Heart Failure Prediction using Different Machine Learning Techniques

Prof. Pritesh Patil[1], Rohit Bharmal[2], Shravani Ghadge[3], Dhanashri Gundal[4], Ankita Kawade[5]

[1]Prof. Information Technology, AISSMS Institution of Information Technology, Pune, Maharashtra, India
[2]Student, Information Technology, AISSMS Institution of Information Technology, Pune, Maharashtra, India
[3]Student, Information Technology, AISSMS Institution of Information Technology, Pune, Maharashtra, India
[4]Student, Information Technology, AISSMS Institution of Information Technology, Pune, Maharashtra, India
[5]Student, Information Technology, AISSMS Institution of Information Technology, Pune, Maharashtra, India

---***---

Abstract - *A This study compares the effectiveness of four well-known machine learning methods for predicting heart failure using a publically accessible dataset from kaggle.com: Random Forest (RF), K-nearest neighbors (KNN), Naive Bayes (NB), and Logistic Regression (LR). LR, or regression. They were chosen as a result of their successful applications in the field of medicine. The care of heart failure situations has to be improved in order to raise the survival rate because it is a widespread public health issue. The availability of sophisticated computational systems and the abundance of medical data on heart failure allow researchers to carry out more tests. Accuracy, precision, recall, f1-score, sensitivity, and specificity were used to evaluate the effectiveness of the machine learning algorithms in predicting heart failure using 14 symptoms or characteristics. In comparison to KNN, NB, and LR, experimental investigation revealed that RF delivers the greatest performance score (90.16). The findings of additional RF trials to identify the key indicators of heart failure prediction showed that each of the 14 symptoms or traits is crucial.*

Key Words: **Machine Learning, Heart Failure Prediction, Logistic Regression, Naive Bayes, Random Forest, K-nearest neighbors**

1. INTRODUCTION

The human body's most vital organ is the heart. The effective functioning of the heart is absolutely essential to human life. The heart delivers blood through blood vessels to the various bodily areas, with enough oxygen and other necessary nutritional elements for the organism's efficient operation. A healthy heart leads to a healthy life. But in the modern world, heart disease has emerged as a major factor in both male and female fatalities. cardiac failure results from corona virus induced cardiac muscle inflammation. Regardless of respiratory symptoms, experimental data indicates that 1 in 5 patients had cardiac damage caused by the Corona virus. The most prevalent kind of cardiac illness is coronary artery disease. Heart failure is a major issue that significantly affects people's lives. Most individuals consistently disregard their health due to the faster pace of life, larger portion sizes, and inactivity. Additionally, because of the deterioration of the environment, those factors may contribute to the problem of heart failure, which Future times may see an increase in their frequency. Heart failure might eventually result in death if people did not pay attention to it. The patient's medical and family histories, a physical examination, and test findings are often the foundation of the diagnostic process for heart failure. Due to several risk factors, including diabetes, high blood pressure, high cholesterol, an irregular pulse rate, and many other conditions, it can be challenging to diagnose heart disease. cardiac failure is a severe symptom or advanced stage of several cardiac disorders. Typically, cardiac ejection would be insufficient in patients with heart failure.

Heart failure has a high mortality rate and is expensive to treat. Since heart disease is the most prevalent, it is urgent to develop very accurate and early methods of diagnosing heart disease, which can help many patients survive. There are several scanners available to identify heart illness, however detecting a cardiac ailment before it manifests itself can save many lives. By utilizing a tool that enables the administrator to visually assess the patient's data, we are giving further information to the administrator. Early detection and analysis of the existence of arrhythmia is crucial to preventing patients from developing heart problems. In many circumstances, the existence of a stroke or heart failure may be caused by the small levels of cardiac rhythm. The healthcare sector has a lot of promise with data mining since it can help health systems assess and diagnose diseases by using the data. The cost and time savings result from our ability to evaluate data and forecast illness.

The World Health Organization predicts that heart fragility will cause the deaths of almost 23.6 million people between now and 2030. Therefore, anticipating a coronary illness should be avoided in order to lower the risk. There are two primary categories for heart disease risk factors. We cannot modify the risk variables in the first category, which includes things like age, gender, and family history. Risk factors in the second category include things like smoking, poor eating habits, and high cholesterol; we improve this second group. Therefore, by using the medical data mining classification algorithms, which is a crucial part for identifying the possibility of heart attack, the risk factors belonging to the second class can be eliminated or controlled by changing lifestyle and through medication. Medical professionals most frequently employ angiography to diagnose CAD, however this procedure has some serious drawbacks, chief among them its cost.

2. LITERATURE SURVEY

The authors proposed the predicting systems by using various Classification and prediction algorithms, and there were several prediction systems that were already in use. Among the current systems are In order to demonstrate the approach's suitability for disease prediction, Kaur K. suggested a method for the prediction of heart disease that is based on the principals of component analysis and SVM classification. A hybrid neural network that combines self-encoding and two-way long-term memory was proposed by Ren Y. The model was applied to predict renal disease in hypertension patients, and the final prediction accuracy was 89.7% using the data of 35,332 hypertensive patients.

An intelligent system-based support vector machine and a radial basis function network were proposed by Shashikant Ghumbre and Chetan Patil to reflect the patient's diagnosis. Clinical symptoms will be used to determine the type of heart disease that may manifest in a patient, including whether or not a heart attack is imminent. The patient data set is subjected to the support vector machine with sequential minimum optimization technique. The same data set is then used to make predictions using the Radial Basis Function (RBF) network structure trained by the Orthogonal Least Square (OLS) technique.

CART is able to handle and analyse high dimensional categorical data and employs the Gini index as a measure of the contamination of a partition or a collection of training tuples. Although categorical data must first be converted to continuous data, decision trees can handle continuous data (like in regression). According to B. Jin, C. Che, Z. Liu, S. Zhang, X. Yin, and X. Wei, heart disease may be predicted using some patient-provided features. They also gathered patient health data and prior patient history to do so. They made advantage of the pictures from the electronic health records (EHR). EHR is made up of hospital records, physician records, and patient diagnostic records. By detecting and evaluating the connections between events, they can finally forecast when a patient will be diagnosed based on the output they received from the EHR records, which is some unstructured data in picture format. As a result, using the current data directly will be challenging due to their sparseness.

The data mining approach was proposed by Raj Kumar and Sophia, and they achieved an accuracy of 52.33% in their heart disease detection. To diagnose cardiac illness, they combined characteristics. This system employs the Naive Bayes method, the Decision list algorithm, and the KNN algorithm. This method is inaccurate and produces inaccurate results.

Anbarasi et al. also apply a genetic algorithm in a different strategy. By identifying the characteristics involved in the prediction of heart disease, the patient's required number of tests can be decreased. This method uses three classifiers and requires more time to build models because the classifiers were supplied with fewer characteristics.

Support vector machine and multilayer perceptron neural network architecture were suggested by Gudadhe et al. as a method for diagnosing cardiac illness. To illustrate the presence or absence of heart disease, they used the Support Vector Machine to split the information into two groups. They were 80.41% accurate, whereas the artificial neural network classified the heart disease data into 5 categories with an accuracy of 97.5%.

The hybrid strategy was used to offer a solution by Kanika Pahwa and Ravinder Kumar for choosing the characteristics on the heart disease dataset for prediction. The SVM-RFE and gain ratio were used by the author in their feature selection strategy to get rid of extraneous and unnecessary characteristics. Prediction requires identifying the characteristics. On the subset of features for categorizing the data set into the presence or absence of heart disease, they applied Nave Bayes and Random Forest. When certain characteristics were used, they more accurately attained the results. The accuracy obtained using the Naive Bayes algorithm is 84.15 percent, whereas the accuracy obtained using the Random Forest algorithm is 84.16%. In order to anticipate short-term time series data more accurately and make it acceptable for numerical sequences, the ARIMA is utilized. A neural network may be built to handle the issue for non-numerical time series, however this system is inefficient and does not produce reliable results. Data mining Methods for Heart Disease Years of practice and rigorous medical examinations may be required to determine whether a person has heart disease. It takes more time and money to complete this process because there are numerous tests that must be performed. A system that can more accurately predict the possibility is what we need.

3. PROPOSED WORK (METHODOLOGY)

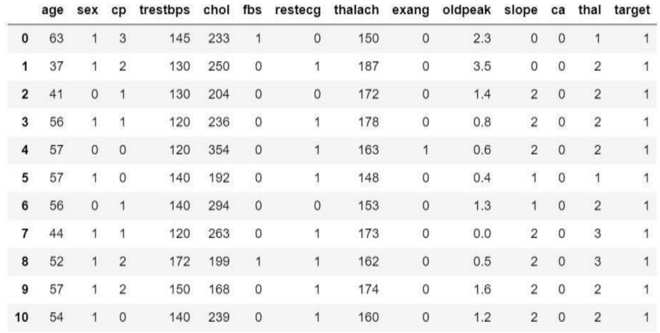
Construction of computer systems that can automatically improve based on experience is the focus of the field of machine learning. It has become more well-liked in the medical field because to its capacity to handle enormous, complicated, and uneven data, one of which is prediction. Different machine learning techniques have been used to foresee heart failure issues. To predict cardiac disease, a comparison of the four machine learning methods Logistic Regression (LR), K-Nearest Neighbor (KNN), Random Forest (RF), and Naive Bayes (NB) has been done. But when RF, NB, KNN, and LR were compared for heart disease prediction, RF outperformed the other ML techniques with an accuracy rate of 90.16%. RF outperformed other categorization methods in terms of heart disease prediction.

3.1 Dataset and Features

The heart disease dataset from Kaggle, which includes 303 cases, is where the dataset is found (https://www.kaggle.com/ronitf/heart-disease-uci). Age, sex, cp (chest pain), trestbps (resting blood pressure), chol (cholesterol), fbs (fasting blood sugar), restecg (resting electrocardiographic), thalach (maximum heart rate), exang (exercise-induced angina), oldpeak, slope, ca (number of major vessels), and thal (thalassemia) are among the 14 variables in this dataset. They act as features or input variables, and the output variable indicates whether the patient exhibiting the listed symptoms is experiencing heart failure. An output variable, target, in this experiment with values of 0 and 1, respectively, denotes the absence and existence of heart failure. Figure 1 provides some instances of the heart disease prediction data utilized in this study, and Table 1 provides a summary for each variable.

Table 1. Description of Features

Age	Age of the person in years
Sex	Sex of the person
Cp	Chest pain experience
fbs	Fasting blood sugar of the person
restecg	Resting electrocardiographic measurement
thalach	Maximum heart rate achieved by the person
exang	exercise-induced angina
oldpeak	ST depression induced by exercise relative to rest
slope	The slope of the peak exercise ST segment
ca	Number of major vessels
thal	Thalassemia
trestbps	Resting blood pressure of the person
chol	Cholesterol measurement of the person
target	Heart failure problem



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
5	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
6	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
7	44	1	1	120	263	0	1	173	0	0.0	2	0	3	1
8	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
9	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
10	54	1	0	140	239	0	1	160	0	1.2	2	0	2	1

Figure 1. Example of Dataset for Heart Disease Prediction

3.2 Machine Learning Techniques

a) Random Forest

A supervised machine learning technique known as random forest is utilized for both classification and regression applications. It is an ensemble learning technique that creates a number of decision trees and then combines their outputs to predict something. A subset of characteristics and data points from the training set are randomly chosen to generate a number of decision trees in a random forest. They are independent of one another since each decision tree is built using a separate subset of the characteristics and data points. Each tree in the forest separately predicts something during prediction, and the combined forecasts of all the trees in the forest lead to the final prediction. This method enhances the model's accuracy while reducing overfitting.

b) Logistic Regression

Binary classification challenges are handled by supervised machine learning techniques such as logistic regression. The likelihood of a binary result as a function of one or more input factors is modeled by logistic regression.The linear combination of the input variables and their weights is transformed into a probability value between 0 and 1 using the logistic function, sometimes referred to as the sigmoid function.

The formula for logistic regression can be expressed as follows:

$$p = 1 / (1 + e^{(-z)})$$

The logistic regression model may take into account interactions between variables and can handle categorical and continuous input variables.To discover the weights or coefficients that minimize the error between the predicted probabilities and the actual binary outcomes in the training data, the logistic regression model is trained using maximum likelihood estimation or gradient descent optimization.

c) K-Nearest Neighbor

The supervised machine learning method K-nearest neighbors (KNN) is utilized for both classification and regression applications. Finding the k closest neighbors to a data point in the training set and using their labels to predict the label of the data point is the foundation of the method.

KNN's mathematical formula is as follows:

Calculate the distance between each data point in the test set and each data point in the training set using a distance metric, such as the Manhattan distance or the Euclidean distance.

Based on their distance from the test data point, choose the k neighbors that are closest to it. The user selects the value of the hyperparameter k.

In classification tasks, choose the class label for the test data point that receives the most votes from its k closest neighbors. Take the average of the output values of the test data point's k closest neighbors to determine the output value for regression tasks.

d) Naive Bayes

In supervised machine learning, the Naive Bayes method is a sort of classification algorithm. Based on the Bayes theorem and the supposition that the input variables are conditionally independent given the class label, the model is constructed. The formula for Naive Bayes can be expressed as follows:

P(y|x1, x2, ..., xn) = P(y) * P(x1|y) * P(x2|y) * ... * P(xn|y) / P(x1, x2, ..., xn)

Naive Bayes makes the "naive" assumption that the input variables are conditionally independent given the class label, allowing for the estimation of each input variable's probability given the class label without taking into account the other input variables. This presumption makes the procedure computationally efficient and makes the calculation of the conditional probabilities easier.

4. ARCHITECTURE OF PROPOSED SYSTEM

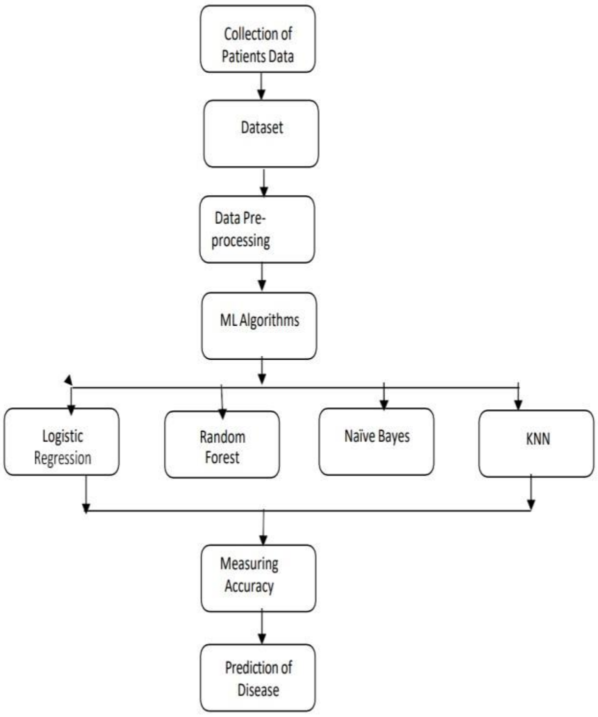


Figure 2. Architectural Model of Proposed system

Figure 2 shows us Architectural Model of Proposed system

1. Data collecting is the first step in this project's process. We have gathered the readily available, opensource data collection from kaggle.
2. Data preprocessing comes after data collecting. The data is cleaned up in this stage by getting rid of pointless values. Additionally, it eliminates corrupted, missing, or null values.
3. Once the data has been cleaned, the next step is to separate it into two sets: training data and testing data. Values must be dealt with before we can build the training model. Using training data, we create a prediction model.
4. We choose the SVM algorithm since it is effective and has higher accuracy. We now need to determine the model's accuracy. Predicting the illness is the last stage. In the final result, 1 will indicate "yes" and 0 will indicate "no."

5. IMPLEMENTATION/ EXPERIMENTAL SETUP

I. Importing essential libraries import numpy as np import pandas as pd import matplotlib.pyplot as plt import seaborn as sns

II. Importing and understanding our dataset dataset = pd.read_csv("heart.csv")

III. Exploratory Data Analysis (EDA)

To comprehend the underlying structure, trends, and relationships in the data, EDA entails the analysis of the data that you have gathered. Before using the data to train your ML models, EDA lets you find any problems or abnormalities in the data that can be fixed.

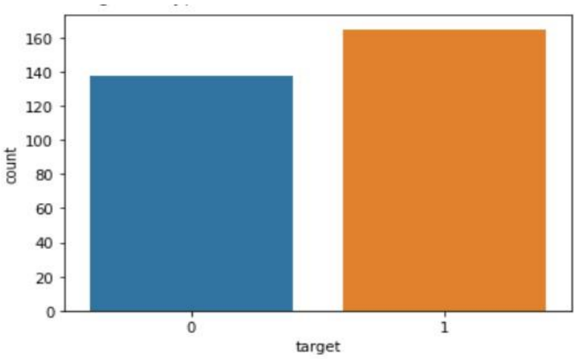


Figure 3. Patient with and without Heart Problems

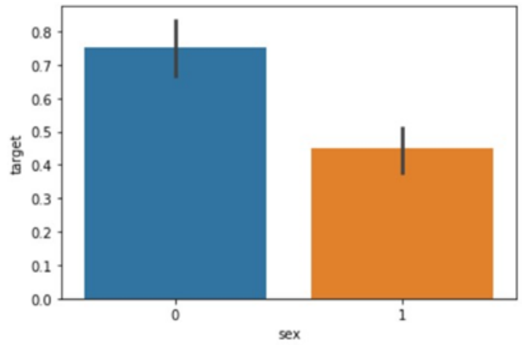


Figure 4. Female Patients and Male Patients

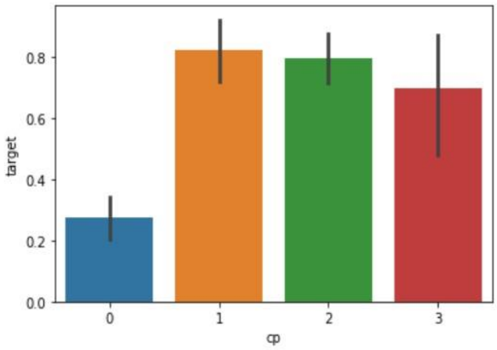


Figure 5. Patient with Typical Anginal, Atypical Angina, Non-Anginal Pain, and Symptomatic

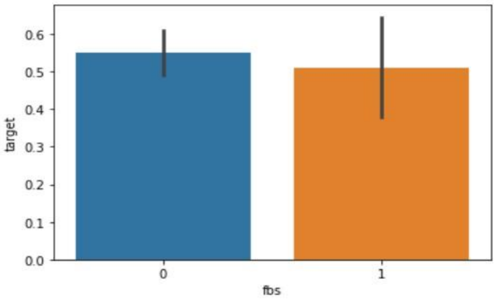


Figure 6. Fasting Blood Sugar of the Patients

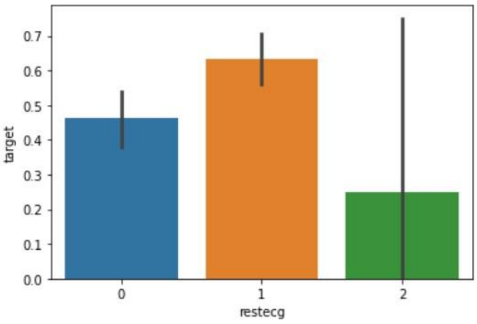


Figure 7. Resting Electrocardiographic Measurement

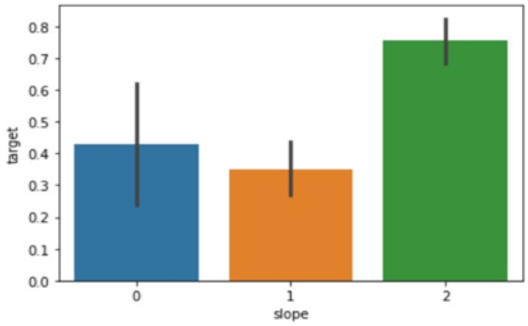


Figure 8. The Peak Exercise ST Segment

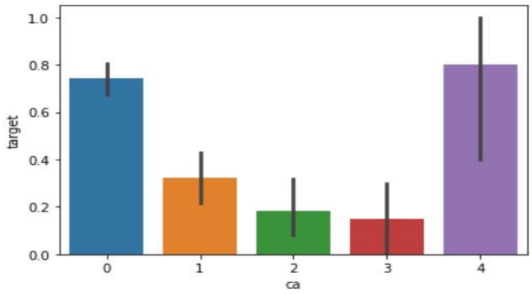


Figure 9. The Number of Major Vessels

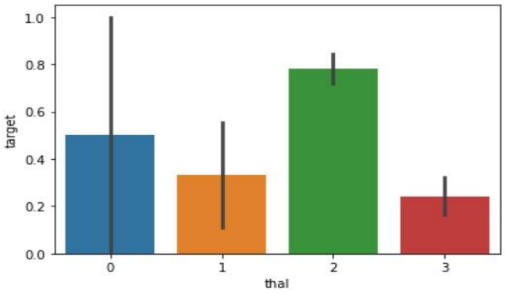


Figure 10. Thalassemia Patients

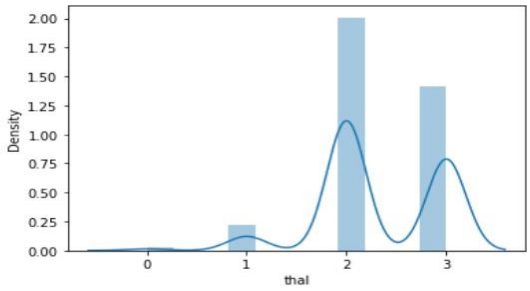


Figure 11. Maximum heart rate achieved by the person

IV.Train Test split

```
from sklearn.model_selection import train_test_split
predictors = dataset.drop("target",axis=1)
 target = dataset["target"]
X_train,X_test,Y_train,Y_test=
train_test_split(predictors,target,test_size=0.20,ra
```

V. Model Fitting

In order to train a model to generate predictions on fresh data, a fundamental stage in machine learning is model fitting. Using Python's scikit-learn module.

from sklearn.metrics import accuracy_score

VI. Output final score

The below Figure 12 shows us the comparison all models of machine learning. The highest accuracy score achieved using Random Forest is: 90.16 %.

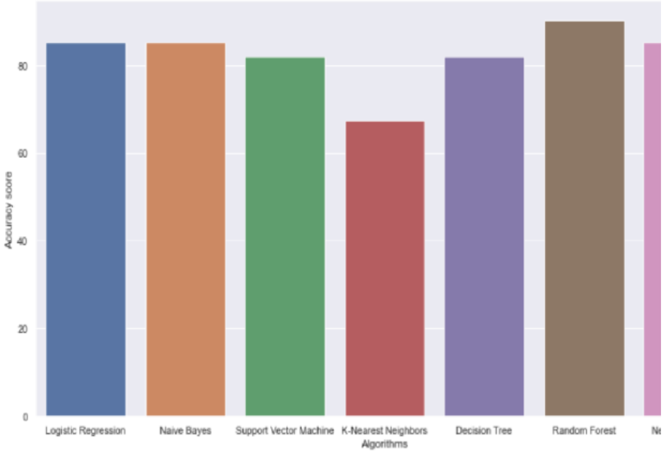


Figure 12. Comparison of all algorithms.

6. RESULT ANALYSIS

As indicated in Table 1, the data set used for this study includes a variety of clinical variables for the patients, including age, sex, chest discomfort, Fbs, and more. The data set is then split into two groups, with the training set being 70% and the testing set being 30%.Testing is done to determine whether the model is accurate once the training set has been constructed. The data set is run through five different algorithms as part of this research project, and the outcomes are compared. This study effort was able to predict with 90.16% accuracy if a patient had heart disease or not using this strategy. When compared to other techniques, this model's accuracy of prediction utilizing the Random Forest was the greatest at roughly 90.16%.

The following are the outcomes after using the algorithms:

ML model/results	Average Score Performance
Random Forest	90.16 %
Naive Bayes	85.25 %
K-nearest neighbors	67.21 %
Logistic Regression	85.25 %

Table 2 lists the performance score of all the four techniques by computing the average for all the evaluation criteria, and it shows that RF achieves the highest average score, followed by LR, NB, and KNN.

7. CONCLUSION

The human heart is the most important organ in the body, and heart failure is responsible for an exponential rise in the loss of human life every day. Experimental research has shown that the Global Pandemic Corona Virus hurts the hearts of many sufferers. Thus, it is urgently necessary for research to concentrate on the causes of heart failure and to develop a reliable early detection system in order to prevent loss of life. Medical officers' time and effort spent on early forecasts for healthcare management objectives are reduced thanks to machine learning technology.

A machine learning approach that can assist in reliably and effectively predicting heart failure as the frequency of fatal heart failures rises. This study demonstrates the potential for machine learning to enhance the healthcare management system by using early heart failure predictions. RF appears to provide the highest performance score among the approaches tested in this trial. It could result in an effective method of controlling the condition that could slow its progression. The accuracy will be increased in future studies by combining machine learning approaches with optimisation algorithms and more data.

8. FUTURE WORK

Future scope for heart disease prognosis reports might be in a number of areas, including:

Expanding the Size and Diversity of the Dataset: Increasing the size and diversity of the dataset can aid in increasing the precision of the machine learning models. To make the dataset more typical of the community, it might contain people with a range of ages, regions, and health issues.

Integration with Electronic Health Records (EHR): When machine learning models are integrated with electronic health records (EHR), it is possible to forecast heart failure accurately and in real time, which improves patient outcomes. Informed choices concerning the patient's treatment may be made by doctors and other healthcare professionals because to this.

Using Multi-Modal Data: Machine learning models for heart failure prediction can be improved upon by using multi-modal data, such as photos, audio, and video. Medical scans, for instance, might offer more details on the anatomy and operation of the heart, which can increase the models' accuracy. The creation of mobile applications can give people a simple and handy method to keep track of their heart health. Mobile applications for heart failure prediction are one such example. To anticipate the possibility of heart failure and prompt patients to seek medical assistance when necessary, these applications can use machine learning models.

ACKNOWLEDGMENT

We appreciate the support of the Information Technology Department of our institution, the AISSMS Institute of Information Technology, and our mentor, Mr. Pritesh Patil, an associate professor there. We would especially want to express our thanks to Dr. Meenakshi Thalor, Head of the Information Technology and Engineering Department of the AISSMS IOIT College of Engineering, for her unwavering direction and support of our project effort.

REFERENCES

[1] Jiang W, Luo J. Graph Neural Network for Traffic Forecasting: A Survey[J]. arXiv preprint arXiv:2101.11174, 2021.

[2] Kim, Young-Tak, et al. "A Comparison of Oversampling Methods for Constructing a Prognostic Model in the Patient with Heart Failure." 2020 International Conference on Information and Communication Technology Convergence (ICTC). IEEE, 2021.

[3] C. for Disease Control and Prevention, "Atrialfibrillation| cdc.gov, " Centers for Disease Control and Prevention, May 2020, Accessed: Mar, 23, 2021. [Online]. Available: https://www.cdc.gov/heartdisease/atrial_ fibrillation.htm

[4] Olsen, Cameron R., et al. "Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure." American Heart Journal (2020).

[5] Fang, Hao, Cheng Shi, and Chi-Hua Chen. "BioExpDNN: Bioinformatic Explainable Deep Neural Network." 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2020.

[6] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor and R. Nour, "An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection," in IEEE Access, vol. 7, pp. 180235-180243, 2019.

[7] Guo, Aixia, et al. "Heart Failure Diagnosis, Readmission, and Mortality Prediction Using Machine Learning and Artificial Intelligence Models." Current Epidemiology Reports (2020).

[8] 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.

[9] 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

[10] Rajalakshmi, S., & Madhav, K. V., A Collaborative Prediction of Presence of Arrhythmia in Human Heart with Electrocardiogram Data using Machine Learning Algorithms with Analytics. 278 287. doi:10.3844/jcssp.2019.278.287, 2019.

[11] 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. 92569261,2018.

BIOGRAPHIES

Prof. Pritesh Patil
Information Technology
Professor
AISSMS Institute of Information Technology

Rohit Bharmal
BEIT Student
AISSMS Institute of Information Technology

Dhanashri Gundal
BEIT Student
AISSMS Institute of Information Technology

Shravani Ghadge
BEIT Student
AISSMS Institute of Information Technology

Ankita Kawade
BEIT Student
AISSMS Institute of Information Technology