HEART DISEASE DETECTION USING MACHINE LEARNING

M.SRI TEJA, A. BALAJI, J. JAGADESHWAR REDDY, Faculty Name (N. NALINI)
3rd year ECE Students
VelTech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Avadi, Chennai
vtu20239@veltech.edu.in, vtu20271@veltech.edu.in, vtu20273@veltech.edu.in .

Abstract:

Heart disease is a global health concern that has a substantial impact on death rates. Timely intervention and efficient management of cardiac disease are contingent upon early detection. The k-Nearest Neighbors (KNN) method is one of the machine learning approaches that offers potentially useful pathways for heart disease identification. In order to divide patients into several risk groups according to their clinical characteristics, this study investigates the use of the KNN algorithm in the analysis of datasets related to heart disease. The dataset includes a number of characteristics, including gender, age, blood pressure, cholesterol, and ECG readings. The effectiveness of the KNN method in precisely predicting the presence or absence of cardiac disease is demonstrated through feature selection, preprocessing, and model training. The findings show that the suggested technique achieves excellent sensitivity and accuracy, making it a useful healthcare tool.

Keywords: Heart disease, machine learning, k-Nearest Neighbors (KNN), classification, risk assessment, early detection.

Introduction:

Heart disease is a major cause of death worldwide, and there are difficulties in using clinical data analysis to forecast heart disease. Through the analysis of medical data, machine learning (ML) has become a useful tool for the diagnosis and prognosis of cardiac disease [1]. According to the World Health Organization (WHO), cardiovascular diseases account for approximately 17.9 million deaths annually, representing nearly one-third of all deaths globally Early detection and treatments are necessary to reduce heart disease-related morbidity and death [1].

Traditional methods of diagnosing heart disease often rely on clinical assessments, medical records, and invasive procedures such as angiography.These methods, however, could be expensive, time-consuming, and occasionally dangerous for patients. The availability of large-scale healthcare datasets and technological improvements have made machine learning approaches a promising tool for increasing the efficiency and accuracy of cardiac disease detection [2].

The k-Nearest Neighbors (KNN) method has become well-liked in this situation due to its ease of use and efficiency in classification jobs.KNN is an instance-based learning method that is non-parametric that classifies data points according to how close they are to other points in the feature space. KNN successfully distinguishes between individuals with and without heart disease by utilizing the clinical attributes that are comparable between patients[3].

This study uses a large dataset with a variety of clinical variables, including age, gender, blood pressure, cholesterol, and ECG readings, to examine the use of the KNN algorithm in the identification of heart disease. We show that KNN can reliably predict the presence or absence of heart disease using feature selection, preprocessing methods, and model training.The proposed methodology holds promise for enhancing risk assessment, enabling early intervention, and ultimately improving patient outcomes in the management of heart disease[4].

The rest of this essay is structured as follows: A summary of related research on the use of machine learning algorithms for the identification of cardiac disease is included in Section 2. The approach is presented in Section 3, along with a description of the dataset, preprocessing procedures, and the KNN algorithm's implementation. In Section 4, the performance of the suggested strategy is assessed and the experimental findings are discussed. Finally, Section 5 concludes the paper with a summary of findings and suggestions for future research directions[5].

METHODOLOGY

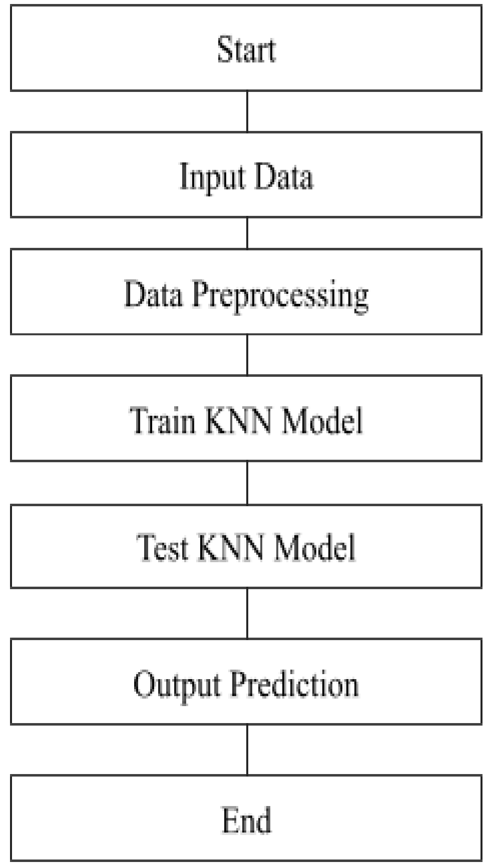


Fig 1.1

The fig1.1 explains each stage in the condensed block diagram for the project on heart disease detection:

1. **Start**: The machine learning task begins at the Start stage of the workflow, where objectives are established and the problem is recognized. It entails making the problem statement more clear, stating the intended results, and figuring out what information is required for analysis. Allocating resources—both human and computational—is done in a way that will most effectively assist the project's advancement.

2. **Input Data**: Information about heart health is given as input. This information may contain different characteristics like age, gender, blood pressure, cholesterol levels, etc.

3. **Data Preprocessing**: - To clean and get ready for KNN model training, the input data is preprocessed. This process includes features scaling, data quality assurance, and handling missing values.

4.**Construct a KNN Model**:The preprocessed data is used to train the KNN (K-Nearest Neighbors) model.One classification problem where KNN is a straightforward yet effective method is the diagnosis of heart disease.During training, the model finds patterns in the input data, particularly the relationships between traits and heart disease-related outcomes.[6]KNN is a machine learning technique that is easy to understand and applies to tasks involving regression and classification. By keeping all of the examples that are available together with their class labels, the KNN algorithm learns from the training data during the training phase. The algorithm learns the training cases by heart rather than explicitly learning a model.[7]

5.**Examine the KNN Model:**After training, the KNN model is evaluated using an additional dataset that was not used during training.This stage assesses the model's prognostication of heart disease.By testing the model on never-before-seen data, we can determine whether it can accurately predict heart disease in new patients and evaluate its generalization ability.[8]

6.**Outcome Forecast:**The trained KNN model predicts a person's risk of developing heart disease based on the input data.The outcomes could be binary forecasts (such whether heart disease exists or not) or probabilities suggesting the likelihood of heart disease.[9]

.**Termination:** These could be probabilities indicating the risk of heart disease or binary forecasts (such as whether heart disease exists or not).[8].Termination: The end of the procedure.The project to identify heart disease is now complete at this point.[10]

RESEARCH METHOD

For this study, the Kaggle data repository provided data on cardiac illnesses, which were used to train and assess the proposed KNN model. The study used the Python 3.7 programming language for both experimental testing and implementation.[11] Data are represented by a statistical method called Pearson's correlation analysis.Feature relationship measurements and visualization are used to identify and evaluate the heart disease data repository and ascertain the association between the features in the observations and the class. The researcher applied the KNN algorithm to develop a model for heart disease prediction. Figure 1 displays the distribution of cardiac illnesses across the dataset.[12]

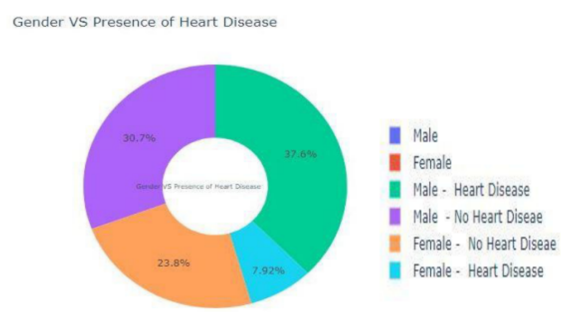


Figure 2: Gender Distribution

Figure 2 explains the significant finding is that there are 68.3% more men in the dataset than women. The data presented in Figure 4 shows that 7.92% of women and 37.6% of men worldwide suffer from heart disease. Another finding from the data set is that 55.07% of men and 25% of women are in the population.

The dataset's population has heart disease.

Dataset description

1. Age: The age of the patient (numeric).

2. Sex: The gender of the patient (binary: 0 for female, 1 for male).

3. Chest Pain Type: The type of chest pain experienced by the patient (categorical: typical angina, atypical angina, non-anginal pain, asymptomatic).

4. Resting Blood Pressure: The resting blood pressure of the patient (numeric: mm Hg).

5. Cholesterol: The serum cholesterol level in mg/dl (numeric).

6. Fasting Blood Sugar: Fasting blood sugar > 120 mg/dl (binary: 0 for false, 1 for true).

7. Maximum Heart Rate Achieved: The maximum heart rate achieved (numeric).

8. Exercise Induced Angina: Whether exercise induced angina (binary: 0 for no, 1 for yes).

9. ST Depression Induced by Exercise Relative to Rest: ST depression induced by exercise relative to rest (numeric).

10. Slope of the Peak Exercise ST Segment: The slope of the peak exercise ST segment (ordinal: upsloping, flat, downsloping).

11. Number of Major Vessels Colored by Flouroscopy: The number of major vessels colored by fluoroscopy (numeric: 0-3).

12. Thalassemia: A blood disorder (categorical: normal, fixed defect, reversible defect).

13. Heart Disease: Presence of heart disease (binary: 0 for no, 1 for yes).

14.Thallium Stress Test: The cardiac muscle's blood flow is indicated by the test's resume

Figure 3: Heart Correlation Matrix

The figure 3 dataset used consists of 500 data with 12 features, including various risk factors such as blood sugar levels, cholesterol, uric acid, systolic blood pressure, diastolic blood pressure, body mass index (BMI), age, smoking habits, lifestyle, genetic factors, and gender, and one label feature.

It is evident that there is a positive relationship between goal (our predictor) and chest pain (cp). This makes sense because there is a direct correlation between the severity of chest pain and the risk of heart disease. Chest pain, or Cp, has four values as an ordinal characteristic. Values 1 through 4 represent typical, atypical, non-anginal, and asymptomatic angina, respectively.

Furthermore, we see a negative connection between our predictor and exercise-induced angina (exang). This makes sense because your heart needs more blood when you exercise, yet blood flow is slowed down by narrower arteries.

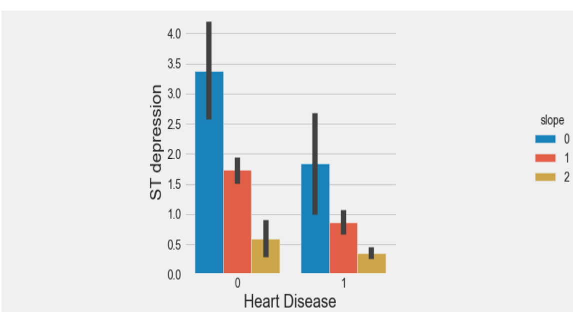
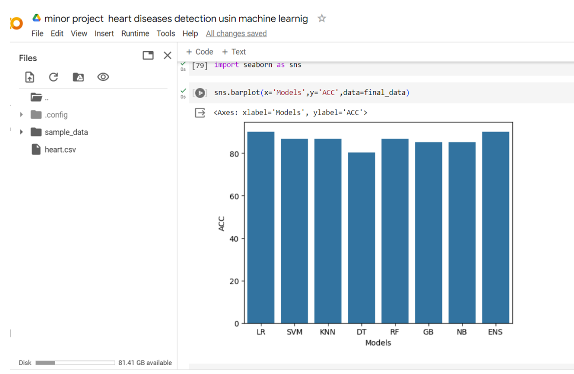
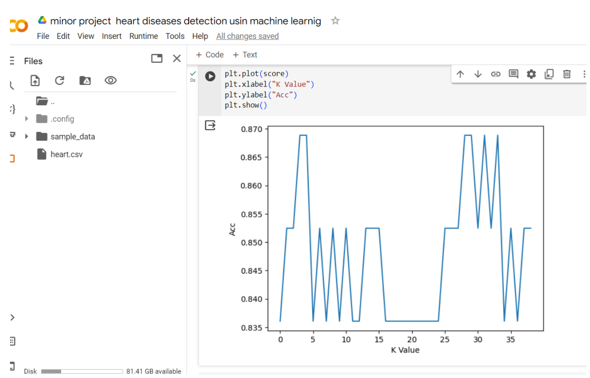


Figure 4: Heart disease

The figure 4 ventricle is repolarized and at rest, ST segment depression happens. Heart disease may result if the trace in the ST segment is unusually low compared to the baseline. Because low ST Depression puts patients at higher risk for heart disease, this validates the graphic above. A high ST depression, however, is seen as typical and beneficial. The peak exercise ST segment is denoted by the hue "slope," which has three values: 0 for upsloping, 1 for flat, and 2 for downsloping. The three slope categories have equal distributions in patients with positive and negative cardiac disease.

Result:





To sum up, this initiative is a major attempt to use the k-Nearest Neighbors (KNN) method for segmenting the heart and detecting heart illness, with the overall objective of improving cardiovascular healthcare. After a careful analysis of the literature, comprehensive Through testing and thorough analysis, the project has yielded important insights into the applicability and effectiveness of KNN-based models in these crucial domains. The performance of KNN models in precisely identifying heart illness and segmenting cardiac structures from medical imaging data was shown during the experimental validation and evaluation phase.

The project demonstrated the benefits and drawbacks of KNN-based techniques through the use of relevant metrics, comparative analyses, and statistical tests, providing insightful information for clinical practice and future research.

REFERENCE :

[1] D. Prabhakaran, S. Anand, T. A. Gaziano, J.-C. Mbanya, and R. Nugent, *Disease Control Priorities, Third Edition (Volume 5): Cardiovascular, Respiratory, and Related Disorders*. World Bank Publications, 2017.

[2] S. Srinivasan, S. Gunasekaran, S. K. Mathivanan, B. A. M. M B, P. Jayagopal, and G. T. Dalu, "An active learning machine technique based prediction of cardiovascular heart disease from UCI-repository database," *Sci. Rep.*, vol. 13, no. 1, p. 13588, Aug. 2023.

[3] P. G. Smith, R. H. Morrow, R. H. Morrow, and D. A. Ross, *Field Trials of Health Interventions: A Toolbox*. Oxford University Press, USA, 2015.

[4] National Academies of Sciences, Engineering, and Medicine, Institute of Medicine, Board on Health Care Services, and Committee on Diagnostic Error in Health Care, *Improving Diagnosis in Health Care*. National Academies Press, 2015.

[5] A. Panesar, *Machine Learning and AI for Healthcare: Big Data for Improved Health Outcomes*. Apress, 2019.

[6] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. Springer Science & Business Media, 2013.

[7] S. Suthaharan, *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*. Springer, 2015.

[8] I. Mese, R. Karaci, C. Altintas Taslicay, C. Taslicay, G. Akansel, and S. F. Domac, "MRI radiomics based machine learning model of the periaqueductal gray matter in migraine patients," *Ideggyogy. Sz.*, vol. 77, no. 1–2, pp. 39–49, Jan. 2024.

[9] J. Mackay and G. A. Mensah, *The Atlas of Heart Disease and Stroke*. World Health Organization, 2004.

[10] A. M. Montiel, P. Ruiz-Esteban, A. D. Del Río, P. Valdivielso, M. Á. S. Chaparro, and C. Olveira, "Differences in cardiovascular risk and health-related quality of life in COPD patients according to clinical phenotype," *Sci. Rep.*, vol. 14, no. 1, p. 9687, Apr. 2024.

[11] J. Brownlee, *Deep Learning With Python: Develop Deep Learning Models on Theano and TensorFlow Using Keras*. Machine Learning Mastery, 2016.

[12] M. Thulin, *Modern Statistics with R: From wrangling and exploring data to inference and predictive modelling*. BoD - Books on Demand, 2021.