



A SYSTEM FOR PREDICTING HEART FAILURE USING MACHINE LEARNING

Ms.M.ANITHA¹, Ms.K.PAVANI², Ms. V. JAYASRI³

#1 Assistant professor in the Master of Computer Applications in the SRK Institute of Technology, Enikepadu, Vijayawada, NTR District

#2 Assistant professor in the Master of Computer Applications SRK Institute of Technology, Enikepadu, Vijayawada, NTR District

#3 MCA student in the Master of Computer Applications at SRK Institute of Technology, Enikepadu, Vijayawada, NTR District

ABSTRACT Heart failure prediction is a critical area of research in healthcare, as early detection and intervention can significantly improve patient outcomes and reduce healthcare costs. Machine learning algorithms have been widely used for heart failure prediction, utilising various features such as patient demographics, medical history, vital signs, and lab results. Heart failure prediction is a challenging task due to the complex and heterogeneous nature of the disease. In this study, we aimed to develop and validate a machine learning model for predicting heart failure in a large population of patients using electronic health records (EHRs).

According to the World Health Organization, 12 million deaths occur yearly due to heart disease. Load of cardiovascular disease is rapidly increasing all over the world in the past few years. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients every day in all cases accurately and consultation with a patient for 24 hours by a doctor is not available since it requires more patience, time and expertise. Our Heart Failure Prediction System is intended to assist patients in recognizing their heart state early and receiving treatment at an earlier stage, allowing them to avoid any serious condition

1.INTRODUCTION

Over 70% of all deaths worldwide are caused by heart disease, more specifically cardiovascular disease (CVDs), which is a leading cause of morbidity and mortality. More than 43% of all fatalities in 2017 were caused by CVD, according to the Global Burden of Disease Study. In high-income countries, unhealthy food, tobacco use, too much sugar, and being overweight or having extra body fat are common risk factors for heart disease. However, the prevalence of chronic diseases is also rising in low- and middle-income nations. Between 2010 and 2015, it is predicted

that CVDs will cost the global economy about USD 3.7 trillion.

Additionally, some consumers cannot afford or use certain technologies, such as electrocardiograms and CT scans, which are essential for diagnosing coronary heart disease. 17 million people have died as a result of the aforementioned reason alone. Employees with cardiovascular disease were responsible for 25–35% of a company's annual medical costs. In order to reduce the financial and physical costs of heart disease for individuals, institutions, and society as a whole, early detection is crucial. By 2030, the WHO



projects that there will be 23.6 million CVD-related deaths worldwide, with heart disease and stroke being the main contributors. Applying data mining and machine learning techniques to predict the likelihood of having heart disease is essential in order to save lives and lessen the financial burden on society. Over 70% of all deaths worldwide are caused by heart disease, more specifically cardiovascular disease (CVDs), which is a leading cause of morbidity and mortality. More than 43% of all fatalities in 2017 were caused by CVD, according to the Global Burden of Disease Study. In high-income countries, unhealthy food, tobacco use, too much sugar, and being overweight or having extra body fat are common risk factors for heart disease. However, the prevalence of chronic diseases is also rising in low- and middle-income nations. Between 2010 and 2015, it is predicted that CVDs will cost the global economy about USD 3.7 trillion.

Furthermore, many low- and middle-income countries find it difficult and often prohibitively expensive to access diagnostic tools like electrocardiograms and CT scans, which are crucial for finding coronary heart disease. The physical and financial toll that heart disease takes on people and organisations must therefore be reduced through early detection. A WHO report estimates that by 2030, there will be 23.6 million CVD-related deaths worldwide, mostly from heart disease and stroke. Therefore, in order to save lives and lessen the financial burden on society, it is essential to use data mining and machine learning techniques to predict the likelihood of developing heart disease

2.LITERATURE SURVEY

2.1 Narain, A.; Isler, Y.; Ozer, M. Early prediction of Paroxysmal Atrial Fibrillation using frequency domain measures of heart rate variability. In Proceedings of the 2016 Medical Technologies National Congress (TIPTEKNO), Antalya, Turkey, 27–29 October 2016.

In order to improve the accuracy of the popular Framingham risk score, Narain et al(2016) .'s study aims to develop an innovative machine-learning-based cardiovascular disease (CVD) prediction system (FRS). The proposed system—which employs a quantum neural network to learn and recognise patterns of CVD—was experimentally validated and compared with the FRS using data from 689 people who had symptoms of CVD and a validation dataset from the Framingham research. The accuracy of the proposed system in predicting the risk of CVD was found to be 98.57%, which is significantly higher than the FRS's accuracy of 19.22% and other methods currently in use. The suggested method may be a helpful tool for doctors in predicting CVD risk, helping to create better treatment plans, and facilitating early diagnosis, according to the study's findings.

2.2 Shah, D.; Patel, S.; Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. *SN Computer Science* 2020, 1, 345.

Shah et al(2020) .'s study sought to create a model for the prediction of cardiovascular disease using machine learning methods. The 303 instances and 17 attributes of the Cleveland heart disease dataset, which was sourced from the UCI



machine learning repository, were used to generate the data for this project. The authors used a range of supervised classification techniques, including k-nearest neighbour, naive Bayes, decision trees, and random forests (KKN). The study's findings showed that, at 90.8% accuracy, the KKN model had the highest level of precision. The study emphasises the potential value of machine learning methods for predicting cardiovascular disease and stresses the significance of picking the right models and methods to get the best outcomes.

2.3 Drożdż, K.; Nabrdalik, K.; Kwiendacz, H.; Hendel, M.; Olejarz, A.; Tomasiak, A.; Bartman, W.; Nalepa, J.; Gumprecht, J.; Lip, G.Y.H. Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: A machine learning approach. *Cardiovascular Diabetol.* 2022, 21, 240.

In a study by Drod et al. (2022) the goal was to identify the most important risk factors for cardiovascular disease (CVD) in patients with metabolic-associated fatty liver disease using machine learning (ML) techniques (MAFLD). 191 MAFLD patients had their blood biochemically analysed, and subclinical atherosclerosis was evaluated. Using ML techniques, such as multiple logistic regression classifier, univariate feature ranking, and principal component analysis, a model to identify those with the highest risk of CVD was created (PCA). The most important clinical traits, according to the study, were hypercholesterolemia, plaque scores, and length of diabetes. With an AUC of 0.87, the ML technique performed well, correctly classifying 114/144 (79.17%) low-risk patients and 40/47 (85.11%) high-risk patients. Based on straightforward

patient criteria, the study's findings suggest that an ML method is useful for identifying MAFLD patients with widespread CVD.

3. PROPOSED SYSTEM

We used training data, trained and evaluated a model instance, adjusted hyper-parameters, and exported the final model to get the maximum accuracy for predicting heart failure. Our model was built using Naive Bayes for high accuracy. It is critical to keep in mind that the quality and quantity of the training data utilised can affect the model's accuracy. It may also be necessary to conduct additional research on and testing of the model's capacity to forecast heart failure using new data sets.

3.1 IMPLEMENTATION

3.1.1 Data collection:

The data was originally taken from kaggle.com, an open-source website with many data sets. However, after some investigation, it was discovered that Oracle had published the dataset. Despite being aware of the fact that we were unable to locate the exact dataset's release link or point of origin, we were still able to find some related links from Oracle that contained the data and provided additional information about its source. This dataset was produced by combining various datasets that were previously available separately but had never been combined. The largest heart disease dataset currently available for research purposes, this dataset combines five heart datasets using eleven features in common. The following five datasets were used to curate it: 303 observations for Cleveland 294 observations in Hungarian 123 observations from Switzerland 200 observations from Long Beach, VA Data

set for Stalog (Heart): 270 observations
This dataset can be used for a number of heart disease-related research projects, including those involving prevention, diagnosis, and treatment. Based on patient medical histories and other factors, researchers can use this dataset to create machine learning models that can precisely predict the development of heart diseases in patients.

3.1.2 Data Cleaning:

The data set must be free of any flaws that might obstruct testing or, more seriously, result in inadequate analysis. Effective solutions must be found for these flaws or issues brought on by redundant records, missing values, or loss of dimension. Bad data will therefore be removed in this step, and missing data will be added. We need to cut out any extraneous information and possibly add any that is missing from the information we currently have, which is a comprehensive general information.

```
Age          0
Sex          0
ChestPainType  0
RestingBP    0
Cholesterol  0
FastingBS    0
RestingECG   0
MaxHR        0
ExerciseAngina 0
Oldpeak      0
ST_Slope     0
HeartDisease 0
dtype: int64
```

Fig 1 . Sum of null values

Since there are no missing values in the data set, handling null values is an important step in the data cleaning process. because it makes handling problems that come up during subsequent procedures easier. There are numerous ways to deal with null values and missing values. Regression, mean, and median techniques can be used to replace missing values or to remove the entire data set.

3.1.3 Exploratory Dataset:

To better understand and investigate the information contained in the data, a lot of information must first be discovered. By visualising the data, we can gain a better understanding of the information contained in the data.

```
#      Column      Non-Null Count  Dtype
---  -
0     Age          918 non-null    int64
1     Sex            918 non-null    object
2     ChestPainType  918 non-null    object
3     RestingBP      918 non-null    int64
4     Cholesterol    918 non-null    int64
5     FastingBS      918 non-null    int64
6     RestingECG     918 non-null    object
7     MaxHR          918 non-null    int64
8     ExerciseAngina 918 non-null    object
9     Oldpeak        918 non-null    float64
10    ST_Slope       918 non-null    object
11    HeartDisease   918 non-null    int64
dtypes: float64(1), int64(6), object(5)
memory usage: 86.2+ KB
```

Fig 2. Heart failure data

The heart failure data is made up with 11 features out of 11 features only 7 are numerical features and remaining all are categorical features.

3.1.4 Data transformation:

In this stage, data will be arranged or managed to make it useful for achieving the specified objective.

<Axes: >

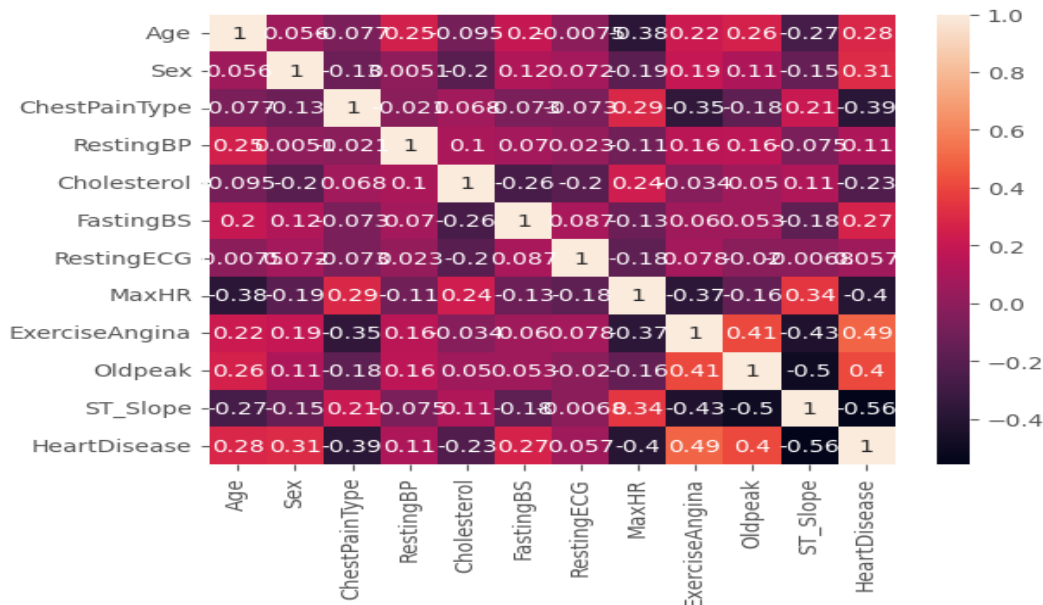


fig 3. Relation Between dependent and independent variables.

We must utilise supervised machine learning techniques to forecast the result. When we have a small amount of data, we can sometimes work with it easily, but as the data volume grows, it gets harder to identify predictors or variables. When it comes to this situation, using all the data can frequently be detrimental, which has an impact on both the computational resources and the model's accuracy. As we examine the relationship between dependent and independent features and then choose the features that are crucial for prediction, the idea of correlation enters the picture. The above figure illustrates the connection between each feature and how they correlate with one another.

3.1.5 Data Exploration:

The necessary data is extracted from the initially loaded data using Pandas data frames. Data is thoroughly examined, information is found, and then it is drawn to a conclusion to produce the report. 36 Line graphs are used to present the data so that the user can analyse it using matplotlib and seaborn. The final graph displays all of the line graphs for easier comparison. The data is displayed on different graphs to show trends in the various dataset values obtained using data visualisation before being combined into one big line graph for analysis and comparison with the same date.

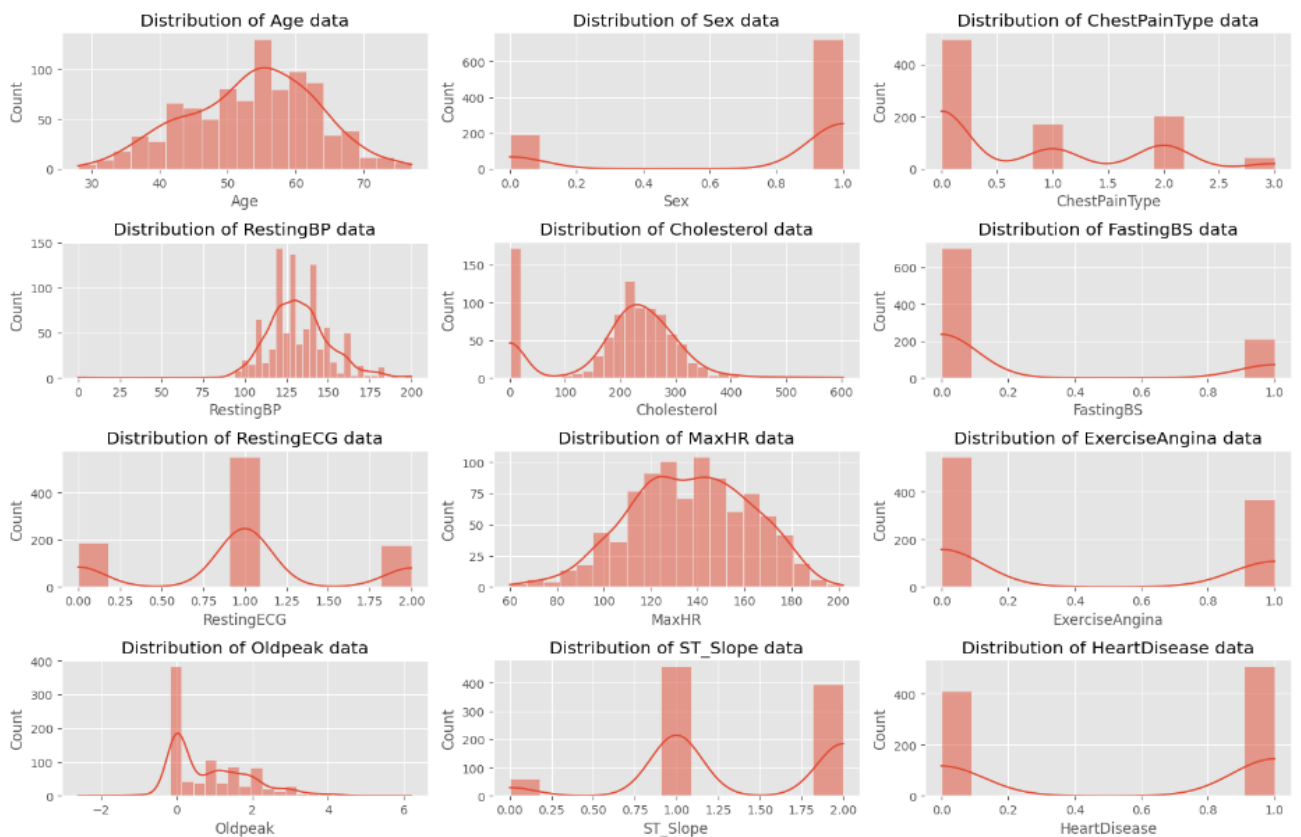


Fig 4 : showing the pairwise relationships between the categorical features that are present in the dataset.

4.RESULTS AND DISCUSSION

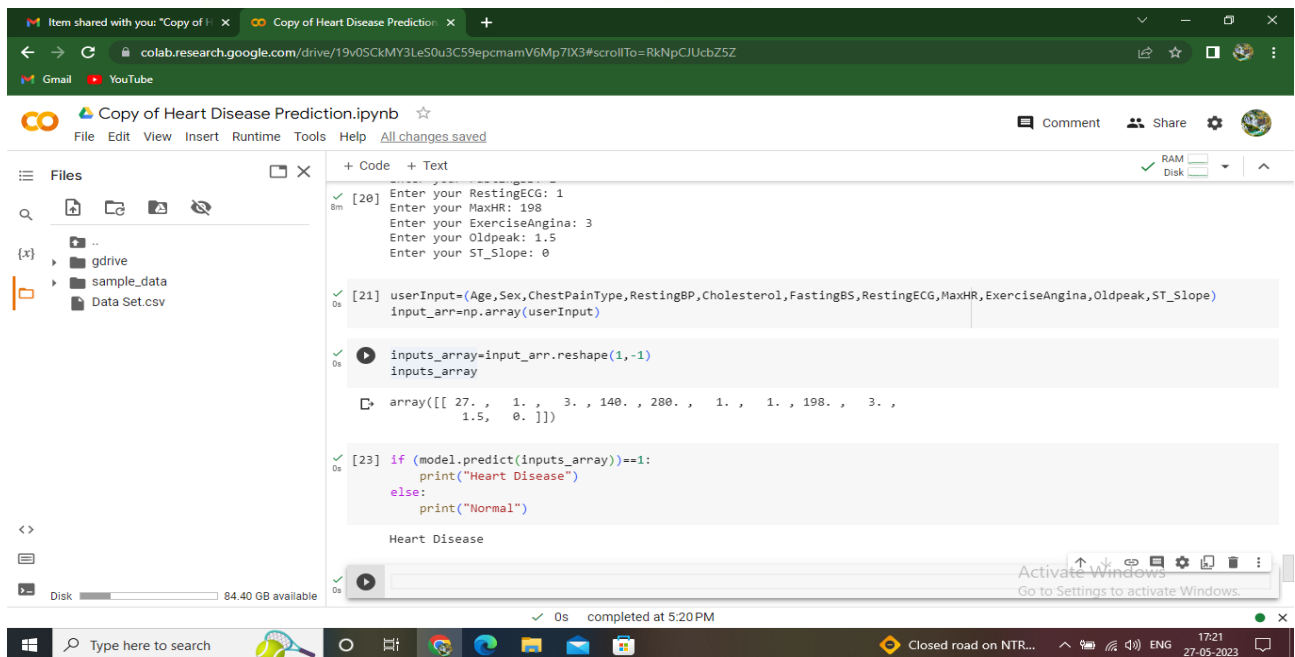
```

Age = int(input("Enter your age: "))
Sex = int(input("Enter your gender: "))
ChestPainType = int(input("Enter your ChestPainType: "))
RestingBP = int(input("Enter your RestingBP: "))
Cholesterol = int(input("Enter your Cholesterol: "))
FastingBS = int(input("Enter your FastingBS: "))
RestingECG = int(input("Enter your RestingECG: "))
MaxHR = int(input("Enter your MaxHR: "))
ExerciseAngina = int(input("Enter your ExerciseAngina: "))
Oldpeak = float(input("Enter your Oldpeak: "))
ST_Slope = int(input("Enter your ST_Slope: "))

userInput=(Age,Sex,ChestPainType,RestingBP,Cholesterol,FastingBS,RestingECG,MaxHR,ExerciseAngina,Oldpeak,ST_Slope)
input_arr=np.array(userInput)
    
```

Enter your age: 27
 Enter your gender: 1
 Enter your ChestPainType: 3
 Enter your RestingBP: 140
 Enter your Cholesterol: 280
 Enter your FastingBS: 1
 Enter your RestingECG: 1
 Enter your MaxHR: 198
 Enter your ExerciseAngina: 3
 Enter your Oldpeak: 1.5
 Enter your ST_Slope: 0

Fig 5: Here we are giving the inputs based on that only we are getting the result of the patients.



```

[20] Enter your RestingECG: 1
      Enter your MaxHR: 198
      Enter your ExerciseAngina: 3
      Enter your Oldpeak: 1.5
      Enter your ST_Slope: 0

[21] userInput=(Age,Sex,ChestPainType,RestingBP,Cholesterol,FastingBS,RestingECG,MaxHR,ExerciseAngina,Oldpeak,ST_Slope)
      input_arr=np.array(userInput)

[22] inputs_array=input_arr.reshape(1,-1)
      inputs_array
      array([[ 27. ,  1. ,  3. , 140. , 280. ,  1. ,  1. , 198. ,  3. ,
            1.5,  0. ]])

[23] if (model.predict(inputs_array))==1:
      print("Heart Disease")
      else:
      print("Normal")
      Heart Disease
  
```

Fig 6: This is the output for which we have given the inputs.

5.CONCLUSION

The expert model foretells whether a person will experience heart failure or heart disease. By identifying patients who are at a high risk of developing heart failure or heart disease, this model can be a useful tool for doctors and other healthcare professionals, enabling early intervention and prevention. The model should be used in conjunction with other clinical assessments because it may have limitations, but it should not serve as the sole basis for decisions regarding diagnosis and treatment.

So by using this machine learning algorithms we can say that the persons heart condition is in normal stage or in critical stage. If the patient heart condition is in normal state means then it is good . If the patients heart condition is in critical stage means then that particular patients have to under gone with the treatment which has to be done to the patients

REFERENCES

- Narin, A.; Isler, Y.; Ozer, M. Early

prediction of Paroxysmal Atrial Fibrillation using frequency domain measures of heart rate variability. In Proceedings of the 2016 Medical Technologies National Congress (TIPTEKNO), Antalya, Turkey, 27–29 October 2016.

- Shah, D.; Patel, S.; Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. *SN Comput. Sci.* 2020, *1*, 345.
- Drożdż, K.; Nabrdalik, K.; Kwiendacz, H.; Hendel, M.; Olejarz, A.; Tomasik, A.; Bartman, W.; Nalepa, J.; Gumprecht, J.; Lip, G.Y.H. Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: A machine learning approach. *Cardiovasc. Diabetol.* 2022, *21*, 240.
- Subahi, A.F.; Khalaf, O.I.; Alotaibi, Y.; Natarajan, R.; Mahadev, N.; Ramesh, T. Modified Self-Adaptive Bayesian Algorithm for Smart Heart Disease Prediction in IoT System. *Sustainability* 2022, *14*, 14208.



AUTHOR PROFILES



Ms.M.ANITHA completed her Master of Computer Applications and Masters of Technology. Currently working as an Assistant professor in the Department of Masters of Computer Applications in the SRK Institute of Technology, Enikepadu, Vijayawada, NTR District. Her area of interest includes Machine Learning with Python and DBMS.



Ms.K.Pavani completed her Master of Computer Applications. Currently working as an Assistant professor in the department of MCA at SRK Institute of Technology, Enikepadu, NTR (DT). His areas of interest include Artificial Intelligence and Machine Learning.



Ms. V. Jayasri is an MCA student in the Department of Computer Applications at SRK Institute of Technology, Enikepadu, Vijayawada, NTR District. She had a Completed Degree in B.Sc.(computers) from Andhra Loyola College(Autonomous) Vijayawada. Her areas of interest are DBMS, and Machine Learning with Python.