Cardiovascular Disease Prediction: Performance Analysis and Comparison of Various Supervised Machine Learning Algorithms

Ritika Garg

Chitkara University Institute of

Engineering & Technology, Chitkara University, Punjab, India

ritika0348.cse19@chitkara.edu.in

Pradeepta Kumar Sarangi

Chitkara University Institute of

Engineering & Technology, Chitkara University, Punjab, India

pradeeptasarangi@gmail.com

Ashok Kumar Sahoo

Graphic Era Hill University, Dehradun, India

ashoksahoo2000@yahoo.com

Jayant Jha

Presight, UAE

jayant.jha@presight.ai

2023 International Conference on IoT, Communication and Automation Technology (ICICAT) | 979-8-3503-0282-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICICAT57735.2023.10263671

***Abstract- C*ardiovascular disease also termed as heart disease includes many types of disorders related to heart. Heart disease kills more people annually than all other diseases combined, making it the leading cause of morbidity and mortality worldwide. The prevalence of heart disease is increasing at an alarming rate, making early illness prediction essential. The main objective of this research is to investigate the most suitable machine learning algorithm for heart disease prediction. Hospitals may now easily do automatic diagnoses since they are playing a crucial part in resolving these issues with heart disease forecasts thanks to data science algorithms and data mining approaches. This research paper is based on supervised learning methods like Logistic Regression (LR) and Gaussian Naïve Bayes (NB). This work also explores the applicability of other methods such as K-Nearest Neighbours (KNN), Decision Trees (DT) and Random Forest (RF). It displays numerous heart disease-related features. The findings show that accuracy produced by the Random Forest is 92% which is better than other models.**

Keywords- Cardiovascular disease, Morbidity, Naïve Bayes, K-Nearest Neighbour, Data Mining.

I. INTRODUCTION

One of the most hazardous illnesses afflecting people today is cardiac arrest (heart disease) and has been widespread since a long period. The main cause of death worldwide is cardiovascular disease (CVD), according to the WHO. One-third of Cardiovascular deaths occur before the age of 70, compared to strokes and cardiac arrests, which contribute up to four each of the five fatalities

Medical diagnosis is seen as an important but challenging process that must be completed well. This task's automation is quite useful. Regrettably, aside from some areas with a lack of resources, not all doctors are experts in every field. Finding hidden patterns and information that could aid in making wise decisions can be done through data mining. In order to make informed judgements and offer the public high-quality services, this is crucial for healthcare workers.

Based on the medical history, our research can determine who has a higher chance of being diagnosed with heart disease. It generates predictions based on variables including blood pressure, cholesterol, chest discomfort, and blood

sugar levels. This will help people better understand themselves in advance and enable them to take the required safety measures.

Machine learning applications have shown promising results in various fields of the health care sector [1, 15, 16]. Some notable application areas are cancer detection [2], bone fracture detection [3] and tumour detection.

This work uses a variety of data science algorithms, like KNN, LR, DT and Bayesian Network, synthetic neural systems and Random Forest, to forecast the onset of cardiac arrest on the basis of different well-being indicators. Both KNN and logistic regression models were used to evaluate accuracy and investigate various models. For the purpose of predicting heart disease, we analysed data on individuals with and without the illness. To determine whether each patient was at risk for heart disease, 14 distinct aspects of them were looked at. The most efficient algorithm among them is RF, which offers an accuracy of 92 percent.

II. RELATED WORKS

A substantial amount of research on the application of data science techniques for CVD detection served as the basis for this effort. ML algorithms have been applied to create a number of accurate forecasts of cardiac disease. Using both the new and old data science models, the IHDPS model was able to predict a person's likelihood of acquiring cardiac attack rather precisely.

In their work, Kaur et al. [4] have used four different data science algorithms for forecasting cardiac attacks. The authors claim to achieve the highest accuracy of 95% from their experimental analysis.

Bhatt et al. [5] have proposed a method of k-modes

clustering to improve the classification accuracy. Using Kaggle data set, the authors report an accuracy of 87.28%.

focus on the importance of feature
selection in cardiovascular disease prognosis
different data mining algorithm .Using proner
selection technique. any classification algorithm
improved significantly.

Employing various data science algorithms, Sabab et al. [6] concentrated on the significance of feature selection in the therapy of cardiac disease prediction. Every classification method may be considerably enhanced by using the right attribute selection approach. In their research, they discovered that Bayesian produced the greatest results prior to attribute selection.

Ali at el. [7] in their work have reported the implementation of machine learning models for the prediction heart disease. The authors have used three machine learning algorithms such as KNN, DT and RF and claim to achieve 100% accuracy for the RF method.

In another work by Umair et al. [8], four tests were run in two separate circumstances, with the first case was using all characteristics and the second case was using only a subset of the attributes. The data set was in the Weka-compatible ARFF format. According to the trials, Bayesian classification algorithms had the best accuracy of all, with a score of 82.914%. This study demonstrates how efficiently and successfully cardiac disease may be predicted using data science.

Oliveira et al. [9] discussed the creation and testing of an NLP tool that can be used to classify and extract data from reports on the cytology and histology of the cervix and anus. Their NLP system can classify with an accuracy of 91% whether a cytology or histology specimen was abnormal and whether any HPV tests had shown positive results based on these preliminary data. This algorithm's average recall (also known as sensitivity) and accuracy (also known as positive predictive value) at the document level were 96% and 94%, respectively.

Using a public data collection that included details on 520 patients aged 16 to 90, Chaves at el. [10] applied standardization. The findings have demonstrated that neural networks are a useful tool for diagnosing diabetes in its early stages. Out of 520 cases, 510 were correctly predicted, a 98.1% accuracy rate. To prevent overfitting, a 10-fold bridge was used for the experimental validation.

Priyanka et al. [11] in their paper have applied Naïve Bayes & Decision tree classifiers for prediction of Heart disease. Finally, they have compared the efficiency & accuracy of the two techniques to decide the best.

Jindal et al. [12] in their work have suggested method for heart disease prediction. The authors have used the medical history of the patient. Finally, the authors have applied a KNN classifier to classify the patient with heart disease or not.

In another work, Mustaqeem et al. [13] have implemented machine learning model to identify the disease of a patient. Their model is able to classify the disease into four different classes including angina and myocardial infarction.

The paper by Boukhatem et al. [14] implements four classifiers such as MLP, SVM, RF and NB for heart disease prediction. Out of these four models, the authors report that SVM model performed better than others with 91.67% accuracy.

III. METHODOLOGY

The primary objective of the suggested technique is to anticipate the onset of heart disease in order to swiftly and accurately diagnose the condition. We are utilising a variety of data mining approaches and data science techniques, such as K - means, Logistic Regression, Decision Tree, Bayesian Network and Random Forest, to anticipate cardiac attack on the basis of wellness indicators. The three phases of the suggested approach are as follows: data gathering, feature extraction and data exploration. Data preparation deals with missing values, cleaning up the data, and normalising it, depending on the techniques employed. After post, the proposed techniques' classifier is employed to identify the post data. The recommended model was then examined in detail, and a range of benchmarks were used to gauge its efficacy and accuracy. 30% of the

total dataset is used to evaluate the models. The various attributes used are given in figure-1.

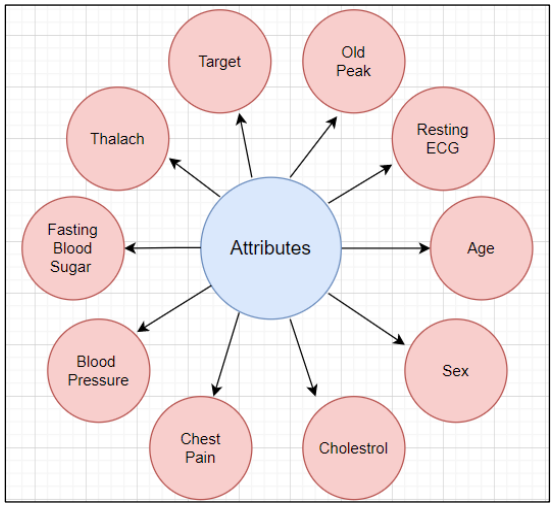

Figure 1: Attributes used for the dataset taken

IV. DATASET VISUALIZATION

For the model's training, Kaggle's dataset has been downloaded. Several instances were used to collect the data. The UCI repository is where the database is located. It originally contained 76 properties, but only a subset of 14 of those were actually used. The collection contains information on a wide range of people, including their history of heart disease and other illnesses. The medical histories of 403 distinct patients with a variety of features make up the dataset. The patient's medical features, including age, the type of chest discomfort experienced, blood pressure, sugar levels, angina, and other factors, are well-detailed in this dataset, allowing us to determine regardless of whether the person has indeed been identified with cardiac attack.

Training data and testing data are separated from the data. Data is used for testing 30% of the time, and training 70% of the time. For the purpose of removing Nan values, normalised the data.

	age	sex	cp	trestbps	chol
count	402.000000	402.000000	402.000000	402.000000	402.000000
mean	54.425373	0.699005	0.912935	129.927861	245.189055
std	8.816820	0.459262	1.033060	17.021250	52.880367
min	29.000000	0.000000	0.000000	94.000000	126.000000
25%	48.000000	0.000000	0.000000	120.000000	209.000000
50%	55.000000	1.000000	0.000000	128.500000	234.000000
75%	60.000000	1.000000	2.000000	140.000000	282.000000
max	77.000000	1.000000	3.000000	180.000000	564.000000

Figure 2: Statistical outline of attributes

This figure-2 displays the statistical summary of the subset properties for 403 cases. The count informs us of the number of nonempty rows in a feature. The value of "mean" denotes the average value of the attribute. Normative Deviation of the feature is represented by the value of "std". The "min" denotes the feature's lowest possible value. The percentile/quartile for each feature is 25%, 50%, and 75%. The word "max" denotes the attribute's maximum value. Statistical testing can be used to determine which links between the attributes and the performance measure are the strongest.

V. IMPLEMENTATION OF ALGORITHMS

The implementation design involves various techniques such as LR, RF, DT, Bayesian network and KNN. The next parts provide a detailed implementation and outcome analysis. The classification results are shown through a confusion matrix for each of the implementations as shown in figure-3.

Actual Positive Predicted Positive	Actual Negative Predicted Positive
Actual Positive Predicted Negative	Actual Negative Predicted Negative

Figure 3: Confusion matrix format

A. Implementation of Logistic Regression

Techniques for supervised learning like logistic regression are used to calculate the likelihood of categorization problems resulting in two probabilities. Another application is to forecast results for several classes. The formula used in our Logistic Regression model is (z)= 1/(1 + ez). Using this formula, which easily transforms any integer in a value ranging from zero to one, the chance of correctly classifying groups was computed. Those with and without cardiac disease, for instance, can be categorised into two groups. The confusion matrix for the predicted values is shown in figure-4.

Figure 4: Confusion Metrics of Logistic Regression

The statistical values such as precision, recall, f1-score, support and accuracy generated through the implementation are shown in the figure below.

	precision	recall	f1-score	support
0	0.88	0.85	0.87	62
1	0.85	0.88	0.87	59
accuracy			0.87	121
macro avg	0.87	0.87	0.87	121
weighted avg	0.87	0.87	0.87	121

Figure 5: Logistic Regression Model Results

B. Implementation of Decision Tree

A simple to comprehend supervised learning algorithm classifier is a decision tree. They manage numerical as well as categorical data. The core nodes, branches, and subtrees of the tree structure each reflect a specific data set's value, much like the branching structure of a tree. The class's ability to predict or convey the outcome is demonstrated by experiments on a certain attribute and its hidden layers. The categorization rule moves from the cluster head to the cluster members based on the expected characteristic and the specified rules. The confusion matrix for the predicted values is shown in figure-6.

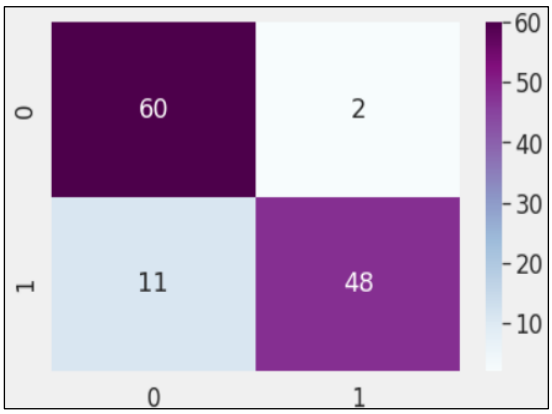

Figure 6: Confusion Metrics of Decision Tree

The statistical values such as precision, recall, f1-score, support and accuracy generated through the implementation are shown in the figure-7 below.

	precision	recall	f1-score	support
0	0.85	0.97	0.90	62
1	0.96	0.81	0.88	59
accuracy			0.89	121
macro avg	0.90	0.89	0.89	121
weighted avg	0.90	0.89	0.89	121

Figure 7: Decision Tree Model Result

C. Implementation of Random Forest

An alternative monitored learning approach is random forest. Classification as well as regression are compatible with its use. It is the algorithm that is also most user-friendly and versatile. A forest is created by trees. More trees are supposed to make a forest stronger. Utilizing data samples selected at random, random forests create decision trees, obtain predictions from each tree, and then cast votes for the best solution. The confusion matrix for the predicted values is shown in the figure-8.

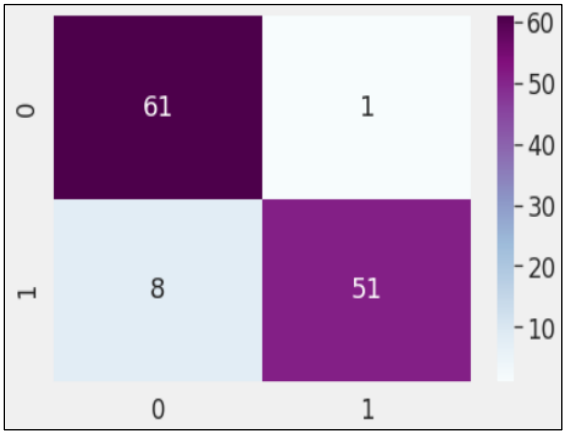

Figure 8: Confusion Metrics of Random Forest

The statistical values such as precision, recall, f1-score, support and accuracy generated through the implementation are shown in the figure-9 below.

	precision	recall	f1-score	support
0	0.88	0.98	0.93	62
1	0.98	0.86	0.92	59
accuracy			0.93	121
macro avg	0.93	0.92	0.93	121
weighted avg	0.93	0.93	0.93	121

Figure 9: Random Forest results

D. Implementation of Gaussian Naïve Bayes

Based on the Bayesian network, Naive Bayes serves to categorise data. The Naïve Bayesian classifier principle says that the presence or absence of other qualities has no bearing on the occurrence of certain class features. For predicting cardiac attacks, it is a reliable classifier. Based on the possible outcomes of categorising data sets, The Bayesian network is used to calculate the cumulative probability in every category. The equation is displayed below. P(C) = P X|C P(C|X) P(C) P (X) (X) The aforementioned formula or equation, where X seems to be the object that has to be anticipated, and C is its values that guide, establishes the class in which feature is expected to categorise. The confusion matrix for the predicted values is shown in figure-10.

Figure 10: Confusion Metrics of Gaussian NB

The statistical values such as precision, recall, f1-score, support and accuracy generated through the implementation are shown in the figure-11 below.

	precision	recall	f1-score	support
0	0.82	0.82	0.82	62
1	0.81	0.81	0.81	59
accuracy			0.82	121
macro avg	0.82	0.82	0.82	121
weighted avg	0.82	0.82	0.82	121

Figure 11: Gaussian NB Model Result

E. Implementation of K-Nearest Neighbour

By foretelling the nearest neighbour, classification in KNN is accomplished. As a result of its efficiency and speed, it is favoured to other categorization algorithms. Regression and classification problems can both be solved with it. An individual's status as having or not having heart disease is determined by the algorithm using the heart disease data set. Calculating the separation between points on a graph allows KNN to accurately represent the concept. Based on variables like age, sex, and more, we utilised KNN to categorise and forecast persons with heart disease. The confusion matrix for the predicted values is shown in figure-12.

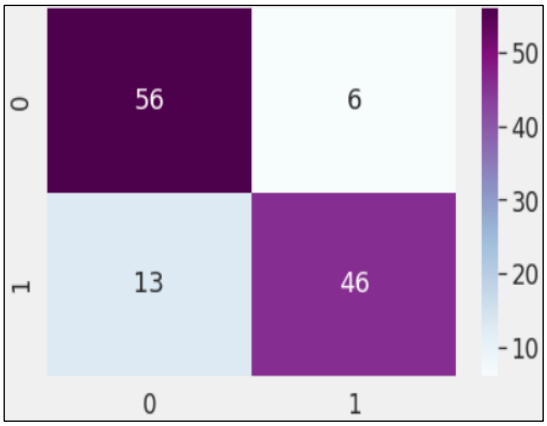

Figure 12: Confusion Metrics of KNN

The statistical values such as precision, recall, f1-score, support and accuracy generated through the implementation are shown in the figure-13 below.

	precision	recall	f1-score	support
0	0.81	0.90	0.85	62
1	0.88	0.78	0.83	59
accuracy			0.84	121
macro avg	0.85	0.84	0.84	121
weighted avg	0.85	0.84	0.84	121

Figure 13: KNN Model Result

VI. RESULTS AND ANALYSIS

Our work concentrated on the application of data mining methods to cardiovascular care. With the help of five data mining techniques, we ran various tests on our database of heart illness. Through the use of multiple algorithms, we try to ascertain which classification algorithm works the best in predicting cardiac disease.

The following stage, after constructing a number of algorithms, is to contrast the various data science techniques employed in these experiments and select which of them provides the highest degree of accuracy. To compare these tests, a number of performance metrics including Accuracy. The True Positive rate, True Positive, False Positive, False Negative, and True Negative values are calculated using it. The formula %Accuracy = (TP + TN) / (TP + TN + FP + FN) * 100 is used to calculate the reliability of the categorization model using the premise on the fact from the confusion matrix. The following table displays a summary of the algorithms.

TABLE 1: COMPARISON OF RESULTS

Algorithms	**Accuracy**		**TN**	**FP**	**FN**	**TP**
LR	0.86		53	9	7	52
DT	0.89		60	2	11	48

RF	0.92		61	1	8	51
NB	0.82		51	11	11	48
KNN	0.85		56	6	13	46

The graphical representation of table-1 is shown in figure-14.

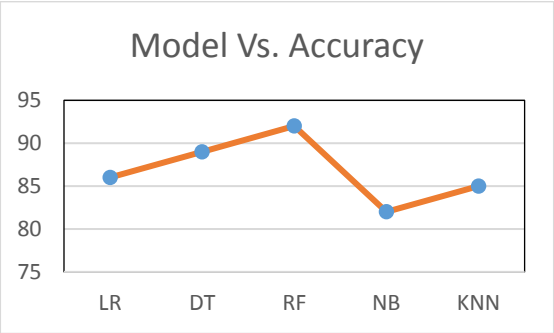

Figure 14: Comparative analysis of all models

The submitted dataset has an accuracy range of 82% to 92%, as seen in the table above. Random Forest achieved the highest accuracy while Naïve Bayes has lowest one. As everyone is aware, cardiac arrest is a delicate and serious condition that claims millions of lives every year. As a result, we must maintain a high TP rate and a low FP rate.

VII. CONCLUSION

The utilisation of data mining techniques in healthcare, particularly the early diagnosis of heart disease, was the main focus of our work. K - means, Linear Regression, Decision Tree, Bayesian Network, and Random Forest were utilised to construct data mining algorithms. We assessed performance using some algorithms, accuracy, TN, FP, FN, and TP rate.

We conducted five experiments to predict heart disease utilizing a single data set. All of the implemented algorithms' results are displayed in a table format for simpler comprehension and comparison. The results signify that Random Forest offers 92% accuracy, which is the highest. Our findings demonstrate the potential for data mining in the medical industry to foresee and detect illness in its infancy.

REFERENCES

[1] Kumar, M., Singhal, S., Shekhar, S., Sharma, B., & Srivastava, G., "Optimized Stacking Ensemble Learning Model for Breast Cancer Detection and Classification Using Machine Learning". Sustainability, 14(21), 13998, 2022.

[2] Sharma, A., Yadav, D. P., Garg, H., Kumar, M., Sharma, B., & Koundal, D., "Bone cancer detection using feature extraction-based machine learning model". Computational and Mathematical Methods in Medicine, 2021.

[3] Yadav, D. P., Sharma, A., Athithan, S., Bhola, A., Sharma, B., & Dhaou, I. B., "Hybrid SFNet model for bone fracture detection and classification using ML/DL". Sensors, 22(15), 5823.2022.

[4] Kaur, B., Kaur, G. "Heart Disease Prediction Using Modified Machine Learning Algorithm", International Conference on Innovative Computing and Communications. Lecture Notes in Networks and Systems, vol 473. Springer, Singapore, (2023).

[5] Bhatt, C.M.; Patel, P.; Ghetia, T.; Mazzeo, P.L. Effective Heart Disease Prediction Using Machine Learning Techniques. *Algorithms* **2023**, *16*, 88 (2023).

[6] Sabab, S.A. "Cardiovascular disease prognosis using effective classification and feature selection technique". International Conference on Medical Engineering, Health Informatics and Technology (MediTec), IEEE. (2016)

[7] Ali M., Paul B. K., Ahmed K., Francis M., Julian M.W., Mohammad A. M., "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison", Computers in Biology and Medicine, Volume 136, 2021.

[8] Umair S., Irfan U Mustafa, Haseeb Qaiser, Fiaz Majeed,"Data Mining in Healthcare for Heart Diseases", March 2015.

[9] Oliveria, C.R. "Development and Validation of a Natural Language Processing Algorithm for Surveillance of Cervical and Anal Cancer and Precancer: A Split-validation Study". JMIR Medical Informatics,2020.

[10] Chalves, L. and Marques, G. "Data Mining Techniques for Early Diagnosis of Diabetes: A Comparative Study". Applied Sciences, 11(5), pp. 2218, 2021.

[11] Priyanka, N., and Ravikumar, P. "Usage of data mining techniques in predicting the heart diseases—Naïve Bayes & decision tree". International Conference on Circuit, Power and Computing Technologies (ICCPCT), IEEE, 2017.

[12] Jindal, H., Agrawal, S., Khera, R., Jain, R. and Nagrath, P., "Heart disease prediction using machine learning algorithms", IOP *Conf. Ser.: Mater. Sci. Eng., 2020.*

[13] Mustaqeem, A., Anwar, M., Khan, A. R., Majid, M. "A statistical analysis-based recommender model for heart disease patients". International journal of medical informatics, 108, pp. 134-145. 2017

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

[15] Bhatt, C., Kumar, I., Vijayakumar, V. et al. The state of the art of deep learning models in medical science and their challenges. Multimedia Systems 27, 599–613 (2021). https://doi.org/10.1007/s00530-020-00694-1

[16] Kumar, I., Mohd, N., Bhatt, C., & Sharma, S. K. (2020). Development of IDS using supervised machine learning. In Soft Computing: Theories and Applications: Proceedings of SoCTA 2019 (pp. 565-577). Springer Singapore.