International Journal for Multidisciplinary Research (IJFMR)

E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Heart Disease Risk Assessment by Using LightGBM Technique

N.M.K. Ramalingamsakthivelan[1], V. Silambarasan[2], S. Thavasi[3], P. Vijaya Shankar[4]

[1]Associate Professor, Department of Computer Science and Engineering, Paavai Engineering College, Namakkal.

[2,3,4]Student, Department of Computer Science and Engineering, Paavai Engineering College, Namakkal.

Abstract

Heart disease is a term that covers various illnesses affecting the heart, which may involve blood vessels, heart rhythm, or congenital abnormalities. The World Health Organization (WHO) has identified heart disease as one of the primary causes of death globally. Heart disease includes conditions such as coronary artery disease and arrhythmias, among others. Predicting Heart disease risk can help doctors accurately assess patients' health, allowing for early intervention and lifestyle changes or medical treatment as needed. Machine learning (ML) can aid in understanding and reducing symptoms of heart disease by using key parameters such as heart rate, body temperature, and blood pressure. To more accurately predict heart disease risk and alert medical professionals and caregivers to patient location, the Light Gradient Boosting Machine (LightGBM) technique is proposed. LightGBM modelling is a promising classification strategy for predicting medication adherence in heart disease patients, assisting with patient stratification and decision-making based on the best available data. The chi-square statistical test is utilized to select specific features from the Cleveland heart disease (HD) dataset, and data visualization is used to depict feature relationships. In 303 instances of Cleveland HD dataset attributes, the random forest algorithm achieved an 88.5% accuracy rate during validation, according to experiment results.

Keywords: Heart disease, Machine learning, Light Gradient Boosting Machine (LightGBM) technique, chi-square statistical test.

1.Introduction

The circulatory system, also referred to as the cardiovascular system or blood-vascular system, comprises the heart, a muscular pump, and a closed network of blood vessels that include arteries, veins, and capillaries. The heart pumps blood in a closed circle or circuit through the different circulations of the body. The flow of blood through the capillaries is crucial for maintaining homeostasis in every tissue and cell of the body.

The various activities and components of the cardiovascular system must be integrated, regulated, and coordinated to provide blood to specific body locations as needed. Detecting heart disease is a major challenge. Although there are tools available to forecast heart disease, they may be expensive or ineffective at predicting the likelihood of heart disease in an individual.

Early identification of heart conditions can reduce mortality rates and overall consequences. However, accurately monitoring patients on a daily basis requires significant resources, time, and expertise. In the modern era, large amounts of medical data can be analyzed using machine learning algorithms to uncover hidden patterns.

These patterns may be used to improve health diagnosis and treatment based on medical data. To interpret CHD, it is recommended to employ the machine learning technique Support Vector Machine (SVM). SVM is known for its ability to automatically learn a hierarchical feature representation from raw data, eliminating the need for manual feature engineering.

This system employs wearable devices like smartwatches or fitness trackers to monitor various physiological parameters such as heart rate, blood pressure, and activity levels. The data generated by these devices is then transmitted to a central database, where it undergoes analysis using machine learning algorithms to identify patterns or anomalies that may indicate a potential Heart problem.

Upon detecting any issue, the system sends alerts to the patient's mobile device and healthcare provider, enabling early intervention and treatment. Furthermore, the system can furnish patients with tailored feedback on their lifestyle choices, such as diet and exercise, to encourage healthy habits and prevent the onset of heart disease.

1.1 Objective

The objective of developing a machine learning framework for diagnosing heart disease is to enhance the system's capability to predict heart disease and improve patient survival rates by enabling accurate, precise, and early detection of the disease. The framework should also provide a mechanism for alerts and recommendations.

2.Literature Survey

K. Polaraju et al conducted a study using the Multiple Regression Model to predict the likelihood of heart disease, using a training dataset consisting of 3000 instances with 13 attributes. The dataset was divided into a 70% training and 30% testing set, and the results showed that the regression algorithm outperformed other algorithms in terms of classification accuracy.

S. Seema et al focused on predicting chronic diseases by mining historical health records using various classification algorithms, including Naïve Bayes, Decision Tree, Support Vector Machine (SVM), and Artificial Neural Network (ANN). The study found that SVM had the highest accuracy rate for predicting chronic diseases, while Naïve Bayes performed best for predicting diabetes.

Purushottam et al proposed an efficient heart disease prediction system that assists medical practitioners in making effective decisions based on specific parameters. The system achieved 86.3% accuracy in the testing phase and 87.3% accuracy in the training phase.

K. Gomathi et al suggested using data mining techniques to predict multiple diseases, including heart disease, diabetes, and breast cancer, to reduce the number of tests required.

3.Existing System

The previous system utilized wearable sensors to collect medical attributes such as blood pressure, temperature, and glucose levels to predict heart disease. Various machine learning algorithms including Naïve Bayes, Decision Tree, and Support Vector Machine were employed in the data mining process.

Naïve Bayes was used to extract relevant information from the dataset, Decision Tree was utilized for predicting cardio arrest, and Support Vector Machine was employed to generate alerts when certain patterns were detected in the medical attribute graph. The entire project was developed based on the principles of data mining.

4. Proposed System

The proposed system aims to predict cardiac arrest and send alerts to the patient's guardian and doctor with location and medical report. The system uses the Support Vector Machine algorithm to train the dataset and the X2 statistical model to eliminate irrelevant features. The Light GBM algorithm is used as a classifier to predict the cardiac arrest.

Google Maps is used to locate the user during a cardiac arrest. The doctor can monitor the patient's medical history and provide the necessary prescriptions. The guardian can also monitor the patient's data and receive alerts during a cardiac arrest.

5. ARCHITECTURE

Here the system divided into two phases such as training phase and testing phase. The training phase contains dataset pre-processing, feature selection, feature extraction, classifier (LightGBM), database. The testing phase contains pre-processing, feature selection, feature extraction, prediction, alert, recommendation system. Here we are using WAMP for server and Python for backend work. The HTML, CSS, JavaScript for frontend work. The flask framework was help to connect Machine learning in this project.

Here we are using 28 attributes and 499 datasets for predict the disease. The system is connected with IOT device (smart watches). That device monitors the data and send to the system. The system classifies the given data for produce result (either cardiac arrest found or not). That time system provides alert message (SMS) for the guardian who is register for the respective patient.

The guardian also gets the current location of the patient. The medical practitioner can access the patient's information through the system. Doctor can suggest the tablet for each patient as per the current data. The patient can get suggestion from the doctor by the system. The patient can get rescue from the guardian at right time.

5.1 Patient Module

The Registration module serves as a comprehensive patient management system that captures relevant patient information. It enables new users to create an account and log in to the web UI.

The Login module focuses on user security, handling user logons and authentications.

In this system, patients can update their guardian details and provide disease-related attributes for prediction. The system generates prediction results and provides medical suggestions from doctors. This feature helps avoid travel due to a patient's illness.

5.2 Doctor Module

The doctor's system holds the primary database of all the patients, allowing for monitoring of all the patients. The module provides the doctor with patient information, enabling them to give suitable suggestions and prescribe medications based on the patient's current medical condition.

Live monitoring and guidance assist the doctor in reducing the patient's risk of death. Immediate action to the disease is possible through the web UI, including suggestions and prescription of medicine.

5.3 Guardian Module

The Patient Caretaker module is designed for the patients' caretakers, who can access their patients' health status by logging into the web UI. This module also receives emergency alert messages. The web UI displays the patient's current medical condition and location using Google Maps.

If the patient is unconscious, the caretaker can easily locate the patient through the map and provide immediate attention. The alert message notifies the caretaker about the patient's situation, ensuring timely care.

5.4 Admin Module

The admin module serves as the central control unit for the entire system. It maintains a database of all available doctors in the web UI. It also includes the dataset used for prediction, as well as data preprocessing, feature selection, and feature extraction.

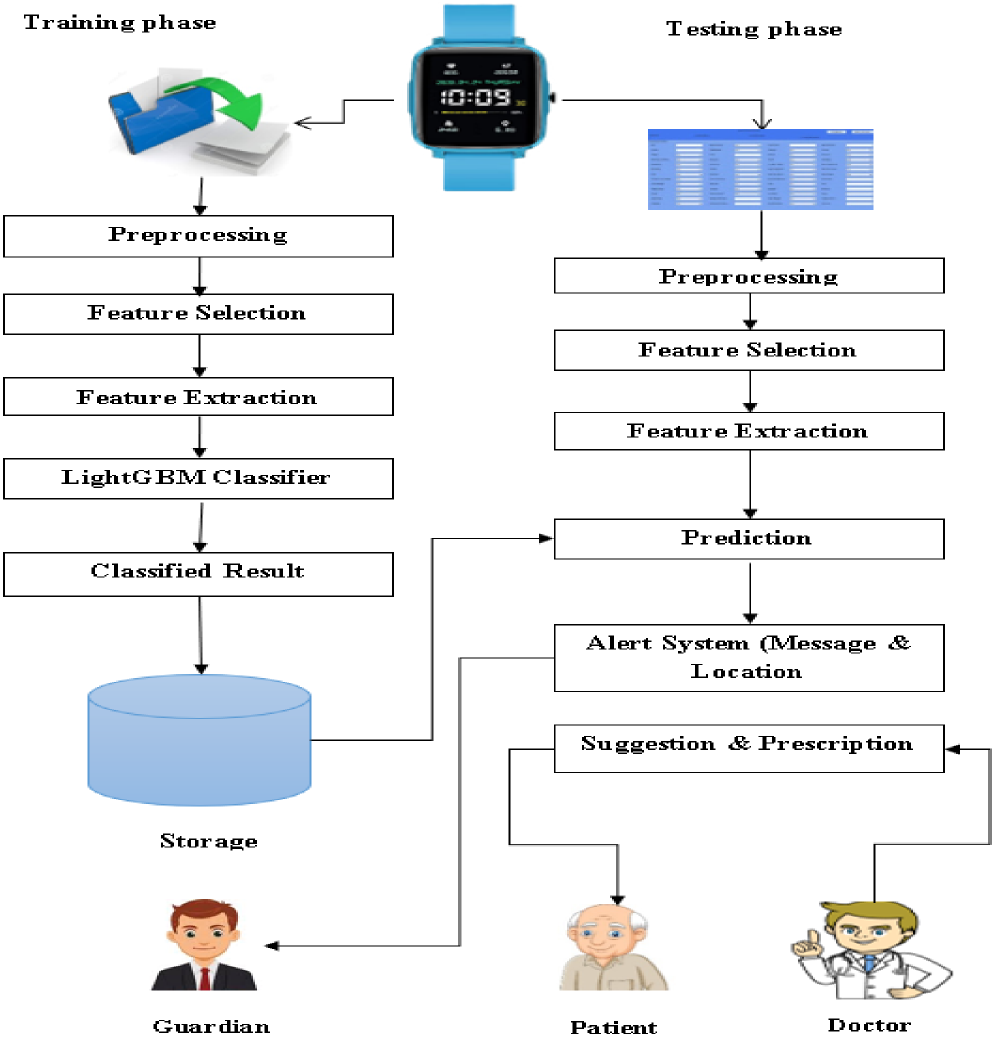
The classification process is also managed within the admin module. Overall, the admin module is responsible for coordinating and managing the different components of the system to ensure smooth and efficient operation.

5.5 Alert Module

The system incorporates an alert service to remind patients to take their medication at the scheduled time, which can be sent to their guardian's mobile device. This feature is particularly beneficial for busy individuals or the elderly who may forget their medication schedule.

In addition, the system allows patients to request medical advice from their doctor through Short Message Service (SMS), which includes their blood pressure, blood glucose, temperature, and heartbeat readings. Patients are required to manually attach their readings to the message for the doctor's reference.

Figure 1: Heart Disease Monitoring and Alert System



6. Methodologies

6.1 Feature Selection

The network is designed to incorporate only the most significant or relevant features while removing irrelevant ones. Including all features in the network can lead to overfitting and negatively impact generalization performance.

Therefore, identifying the ideal feature subset by removing noisy features can improve the performance of the network on both training and testing data. To achieve this, the X2 statistical model is used in this module to eliminate irrelevant attributes.

X2 statistics are computed between each non-negative feature and the class (y) as part of the feature reduction process. The X2 model performs an X2 test to determine the dependency of the features on the class, allowing it to remove features that are likely to be class-independent and therefore considered unimportant for categorization.

Figure 2: Chi-Square in Dataset

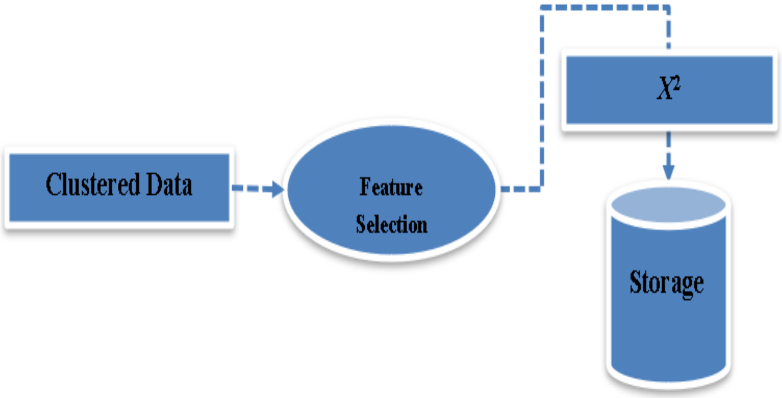
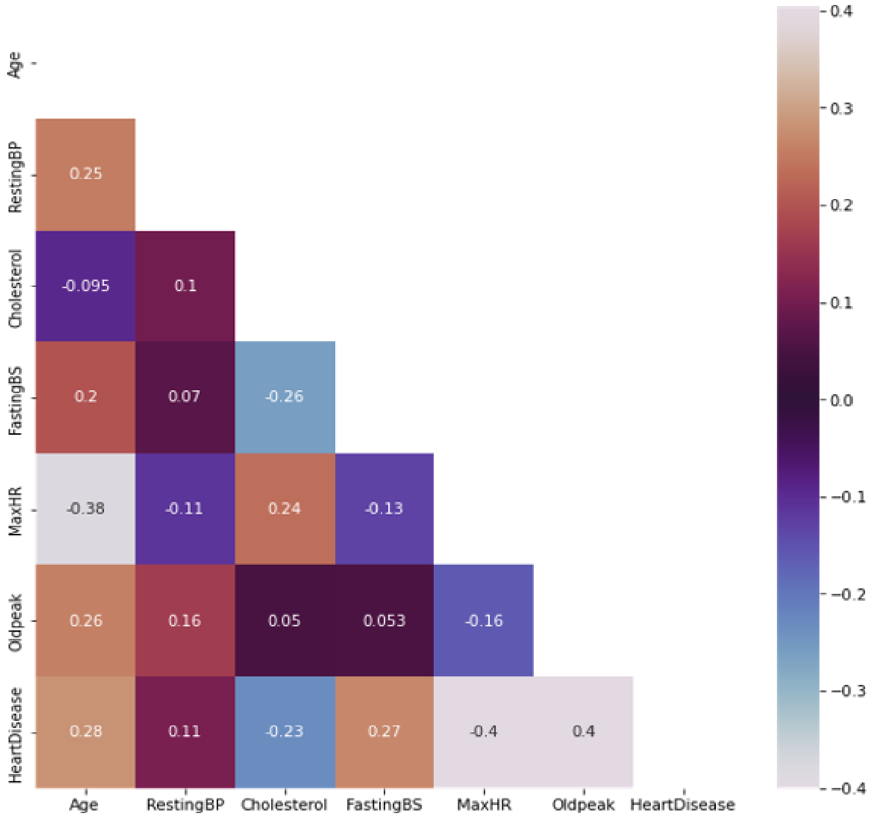


Figure 3: LightGBM's Important Attributes



6.2 LightGBM CVD Classification

The LightGBM gradient boosting framework is a tree-based learning method that can be used to address various machine learning problems, including predicting heart diseases. To use LightGBM for predicting heart diseases, a dataset containing features and target variables would be required.

These features might include age, sex, blood pressure, cholesterol levels, smoking status, family history of heart disease, and other relevant information. The target variable would be whether or not a person has heart disease. Once the dataset is prepared, LightGBM would use gradient boosting to build a model that can determine if a person has heart disease based on the provided attributes.

During the training process, LightGBM would use the features to build decision trees and then aggregate the output of these trees to make a prediction. During the prediction phase, the model would consider the characteristics of a new patient and use the decision trees to determine whether or not the patient has heart disease. The following table shows important features consider by the LightGBM algorithm:

Table 1: LightGBM’s Important Attributes

Weight	Feature
0.0838 ± 0.0122	num_major_vessels
0.0676 ± 0.0337	chest_pain_type
0.0171 ± 0.0424	st_slope
0.0018 ± 0.0177	max_heart_rate_achieved
0 ± 0.0000	thalassemia
0 ± 0.0000	exercise_induced_angina
0 ± 0.0000	rest_electrocardiographic
0 ± 0.0000	fasting_blood_sugar
0 ± 0.0000	resting_blood_pressure
0 ± 0.0000	sex
-0.0045 ± 0.0057	st_depression
-0.0063 ± 0.0092	cholesterol

Weight	Feature
-0.0072 ± 0.0044	age

6.3 Prediction

In machine learning algorithms, it is common to calculate the distance between data points, which is measured using various distance metrics in n-dimensional space. The Euclidean distance is one of the popular distance measures used to cluster or group together data points with similar characteristics.

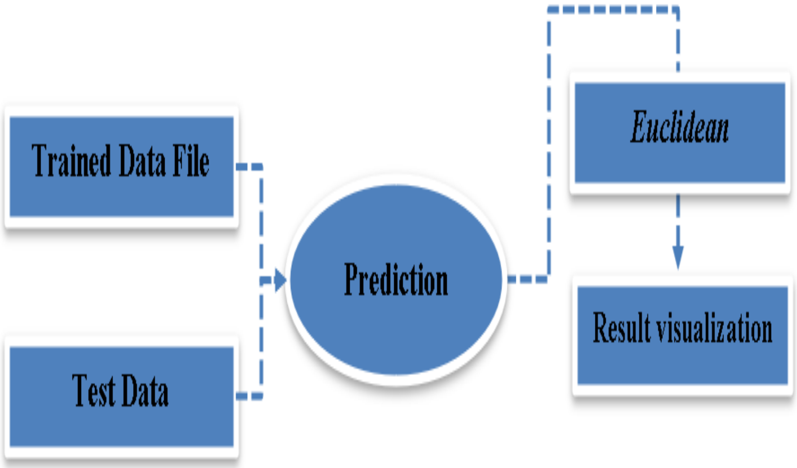
However, in the context of LightGBM, a gradient boosting technique that predicts outcomes using decision trees, the Euclidean distance is not directly used for predicting Heart illness. Instead, it is utilized to group similar data points together based on their features during the training phase. LightGBM divides the dataset into smaller subgroups and optimizes the difference between the target variable values in these subsets using a loss function to modify the decision trees' weights.

During this process, LightGBM leverages several distance metrics, including the Euclidean distance, to measure the similarity between data points and group them together. By doing so, the algorithm gathers relevant data from these groups to produce more accurate predictions.

However, LightGBM uses other distance measures as well, and the choice of distance metric depends on the type of data being used and the problem being solved. For instance, when predicting heart disease, features such as age, sex, blood pressure, cholesterol level, smoking status, family history, and others would be used to train the model and make predictions.

To predict the results, the training and testing phases are synchronized using the LightGBM technique. The Euclidean method is utilized to identify the most significant attribute among the 28 attributes present in the dataset.

Figure 4: Prediction model



7. Result and Discussion

The heart disease risk assessment and monitoring system achieved an accuracy of 83.67% in predicting the occurrence of heart disease using the testing dataset. The system also showed an AUC and ROC accuracy of 96.7%, while the precision, recall, and f1-score were all 97.5%. These results are superior to those of recently published studies.

The LightGBM technique was chosen for this project due to its high accuracy, efficient memory usage, and fast training times. The system could handle missing variables and was easy to use. The developed system used 28 input variables from the clinical dataset of Cleveland Clinic patients for training and testing.

During the testing phase, the system classified the given data to determine whether a cardiac arrest was present or not. Additionally, the system was connected to an IoT device, such as a smartwatch, which monitored the patient's data and sent it to the system.

When cardiac arrest was detected, the system sent an alert message to the registered guardian, providing the patient's current location.

The system also provided the patient details to the doctor, who could suggest medication based on the current data.

Furthermore, the patient could receive suggestions from the doctor via the system, and the guardian could provide timely assistance to the patient.

7.1. ROC and AUC

The area under the receiver operating characteristic (ROC) curve is commonly referred to as the AUC, which represents the classifier's performance in distinguishing between positive and negative instances. In the ROC curve, the x-axis corresponds to the false positive rate (FPR), and the y-axis corresponds to the true positive rate (TPR).

The ROC curve offers an intuitive way to visualize the trade-off between sensitivity and specificity. A high AUC value, close to 1, indicates a better classifier performance, where the FPR is low, and the TPR is high. Therefore, improving the AUC value is a desirable goal in developing a predictive model.

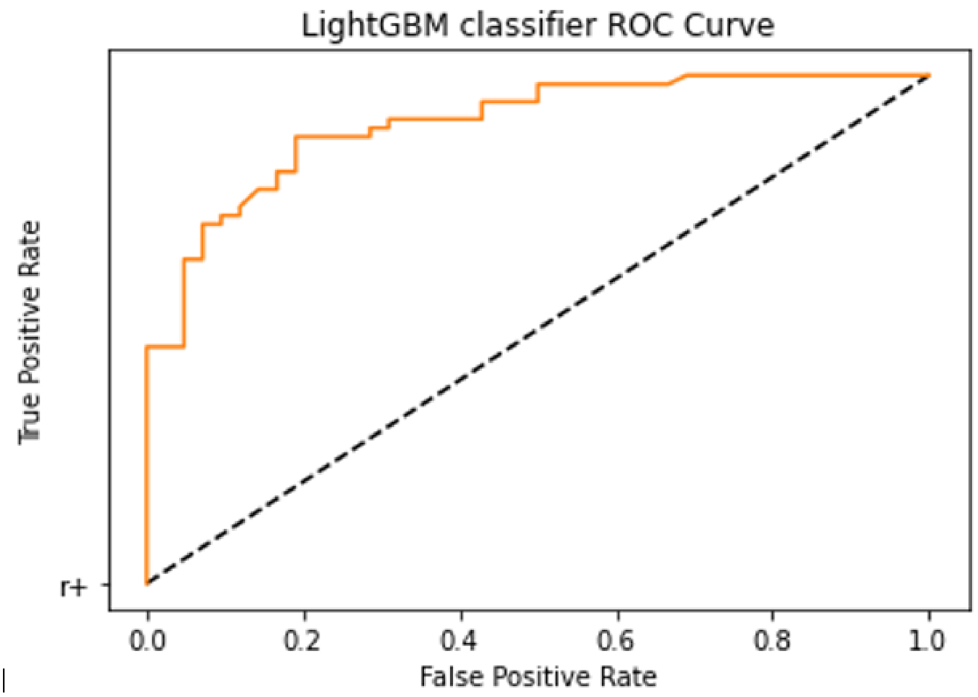
Table 2: Value for Prediction Model result

Classifiers	**Accuracy**	**Recall**	**F1**
lightGBM	**0.97 ± 0.060**	**0.922 ± 0.068**	**0.980 ± 0.038**

Table 3: Value for Prediction Model result

Classifiers	**Precision**	**Specificity**	**AUC**
lightGBM	**0.963 ± 0.041**	**0.881 ± 0.095**	**0.93 ± 0.05**

Figure 5: AUC and ROC



The optimal model is the LIGHTGBM tree model, which achieved an accuracy rate of 96.7% and an f1-score, recall, and precision of 97.5%.

8.Conclusion

LightGBM is a machine learning algorithm known for its high accuracy, fast training times, efficient memory usage, ability to handle missing data, and ease of use. In this study, a LightGBM-based classification and prediction model was developed to predict the risk of heart disease using a clinical dataset from Cleveland Clinic patients. The model was trained and tested using 28 input variables.

The results show that the LightGBM model has an accuracy of 83.67% in diagnosing heart disease, with a misclassification rate of 16.33% based on the testing data. The model also has an AUC and ROC accuracy of 96.7%, and a precision, recall, and f1-score of 97.5%, which is superior to the results of recent studies.

The model can provide reliable and accurate diagnoses for coronary heart disease, reducing the potential harm caused by incorrect diagnoses. The machine learning model can benefit patients and healthcare professionals worldwide in improving public and global health, especially in under-resourced regions where there is a shortage of cardiac specialists.

9.Future Enhancement

In future studies, we plan to investigate additional techniques that may further enhance the diagnostic performance and accuracy of LightGBM models in detecting heart disease among patients worldwide.

Our ultimate objective is to integrate this system with smartwatch technology, leveraging its sensor capabilities and expanding its functionality to include features such as alert notifications, as well as the ability to identify nearby hospitals and ambulance drivers for emergency response.

References

1. K. Polaraju, D. Durga Prasad, "Prediction of Heart Disease using Multiple Linear Regression Model", International Journal of Engineering Development and Research Development, ISSN:2321-9939, 2017.
2. Marjia Sultana, Afrin Haider, "Heart Disease Prediction using WEKA tool and 10-Fold cross-validation", The Institute of Electrical and Electronics Engineers, March 2017.
3. Dr.S.Seema Shedole, Kumari Deepika, "Predictive analytics to prevent and control chronic disease",https://www.researchgate.net/punlication/316530782, January 2016.
4. Ashok kumar Dwivedi, "Evaluate the performance of different machine learning techniques for prediction of heart disease using ten-fold cross-validation", Springer, 17 September 2016.
5. Megha Shahi, R. Kaur Gurm, "Heart Disease Prediction System using Data Mining Techniques", Orient J. Computer Science Technology, vol.6 2017, pp.457-466.
6. Mr. Chala Beyene, Prof. Pooja Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques", International Journal of Pure and Applied Mathematics, 2018.
7. R. Sharmila, S. Chellammal, "A conceptual method to enhance the prediction of heart diseases using the data techniques", International Journal of Computer Science and Engineering, May 2018.
8. Jayami Patel, Prof. Tejal Upadhay, Dr. Samir Patel, "Heart disease Prediction using Machine Learning and Data mining Technique", March 2017.
9. Purushottam, Prof. (Dr.) Kanak Saxena, Richa Sharma, "Efficient Heart Disease Prediction System", 2016, pp.962-969.
10. K Gomathi, Dr.D. Shanmuga Priyaa, "Multi Disease Prediction using Data Mining Techniques", International Journal of System and Software Engineering, December 2016, pp.12-14.
11. Mr.P.Sai Chandrasekhar Reddy, Mr.Puneet Palagi, S.Jaya, "Heart Disease Prediction using ANN Algorithm in Data Mining", International Journal of Computer Science and Mobile Computing, April 2017, pp.168- 172.
12. Ashwini Shetty A, Chandra Naik, "Different Data Mining Approaches for Predicting Heart Disease", International Journal of Innovative in Science Engineering and Technology, Vol.5, May 2016, pp.277- 281.