# Galla . Jyothsna (322103282114)

# Karakavalasa . Likitha (322103282122)

***Gayatri Vidya Parishad College Of Engineering For Women (GVPCEW)***

**HEART FAILURE PREDECTION**

**USING**

**MACHINE LEARNING**



**T**

**eam**

**:**

**Nukapeyyi Latha Kumari (322103282071)**



**Sathi Sanvitha Reddy**

**(322103282088)**



***Vishakapatnam,530048***

***Andhra Pradesh***

# Abstract:

The objective of this research is to design a reliable decision-support system for the survival prediction of heart failure patients based on their clinical records and laboratory test results. In the modern era, people are very busy and working hard in order to satisfy their materialistic needs and cannot spend time for themselves, which leads to physical stress and mental disorder. As a result Heart disease has become one of the most important factors for death of men and women in the so called material world. It has emerged as the top killer that has affected both urban and rural population. This study investigates the application of machine learning techniques for heart disease prediction using a comprehensive dataset of 918 patients. The research implements four different algorithms, namely Logistic Regression, Random Forest, Support Vector Machine (SVM), and Decision tree, to construct predictive models that make use of 12 clinical features.

# Keywords:

Heart disease prediction, detection, machine learning, Support Vector Machine (SVM), clinical decision support, predictive Modeling.

# Introduction:

Heart failure (HF) is a condition wherein the heart cannot pump enough blood throughout the body, and it usually develops as a result of chronic conditions like coronary heart disease, high blood pressure, and other heart conditions or diseases. Predicting survival in heart failure is of utmost importance to healthcare providers, patients, and health systems. Prediction of this disease before infection is part of the prevention methods, or the computer tools are the most used means in it, more precisely the Machine Learning algorithms. On the one hand, accurate predictions provide an informed basis for which healthcare providers decide on treatment pathways, intensify intervention, apply advanced life-prolonging therapies, or plan for end-of-life care. Reliable survival predictions can make the use of resources much more efficient so that high-risk patients receive appropriate care and followup. Our study toward this problem is part of applications in data science, in which we de tect cardiac patients based on well-de ned attributes such as (age, sex, choles terol, blood pressure). The use of data collected from patients is very important to train the learning algorithms.

The SVM concept is referred to as an attempt to find the best hyperplane that will separate data into two classes in the input space. The primary aim of the training process on the SVM concept is to determine the location of the hyperplane. The SVM method uses the dot product function. The hyperplane is the line used to separate the dataset. Hyperplane can be a line in two dimensions and can be a flat plane in multiple planes. Here is the illustration of finding out the best hyperplane of the SVM algorithm. Here under is the illustration of best hyperplane of SVM.

**Linear SVM:**

Linear classification is commonly used on datasets of low dimensions. The lower dimension of a dataset means that it has fewer features to classify. Hyperplane in both images can be obtained by measuring the distance (margin) between the hyperplane and the closest point in each class.

**Non-Linear SVM:**

Another working principle on SVM that is used for data which cannot be separated linearly due to its high dimensions is Non-Linear SVM. In the non-linear case, classification is done using the concept of a kernel. In the non-linear case, the kernel concept plays the role of determining the classification limits that are used as a model.

Here we used Non-linear SVMs. So the Non-linear SVMs use kernel functions to transfer the input data into a higher-dimensional space such that a linear separation is possible in order to address this problem.

**The Non-Linear SVM: Why Choose It?**

Non-linear SVM works especially well with data that has curved, complex boundaries. The technique may handle complex relationships in the data by utilizing kernel functions as the sigmoid, polynomial, or Radial Basis Function (RBF). The RBF kernel is one of the most often used of these because of its adaptability and capacity to establish localized decision limits.

# Literature Survey:

Heart disease remains a leading cause of death globally, prompting significant research into prediction systems using machine learning techniques. Various approaches have been proposed, leveraging different algorithms and datasets to enhance accuracy and address challenges like imbalanced data and feature selection.

**2.1. Prediction Models and Algorithms:**

* Sahoo et al. evaluated multiple algorithms, including SVM, Naïve Bayes, and Decision Trees, for predicting heart disease using UCI datasets. They concluded SVM achieved the highest accuracy of 85.2%, proving its robustness for classification tasks(ssrn-3759562).
* Salhi et al. tested neural networks, SVM, and KNN, achieving an accuracy of 93% with neural networks on Algerian datasets. They emphasized the importance of preprocessing and feature selection for optimal results(Using Machine Learning).
* Newaz et al. introduced a Balanced Random Forest (BRF) classifier for heart failure survival prediction. By integrating sampling and feature selection techniques like Chi-square tests and Recursive Feature Elimination, they achieved a G-mean score of 76.83%, addressing class imbalance effectively(article for ml).
  1. **Feature**

Feature selection plays a crucial role in improving model performance by identifying relevant attributes. Recursive Feature Elimination (RFE) and Chi-square tests have been widely used to refine datasets and enhance prediction accuracy. Salhi et al. utilized a correlation matrix to select significant features, while Newaz et al. achieved higher sensitivity using RFE( Using\_Machine\_Learning\_… )

(article for ml).

* 1. **Addressing**

Imbalanced datasets, common in medical predictions, lead to biased models favoring the majority class. Newaz et al. incorporated undersampling within the RF classifier to mitigate bias and improve predictions for minority classes(article for ml). SMOTE and other oversampling methods have also been proposed in similar studies.

* 1. **Comparative Analysis:**
* Algorithms like SVM and Random Forest are frequently preferred for their accuracy and stability. Studies highlight SVM's ability to handle non-linear data efficiently using kernel functions, while RF's ensemble nature reduces overfitting(ssrn-3759562)(article for ml).
* Neural networks, despite their high accuracy, require careful tuning and larger datasets, as demonstrated by Salhi et al.(Using\_Machine\_Learning\_…).

This review underscores the significance of tailored machine learning approaches in predicting heart diseases. Future research should focus on enhancing interpretability, addressing imbalances, and exploring hybrid models for improved outcomes.

# Methodology:

**3.1. Dataset collection:**

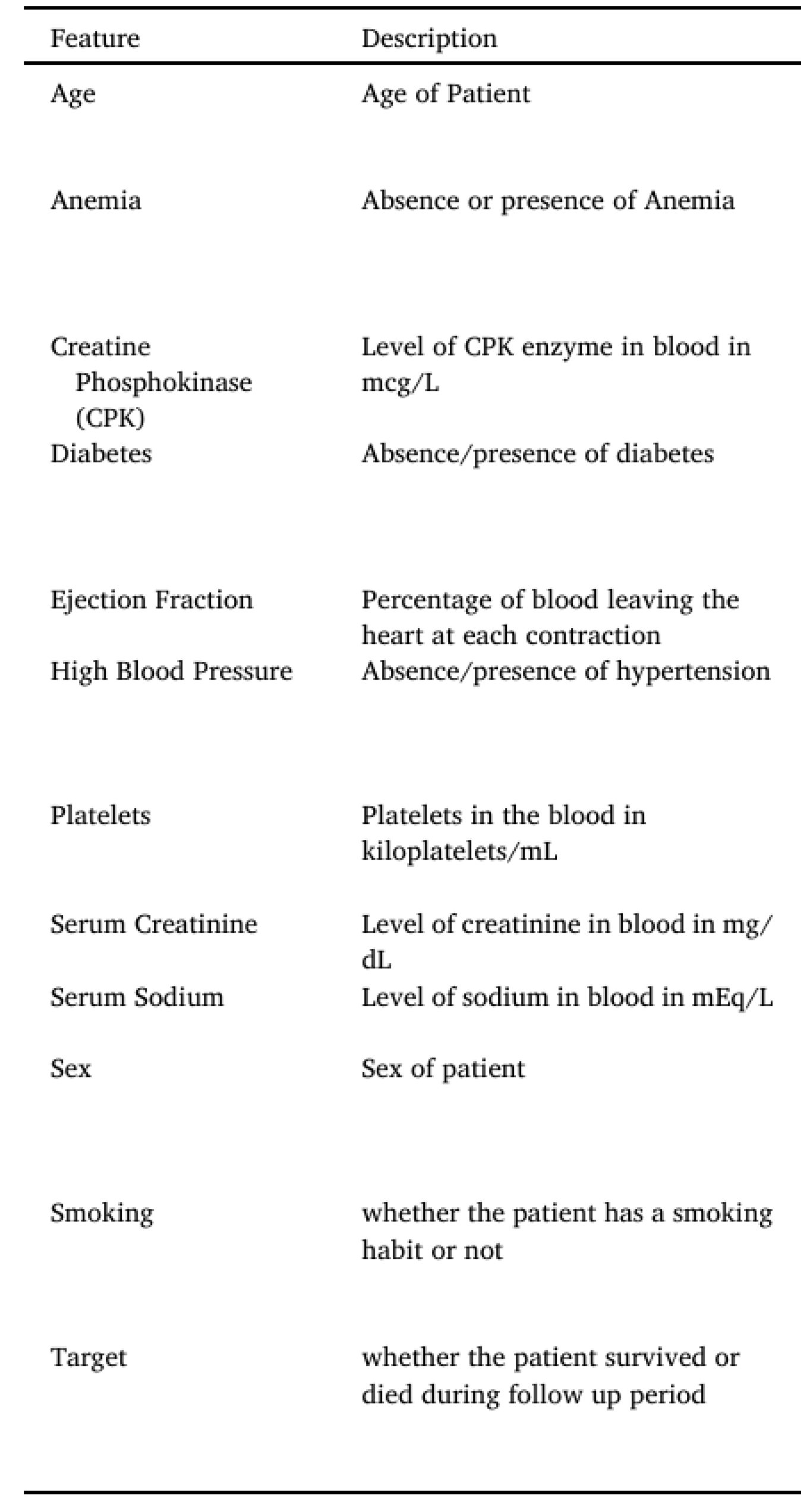
Our application involves a data set, consisting of people who have gone for analysis and tests to test their chances of suffering heart diseases. The data set is essentially a matrix wherein rows correspond to the patients while the columns correspond to the factors or attributes (features) that have to be tested.

The inclusion criteria in the dataset ensured that all the patients had records for the features such as age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate, exercise-induced angina, old peak, ST slope, definite diagnosis on the presence or absence of heart disease.

**3.2. Preprocessing of data:**

The data set was then divided into a train and test. This was achieved by the 'train\_test\_split' function of the sklearn library. Features were further categorized into numerical and categorical types. Numerical comprised Age, RestingBP, Cholesterol, FastingBS, MaxHR, and Oldpeak. On the other hand, categorical comprised Sex, ChestPainType, RestingECG, ExerciseAngina, and ST\_Slope. This was for ensuring the respective preprocessing techniques applied to these feature types.

Pre-processing was required to be prepared with two different pipelines: the numerical pipeline and the categorical one. In the former, an imputation by the median if there were missing data applied while in the latter with subsequent standardization. This is in contrast to another pipeline where missing values would have been replaced by most common ones and then one-hot encoding applied. These would have been merged using ColumnTransformer.



**3.3. Machine Learning Algorithms :**

This work used four different machine-learning algorithms. All of these algorithms were chosen for the different strengths and insights they can provide. Logistic regression, being a basic algorithm for binary classification problems, was chosen due to its interpretability and ease of implementation. This algorithm is a good starting point for determining the relationship between the feature and target variables.

It has been coupled with a Support Vector Machine (SVM) that works well in high dimensional spaces and is quite versatile in terms of kernel functions used. An SVM must be tuned for probability estimation, which would then enable comparing the results with other models. SVM is capable of identifying optimal decision boundaries in complex feature spaces, thus potentially making it a solid tool for the classification task in hand.

**3.4. Model Building :**

During the model development phase, five different machine learning algorithms related to heart disease prediction have been implemented and trained. The development process starts with preparing the dataset, and it is divided into two parts: 70% for training and 30% for testing, both of which are used for unbiased evaluation.

**3.5. Evaluation Metrics :**

Model performance was measured using a set of metrics, all of which gave particular insights into different aspects of the predictive powers of models. Among the chosen metrics were:

**Accuracy:** This measures the proportion of correct predictions compared to total predictions. It is the ratio of correct predictions, including true positives and negatives, to total predictions. Although this is useful generally, it may sometimes mislead, especially in cases of class imbalance.

**Precision:** This measures the accuracy of the positive predictions. It is calculated by dividing the number of true positive predictions by the sum of true positives and false positives. High precision means that it will likely be right when your model makes a positive prediction. In predicting heart diseases, high precision means that when this model projects a patient to have heart disease, the prediction, in most cases, would be right.

**Recall (Sensitivity):** It refers to a model's capability to classify all the positive instances correctly. It can be expressed as a ratio of true positive predictions against the total number of positive instances. Recall is important in a model predicting heart disease since it will reflect that it yields most of the actual cases of heart disease, reducing the chances of missing a diagnosis.

**F1-Score:** The metric gives a balanced model performance measure because it combines precision and recall into one value—the harmonic mean of precision and recall. This will be useful in class imbalance cases since the F1 score will consider both false positives and false negatives.

**ROC-AUC:** This metric tells users a model's ranking power about classes at varying threshold settings. It is obtained by thresholding the plot of the True Positive Rate against the False Positive Rate at various threshold settings and then computing the area under the resulting curve. The higher the ROCAUC, the better the model is at differentiating patients with and without heart disease.

# Implementation and Results :

**Explanation of key functions:**

This work used Python programming for this project, as it is a high level programming language and it has vast libraries and Python automates tasks and makes it efficient.Firstly, we need to install Python then we need to import some libraries, they are:

**1.Numpy:** Numpy is used for multi-dimensional arrays, It does element to element operations and it also has different methods for processing arrays.

1. **Panda:** Pandas is one of the highly used python library, it provides high performance. It manipulates data and it makes data analysis fast and easy.
2. **Sklearn:** It is most useful library, This library contains lot of efficient tools, It is used to build models like statistical modeling including classification,regression,clustering.After loading required packages, we divide dataset as training and testing as follows, here 70 % of dataset is taken as training and remaining 30 % as to perform test.

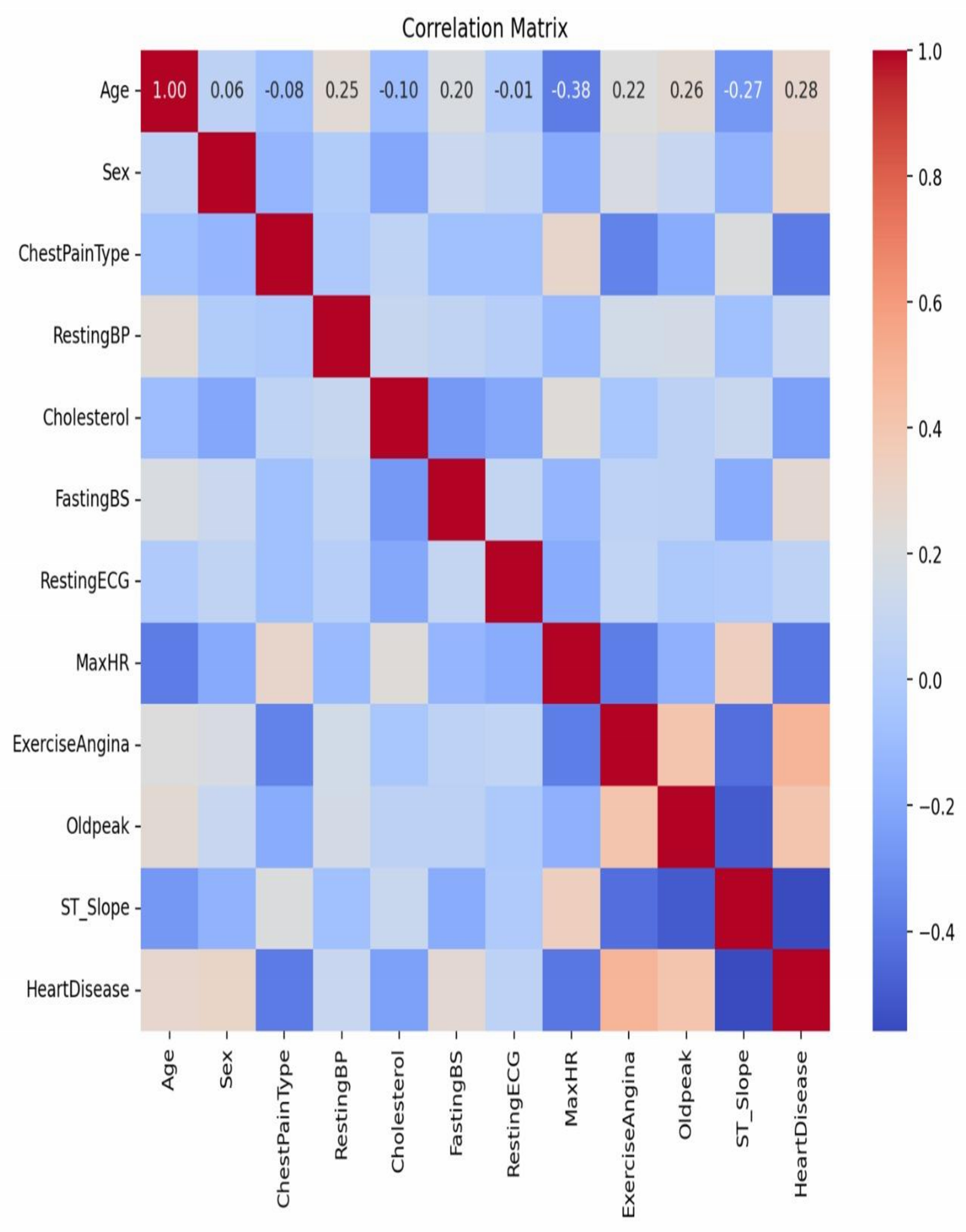
All models performed well; the SVM showed better performance in balanced precision and recall, and also got the highest ROC-AUC score.On at least one level, high performance across these very different model architectures would suggest that the selected featuresThey are very potent predictors of heart disease.

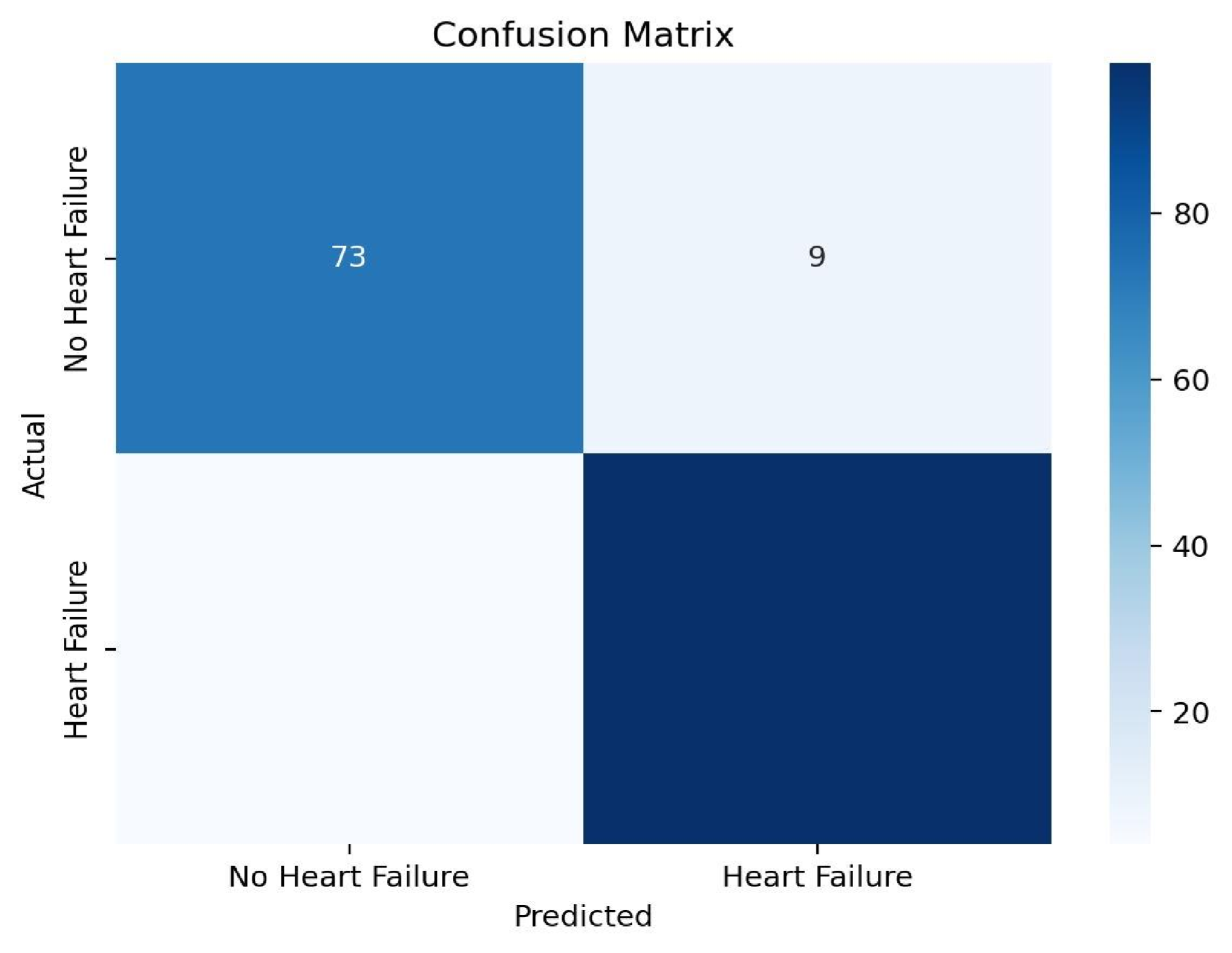
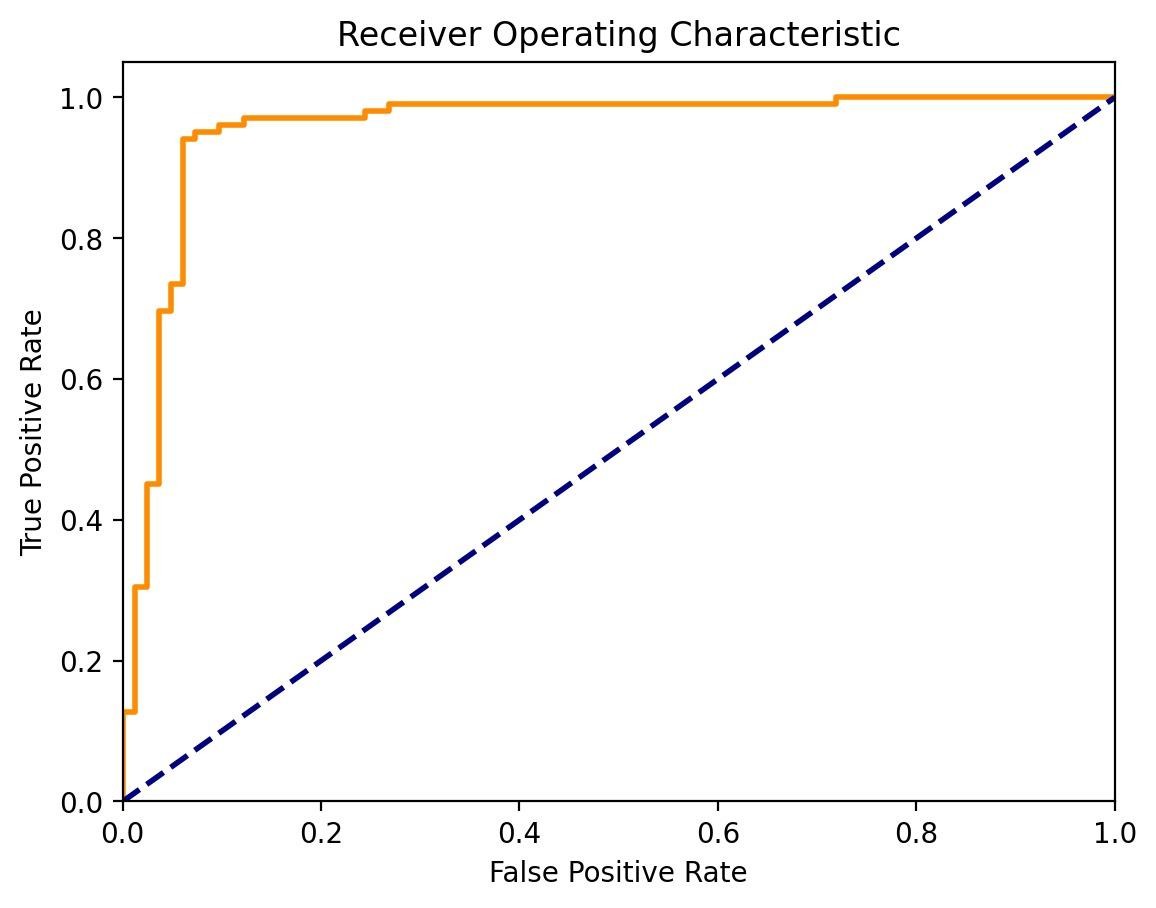
**Results:**

Displaying Results:

* Displays the prediction result (Heart Failure Risk Detected / No Risk) and confidence score.
* Provides an option to view:
* The processed dataset.
* Model information (e.g., SVM with RBF kernel).

**Method of Implementation:**





**Result Analyser:**

This research work uses the data set comprises of several clinical attributes like patient Age, Sex, Chest pain, Fbs and etc as shown in table 2. Next step is to divide the data set into two sets, Training set is 70% and testing set is 30%. Training set is used to construct the model and testing to evaluate the correctness of the model. This research work is carried out by implementing the data set over five different algorithms and results are compared. By using this method this research work could able to predict whether a patient is suffering from heart disease or not (0 is absence and 1 is presence) with 85.2% accuracy. By using the Support vector machine, this model could able to predict with an accuracy of about 85.2%, which is highest as compared to other algorithms.

# Conclusion:

We develop the model to build a robust decision-support system for the prediction of survival in heart failure patients. The performed performance on the public database of HF as well as the versatility of the suggested approach indicate that it would have the potential to act as a reliable tool, making it applicable in clinical practices for assisting clinicianscians and practitioners to take decisions. Serum Creatinine and Ejec\-tion Fraction are the essential risk predictors. However, an age of the patient is of a great concern. Now with only these 3 parameters the heart failure patients present at risk can be classed accurately using our presented approach. Limitation. This study has a problem since the model is made for acomparatively smaller dataset due to the absence of publicly availabledataset for this. A large dataset from another geographicalregion would definitely improve the robustness of the model andwould also give a better insight into the features that are most likelyto be the cause of mortality in heart failure patients.In our model, we have incorporated undersampling approach intothe construction of bootstrapped samples to build our balanced random forest classifier. Other sampling techniques, however, can also be used and their performance compared with the model presented in this study.

Something we plan to look into in future developments.