Optimal Medical diagnosis of Human Heart Disease by K-Nearest Neighbors And Decision Trees Classifiers Algorithms

Ghulab Nabi Ahmad[1], Hira Fatima[2], Shafiullah[3], Nazish Laeiq[4], Syed Md Humayun Akhter[5]

Department of Mathematics, Mangalayatan University, Aligarh, U.P.[1,2]
Department of Mathematics, K.C.T.C College, Raxual, BRA, Bihar University, Muzaffarpur[3]
Department of computer science, Institute of Technology Management Aligarh, UP[4]
Department of Applied Sciences and Humanities, Institute of Technology and Management, Aligarh, UP[5]

Corresponding author: Ghulab Nabi Ahmad, Email: ghulamnabiahmad@gmail.com

Because of its relevance in the growth of tremendous applications in the medical area, data mining is a hugely important domain for exploration. When it comes to fatalities throughout the world, heart disease appears to be the leading cause of death. Recognizing a person's risk of heart disease, it is a difficult assignment for health professionals since it takes very much time and extensive medical testing in the existing models. For the early prediction of cardiac disease, an improved version of the K-Nearest Neighbors (KNN) and Decision Tree (DT) classifier are applied, which ensures greater accuracy than existing models. Enhanced KNN is used for the correct treatment of patients and preserve consistency in the heart disease prediction system. We have calculated feature relevance, which assigns a score to independent factors based on how well they predict the difference walking. The variables difference walking, stroke and diabetes were the most relevant aspects, as shown in figure 2. We have compared to other existing models, the proposed model outperforms the existing models in terms of heart disease prediction. And the suggested work's KNN algorithm has the greatest accuracy of 90.97 percentage and the highest ROC Curve of 69.97 percentage in predicting heart disease in comparison to Decision Tree (DT). The expenses and duration of treatment are reduced when the condition is detected early.

Keywords: Machine learning, heart disease, K-Nearest Neighbors (KNN), Decision Tree (DT)

2023. In Saroj Hiranwal & Garima Mathur (eds.), *Artificial Intelligence and Communication Technologies*, 1031–1037. Computing & Intelligent Systems, SCRS, India.
https://doi.org/10.52458/978-81-955020-5-9-98

1 Introduction

Several medical institutions have now implemented various types of information frameworks to handle their social insurance or patient's data. These data frameworks often create a large amount of data that can be uniquely structured in a flawless manner. This database of rich data is occasionally employed for clinical decision-making. Medical data repositories are huge in scale; hence data mining approaches have been widely employed to extract enlightened knowledge from them. Classification may be less accurate if the data collection contains duplicate and unrelated characteristics [1]. Heart disease is one of the leading causes of death in the World for people of all races, and according to centre of Disease Control and Protection (African Americans, American Indians and Alaska Natives, and white people). Heart disease is caused by three major risk factors: excessive blood pressure, high cholesterol, and smoking, which affect more than half of all Americans (47 percent). Diabetes, obesity (high BMI), a lack of physical activity, and excessive alcohol use are all key markers. Identifying and avoiding the factors that have the greatest impact on heart disease is crucial in medicine. Machine learning algorithms may be used to detect "patterns" in material that might predict a patient's state as a consequence of computer improvements [2].

Explanation of the variables of the dataset

- Heart disease: Those who have ever had heart disease or a myocardial infarction (MI)
- Smoking: Have you ever smoked more than 100 cigarettes in your life? (The solution is either Yes or No.)
- Alcohol consumption: Problem drinkers are adult males who drink more than 14 drinks per week and adult females who drink more than 7 drinks per week.
- Stroke: Have you ever been told you've had a massive heart attack.
- Physical health: Now consider your personal condition, which involves physical disease and injury. Number of times days in the last 30 days have you been in poor physical condition? 0 to 30 days
- Mental Health: When you think about your mental health, how many days in the last 30 days did you have bad mental health? 0 to 30 days
- Difference Walking: Do you find it difficult to walk or climb stairwells?.
- Sex: Are you a male or a female ; Age category: Fourteen-year-old age group.
- Race: The value that has been assigned to race/civilization.
- Diabetic: Have you just been told that you have diabetes?
- Physical Activity: Adults who reported conducting physical activity or exercise outside of their usual work in the previous 30 days.
- Sleep Time: In a 24-hour period, how many hours of sleep do you get on average.
- Asthma: Have you ever been told you have asthma.
- Kidney Disease: Were you ever informed you had kidney illness, excluding kidney stones, fecal impaction, or urinary.
- Skin Cancer: Have you ever been informed you have skin cancer.

The primary goal of exploration and processing after dividing the dataset into 80:20, with 80 percent for training and 20 percent for testing, classification machine learning models such as K-Nearest Neighbours (KNN), Decision Tree (DT), and others can be used with this data set. With the fast advancement of database technology, it has intruded into a variety of sectors. The majority of hospitals now have their own hospital information system (HIS), complete with apps. The database's data volume grows fast, the database's scale extends on a regular basis, and the database's complexity grows [3]. As a result, data mining techniques are employed to cope with massive amounts of medical data. Heart disease diagnosis is a critical issue in the health-care business. It has been determined that using a machine learning method yields superior outcomes [4].

This research provides a model for predicting heart disease using K-Nearest Neighbours (KNN) and Decision Tree. Section II explained the literature review throughout the remainder of the paper. Section III proposed methodology Section IV, as well as the Exploratory and experimental data analysis results, Section V describe the conclusion and future work.

2 Literature review

There have been numerous trainings done on heart disease prediction systems; the following paragraphs discuss some of the publications on heart disease in order to offer a complete survey of literature that may aid in gaining knowledge into various machine learning [5] techniques and their applications. Dun et al. [6] used a variety of machine learning and deep learning approaches to diagnose cardiac illness, as well as hyperparameter tweaking to improve the accuracy of the results. The other models were logistic regression, SVM, and ensemble approaches, and neural networks attained a high accuracy of 78.3 %. Ram Prakash et al. [7] described that healthcare is an essential component of human life. Machine learning models were employed to produce effective conclusions in the heart disease prediction since the healthcare business possesses a significant quantity of psychiatric data. Using machine learning approaches, it is possible to consistently classify people as healthy or unhealthy. In this investigation, we built a framework to comprehend the concepts of forecasting the risk profile of patients using clinical data characteristics. G. N. Ahmad et at. [8] have provided a Comparative Study of Optimal Medical Diagnosis of Human Heart Disease was performed using Machine Learning Techniques with and without Sequential Feature Selection.. Linda et al. [9] had described for the heart disease patients who were given exercise, we developed a one-of-a-kind health information system. According to early data, clinicians are unsure how to construct an exercise prescription for patients with a variety of CVD risk factors. The supplied method is beneficial to patients since it is evidence-based, easy to use, guided, and time-saving. G. N. Ahmad et at. [10] have used Machine Learning Techniques for Efficient Medical Diagnosis of Human Heart Diseases with and without GridSearchCV

3 Methodology

On a heart disease database, two data mining approaches, namely KNN and DT, are examined in this work. Prediction techniques include [10-11].

3.1 K Nearest Neighbours

K Nearest Neighbours (Data points) are used to anticipate the class or continuous value for a new Data point, as the name indicates. The algorithm is in the process of becoming more intelligent. Decide on the number K of neighbours. And Calculate the Euclidean distance between K of your neighbours, Select the K closest neighbours using the estimated Euclidean distance, count how many data points there are in each category among the k neighbours, Assign the newly added data points to the category with the most neighbours.

3.2 Decision Tree

J48 is the most common decision tree algorithm for prediction because it employs a pruning strategy to create a good decision tree. Pruning helps to eliminate elements that aren't necessary for prediction and may result in incorrect judgments. It has a binary tree representation, which makes it simple to learn.

3.3 Proposed Algorithm

Select the data set(2020_cleaned_dataset), Remove the rows with missing values from the dataset as a pre-processing step:

- Split the dataset into two halves (80 % trained data & 20 % test data).
- Train the classical by applying a KNN and DT.
- 20% of the dataset is used to test the model.
- Output the expected result.
- End figure 1 is a flowchart illustration of the suggested technique.

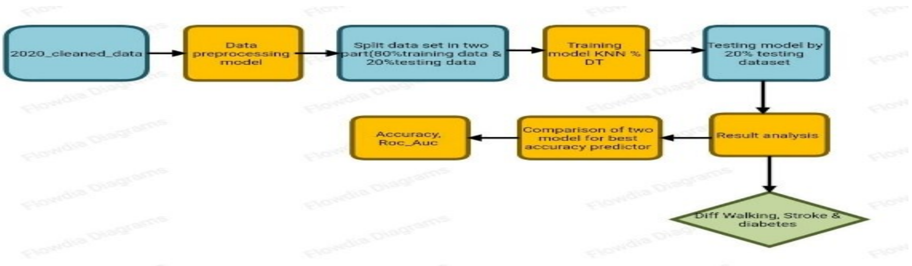
2020_cleaned_data
Data preprocessing model
Split data set in two part(80%training data & 20%testing data
Training model KNN % DT
Testing model by 20% testing dataset
Result analysis
Comparison of two model for best accuracy predictor
Accuracy, Roc_Auc
Diff Walking, Stroke & diabetes

Fig 1. Block diagram proposed method

3.4 Evaluation Model

$$Accuracy = \frac{\text{True Possitive(TP)+True Negative(TN)}}{True\ Possitive(TP)\text{+True Negative}(TN)+False\ Possitive\ (FP)+False\ Negative(FN)} \qquad (1)$$

$$Precision = \frac{\text{True Possitive(TP)}}{True\ Possitive(TP)+False\ Possitive(FP)} \qquad (2)$$

$$Recall = \frac{\text{True Possitive(TP)}}{true\ possitive(tp)+false\ negative(fn)} \qquad (3)$$

$$F1 - Score = 2 * \frac{\text{Precision}*\text{Recall}}{Precision+Recall} \qquad (4)$$

4. Result discussion

4.1 Exploratory result discussion

The first step is to choose a dataset from a variety of machine learning resources available online. The 2020 _cleaned _dataset is one of several online repositories with 18 variables relevant to patients' vital signs in connection to heart disease. We have roughly taken 319795 items with 18 columns, according to Kaggle's output and there are 14 numeric and four categorical characteristics. We can convert string properties that only have two possible unique values, but first, let's make sure there aren't any out-of-the-ordinary values. Because there were no biases in the data, this strategy had no influence on the rest of the data utilised in the experiment. After the splitting of the datasets into a training: testing ratio of 255836 (training):63959 (testing), with 80 % for training and 20 % for testing, **Figure 2** describe the heat map correlation features for the chosen parameters which is used to predict heart and **Figure 3** gives the distribution correlation to the greatest and lowest characteristics in the form of a Heat Map. **Figures 4 and 5** shows the distributions of BMI and heart disease, both normal and abnormal, as well as the frequency of normal and abnormal heart disease.

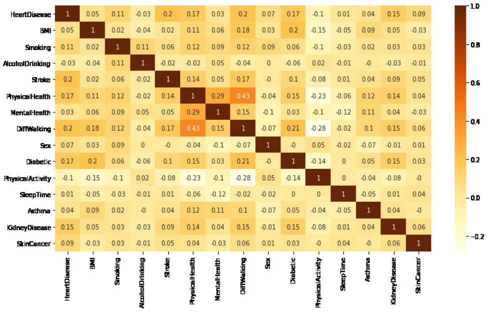

Fig 2. Heat Map Correlation Features

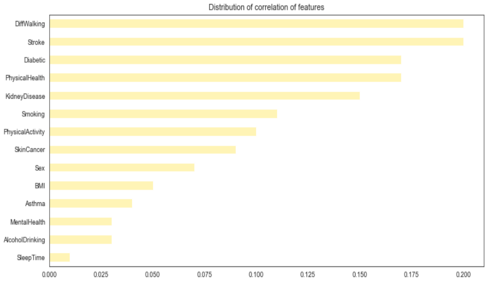

Fig 2. Distribution of Correlation Features

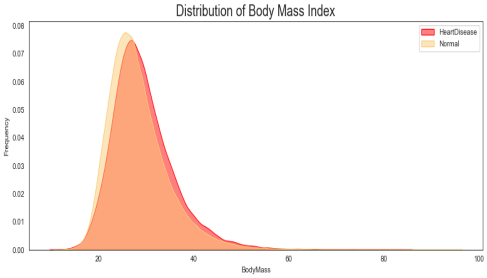

Fig 4. Distribution of body mass

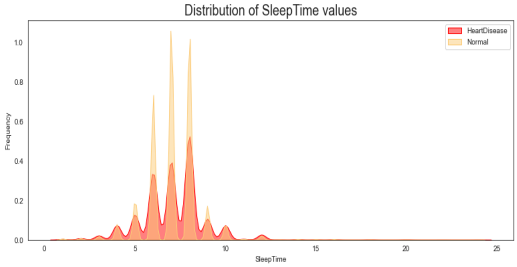

Fig 5. Distribution of sleep time value

4.2 Discussion of Experimental Results

The key features discovered in the 2020_cleaned_ dataset Kaggle dataset in terms of heart disease medical prediction datasets were difficult Walking, stroke, and diabetes). K-Nearest Neighbours (KNN) demonstrated to be the best-performing model across the five metrics of accuracy (90.70 percent), precision (35.39 percent), recall (11.38 percent), f1-score (17.12 percent), and roc-auc in the top-three attribute classification utilising the train-test split approach (69.97 percent), while the decision tree Classifier is lowest in terms of accuracy (86.56%), precision (23.47%), recall (25.59%), f1-score (24.49%), and roc-auc (24.49%), (58.95%). **Table 1** shows the comparative model.

Table 1: Comparison of models performances

Models	Accuracy	precision	Recall	F1-score	Roc-Auc
KNN	90.70	35.39	11.38	17.12	69.97
DT	86.56	23.47	25.59	24.49	58.95

In terms of performance evaluation, the preceding Table 1 compared the results of two data mining algorithms, namely KNN and DT (Precision, Recall, F1 Score and Accuracy). The higher the level of precision, the more exact the output will be with fewer in accuracy. This model incorporates more real positive findings (having heart disease) from all of the patients, as shown by the recall. The harmonic mean of recall and accuracy is used to determine the F1 Score.

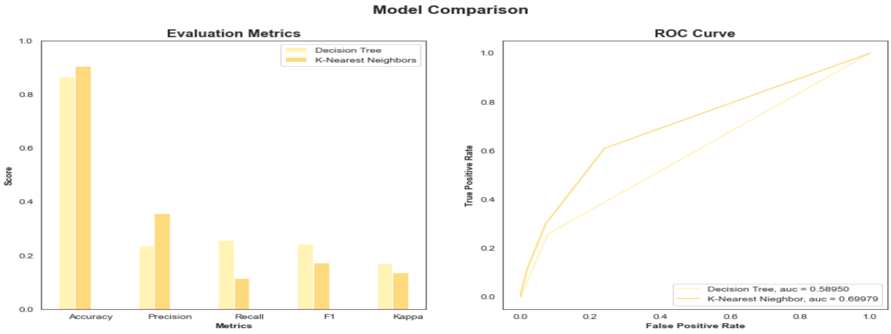

Fig 6: Comparison model of KNN and DT

When compared to DT, **figure 6** shows that among the two algorithms active for heart disease prediction using 2020_clened_ data, the KNN algorithm predicts the illnesses with the best accuracy of 90.97 % and the largest ROC Curve of 69.97 %.

Table-2. Comparison of the Model from the existing previous studies

Authors	Methods	Results
Saqlain *et al.* [13]	MFSFSA + SVM	81.19%
Latha and Jeeva [14]	Naïve Bayes + BN + Random Forest + MLP	85.48%
Mohan *et al.* [15]	RF + Linear Model	88.4%
Ayon *et al.* [16]	RF	87.45%
Proposed model	**DT**	**86.56%**
	KNN	**90.70%**

We also have compared our strategy to a number of other researchers' previously published methodologies. For example, Mohan *et al.* [15] used a grouping of the RF and Linear Model to get a high-accuracy classification, **Table-2** compares the results.

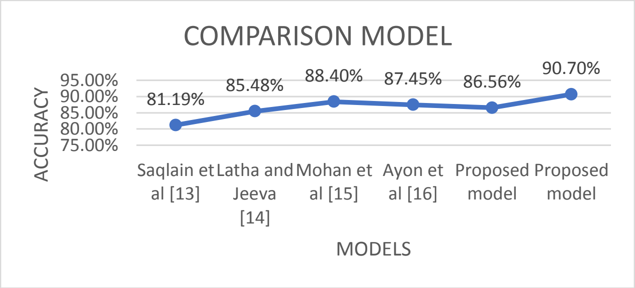

Fig 7. Comparison of the Model from the existing previous studies

Figure 7 shows a comparison of performance measures such as accuracy of the proposed technique to the existing methodologies utilising certain performance assessment measures.

4 Conclusions

We have offered two strategies in this research for comparative analysis and have obtained encouraging findings. In our investigation, we have discovered that machine learning techniques performed better. Many scholars have already recommended that we apply machine learning, which this article supports. In this analysis, we have obtained precision, recall, sensitivity, F1 score, and roc-auc for the comparative methodologies employed for the 18 attributes. There are no null values in the dataset, and there are 14 numeric variables and 4 categorical characteristics. In this analysis, it is obtained that KNN performed better in the ML techniques. The accuracy of the model is assessed using publicly accessible datasets from Kaggle, such as the 2020 _cleaned _dataset, with KNN achieving a 90.70 % accuracy, The accuracy of roc-auc is 69.97%, whereas DT had an accuracy of 86.56 %. In this work, we offer a predictive analytics-based strategy for identifying the factors that affect heart disease. The most important components of heart disease were difference walking, stroke, and diabetes. **Figure 8** compares the proposed technique's performance measurements, such as accuracy and roc auc, to existing approaches that use performance evaluation measures.

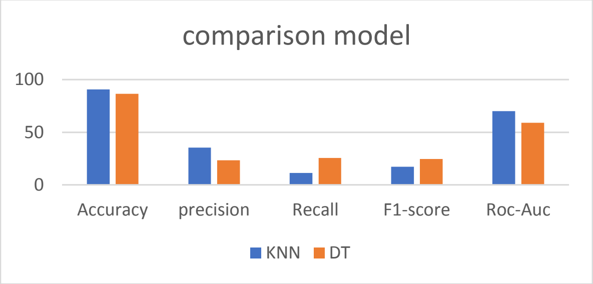

Fig 8. Comparison of models KNN and DT

This model's performs is better than the existing approach with 90.70 % accuracy.

5 Future work

KNN is a Machine Learning Techniques that is used to predict cardiac disease. Heart attacks account for four out of every five deaths caused by cardiovascular disorders, according to the World Health Organization (CVD). In the future work might be enhanced by developing an online application based on KNN and employing a larger dataset than the one utilized in this study, which would assist to deliver better findings and aid health professionals in successfully and efficiently predicting gastrointestinal disease.

References

[1] Mai Shouman, Tim Turner and Rob Stocker, "Using data mining techniques in heart disease diagnosis and treatment", *Japan-Egypt Conference on Electronics, Communications and Computers (JEC-ECC)*, 2012.

[2] Liburd, L. C., Jack Jr, L., Williams, S., & Tucker, P. "Intervening on the social determinants of cardiovascular disease and diabetes." *American journal of preventive medicine*, *29*(5), 18-24. 2005.

[3] K. Subhadra, Vikas B, "Neural Network Based Intelligent System for Predicting Heart Disease", *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* ISSN: 2278-3075, Volume-8 Issue-5, pp. 484-487, March 2019.

[4] Zhi-Gen Hu, Jian-Ping Li, "Research And Application Of Data Warehouse And Data Mining Technology In Medical Field", *12th International Computer Conference On Wavelet Active Media Technology And Information Processing (ICCWAMTIP)*, IEEE 2015.

[5] Jaymin Patel et al., "Heart Disease Prediction Using Machine learning and Data Mining Technique", volume 7, pp. 129-137, March 2016.

[6] B. Dun, E. Wang, and S. Majumder, "Heart disease diagnosis on medical data using ensemble learning," 2016.

[7] P. Ramprakash, R. Sarumathi, R. Mowriya, and S. Nithyavishnupriya, "Heart disease prediction using deep neural network," *in Proceedings of the 2020 International Conference on Inventive Computation Technologies (ICICT)*, pp. 666–670, IEEE, Coimbatore, India, February 2020.

[8] G. N. Ahmad, Shafiullah, A. Algethami, H. Fatima and S. M. H. Akhter, "Comparative Study of Optimum Medical Diagnosis of Human Heart Disease Using Machine Learning Technique with and Without Sequential Feature Selection," in IEEE Access, vol. 10, pp. 23808-23828, 2022, doi: 10.1109/ACCESS.2022.3153047.

[9] P. S. Linda, W. Yin, P. A. Gregory, Z. Amanda, and G. Margaux, "Development of a novel clinical decision support system for exercise prescription among patients with multiple cardiovascular disease risk factors," *Mayo Clinic Proceedings: Innovations, Quality & Outcomes*, vol. 5, no. 1, pp. 193–203, 2021.

[10] G. N. Ahmad, H. Fatima, Shafiullah, A. S. Saidi and Imdadullah, "Efficient Medical Diagnosis of Human Heart Diseases using Machine Learning Techniques with and without GridSearchCV" *in IEEE Access,* 2169-3536,2022, doi: 10.1109/ACCESS.2022.3165792.

[11] A. Kishor & Jeberson, W. "Diagnosis of heart disease using internet of things and machine learning algorithms." In *Proceedings of Second International Conference on Computing, Communications, and Cyber-Security* (pp. 691-702). (2021). Springer, Singapore.

[12] P.Ghosh, S. Azam, M. Jonkman, A. Karim, "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques." *IEEE Access*, *9*, pp.19304-19326. 2021.

[13] S. M. Saqlain, M. Sher, F. A. Shah, I. Khan, M. U. Ashraf, M. Awais, and A. Ghani, "Fisher score and Matthew's correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines," Knowl. Inf. Syst., 58(1), pp. 139_167, Jan. 2019, doi:10.1007/s10115-018-1185-y.

[14] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," Inform. Med. Unlocked, 16, Jan. 2019, Art. no. 100203, doi: 10.1016/j.imu.2019.100203.

[15] Senthil Kumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava, Effective heart disease prediction using hybrid machine learning techniques, IEEE Access 7 (2019) 81542–81554, https://doi.org/10.1109/ACCESS.2019.2923707.

[16] Safial Islam Ayon, Md. Milon Islam, Md. Rahat Hossain, Coronary Artery heart disease prediction: a comparative study of computational intelligence techniques, IETE J. Res. (2020), https://doi.org/10.1080/03772063.2020.1713916.