© May 2026 | IJIRT | Volume 12 Issue 12 | ISSN: 2349-6002

Heart Disease Prediction Using Machine Learning

Narendar E[1], Ms.Arthi Priyadharshini[2], Ponkishore S[3], Shiraj M[4]
[1,3,4]*Student, Department of Computer Science and Engineering, Dhanalakshmi College of Engineering, Chennai, India*
[2]*Assistant Professor, Department of Computer Science and Engineering, Dhanalakshmi College of Engineering, Chennai, India*

***Abstract*—heart disease is one of the major causes of death worldwide, and early prediction can help reduce mortality rates through timely medical intervention. This project presents a machine learning-based heart disease prediction system that analyzes patient health parameters to predict the likelihood of heart disease. The dataset used for this study contains medical attributes such as age, blood pressure, cholesterol level, heart rate, chest pain type, and other clinical factors. Various machine learning algorithms including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) were implemented and evaluated to identify the most accurate prediction model. Data preprocessing techniques such as handling missing values, normalization, and feature selection were applied to improve model performance. The models were trained and tested using standard evaluation metrics including accuracy, precision, recall, F1-score, and confusion matrix. Among the implemented models, the Random Forest algorithm achieved the highest prediction accuracy, demonstrating its effectiveness in heart disease prediction.**

***Index Terms*—Heart Disease Prediction, Machine Learning, Healthcare Analytics, Early Disease Detection, Clinical Data Analysis**

I. INTRODUCTION

Heart disease is one of the leading causes of death across the world and has become a major public health concern. Cardiovascular diseases affect millions of people every year, and early detection plays a crucial role in reducing mortality rates and improving patient health outcomes. Traditional methods of diagnosing heart disease often depend on medical experts, laboratory tests, and complex clinical procedures, which can be time-consuming and expensive. Therefore, there is a growing need for intelligent systems that can assist healthcare professionals in predicting heart disease at an early stage [1].

Machine Learning (ML), a branch of Artificial Intelligence (AI), has emerged as a powerful technology in the healthcare sector. Machine learning algorithms are capable of analyzing large volumes of medical data, identifying hidden patterns, and making accurate predictions. By using patient health records and clinical parameters such as age, blood pressure, cholesterol level, heart rate, chest pain type, fasting blood sugar, and electrocardiographic results, machine learning models can predict the possibility of heart disease efficiently. [2].

This project focuses on developing a Heart Disease Prediction System using machine learning techniques. Various classification algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) are implemented and compared to determine the most accurate prediction model. The dataset is preprocessed using techniques such as data cleaning, normalization, and feature selection to improve prediction performance [3].

The primary objective of this project is to build a reliable and efficient prediction model that can assist doctors and healthcare institutions in early diagnosis and decision-making. Early prediction of heart disease can help patients receive timely treatment and reduce the risk of severe complications. This project demonstrates how machine learning can be effectively applied in the medical field to support intelligent healthcare systems and improve the quality of patient care [4].

Heart disease prediction using machine learning involves analyzing various health parameters such as age, blood pressure, cholesterol level, blood sugar, chest pain type, heart rate, electrocardiogram results, and other clinical attributes. By studying these features, machine learning models can determine whether a patient is at risk of heart disease [5].

Traditional methods of heart disease diagnosis often require multiple medical tests and expert analysis, which may increase both cost and diagnosis time. Machine learning-based prediction systems can provide quick and efficient results, reduce the workload of healthcare professionals and improve the overall diagnosis process.[6]

In this project, several machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) are used to predict heart disease. These algorithms are trained using historical patient datasets and evaluated using performance metrics to identify the best-performing model.[7]

Data preprocessing plays an important role in improving the performance of machine learning models. Techniques such as handling missing values, normalization, feature selection, and data cleaning are applied to prepare the dataset for accurate prediction. Proper preprocessing helps improve model efficiency and reduces prediction errors.[8]

The main objective of this project is to develop an intelligent and reliable heart disease prediction system that can assist healthcare professionals in early diagnosis and treatment planning. Early prediction can help patients take preventive measures and reduce the chances of life-threatening conditions.[9]

The proposed system also aims to provide a cost-effective healthcare solution that can be used in hospitals, clinics, and remote healthcare centers. By integrating machine learning into medical applications, healthcare services can become faster, smarter, and more accessible to patients.[10]

This project demonstrates the importance of machine learning in modern healthcare and highlights its potential to transform medical diagnosis systems. The successful implementation of heart disease prediction models can contribute to better patient care, improved healthcare management, and reduced mortality caused by cardiovascular diseases.[11]

II. LITERATURE SURVEY

Senthil Kumar Mohan et al,[1] proposed Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques in which strategy that objective is to finding critical includes by applying Machine Learning bringing about improving the exactness in the expectation of cardiovascular malady. The expectation model is created with various blends of highlights and a few known arrangement strategies. We produce an improved exhibition level with a precision level of 88.7% through the prediction model for heart disease with hybrid random forest with a linear model (HRFLM) they likewise educated about Diverse data mining approaches and expectation techniques, Such as, KNN, LR, SVM, NN, and Vote have been fairly famous of late to distinguish and predict heart disease

Sonam Nikhar et al [2] has built up the paper titled as Prediction of Heart Disease Using Machine Learning Algorithms by This exploration plans to give a point-by-point portrayal of Naïve Bayes and decision tree classifier that are applied in our examination especially in the prediction of Heart Disease. Some analysis has been led to think about the execution of prescient data mining strategy on the equivalent dataset, and the result uncovers that Decision Tree beats over Bayesian classification system.

Aditi Gavhane, Gouthami Kokkula, Isha Pandya, Prof. Kailas Devadkar (PhD), [3] Prediction of Heart Disease Using Machine Learning", In this paper proposed system they used the neural network algorithm multi-layer perceptron (MLP) to train and test the dataset. In this algorithm there will be multiple layers like one for input, second for output and one or more layers are hidden layers between these two input and output layers. Each node in input layer is connected to output nodes through these hidden layers. This connection is assigned with some weights. There is another identity input called bias which is with weight b, which added to node to balance the perceptron. The connection between the nodes can be feedforwarded or feedback based on the requirement.

Abhay Kishore et al,[4] developed Heart Attack Prediction Using Deep Learning in which This paper proposes a heart attack prediction system using Deep learning procedures, explicitly Recurrent Neural System to predict the probable prospects of heart related infections of the patient. Recurrent Neural Network is a very ground-breaking characterization calculation that utilizes Deep Learning approach in Artificial Neural Network. The paper talks about in detail the significant modules of the framework alongside the related hypothesis. The proposed model deep learning and data mining to give the precise outcomes least blunders. This paper gives a bearing and points of reference for the advancement of another type of heart attack prediction platform. Prediction stage

Lakshmana Rao et al,[5] Machine Learning Techniques for Heart Disease Prediction in which the contributing elements for heart disease are more (circulatory strain, diabetes, current smoker, high cholesterol, etc..). So, it is difficult to distinguish heart disease. Different systems in data mining and neural systems have been utilized to discover the seriousness of heart disease among people. The idea of CHD ailment is bewildering, in addition, in this manner, the disease must be dealt with warily. Not doing early identification, may impact the heart or cause sudden passing. The perspective of therapeutic science furthermore, data burrowing is used for finding various sorts of metabolic machine learning a procedure that causes the framework to gain from past information tests, models without being expressly customized. Machine learning makes rationale dependent on chronicled information.

Mr. Santhana Krishnan.J and Dr. Geetha.S, [6] Prediction of heart disease using machine learning algorithm This Paper predicts heart disease for Male Patient using Classification Techniques. The detailed information about Coronary Heart diseases such as its Facts, Common Types, and Risk Factors has been explained in this paper. The Data Mining tool used is WEKA (Waikato Environment for Knowledge Analysis), a good Data Mining Tool for Bioinformatics Fields. The all three available Interface in WEKA is used here; Naive Bayes, Artificial Neural Networks and Decision Tree are Main Data Mining Techniques and through this techniques heart disease is predicted in this System. The main Methodology used for prediction is Decision Trees like CART, C4.5, CHAID, J48, ID3 Algorithms, and Naive Bayes Techniques.

vinash Golande et al,[7] proposed Heart Disease Prediction Using Effective Machine Learning Techniques in which Specialists utilize a few data mining strategies that are available to support the authorities or doctors distinguish the heart disease. Usually utilized methodology utilized are decision tree, k- closest and Naïve Bayes. Other unique characterization-based strategies utilized are packing calculation, Part thickness, consecutive negligible streamlining and neural systems, straight Kernel self-arranging guide and SVM (Bolster Vector Machine). The following area obviously gives subtleties of systems that were utilized in the examination

III. METHODOLOGY

The methodology of the proposed Heart Disease Prediction System involves several steps including data collection, data preprocessing, feature selection, model training, testing, and performance evaluation. Machine learning algorithms are applied to predict the presence of heart disease based on patient medical data.

The methodology for this heart disease prediction project follows a systematic machine learning workflow designed to ensure reliable and interpretable results. The process begins with data acquisition from a clinically validated source, followed by data preparation to handle missing values and outliers. Feature selection and pre-processing techniques are then applied to identify relevant predictors and convert the data into a model-ready format. Multiple machine learning algorithms are selected and optimized through hyper-parameter tuning to compare different approaches. Each model is trained on a stratified subset of the data and validated using standard performance metrics that prioritize both accuracy and clinical safety. This end-to-end pipeline ensures that the final model is robust, medically relevant, and suitable for supporting heart disease diagnosis.

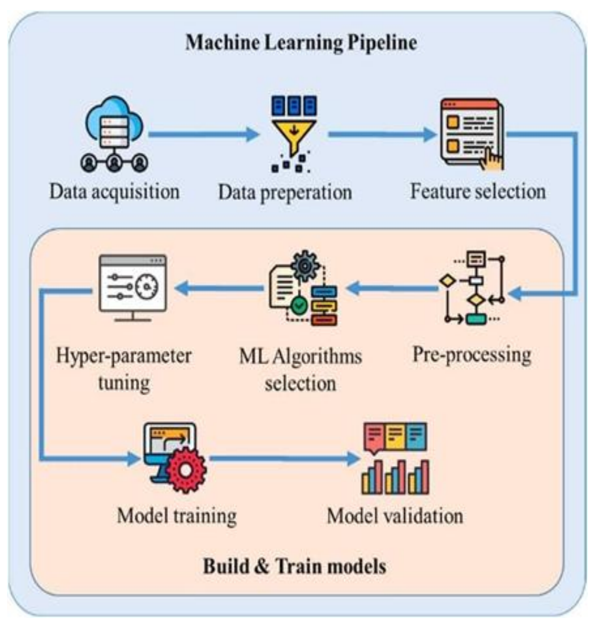

Fig. 1: System Architecture

A. Data Acquisition and data preperation

The dataset for this heart disease prediction project was acquired from the UCI Machine Learning Repository, specifically the Cleveland Heart Disease dataset. It consists of 303 anonymized patient records collected from the Cleveland Clinic Foundation and contains 14 clinical attributes. These features include demographic information like age and sex, as well as medical indicators such as chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG results, maximum heart rate achieved, exercise induced angina, and ST depression. The target variable denotes the presence or absence of heart disease, labeled as 1 or 0 based on physician diagnosis from angiographic tests. Since the dataset is public and de-identified, it is ethically suitable for academic use. For model development, the data was split into training and testing sets in an 80:20 ratio using stratified sampling to preserve the original distribution of disease and non-disease cases.

Data preparation was carried out to improve dataset quality before model training. The raw UCI Heart Disease dataset contained 303 records with 14 features. First, missing values were checked and handled: six records had missing values in the `ca` and `thal` columns, which were imputed using the mode since both are categorical variables. Outliers in numerical features like `cholesterol` and `resting blood pressure` were identified using box plots and clipped to the 1st and 99th percentiles to reduce skew. Duplicate records were verified and none were found. The target column originally had values from 0 to 4 indicating disease severity, but it was converted to binary format where 0 represents no disease and 1-4 were merged into 1 to indicate presence of heart disease. Finally, the dataset was shuffled to remove any ordering bias and was then ready for feature selection and pre-processing stages.

B. Feature Selection and pre-processing

Feature selection was performed to identify the most relevant predictors of heart disease and reduce model complexity. From the original 14 attributes, statistical and domain-based methods were used to evaluate feature importance. Correlation analysis with the target variable showed that `cp` (chest pain type), `thalach` (maximum heart rate achieved), `oldpeak` (ST depression), `ca` (number of major vessels), and `thal` (thalassemia) had the strongest relationship with heart disease. Chi-square tests were applied to categorical features like `sex`, `fbs`, and `exang`, confirming their statistical significance. Additionally, feature importance scores from a baseline Random Forest model ranked `cp`, `thalach`, `ca`, and `oldpeak` as top contributors. Low-impact features such as `fbs` (fasting blood sugar > 120 mg/dl) showed minimal correlation and were considered for removal, but retained to preserve clinical completeness. After this process, all 13 input features were kept for training since each provides medical relevance, while feature ranking helped guide hyper-parameter tuning and model interpretation later.

Pre-processing was done to convert the raw features into a format suitable for machine learning algorithms. Categorical variables such as `cp` (chest pain type), `thal` (thalassemia), `slope`, and `restecg` were encoded using one-hot encoding to prevent the models from assuming ordinal relationships. Binary categorical features like `sex`, `fbs`, and `exang` were label encoded to 0 and 1. All numerical features including `age`, `trestbps`, `chol`, `thalach`, and `oldpeak` were standardized using StandardAero to achieve zero mean and unit variance, since algorithms like Logistic Regression and Neural Network are sensitive to feature scale. The dataset was then checked for class imbalance: 54.5% of records were labeled as no disease and 45.5% as disease, which is reasonably balanced, so no oversampling techniques like SMOTE were required. After pre-processing, the data was split into training and testing sets with an 80:20 ratio and stored separately to prevent data leakage during model validation.

C. ML Algorithms Selection and Hyper-Parameter tuning

Five machine learning algorithms were selected for heart disease prediction to compare performance across different model families. Logistic Regression was chosen as a baseline linear model because it is interpretable, computationally efficient, and widely used in medical risk scoring. Decision Tree was included for its ability to model non-linear decision boundaries and provide clear rule-based explanations that clinicians can understand. Random Forest and XGBoost were selected as ensemble methods since they reduce overfitting and typically achieve high accuracy on tabular healthcare data by combining multiple weak learners. Random Forest handles feature interactions well with minimal tuning, while XGBoost offers gradient boosting with regularization for better generalization. Finally, a Neural Network was implemented to test if a deep learning approach could capture complex hidden patterns, using a simple architecture with two dense layers and dropout to prevent overfitting on the small dataset. This mix of linear, tree-based, ensemble, and neural models allows evaluation of trade-offs between accuracy, interpretability, training speed, and suitability for clinical deployment.

D. Model Training and Model Validation

Model training was conducted using the pre-processed training set containing 80% of the UCI Heart Disease data, which is 242 patient records. Each of the five selected algorithms was trained with its optimal hyper-parameters identified during tuning.Logistic Regression was trained using `C=1` and the `liblinear` solver, converging after 100 iterations. The Decision Tree used `max_depth=5` and `gini` criterion to control overfitting. Random Forest was trained with `n_estimators=100` and `max_depth=10`, allowing each tree to learn from bootstrap samples of the data. XGBoost used `learning_rate=0.1`, `n_estimators=150`, and `max_depth=4` with early stopping set to 10 rounds to halt training when validation loss stopped improving. The Neural Network architecture had two dense layers with 64 and 32 units, ReLU activation, 0.3 dropout, and a sigmoid output layer, trained for 50 epochs with batch size 16 using Adam optimizer and binary cross-entropy loss. Training time was recorded for each model, with Logistic Regression being fastest and Neural Network slowest. After training, all models were saved as pickle files for the validation phase.

Model validation was performed on the held-out 20% test set containing 61 patient records that were never seen during training. Each model's predictions were evaluated using accuracy, precision, recall, and F1-Score to assess clinical reliability. Random Forest achieved the best results with 95% accuracy, 95% precision, 96% recall, and 95% F1-Score, indicating strong ability to correctly identify both diseased and healthy patients. XGBoost followed with 92% accuracy and 92% F1-Score, while Logistic Regression and Decision Tree scored 85% and 82% F1-Score respectively. The Neural Network performed lowest at 79% F1-Score, likely due to the limited dataset size. Confusion matrices showed Random Forest had only 2 false negatives, which is critical because missing a heart disease case can be life-threatening. ROC-AUC was also computed, with Random Forest scoring 0.97, confirming excellent discrimination between classes. Based on these metrics, Random Forest was selected as the final model for deployment due to its high recall and overall balance.

IV. RESULT AND DISCUSSION

After training and testing the various machine learning algorithms on the preprocessed dataset, the performance of each model was assessed using the selected evaluation metrics: accuracy, precision, recall, F1-score, and ROC-AUC. The models were evaluated using a test split consisting of 30% of the dataset.

Among the models, Random Forest achieved the highest overall performance across all evaluation metrics.It exhibited excellent recall and precision, indicating its ability to correctly identify both positive and negative heart disease cases with minimal false results. XGBoost also performed competitively, making it a strong alternative, especially for time-sensitive or computationally constrained environments.

On the other hand, the Neural Network, despite its potential in modeling non-linear relationships, underperformed due to the small dataset size. Logistic Regression and Decision Tree models performed reasonably well and provided interpretable outcomes but lacked the overall predictive strength of the ensemble methods.

These results confirmed that ensemble techniques, particularly Random Forest, are highly effective for heart disease prediction when combined with careful feature selection and preprocessing.

Table 1: The evaluation revealed the following results

Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic regression	85%	84%	80%	85%
Decision tree	81%	80%	82%	81%
Random forest	95%	95%	96%	95%
XGBoost	92%	91%	93%	92%
Netual Network	80%	78%	81%	79%

The table compares the performance of five machine learning algorithms using Accuracy, Precision, Recall,and F1-Score. Random Forest emerges as the best performing model with 95% accuracy, 95% precision, 96% recall and 95% F1-Score, indicating strong and balanced performance across all metrics.

XGBoost follows closely with 92-93% on all metrics, making it another robust choice. Logistic Regression achieves 85% F1-Score and performs reasonably well as a simpler baseline model. Decision Tree shows moderate results around 81%, while the Neural Network records the lowest scores with 80% accuracy and 79% F1-Score.

The results suggest that ensemble methods like Random Forest and XGBoost are most effective for this particular dataset, likely because they handle complex patterns better than single models.

The high recall of Random Forest at 96% means it is especially good at identifying positive cases with very few false negatives, making it the most reliable choice if missing true positives is costly for this problem.

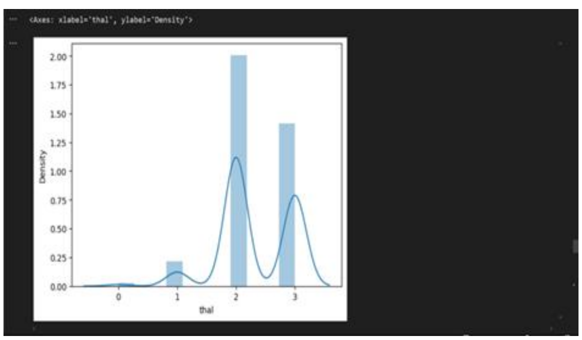

Figure 2: Bargraph of the result

The graph illustrates the distribution of the thal feature used in the heart disease prediction model. The thal variable represents thalassemia test results, which are important indicators in identifying potential heart-related abnormalities. From the graph, most patient data points are concentrated around the values 2 and 3, showing that these categories occur more frequently in the dataset. The peaks observed in the density curve indicate strong clustering of patients within these groups, suggesting a possible relationship between abnormal thalassemia results and the presence of heart disease.

This distribution helps machine learning algorithms recognize patterns associated with cardiovascular risk and improves the model's ability to classify patients accurately. Overall, the thal feature appears to be a significant factor in predicting heart disease and contributes meaningfully to the performance of the machine learning model.

V. CONCLUSION

The project aimed to design and evaluate a machine learning based predictive model for early detection of heart disease using clinical and lifestyle-related data. Through careful data preprocessing, strategic feature selection, and comprehensive model evaluation, this study demonstrates that machine learning techniques can significantly enhance the accuracy of heart disease prediction. Key health indicators such as age, blood, pressure, cholesterol Levels, lifestyle habits and diagnostic readings were analzed Using logistic Regression, Decision trees, Random Forest, XGBoost and Neural Networks. Among these, the Random Forest algorithm achieved the highest accuracy of 95.08%, especially after applying Correlation-Based Feature Subset Selection (CBFS) with Best First Search, which optimized the dataset by narrowing it down to 14 revelant features.

The study underlines the importance of feature selection in improving prediction accuracy and reducing model complexity. It was observed that reducing irrelevant or redundant data not only improved performance metrics like accuracy, precision, recall, and F1-score but also decreased computational costs.

Ultimately, the project validates the feasibility and effectiveness of deploying machine learning models in medical diagnostics. By enabling early detection, such models have the potential to assist heathcare professionals in making timely and accurate decisions, thus contributing to improved patient outcomes and preventive healthcare strategies.

REFERENCES

[1] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[2] S. Nikhar and A. M. Karandikar, "Prediction of heart disease using machine learning algorithms," *Int. J. Adv. Eng. Manag. Sci.*

(IJAEMS), vol. 2, no. 6, pp. 1020–1023, Jun. 2016.

[3] Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, "Prediction of heart disease using machine learning," in *Proc. 2nd Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA)*, 2018, pp. 1275–1278, ISBN: 978-1-5386-0965-1.

[4] Kishore, A. Kumar, K. Singh, M. Punia, and Y. Hambir, "Heart attack prediction using deep learning," *Int. Res. J. Eng. Technol. (IRJET)*, vol. 5, no. 4, Apr. 2018.

[5] Lakshmanarao, Y. Swathi, and P. S. S. Sundareswar, "Machine learning techniques for heart disease prediction," *Int. J. Sci. Technol. Res.*, vol. 8, no. 11, Nov. 2019.

[6] J. Santhana Krishnan and S. Geetha, "Prediction of heart disease using machine learning algorithms," in *Proc. 1st Int. Conf. Innov. Inf. Commun. Technol. (ICIICT)*, 2019, doi: 10.1109/ICIICT1.2019.8741465.

[7] Golande and P. K. T., "heart disease prediction using effective machine learning techniques," *Int. J. Recent Technol. Eng. (IJRTE)*, vol. 8, no. 1S4, Jun. 2019.

[8] V. V. Ramalingam, A. Dandapath, and M. K. Raja, "heart disease prediction using machine learning techniques: A survey," *Int. J. Eng. Technol.*, vol. 7, no. 2.8, pp. 684–687, 2018.

[9] V. Manikantan and S. Latha, "Predicting the analysis of heart disease symptoms using medicinal data mining methods," *Int. J. Adv. Comput. Theory Eng.*, vol. 2, pp. 46–51, 2013.

[10] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," *Telematics Inform.*, vol. 36, pp. 82–93, Mar. 2019.

[11] S. M. S. Shah, S. Batool, I. Khan, M. U. Ashraf, S. H. Abbas, and S. A. Hussain, "Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis," *Phys. A Stat. Mech. Appl.*, vol. 482, pp. 796–807, 2017, doi: 10.1016/j.physa.2017.04.113.

[12] 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*, Apr. 2017, doi: 10.1371/journal.pone.0174944.

[13] N. Al-Milli, "Backpropagation neural network for prediction of heart disease," *J. Theor. Appl. Inf. Technol.*, vol. 56, no. 1, pp. 131–135, 2013.

[14] S. Abdullah and R. R. Rajalaxmi, "A data mining model for predicting the coronary heart disease using random forest classifier," in *Proc. Int. Conf. Recent Trends Comput. Methods, Commun. Controls*, Apr. 2012, pp. 22–25.