Heart disease prediction using Machine learning

Konyala Sai Vardhan Reddy
700745733
CS Department
University of Central Missouri

Mythresh Maddina
700741162
CS Department
University of Central Missouri

Eda Sai Akhil
700747481
CS Department
University of Central Missouri

Stuthi Geetha Muppalla
700745858
CS Department
University of Central Missouri

Abstract— A crucial component of the human body is the heart. Heart, most vital and delicate human bodily organ. The heart's ability to beat determines whether a person will live or die. It performs various roles in our bodies, making its preservation crucial because many illnesses, including heart disease, are linked to it. The heart's job is to pressurize the circulatory system's blood arteries. The heart's proper operation is necessary for human survival.

According to multiple polls, heart disease is one of the major causes of death around the globe, regardless of a person's gender. The health care industry generates a large amount of data; thus it is necessary to process this data using any cutting-edge methods that will help us produce efficient outcomes, make efficient decisions based on the data, and obtain the desired outcomes.

Using machine learning methods including Gaussian Naive Bayes, Random Forest, K-Nearest Neighbor, Support Vector Machine, and Xg-Boost, an effective framework for heart disease prediction is constructed in this paper. 13 features are used by the framework, including age, gender, blood pressure, cholesterol, obesity, and cp. We have some steps in this user-friendly process.

The dataset file is uploaded, and the algorithm to apply to the chosen dataset is chosen, in the first stage. As a result of training the dataset on the method with the highest frequency, the modal is formed. Next, the accuracy is predicted for each chosen algorithm along with a graph. The sick stage of the heart is predicted during the following phase, which involves providing input for each heart parameter and basing the modal derived on that information. According to the patient's state, we subsequently take the necessary safeguards. We can successfully predict a victim's heart disease using our method. The Heart Disease Prediction Framework that has been developed in this view is a one-of-a-kind methodology that can be used within the class of heart disease.

Keywords: Heart diseases, feature extraction, machine learning techniques, Logistic Regression, Decision tree, Random Forest, Hybrid Model.

I. INTRODUCTION

Our project is upon detecting the heart diseases using machine learning algorithms which helps in early diagnosis and take precautions that helps the individual to remain healthy. Machine Learning techniques are used in solving the complex problems that are difficult to determine by human.

Usage of Machine Learning in the field of medical industry is increasing drastically and can reduce manual error and python provides the libraries which has data mining techniques get large amounts of data as input and process them.

Challenges:

- Factors caused by heart diseases: There are large number of factors that affect the heart such as the body mass index, type of food, medical history thalassemia, blood pressure etc.

Github Link: https://github.com/vardhan141/machine-learning-project

- Time-consuming and expensive traditional methods: Traditional methods of uses ECG and heavy equipment that and skilled personal is needed to determine the possibility of heart disease. In addition, the equipment cost is expensive.
- Need for early detection: Early detection of heart disease detection is necessary to reduce the risk of fatal.
- Large volumes of data: The analysis of large volumes of data, medical dataset of all the parameters affecting can be challenging and requires specialized skills and tools.

The above challenges led to the project objectives of developing and implementing machine–learning algorithms for the heart disease detection.

Objectives:

- More Accuracy and efficient model
- Use the large amount of data.

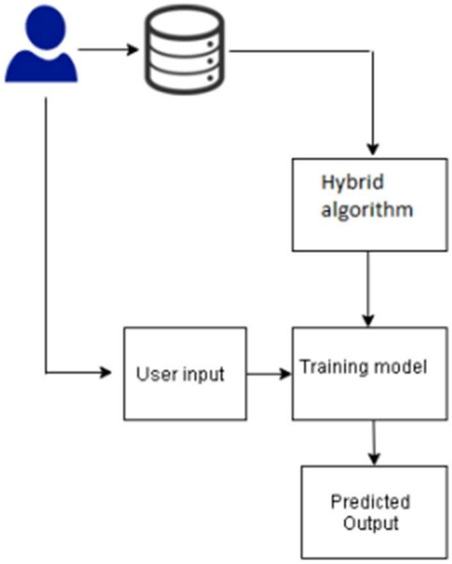

Fig 1: Block Diagram

Here the dataset is taken from UCI repository with 1024 instances and 72 attributes and them the dataset is preprocessed, where the dropping for null values are done and necessary features are extracted from the dataset and train the model with different ml algorithms and make a hybrid model which improve the efficiency further.

II. Motivation

Heart disease is a complex and multifactorial condition, which poses a significant challenge to healthcare providers globally. Despite significant progress in medical knowledge and technology, predicting the risk of heart disease remains a challenging task due to the heterogeneity of the disease and the complexity of the underlying risk factors.

Machine learning has demonstrated great potential in addressing this challenge by utilizing large-scale patient data to develop predictive models that can accurately predict the likelihood of heart disease in individuals.

These models utilize advanced algorithms such as deep learning, decision trees, and support vector machines to identify complex patterns and relationships in the data to develop accurate risk predictions.

Moreover, machine learning models can incorporate a vast array of patient-specific data, including demographic information, medical history, lifestyle factors, and genetic information. This enables the development of personalized risk assessments that can help clinicians tailor interventions to individual patients and optimize treatment outcomes.

. In addition, machine learning models can overcome limitations associated with traditional risk prediction methods, such as logistic regression and decision trees, which often have limited predictive power due to their simplistic assumptions about the data and the underlying relationships between risk factors and outcomes.

Overall, the development of a machine learning-based heart disease prediction system holds the potential to significantly enhance risk assessment and management for patients with cardiovascular disease. By providing accurate, personalized risk assessments and treatment recommendations, these models can help clinicians make more informed decisions, ultimately improving patient outcomes. Machine learning algorithms in healthcare can revolutionize the industry by providing advanced solutions to complex problems.

III. Main Contributions & Objectives

Action item	Team member	Description
Requirement Analysis	Geetha	Necessary libraries and IDE tools to implement, required ML Techniques
Project flow design	Sai Vardhan	Project flow is designed and delegated the work to each team member
Data collection and cleaning	Mythresh	Data is collected from the GitHub repository and cleaned the data accordingly
Data exploration-EDA	Vardhan, Akhil	Analyzed data on different factors

IV. Related Work

Machine learning is the part of artificial intelligence which helps humankind in prediction, classification and analysis of security threats, diseases, weather conditions etc... These technologies can be used in the medical field in detection of complex and life-threatening diseases such as cancer detection, heart disease prediction in early stages and decreases manual error and improve accuracy. Some of the risk factors that affect the heart are diabetes, high blood pressure, more cholesterol, the physical activity done by the individual. Here the dataset is obtained from UCI repository, since most of the research works use this dataset, we also use the data with 14 attributes and 1024 instances.

Some of the commonly used machine learning algorithms in the field of disease prediction of [1] classification algorithms such as support vector machine(SVM), decision tree and naïve bayes, similarly regression algorithms such as lasso regression, logistic regression and Random forest is also used. In this article, the author utilizes decision tree, random forest and hybrid model of the decision tree and random forest which attains an accuracy of 79%, 81% and 88% which can be concluded the hybrid models attain more accuracy. In article [2] the author demonstrates data mining technique with a feature selection of chi Square and genetic algorithms for classification and analyzed the factors that affect the heart diseases.

Data mining and machine learning accepts the large amount of data which are preprocessed for removal of unwanted data and fed to the classification algorithms such as xg-Boost which uses a gradient boosting framework of 304 records along with KNN, SVM, Random Forest and model is trained on dataset[3].Classification rule mining is an important tasks in data mining in which PSO(Particle Swarm Optimization) is used as classification rule with 10 fold approach by the author in article [4] and compared against decision tree. In [5] article the author proposes the rule-based algorithms where they used Classical prepositional logic(C-rules) and this provides combination of SVM, logistic regression and decision tree. In [6] the author displays the ROC area and accuracies of SVM, Random Forest, Gaussian Naïve Bayes and Decision tree with accuracies of 99.5, 99.7 85.1 and 90.4 respectively.

In [7] the author used machine learning algorithms using sdknl and Jaccard index is used to improve the performance of the model. In [8] prediction of heart diseases is done using statistical model multiple linear regression model using c# as the programming language. In [9] review of heart disease prediction system using data mining of naïve bayes, decision list, K-NN and intelligent techniques of neural networks and concluded the neural network with offline training is a good for disease prediction in early stage and the classification accuracy can be improved by reduction in features.

The article [10] shows a framework is created with feature extraction, outlier detection using PCA. The feature subset uses a wrapper filter as a classifier and the performance is improved. In [11] the review on heart diseases prediction using ml and data analytics approaches using 26 different research papers with comparative analysis which concludes only marginal success is attainable with normal classifiers whereas, more accuracy attained by hybrid models. The author in [12] uses the deep learning neural network such as K-NN, SVM, Hyper-parameter optimization (Talos), where Hyper-parameter optimization (Talos) provide with an accuracy of 90.76%.In [13] R. Sharmila proposes the conceptual method to enhance the prediction of heart diseases using the data techniques of (SVM) where the SVM in parallel fashion provides more accuracy than sequential SVM with 85% and 82.35% accuracy respectively.

In [14], the author proposed modified k-means and Naïve bayes where the k-means is compared with other classifiers and 93% the disease is detected whereas, 89% the heart disease is not detected. In Naïve bayes, the output is resulted in probability of presence of heart disease.[15] The author concludes it is observed that accuracy is more when the data is pruned and properly cleaned dataset. The authors in [16] proposed different models such as Naïve bayes, Generalized Linear Model, Logistic Regression, Deep learning, Decision tree, Random Forest, Gradient Boosted Trees, SVM, VOTE, HRFLM(Hybrid RF and Linear Model) with accuracy of 88.4%.Here the evaluation metrics such as classification error, precision, F-measure, sensitivity, specificity is also compared.

In [17] the author concluded that the ex, cp, fbs, restecg, exang, oldpeak, slope, ca and thal are the most significant features in dataset and top three data mining models are VOTE, Naïve Bayes and SVM in the descending order of high accuracy with 87.41%, 86.76% and 86.50% respectively. The VOTE model proposed here is the combination of naïve bayes and Linear Regression hybrid model.

[18] In this article, the author proposes the modified differential equation with combination fuzzy ahp and feed-forward neural network works with reduced prediction time and better accuracy and the 9 critical values is more efficient than the 13 critical values.

In [19] the author predicted the heart disease using multi-layer perceptron(MLP) with an average precision value of 91%.In which 92% for yes and 91% for No.

In [20] article the hybrid model of Logistic Regression and MARS (Multivariate adaptive regression spline) with an accuracy of 83.33% and high error rate. The model MARS-ANN have 82% approximately but with less number of total number of errors.

In [21] article Jaymin uses Linear Model Tree (LMT), J48, Random Forest Algorithm are used in which J48 with error.

In conclusion, different algorithms are studied such as SVM, Naïve bayes, PCA, Logistic Regression, KNN, Decision tree, Random Forest and the hybrid models.

V. Proposed Framework

Data visualization: The first step we did was to visualize the data to understand the similarities between the data features by using univariate analysis and bivariate analysis which comprises of histogram, barcharts, correlation matrix to find out if features have high correlation to delete one of them

Model Selection: The model selection is very important here for baseline performance we used logistic regression since this a binary classification problem followed by naïve bayes,k-nearest neighbours,support vector machine,random forest,hybrid model which is the combination of decision tree and random forest. The selection of the algorithm depends on the size of the dataset, the complexity of the problem, and the availability of computational resources.

Model training: The selection of the algorithm depends on the size of the dataset, here we divide the data into 80% train and 20% test. The training data consists of all the features required to predict the output class if the heart disease Is present or not

Model Evaluation: Once the model is trained, it is evaluated using a test dataset that is separate from the training dataset. The evaluation metrics may include accuracy, precision, recall, and F1-score here accuracy is the main metric for our project.

Seven different types of machine learning models are used:

1. Logistic regression
2. K nearest neighbors
3. Support vector machines
4. Decision tree
5. Random Forest
6. Naïve bayes
7. Hybrid

Logistic regression is a statistical method used to model the relationship between a binary dependent variable (i.e., a variable that can take on only two values) and one or more independent variables. It is a type of regression analysis that is commonly used in machine learning, statistics, and other fields to model the probability of a certain event occurring.

K-Nearest Neighbors (KNN) is a simple and popular machine learning algorithm used for classification and regression tasks. It is a non-parametric method, which means it doesn't make any assumptions about the underlying data distribution.

Support Vector Machines (SVMs) are a type of supervised learning algorithm used for classification and regression analysis. SVMs find a hyperplane (i.e., a decision boundary) that separates the input data into different classes. The goal of SVM is to find the hyperplane that maximally separates the data points of different classes.

The decision tree method is one of the most used machine learning classification techniques, which picks the best root and derives the data set into various partitions. By splitting and iterating over the data set we find the entropy measures, When the entropy is zero then the instance is the same class.

Name	Accuracy score
Logistic Regression	0.867
Naive Bayes	0.85
Decision tree	0.7167
SVM	0.767
Hybrid model	0.883
Random Forest	0.85
K Nearest Neighbours	0.733

$$E = Xc \; i=1 \; -pi \; log2 \; pi$$

Information gain is used to select the best attribute which helps to choose the next best variable attribute.

$$Gain\ (S, A) = Entropy(S) - X \; |Sv| \; |S| \; Entropy\ (Sv)$$

Naïve Bayes classifier is used to find the best hypothesis, it’s a probabilistic algorithm.

$$y\hat{} = argmaxP(y) \; Yn \; i=1 \; (P(xi\ |y))$$

Hybrid model uses combination of both decision tree and random forest by using combined weights we predict the output to increase the performance of model.

Cross-validation is a technique implemented when we have a limited data set. It is a technique for evaluating machine learning models by training several models on subsets of the available input data and evaluating them on the complementary subset of the data.

The logical reason for using machine learning algorithms for heart disease detection is that they can learn from large amounts of data, identify complex patterns, and make accurate predictions.

Traditional methods of disease detection involve visual inspections by experts, which can be time-consuming, expensive, and prone to human error. Machine learning algorithms can automate the process and provide real-time detection of heart diseases, allowing medical experts to take immediate action to prevent risks associated to heart disease.

Data Description:

The dataset for the project was collected from the open-source repository Kaggle. It contains of total 76 features below are the features I deemed important.

1. age: Number of years a person has lived
2. sex: Gender of patient (Male:1/Female:0)
3. Cp: Chest Pain type (4 values)
4. Trestbps: Resting Blood Pressure
5. Chol: serum cholestoral in mg/dl
6. Fbs:Fasting Blood Sugar > 120 mg/dl
7. Restecg: Resting Electrocardiographic (ECG) results (values 0,1,2)
8. Thalach:Maximum Heart Rate Achieved
9. Exang:Exercise Induced Angina
10. Oldpeak:oldpeak = ST depression induced by exercise relative to rest
11. Slope:the slope of the peak exercise ST segment
12. Ca: number of major vessels (0-3) colored by flourosopy
13. Thal:Thalium stress

Target variable: -

Condition: diagnosis of heart disease (angiographic disease status)

Value 0: < 50% diameter narrowing (negative for disease)

Value 1: > 50% diameter narrowing (positive for disease)

VII. Results and Conclusion

The dataset was preprocessed. Next, the model is trained and undergoes logistic regression, KNN, decision tree classifiers, random forest classifiers, naive Bayes, and SVM and hybrid model. Experiment analysis led us to conclude that the hybrid classifier provides higher accuracy than the alternative classification approaches as shown in the research paper, so we have successfully implemented the research paper.

We have built the confusion matrix for the hybrid algorithm to analyze its efficacy. A confusion matrix is a tabular representation of a classifier's accuracy and error rates. It is a metric for gauging how well a classification model does its job. To measure the efficacy of a machine learning model, we can look at its performance indicators, which can be found in a report called a classification report.

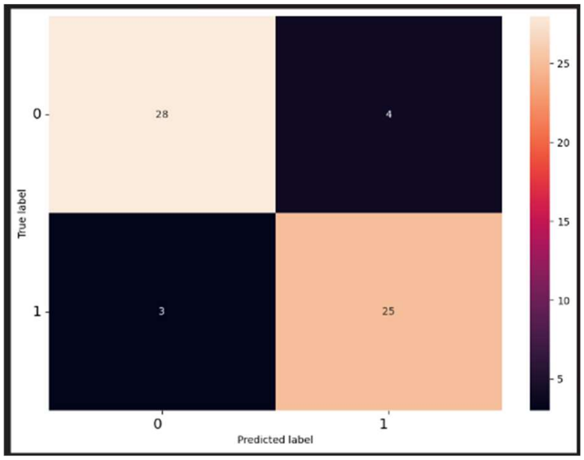

Fig: Confusion Matrix

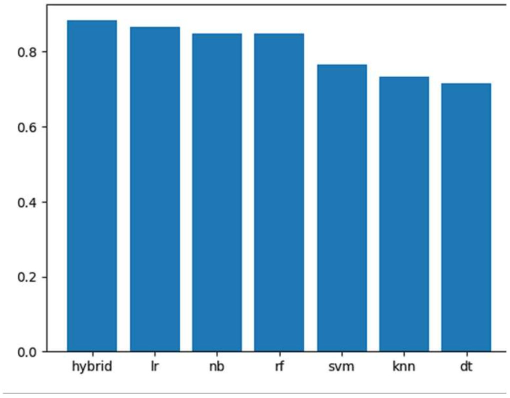

Fig: Comparison of accuracy scores

REFERENCES

[1] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2021, pp. 1329-1333, doi: 10.1109/ICICT50816.2021.9358597.

[2] Jabbar, M. A., B. L. Deekshatulu, and Priti Chandra. "Intelligent heart disease prediction system using random forest and evolutionary approach." Journal of Network and Innovative Computing 4.2016 (2016): 175-184.

[3] S. Farzana and D. Veeraiah, "Dynamic Heart Disease Prediction using Multi-Machine Learning Techniques," 2020 5th International Conference on Computing, Communication and Security (ICCCS), Patna, India, 2020, pp. 1-5, doi: 10.1109/ICCCS49678.2020.9277165.

[4] Alkeshuosh, Azhar Hussein, et al. "Using PSO algorithm for producing best rules in diagnosis of heart disease." 2017 international conference on computer and applications (ICCA). IEEE, 2017.

[5] Mythili, T., Dev Mukherji, Nikita Padalia and Abhiram Naidu. “A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL).” International Journal of Computer Applications 68 (2013): 11-15.

[6] V. Sharma, S. Yadav and M. Gupta, "Heart Disease Prediction using Machine Learning Techniques," 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2020, pp. 177-181, doi: 10.1109/ICACCCN51052.2020.9362842.

[7] G. Choudhary and S. Narayan Singh, "Prediction of Heart Disease using Machine Learning Algorithms," 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), Bengaluru, India, 2020, pp. 197-202, doi: 10.1109/ICSTCEE49637.2020.9276802.

[8] K. Polaraju, D. Durga Prasad, Prediction of Heart Disease using Multiple Linear Regression Model, IJEDR,vol 5, ISSN:2321-9939, 2017

[9] R. Kavitha and E. Kannan, "An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining," 2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS), Pudukkottai, India, 2016, pp. 1-5, doi: 10.1109/ICETETS.2016.7603000.

[10] Mai Shouman, Tim Turner, Rob Stocker 2012 “Using Data Mining Techniques In Heart Disease Diagnoses And Treatment“ Electronics, Communications and Computers (JECECC), 2012 Japan-Egypt Conference March 2012, pp 173-177.

[11] M. Marimuthu, M. Abinaya, K.S. Hariesh, K. Madhankumar, A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach, International Journal of Computer Applications, Vol 181- No. 18, September 2018.

[12] Heart Diseases Prediction using Deep Learning Neural Network Model sumit sharma. DOI: 10.35940/ijitee.C9009.019320

[13] R. Sharmila, S. Chellammal, “A conceptual method to enhance the prediction of heart diseases using the data techniques”, International Journal of Computer Science and Engineering, May 2018.

[14] Sairabi H.Mujawar, P.R.Devale, “Prediction of Heart Disease using Modified K-means and by using Naïve Bayes”, International Journal of Innovative research in Computer and Communication Engineering, vol.3, October 2015, pp.10265-10273.

[15] Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee and Asmita Mukherjee Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techiniques: A Review, ISSN vol 10, No. 7, pp. 2137- 2159.

[16] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." IEEE Access 7 (2019): 81542-81554.

[17] Mohammad Shafenoor Amin, Yin Kia Chiam, Kasturi Dewi Varathan,Identification of significant features and data mining techniques in predicting heart disease, Telematics and Informatics, Volume 36,2019,Pages 82-93,ISSN 0736-5853,https://doi.org/10.1016/j.tele.2018.11.007.

[18] T. Vivekanandan and N. C. S. N. Iyengar, ‘‘Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease,’’ Comput. Biol. Med., vol. 90, pp. 125–136, Nov. 2017.

[19] A. Gavhane, G. Kokkula, I. Pandya and K. Devadkar, "Prediction of Heart Disease Using Machine Learning," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2018, pp. 1275-1278, doi: 10.1109/ICECA.2018.8474922.

[20] Y. E. Shao, C.-D. Hou, and C.-C. Chiu, ‘‘Hybrid intelligent modeling schemes for heart disease classification,’’ Appl. Soft Comput. J., vol. 14, pp. 47–52, Jan. 2014. doi: 10.1016/j.asoc.2013.09.020

[21] Jaymin Patel, Prof. Teja;Upadhyay, Dr.Samir Patel, Heart Disease Prediction using Machine Learning and Data Mining Technique, IJCSC, vol 7, pp- 129- 137.

[22] johnsmith88. (2019,June). Heart Disease Dataset, Version 1. Retrieved February 23, 2023 from https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset?select=heart.csv.

[23] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.