Journal of Current Research in Engineering and Science
JCRES
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume7-Issue1, January 2024
Paper: 20

DIAGNOSIS OF HUMAN HEART DISEASE USING MACHINE LEARNING

Mrs.P.Manju Bala
DeptofCSE
IFET College of Engineering
Villupuram,India.
pkmanju26@gmail.com

Elamparithi.T
Deptof CSE
IFET College of Engineering
Villupuram,India.
parithithulasi@gmail.com

***Abstract*— Human heart disease prediction is one of the most difficult task in the medical field. It takes more time to diagnose and find the causes of the heart disease, especially for doctors and medical experts. In this paper, the heart disease could be predicted by various Machine Learning algorithms such as Logistic Regression, K-Nearest Neighbor, Support Vector Machines and Gradient Booster Classifier with RandomizedSearchCV. This system uses five-fold cross-validation technique for verification. The diagnosis of the human heart disease implemented in a jupyter notebook platform. The machine learning algorithms are implemented in thejupyter notebook to identify the presence of theheart disease. The extreme gradient booster optimizes the machine learning algorithms to provide the accurate results. With the help of this project, we can easily diagnose the heart disease. It is tested that the gradient booster with the RandomizedSearchCV providing high accuracy results.**

I. INTRODUCTION

Heart Disease has been proven one of the leading causes of death that is why an accurate and timely prediction of heart disease is extremely essential. Machine learning proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the health care industry. The availability of various medical data prompts us to consider whether there are any efficient and effective methods for analysing this data and deriving potentially innovative and applied knowledge. One of the most significant difficulties for data analytics is the diagnosis of various diseases. Heart-related disorders can be reduced with the use of medical data and machine learning techniques. A machine learning model learns from the historical data it receives and then creates prediction algorithms to forecast the outcome for fresh data that enters the system as input. The accuracy of these models is determined on the quality and amount of input data. In this situation, the heart typically struggles to provide enough blood to the other areas of the body for them to operate normally. The ability to diagnose this issue quickly and accurately is crucial for sparing patients' lives and limiting further damage. Although angiography is thought to be the most well-known invasive-based approach for identifying heart issues, it has several drawbacks. On the other hand, non-invasive based methods, such as computational techniques based on intelligent learning, are shown to be more reliable and effective for the diagnosis of cardiac disease.

Coronary artery blockage and narrowing cause heart failure. The heart's own blood supply is provided via the coronary arteries. The most typical signs of cardiac illness include fatigue, physical body weakness, shortness of breath, swollen feet, and weariness. A person's lifestyle choices, such as smoking, eating poorly, having high cholesterol or blood pressure, not exercising enough, and being physically inactive, can all raise their chance of developing heart disease.

Heart illness comes in a variety of forms, the most common of which is coronary artery disease (CAD), which can cause heart attacks, strokes, and chest pain. Heart rhythm issues, congestive heart failure, congenital heart disease (birth-related heart disease), and cardiovascular disease are among the additional types of heart illness (CVD). Heart disease was initially diagnosed using conventional investigation procedures, but they proved to be difficult. The diagnosis and treatment of heart disease are exceedingly complicated since medical diagnostic equipment and specialists, particularly in underdeveloped nations, are not readily available. To save the patient from suffering further harm, a proper and accurate diagnosis of heart disease is crucial.

Diagnosis of heart disease typically involves a combination of medical history, physical examination, and diagnostic tests. Medical history involves asking the patient about their symptoms, medical history, and any risk factors for heart disease, such as high blood pressure, high cholesterol, and diabetes. A physical examination may include listening to the heart with a stethoscope and checking for any signs of heart disease.

Diagnostic tests that may be used to diagnose heart disease include electrocardiograms (ECGs), echocardiograms, stress tests, cardiac catheterization, and blood tests. These tests can provide valuable information about the structure and function of the heart, as well as any abnormalities or damage that may be present.

Prompt diagnosis and treatment of heart disease are essential for managing symptoms, preventing complications, and reducing the risk of serious health problems. With early detection and appropriate care, many people with heart disease can live long and healthy lives.

Machine learning algorithms have been used in recent years to aid in the diagnosis of heart disease. These algorithms can analyze large datasets of patient information and medical records to identify patterns and predict the likelihood of heart disease.

Machine learning algorithms can be trained on large datasets of patient information to predict the likelihood of heart disease based on a combination of risk factors such as age, sex, blood pressure, and cholesterol levels. These models can be used to identify patients who are at high risk of developing heart disease and may benefit from closer monitoring or early interventions.Overall, machine learning has the potential to improve the accuracy and efficiency of heart disease diagnosis which can ultimately lead to better patient outcomes.

II. Methodology

This project's key contribution was the development of an understandable medical prediction system for the diagnosis of heart disease using cutting-edge machine learning algorithms. In this study, a variety of machine learning classifier algorithms, such as logistic regression (LR), K-nearest neighbours (K-NN), support vector machines (SVM), and XGBoost with RandomizedSearchCV, were trained with the goal of identifying the most effective predictive model for precise early heart disease diagnosis.Machine learning can be used to aid in the diagnosis of heart disease by analyzing large amounts of data and detecting patterns that may not be easily noticeable to humans. It involves data collection, data pre-processing, feature extraction, training, etc.

Four model selection procedures were utilised, including the correlation-based feature subset evaluator, to find the perfect set of characteristics that significantly influenced the classifiers' abilities to predict the target class. Finally, the hyperparameter "RandomizedSearchCV" in the XGBoost was tuned using the complete attribute set and optimal sets discovered using attribute evaluators. The system verifies itself using a 5-fold cross-validation method.

Large amounts of data, including medical records, imaging studies, and laboratory results, are collected from patients with and without heart disease.The collected data is pre-processed to remove any missing values, outliers, or errors.Relevant features are extracted from the pre-processed data, such as age, blood pressure, cholesterol levels, and electrocardiogram readings.A machine learning model is trained using the pre-processed data and the extracted features. Various algorithms can be used for this purpose, such as logistic regression, decision trees, or neural networks.

The performance of the machine learning model is evaluated on a separate dataset to assess its accuracy, sensitivity, and specificity in diagnosing heart disease.Once the machine learning model is validated, it can be deployed in clinical settings to assist healthcare providers in making a diagnosis of heart disease.

The key focus of this system is to make a heart disease diagnosis system to identify the risk of the heart disease. It helps the heart disease affected patients to identify the heart disease risk earlier by using this diagnosis system. One can easily know the risk of theirs heart condition with this diagnosis system.

It is implemented in the jupyter notebook with the help of the machine learning algorithms. The machine learning algorithms such as logistic regression, K-nearest neighbour, Support vector machines, XGBoost were used to train the model and to analyze the dataset to check the number patients who are having the heart disease and classify the gender i.e. it classifies the number of males with the heart disease and the number of females having heart disease. The model also shows the heart disease affected persons and also the non-affected people.

Data collection and pre-processing

Data Collection:

The first step is to collect the data required for the machine learning model.

Data Cleaning: Once the data is collected, it needs to be cleaned to remove any errors or inconsistencies. This may involve removing duplicate or irrelevant data, handling missing values, and correcting any errors or outliers.

Feature Selection:

Feature selection involves selecting the most relevant features (or variables) from the dataset. This can be done using statistical methods, domain knowledge, or automated feature selection algorithms.

Feature Scaling:

Machine learning algorithms often require features to be scaled to a common range to ensure that they have equal importance. This can be done using normalization or standardization techniques.

Data Splitting:

Once the data is preprocessed, it is split into training, validation, and test sets. The training set is used to train the machine learning model, while the validation set is used to fine-tune the model and select hyperparameters. The test set is used to evaluate the performance of the model on new, unseen data.

Data Augmentation:

In some cases, data augmentation techniques may be used to increase the size of the dataset and improve the performance of the machine learning model.

A. Project Specification:

The main components being used in making the heart disease diagnosis model have been discussed below.

Hardware requirements:

- Monitor 15.6”
- Intel i3 or above processor
- 250 Gb Hard Disk

Software requirements:

- Python
- Jupyter notebook(Anaconda)

B. Machine Learning Algorithms and Technique:

Logistic regression:

It is a statistical method used for binary classification, which is the task of predicting a binary outcome (such as yes/no, true/false, or 0/1) based on one or more predictor variables. It is a type of generalized linear model that models the probability of the binary outcome as a function of the predictor variables.The logistic regression model uses the logistic function (also known as the sigmoid function) to model the probability of the binary outcome. The logistic function takes any input value and maps it to a value between 0 and 1, which can be interpreted as a probability.

K-nearest neighbor:

K-nearest neighbor (KNN) is a supervised machine learning algorithm used for classification and regression. It is a non-parametric method that makes predictions based on the k closest (or nearest) neighbors in the training set.For classification, KNN determines the class of an input sample by finding the K nearest neighbors in the training set and assigning the class that is most frequent among those neighbors. In other words, the predicted class of an input sample is based on the majority class of its k-nearest neighbors.

Support Vector Machine:

Support Vector Machines (SVM) is a supervised machine learning algorithm used for classification and regression tasks. The goal of SVM is to find a hyperplane in a high-dimensional space that separates the input data into different classes with the largest possible margin.In the case of a binary classification problem, SVM finds the hyperplane that best separates the two classes by maximizing the distance (or margin) between the closest points from each class, known as support vectors. The distance between the hyperplane and the support vectors is called the margin, and SVM aims to find the hyperplane with the maximum margin.

SVM can also handle non-linear classification problems by using a technique called the kernel trick, which maps the input data into a higher-dimensional space where it can be linearly separated. The kernel function measures the similarity between two points in the input space, and the choice of kernel function can have a significant impact on the performance of the SVM model.In addition to classification, SVM can also be used for regression tasks by finding a hyperplane that best fits the input data while minimizing the number of training errors.SVM is a powerful algorithm that can handle complex and non-linear relationships between the input and output variables, and it is known for its ability to generalize well on unseen data. However, SVM can be sensitive to the choice of hyperparameters, such as the kernel function and the regularization parameter, and can be computationally expensive for large datasets.

XGBoost:

XGBoost (eXtreme Gradient Boosting) is a popular and effective machine learning algorithm for both regression and classification tasks. It is an ensemble method that combines multiple decision trees to make predictions.XGBoost builds decision trees sequentially, with each subsequent tree attempting to correct the errors of the previous tree. It uses a gradient descent algorithm to minimize a loss function that measures the difference between the predicted and actual values of the target variable. In each iteration, XGBoost fits a new decision tree to the negative gradient of the loss function, which helps it to focus on the data points that are hard to classify.One of the key strengths of XGBoost is its ability to handle missing values, outliers, and skewed distributions in the data.

It also provides options for regularization, such as L1 and L2 regularization, which can help to prevent overfitting and improve the generalization performance of the model.XGBoost uses a number of hyperparameters that can be tuned to optimize the performance of the model. These include the learning rate, the number of trees, the depth of the trees, and the minimum number of samples required to split a node.XGBoost has become a popular choice for machine learning competitions and is widely used in industry for a variety of applications, such as fraud detection, recommender systems, and image classification.

RandomizedSearchCV:

It is a technique used for hyperparameter tuning in machine learning. It is a variant of grid search, which is a method for searching through a specified subset of hyperparameters to find the best combination of hyperparameters for a given model. However, unlike grid search, which searches through all possible combinations of hyperparameters, RandomizedSearchCV searches through a randomized subset of hyperparameters.RandomizedSearchCV works by specifying a range or distribution for each hyperparameter to be tuned, and then randomly selecting a combination of hyperparameters to be tested. This helps to reduce the computational cost of hyperparameter tuning, particularly when the hyperparameter space is large.The algorithm performs a specified number of iterations, each time randomly selecting a combination of hyperparameters from the specified ranges or distributions.

The performance of the model is then evaluated using cross-validation, and the hyperparameters that give the best performance are recorded. After all iterations are completed, the hyperparameters that result in the best performance are returned as the final set of hyperparameters for the model.RandomizedSearchCV is particularly useful when the search space for hyperparameters is large and it is computationally expensive to search through all possible combinations. It also helps to avoid overfitting to the validation set by evaluating the performance of the model on a different set of data during each iteration.

C. Implementation:

Anaconda installation:

The first step is to install Anaconda, which is a popular Python distribution that includes Jupyter Notebook and many useful libraries.

Launching Anaconda Navigator: Once the Anaconda is installed, launch Anaconda Navigator. This will open a graphical interface that allows a user to launch Jupyter Notebook and other tools.

Create a new environment:

It's good practice to create a new environment for each machine learning project. This allows a user to isolate the dependencies and avoid version conflicts. To create a new environment, click on the "Environments" tab in Anaconda Navigator, then click on the "Create" button. Give your new environment a name and select the Python version you want to use.

Installing necessary libraries:

Once the environment is created, the user will need to install the required libraries such as NumPy, Pandas, Scikit-learn, etc.

Launch Jupyter Notebook:

Once the libraries are installed, go back to Anaconda Navigator and click on the "Home" tab. From here, you can launch Jupyter Notebook by clicking on the "Launch" button.

Create a new notebook:

Once Jupyter Notebook is launched, you can create a new notebook by clicking on the "New" button on the top right corner of the page and selecting "Python 3" to create a new notebook.

Import libraries:

The libraries are imported in the first cell of the notebook so that it can be used in the model to diagnose the heart disease.

Load data:

Load the dataset which contains the patient's data such as age, sex, cp, tresbps, chol, fbs, restecg, thalach,exang, oldpeak, slope, ca, thal, num are the attributes which will be used by the model to predict the heart disease.

Split the data:

Splitting your dataset into training and testing sets is an important step in machine learning. It allows you to train your model on a subset of the data and evaluate its performance on a different subset. This helps you to assess how well your model will generalize to new, unseen data.X_train and y_train to train your model, and X_test and y_test to evaluate its performance.

Build the model:

Build the machine learning model using the training data. The algorithms such as logistic regression, k-nearest neighbour, support vector machine and XGBoost were used to build the model.

Make predictions:

The trained model analyse the dataset and give the results which shows the total number of people are having heart disease and the people who don't have heart disease.

Evaluate the model: The performance of the model was evaluated by various metrics such as accuracy, precision, recall, etc.

C. Dataset Description

The attributes in the dataset is explained as follows:

Table 1: Description of Attributes of Heart Diseases

S.No	Attribute	Description	Values
1.	Age	Age of the person	29 to 79
2.	Sex	Gender[male:1,female:0]	0,1
3.	Cp	Chest pain type [1-Typical type angina 2-Atypical type angina 3-Non-angina pain 4-Asymptomatic]	1,2,3,4
4.	Tresbps	Resting blood pressure in mm/hg	94 to 200
5.	Chol	Serum Cholesterol level in mg/dl	126 to 564
6.	Fbs	Fasting blood sugar in mg/hg	0,1
7.	Restecg	RestingElectrocardiograph results	0, 1, 2
8.	Thalach	Maximum heart rate achieved	71-202
9.	Exang	Exercise induced angina	0, 1
10.	OldPeak	ST depression induced by exercise relative to rest	1 to 3
11.	Slope	Slope of the peak exercise	1, 2, 3

		ST segment	
12.	Ca	Number of major vessels coloured by fluoroscopy	0 to 3
13.	Thal	3-Normal, 6-Fixed defect, 7-reversible defect	3, 6, 7
14.	Num	Class attribute	0 or 1

Using a few of the dataset's characteristics as a summary, we can say that the patients in the dataset ranged in age from 29 to 79; male patients received a value of 1, while female patients received a value of 0. Four categories were established to represent various forms of heart illness. Type 1 is characterized by typical angina, which occurs when the heart's blood flow is compromised and causes chest pain. Type 2 refers to Atypical Angina, Type 3 to Non-Angina Pain, and Type 4 to be Asymptomatic. The fourth feature of the dataset was Trestbps which is the resting blood pressure measure ranging from 94 to 200.The following characteristic, Chol, has a value between 126 and 564. If the blood sugar level was above 120 mg/dl, it was marked as 0, and if it was below, it was denoted as 1. The thalass, which ranges from 71 to 202, is the highest cardiac rate ever attained. Exang was assigned a value of 0 in the absence of pain and 1 in the presence of pain. If a patient has been diagnosed with heart disease, the target or num attribute is indicated as 1, and 0 for healthy individuals.

II. Motivation

The development of a model for predicting the presence of heart disease is the primary driving force behind this endeavour. Also, this research endeavour tries to find the classification algorithm with the highest accuracy for predicting heart disease. By comparing four algorithms—Logistic Regression, K-Nearest Neighbor, Support Vector Machine, and XGBoost classifier—this work is justified. These algorithms have been applied and evaluated based on their performance at various levels of analysis and analysis methodologies. With the help of this project, the model will be improved in order to find the best approach and better way to find theheart disease.

III. Application

Electrocardiogram (ECG) is a standard diagnostic test for heart diseases. Machine learning algorithms can analyze the ECG signal to identify patterns and abnormalities. For example, deep learning models can detect arrhythmias, heart blocks, and ST segment abnormalities.

Cardiac imaging techniques like echocardiography and magnetic resonance imaging (MRI) can provide detailed images of the heart structure and function. Machine learning algorithms can analyze these images to identify the presence of heart disease, measure cardiac function, and track disease progression.

Machine learning algorithms can analyze patient data, including demographics, medical history, and laboratory results, to predict the risk of heart disease. These algorithms can help identify patients who are at high risk of developing heart disease, allowing for early intervention and better outcomes.

Machine learning algorithms can assist healthcare providers in making diagnostic and treatment decisions. For example, a machine learning algorithm can analyze patient data and provide recommendations on the appropriate diagnostic tests or treatment options based on the patient's specific condition.

IV. Conclusion

Machine learning has the potential to aid in the diagnosis of heart disease by analyzing large amounts of data and detecting patterns that may not be easily noticeable to humans. By using machine learning, healthcare providers can improve the accuracy and efficiency of the diagnostic process, leading to better patient outcomes. However, the use of machine learning should be accompanied by clinical judgment and expertise to ensure the best possible patient care. Further research and development in this field will continue to improve the capabilities and effectiveness of machine learning in the diagnosis of heart disease.As the field of machine learning continues to advance, we can expect further developments and improvements in the diagnosis of heart disease, leading to better outcomes for patients.

Machine learning algorithms can help healthcare providers to detect heart disease at an early stage, identify patients at high risk of developing heart disease, and customize treatment plans based on individual patient characteristics. This can lead to better outcomes for patients, including reduced morbidity and mortality rates, improved quality of life, and more cost-effective healthcare delivery.

The model incorporates the four methods. The datasets were trained separately for each algorithm. All of them were then put to the test. On the basis of a number of factors, the most effective algorithm had to be chosen. With an accuracy of 82.89%, I discovered that the Logistic Regression technique is the most effective. The accuracy of the classifiers K-Nearest Neighbor, Gradient Boosting, and SVM were 80.43%, 80.43%, and 81.57%, respectively. The model gives an high accuracy of 87% with the help of RandomizedSearchCV. The users would benefit from receiving a preliminary assessment of their heart's health in this way. Application of a promising technology, such as machine learning, to the initial prediction of heart problems will have a significant impact on the situation because heart diseases are a leading cause of death in India and throughout the world.As such, the ongoing integration of machine learning into clinical practice is likely to play an increasingly important role in the prevention, diagnosis, and treatment of heart disease in the years ahead.

V. Future Work

The future work is to develop a web application to get the data from the user to diagnose the heart disease. It works in the backend using python with machine learning algorithms. In the frontend Html, CSS and Php will be used to build the application. The Html and CSS are helpful in creating the webpage and make them interactive for better user experience. The web application can be created by using the flask with python. It also has a separate section for each users where the users can register with their personal details such as email, username, mobile number and password and then they can login with the registered details.This web application takes input from the users such as age, gender, blood sugar, stress level, cholesterol level, blood pressure, etc. The inputs goes under the machine learning process to predict the precence of the heart disease.

Once a machine learning model has been developed, it must be validated in clinical settings to ensure that it can accurately predict heart disease in real-world scenarios. This would involve testing the model on new patient data and comparing its predictions to those made by human experts. Researchers would also need to determine how the model can best be integrated into clinical workflows to optimize patient outcomes.

Reference

[1] E. Nasarian, M. Abdar, M. A. Fahami, R. Alizadehsani, S. Hussain, M. E. Basiri, M. Zomorodi-Moghadam, X. Zhou, P. Pławiak, U. R. Acharya, R.-S. Tan, and N. Sarrafzadegan, ‘‘Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach,’’ Pattern Recognit. Lett., vol. 133, pp. 33–40, May 2020.

[2] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, ‘‘Prediction of heart disease using a combination of machine learning and deep learning,’’ Comput. Intell. Neurosci., vol. 2021, pp. 1–11, Jul. 2021.

[3] A. Kishore, A. Kumar, K. Singh, M. Punia, and Y. Hambir, ‘‘Heart attack prediction using deep learning,’’ Int. Res. J. Eng. Technol., vol. 5, no. 4, p. 2395, 2018.

[4] L. Wang, W. Zhou, Q. Chang, J. Chen, and X. Zhou, ‘‘Deep ensemble detection of congestive heart failure using short-term RR intervals,’’ IEEE Access, vol. 7, pp. 69559–69574, 2019.

[5] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, ‘‘Heart disease prediction system using associative classification and genetic algorithm,’’ in Proc. Int. Conf. Emerg. Trends Elect., Electron. Commun. Technol. (ICECIT), 2012, pp. 40–46.

[6] A. Gupta, L. Kumar, R. Jain, and P. Nagrath, ‘‘Heart disease prediction using classification (naive bayes),’’ in Proc. 1st Int. Conf. Comput., Commun., Cyber-Secur. (ICS). Singapore: Springer, 2020, pp. 561–573. A. Ahmed and S. A. Hannan, ‘‘Data mining techniques to find out heart diseases: An overview,’’ Int. J. Innov. Technol. Exploring Eng., vol. 1, no. 4, pp. 18–23, 2012.

[7] A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, ‘‘Classification models for heart disease prediction using feature selection and PCA,’’ Informat. Med. Unlocked, vol. 19, Jan. 2020, Art. no. 100330.

[8] Z. Sani, R. Alizadehsani, J. Habibi, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, F. Khozeimeh, and F. Alizadeh-Sani, ‘‘Diagnosing coronary artery disease via data mining algorithms by considering laboratory and echocardiography features,’’ Res. Cardiovascular Med., vol. 2, no. 3, p. 133, 2013.

[9] E. O. Olaniyi, O. K. Oyedotun, and K. Adnan, ‘‘Heart diseases diagnosis using neural networks arbitration,’’ Int. J. Intell. Syst. Appl., vol. 7, no. 12, p. 72, 2015.

[10] D. Swain, S. K. Pani, and D. Swain, ‘‘A metaphoric investigation on prediction of heart disease using machine learning,’’ in Proc. Int. Conf. Adv. Comput. Telecommun. (ICACAT), Bhopal, India, Dec. 2018, pp. 1–6.

[11] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, ‘‘Can machine learning improve cardiovascular risk prediction using routine clinical data?’’ PLoS ONE, vol. 12, no. 4, Apr. 2017, Art. no. e0174944.