IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 9, May 2023

Predict Cardio Disease using Supervised ML Algorithm

K Madhu Babu, Pratham Kabra, M. Srujan, Nithin Sai
Department of ECM
Sreenidhi Institute of Science and Technology, Hyderabad, India

Abstract: *Cardiovascular disease is the leading cause of death worldwide, with deaths due to heart disease becoming a significant issue, claiming approximately one life per minute. Early detection of heart disease is a major challenge, and data science is increasingly being used to process large amounts of healthcare data. Automating the prediction process is critical to minimize the risks associated with heart disease and alert patients in advance. In this study, we propose a heart disease prediction system that classifies patients' risk levels using Naive Bayes, Decision Tree, Logistic Regression, and Random Forest algorithms. By analyzing patient data, our system predicts the likelihood of heart disease and provides effective prescription details based on the risk level. We have implemented a robust Machine Learning algorithm - Random Forest - to design an effective heart attack prediction system. Our system aims to identify different risk levels of heart attack, including normal, low, or high, at early stages, thus preventing the loss of lives.*

Keywords: Random Forest, Naïve Bayes, Classification and Regression, Cleveland Dataset

I. INTRODUCTION

Heart disease is a major public health concern globally, with it being the leading cause of death in many countries. Early detection and prevention of heart disease are crucial in reducing the number of deaths caused by the disease. With the advancement of computer science has brought vast opportunities in different areas, it is possible to predict heart diseases employing various Machine Learning Algorithms.[3] In this context, this paper presents a heart disease prediction system that uses various ML algorithms, such as Random Forest, SVM, Decision Tree, and Logistic Regression, to classify patients into different risk levels of heart disease. However, complete and frequent physical evaluations would lead to data overload. Heart failure patients and society would benefit if we could provide an accurate, systematic diagnostic service for the population. To this end, this paper develops a new approach to this vital task using an enhanced long shortterm memory networks (LSTM) method and a data-driven framework.[1]

The motto of the project is to ease the way of testing and prevention of heart disease. Proposed model is used as the tool to make complex test simple and accurate.[4]

II. LITERATURE REVIEW

Bo Jin, Chao Che (2018) Introduced a “Predicting the Risk of Heart Disease With EHR” model designed by applying Artificial neural networks. In this study, real-world electronic health record data of patients with heart disease was utilized to analyze and predict the occurrence of heart disease. A one-hot encryption model was implemented to diagnose heart failure events based on the principles of a neural network model with expanded memory. The analysis of results revealed the significance of considering natural patterns in the records.[1]

Edward Choe, Jimeng Sun(2017),”Using recurrent neural network models for early detection of heart failure onset” model designed by using recurrent neural networks. In comparison to well-known techniques like logistic regression, MLP, SVM, and KNN, the GRU models displayed remarkable performance in forecasting the diagnosis of HF. The outcomes of the study emphasized the significance of considering the sequential nature of medical records. Subsequent research will involve integrating expert knowledge into our methodology and extending our approach to other healthcare domains. [2]

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 9, May 2023

AKINROTIMI Akinyemi, MABAYOJE Modinat(2021)," A Machine Learning Approach to Diagnosing Heart Diseases" uses boosting algorithms. a performance comparison amongst four machine learning algorithms for detecting the presence or absence of heart diseases in patients has been carried out. The accuracy result of each of the machine learning algorithms and their predicted outputs, suggest that the lower the rate of the false negatives obtained in the use of a classification algorithm, in carrying out the classification of a dataset.[3]

III. WORKING AND ALGORITHM

In this paper the proposed algorithm uses supervised machine learning algorithms which works on the principle of classification and regression model. To improvise our algorithm we also use Feature selection algorithm like K-Features, Lasso Feature Selection model so that we can reduced number of variables and increase the efficiency of our algorithm. The proposed model overview

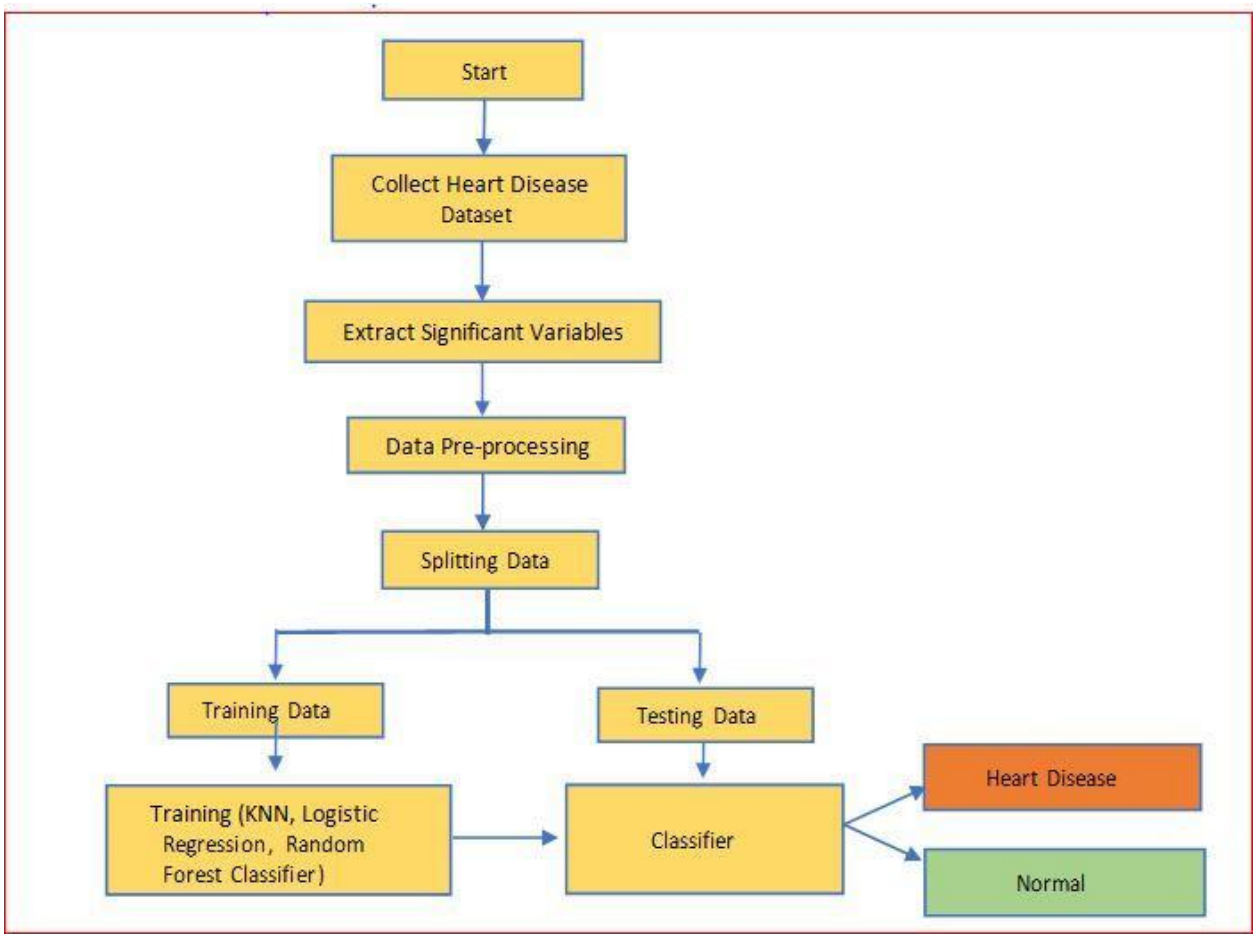

Fig 1 Proposed Model

To implement the above flow we required several steps to be followed. Firstly we need to dig all the variables which effect our algorithm. Then datasets of corresponding variables is collected and cleansed for effective results. The data source and the factors or measures used for calculating the effectiveness of algorithm are given as follows

A. Data Collection

The Cleveland dataset is a frequently employed dataset in the domain of heart disease prediction. It is commonly used by machine learning researchers, and contains a total of 303 instances and 76 attributes, however, only 14 attributes are commonly referred to in published research. The output attribute, "Num," has varying values ranging from 0 to 4, indicating the degree of the disease's presence. The dataset includes some missing values, which are filled with interpolation values. To perform experiments, the dataset is split into two parts - 85% for training and 15% for

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 9, May 2023

validation. The training dataset is comprised of 258 instances and 13 attributes, while the validation set contains 45 instances and 13 attributes. The experiment utilizes 13 input parameters, and the output parameter is "Num," which can range from 0 to 4. A value of 0 signifies the absence of the disease, while other values indicate various degrees of disease presence. [3]

Features	Description
Age	Age
Ca	This feature describes the number of major blood vessels (ranging from 0 to 3) that were colored by a special X-ray imaging technique called fluoroscopy.
Chol (mg/dl)	Serum Cholesterol
Cp	Chest Pain type
Exang	Exercise-induced angina
Fbs	Fasting blood sugar
Num	Diagnosis of heart disease
Oldpeak	ST depression induced by exercise relative to rest
Restecg	Resting electrocardiographic results
Sex	Gender
Slope	The slope of the peak exercise ST segment
Thal	3=normal; 6=fixed defect; 7= reversible defect
Thalach	Maximum heart rate achieved
TrestBPS(mmHg)	Resting Blood Pressure

Table 1 Attributes Used in experiment

B. Factors and Measurements Criteria

To evaluate the performance of the proposed system, three metrics, namely accuracy, precision, recall were used. The three metrics were computed based on the number of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) predictions. TP refers to the number of instances that were correctly predicted, FP represents the number of instances that were incorrectly predicted, TN represents the number of instances that were correctly predicted as not required, and FN refers to the number of instances that were incorrectly predicted as not required.

$Accuracy = TP+TN/TP+FP+TN+FN$

$Recall = TP/TP+FN$

$Precision = TP/\ TP+FP$

The success of the heart disease prediction system is expressed using the terms Accuracy, Precision, Recall. Utilizing merely accuracy can occasionally be deceptive. In certain cases, using a model with lower accuracy is preferable since it offers a more reliable prediction for the issue. When there is a significant class imbalance in the issue domain, the model can anticipate all predictions as the value of the dominant class. In order to obtain findings that are more accurate, we favor the three different factors. [3]

C. Feature Selection Algorithm

Feature selection is a crucial step in the development of machine learning models. It involves identifying and selecting the most important features from the input data that have the greatest impact on the model's performance. There are various algorithms and techniques available for feature selection, including filter methods, wrapper methods, and embedded methods. These algorithms evaluate the relevance and redundancy of each feature and select the optimal subset of features for the model. The use of feature selection algorithms not only improves the performance of machine learning models but also reduces their computational complexity and training time. In proposed model we use lasso feature selection for high efficient model.

To choose the most pertinent features in a dataset, the regularisation technique known as lasso feature selection is frequently employed in machine learning. It is a linear regression model with a penalty term that is added to the cost function, resulting in decreased coefficient values for unimportant features. By choosing the coefficients with non-zero

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 9, May 2023

values, the Lasso algorithm can be used to find the most important features. This technique can help to reduce overfitting and enhance the performance of the model in high-dimensional datasets where there may be a lot of characteristics that are irrelevant to the model.[5]

D. Supervised Machine Learning Algorithms

Supervised machine learning algorithms are a set of methods used to train models that learn from labeled data to make predictions or decisions on new, unseen data. These algorithms use labeled examples to infer relationships between input features and output variables. The labeled data acts as a teacher to the model, enabling it to learn how to map inputs to outputs. Popular supervised machine learning algorithms include regression, decision trees, random forests, support vector machines (SVM), and artificial neural networks. These algorithms are widely used in various applications such as image recognition, speech recognition, natural language processing, and many more. In proposed model six supervised algorithms are used to find more efficient algorithm which fits more to proposed algorithm. After comparing all the algorithms the best algorithm can be used by developers to implement and predict the probability of heart disease.

IV. RESULTS AND GRAPHS

In this proposed model six algorithms are implemented and the factors like precision, accuracy and recall are calculated and plotted. The graph below shows three parameters for every algorithm.

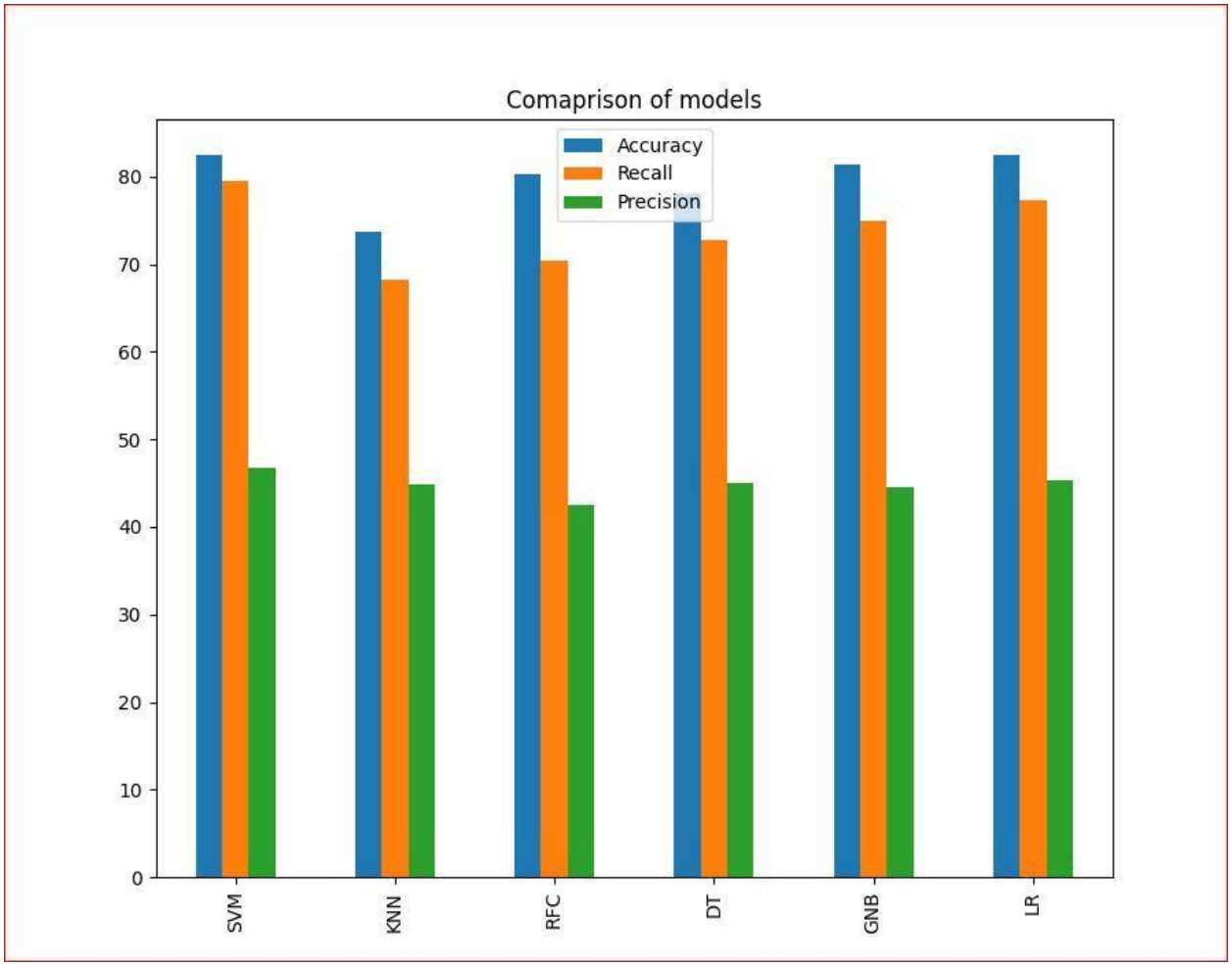

Fig 2 Comparison of all algorithm

From the above graph we can tabulate the accuracy for each algorithm and find the best and suitable algorithm.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 9, May 2023

Algorithms	Accuracy	Recall	Precision
Support Vector Machine	82.4%	79.4	46.67
K-Nearest Neighbour	73.6%	68.18	44.77
Random Forest	80.2%	70.45	42.46
Decision Tree	78.0%	72.72	45.07
Naïve Bayes	81.3%	75	44.95
Logistic Regression	82.4%	77.27	445.33

Table 2 Comparing accuracy of algorithms

From the above it can be concluded that Support Vector machine performs more efficiently with our proposed model. The Accuracy is around 82% with recall and precision are 79.% and 46.67%. So SVM improved the efficiency of prediction using Lasso Feature Selection model and classification and regression technique.

V. CONCLUSION

Support Vector Machine are an ensemble learning method for classification and regression techniques. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. So accuracy of prediction at early stages is achieved effectively. Identifying the processing of raw healthcare data of heart information will help in the long term saving of human lives and early detection of abnormalities in heart conditions. Machine learning techniques were used in this work to process raw data and provide a new and novel discernment towards heart disease. However, the Casualties can be reduced if the disease is detected at the early stages and preventative measures are adopted as soon as possible. The potential benefits of applying Machine learning methods with the suitable algorithm will help people be more cautious and preventive measure can be taken at earliest.

VI. FUTURE SCOPE

The careful selection and implementation of mining techniques on the dataset is crucial for achieving a fast and accurate system for heart disease management. In the future, ML applications using various algorithms will continue to improve disease prediction and diagnosis, as well as other fields of bioinformatics. This project's success highlights the potential of data science in healthcare and the need for further research in this area. By incorporating advanced ML algorithms and techniques, we can continue to improve the accuracy and speed of disease prediction and diagnosis, ultimately improving patient outcomes. More Improvements in future technology can help to identify the type of disease and predict with more specification and accuracy.

REFERENCES

[1]. Bo Jin, Chao Chi, Zen Liu, Shulong Zhang, Xiaomeng Yin, Xiaopeng Wi , "Predicting The Risk of Heart Failure With EHR Sequential Data Modeling", IEEE Access,(6:9256-9261**),** January, 2018.

[2]. Edward Choe, Andy Schuetz, Walter F Stewart, Jimeng Sun,"Using recurrent neural network models for early detection of heart failure onset", Journal of the American Medical Informatics Association(24(2):361-370),March,2017.

[3]. AKINROTIMI Akinyemi, MABAYOJE Modinat," A Machine Learning Approach to Diagnosing Heart Diseases", Journal of Computer Science and Control Systems(14(1):5-11),may,2021

[4]. Raunak Verma , Shashank Tandon , Mr. Vinayak," Heart Disease Prediction using Machine Learning", International Journal for Research in Applied Science & Engineering Technology(10(5):1872-1876), May,2022.

[5]. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.