IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

"Review On Heart Disease Detection Using Machine Learning"

Ms. Devyani V. Pakhale[1], Ms. Revati D. Tantrapale[2], Mr. Ankush K. Kautikkar[3], Mr. Sarvesh N. Madhapure[4], Prof. Samata V. Athawale[5]

Department of Computer Science and Engg,

DRGITR, Amravati, Maharashtra, India

Abstract: Day by day the cases of heart diseases are increasing at a rapid rate and it's very Important and concerning to predict any such diseases beforehand. Heart disease cases are increasing rapidly every day, so it's really important to predict and detect them early. Catching the disease in time can help save lives and improve treatment. This research project focuses on predicting whether a person is likely to have heart disease, based on their medical information and history. To do this, we built a heart disease detection system that uses machine learning. We used several machine learning algorithms like Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, and Random Forest to analyze the data and predict if a patient has heart disease or not. The system looks at patterns and trends in the collected medical data to make smart predictions. After testing the performance of these models, we found that they can help improve medical care, make early diagnoses, and even lower healthcare costs. This project helped us gain valuable knowledge about how machine learning can be used to detect heart disease. The system was developed using a pynb (Jupyter Notebook) file format.

Keywords: Machine learning, Heart Disease, Diagnosis, Algorithm

1. INTRODUCTION

Heart disease includes different types of conditions that affect the heart. Right now, it is the leading cause of death in the world, causing around 17.9 million deaths each year, according to the World Health Organization (WHO). Several unhealthy habits increase the risk of heart disease, like having high cholesterol, obesity, high triglycerides, and high blood pressure. The American Heart Association mentions certain warning signs such as trouble sleeping, irregular heartbeat, swollen legs, and rapid weight gain (sometimes 1–2 kg per day). These symptoms can also appear in older adults or people with other health problems, which makes it hard for doctors to diagnose heart disease correctly. Because of this, some cases are missed, leading to serious health issues or even death. Thankfully, a lot of patient data and research is now available from

hospitals and open sources. This allows researchers to use computer technologies like machine learning (ML) and artificial intelligence (AI) to help diagnose heart disease more accurately. Machine learning can analyze large amounts of health data and learn to find patterns, helping to predict and detect diseases. It has even been helpful in predicting health issues during pandemics. Many research studies have tested different machine learning models to predict and classify heart disease.

A complete genomic data analysis can easily be done using machine learning models. Models can be trained for knowledge pandemic predictions and also medical records can be transformed and analyzed more deeply for better predictions. Many studies have been performed and various machine learning models are used for doing the classification and prediction for the diagnosis of heart disease. An automatic classifier for detecting congestive heart failure shows the patients at high risk and the patients at low risk by Melillo et al; they used machine learning algorithm as CART which stands for Classification and Regression in which sensitivity is achieved as 93.3 percent and specificity is achieved as 63.5 percent. Then for improving the performance electrocardiogram (ECG) approach is suggested by Rahhal et al in which deep neural networks are used for choosing the best features and then using them. Then, for detecting heart failures, a clinical decision support system is contributed by Guidi et al. for preventing it at an early stage.

Objectives:

- Develop a machine learning model to classify heart disease risk.
- Design a seamless web-based system using Django for easy interaction.
- Provide an admin interface for managing users and datasets.
- Allow doctors to manage patient data and make informed predictions.
- Enable patients to check their health status using an UI.

Key components and motivations for adopting AI in heart disease detection include:

- **Data Collection & Sensors:**
- ECG/EKG, echocardiograms, blood pressure monitors
- Wearable devices (e.g., smartwatches, Holter monitors)
- Electronic Health Records (EHRs)
- Data Preprocessing & Feature Extraction:
- Noise removal and normalization of physiological signals
- Extraction of relevant features (e.g., heart rate variability, QRS complex)

AI Models & Algorithms:

- Machine Learning: SVM, Decision Trees, Random Forest
- Deep Learning: CNNs (for imaging), RNNs/LSTMs (for time-series data)

- Ensemble models and hybrid techniques for improved accuracy.
- **Early Detection and Prevention:**
- Catching heart conditions before they become critical
- Predicting risk of stroke, heart attack, or cardiac arrest.
- **Improved Diagnostic Accuracy:**
- Reducing false positives/negatives compared to manual interpretation
- Standardizing diagnosis and minimizing human bias.

Algorithm Use in Heart Disease Detection

1. Logistic Regression (LR)

- Use: Predicting binary outcomes (e.g., disease vs. no disease)
- Why: Simple, interpretable, and effective for small to medium datasets
- Example: Predicting presence of heart disease based on age, cholesterol, and blood pressure

2. Support Vector Machine (SVM)

- Use: Classifies patients into disease/no-disease categories by finding optimal decision boundaries
- Why: Good with high-dimensional data and small sample sizes
- Example: Detecting arrhythmias or heart failure risk from ECG features

3. Decision Tree (DT)

- Use: Predictive modeling with a flowchart-like structure
- Why: Easy to understand and visualize; handles non-linear relationships well
- Example: Risk stratification based on cholesterol, glucose, and family history.

2. LITERATURE REVIEW

Ijaz Bo Jin, Chao Che et al. (2018) proposed a "Predicting the Risk of Heart Failure with EHR Sequential Data Modelling" model designed by applying neural network. This paper used the electronic health record (EHR) data from real-world datasets related to congestive heart disease to perform the experiment and predict the heart disease before itself. We tend to used one-hot encryption and word vectors to model the diagnosing events and foretold coronary failure events victimization the essential principles of an extended memory network model. By analysing the results, we tend to reveal the importance of respecting the sequential nature of clinical records.

Aakash Chauhan et al. (2018) presented "Heart Disease Prediction using Evolutionary Rule Learning". This study eliminates the manual task that additionally helps in extracting the information (data) directly from the electronic records. To generate strong association rules, we have applied frequent pattern growth association

mining on patient's dataset. This will facilitate (help) in decreasing the number of services and shown that overwhelming majority of the rules helps within the best prediction of coronary sickness.

Ashir Javeed, Shijie Zhou et al. (2017) designed "An Intelligent Learning System based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection". This paper uses random search algorithm (RSA) for factor selection and random forest model for diagnosing the cardiovascular disease. This model is principally optimized for using grid search algorithmic program. Two forms of experiments are used for cardiovascular disease prediction. In the first form, only random forest model is developed and within the second experiment the proposed Random Search Algorithm based random forest model is developed. This methodology is efficient and less complex than conventional random forest model. Comparing to conventional random forest it produces 3.3% higher accuracy. The proposed learning system can help the physicians to improve the quality of heart failure detection. "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" proposed by Senthilkumar Mohan, Chandrasegar Thirumalai et al. (2019) was efficient technique using hybrid machine learning methodology. The hybrid approach is combination of random forest and linear method. The dataset and subsets of attributes were collected for prediction. The subset of some attributes was chosen from the pre-processed knowledge(data) set of cardiovascular disease. After prep-processing, the hybrid techniques were applied and disgnosis the cardiovascular disease.

3. DATASETS USED IN RESEARCH

The dataset used for this research purpose was the Public Health Dataset and it is dating from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "target" field refers to the presence of heart disease in the patient. It is integer-valued 0 = no disease and 1 = disease. The first four rows and all the dataset features are shown in Table 1 without any preprocessing. Now the attributes which are used in this research purpose are described as follows and for what they are used or resemble:

- Age
- Gender
- Blood pressure
- Cholesterol levels

- Blood sugar
- ECG results
- Heart rate
- Chest pain type
- Exercise-induced angina
- Family history, etc.
- Popular dataset: UCI Heart Disease Dataset

4. MACHINE LEARNING TECHNIQUES

Machine learning is used to provide the good learning to the machines and analyze some pattern for handling the data in extra efficient manner. Sometimes, it may happens that after viewing the data, we even unable to predict the actual pattern or acquire the valuable information from the data. In this condition, we have to go for machine learning. The motive of machine learning is to grasp some knowledge from the data by themselves. Even, many studies has been terminated which highlights the purpose of machine learning that how do machines learn by its.

Machine Learning Technique: The main ML techniques can be classified as follows.

Supervised Learning: The supervised machine learning alg0rithms are those which demand some external assistance. The input dataset splits into training and test dataset. The trained dataset composed of output variable which is t0 be predicted or classified. Each algorithm get to know a specific pattern from the training dataset and just apply them to the test dataset for prediction or classification purposes. This algorithm is named as supervised learning in view of the fact that the process of algorithm learning from the training dataset can be thought 0f as a teacher supervising the learning process.

Support Vector Machine: SVM is a preferred ML algorithm because it is resistant to outliers and gives good results when the data size grows. SVM represents data points in an n-dimensional space and tries to find the best hyperplane separating samples belonging to different classes. However, in some cases, data points cannot be separated linearly. In these cases, the SVM's solution is found using more complex hyperplanes. The kernel trick allows the SVM to work with data that can be separated more easily in higher dimensional spaces by moving the data to higher dimensional spaces (kernel space). This allows it to perform the separation using more complex hyperplanes for the non-linearly separable dataset. The kernel trick works by using different kernel functions, especially the radial basis function (RBF) and the polynomial kernel. These kernel functions operate based on the properties of data points (distance, similarity, inner product, etc.) and allow the SVM to find an appropriate hyperplane that it can use to separate data in higher dimensional spaces.

Naive Bayes: Naive Bayes is a surprisingly powerful algorithm for predictive modeling. It is a statistical classifier which assumes no dependency between attributes attempting to maximize the posterior probability in determining the class. Theoretically, this classifier has the minimum error rate, but may not be the case always. Inaccuracies are caused by assumptions due to class conditional independence and the lack of available probability data.

Random Forest: Random Forest is one of the most renowned and most powerful machine learning algorithms. It is one kind of machine learning algorithm that is called Bagging or Bootstrap Aggregation. In order to estimate a value from a data sample such as mean, the bootstrap is a very powerful statistical approach. Here, lots of samples of data are taken, the mean is calculated, after that all of the mean values are averaged to give a better prediction of the real mean value. In bagging, the same method is used, but instead of estimating the mean of every data sample, decision trees are generally used. Here, numerous samples of the training data are considered and models are generated for every data sample. While a prediction for any data is needed, each model gives a prediction and these predictions are then averaged to get a better estimation of the real output value.

Logistic Regression: Logistic regression is a technique of machine learning which is taken from the field of statistics. This method can be used for binary classification where values are distinguished with two classes. Logistic regression is similar to linear regression where the goal is to calculate the values of the coefficients within every input variable. Unlike linear regression, here the prediction of the output is constructed using a non-linear function which is called a logistic function. The logistic function transforms any value within the range of 0 to 1. The predictions made by logistic regression are used as the probability of a data instance concerning to either class 0 or class 1. This can be necessary for problems where more rationale for a prediction is needed. Logistic regression works better when attributes are unrelated to output variable.

5. Working

Heart disease detection using machine learning involves using various algorithms to analyze medical data and predict the likelihood of a person having heart disease.

- The process usually includes collecting the data, cleaning it, choosing the most important features, training the model, checking how well it works, and finally putting it into use.
- Heart disease detection using machine learning involves the collected data is cleaned and modify data into suitable format. And it apply an dataset for decision making mechanism
- After it apply machine learning classifier that is ML algorithm including logistic regression, decision tree, support vector machine. Which then learns patterns to predict the likelihood of a person developing heart disease, allowing for early identification and intervention based on the calculated risk factors.

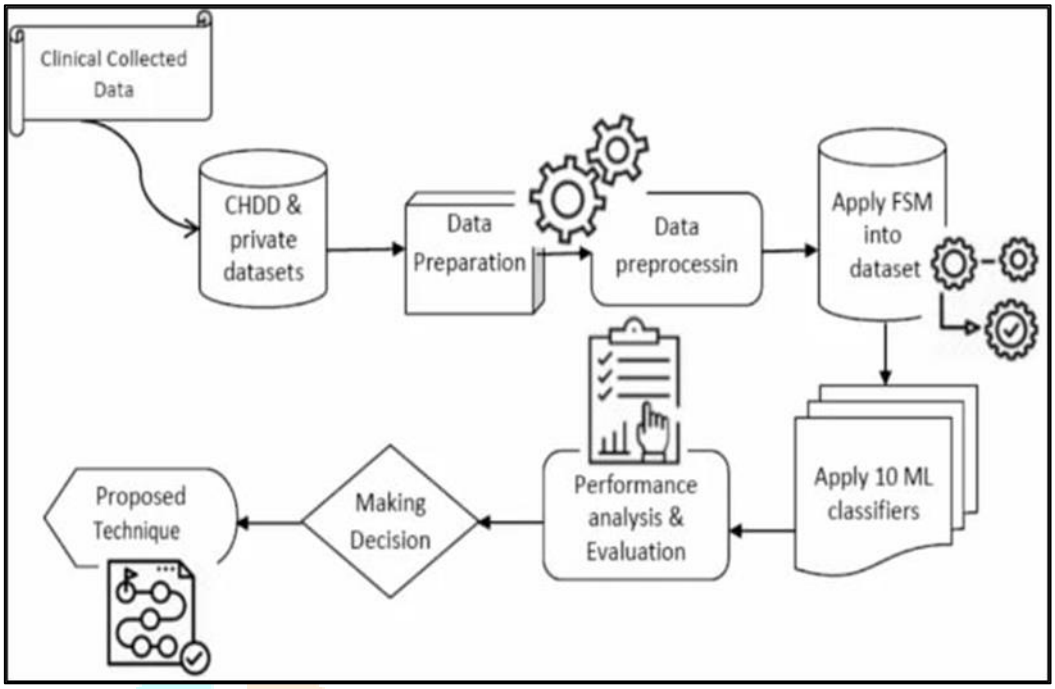

Fig. 4.1 Working of Heart Disease Detection

1. Data Collection

Datasets typically include patient medical records.

Common features:

- Age
- Gender
- Blood pressure
- Cholesterol levels
- Blood sugar
- ECG results
- Heart rate
- Chest pain type
- Exercise-induced angina
- Family history, etc.
- Popular dataset: UCI Heart Disease Dataset

2. Data Preprocessing

- Cleaning: Handling missing or inconsistent data.
- Normalization/Scaling: Ensuring all features are on a similar scale.
- Encoding: Converting categorical variables (like sex or chest pain type) into numerical formats.
- Splitting: Dividing data into training and testing sets (e.g., 80% train, 20% test).

3. Feature Selection

- Choosing the most relevant inputs (features) that affect heart disease risk.
- Helps reduce complexity and improve model accuracy.
- Methods: Correlation matrix, Recursive Feature Elimination (RFE), etc.

4. Model Selection

- Common ML algorithms used:
- Logistic Regression – good for binary classification (disease/no disease).
- Decision Trees / Random Forest – interpretability and performance.
- Support Vector Machines (SVM) – effective in high-dimensional space.
- K-Nearest Neighbors (KNN) – simple, non-parametric.
- Neural Networks – for more complex patterns.
- XGBoost / LightGBM – for optimized performance.

5. Training the Model

- The chosen model is trained on the historical data.
- It learns the patterns and relationships between feature sand the target (presence of heart disease).

6. Evaluation

- Performance is evaluated using metrics like:
- Accuracy
- Precision & Recall
- F1-score
- ROC-AUC Curve
- Helps understand how well the model distinguishes between patients with and without heart disease.

7. Prediction

- Once validated, the model is used to predict heart disease in new patients using their medical data.

8. Deployment (Optional)

- Models can be deployed in a hospital's system or a web/mobile app.
- Doctors can input patient data and receive risk predictions instantly.

9. Ethical Considerations

- Privacy: Patient data must be handled with confidentiality (e.g., HIPAA, GDPR).
- Bias: Ensure data is representative of diverse populations.
- Explainability: Clinicians often need to understand why a model made a decision.

10. Simple Tools/Libraries

- Python libraries: scikit-learn, pandas, matplotlib, tensorflow, xgboost
- Platforms: Google Colab, Jupyter Notebooks

6. Challenges in Heart Disease Detection

1. Data Quality and Availability

- Limited datasets: Most publicly available datasets (like UCI's Heart Disease dataset) are small and not diverse.
- Missing or noisy data: Medical records often have missing values, inconsistencies, or errors.
- Labeling problems: Ground truth labels (e.g. confirmed heart disease) can be uncertain or depend on expert **opinions.**

2. Data Imbalance

- Heart disease datasets often have imbalanced classes (more healthy individuals than those with heart disease), making it hard for models to learn to detect the minority class.

3. Generalization and Bias

- Overfitting to training data is a risk, especially with small datasets.
- Bias can arise due to demographic limitations (e.g. age, gender, ethnicity), leading to models that don't generalize well to all patient populations.

4. Interpretability

- Clinicians need to understand how the model makes decisions, especially for life-and-death scenarios.
- Complex models like deep learning often act like "black boxes," which makes them hard to trust in clinical settings.

7. Future Research Directions

1. Real-Time Monitoring with Wearable Technology

Objective: Use continuous data from smartwatches, ECG patches, and fitness trackers to detect heart irregularities early.

Research Focus: Develop lightweight ML models for on-device predictions. Handle noisy, streaming data in real-time.

2. Deep Learning on Medical Imaging

Objective: Use deep learning (CNNs, Transformers) for automated analysis of:

Echocardiograms, CardiacMRI, CT angiograms

Potential: Early detection of structural heart problems, plaque buildup, and valve disorders.

3. Explainable and Transparent AI Models

Challenge: Most powerful models (like deep learning) are black boxes.

Research Goal: Make AI decisions explainable to clinicians.

Approaches: SHAP (SHapley Additive exPlanations) LIME (Local Interpretable Model-Agnostic Explanations) Interpretable neural networks

4. Multimodal Data Fusion

Aim: Combine data from various sources for improved accuracy:

Electronic Health Records (EHR), ECG signals

Blood test results, Imaging data, Genetic profiles

Research Potential: Create holistic patient profiles for risk prediction.

5. Federated Learning and Data Privacy

Problem: Healthcare data is sensitive and often decentralized.

Future Direction: Use federated learning to train models across hospitals without sharing raw data.

Benefits: Enhanced privacy

Collaboration across institutions

Larger and more diverse training data

6. Early and Pre-Symptomatic Detection

Focus: Predict future risk of heart disease before symptoms arise.

Techniques: Time series forecasting, Survival analysis, Longitudinal patient monitoring

7. Personalized and Preventive Cardiology

Goal: Shift from generalized diagnosis to individualized risk assessment.

Research Focus: Models tailored to lifestyle, age, gender, genetics. AI-guided lifestyle intervention recommendations.

REFERENCES

1. BoJin , Chao Che,Zhen Liu ,Shulong Zhang ,Xiaomeng Yin And Xiaopeng Wei “Predicting the Risk of Heart Failure with EHR Sequential Data Modelling”. IEEE Access 2018.
2. Aakash Chauhan, Aditya Jain, Purushottam Sharma, Vikas Deep, “Heart Disease Prediction using Evolutionary Rule Learning”, “International Conference on "Computational Intelligence and Communication Technology” (CICT 2018).
3. Ashir Javeed, Shijie Zhou, Liao Yongjian, Iqbal Qasim, Adeeb Noor. “An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection”. IEEE Access (Volume: 7) 2019 .
4. Senthilkumar Mohan, Chandrasegar Thirumalai and Gautam Srivastava. “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques”. IEEE Access (Volume: 7) 2019.
5. K. Prasanna Lakshmi, Dr. C.R.K.Reddy. “Fast Rule-Based Heart Disease Prediction using Associative Classification Mining”. International Conference on Computer, Communication and Control (IC4) 2015
6. M. Satish, D Sridhar, “Prediction of Heart Disease in Data Mining Technique”, International Journal of Computer Trends & Technology (IJCTT), 2015.
7. Lokanath Sarangi, Mihir Narayan Mohanty, Srikanta Pattnaik, “An Intelligent Decision Support System for Cardiac Disease Detection”, IJCTA, International Press 2015.
8. Boshra Bahrami, Mirsaeid Hosseini Shirvani, “Prediction and Diagnosis of Heart Disease by Data Mining Techniques”, Journal of Multidisciplinary Engineering Science and Technology (JMEST) ISSN: 3159-0040 Vol. 2 Issue 2, February–2015.
9. Mamatha Alex P and Shaicy P Shaji, “Prediction and Diagnosis of Heart Disease Patients using Data Mining Technique”, International Conference on Communication and Signal Processing 2019.
10. Dangare Chaitrali S and Sulabha S Apte. "Improved study of heart disease prediction system using data mining classification techniques." International Journal of Computer Applications 47.10 (2012): 44-8.