iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** I **Month of publication:** January 2025

DOI: https://doi.org/10.22214/ijraset.2025.66292

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue I Jan 2025- Available at www.ijraset.com

Heart Failure Prediction Using Machine Learning Techniques

Maroti Biradar[1], Savita S. Wagre[2], Avadhoot Patil[3], Shaikh Affanuddin[4], Mohd Muddabiruszama[5]
[1, 2, 3, 4]*B.Tech Student, Computer Science Department, MGM's College of Engineering*
[5]*Guide, Asst. Prof. (M.Tech.B.E.), Dept. of Computer Science & Engg., MGM's College of Engineering*

Abstract: ***Heart failure is a complex cardiovascular condition defined by the heart's impaired ability to pump blood efficiently, which leads to a great deal of morbidity, mortality, and health care costs internationally. Predicting and intervening in time can significantly improve patient outcomes and reduce the pressure on healthcare systems. This project focuses on building a predictive model using machine learning techniques to detect individuals that are at a higher risk of developing heart failure. In the medical field predicting the heart disease has become a very complicated and challenging task, requires patient previous health records and in some cases, they even need Genetic information as well. So, in this contemporary life style there is an urgent need of a system which will predict accurately the possibility getting heart disease. Predicting a heart failure in early stage will save many people's Life.***

I. INTRODUCTION

A. Background

Human Life is completely depending on the efficient working of Heart. The heart pumps blood over blood vessels to the different body parts of the body, with enough oxygen and other essential nutritional components that are required for smooth functioning of the body. Healthy Heart leads a Healthy Life. But, in today's world heart failure has become vital cause of death for both men and women in the world. Corona virus causes inflammation of the heart muscle leading to heart failure. Experimental evidences suggest that 1 in every five-patient having heart injury due to Corona Virus irrespective of respiratory symptoms. Coronary heart disease is the most common type of heart disease. About 630,000 dies from heart disease each year that's 1 in every 4 deaths. Heart Failure is a condition that occurs when the heart is unable to pump enough blood to the body, and it is usually caused by chronic condition such as coronary heart disease, high blood pressure, or other heart conditions or diseases.

B. Objective

The focus of a Heart failure project is to create a predictive system that is consistent and accurate at identifying likely to suffer from heart failure in the future, so that timely intervention can be sought and the possibility of better outcomes are increased. The project seeks to utilize machine learning algorithms to analyze complex healthcare data, including patient demographics, clinical histories, and diagnostic test results, in order to identify patterns and relationships that may be missed by traditional methods.

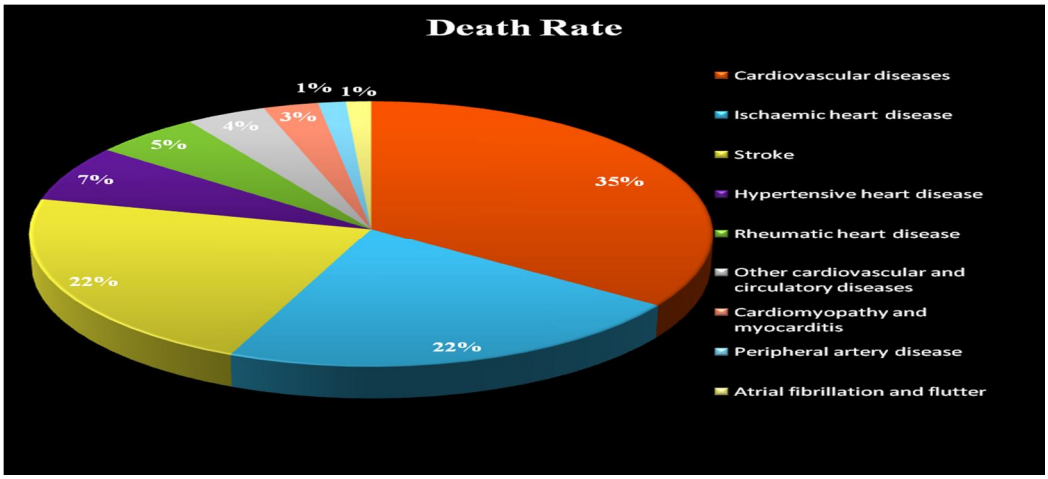


Figure:1. Death rate percentage of heart failure

II. LITERATURE REVIEW

Some prediction systems already existed, and the Authors have researched successfully and proposed the predicting systems through different Classification and prediction algorithms. Some of the existing systems are Kaur K. proposed decision tree-based system for heart disease prediction and that method was proved to be effective for disease prediction. This is predicting the maximum 90% accuracy

Intelligent System based Support vector machine-objective function scheme (used to represent the diagnosis of the patient along with a radial basis function network). It will predict what type of heart disease could occur for a patient Wheater it is heart attack or not based on clinical symptoms. The intelligent machine classifier (support vector machine with sequential minimal optimization algorithm) implemented to a dataset of the patients in India**.** B. Jin, C. Che, Z. Liu, S. Zhang, X. Yin and X. Wei [1] the heart disease is predicted with some attributes collected from the patient and they have also collected patient health records and previous patient history to predict the heart disease. They utilized the (EHR) Electronic Health Records images. EHR is a combination of patient diagnosis record, physician record and hospital record. They can finally predict when a patient will be diagnosed by correlating the links between events and analyzing it using EHR records they have some output which exists in image format unstructured data. We have sparse data of electronic health records, with that, we cannot analyze and predict. Non-structured Unstructured data obtained through Electronic Health Record (EHR).

S. Aditya Varun, G. Mounika, Dr. P. K. Sahoo, K. Eswaran have proposed by using the Logistic regression for heart failure prediction and they got with accuracy of 70.45%. The algorithm used by this system are naïve bayes algorithm, Decision tree and KNN algorithm. This system gives only better accuracy to decision tree and other algorithm lacks accuracy; it does not provide accurate results.

III. METHODOLOGY

A. Architecture Of Proposed Work

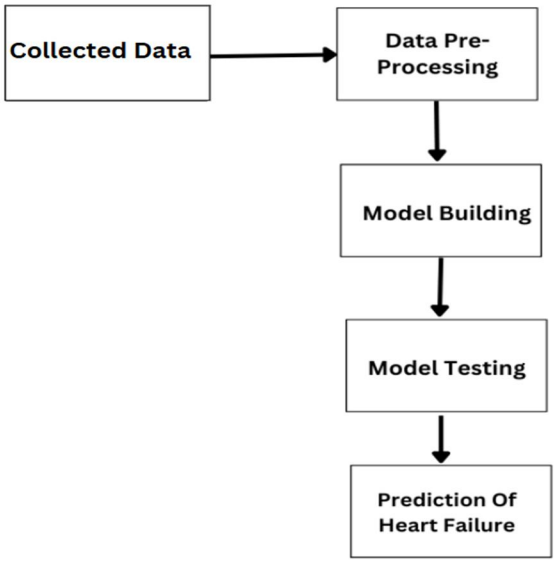


Figure: 2. Architectural Model of Proposed System

1) Collection of data is the first step of the process for this project. We had collected data set from Kaggle, it is available, it is open source.

2) Next task after data collection is data pre- processing. Then they are cleansed in this phase of data and unnecessary values are eliminated. It also takes out the None/ null/ corrupted values.

3) The next step is to split the Data after cleansing the data, we split the data into balanced two sets Training data and test data. We need to work with these missing entries before we go back to build the training model. And from training data we construct a prediction model.

4) We use Decision tree as it is accurate and efficient. Now, we have to calculate the accuracy of the model.

5) The final Step is predicting the heart failure. The final output in percentage how much chance heart failure.

B. Data Set

1) *Data Set Description:* We have collected the data set for this research work from UCI repository. This dataset has 13 major attributes which helps in determining the presence or absence of the heart disease. These preprocessed 13 clinical attributes are trained to predict the absence and presence of heart failure. It means the attributes are in the format of numeric type which constitute the age of the patient in the numeric values, Age is the major risk factor in increasing the heart failure; In ADOLENCE it doubles the risk.
2) *Sex:* Men are at greater risk heart failure than woman. The 1 is represents the Male; whereas 0 represents the female. The Age of the patient is recorded in years, ranging from 40 to 95. Anemia is a Boolean feature (0 or 1) that indicates whether the patient has a decrease in red blood cells or hemoglobin levels. High blood pressure is also a Boolean feature (0 or 1), representing whether the patient has hypertension. Creatinine phosphokinase (CPK) refers to the level of the CPK enzyme in the blood, measured in micrograms per liter (mcg/L), with values ranging from 23 to 7861. Diabetes is a Boolean feature (0 or 1), indicating whether the patient has diabetes. Ejection fraction is a percentage measurement (ranging from 14 to 80) that shows the percentage of blood leaving the heart with each contraction. The Sex of the patient is a binary variable, with 0 representing a woman and 1 representing a man. Platelets are measured in kilo platelets per milliliter (kilo/mL), with values ranging from 25.01 to 850. Serum creatinine indicates the level of creatinine in the blood, measured in milligrams per deciliter (mg/dL), with a range of 0.50 to 9.40. Serum sodium is the level of sodium in the blood, measured in milliequivalents per liter (mEq/L), with a range of 114 to 148. Smoking is another Boolean feature (0 or 1), indicating whether the patient is a smoker. Time refers to the follow-up period in days, ranging from 4 to 285 days. Finally, the (target) death event is a Boolean value (0 or 1) indicating whether the patient died during the follow-up period. This dataset includes both numerical and categorical data, covering a wide range of clinical and demographic features important for health analysis.
3) *Data pre-processing:* is a critical step in preparing data for analysis and mining. It involves cleaning data by removing inaccuracies such as outliers and filling in missing values appropriately. Data integration ensures that related attributes are combined meaningfully, while normalization ensures uniform data scaling. The process also includes removing feature noise, extracting relevant features, and transforming the data into a suitable format for mining procedures. These steps collectively enhance the quality of the data, ensuring more accurate and reliable results.

Feature	Explanation	Measurement	Range
Age	Age of the patient	Years	[40,..., 95]
Anaemia	Decrease of red blood cells or hemoglobin	Boolean	0, 1
High blood pressure	If a patient has hypertension	Boolean	0, 1
Creatinine phosphokinase	Level of the CPK enzyme in the blood	mcg/L	[23,..., 7861]
(CPK)			
Diabetes	If the patient has diabetes	Boolean	0, 1
Ejection fraction	Percentage of blood leaving	Percentage	[14,..., 80]
	the heart at each contraction		
Sex	Woman or man	Binary	0, 1
Platelets	Platelets in the blood	kiloplatelets/mL	[25.01,.., 850.00]
Serum creatinine	Level of creatinine in the blood	mg/dL	[0.50,..., 9.40]
Serum sodium	Level of sodium in the blood	mEq/L	[114,..., 148]
Smoking	If the patient smokes	Boolean	0, 1
Time	Follow-up period	Days	[4,...,285]
(target) death event	If the patient died during the follow-up period	Boolean	0, 1

Table1: shows data set attributes

C. Machine Learning Algorithms

1) Random Forest: Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue I Jan 2025- Available at www.ijraset.com

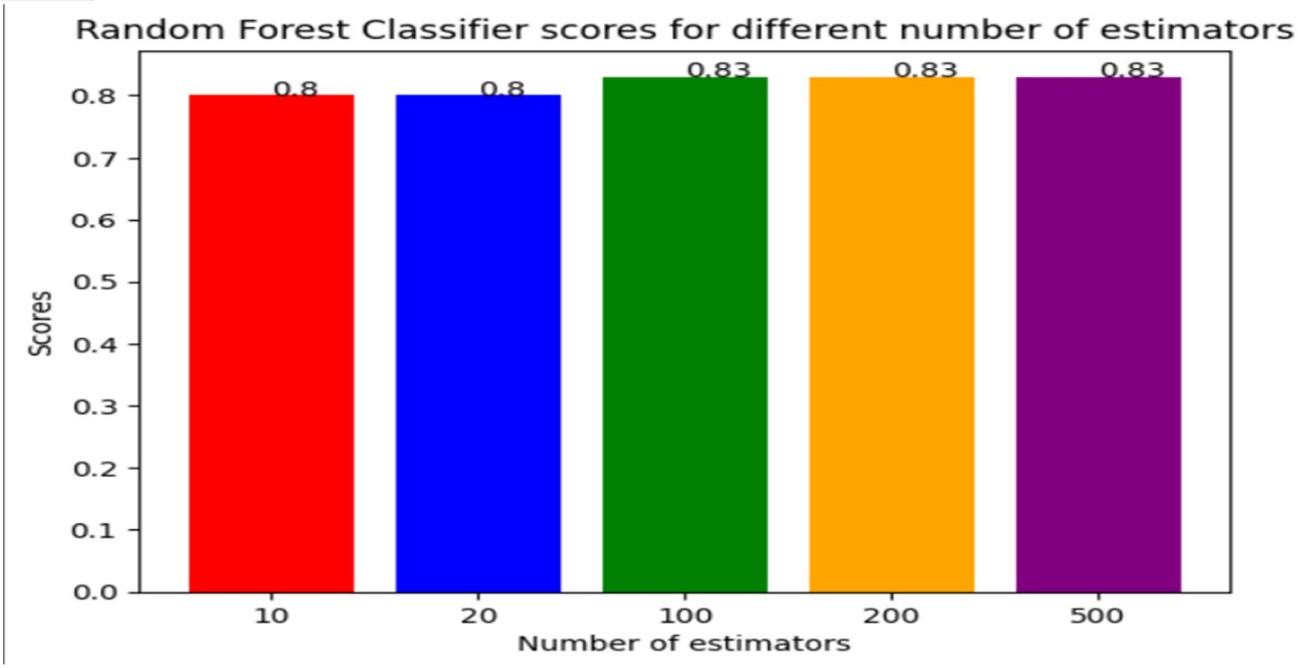


Figure 3: shows accuracy of Random Forest

2) K-Nearest Neighbors: K-Nearest Neighbors is the simplest of all Machine learning algorithms which stores all available cases and the class of the cases. K-NN algorithm assumes the similarity between the new case/data and available cases and assign the new case into the category that is closest to the available categories. K-NN algorithm keeps all of the existing data and classifies a new data point according to the proximity. That is, if new data comes then it can be classified easily into the well-suited category by using K- NN algorithm., as it directly uses the training data for making predictions. However, KNN can be computationally expensive, particularly as the size of the dataset grows, because it needs to compute the distance to every training point for each prediction. It is also sensitive to the choice of K and the scale of the features, which requires careful tuning and preprocessing, such as feature scaling, to improve performance.

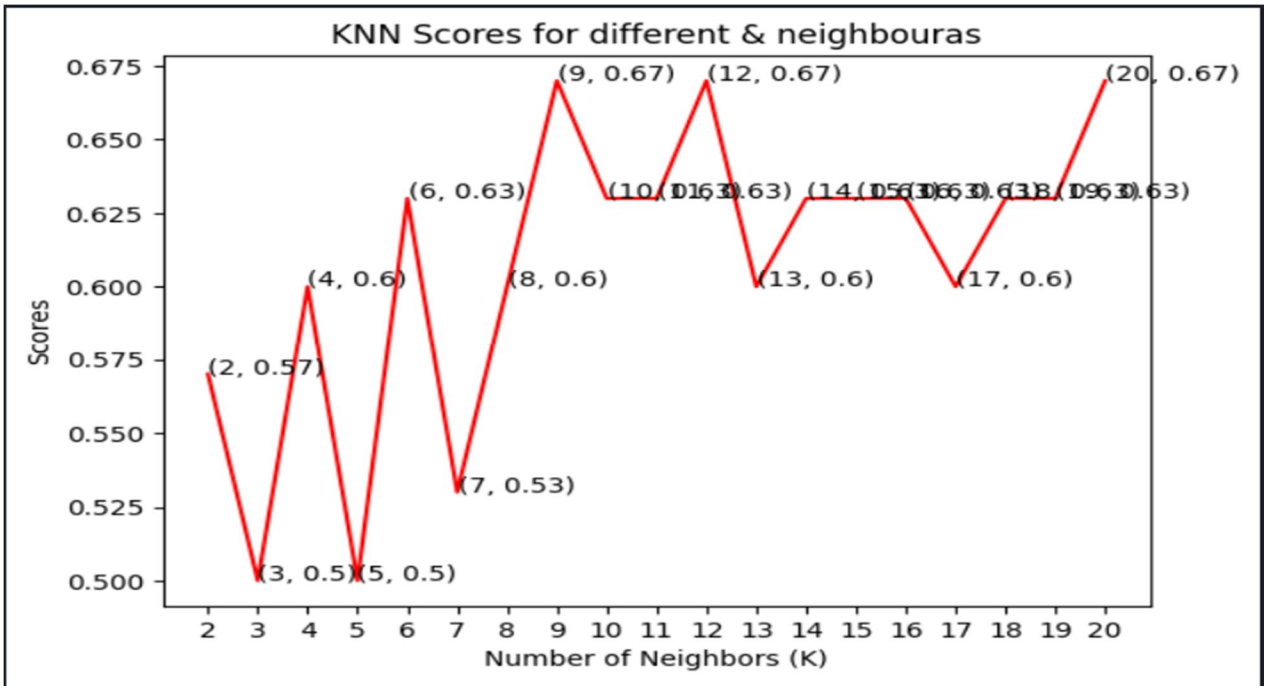


Figure 4: shows accuracy of KNN

©IJRASET: All Rights are Reserved

3) Logistic Regression: Logistic Regression comes under Supervised Learning technique and is one of the most popular Machine Learning algorithms. It is applied for making predictions of the categorical dependent variable based on a known set of independent variables. The output of a categorical dependent variable is predicted by logistic regression. This means that the result has to be 12 categorical or discrete value. This can be Yes or No, 0 or 1, true or False, etc. but instead of returning the exact road or 1-specific values, it returns the values of probability which are bound between 0 and 1.

	precision	recall	f1-score	support
0	0.77	1.00	0.87	20
1	1.00	0.40	0.57	10
accuracy			0.80	30
macro avg	0.88	0.70	0.72	30
weighted avg	0.85	0.80	0.77	30

Figure 5: shows Confusion Matrix of Logistic Regression

4) Decision Tree: Now a decision tree is a Machine learning model which helps you to get organized and reach a decision more easily its tree structure helps you to understand better how the decision- making process works for a classification and regression problem. it starts with a root node containing the complete data set which splits it into branches based on values of features and ultimately reaches leaf nodes which determine the outcome. The splits are determined based on criteria, such as Gini index, entropy, or variance reduction, to ensure that the purpose of dividing the data, at each step, is maximized.

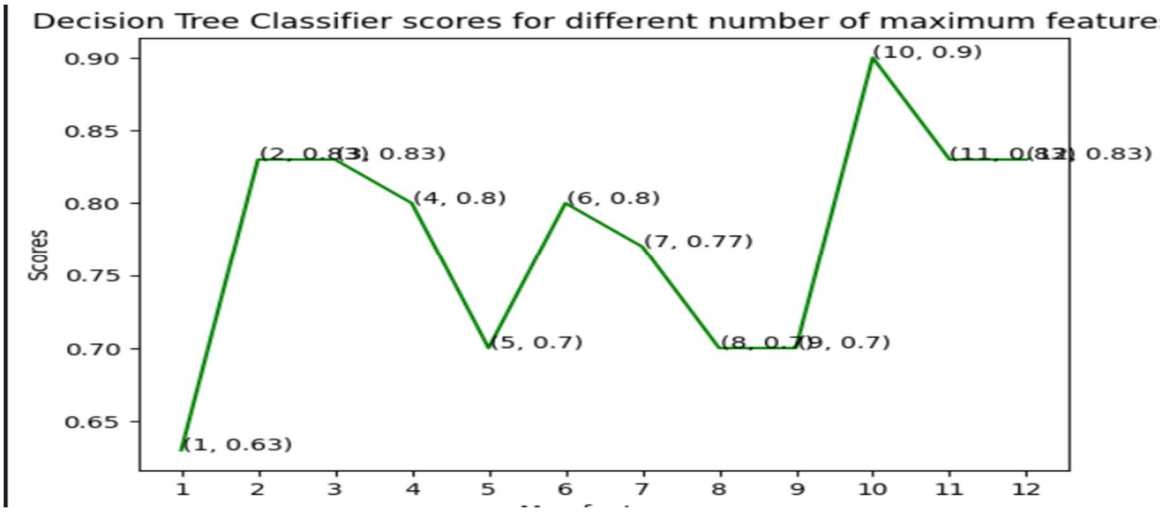


Figure 6: shows accuracy of Decision Tree

D. Implemention

This project was implemented using Python, a high-level programming language that is widely used due to its versatility and extensive libraries. Python automates tasks efficiently and is ideal for data analysis and machine learning projects. Below are the steps and tools used:

Installing Python: The first step is to install Python. After installation, the following libraries are imported:

NumPy: NumPy is used for working with multi-dimensional arrays. It performs element-wise operations and provides various methods for processing arrays efficiently.

Pandas: Pandas is a popular Python library for data manipulation and analysis. It offers high-performance tools for handling and analyzing data, making the process fast and easy.

Sklearn: Sklearn (Scikit-learn) is an essential library for building machine learning models. It provides efficient tools for tasks such as classification, regression, and clustering.

Dataset Splitting: After importing the necessary libraries, the dataset is divided into training and testing sets. In this project, 90% of the dataset is used for training the model, and the remaining 10% is used for testing its performance.

Performance Metrics Analysis

Accuracy

Accuracy measures the overall correctness of the model's predictions. It is calculated as the ratio of correctly predicted values (both positive and negative) to the total number of predictions:

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}}$$

Precision

Precision indicates how often the model's positive predictions are correct. It is calculated as the ratio of true positives to the total predicted positives:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Recall

Recall measures the model's ability to identify actual positive values. It is the ratio of true positives to the total actual positives:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

F1Score

The F1 score combines both precision and recall into a single metric. It is particularly useful when the balance between precision and recall is important. The F1 score is the harmonic mean of precision and recall:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

These metrics are used to evaluate the performance of the machine learning model and ensure its effectiveness in predicting outcomes.

IV. RESULTS

This research used a dataset with clinical attributes such as Age, Sex, Chest Pain, and Fasting Blood Sugar (FBS), among others, as shown in Table 2. The dataset was divided into two parts: 70% for training and 30% for testing. The training set was used to build the model, while the testing set was used to evaluate its accuracy. The study applied the dataset to five different algorithms and compared the results. The model predicts whether a patient has heart disease (0 = no, 1 = yes) with an accuracy of 90%. Among the algorithms, the Decision tree performed the best, achieving the highest accuracy 90%.

V. CONCLUSION

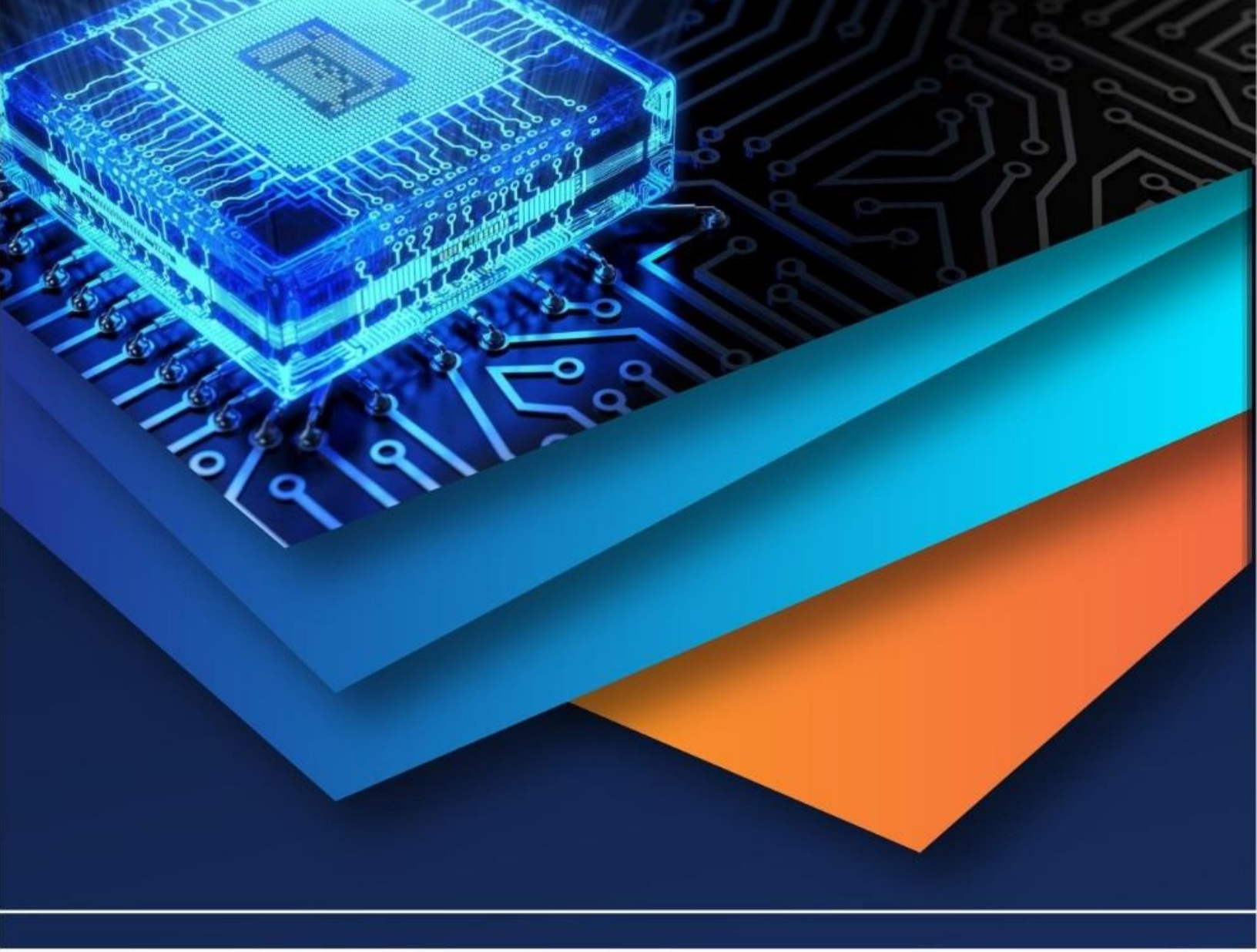
The heart is a vital organ, and the number of deaths due to heart failure is rising rapidly. Studies have also shown that the COVID-19 pandemic has led to heart injuries in many patients, highlighting the urgent need for early detection systems to prevent loss of life. While there are several heart disease prediction systems available, each has its own limitations. The goal of this research was to address these challenges and develop a reliable system to predict the risk of heart failure at an early stage. This study successfully created a robust and accurate model, using the Decision Tree algorithm, which achieved the highest accuracy of 90% among the tested methods. This system can significantly help in the early detection of heart failure and potentially save lives.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

REFERENCES

[1] B. Jin, C. Che, Z. Liu, S. Zhang, X. Yin and X. Wei, "Predicting the Risk of Heart Failure with EHR Sequential Data Modeling," in IEEE Access, vol. 6, pp. 9256-9261, 2018.
[2] B. Wang et al., "A Multi-Task Neural Network Architecture for Renal Dysfunction Prediction in Heart Failure Patients With Electronic Health Records," in IEEE Access, vol. 7, pp. 178392-178400, 2019.
[3] S.Adithya Varun, G.Mounika, Dr. P.K. Sahoo, K. Eswaran, "Efficient system for Heart disease prediction by applying Logistic regression. ijcst vol 10, issue 1, march 2019.
[4] Ahmad, M., Ali, M. A., Hasan, M. R., Mobo, F. D., & Rai, S. I. (2024). Geospatial Machine Learning and the Power of Python Programming: Libraries, Tools, Applications, and Plugins. In Ethics, Machine Learning, and Python in Geospatial Analysis (pp. 223-253). IGI Global.
[5] Ouwerkerk, W., Voors, A. A., & Zwinderman, A. H. (2017). Factors influencing the predictive power of models for predicting mortality and heart failure hospitalization in patients with heart failure. JACC: Heart Failure, 5(5), 377-384.
[6] Bell, J., "Machine learning: Hands-on for developers and technical professionals", INpolis, IN: John Wiley & Sons, Inc, 2020, ISBN: 9781118889060.
[7] Boshra Bahrami, Mirsaeid Hosseini Shirvani, "Prediction and Diagnosis of Heart Disease by Data Mining Techniques", Journal of Multidisciplinary Engineering Science and Technology (JMEST) ISSN: 3159-0040 Vol. 2 Issue 2, February–2015.
[8] Pandey and Rautaray, "Machine learning: Theoretical foundations and practical applications", Singapore: Springer, 2021, ISBN: 978981336517

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)