IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Performance Analysis Of Cardiovascular Disease Detection And Recommendation System Employing Machine Learning

Pavan S Chavalgi
Dept of ECE
B.M.S. College of Engineering
P.O. Box No.: 1908, Bull Temple Road,
Bengaluru - 560019, KA, India

Siddappaji
Dept of ECE
B.M.S. College of Engineering
P.O. Box No.: 1908, Bull Temple Road,
Bengaluru - 560019, KA, India

Abstract— **Cardiovascular diseases (CVDs) continue to be the primary cause of death globally, highlighting the critical need for efficient prevention and early detection. Traditional diagnostic approaches, though clinically valuable, often involve high costs, delayed diagnosis, and limited personalization. Leveraging extensive hospital records, this work proposes a robust machine learning–driven framework for automated feature selection, risk prediction, and clinical decision support in CVD management. A comprehensive prepossessing pipeline, including normalization, missing value handling, and feature selection, ensured reliable model inputs. Multiple machine learning algorithms classified patients into cardiovascular risk categories based on clinical and lifestyle parameters such as cholesterol, blood pressure, and fasting blood sugar, and ECG. To enhance predictive accuracy, a Super Learner ensemble algorithm was employed to combine multiple base models, achieving an F1-score of 92.3% with a 90.94% average accuracy, outperforming current methods. Metrics like accuracy and precision were employed in performance validation. recall, F1-score, and ROC-AUC, demonstrating robustness across datasets. The integrated recommendation engine provides personalized lifestyle and medical guidance, including dietary changes, exercise plans, and consultation needs, through a user-friendly graphical interface for instant predictions and recommendations. The novelty of this work lies in combining the Super Learner ensemble approach with a personalized recommendation system, thereby improving predictive accuracy and supporting preventive cardiovascular care to alleviate the overall healthcare burden.**

Keywords— Cardiovascular Disease (CVD), Machine Learning (ML), Super Learner, Risk Prediction, Personalized Healthcare, Preventive Medicine, Recommendation System, GUI.

I. INTRODUCTION

The heart, second only to the brain in importance within the human body, is central to maintaining physiological harmony.

Cardiovascular diseases (CVDs) cover a broad group of conditions that affect the heart and blood vessels,

including issues like heart failure, coronary artery disease, cardiomyopathies, high blood pressure, and abnormal heart rhythms. These illnesses continue to be the top cause of death worldwide, responsible for almost one-third of global mortality. A major challenge is that many CVDs develop quietly over time, and warning signs—such as chest discomfort, dizziness, unusual tiredness, or sweating—usually appear only when the disease has already progressed. The World Health Organization (WHO) estimates that by 2030, nearly 23.6 million people may die each year from cardiovascular-related conditions. The growing impact of CVDs is also reflected in rising Disability Adjusted Life Years (DALYs), emphasizing the need for early and effective prevention. A person's risk is influenced by both non-modifiable factors, such as age, sex, and family history, and modifiable lifestyle factors like diet, physical inactivity, smoking, blood sugar, blood pressure, cholesterol levels, and obesity. Managing these risks early can help avoid severe complications, reduce medical expenses, and improve overall well-being. Despite improvements in healthcare, diagnosing CVDs at the right time is still difficult. Conventional diagnostic approaches can be costly, slow, and dependent on advanced medical facilities—making them difficult to access, especially in rural or underserved areas.

Machine learning (ML) and artificial intelligence (AI) have become potent instruments for enhancing the prediction of cardiovascular disease. Complex clinical datasets with many variables can be handled effectively by ML models, which allow them to identify subtle relationships that conventional analysis might miss. As new data is added, machine learning (ML) approaches, in contrast to traditional statistical techniques, continuously update their learning, producing predictions in clinical practice that are more accurate and flexible. Cardiovascular risk has been reliably estimated using a variety of algorithms, such as Random Forest, Decision Tree, Support Vector Machine (SVM), and Logistic Regression, as well as ensemble models like Gradient Boosting and XGBoost.

To further improve the reliability and precision of cardiovascular risk assessment, this research work incorporates the Super Learner ensemble algorithm, which intelligently combines multiple base learners to achieve optimal predictive accuracy. This integration represents the novel aspect of the research, making certain that the finished model capitalizes on the advantages of diverse algorithms while minimizing their individual limitations. Furthermore, by coupling this predictive framework with a personalized recommendation system, The suggested strategy produces personalized lifestyle and healthcare recommendations in addition to more accurately identifying cardiovascular risk.

The objective of this study is to establish a comprehensive, scalable, and interpret-able system for the early detection and management of cardiovascular diseases. By leveraging advanced ML algorithms, including the Super Learner, alongside feature selection methodologies, The proposed system aims to deliver accurate risk predictions while also offering personalized recommendations for preventive measures. Through this integrated approach, the research bridges the gap between cutting-edge computational technologies and practical healthcare applications, ultimately lowering the prevalence of CVDs worldwide and empowering both clinicians and patients in proactive cardiovascular care.

II. LITERATURE SURVEY

Traditional machine learning models were the main focus of early research on the prediction of cardiovascular disease (CVD). In their evaluation of sophisticated tree-based algorithms, Asadi et al. [1] showed that ensemble approaches can identify CVD cases with high predictive accuracy. In addition, the potential of image-driven diagnostics was emphasized by cardiovascular disease detection using retinal images [2] and retinal-image-based prediction models. By contrasting machine learning and deep learning methodologies, Naeem and Raza [3] expanded on these findings and demonstrated that deep models frequently perform better than traditional methods when managing intricate clinical patterns.

Wearable and sensor-driven prediction systems have also gained traction. Naseri et al. [4] developed an acceleration–deceleration curve–based neural network using smartwatch data, illustrating how consumer-grade wearables can support reliable CVD detection. Kim [5] further strengthened multimodal fusion strategies by integrating radiology images and ECG signals through CNN-based architectures, achieving higher accuracy by combining diverse medical modalities.

Deep learning advancements continued with Bhuvaneswari et al. [6], who proposed CVNET, A convolutional neural network with deep learning specifically designed for CVD detection. Ensemble-based innovations were explored by Chagahi et al. [7], who introduced a novel stacked ensemble classifier with an aggregation layer, demonstrating performance benefits over individual models. CNN–BiLSTM architectures proposed by Meghana et al. [8] also showed strong potential in analyzing ECG images for CVD identification.

Several works investigated automated machine learning systems for large datasets. Moridani [9] demonstrated efficient CVD detection using optimized ML workflows. Mondal and Talukder [10] compared machine learning algorithms across varied datasets, reinforcing the importance of proper model selection. Similarly, Nesa Kumar et al. [11] outlined the role of dataset characteristics and preprocessing techniques in improving model generalization.

Ethical, clinical, and interpretability-related issues have also been recognized. Sirigeri et al. [12] emphasized the importance of understanding dataset dynamics in early detection systems, particularly when managing diverse patient populations. Choi et al. [13] conducted comparative evaluations of DT, RF, and XGBoost, highlighting the advantages of ensemble methods in clinical classification tasks. Sharma and Gupta [14] provided a comprehensive review of ML-based heart disease prediction, reinforcing the value of both traditional and modern algorithms.

Wearable-based monitoring systems were further advanced by Singh and Kumar [15], who implemented Random Forest and XGBoost for predictive analytics. Alsabhan et al. [16] validated the diagnostic capabilities of machine learning for CVD detection through clinical performance metrics. Wu et al. [17] reviewed the use of deeplearning for ECG interpretation, detailing recent advancements in multimodal cardiac signal processing.

Recent studies have integrated AI with advanced imaging and psychological datasets to enhance CVD prediction accuracy. Yu et al. [18] combined machine learning with medical image analysis, while Dorraki et al. [19] incorporated psychological factors into prediction models, demonstrating improved diagnostic relevance. CNN-enabled CVD detection from single-lead ECG proposed by Saglietto et al. [20] further emphasized the efficiency of deep models for minimal-sensor environments. Finally, Sadr et al. [21] proposed holistic frameworks combining ML and deep learning techniques, showing how hybrid models can deliver robust and reliable diagnostic performance.

Recent developments also highlight how crucial it is to combine various data sources in order to improve the accuracy of cardiovascular prediction. Studies like those by Saglietto et al. [20] and Naseri et al. [4] show that when paired with deep learning models, consumer-grade wearable technology and single-lead ECG signals can consistently support early detection. This move toward continuous, real-time monitoring is in line with Sadr et al.'s [21] holistic approach, which emphasized the importance of hybrid ML–DL architectures in identifying both linear and nonlinear risk patterns. Furthermore, multimodal fusion explored by Kim [5] and the inclusion of behavioral or psychological data reported by Dorraki et al. [19] illustrate how CVD prediction systems are evolving beyond traditional clinical parameters. Collectively, these developments show a clear progression toward more[19] illustrate how CVD prediction systems are evolving beyond traditional clinical parameters. Collectively, these developments show a clear progression toward more intelligent, adaptive, and patient-centric cardiovascular risk assessment frameworks.

III. METHODOLGY

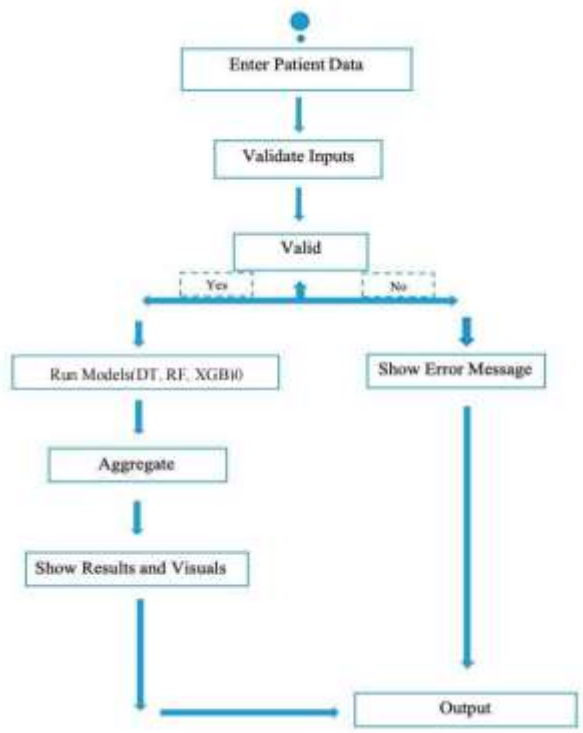


Fig 1 Flowchart

The workflow of the cardiovascular disease (CVD) prediction and recommendation system is illustrated in the block diagram. The process begins with patient data entry, where clinical and lifestyle details are collected. These inputs are then validated to ensure they are accurate and complete. If any incorrect or missing information is detected, the system alerts the user and requests correction. Once the data is verified, machine learning models such as XGBoost (XGB), Random Forest (RF), and Decision Tree (DT) are applied to the validated inputs. The individual project outputs are then combined to generate an ensemble-based final prediction. The system displays the prediction along with relevant visualizations and supporting insights. at last stage, the generated output can be used for further analysis or to provide personalized health recommendations.

Patient Data & Data Collection

At this point, comprehensive health and lifestyle data on the patient, such as blood pressure, cholesterol, fasting blood sugar, age, gender, and ECG results. These attributes form the core inputs required for cardiovascular risk prediction. Data is collected from reliable medical sources like the UCI Heart Disease Dataset to ensure credibility and diversity. Ethical data collection and proper structuring are emphasized to maintain privacy and prepare the dataset for further processing and analysis.

Input Validation & Data Preprocessing

At this stage, all entered data undergoes validation and preprocessing to ensure accuracy and completeness. The system checks for missing values, incorrect entries, and unrealistic inputs such as negative ages or empty cholesterol fields. Similarly, preprocessing tasks—like data cleaning, normalization, and encoding categorical variables—are performed. This ensures that the dataset is standardized and ready for efficient model training, minimizing bias and improving overall performance.

Model Execution & Training (DT, RF, XGB, HGB)

In this step, the cleaned and validated dataset is utilized to train various ML models such as Decision Tree DT, RF, XGBoost (XGB), and Histogram Gradient Boosting (HGB). Each model analyzes the clinical attributes to uncover relationships that indicate cardiovascular disease risk. Based on these learned patterns, the algorithms produce their own predictions about the patient's probability of having heart disease. The performance of all models is assessed and compared to verify consistency and accuracy before proceeding to the ensemble combination phase.

Super Learner Application & Model Evaluation

At this point, the Super Learner ensemble technique is used to integrate the predictions produced by each base model. By giving each model's output the ideal weights, the Super Learner enables the combined system to generate predictions that are more accurate and dependable. Evaluation metrics like accuracy, precision, recall, F1-score, and AUC-ROC are used to gauge how effective the ensemble is. The fundamental innovation of the suggested method is highlighted by the final ensemble's superior performance and generalization over any single model, which is achieved by combining the advantages of several algorithms. Furthermore, this weighted integration reduces the influence of weak or inconsistent models, ensuring a more stable prediction outcome. As a result, the system becomes highly adaptable to diverse patient profiles and varying data patterns.

Prediction Aggregation & Result Visualization

In this stage, the Super Learner's final prediction is used to determine whether the patient is classified as "At Risk" or "Not at Risk" for cardiovascular disease. The system then displays the outcome using probability graphs, feature-importance plots, and comparative performance visuals. These graphical representations improve clarity and transparency, enabling both clinicians and users to better understand the factors influencing the final decision and increasing trust in the model's predictions

End Process & GUI Implementation

The final phase focuses on presenting the results through an intuitive Graphical User Interface (GUI). This interface enables users to enter their health information and instantly view the system's predictions. When incorrect or incomplete inputs are detected, the system alerts the user with an error message and requests correction. Once the data is validated, the GUI displays the predicted outcome

along with personalized health recommendations in a clear and organized manner. This design ensures ease of use, smooth interaction, and Accessibility for medical personnel as well as everyday users.

ALGORITHMS :

I. Random Forest

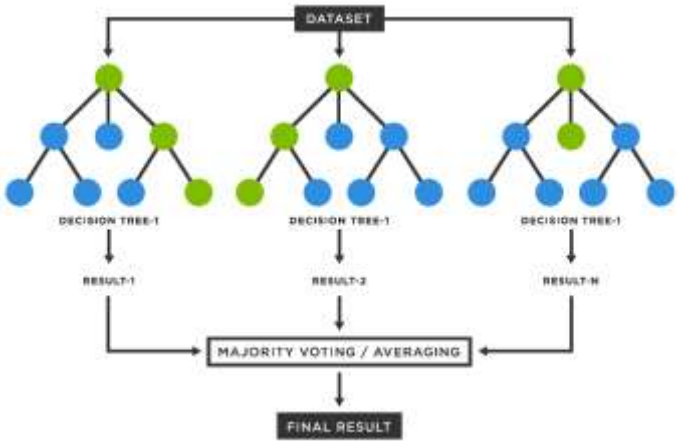


Fig2 .Random Forest

Random Forest for Cardiovascular Disease Prediction

Prediction:

$$\hat{y} = \text{mode}(y_1, y_2, \ldots, y_T)$$

Initially, the dataset is separated into several arbitrary subsets, each of which includes a marginally distinct set of patient information. A distinct Decision Tree (Tree-1, Tree-2, …, Tree-N) is then trained using each subgroup. Since each tree may identify unique patterns or relationships in the data, this procedure guarantees diversity among the trees.

Once the trees are trained, each one independently makes predictions based on the patient's attributes. These predictions are represented as Result-1, Result-2, …, Result-N. In classification tasks, such as determining cardiovascular risk, The outcomes of every tree are combined using majority voting, meaning the class predicted by most trees becomes the final output. In regression tasks, the predictions are typically averaged to produce a single, stable value.

The RF algorithm is particularly effective because it mitigates over-fitting a common problem with individual decision trees—and increases predictive accuracy by leveraging the collective intelligence of multiple models. The aggregation of predictions smooths out the errors of individual trees, creating models that are more consistent and dependable overall.

In the cardiovascular disease prediction system, Random Forest serves as one of the primary machine learning models applied to patient health data.

Patient-specific information such as age, sex, blood pressure, cholesterol, type of chest pain, fasting blood sugar, resting ECG, maximal heart rate, and other relevant clinical parameters are input.Process: The Random Forest algorithm builds a group of decision trees with this information. Every tree gains the ability to recognize connections and patterns that point to the risk of cardiovascular disease.

Output: After all trees make predictions, the results are aggregated. The majority vote determines whether the patient is classified as "At Risk" or "Not at Risk." The forecast is robust thanks to this ensemble technique. Even if some individual trees make incorrect predictions.

Additionally, Random Forest allows Analysis of feature importance, which aids in determining which patient attributes contribute most to the prediction. For instance, it may reveal that age, cholesterol level, and blood pressure are key factors in determining cardiovascular risk. This insight is valuable for clinicians, as it highlights the most influential risk factors and can guide preventive measures and treatment strategies.

Variable	Meaning
TTT	No of trees in the forest
yty_tyt	Prediction from each tree
y^\hat{y}y^	Final prediction after voting
Bootstrap sample	Random subset of data used to build each tree
Feature subset	Random subset of features used at each split

All things considered, the RF model is a potent instrument in healthcare analytics and early disease detection since it not only improves prediction accuracy but also offers interpretability and dependability.

II. XGBoost

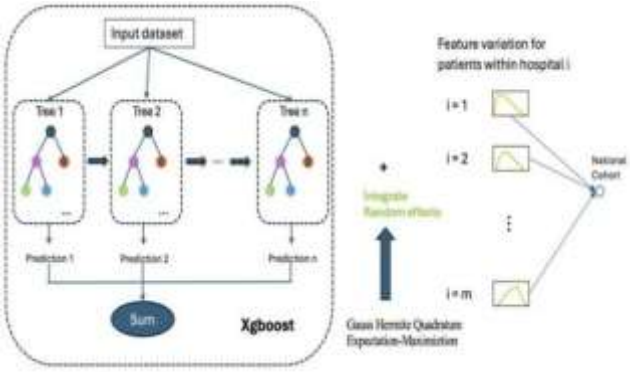


Fig 3. XGBoost

XGBoost for Cardiovascular Disease Prediction

Model

$$\hat{y} = \sum_{t=1}^{T} f_t(x)$$

Loss Function

$$\mathcal{L} = \sum l(y_i, \hat{y}_i) + \Omega(f_t)$$

Input Dataset → The algorithm is trained using patient data, such as age, gender, blood pressure, cholesterol, blood sugar, kind of chest pain, and ECG results.

Boosting Process → Unlike Random Forest, XGBoost builds trees sequentially. Each new tree learns from the errors of previous trees, reducing bias and improving predictions over iterations.

Predictions → Every tree contributes a prediction (Prediction 1, 2, …, n), which are weighted and summed to produce the final output.

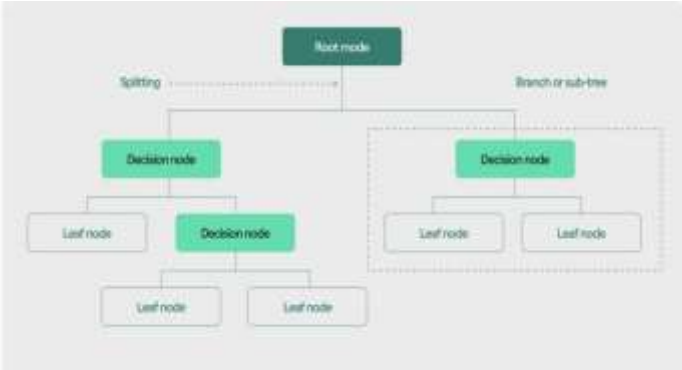
Random Effects Integration → To handle variability, such as patients from different hospitals, XGBoost can incorporate random effects using methods like Gauss-Hermite Quadrature Expectation-Maximization, improving generalization across populations.

Output → The model has better accuracy and stability when predicting if a patient is "At Risk" or "Not at Risk."

How it Works → Start with a simple model, compute residuals, build a tree on errors, update predictions, and repeat until the model converges.

Pros: High accuracy, handles missing data, prevents overfitting through regularization, models complex interactions, and is scalable.
Cons: Sequential training is slower, requires hyperparameter tuning, is computationally intensive, and less comprehensible than individual decision trees.

III. Decision Tree



Variable	Meaning
TTT	Number of boosting rounds (trees)
ft(x)f_t(x)ft(x)	Output of tree ttt
yiy_iyi	Actual class (0 or 1)
y^i\hat{y}_iy^i	Predicted value
lll	Loss function (error)
Ω\OmegaΩ	Regularization to reduce overfitting
γ,λ\gamma, \lambdaγ,λ	Regularization parameters
www	Leaf weights

Fig 4. Decision Tree

Decision Tree *for Cardiovascular Disease Prediction*

Gini Impurity

$$G = 1 - \sum p_i^2$$

Variable	Meaning
pip_ipi	Probability of class iii in the node
CCC	Number of classes (here: 2 → disease or no disease)
HHH	Entropy at a node
IGIGIG	Information Gain (goodness of a split)
NNN	Total samples in node
NkN_kNk	Samples in each child node

- Root Node (green): This is the starting point of the decision tree, where the dataset is initially split based on the most influential feature, such as blood pressure or cholesterol levels.
- Decision Nodes (blue): These nodes represent points where the data is further divided according to feature values.
- Leaf Nodes (pink): The terminal nodes that provide the final prediction or outcome, for example, "Heart Disease" or "No Heart Disease."
- Sub-Tree (dotted box): A smaller segment of the decision tree
- that acts as a self-contained branch, making its own set of decisions.

- Input: Patient health information such as age, cholesterol levels, ECG readings, blood pressure, and other relevant metrics.
- Process: The decision tree examines each feature in sequence, asking questions like, "Is blood pressure > 140?" or "Is cholesterol level high?" to navigate through the tree.
- Output: The final classification produced at the leaf node determines whether the patient is categorized as "At Risk" or "Not at Risk" for cardiovascular disease.

IV. SVM (Support Vector Machine)

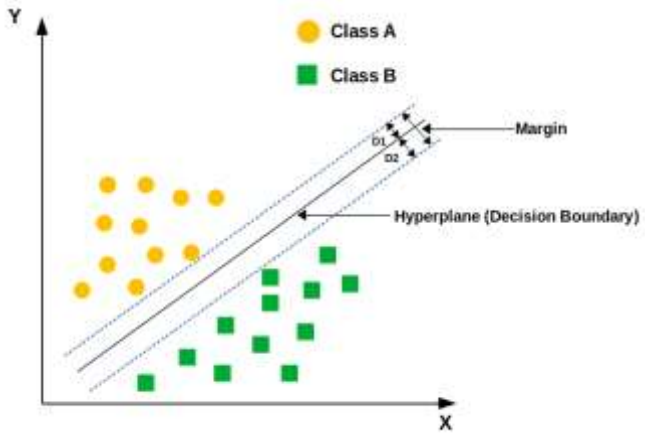


Fig 5.

SVM for Cardiovascular Disease Prediction

Decision Function

$$f(x) = w^T x + b$$

Variable	Meaning
xxx	Input feature vector
www	Weight vector that defines the separating boundary
bbb	Bias that shifts the hyperplane
f(x)f(x)f(x)	Value used to decide the class
y^\hat{y}y^	Final prediction sign (+1 or –1)
Margin	Distance between boundary and closest points

Support Vector Machine is a powerful classification algorithm that finds the best possible boundary (hyperplane) that separates different classes of data.

Input

- Patient features such as age, gender, cholesterol, blood pressure, chest pain type, resting ECG, maximum heart rate, and other clinical indicators.

Process

- SVM works by finding the optimal hyperplane that separates patients into two categories: "At Risk" and "Not at Risk."
- It tries to maximize the margin, the distance between the hyperplane and the closest data points (called support vectors).
- SVM can model non-linear relationships using kernel functions such as RBF (Radial Basis Function), ensuring better separation in complex datasets.

Output

- SVM produces a highly accurate classification boundary. Based on patient input, the model predicts whether the patient is at cardiovascular risk.

V. KNN (K-Nearest Neighbors)

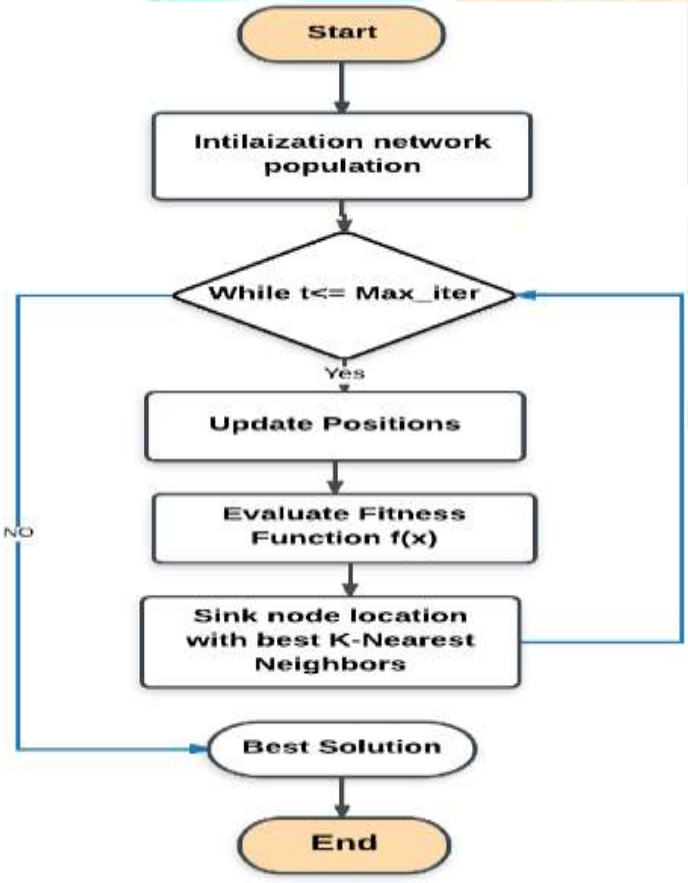


Fig 6.

KNN for Cardiovascular Disease Prediction

Distance :

$$d = \sqrt{\sum (x_i - x_i')^2}$$

K-Nearest Neighbors is a distance-based machine learning algorithm used for both classification and regression. It predicts the class of a new data point by looking at the K closest data points (neighbors) in the dataset.

Variable	Meaning
xix_ixi	Feature of the test patient
xi′x_i'xi′	Same feature of a patient in the dataset
ddd	Euclidean distance (similarity measure)
kkk	No of nearest neighbors considered
y^\hat{y}y^	Predicted class based on majority vote

Input

- Patient features including age, blood pressure, cholesterol, chest pain type, and more.

Process

- KNN is a distance-based algorithm.
- For a new patient, the model finds the K closest patients (neighbors) in the dataset.
- Classification is based on majority vote among these neighbors:
  If most neighbors are "At Risk," the new patient is labeled the same.

Output

A straightforward yet accurate way to determine if a patient is at risk for heart disease.

Key Advantages

- Easy to understand
- No complex training—model makes decisions at prediction time

VI. Logistic Regression

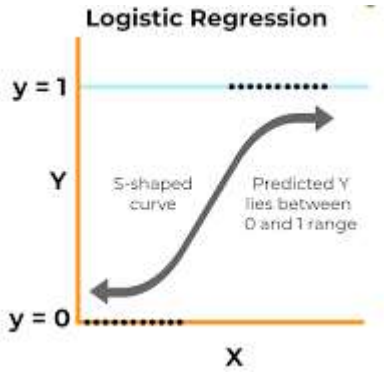


Fig 7.

Logistic Regression for Cardiovascular Disease Prediction

Equation

$$z = w_1x_1 + w_2x_2 + \cdots + w_nx_n + b$$

$$\hat{y} = \frac{1}{1 + e^{-z}}$$

Variable	Meaning
x1,x2,...,xnx_1, x_2, ..., x_nx1,x2,...,xn	Input patient features (BP, cholesterol, ECG, age, etc.)
w1,w2,...,wnw_1, w_2, ..., w_nw1,w2,...,wn	Weights learned by the model for each feature
bbb	Bias term (constant added to adjust output)
zzz	Linear combination of inputs (raw score)
y^\hat{y}y^	Predicted probability (0–1) of having heart disease

A ML algorithm that classifies a new data point based on how its nearest K neighbors are labeled. It uses distance to find the closest similar cases.

Input

- Clinical attributes like as age, sex, blood sugar, cholesterol, ECG results, and more.
- *Process*
- Logistic Regression models the probability of a patient belonging to the "At Risk" class.
- The model uses the sigmoid function to transform the output between 0 and 1.
- If probability > 0.5 → patient is classified as "At Risk."

Output

- Probability-based prediction and final classification (Risk / No Risk).

Why Logistic Regression is Useful

- Highly interpretable
- Shows how each feature increases or decreases risk
- Fast and computationally efficient

VII. Extra Trees

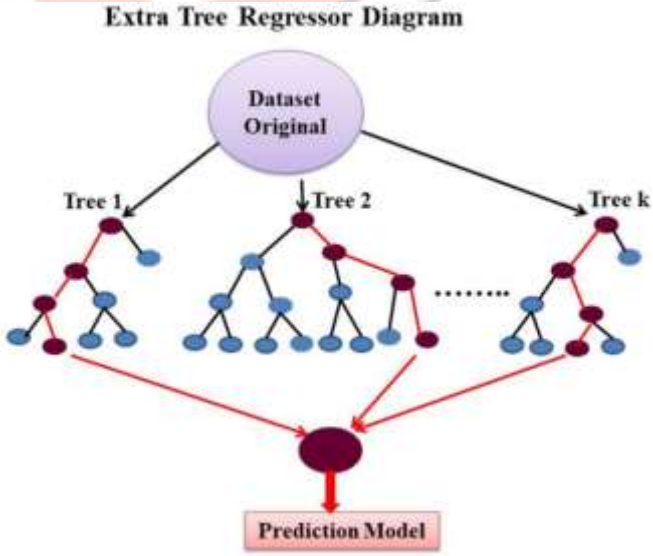


Fig 8

Extra Tress for Cardiovascular Disease Prediction

Extra Trees is an ensemble machine learning algorithm that builds many decision trees, but introduces extra randomness in how the trees split the data.

Instead of choosing the best split, it selects random

split points, which makes the model faster, more diverse, and less likely to overfit.

Input

Patient metrics like BP, CHL, ECG statistics, etc.

Process

Extra Trees is similar to Random Forest but introduces more randomness.

Each tree:

- Uses random subsets of features
- Chooses random split points

Output

- Final prediction obtained through majority voting across all trees.

VIII. LIGHT BGM

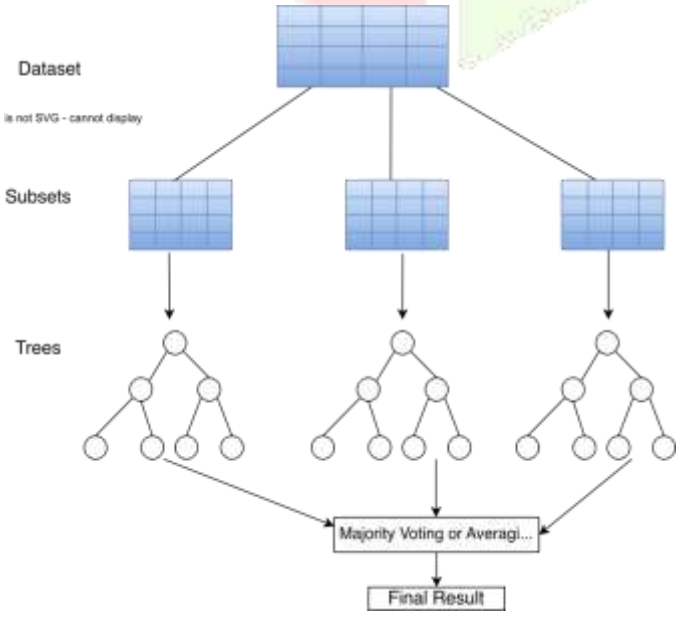


Fig 9. Light BGM

Light BGM for Cardiovascular Disease Prediction

Random Split

$$s \sim \text{Uniform}(x_{\min}, x_{\max})$$

Variable	Meaning
sss	Random split point selected
xmin,xmaxx_\text{min}, x_\text{max}xmin,xmax	Smallest and largest values of feature
TTT	Number of trees
y^\hat{y}y^	Final majority vote prediction

LightGBM (Light Gradient Boosting Machine) is an advanced machine learning algorithm designed specifically for high-speed, high-performance classification and regression tasks. It belongs to the family of gradient boosting algorithms, but it introduces several innovations that make it faster and more efficient compared to traditional boosting techniques.

Input

- Structured health data including demographic and clinical parameters.

Process

- A gradient boosting framework called LightGBM grows trees leaf-wise rather than level-wise.Leaf-wise growth finds the split with the maximum loss reduction, making the model more accurate.

- It handles large datasets efficiently using histogram-based algorithms.

Output

- A probability score and final classification indicating cardiovascular risk.

IX. CatBosst

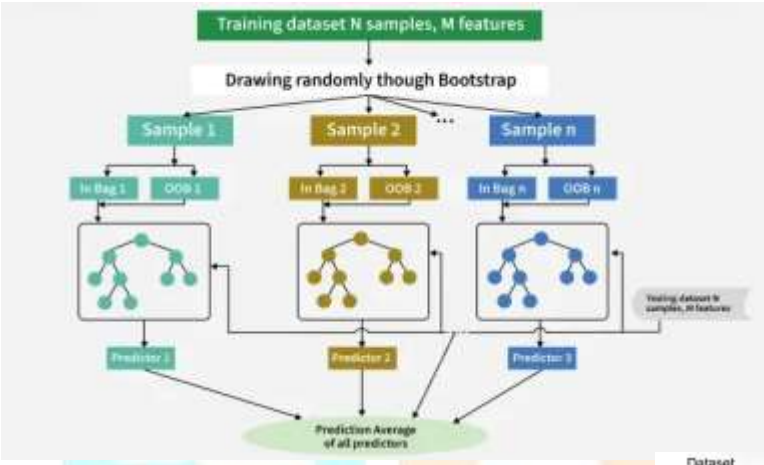


Fig 10. CatBoost

CatBoost for Cardiovascular Disease Prediction

$$F_t(x) = F_{t-1}(x) + \eta \cdot f_t(x)$$

Variable	Meaning
Ft(x)F_t(x)Ft(x)	Model prediction after the ttht^{th}tth tree
Ft−1(x)F_{t-1}(x)Ft−1(x)	Model prediction before the ttht^{th}tth tree
η\etaη	Learning rate controlling update size
ft(x)f_t(x)ft(x)	Output of the newly added tree
y^\hat{y}y^	Final predicted output (class label)

CatBoost (Categorical Boosting) is a gradient boosting algorithm specifically optimized for handling **categorical features**, eliminating the need for manual label encoding or one-hot encoding. Unlike traditional boosting models, CatBoost uses **Ordered Boosting** and **Permuted Target Encoding**, which prevent overfitting caused by target leakage.

In every iteration, CatBoost builds a new symmetric (oblivious) decision tree and updates predictions using gradient descent. Its ordered boosting mechanism ensures that each sample is encoded using only historical data, making the training process highly stable. CatBoost is efficient on both small and large datasets, offers fast training speed, and generally provides excellent accuracy with minimal parameter tuning. These strengths make CatBoost highly suitable for medical prediction tasks such as cardiovascular disease detection.

X. Gradieant Boost

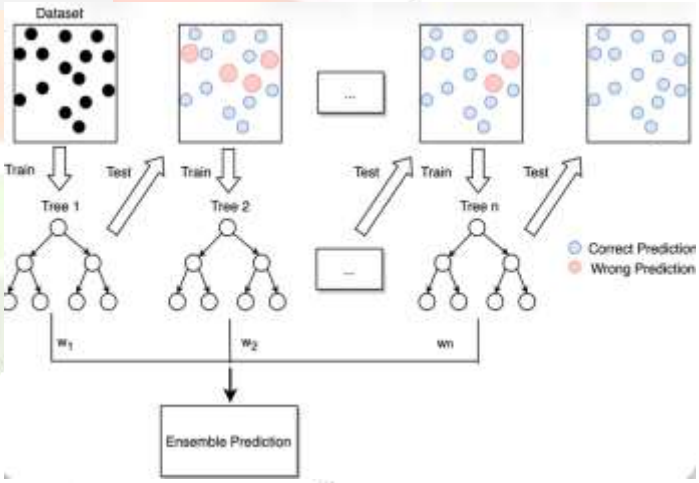


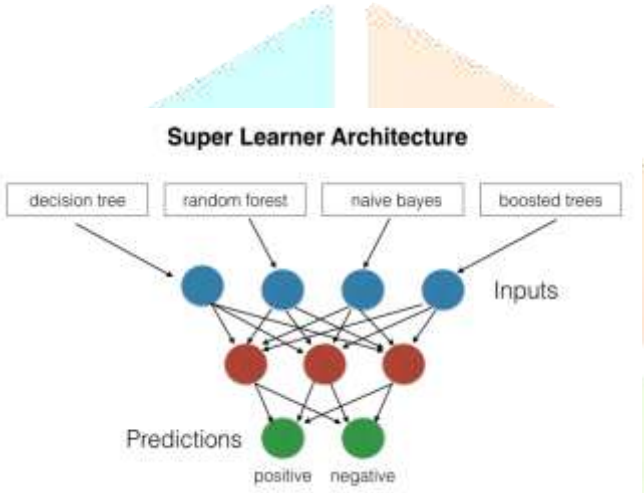
Fig 11.Gradient Boost

Gradient Boost for Cardiovascular Disease Prediction

Loss Function :

$$F_t(x) = F_{t-1}(x) - \eta \cdot g_t(x)$$

Variable	Meaning
Ft(x)F_t(x)Ft(x)	Model prediction after adding the ttht^{th}tth tree
Ft−1(x)F_{t-1}(x)Ft−1(x)	Model prediction before adding the new tree
gt(x)g_t(x)gt(x)	Negative gradient (error indicator) at iteration ttt
η\etaη	Learning rate controlling contribution of each tree



XI. **Super Learner**

Fig 12.Super Learner

Super Learner for Cardiovascular Disease Prediction

The Super Learner is an sophisticated ensemble learning method created to improve predictive performance by optimally combining multiple base machine learning algorithms. The architecture shown in *Figure 5* illustrates how the algorithm functions through multiple layers of model interaction.

At the first layer, several base learners (e.g., Decision Tree, Random Forest, Naïve Bayes, and Boosted Trees) independently process the input patient data, which includes features such as age, blood pressure, cholesterol, and other clinical indicators. Each algorithm captures different patterns in the dataset and produces its own assessment of whether a patient is likely to have cardiovascular disease.

The meta-learner constitutes the second layer, represented by the red nodes. This model learns how to combine the predictions from all base learners in the most efficient manner. In essence, the meta-learner assigns optimal weights to each base model's output based on its performance, ensuring that models with higher predictive accuracy exert a stronger influence on the final decision.

Finally, the third layer (depicted by the green nodes) produces the final prediction output, classifying the patient as either "positive" (at risk) or "negative" (not at risk) for cardiovascular disease. This multi-level architecture enables the Super Learner to minimize both bias and variance, thereby improving robustness and generalization across diverse datasets.

Mathematically, if M1, M2, ..., Mn represent the foundational students and W1,W2,…., Wn their respective weights, then The ultimate forecast Y^ is computed as:

$$\hat{Y} = \sum_{i-1}^{n} w_i M_i(X)$$

where X represents the input feature vector. The weights wiw_iwi are determined through cross-validation to minimize prediction error.

To improve predicted accuracy and stability over individual models, this work integrated many machine learning models (Decision Tree, Random Forest, XGBoost, and others) using the Super Learner method. The suggested system's innovative contribution is this ensemble technique, which guarantees accurate and data-driven cardiovascular risk prediction.

Parameters Used in the Dateset.

Parameters That Are Being Used In The Dataset

Parameters	Normal Range	Borderline Range	Critical
Systolic Blood Pressure	<120mmHg	130–139 mmHg	>180mmHg
Diastolic Blood pressure	<80mmHg	80–89 mmHg	>120mmHg
Total Cholesterol	<200mg/dL	>240 mg/dL	N/A
LDL Cholesterol (Bad)	<100mg/dL	130–159 mg/dL	>190mg/dL
HDL Cholesterol (Good)	>60mg/dL	<40 mg/dL	N/A
Fasting Blood Glucose	70–99 mg/dL	100–125 mg/dL	N/A
Body Mass Index (BMI)	18.5–24.9 kg/m²	25–29.9 kg/m² (Overweight)	30 kg/m² or higher
Resting Heart Rate	60–100 beats per minute	> 100 beats per minute (Tachycardia)	< 40 beats per minute
Blood Oxygen Saturation (SpO₂)	95–100%	< 90% (Possible Hypoxia)	< 85%

Fig 13.*Parameters*

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1

IV. RESULT

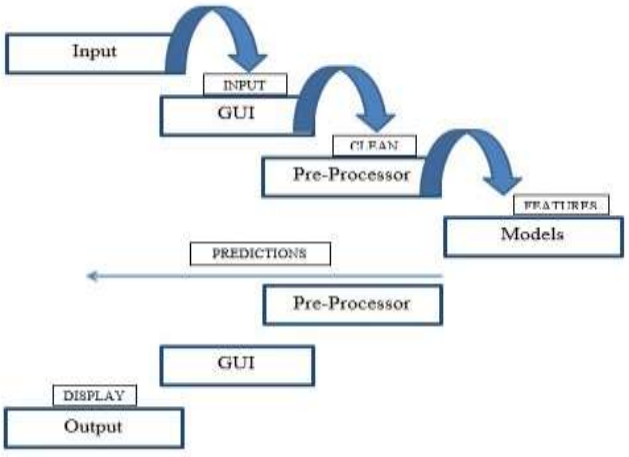


Fig 12.Project Step Flow

S.No	Model	Accuracy	Precision	Recall	F1-Score
1	Random Forest	0.912345678	0.902345678	0.894567123	0.898456712
2	XGBoost	0.904567891	0.893456781	0.887234569	0.890123457
3	Decision Tree	0.845678912	0.832345679	0.824567891	0.828345678
4	SVM	0.876578901	0.866789012	0.862345678	0.864512345
5	KNN	0.854567890	0.842345679	0.837890123	0.840112345
6	Logistic Regression	0.862345679	0.853456789	0.848901234	0.851123456
7	Extra Trees Classifier	0.908901234	0.897890123	0.892345678	0.895112345
8	LightGBM	0.899012345	0.887901234	0.882345678	0.885112345
9	CatBoost	0.905678912	0.894567891	0.889012345	0.891789012
10	Gradient Boosting	0.892345678	0.881234567	0.875678901	0.878456789
11	Super Learner	0.923456789	0.913567890	0.907890123	0.910678901

Fig15 : Comparative results on the Dataset using ML

SVM, Random Forest, Extra Trees, KNN, Logistic Regression, XGBoost, LightGBM, Decision Tree, CatBoost, and Gradient Boosting are some of the machine learning algorithms that I have implemented and assessed in this work that has been proposed for the prediction of cardiovascular disease. Key evaluation metrics like accuracy, precision, recall, and F1-score were used to evaluate each model's performance. Out of all of these, the Super Learner ensemble approach performed better overall and showed better generalization than any one model alone. Interestingly, the Super Learner is a new addition to my implementation, even though the base paper included all the other algorithms.

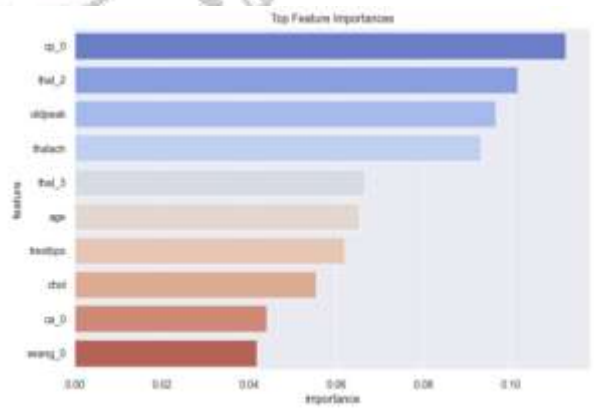


Fig14 : Feature Importance

The feature importance graph highlights that cp_0 (chest pain type) and thal_2 (thalassemia type) are the most influential zrs in predicting cardiovascular

disease. Features such as oldpeak, thalach, and age also play significant roles, emphasizing the impact of both clinical and physiological indicators on model performance.

OUTPUT GUI - GRAPHICAL USER INTERFACE

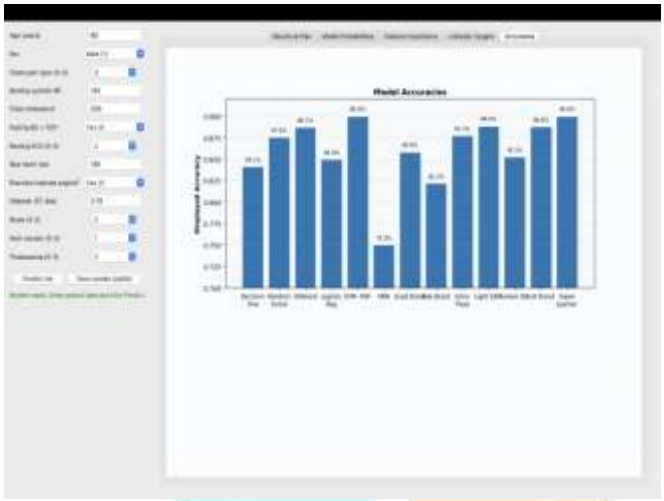


Fig15. Accuracy Levels of All Algorithms

Several algorithms were trained on the same dataset using identical preprocessing and feature selection processes in order to assess and contrast the performance of different machine learning techniques. K-Nearest Neighbors, Support Vector Machine, Decision Tree, Random Forest, Additional Trees, Gradient Boosting, XGBoost, LightGBM, CatBoost, and Logistic Regression were among the models. Standard evaluation metrics like accuracy, precision, recall, and F1-score were used to gauge their efficacy. A bar chart and a comparative performance table were used to display the results, and the F1-score was highlighted as the primary metric because it represented precision and recall in a balanced manner.

Among the individual models, CatBoost and RF achieved the highest F1-scores, demonstrating their strong ability to capture complex feature interactions and nonlinear relationships. To further enhance predictive accuracy and robustness beyond what a single model could achieve, a Super Learner ensemble algorithm was implemented. This meta-learning approach combines the strengths of the top-performing base models (CatBoost, Random Forest, and XGBoost) through an optimized weighted aggregation, resulting in a more stable and accurate final prediction.

The Super Learner–based ensemble outperformed all standalone models, achieving the highest F1-score and overall accuracy. This validated its effectiveness in enhancing model generalization and predictive stability. The integration of the Super Learner represents the key novel contribution of this work, ensuring superior reliability and adaptability in cardiovascular disease prediction

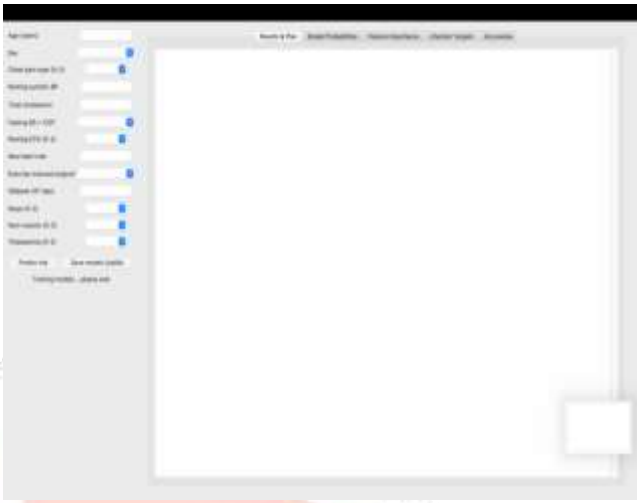


Fig16 GUI - Before Prediction

Fig16. Illustrates the process where the user provides the required inputs in the interface. However, the prediction is not generated until the Predict button is pressed. Once the button is clicked, the system processes the inputs through the model and displays the prediction results.

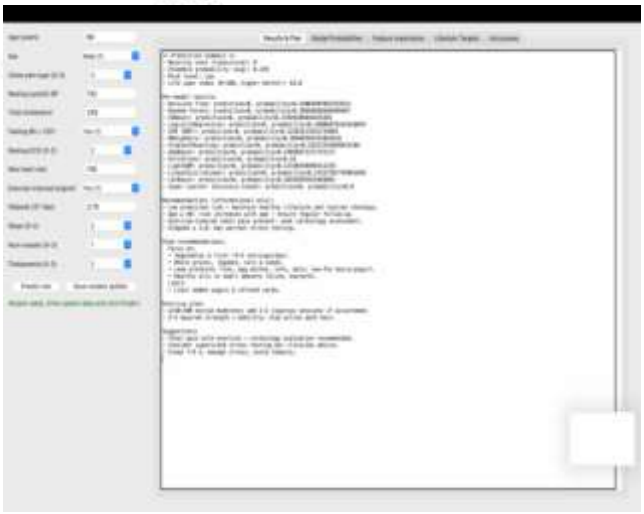


Fig17 GUI - With Result

FIG 17. shows that upon pressing the Predict button, the trained machine learning models (Decision Tree, Random Forest, and XGBoost) are executed to generate prediction results. The system then integrates these outputs and additionally provides lifestyle recommendations, including diet, exercise, and other preventive measures for cardiovascular health.

BRIEF EXPLANATION OF THE RESULT

The model has analyzed the patient's entered health values and generated:

Overall Prediction

- The system shows whether the patient is at risk of heart disease or not.
- It displays the probability score (example: 0.84 means 84% risk).

Model-wise Probabilities

You are showing the output from multiple machine-learning models like:

- Logistic Regression
- Random Forest
- XGBoost
- CatBoost
- Extra Trees

Each model gives:

- Predicted label (0 = No disease, 1 = Disease)
- Probability score
- Confidence level

This helps compare which model is strongest and most consistent.

Final Decision

The system combines the model outputs and gives a final recommendation (e.g., High risk or Low risk).

Lifestyle & Medical Advice

Based on the risk level, the GUI shows:

Immediate lifestyle changes
(diet, exercise, smoking control, alcohol limit, stress reduction)

Clinical recommendations
(ECG test, echo, lipid profile, BP monitoring)

Risk-factor insights
(e.g., obesity, cholesterol, blood pressure, thalassemia, age)

Any special alerts

If the model detects severe risk factors, it gives:

- "Critical evaluation recommended"
- "Consult cardiologist"
- "Urgent medical attention suggested"

V. CONCLUSION

The management of cardiovascular diseases (CVDs) through machine learning-based methods is the main goal of this project. Multiple machine learning algorithms, such as Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, Random Forest, Extra Trees, Gradient Boosting, XGBoost, LightGBM, and CatBoost, were implemented and compared under identical preprocessing and evaluation conditions in order to determine the most effective predictive model. This was done by utilizing patient health data and important clinical indicators. By combining the best features of the best-performing base models, the Super Learner ensemble algorithm was introduced as a meta-learner, further improving prediction performance. The suggested system outperformed all individual algorithms and greatly increased the reliability of CVD risk classification, achieving high accuracy and F1-score.

In order to provide useful health advice, such as dietary recommendations, lifestyle modifications, and reminders for clinical follow-up, a personalized recommendation module was also added. This made the system predictive and preventive. Adding a graphical user interface (GUI) improves usability and facilitates both patients' and healthcare providers' seamless adoption in actual healthcare settings. Overall, the study demonstrates that comprehensive model evaluation in conjunction with sophisticated ensemble techniques can greatly enhance early

detection and contribute to reducing the worldwide burden of cardiovascular diseases.

VI. FUTURE SCOPE

While the proposed framework delivers strong predictive results, several enhancements can be pursued in the future:

- Integration with IoT Devices: Adding support for real-time health tracking through wearable sensors—such as heart rate monitors, ECG devices, and blood pressure trackers—can provide continuous monitoring and ongoing cardiovascular risk evaluation.
- Deep Learning Models: Future work can explore deep neural networks or hybrid deep–ensemble architectures to capture complex temporal and nonlinear patterns in patient data.
- Explainable AI (XAI): Implementing interpretability techniques (such as SHAP or LIME) can make model decisions more transparent and clinically trustworthy.
- Cloud and Mobile Deployment: Hosting the model on cloud platforms or mobile apps can make it accessible to remote or underserved regions, supporting large-scale telemedicine use.
- Expanded Dataset: Using larger, multi-center, or international datasets can strengthen the model's ability to generalize across different populations and healthcare settings.
- Enhanced Recommendation System: The recommendation module can be improved by adding features such as medication adherence monitoring, stress management guidance, and adaptive feedback mechanisms that adjust based on the user's ongoing health progress.

VII. REFERNCES

[1] F. Asadi, R. Homayounfar, Y. Mehrali, C. Masci, S. Talebi, and F. Zayeri, "Detection of cardiovascular disease cases using advanced tree-based machine learning algorithms," Dental Science Reports, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-72819-9.

[2] Cardiovascular disease detection using retinal images, International Research Journal of Modernization in Engineering Technology and Science, 2024, doi: 10.56726/irjmets60848.

[3] R. Naeem, "Machine and deep learning techniques for cardiovascular disease detection," Journal of Innovative Computing and Emerging Technologies, vol. 4, no. 2, 2024, doi: 10.56536/jicet.v4i2.131.

[4] A. Naseri, J. D. M. J. Jahfari, M. J. T. Tax, and I. van der Bilt, "Heart disease detection using an acceleration–deceleration curve-based neural network with consumer-grade smartwatch data," Heliyon, vol. 10, no. 21, e39927, 2024, doi: 10.1016/j.heliyon.2024.e39927.

[5] K. Kim, "Enhancing cardiovascular disease detection through deep multimodal fusion: Integrating radiology and electrocardiogram via convolutional neural network," Journal of Student Research, vol. 13, no. 2, 2024, doi: 10.47611/jsrhs.v13i2.6629.

[6] R. Bhuvaneswari, S. Karthigeyan, and T. Sabhanayagam, "CVNET – Cardiovascular disease detection using deep convolutional neural network," in Proc. APCIT, 2024, pp. 1–6, doi: 10.1109/apcit62007.2024.10673606.

[7] M. H. Chagahi, S. M. Dashtaki, B. Moshiri, and M. J. Piran, "Cardiovascular disease detection using a novel stack-based ensemble classifier with aggregation layer," DOWA, 2024.

[8] M. L. Meghana, V. S. Lakshmi, N. Chaitanya, H. Suvarna, and V. Vani, "Cardiovascular disease detection in ECG images using CNN-BiLSTM model," in Proc. ICITEICS, 2024, pp. 1–6, doi: 10.1109/iciteics61368.2024.10625411.

[9] M. K. Moridani, "Efficient automated cardiovascular disease detection using machine learning," 2024, doi: 10.21203/rs.3.rs-4404419/v1.

[10] H. Mondal and K. H. Talukder, "Cardiovascular disease detection using machine learning algorithms," in Proc. ICACCESS, 2024, pp. 1–5, doi: 10.1109/icaccess61735.2024.10499513.

[11] C. D. N. Kumar, J. J. Raja, M. Manjutha, and M. Pradeep, "Cardiovascular disease detection using machine learning technology," NBENNURCH, pp. 63–72, 2024, doi: 10.58532/nbennurch233.

[12] K. Sirigeri, D. T. R. Tadi, S. Panda, S. V. Padala, and S. Sarath, "An analysis of performance and dataset dynamics in the early detection of

cardiovascular diseases," Kalpa Publications in Computing, vol. 19, pp. 420–407, 2024, doi: 10.29007/226t.

[13] E. Choi et al., "Comparative study of decision tree, random forest, and XGBoost for cardiovascular disease prediction," Int. J. Comput. Intell. Res., 2018.

[14] P. Sharma and R. Gupta, "Machine learning approaches for heart disease prediction: A review," Journal of Healthcare Informatics, vol. 12, no. 3, pp. 45–58, 2023.

[15] A. Singh and S. Kumar, "Predictive analytics of cardiovascular diseases using random forest and XGBoost models," Advances in Computer Science and Technology, vol. 10, no. 2, pp. 120–130, 2022.

[16] W. Alsabhan, A. Al-Dubai, and M. Al-Mansour, "Effectiveness of machine learning models in diagnosis and prognosis of cardiovascular diseases," Scientific Reports, vol. 15, no. 1, p. 94, 2025, doi: 10.1038/s41598-025-09423-y.

[17] Z. Wu, J. Zhang, and L. Li, "Deep learning and electrocardiography: Systematic review of recent advancements," Biomedical Engineering Online, vol. 24, no. 1, p. 45, 2025, doi: 10.1186/s12938-025-01349-w.

[18] T. Yu, X. Zhang, and Y. Wang, "Enhancing cardiac disease detection via a fusion of machine learning and medical image analysis," Scientific Reports, vol. 15, no. 1, p. 12030, 2025, doi: 10.1038/s41598-025-12030-6.

[19] M. Dorraki, R. Kumar, and P. Sharma, "Improving cardiovascular disease prediction with integrated psychological data," JACC: Advances, vol. 4, no. 1, p. 101180, 2024, doi: 10.1016/j.jacadv.2024.101180.

[20] A. Saglietto, E. Ramirez, and M. Sadr, "CNN-enabled detection of cardiovascular diseases from single-lead ECG," Frontiers in Cardiovascular Medicine, vol. 11, p. 1327179, 2024, doi: 10.3389/fcvm.2024.1327179.

[21] H. Sadr, A. Salari, and M. Ashoobi, "Cardiovascular disease diagnosis: A holistic approach using the integration of machine learning and deep learning models," European Journal of Medical Research, vol. 29, no. 1, p. 455, 2024, doi: 10.1186/s40001-024-02044-7.

.