Diabetes, Smoking, Gender, Blood Pressure, Glucose Levels, And Alcohol Consumption of over 72 thousand patients. The dataset is extracted from publicly available websites: https://www.kaggle.com/datasets/thedevastator/exploring-risk-factors-for-cardiovascular-diseases. The dataset also shows the relationship across the several risk factors and CVD that lead to improvement and understanding of the serious health issue and the design of better preventive measures [36].

Data preprocessing

This step starts with data cleaning, and missing values were handled using imputation techniques. The outliers were identified and treated to avoid skewing the results. The categorical parameters like patient demographics encoded into numerical scale using one-hot encoding strategy. The continuous variables were transform into uniform scale across each features.

Feature Engineering

This step selects the most relevant features to improve model performance. Initially, obtained the meaningful features and measure the risk scores parameters such as age, blood pressure, and cholesterol levels for diabetes or heart failure. The temporal features such as duration of the diagnosis were obtained and also capture nonlinear effects. The existing features were transformed to scaling for skewed variables such as age groups. The relation between features such as age and smoking history that retrieve the uncover complex relationships. Creating an ML model starts with collecting and analyzing raw input. In this study, it is used to determine how closely connected items are to one another.

Figure 2 shows the Univariate Analysis of Categorical parameters. The around 57.00% of the population have symptoms of Anemia that indicate the deficiency of red blood cells and around 43% do not exhibit Anemia symptoms. The 65% of the population has hypertension that they suffer from high blood pressure and approximately 35.00% have normal blood pressure. About 58.00% of the population is identified as diabetic and approximately 42.00% are non-diabetic. Around 65% of the population is male, and about 35% is female that shows the gender distribution in dataset. The approximately 68.00% of the population has smoke habits and about 32% are non-smokers persons. The 299 cases of heart failure, 96 individuals unfortunately did not survive the condition, whereas 203 cases have survived. The percentages that translate to 32.11% of the cases experiencing an unfavorable outcome. The majority precisely 67.89% of cases has positive result to survive the heart failure condition. The percentages shows the overview of the distribution of outcomes in the studied population that indicate the higher proportion of individuals who managed to overcome heart failure than those who did not.:

Figures 3 and 4 show the distribution of age with gender. The minimum age for both males and females is set at 40. On the other hand, the maximum age varies between genders, with males reaching up to 95, while females have a maximum age of 90. These age specifications provide insights into the age range covered in the research, indicating that individuals below 40 are omitted, and the maximum ages differ slightly for males and females in the studied cohort.

The impact of patient aging on the probability of survival is depicted in Figure 5. The 50–70 age category has

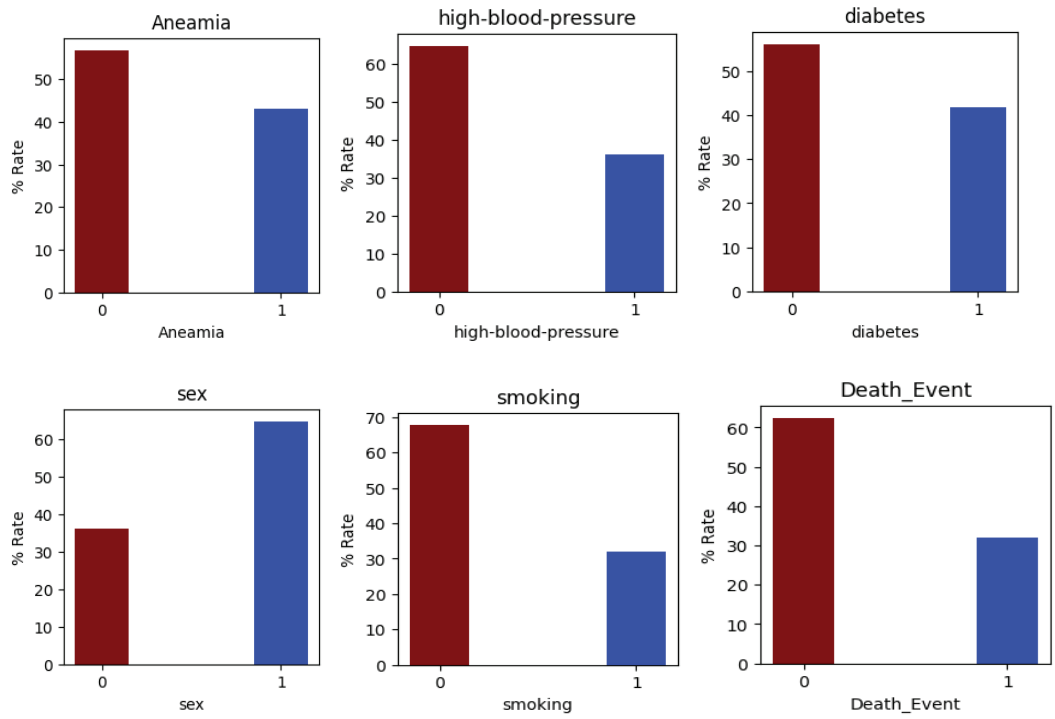


Figure 2. Univariate analysis of categorical variables.

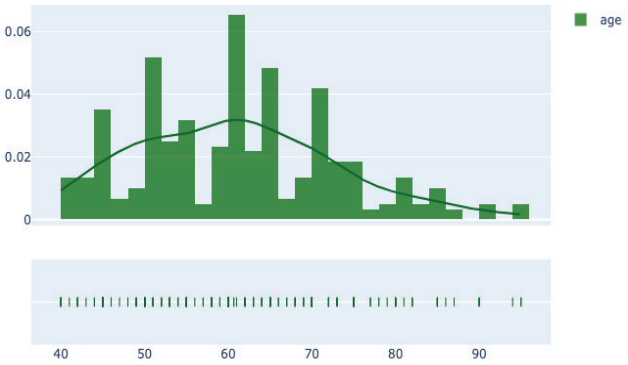


Figure 3. Age distribution.

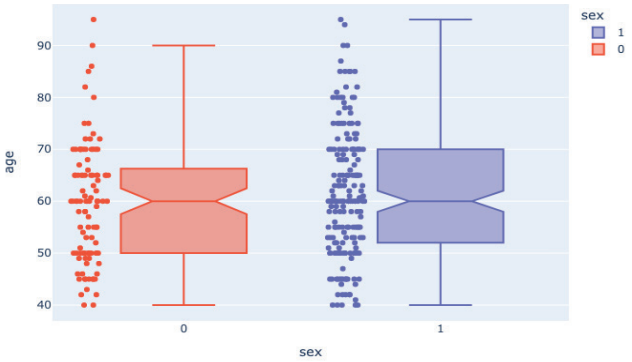


Figure 4. Distribution of Age w.r.t Gender.

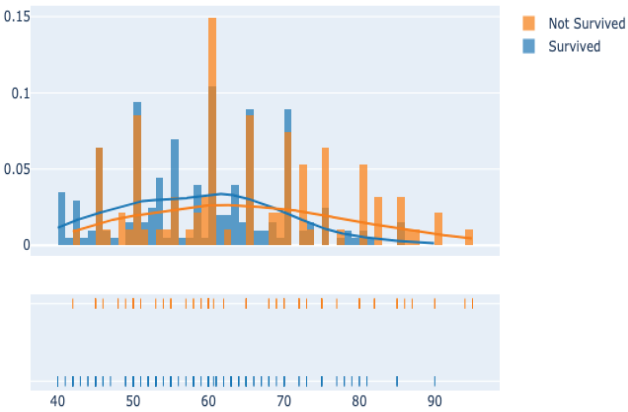


Figure 5. Age's Impact on Patients Survival Probability.

a significantly greater survival probability than the population. Every age category still carries some chance of not suffering a cardiovascular occurrence; the risk is most significant in the range of 60-65 age. Beyond the age of 80, the chances of survival sharply decline. The patterns suggest that age plays a significant role in survival outcomes after a heart failure event.

Figure 6 shows the distribution of various parameters and their survival rate. It is observed that survival outcomes in the studied population are based on gender. For the male population, 44.1% (132 individuals) have survived heart failure, while 20.7% (62 individuals) unfortunately did not survive. In the female population, 23.7% (71 individuals) survived the heart failure event, and 11.4% (34) did not.

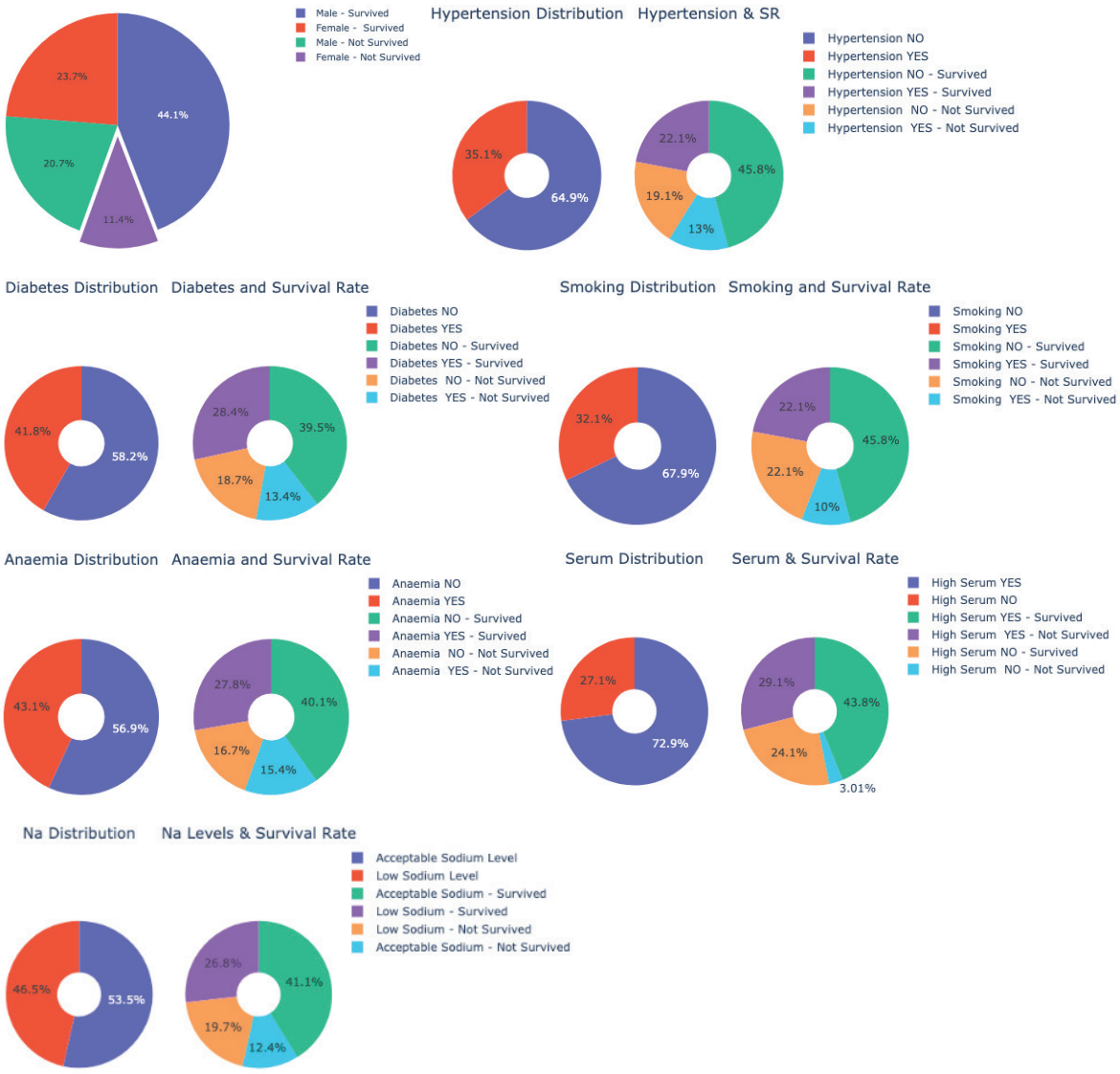


Figure 6. Distribution of parameters and their survival rate.

These percentages highlight the gender-specific variations in survival rates after experiencing heart failure, indicating a higher survival rate for males compared to females in the studied cohort. Around the 35.00% develop hypertension that excessive blood pressure. With this subgroup, 22% have sudden cardiac arrest events, and 13% unfortunately did not make it. Turning to the 65% of the population without hypertension, 45.8% have successfully survived heart failure, underscoring a higher survival rate compared to the hypertensive group, where 19% succumbed to the condition. Around 42% of individuals are identified as having diabetes, while approximately 58% do not have diabetes. 28.4% of people with diabetes have survived a cardiovascular attack, but unfortunately, 13.4% did not survive.

In contrast, among those without diabetes, a higher percentage, precisely 39.5%, have successfully survived after a heart attack, even though 18.7% have unfortunately given in to the medical condition. 32% of individuals have smoking habits, while around 68% do not smoke. Among those who smoke, 22.1% of people have prevented cardiac arrest, and unfortunately, 10% did not survive. On the other hand, among individuals without smoking habits, a higher percentage, precisely 45.8%, have successfully survived after cardiac arrest, while 22.1% have unfortunately given an approach to the illness. 43.1% of individuals exhibit symptoms of anemia, while around 56.9% do not show any signs of anemia. Among those with anemia, 27.8% survived after cardiac arrest, while 15.4%, unfortunately, failed to survive.

Conversely, among individuals without anemia, a higher percentage, precisely 40.1%, have successfully survived after cardiac arrest, while 16.7% have unfortunately succumbed to the condition. 46.5% of the entire population has lower blood sodium levels, while 53.5% have levels within the acceptable range. Among those with low

sodium levels, 26.8% after cardiac arrest, while 19.7% did not survive. In contrast, some patients whose sodium levels are below the appropriate limit have a higher percentage, precisely 41.1%, after cardiac arrest, with a smaller number, 12.4%, succumbing to the condition. These findings suggest a potential correlation between blood sodium levels and heart failure outcomes, with individuals within the acceptable range demonstrating a higher survival rate than those with lower sodium levels in the studied population.

Figure 7 shows information on various factors and laboratory test results related to heart failure outcomes in the studied population. First, it notes that people with cardiac failure who failed to survive generally have higher levels of CPK (creatine phosphokinase) enzyme. The violin histogram indicates outliers with high CPK levels in survival and death events. Additionally, it mentions that individuals who passed away with cardiac arrest often exhibited less than typical values for the ejection proportion, indicating inadequate pumping of blood from the heart. This shows outliers that observed in patients who survived heart failure. Serum creatinine levels 72.9% of cases reported elevated levels. The 43.8% are survived, and 29.1% succumbed to heart

Figure 7. Distribution of parameters and their effect on survival rate.

failure. Cases with serum levels in the normal range show a higher survival rate of 24.1%, contrasting with a lower percentage of 3.01% succumbing to the condition. However, 96 instances have passed away from cardiovascular disease. Of those instances, 59 people had sodium values below the normal range.

Table 1 presents the most correlated values associated with death in the dataset. The age shows the positive correlation of 0.253729 that indicate modest relationship across increase age and the likelihood of death. On the other hand, ejection fraction shows a negative correlation of -0.268603, suggesting that a lower ejection fraction, representing the proportion of blood that the coronary artery pumps continuously per contraction, is associated with a higher likelihood of death. Serum creatinine exhibits a positive correlation of 0.294278, indicating that higher serum

Table 1. Most correlated values of death

Parameter	Correlated Values
Age	0.253729
Ejection_fraction	-0.268603
Serum_creatinine	0.294278
Time	-0.526964

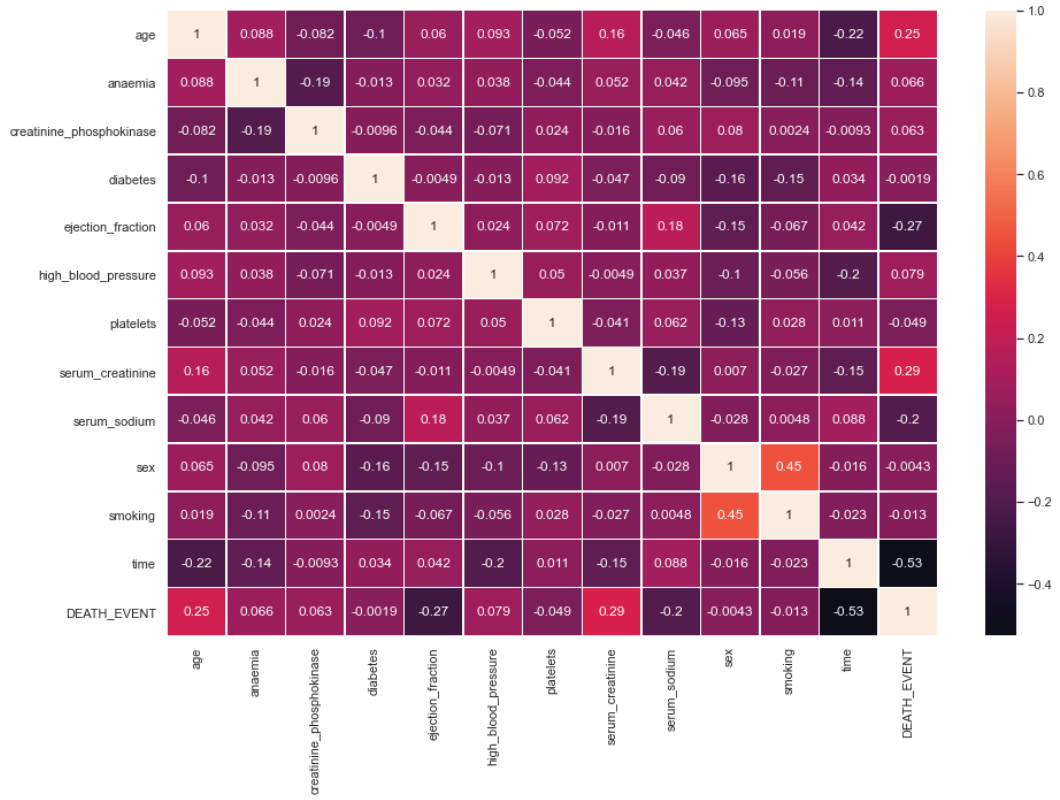


Figure 8. Heat map correlation of between the features.

creatinine levels, a marker of kidney function, are connected to a higher chance of passing away.

Build Baseline Classifiers

The CVD detection using ML based classifiers such as LR, RF, XG Boost, and NB serve as initial models to measure the performance of baseline models. The RF is a simple linear model measure the probabilities. The RF is a group of decision trees that collect correlations between the features. The XG Boost algorithm known for its accuracy and efficiency; and NB for feature independence. The classifiers provide a starting point for evaluating more sophisticated models, allow us to compare their performance against these more straightforward approaches, and measure the effectiveness of their ML classifiers in predicting cardiovascular diseases.

Ensemble Model

Combine the predictions of multiple individual classifiers to improve the overall prediction accuracy. We designed ensemble classifiers for CVD prediction using ML baseline classifiers such as LR, RF, XG Boost, and NB. Ensemble classifiers are often more robust and accurate than individual classifiers that effectively used the strengths of different models and mitigate their weaknesses, ultimately improving the performance of CVD prediction models [37]. The suggested ensemble models develops the outcomes using the weighted majority to combine the predicted results of many ML models. The most favorable results are shown following the fine-tuning of each categorization model. Equation 1 show to obtained the maximum votes.

$$C = argmax\ (X1\ (C_i^1), X2\ (C_i^2), \ldots\ldots Xn(C_i^n)) \quad (1)$$

Normalize and specify the loss function that provide an objective function to measure the performance of the ensemble model

$$C'(\oplus) = Loss\ (\oplus) + \alpha \cdot (\oplus) \quad (2)$$

Equation 2 shows the variables measured using the plus operator and derived from the supplied inputs. The normalization factor (Theta′) and the loss function used to calculate the model generalization.

Figure 9 shows the suggested ML-based ensemble model for predicting CVD. This model uses a voting approach by allowing each model to "vote" on the final prediction. Each model independently makes its prediction, and the most common prediction among the models is selected as the final. By combining these models, the ensemble can achieve higher accuracy and robustness in predicting CVD than any single model.

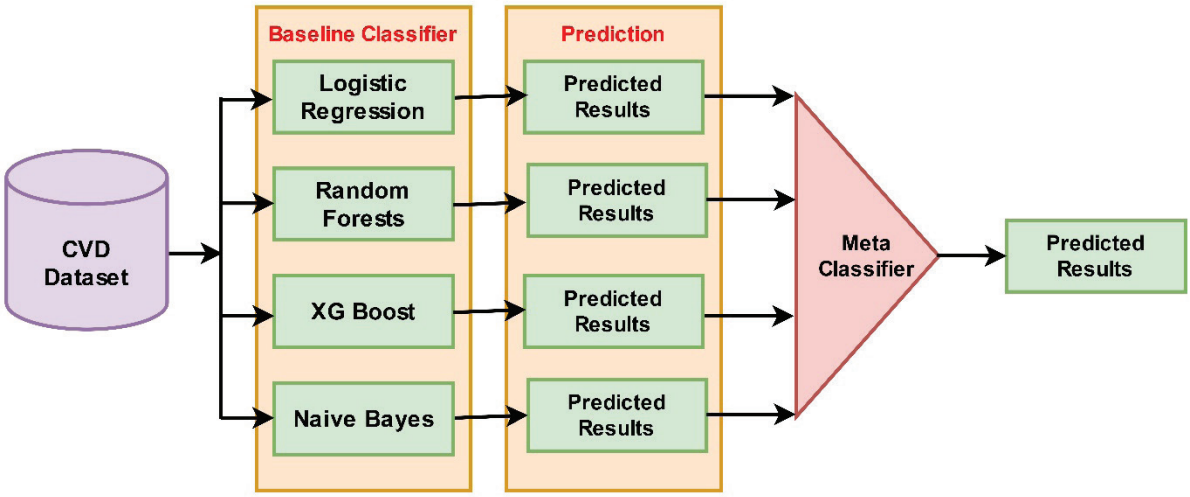


Figure 9. Ensemble Model.

Pseudo Code: Ensemble Model

Input: Training Dataset $CVD = \{(p1, q1), (p2, q2),...(pn, qn)\}$
Baseline Model $M = (LR, RF, XGB, NB)$
Meta Classifier DT
Output: Learn Ensemble Model EM
Start
Step-1: Learn the Baseline classifiers M on CVD
for $i = 1$ *to* n *do*
  $Bi = Mi(CVD)$
end for
Step-2: Construct new CVD Dataset for prediction CVD'
for $j=1$ *to* n' *do*
  for $i = 1$ *to* n *do*
Used Bi to classify training parameters pj
$xij = Bi(pj)$
end for
$CVD = (xj, qj)$, *where* $xj = \{xij, x2j,.... xnj\}$
end for
Step-3: Learn Meta Classifier DT
$EM = DT(CVD)$
Return EM
END

Table 2. Hyperparameter settings of ML models in ensemble

Model	Parameters	Values
LR	C	0.1
	penalty	L1
	solver	liblinear
	max_iter	100
RF	n_estimators	100
	max_depth	10
	min_samples_split	2
	min_samples_leaf	2
	criterion	Gini
	max_features	Sqrt
XGBoost	eta	0.01
	max_depth	3
	n_estimators	100
	subsample	0.7
	colsample_bytree	0.7
	alpha	0
	lambda	1
Naive Bayes	var_smoothing	1e-9
	alpha	0.5

Hyperparameter settings

Hyperparameter optimization plays a crucial role in improving the performance of ensemble models for cardiovascular disease detection by fine-tuning the parameters of individual ML models included in the ensemble. Table 3 shows the hyperparameter setting of ML models used to build the ensemble model

RESULTS AND DISCUSSION

The following section examines the baseline and proposed ensemble classifier performance and briefly overviews the outcomes. The main goal of this investigation is to investigate the effectiveness of suggested ML algorithms for identifying cardiovascular illness. We used the CVD dataset in the tests conducted for this study. We divided the CVD dataset into 70% training and 30% testing. We

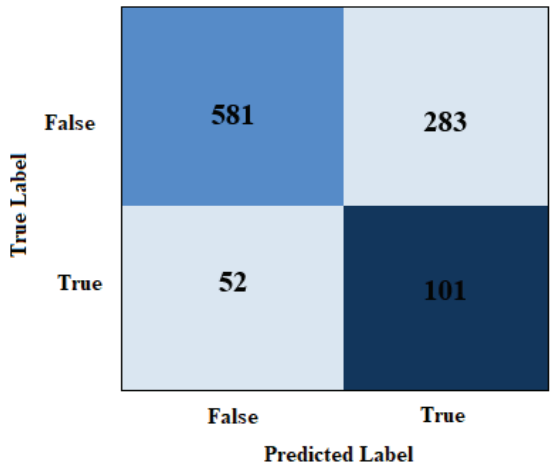
Table 3. Comparative result analysis of baseline model with ensemble model

Classifiers	Accuracy	Precision	Recall	F1-Score
Logistic Regression	67.00	82.00	67.00	72.00
Random Forests	65.00	81.00	65.00	70.00
XG Boost	70.00	78.00	70.00	73.00
Naive Bayes	70.00	81.00	70.00	74.00
Ensemble Model	82.00	85.00	80.00	82.00

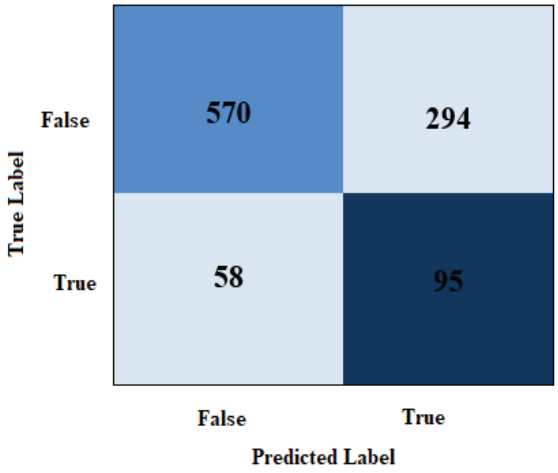
simulated the ensemble model using Google Colab with Python 3, running on an Intel Core i5 GPU @4.20 GHz, 16 GB RAM, and 4 GeForce RTX graphics cards. Several evaluation parameters are shown in equation 3 to 6.

$$Accuracy = \frac{T\_P + T\_N}{T\_P + T\_N + F\_P + F\_N} \quad (3)$$

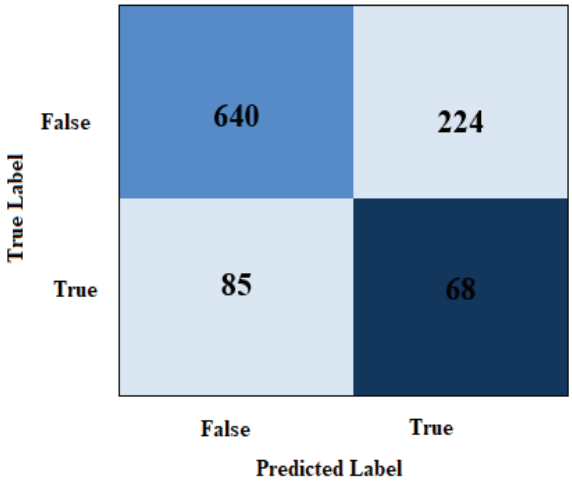
$$Precision = \frac{T\_P}{T\_P + F\_P} \quad (4)$$



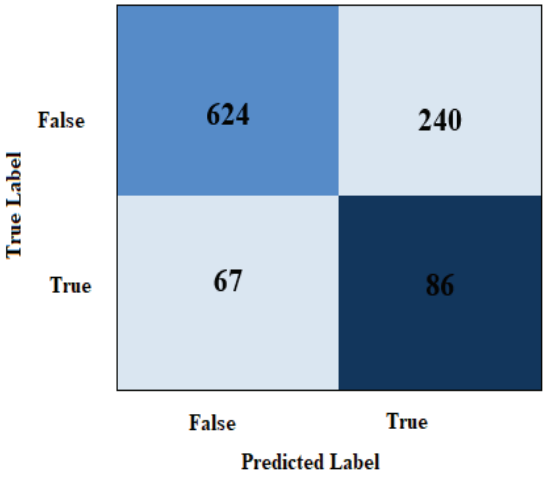
(1) K Nearest Neighbors



(2) Random Forests



(3) XG Boost



(4) Naive Bayes

Figure 9. Confusion Matrix of Baseline Classifiers

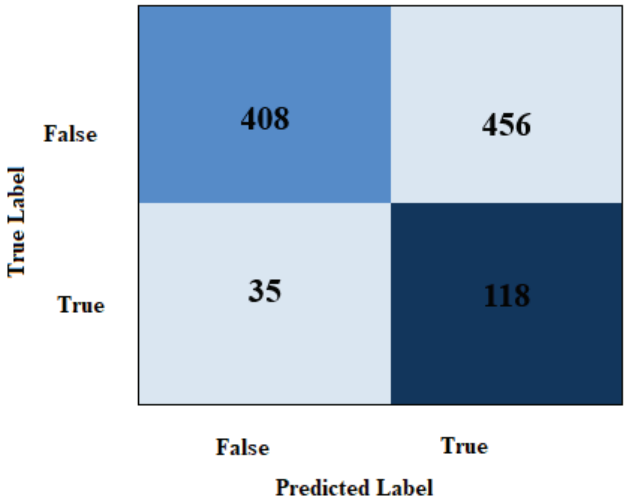


Figure 10. Confusion Matrix of Ensemble Model.

$$Recall = \frac{T\_P}{T\_P + F\_N} \quad (5)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

Table 2 presents a comparative analysis of baseline models, including Logistic Regression, Random Forests, XG Boost, Naive Bayes, and an Ensemble Model, based on their performance metrics. The finding shows the ensemble model performed well ad compared to individual models and obtained the accuracy of 82.00%, precision of 85.00%, Recall of 80.00%, and F1-Score of 82.00%. The suggested model is more effective due to ability to combine the strengths of different ML classifiers.

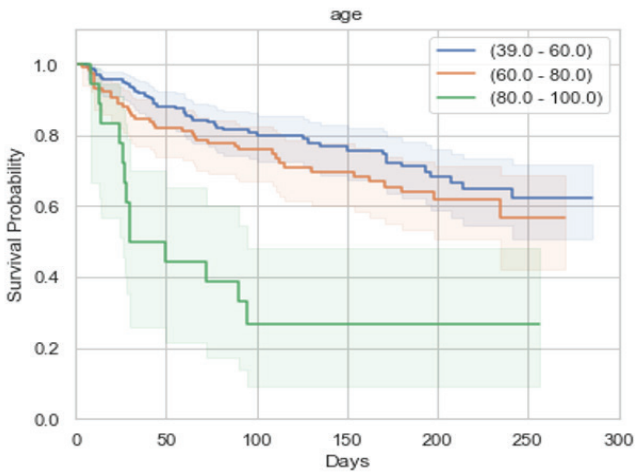
Table 4 shows different machine learning models' training and test recall scores for cardiovascular disease prediction. Recall, sometimes called sensitivity, indicates the percentage of actual positive instances the model accurately recognized. A higher recall indicates better performance in identifying positive cases. The ensemble model obtained the recall scores of training and testing of 86.90% and 82.53% that shows the superior ability to identify positive cases.

Table 4. Training and test recall of ML models

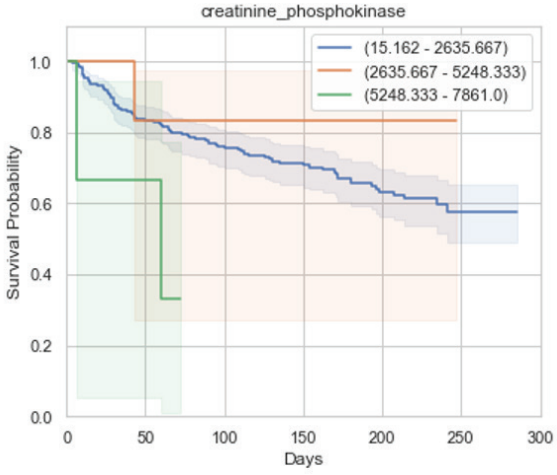
Classifiers	Train Recall	Test Recall
Logistic Regression	69.6352	67.0130
Random Forests	69.3282	63.0915
XG Boost	82.3399	44.4444
Naive Bayes	52.1910	56.2091
Ensemble Model	**86.9057**	**82.5363**

Survival Prediction Using Kaplan Meier

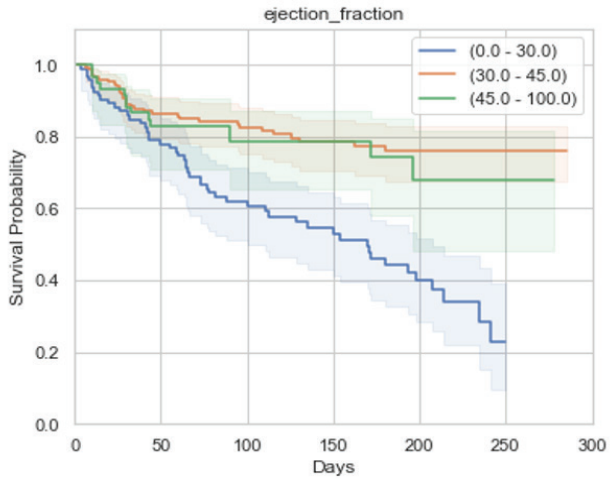
This study uses the Kaplan-Meier estimator to predict the survival rate of cardiovascular diseases for continuous variables, such as Age, Creatinine Phosphokinase, Ejection Fraction, Platelets, Serum Creatinine, Serum Sodium, and Time. The first step is to discretize each continuous variable into intervals. The proportion of observations that survive beyond each time interval is calculated for each combination of intervals. This is done similarly to the standard Kaplan-Meier estimator but using the intervals for the continuous variables. Figure 11 is a powerful tool that visually represents the estimated survival probabilities against each interval of parameters, allowing us to quickly grasp the survival function over the range of the continuous variables. Another name for the Kaplan-Meier estimate is "product limit estimate." This mesure the possibility of an event occurred at a specific moment [38,39]. Combine this sequential probability with all previously estimated possibilities to obtain the highest possible estimation.



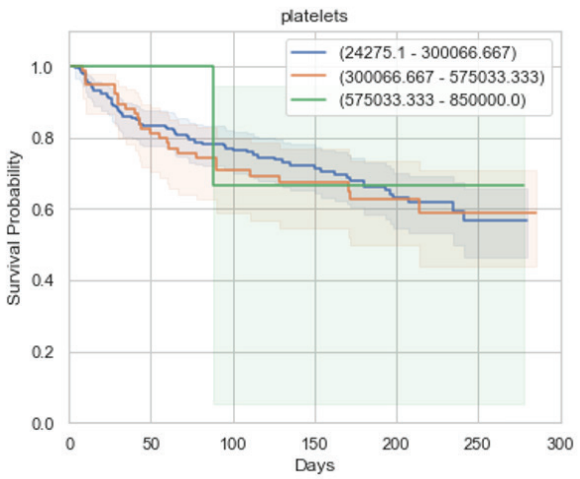
(1) Age



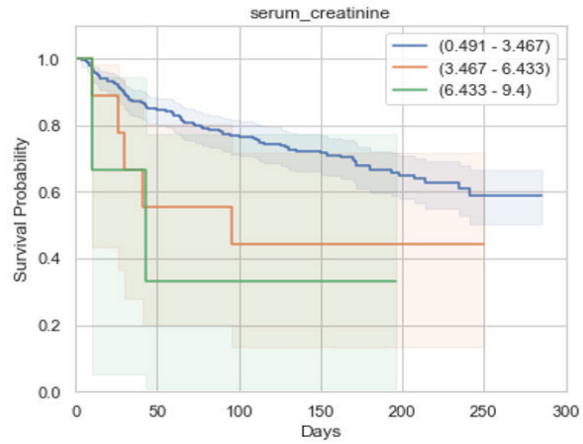
(2) Creatinine Phosphokinase



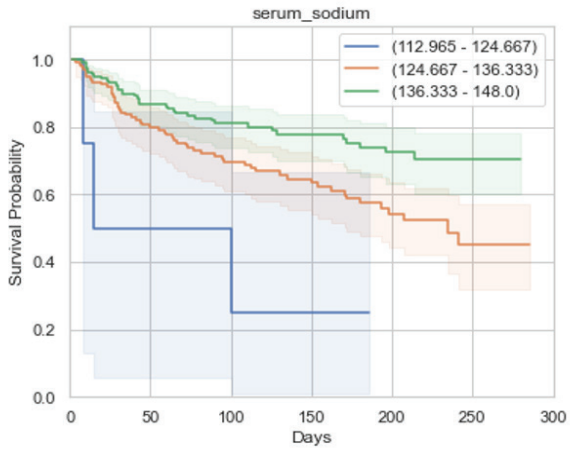
(3) Ejection Fraction



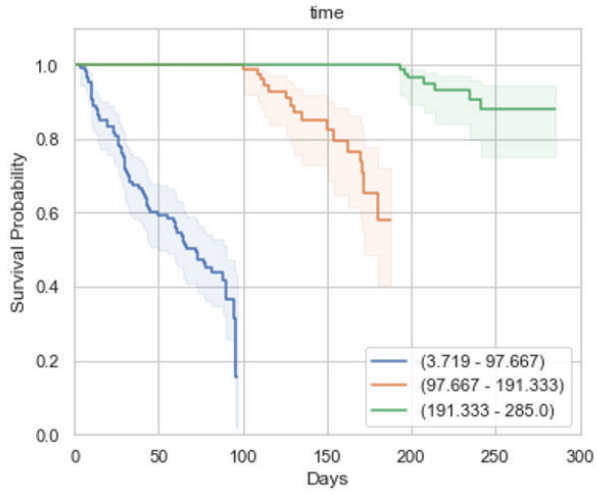
(4) Platelets



(5) Serum Creatinine



(6) Serum Sodium



(7) Time

Figure 11. Survival prediction analysis

$$Kaplan\ Meier\ Estimate = \frac{No\ of\ Patients\ living\ at\ the\ Start - No\ of\ Patients\ died}{No.of\ Patients\ Living\ at\ the\ start} \quad (7)$$

The chance of surviving risk for every period is determined by dividing the total number of vulnerable people by the total number of patients who survive. Individuals who have passed away, stopped participating, or moved away are not included in the denominator and are not regarded as "under threat.

p Values and Statistical Significance of Parameters

The summary's p-values shows risk exhibit considerable significance, with p-values below 0.0005 that shows statistical significance at a confidence level of 99.9995% or higher. The attributes demonstrate a strong correlation with the occurrence of the death event. Conversely, Smoking, Sex, Platelets, and Diabetes yield notably high p-values, making it uncertain whether they hold statistical significance. Consequently, their impact on the hazard rate may be disregarded in the analysis. This observation is supported by their substantial standard errors and the consequent broad confidence intervals [40].

Anemia is a binary variable represented by 1 or 0, indicating whether the subject is anemic. The coefficient of 0.481 is interpreted as:

$$\frac{HazardRateforPatientswithAnemia}{HazardRateforPatientswithoutAnemia}$$

$$\frac{e^{0.481\times1}}{e^{0.481\times0}} = e^{0.481\times(1-0)} = e^{0.481} = 1.618$$

The hazard ratio for anemia is 1.618. This indicate that the patient has anemia, the risk of death increases by 61.8%.

High BP represented by 1 and 0 that shows that subject has high blood pressure (hypertension) or not. The coefficient of 0.406 is interpreted as:

$$\frac{HazardRateforPatientswithHighBP}{HazardRateforPatientswithoutHighBP}$$

$$\frac{e^{0.406\times1}}{e^{0.406\times0}} = e^{0.406\times(1-0)} = e^{0.406} = 1.50$$

The HR for high BP is 1.5. This shows that the patient has hypertension. The risk of death increases by 50%.

Hazard ratio (HR)

The HR shows the impact of a covariate on the hazard rate. An HR of 1 suggests no effect that indicate the covariate does not influence the hazard rate. An HR more significant than 1 increases the hazard rate as the covariate value increased that indicate the higher risk of the event occurring [40]. The HR less than 1 shows to decrease in the hazard rate as the covariate value increases that suggest lower risk of the event.

Figure 12 represents the coefficients (i.e., log hazard ratios) for predicting the survival rate of cardiovascular diseases based on various factors. Each factor, such as anemia, high blood pressure, serum creatinine, diabetes, smoking, age, creatinine phosphokinase, platelets, serum sodium, ejection fraction, and sex, is listed along the vertical axis. The horizontal bars extending from the central axis represent the magnitude and direction of the coefficients. A bar extending to the right indicates a positive effect on the hazard ratio, meaning an increased risk of cardiovascular disease survival, while a bar extending to the left indicates a negative impact, implying a decreased risk.

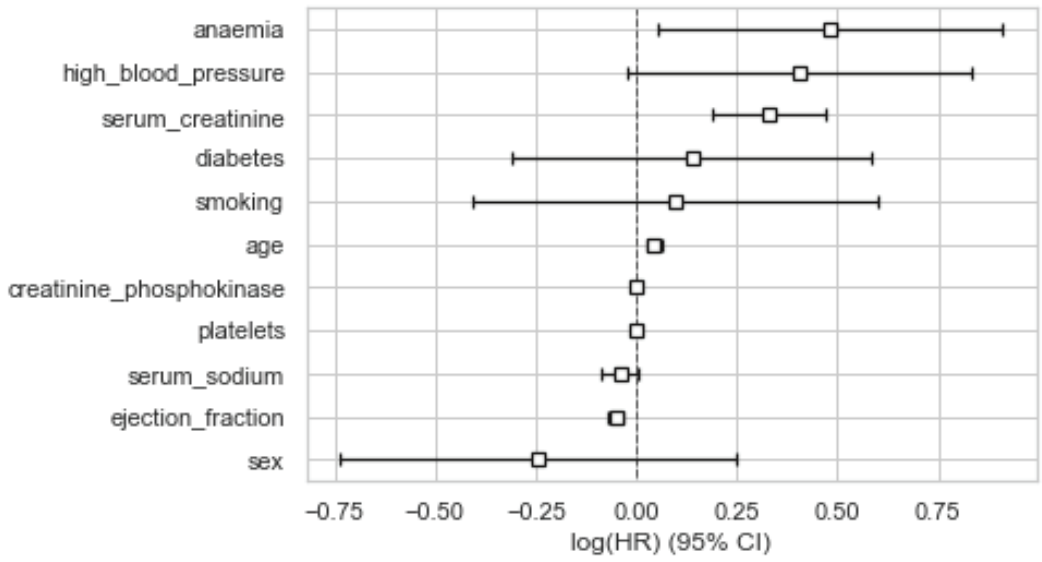


Figure 12. Graphical Visualization of the Coefficients (i.e. log hazard ratios).

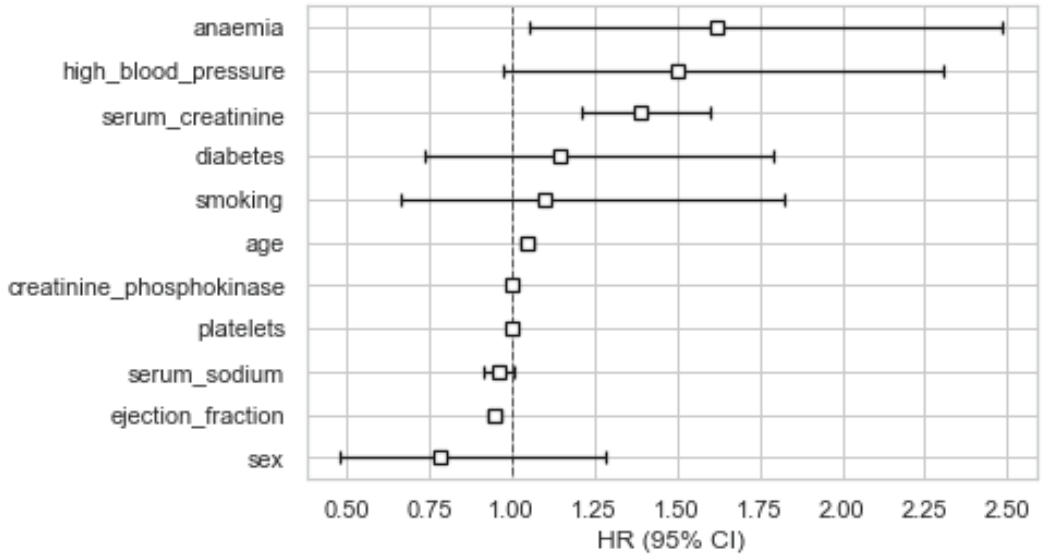


Figure 13. Visual representation of the hazard ratios.

The impact of covariates on survival outcomes has been understood through the HR which measures the change in hazard rate with a one-unit change in the covariate. For Ejection Fraction, a one-unit change results in a 5.2% increase in survival time, as indicated by an HR of 0.95. Conversely, a one-unit rise in Serum Creatinine causes a 28.1% decrease in survival time, given its HR of 1.392 [41]. Age shows a 3.9% decrease in survival time per unit increase, with an HR of 1.046. However, Creatinine Phosphokinase and Platelets have HRs of 1, suggesting no effect on the probability of the death occurrences.

Figures 12 and 13 provide the graphical layout of coefficients, log hazard ratios, and hazard ratios, respectively, encompassing their sizes and standardized errors. Anemia, High Blood Pressure, Serum Creatinine, Age, and the amount of Ejection are all within the 95% confidence interval of influencing the death incident.

Figure 14 shows that with age, the survival probability decreases for any complication arising from a heart failure condition. Increasing age has a significant impact on the likelihood of surviving. Figure 15 shows that the volume of blood pumped out of the heart increases with increasing ejection fraction percentage. Consequently, the probability of survival also increases during any phase of heart failure. Increasing EF levels has a beneficial significant impact on the likelihood of survival.

Figure 16 shows that with increasing serum creatinine levels in the blood, the survival probability decreases for

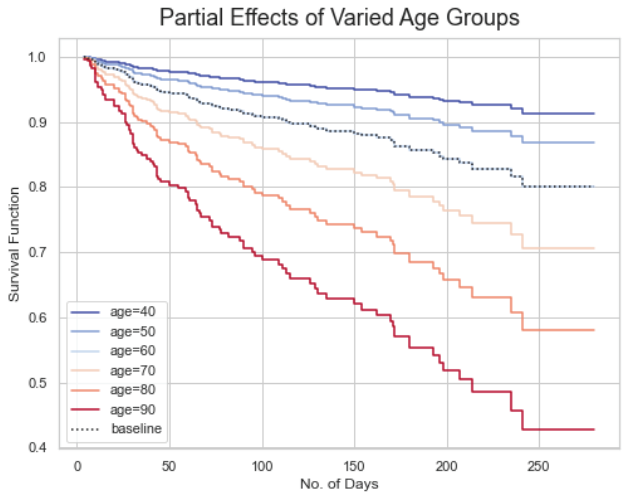


Figure 14. Partial effect of varied age.

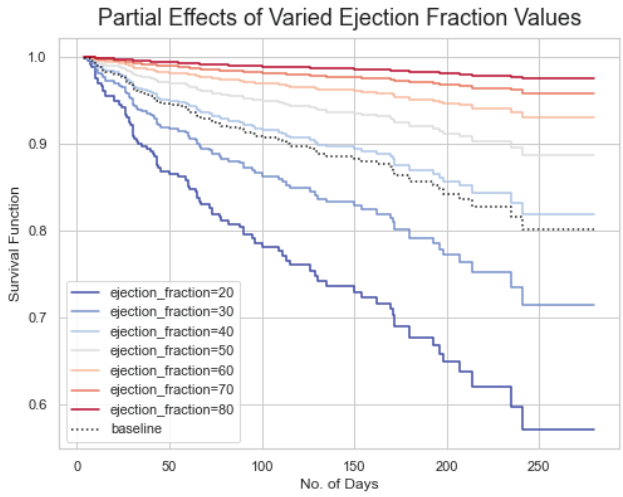


Figure 15. partial effect of varied ejection fraction.

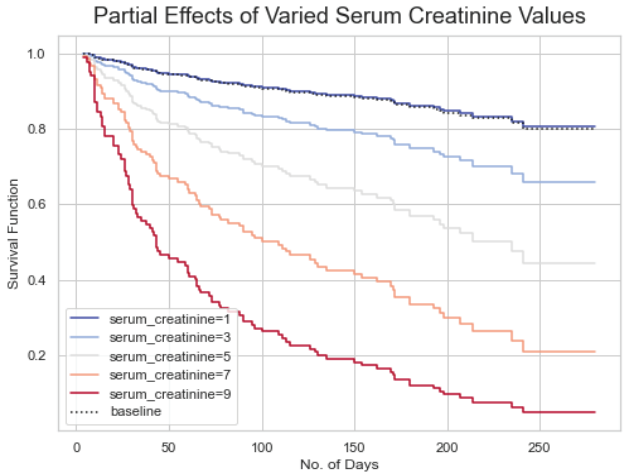


Figure 16. Partial effect of varied serum creatinine vs non-anaemia.

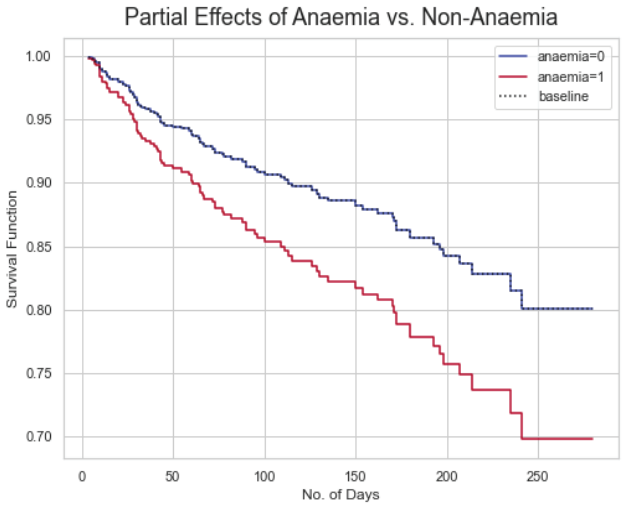


Figure 17. Partial effect of partial effect of anaemia

any complication arising out of a heart failure condition. Increasing creatinine levels have a significant impact on the likelihood of survival. Figure 17 shows that anaemic patients are more likely to encounter a hazard due to heart failure condition—survival Probabilities over 280 days.

Table 5 analyzes the survival probabilities of the patient in the test cohort over 280 days. The analysis assumes that the subjects have just entered the study without considering how long they have lived. It is observed that initially, each subject in the test cohort has high survival chances, hovering around the 98-99% mark. For Patient -298 and Patient -179, the survival probabilities remain consistent throughout the period, at 88% and 85.5%, respectively, by the 280th day. For Patient-42 and Patient -193, the survival probabilities hover around 61% and 34%, respectively, at the end of 280 days. However, for patient -5, the chances of survival show a decreasing trend. By day 15, the survival chance is approximately 75%; by day 38, it hovers around the 50% mark; by the end of 180 days, it falls below 10%.

Table 6 shows the comparative perfromance analysis of the proposed ensemble models and existing model to detect and classify the CVD. The proposed ensemble model obtained the accuracy of 82.53% that performed well as compared to existing individual ML models that obtained the accuracy score of DT has 81.23%, K-means of 78.00%, ANN of 82.10%, and Stacking Models of 82.35%. The XGBoost obtained the 82.44%, and SVM achieved the highest accuracy at 83.12%. This comparison demonstrates that the ensemble approach offers a balanced trade-off between simplicity and performance, positioning it as a competitive alternative among state-of-the-art techniques for this task.

The final results suggest whether the various CVD parameters are used to forecast a patient's survival rate suffering heart disease. They also indicate that forecasts based

Table 5. Survival probabilities of patients last 10 days

Days	Patient - 42	Patient - 298	Patient - 5	Patient - 193	Patient - 179
246.0	0.616569	0.880709	0.033281	0.346240	0.855772
247.0	0.616569	0.880709	0.033281	0.346240	0.855772
250.0	0.616569	0.880709	0.033281	0.346240	0.855772
256.0	0.616569	0.880709	0.033281	0.346240	0.855772
257.0	0.616569	0.880709	0.033281	0.346240	0.855772
258.0	0.616569	0.880709	0.033281	0.346240	0.855772
270.0	0.616569	0.880709	0.033281	0.346240	0.855772
271.0	0.616569	0.880709	0.033281	0.346240	0.855772
278.0	0.616569	0.880709	0.033281	0.346240	0.855772
280.0	0.616569	0.880709	0.033281	0.346240	0.855772

Table 6. Comparative analysis of proposed method with existing methods

Author	Methods	Accuracy
Bhatt, C.M. et. al. (2023) [42]	DT	81.23%
Subramani S. et. al. (2023) [43]	Stacking Model	82.35%
Baghdadi, N.A. et. al. (2023) [44]	XGBoost	82.44%
Ahmad AA, et. al. (2023) [45]	SVM	83.12%
Khdair, H. et. al. (2021) [46]	K-means	78.00%
M. S. Gangadhar et. al. (2023) [47]	ANN	82.10%
Proposed	**Ensemble**	**82.53%**

simply on both variables may be more reliable than those based on the entire dataset. This is especially required for healthcare facilities environments: physicians might still be capable of forecasting the survival of patients using Age, smoking, serum creatinine amounts, and ejection fraction alone, even if the individual technological medical care contained a lot of not present clinical information and test examination from laboratories. However, invetigation that need for further study to design the strong models to daignosis the CVD ealier and implementable in healthcare settings.

Further investigation offered several interesting results that were not discovered in the researchers' earlier dataset analysis [41]. Ahmad et al. found that anemia, high blood pressure, ejection fraction, age, and serum creatinine a sign of kidney disease were actually the most common characteristics. The characteristics are important since are highly predictive of patient survival rates.

CONCLUSION

This study validated the importance of relevant feature extraction with machine learning by demonstrating that conventional statistical analysis identified cardiovascular disease factors as the most significant features. Furthermore, the proposed method showed that machine learning was applied successfully to the binary categorization of electronic healthcare of individuals with coronary artery disease disorders related to the heart system. This study demonstrates the effectiveness of machine learning techniques in predicting cardiovascular diseases and forecasting survival rates of patients. The ensemble model performed well as compared to baseline models that indicate its superiority in disease prediction. This study also identifies age, serum creatinine, and ejection fraction as significant factors affecting the death event, while smoking, sex, platelets, and diabetes may not hold statistical significance. The existing studies that focus on individual models or limited feature sets. The proposed model combined the diverse risk factors and shows superior accuracy compared to existing models. This work is unusual because it provides interpretable insights into important risk factors including age, serum creatinine, and ejection fraction by bridging the gap between ML and medical survival analysis. This method offers useful applications for early treatment in cardiovascular care while also advancing predictive modeling. In order to enhance preventative and therapeutic approaches, future studies might investigate the connections among risk variables and CVD prognosis.

ACKNOWLEDGMENT

We acknowledge the support received from Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India. We are deeply grateful to the management and Principal Dr. U. P. Waghe for their invaluable support and encouragement.

AUTHOR CONTRUBITIONS

Bharati Karare handled conceptualization and data collection. Anushree Anand Pande was responsible for data validation, software tools, and data collection. Prof. Pratibha Waghale focused on data analysis, identifying previous research gaps, and methodology. Prof. Amruta Tapas Paul contributed to methodology and software. Prof. Chanchla Tripathi took care of validation and data visualization. Dr. Lalit Damahe was in charge of writing the original draft and overall writing. Dr. Prashant Bhokardankar was involved in review and editing.

DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE

Artificial intelligence was not used in the preparation of the article.

REFERENCES

[1] World Health Organization. The top 10 causes of death. Geneva: WHO; 2020 Available at: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death Accessed on Dec 30, 2020.

[2] Fryar CD, Chen TC, Li X. Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999–2010. NCHS Data Brief 2012;103:1–8.

[3] Murphy SP, Ibrahim NE, Januzzi JL Jr. Heart failure with reduced ejection fraction: A review. JAMA 2020;324:488–504. [CrossRef]

[4] Mayo Clinic. Cardiovascular diseases. Rochester (MN): Mayo Clinic; 2020 Available at: https://www.mayoclinic.org/medical-professionals/cardiovascular-diseases Accessed on Dec 29, 2020.

[5] Centers for Disease Control and Prevention. Underlying cause of death 1999–2019 [Internet]. Atlanta (GA): CDC; 2020. Available at: https://wonder.cdc.gov/wonder/help/ucd.html Accessed on Dec 28, 2020.

[6] Bhatt CM, Patel P, Ghetia T, Mazzeo PL. Effective heart disease prediction using machine learning techniques. Algorithms (Basel) 2023;16:88. [CrossRef]

[7] Mahmud I, Kabir MM, Mridha MF, Alfarhood S, Safran M, Che D. Cardiac failure forecasting based on clinical data using a lightweight machine learning metamodel. Diagnostics (Basel) 2023;13:2540. [CrossRef]

[8] Saeedbakhsh S, Sattari M, Mohammadi M, Najafian J, Mohammadi F. Diagnosis of coronary artery disease based on machine learning algorithms: Support vector machine, artificial neural network, and random forest. Adv Biomed Res 2023;12:51. [CrossRef]

[9] Alizadehsani R, Khosravi A, Roshanzamir M, Abdar M, Sarrafzadegan N, Shafie D, et al. Coronary artery disease detection using artificial intelligence techniques: A survey of trends, geographical differences and diagnostic features 1991–2020. Comput Biol Med 2021;128:104095. [CrossRef]

[10] Joloudari JH, Joloudari EH, Saadatfar H, GhasemiGol M, Razavi SM, Mosavi A, et al. Coronary artery disease diagnosis: Ranking the significant features using a random trees model. Int J Environ Res Public Health 2020;17:731. [CrossRef]

[11] Pal M, Parija S, Panda G, Dhama K, Mohapatra RK. Risk prediction of cardiovascular disease using machine learning classifiers. Open Med (Wars) 2022;17:1100–1113. [CrossRef]

[12] Subramani S, Varshney N, Anand MV, Soudagar MEM, Al-Keridis LA, Upadhyay TK, et al. Cardiovascular diseases prediction by machine learning incorporation with deep learning. Front Med (Lausanne) 2023;10:1150933. [CrossRef]

[13] Ishaq A, et al. Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. IEEE Access 2021;9:39707–39716. [CrossRef]

[14] Moreno-Sánchez PA. Improvement of a prediction model for heart failure survival through explainable artificial intelligence. Front Cardiovasc Med 2023;10:1219586. [CrossRef]

[15] Baby PS, Vital TP. Statistical analysis and predicting kidney diseases using machine learning algorithms. Int J Eng Res Technol 2015;4:206–210. [CrossRef]

[16] García-Ordás MT, Bayón-Gutiérrez M, Benavides C, et al. Heart disease risk prediction using deep learning techniques with feature augmentation. Multimed Tools Appl 2023;82:31759–31773. [CrossRef]

[17] Shameer K, Johnson KW, Yahi A, Miotto R, Li LI, Ricks D, et al. Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: A case study using Mount Sinai heart failure cohort. Pac Symp Biocomput 2017;22:276–287. [CrossRef]

[18] Kaddour A. Implementation of an incremental deep learning model for survival prediction of cardiovascular patients. IAES Int J Artif Intell 2021;10:101–109. [CrossRef]

[19] Jackins V, Vimal S, Kaliappan M, Lee MY. AI-based smart prediction of clinical disease using random forest classifier and naive Bayes. J Supercomput 2021;77:5198–5219. [CrossRef]

[20] Ahsan MM, Siddique Z. Machine learning-based heart disease diagnosis: A systematic literature review. Artif Intell Med 2022;128:102289. [CrossRef]

[21] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One 2017;12:e0174944. [CrossRef]

[22] Dimopoulos AC, Nikolaidou M, Caballero FF, Engchuan W, Sanchez-Niubo A, Arndt H, et al. Machine learning methodologies versus cardiovascular risk scores in predicting disease risk. BMC Med Res Methodol 2018;18:179. [CrossRef]

[23] Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning techniques. IEEE Access 2019;7:81542–81554. [CrossRef]

[24] Yang L, Wu H, Jin X, Xu X, Yu W, Yan J. Study of cardiovascular disease prediction model based on random forest in eastern China. Sci Rep 2020;10:5245. [CrossRef]

[25] Hu Z, Qiu H, Su Z, Shen M, Chen Z. A stacking ensemble model to predict daily number of hospital admissions for cardiovascular diseases. IEEE Access 2020;8:138719–138729. [CrossRef]
[26] Ahmed U, Mukhiya SK, Srivastava G, Lamo Y, Lin JCW. Attention-based deep entropy active learning using lexical algorithm for mental health treatment. Front Psychol 2021;12:642347. [CrossRef]
[27] Jerjawi NSE, Naser SSA. Diabetes prediction using artificial neural network. Int J Adv Sci Technol 2018;121.
[28] Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning techniques. IEEE Access 2019;7:81542–81554. [CrossRef]
[29] Au H, Li JP, Memon MH, Nazir S, Sun R. A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mob Inf Syst 2018;2018:3860146. [CrossRef]
[30] Renugadevi G, Asha Priya G, Sankari BD, Gowthamani R. Predicting heart disease using hybrid machine learning model. J Phys Conf Ser 2021;1916:012208. [CrossRef]
[31] Sathwika GJ, Bhattacharya A. Prediction of cardiovascular disease using ensemble learning algorithms. In: Proc 5th Joint Int Conf Data Sci Manag Data (ACM IKDD CODS, COMAD); 2022 Jan; Bangalore, India. p.292–293. [CrossRef]
[32] Zheng H, Sherazi SWA, Lee JY. A stacking ensemble prediction model for the occurrences of major adverse cardiovascular events in patients with acute coronary syndrome on imbalanced data. IEEE Access 2021;9:113692–113704. [CrossRef]
[33] Saw M, Saxena T, Kaithwas S, Yadav R, Lal N. Estimation of prediction for getting heart disease using logistic regression model of machine learning. In: Proc Int Conf Comput Commun Inform (ICCCI); 2020; Coimbatore, India. p.1–6. [CrossRef]
[34] Lutimath NM, Sharma N, Byregowda BK. Prediction of heart disease using random forest. In: Proc Emerg Trends Ind 4.0 (ETI 4.0); 2021; Raigarh, India. p.1–4. [CrossRef]
[35] Jain V, Agrawal M. Heart failure prediction using XGB classifier, logistic regression and support vector classifier. In: Proc Int Conf Adv Comput Comput Technol (InCACCT); 2023; Gharuan, India. p.1–5. [CrossRef]
[36] TheDevastator. Exploring risk factors for cardiovascular disease. Kaggle. Available at: https://www.kaggle.com/datasets/thedevastator/exploring-risk-factors-for-cardiovascular-diseas Accessed on Apr 12, 2023.
[37] Alqahtani A, Alsubai S, Sha M, Vilcekova L, Javed T. Cardiovascular disease detection using ensemble learning. Comput Intell Neurosci 2022;2022:5267498. [CrossRef]
[38] Goel MK, Khanna P, Kishore J. Understanding survival analysis: Kaplan–Meier estimate. Int J Ayurveda Res 2010;1:274–278. [CrossRef]
[39] Rich JT, Neely JG, Paniello RC, Voelker CC, Nussenbaum B, Wang EW. A practical guide to understanding Kaplan–Meier curves. Otolaryngol Head Neck Surg 2010;143:331–336. [CrossRef]
[40] Databricks. Cox proportional hazards notebook Databricks. Available at: https://www.databricks.com/notebooks/telco-accel/03_cox_proportional_hazards.html Accessed on Apr 12, 2023.
[41] Ahmad T, Munir A, Bhatti SH, Aftab M, Raza MA. Survival analysis of heart failure patients: A case study. PLoS One 2017;12:e0181001. [CrossRef]
[42] Bhatt CM, Patel P, Ghetia T, Mazzeo PL. Effective heart disease prediction using machine learning techniques. Algorithms (Basel) 2023;16:88. [CrossRef]
[43] Subramani S, Varshney N, Anand MV, Soudagar MEM, Al-Keridis LA, Upadhyay TK, Alshammari N, Saeed M, Subramanian K, Anbarasu K, Rohini K. Cardiovascular diseases prediction by machine learning incorporation with deep learning. Front Med (Lausanne) 2023;10:1150933. [CrossRef]
[44] Baghdadi NA, Farghaly Abdelaliem SM, Malki A, Gad I, Ewis A, et al. Advanced machine learning techniques for cardiovascular disease early detection and diagnosis. J Big Data 2023;10:144. [CrossRef]
[45] Ahmad AA, Polat H. Prediction of heart disease based on machine learning using jellyfish optimization algorithm. Diagnostics (Basel) 2023;13:2392. [CrossRef]
[46] Khdair H. Exploring machine learning techniques for coronary heart disease prediction 2021. Available at: https://thesai.org/Publications/ViewPaper?Volume=12&Issue=5&Code=IJACSA&SerialNo=5 Accessed on Apr 12, 2023.
[47] Gangadhar MS, Sai KVS, Kumar SHS, Kumar KA, Kavitha M, Aravinth SS. Machine learning and deep learning techniques on accurate risk prediction of coronary heart disease. In: Proc Int Conf Comput Methodol Commun (ICCMC); 2023; Erode, India. p.227–232. [CrossRef]