IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

UTILIZING MACHINE LEARNING TECHNIQUES FOR HEART RISK PREDICTION

Om Deshmukh[1], Fardeen Kachawa[1], Sujal Bhatt[1], Kaif Siddique[1], Neelam Phadnis[2]
[1]Student, [2]Assistant Professor

[1] Final year Computer Engineering Department, Shree L.R. Tiwari College of Engineering, Thane, India
[2]Assistant Professor, Computer Engineering Department, Shree L.R. Tiwari College of Engineering, Thane, India

Abstract: Heart disease is a prominent cause of death among people. These diseases are preventable if treated in time. We propose a system that can allow users to assess their heart health and ensure they realise when to visit the doctor. We utilise Machine Learning algorithms prominently to determine the risk of heart diseases to a particular user. We use Classification algorithms for the prediction as they offer the best result as well as fit the use case. Data mining is crucial in feature selection to build better models. We use the UCI Cleveland dataset of heart disease to train our models.

Index Terms: Machine Learning, Classification Algorithm, Data Mining, feature Selection

I. INTRODUCTION

The rise in cardiovascular diseases has been significant in the world. As stated by the Centers for disease control and prevention, around 695,000 people in the United States of America have passed away due to heart diseases in 2021 [1]. Heart attacks are one of the leading causes of death in the world as per World Health Organisation [2]. Treatment for heart disease is essential because there are more and more cases indicating that this condition is becoming more common among people, especially in developing countries like India [3] [20]. In 2000, there were an estimated 29.8 million people suffering from coronary heart disease (CHD) which equates to 3% at that time [4]. In 2003, that number increased significantly as the CHD affected population number increased to 5% [5]. One of the ways to curb this issue is to utilise Machine Learning algorithms for prediction of heart risks for people and taking necessary precautions.

II. BACKGROUND

Cardiovascular diseases affecting people can be controlled with proper treatment [9]. Many treatments for these diseases require detection of these diseases in early stages. When treated in time, most of these diseases can be fully cured. Some of the most common types of heart diseases can be categorised in following categories as stated below [6]:

A. Ischemic heart disease

These are also known as "coronary heart disease" or "coronary artery disease," which are caused by the blocked coronary arteries that deliver blood to the heart muscle [7].

B. Cerebrovascular Disease

A stroke or cerebrovascular accident (CVA) is a sudden worsening of the cerebral perfusion or vasculature that is directly connected to the cardiovascular network. Approximately 85% of the strokes are ischemic in nature and the rest are hemorrhagic in nature [8].

C. Rheumatic Heart Disease

Rheumatic heart disease is a systemic immunological syndrome that can result from untreated rheumatic fever, an inflammatory illness that can arise from strep throat or scarlet fever. Rheumatic heart disease is a severe kind of acquired heart disease that affects people all over the world, including adults and children [9].

D. Hypertensive heart disease

Hypertensive heart disease is the term for the anatomical and functional abnormalities that result from persistently high blood pressure. The left ventricle, left atrium, and coronary arteries are all impacted by these changes [8].

E. Cardiomyopathy

Cardiomyopathy is the term used to describe anatomical and pathological heart muscle or electrical dysfunction. Cardiomyopathies are a broad category of diseases with high rates of morbidity and mortality that frequently lead to progressive heart failure [2].

F. Atrial fibrillation

Atrial fibrillation (AF), the most prevalent cardiac arrhythmia, affects 1% to 2% of the general population. This syndrome can be identified by an EKG with irregular QRS complexes and no P-wave. It is characterised by rapid and chaotic atrial activation leading to impaired atrial function [10].

III. FEATURE SELECTION

Data mining plays a crucial role in determining the factors that have a high impact on the occurrence of heart disease among people [11].

Data cleaning is the main part of pre-processing which is classified into two main categories:

i) Removal of duplicate data:

A large number of repeatable instances made up of data values in a given dataset. We have to get rid of them in these situations since they can cause inconsistencies in our training model. To make sure that only the relevant data may be selected for the model, we create a number of rules for the data values. Only the data values that follow these rules are used further; all other data values are regarded as errors and removed from the dataset [12] [13].

ii) Error Repairing:

There are many other ways to remedy errors, but one popular technique that is still in use today is handily correcting the wrong data value. Making modifications to the databases is another way to automate this entire procedure and complete the work [13].

With these techniques we simplify the dataset by eliminating the outliers and duplicates. In the heart disease dataset released by UCI Cleveland, we utilise the following parameters:

1) age in years
2) Sex (1 = male, 0 = female)
3) chest pain type
4) resting blood pressure
5) cholesterol
6) fasting blood sugar
7) resting electrocardiographic result
8) maximum heart rate achieved
9) exercise included angina(exhang)
10) old peak
11) slope
12) number of major vessels

These parameters are based on the feature importance before creating a machine learning algorithm. This can be visualised as shown in the figure below.

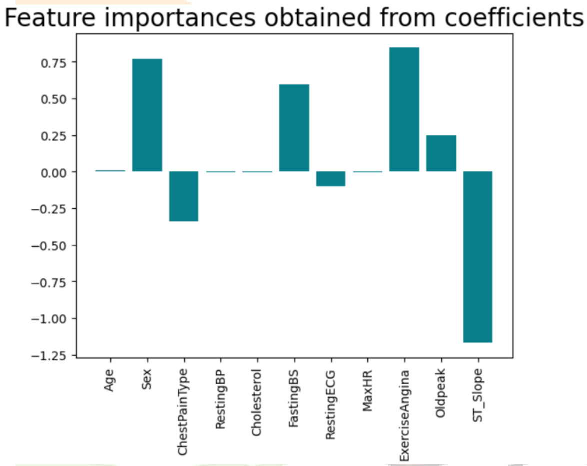

Fig 1. Feature Importance

Learning a function to categorise data items to a subset of a given class set is the process of classification. Among the first kinds of tasks in the classification is to find a reasonable generic that can forecast the class of unknown data items with sufficient accuracy. Finding a simplified and easy-to-understand class model for every class is the next stage [14].

IV. LITERATURE REVIEW

Due to the rise of heart diseases in the world, many researchers stepped forward to find a way to predict these risks using existing algorithms at times with certain extensions.

The authors [14] used the UCI heart disease dataset to research and apply several machine learning algorithms. After data cleaning, which included deleting all incomplete records, the logistic regression technique produced an accuracy of 75.41%.

The naive bayes algorithm was devised and implemented by Vembandasamy K., Sasipriya R., and Deepa E. [15] using the database of one of the top diabetes research institutes in Chennai, which obtained them with a 74% accuracy rate.

A 2005 study by R. Gupta [4], "Burden of coronary heart disease in implementing the decision tree algorithm, namely the j48 and India," Portal Regional da BVS, vol. 57, no. 6, bagging method type of the decision tree algorithm, was conducted by Mai Shouman, Tim Turner, and Rob Stocker [13].

The Cleveland Clinic Foundation Heart Disease data set achieved accuracy of 81.41% on the bagging method and 78.9% on the j48 [9] C. Dass and A. Kanmanthareddy, "Rheumatic Heart Disease," StatPearls Publishing.

A hybrid machine learning model consisting of a random forest with chi squared approach and genetic algorithm was built by M.A. Babbar, B.L. Deekshatulu, and Priti Chandra [16] and applied to a dataset of patients with heart disease at corporate hospitals. The model yielded an accuracy of 83.70%.

Following feature selection based on information gain, Anuradha.P. and Dr. Vasantha Kalyani David [17] applied the XGgradient boosting algorithm on various Cleveland datasets, achieving an accuracy of 88.52%.

V. PROPOSED SYSTEM

After preparing the dataset from the Cleveland heart disease dataset, we implement certain machine learning algorithms to classify and test the accuracy of each algorithm.

The workflow of implementation of our system can be represented with the figure below.

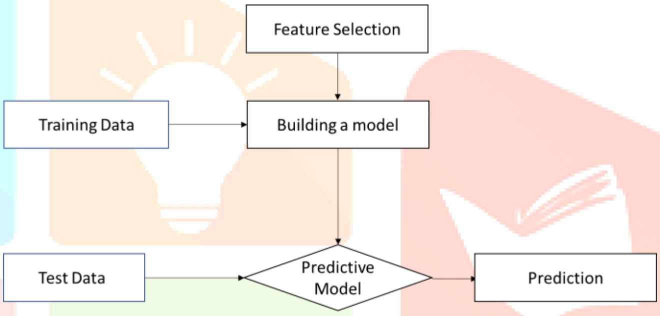

Fig 2. Design of System

We tested the model using Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Decision Tree, Random Forest and Categorical Boosting (CatBoost) machine learning algorithms.

a. Logistic Regression:

We utilise the logistic regression algorithm with the one-vs-rest (OvR) scheme that would allow us to use the multiple parameters present in the dataset for making a binary classification on whether a person has heart risk or not [18]. Mathematically, the sigmoid function for logistic regression can be represented as follows

$$\operatorname{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

b. Support Vector Machine (SVM) :

We use the Support Vector Machine learning algorithm that uses a subset of training points for every parameter making it memory efficient. Since SVM does not support multiclass classification, we break down the problem into multiple binary classification allowing us to predict the risk for each parameter. For this, we utilise the scikit-learn library's LinearSVC classifier which can be represented mathematically as follows [20]

$$\min_{w,b} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \max(0, 1 - y_i(w^T \phi(x_i) + b)),$$

c. K-Nearest Neighbours (KNN):

In this, we use the provided number of values in the dataset in which the machine learning algorithm will find the distance from each point and classify it into clusters. This algorithm is a type of unsupervised learning and hence does not require any training beforehand [18]. Mathematically, we can formulate it as follows

$$\arg\max_{L} \sum_{i=0}^{N-1} p_i$$

Where,

$$p_i = \sum_{j \in C_i} p_{ij}$$

,

$$p_{ij} = \frac{\exp(-||Lx_i - Lx_j||^2)}{\sum_{k \neq i} \exp -(||Lx_i - Lx_k||^2)}, \quad p_{ii} = 0$$

d. Decision Tree:

Decision Trees are a non-parametric supervised learning method which can be used for multiclass classification. We use the CART decision tree algorithm provided by scikit-learn for the classification. mathematically, we represent it as

$$p_{mk} = \frac{1}{n_m} \sum_{y \in Q_m} I(y = k)$$

The above formula is used to produce the output while the formula below is used to calculate the gini index

$$H(Q_m) = \sum_{k} p_{mk}(1 - p_{mk})$$

e. Random Forest:

Random forest is an estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy. It can also reduce the problem of overfitting. For our approach, we use 100 estimators that allowed us the best accuracy out of all the models implemented [18]. mathematically, we formulate it as

$$RFfi_i = \frac{\sum_{j \in all\ trees} normfi_{ij}}{T}$$

Where, norm fi is the normalised feature of the decision tree

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ nodes} ni_k}$$

$$normfi_i = \frac{fi_i}{\sum_{j \in all\ features} fi_j}$$

f. Categorical Boosting (CatBoost):

Categorical Boosting utilises the approach similar to gradient boosting without the need of one hot encoding as compared to other gradient boosting algorithms. This makes CatBoost much more robust and thus easier to implement. It requires the selection of optimal learning rate, depth of tree and number of iterations to avoid problems like overfitting and underfitting [19].

VI. RESULT

We use the following metrics to evaluate the performance of our models.

$$\texttt{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$

$$\text{precision} = \frac{tp}{tp + fp}$$

$$\text{recall} = \frac{tp}{tp + fn},$$

$$F1\ score = 2\left(\frac{1}{\frac{1}{Precision} + \frac{1}{Sensitivity}}\right)$$

The F-1 Score of the model is represented by the figure below

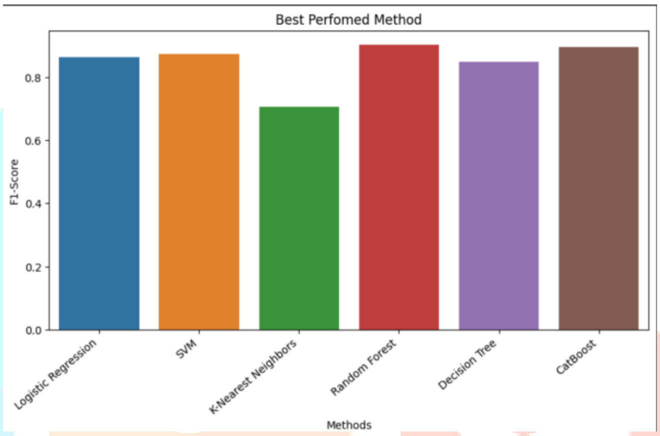

Fig 3. F-1 Score

The Recall can be represented as

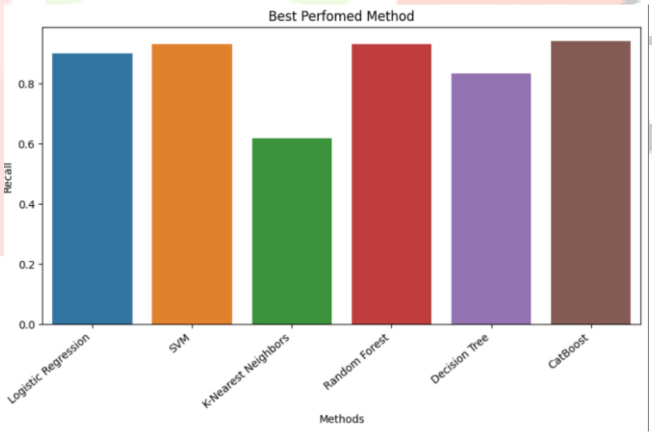

Fig 4. Recall

The accuracy of the models can be represented in the figure below

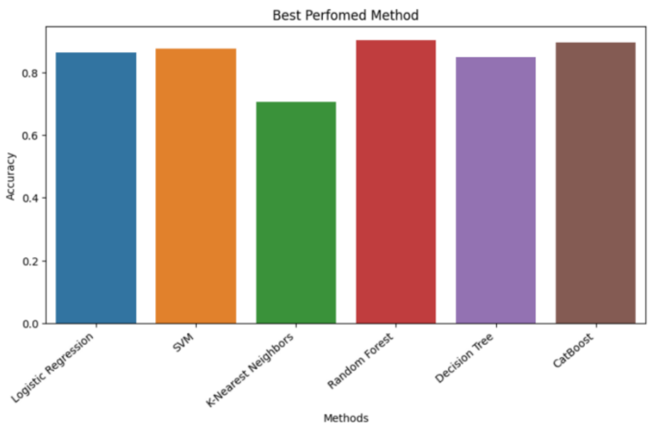

Fig 5. Accuracy

The table below displays the F-1 Score, Accuracy and Recall of every model tested.

Table 1. Model Evaluation

Models	F1-Score	Accuracy	Recall
Logistic Regression	0.863619	0.864130	0.901961
SVM	0.874012	0.875000	0.931373
K-Nearest Neighbours	0.705689	0.706522	0.617647
Random Forest	0.901892	0.902174	0.931373
Decision Tree	0.848206	0.847826	0.833333
CatBoost	0.896150	0.896739	0.941176

VII. CONCLUSION

In conclusion, the development and implementation of a heart risk prediction system using machine learning holds tremendous promise for revolutionising preventive healthcare. By leveraging advanced algorithms and analysing vast datasets, this system can provide accurate and personalised predictions regarding an individual's risk of cardiovascular events. The integration of machine learning techniques allows for the identification of subtle patterns and correlations that may elude traditional risk assessment methods. As per studies of research papers and implementing all the algorithms, random forest provides the highest accuracy of 90.21% as compared to other algorithms.

VIII. REFERENCES

[1] "Centers for Disease Control and Prevention," U.S. Department of Health & Human Services, 2023. [Online]. Available: https://www.cdc.gov/heartdisease/facts.htm. [Accessed 30 November 2023].

[2] "Cardiovascular diseases," World Health Organization, [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1. [Accessed 30 November 2023].

[3] A. Roy, P. Jeemon and D. Prabhakaran, "Cardiovascular Diseases in India," vol. 133, no. 16, pp. 1605-1620, 18 April 2016.

[4] S. Chauhan and B. T. Aeri, "The rising incidence of cardiovascular diseases in India:," *European Journal of Preventive Cardiology*, vol. 4, no. 4, pp. 735-740, May 2015.

[5] R. Gupta, P. Joshi, V. Mohan, K. S. Reddy and S. Yusuf, "Epidemiology and causation of coronary heart disease and stroke in India," *BMJ Journals*, pp. 16-26, 2007.

[6] G. Tackling and M. B. Borhade, "Hypertensive Heart Disease," 26 June 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK539800/. [Accessed 30 November 2023].

[7] R. Wexler, T. Elton, A. Pleister and D. Feldman, "Cardiomyopathy: An Overview," 9 December 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2999879/. [Accessed 30 November 2023].

[8] K. AS and T. P, "Cerebrovascular Disease," *Europe PMC*, 7 August 2022.

[9] C. Dass and A. Kanmanthareddy, "Rheumatic Heart Disease," 25 July 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK538286/. [Accessed 30 November 2023].

[10] N. Jothi, N. A. Rashid and W. Husain, "Data Mining in Healthcare – A Review," *Procedia Computer Science*, vol. 72, pp. 306-313, 2015.

[11] X. Chu, I. F. Ilyas, S. Krishnan and J. Wang, "Data Cleaning: Overview and Emerging Challenges," in *SIGMOD '16: Proceedings of the 2016 International Conference on Management of Data*, New York, 2016.

[12] C. Sowmiya and P. Sumitra, "Analytical Study of Heart Disease Diagnosis," in *IEEE INTERNATIONAL CONFERENCE ON INTELLIGENT TECHNIQUES IN CONTROL, OPTIMIZATION AND SIGNAL PROCESSING*, 2017 .

[13] M. Shouman, T. Turner and R. Stocker, "Using Decision Tree for Diagnosing Heart Disease Patients," in *Proceedings of the 9-th Australasian Data Mining Conference (AusDM'11)*, Ballarat, 2011.

[14] C. Boukhatem, H. Y. Youssef and A. B. Nassif, "Heart Disease Prediction Using Machine Learning," in *Advances in Science and Engineering Technology International Conferences (ASET)*, Dubai, 2022.

[15] E. Deepa, R. Sasipriya and K. Vembandasamy, "Heart Diseases Detection Using Naive Bayes Algorithm," in *International Journal of Innovative Science, Engineering & Technology*, 2015.

[16] M. A. Jabbar, B. L. Deekshatulu and P. Chandra, "Intelligent heart disease prediction system using random forest and evolutionary approach," *Journal of Network and Innovative Computing*, vol. 4, pp. 175-184, 2016.

[17] A. P and V. K. David, "Feature Selection and Prediction of Heart diseases using Gradient Boosting Algorithms," in *Proceedings of the International Conference on Artificial Intelligence and Smart Systems*, Coimbatore, 2021.

[18] F. Pedregosa, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research 12 (2011) 2825-2830*, pp. 2825-2830.

[19] A. Gulin, "CatBoost," Yandex, [Online]. Available: https://catboost.ai/en/docs/. [Accessed 30 November 2023].

[20] R. Gupta, "Burden of coronary heart disease in India.," November 2005. [Online]. Available: https://pesquisa.bvsalud.org/portal/resource/pt/sea-5972. [Accessed 30 November 2023].