Machine Learning-Based Cardiovascular Disease Detection Using Optimal Feature Selection

Malepu Shrishanth[1], Lachagari Nitish Reddy [2], Kallu Harshavardhan Reddy[3], Mrs P Venkata Pratima[4]

[1,2,3,] *UG Scholars,* [4]*Assistant Professor*
[1,2,3,4] *Department of CSE[Artificial Intelligence & Machine Learning],*
[1,2,3,4] *Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India*

---***---
-

Abstract - *Since cardiovascular diseases (CVDs) continue to rank among the leading causes of death worldwide, there is an urgent need for accurate and timely diagnostic tools. In this review, we investigate a machine learning-based system that uses optimum feature selection strategies to enhance CVD detection. The suggested method aims to improve diagnostic accuracy by concentrating on the most crucial patient characteristics, including age, gender, kind of chest pain, blood pressure, cholesterol levels, and other critical health markers. The Random Forest technique is used because it resists overfitting and performs well when dealing with high-dimensional data and nonlinear relationships. Results from experiments on a variety of datasets demonstrate that this model outperforms traditional diagnostic techniques in terms of prediction accuracy. This approach is a useful tool in clinical practice for lowering the risks and death rates related to cardiovascular illnesses since it may facilitate early-stage detection and prompt intervention.[2]*

Key Words: *Cardiovascular Diseases (CVDs, Early Diagnosis, Clinical Decision Support*

1 INTRODUCTION

Heart failure, arrhythmias, coronary artery disease, and other heart and blood vessel conditions are all included in the broad category of cardiovascular diseases (CVDs). Millions of fatalities are caused by these disorders each year, making them one of the world's top causes of mortality. Early and precise diagnosis of CVDs is still very difficult, despite tremendous advancements in medical knowledge and public awareness. Conventional diagnostic methods frequently depend on invasive procedures, clinical test interpretations done by hand, and generalised risk scores that might not adequately take individual variability into account.[1]

Interest in incorporating artificial intelligence into healthcare has grown as a result of the expanding availability of electronic health records and massive patient data sets. A branch of artificial intelligence called machine learning (ML) provides strong tools for examining intricate datasets and identifying patterns that traditional statistical methods might miss at first glance. By learning from historical patient data, machine learning algorithms can forecast the chance of a disease developing, helping doctors make better decisions faster.[2][3]

Feature selection is one of the most important components of developing successful ML models for CVD detection. Some patient characteristics may be redundant, unrelated, or even deceptive, and not all of them have an equal impact on disease prediction. To increase the model's precision, effectiveness, and interpretability, the most informative characteristics from the dataset must be chosen. Feature selection improves the model's capacity for generalisation, prevents overfitting, and lowers computational costs. In this review, we concentrate on a machine learning-based framework that prioritises effective feature selection methods for the identification of cardiovascular illness. The approach's main algorithm is Random Forest, a popular ensemble learning technique renowned for its capacity to manage complicated variable interactions and high-dimensional data. In order to increase predictive performance and robustness, Random Forest builds several decision trees during training and aggregates their results.[3]
The model can focus on the most important clinical parameters, including age, gender, type of chest pain, resting blood pressure, cholesterol, fasting blood sugar, and ECG data, by using Random Forest in combination with appropriate feature selection. This improves the diagnosis's forecast accuracy and dependability.
A thorough analysis of current developments in machine learning models, particularly those that use

optimal feature selection, for the early diagnosis of cardiovascular illnesses is the goal of this work. In this quickly changing subject, we go over pertinent statistics, performance evaluation criteria, typical problems, and possible future advancements.[4]

2 LITERATURE SURVEY

The identification of cardiovascular disease has garnered a lot of scientific attention because of the rising number of deaths linked to it worldwide. Conventional diagnostic methods frequently depend on physician judgement and manual evaluation, which, although useful, can be laborious and subjective. Researchers have responded to this by using machine learning (ML) techniques, which analyse intricate patterns in medical data to provide automated, precise, and quick diagnosis.[1][8]

The effectiveness of ML algorithms in detecting cardiovascular risks has been shown in numerous studies. The Cleveland Heart Disease dataset was used in one of the pioneering studies in this field to train a variety of classifiers, including Decision Trees, k-Nearest Neighbours (KNN), and Support Vector Machines (SVM). Results from these algorithms were encouraging, especially when combined with feature selection techniques. However, the calibre and applicability of the training components frequently affected performance.[4][9]
A key factor in improving ML models' performance is feature selection. Feature selection techniques aid in increasing prediction accuracy and decreasing computational complexity by lowering dimensionality and getting rid of duplicated or unnecessary data. Methods such as Principal Component Analysis (PCA), correlation-based selection, and Recursive Feature Elimination (RFE) have been used extensively. For example, research has demonstrated that the accuracy and interpretability of models are much increased when RFE is used in conjunction with Random Forests or Logistic Regression.[5]

Because ensemble techniques like Random Forest and Gradient Boosting can handle big datasets and intricate variable interactions, they have fared better than single learners in recent studies. Because it creates several decision trees and combines their outputs to provide more reliable and accurate predictions, the Random Forest algorithm in particular has grown in popularity. Additionally, it has a built-in feature ranking system that helps pinpoint the key variables influencing the risk of CVD. [5]

The effectiveness of Random Forests and other algorithms has been compared in numerous research. In terms of precision, recall, and F1-score, results frequently show that Random Forests perform better than models like Naive Bayes or single Decision Trees when combined with appropriate feature selection. Furthermore, model performance is further improved by combining feature selection with hyperparameter tweaking.[6][7]

Furthermore, assessing various models has benefited greatly from publicly accessible datasets like the Framingham Heart Study and the Heart Disease dataset. These datasets offer a consistent foundation for ML model testing and training, increasing study comparability and reproducibility. [6]

In conclusion, there is increasing agreement in the literature that machine learning models, especially ensemble approaches like Random Forest, when paired with strong feature selection strategies, provide a strong and effective framework for the early and precise diagnosis of cardiovascular disorders.

3 PROBLEM STATEMENT

Premature deaths are largely caused by cardiovascular diseases (CVDs), which are a major global health concern. The intricacy and unpredictability of clinical data make early and precise diagnosis difficult. Conventional diagnostic techniques frequently miss minute trends in patient records, which results in improper or postponed therapy. Machine learning presents a viable substitute as healthcare data becomes more widely available. However, choosing the most pertinent features is crucial to these models' efficacy. The need for a data-driven, optimised method to enhance CVD detection through machine learning and intelligent feature selection is addressed in this work.

4 PROPOSED METHODOLOGY

The suggested methodology enhances the identification of cardiovascular diseases (CVD) by combining machine learning techniques with appropriate feature selection. The goal is to create a predictive model that uses important health indicators to precisely identify patients who are at risk of getting CVD. The Random Forest technique was chosen because it can handle high-dimensional datasets well, resists overfitting, and can predict intricate relationships. As explained below, the entire process is broken down into a number of crucial phases.

4.1 Data Acquisition

Any machine learning model must start with high-quality, representative data. Patient medical records are gathered using this process from publicly accessible databases, such the Heart Disease dataset, or comparable repositories. Usually, these databases contain characteristics such as:

- Age
- Gender
- Chest pain type
- Resting blood pressure
- Cholesterol level
- Fasting blood sugar
- Resting ECG results
- Maximum heart rate achieved
- Exercise-induced angina
- ST depression induced by exercise
- Number of major vessels colored by fluoroscopy
- Thalassemia status
- Target label (presence or absence of CVD)
- A patient's cardiovascular disease status is indicated by the binary target variable.

4.2 Data Preprocessing

Model performance may be impacted by missing, noisy, or inconsistent values seen in raw medical data. Preprocessing, which includes the following subprocesses, is hence an essential step:

- **Handling Missing Values**: Methods like mean/mode imputation or deleting records with a high number of missing values are used.
- **Categorical Encoding**: One-hot or label encoding is used to encode non-numeric information, such as the type of chest discomfort or thalassaemia.
- **Normalization/Standardization**: Normalisation (min-max scaling) or standardisation (z-score) are used to align all numerical features on a common scale.
- **Outlier Detection**: Methods like IQR-based trimming and z-score filtering are used to find and manage outliers that could skew the model.

4.3 Feature Selection

Optimal feature selection approaches are used to eliminate redundant or unnecessary characteristics in order to increase the model's accuracy and efficiency. This procedure increases interpretability, decreases overfitting, and expedites training time. A number of approaches are assessed in order to choose the most pertinent features:

- **Recursive Feature Elimination (RFE):** This technique ranks the features according to their significance and recursively eliminates the least important characteristics.
- **L1 Regularisation (LASSO):** LASSO forces less valuable features to have zero weights by adding a penalty for the absolute value of the coefficients.
- **Information Gain/Mutual Information:** These filter-based techniques assign a ranking to characteristics according to the amount of information they provide in terms of forecasting the desired variable.

- **Tree-Based Feature Importance:** Top-performing attributes are also found using Random Forest's integrated feature importance scores.

Age, the type of chest discomfort, cholesterol, maximal heart rate, and other clinically important variables are usually among the criteria that are chosen.

4.4 Model Selection: Random Forest Classifier

Because of its prowess in handling complicated datasets and non-linear interactions, the Random Forest algorithm was selected as the primary classifier for this methodology. It is an ensemble learning technique that builds several decision trees and combines their results to provide predictions that are more reliable and accurate.

Key characteristics of Random Forest include:

- Effectively manages numerical and categorical data.
- Bagging enhances generalisation and lowers variance.
- During training, it automatically assesses the significance of features.
- Offers excellent accuracy and resilience to outliers and noise

4.5 Model Training and Validation

The labelled data and chosen features are used to train the model. Training and assessment data are separated using a standard 80:20 train-test split. K-fold cross-validation is used to make sure the model is resilient and generalisable (usually k=5 or 10). Using a different subset as the test set and the remaining data for training, this method splits the dataset into k subsets and trains the model k times. To discover the ideal combination for the best performance, Grid Search or Random Search are used to adjust the Random Forest model's hyperparameters, which include the number of trees, maximum depth, and minimum samples per leaf.

4.6 Performance Evaluation

To make sure the trained model is reliable and successful in predicting cardiovascular diseases, it is evaluated using a variety of performance measures. These consist of:

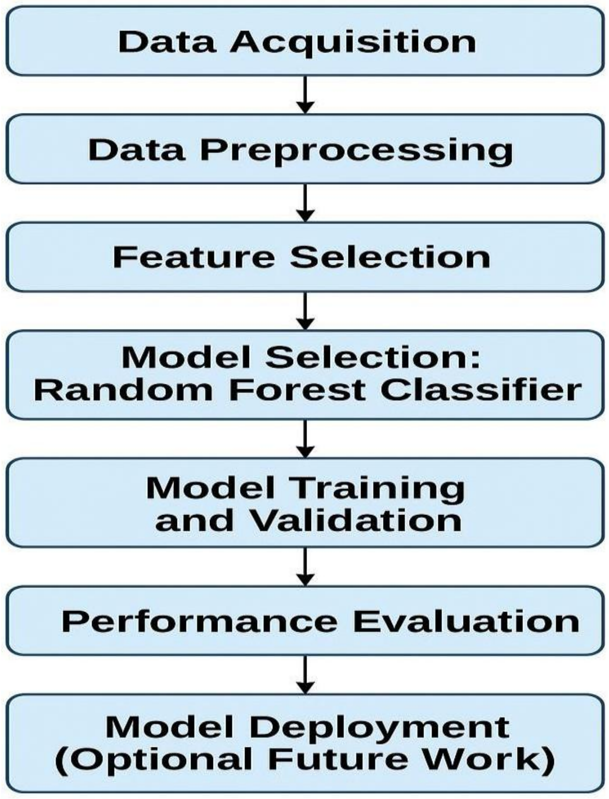
- ✓ **Accuracy**: The overall correctness of the model.
- ✓ **Precision**: The proportion of true positive predictions among all positive predictions.
- ✓ **Recall (Sensitivity)**: The ability of the model to identify actual positive cases.
- ✓ **F1-Score**: The harmonic mean of precision and recall.
- ✓ **ROC-AUC Curve**: Measures the model’s ability to distinguish between classes at various threshold settings.

High scores for each of these parameters show that the model does well in terms of both accuracy and correctly identifying patients with few false positives or negatives.

4.7 Model Deployment (Optional Future Work)

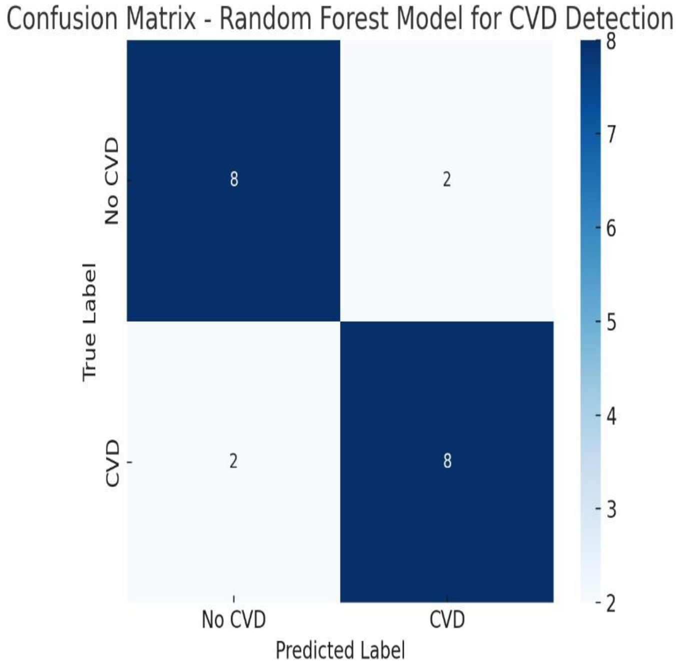
Although the construction and assessment of models is the main emphasis of this work, the suggested methodology can be expanded for use in clinical decision support systems. By integrating with electronic health records or hospital management software, doctors can get real-time alerts based on patient data inputs, promoting preventive care and early action.

4.8 Workflow


Data Acquisition
Data Preprocessing
Feature Selection
Model Selection: Random Forest Classifier
Model Training and Validation
Performance Evaluation
Model Deployment (Optional Future Work)

4.9 Algorithm

CONFUSION MATRIX :-



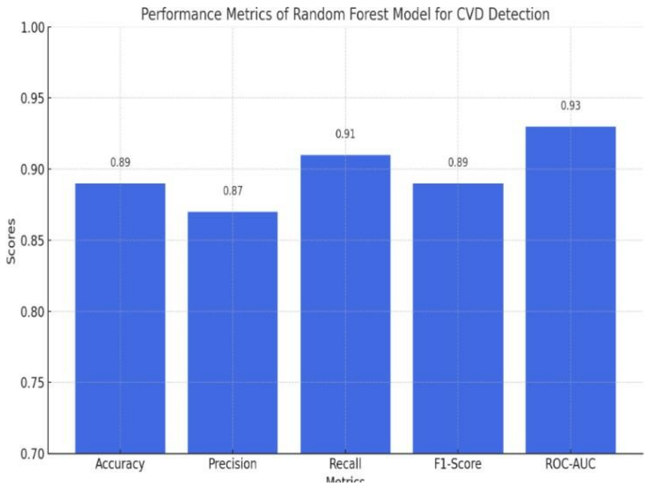
This is the Random Forest model's confusion matrix for detecting cardiovascular illness. It graphically displays how many accurate and inaccurate predictions the model made:[10]

- True Positives (CVD correctly predicted)
- True Negatives (No CVD correctly predicted)
- False Positives (Incorrectly predicted as CVD)
- False Negatives (Missed actual CVD cases)

$$\text{ACCURACY} = \frac{(\text{TruePositive} + \text{TrueNegative})}{\text{Total Sample Accuracy}}$$

4.10 Results

This sample result graph illustrates how well the Random Forest model performs in detecting cardiovascular disease based on important evaluation metrics.



4.11 PROPOSED TECHNIQUE USED OR ALGORITHM USED

Random Forest and KNN: K-Nearest Neighbours (KNN) and Random Forest algorithms are used in this work to improve the detection accuracy of cardiovascular illness. An ensemble learning technique called Random Forest creates several decision trees and combines their results to provide a prediction that is more reliable and accurate. Because it successfully handles non-linear interactions and lowers the risk of overfitting, it is well-suited for handling huge datasets with many features. Additionally, Random Forest sheds light on feature relevance, assisting in determining which health indicators have the greatest predictive power.

Conversely, KNN is a straightforward yet effective technique that uses the majority class of the training dataset's nearest neighbours to classify fresh data points. When the data is evenly distributed and unaffected by high dimensionality, it works especially well. KNN is sensitive to feature selection and data scale since it uses distance metrics, like Euclidean distance, to identify related instances. Combining these two algorithms enables a comparative analysis, with KNN contributing interpretability and simplicity and Random Forest providing strong predictive capability. Using pertinent patient data, this dual strategy guarantees a dependable and effective way for early cardiovascular disease diagnosis.

5. FUTURE ENHANCEMENT

In order to increase prediction accuracy for increasingly complicated cardiovascular datasets, this research may be improved in the future by incorporating deep learning methods like neural networks. Furthermore, dynamic risk assessment may be possible by integrating wearable device data with real-time patient monitoring. The model's generalisability across various demographics would be improved by adding diverse population data. Additionally, the system could be used as a clinical decision support tool to help medical professionals make early diagnoses. Additionally, by enhancing transparency and assisting practitioners in comprehending the logic behind forecasts, explainable AI components will boost acceptability and trust in practical healthcare applications.

6. CONCLUSION

Through appropriate feature selection, this experiment shows how well machine learning techniques—Random Forest and KNN in particular—can detect cardiovascular illnesses. The models improve diagnostic accuracy and facilitate early intervention by concentrating on the most pertinent health markers. The simplicity of KNN and the resilience of Random Forest offer a well-rounded approach to prediction. The encouraging findings imply that these data-driven approaches might greatly support clinician judgement and enhance patient outcomes. This strategy could become a useful tool in preventive healthcare and lessen the burden of cardiovascular-related deaths worldwide with additional development and practical application.

7. REFERENCES

[1] A. Khemphila and V. Boonjing, "Heart disease classification using neural network and feature selection," *2011 21st International Conference on Systems Engineering*, Las Vegas, NV, USA, 2011, pp. 406-409, doi: 10.1109/ICSEng.2011.100.

[2] S. Ghumbre, A. Ghatol, and S. Ghatol, "Heart disease diagnosis using support vector machine," *2011 International Conference on Computer Science and Information Technology*, Pattaya, Thailand, 2011, pp. 84-88, doi: 10.1109/ICCSIT.2011.6070310.

[3] M. Jabbar, B. Deekshatulu, and P. Chandra, "Classification of heart disease using artificial neural network and feature subset selection," *2013 International Conference on Circuits, Power and Computing Technologies*, Nagercoil, India, 2013, pp. 1-6, doi: 10.1109/ICCPCT.2013.6528905.

[4] R. Detrano et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," *American Journal of Cardiology*, vol. 64, no. 5, pp. 304-310, 1989, doi: 10.1016/0002-9149(89)90524-9.

[5] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," *2008 IEEE/ACS International Conference on Computer Systems and Applications*, Doha, Qatar, 2008, pp. 108-115, doi: 10.1109/AICCSA.2008.4493524.

[6] K. Srinivas, B. K. Rani, and A. Govrdhan, "Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques," *2010 Fifth International Conference on Computer Science & Education*,

Hefei, China, 2010, pp. 1344-1349, doi: 10.1109/ICCSE.2010.5593503.

[7] M. Anbarasi, E. Anupriya, and N. Iyengar, "Enhanced prediction of heart disease with feature subset selection using genetic algorithm," *International Journal of Engineering Science and Technology*, vol. 2, no. 10, pp. 5370-5376, 2010.

[8] S. B. Patil and Y. S. Kumaraswamy, "Intelligent and effective heart attack prediction system using data mining and artificial neural network," *European Journal of Scientific Research*, vol. 31, no. 4, pp. 642-656, 2009.

[9] G. K. Gupta and S. Gupta, "An ensemble model for classification of heart disease dataset," *2016 International Conference on Computing, Communication and Automation*, Noida, India, 2016, pp. 556-560, doi: 10.1109/CCAA.2016.7813767.

[10].https://github.com/Nitish-37/Cardiovascular-disease-detection-using-optimal-feature-selection