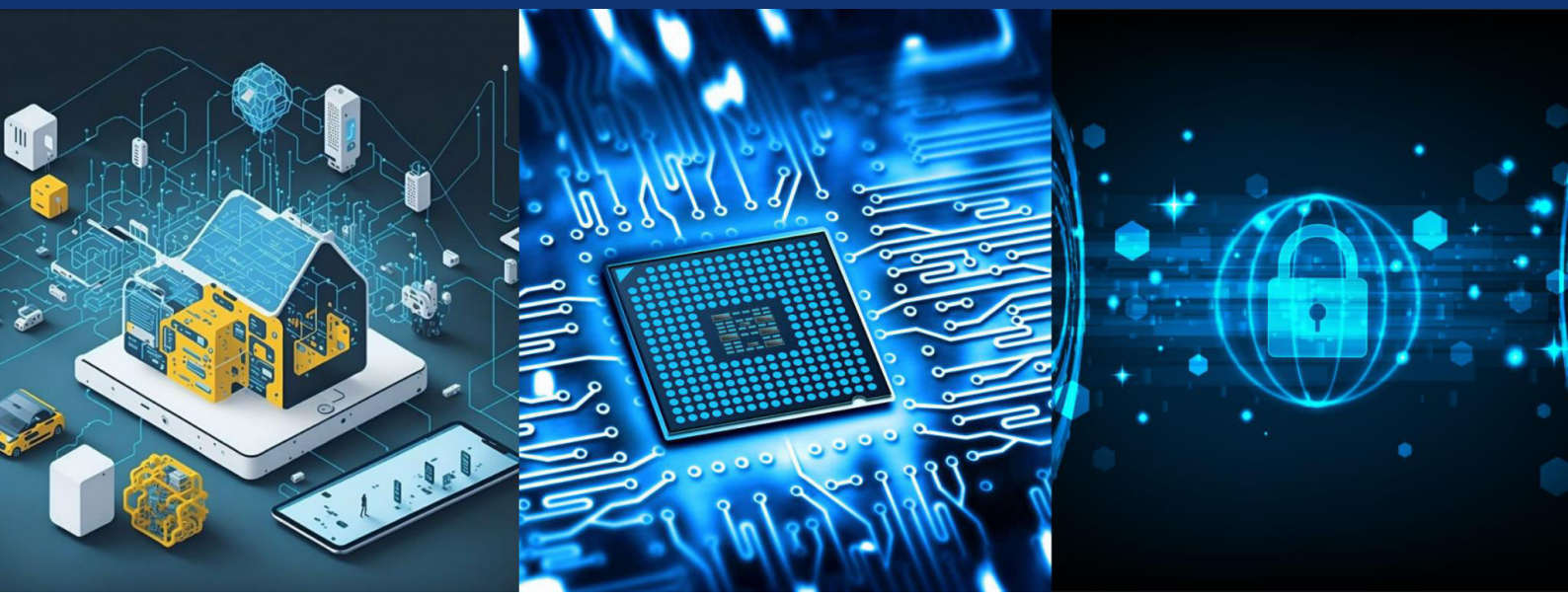
IJIRCCE

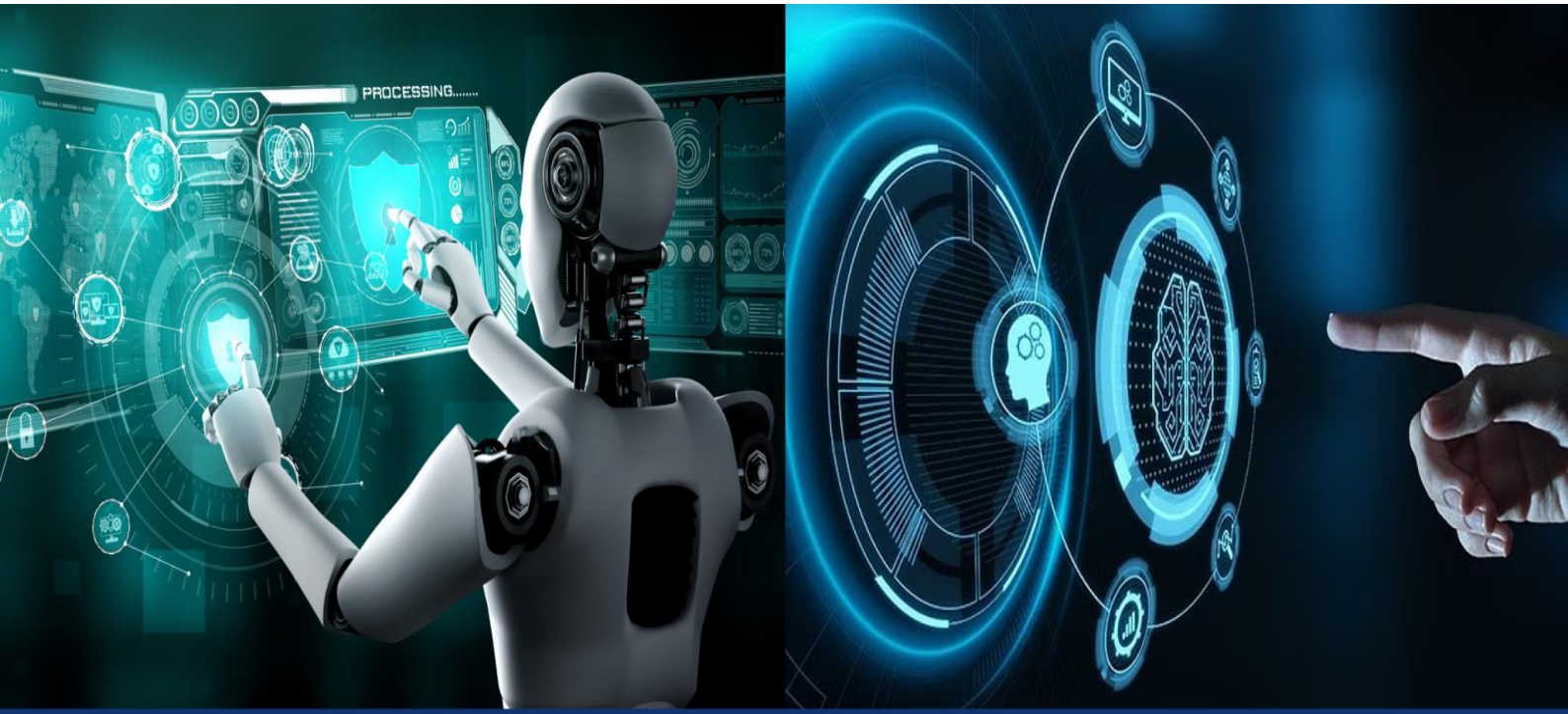
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.625

Volume 13, Issue 1, January 2025

© 2025 IJIRCCE | Volume 13, Issue 1, January 2025| DOI: 10.15680/IJIRCCE.2025.1301047

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

HeartPredict-Leveraging Five ML Algorithms for Cardiovascular Disease (CVD) Forecasting

Dr. Nandini N, Spoorthi Shetty J K, Ambit P M, Shishir B V, Sudeep Swamy Naik

Head, Department of Computer Science and Engineering, Dr. Ambedkar Institute of Technology, Bengaluru, Karnataka, India

Students, Department. of Computer Science and Engineering, Dr. Ambedkar Institute of Technology, Bengaluru, Karnataka, India

ABSTRACT: We live in a modern technological era where technology is being used to address problems, including health issues. With changing lifestyles, health concerns, especially heart diseases, are on the rise. WHO reports over 11 million annual deaths worldwide due to heart-related complications, emphasizing the need for solutions to educate people and enable early diagnosis?

Machine learning, with its recent advancements, offers a scalable solution to address this issue and helps in early prediction of Cardiovascular Disease (CVD). Our project, **"**HeartPredict**",** uses machine learning models to predict the risk of heart disease based on historical medical data. The system runs five machine learning models on user-provided data and offers predictions. It aims to provide a user-friendly, remote-access platform for screening heart disease risk, reducing the burden on the medical system.

KEYWORDS: WHO reports, early prediction of Cardiovascular Diseases, 5 machine learning models user-friendly, remote-access platform**.**

I. INTRODUCTION

The growing population, lifestyle changes, emerging diseases, pandemics, and environmental changes have reshaped how we approach health. Maintaining good health is crucial for individual productivity and is a major concern globally. However, providing quality healthcare remains a challenge, especially for developing and poorer nations.

The strength of healthcare systems varies across countries and can be measured by the healthcare worker-to-population ratio. Developed nations like the USA and Western Europe have strong ratios (~25:10,000), whereas Southeast Asian countries like India fall below the world average, with a ratio of 12.21:10,000 compared to the global average of ~10:1,000 [1].

Increasing the doctor-to-patient ratio is a lengthy process requiring years and significant investment in infrastructure and modern equipment. Many people also avoid visiting doctors for early-stage diseases, emphasizing the need for scalable, software-based systems for pre-screening and diagnosis.

Scalable software solutions can expand healthcare access without requiring additional workforce. They enable early disease detection, speed, and precision in diagnostics, improved patient management, and safer treatment options. Unlike infrastructure and workforce expansion, which are time-consuming and permanent, software systems can scale quickly and affordably by increasing computing power.

Diagnosis often depends on a doctor's experience and expertise, which can be supplemented by computer algorithms. Software systems can assist in decision-making, early disease screening and diagnosis by analyzing previous medical records. These systems are faster, cheaper, and more scalable than human resources, making them essential for meeting current healthcare demands. A simple screening phase using basic questions or symptoms can significantly improve early detection.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Historically, cardiovascular disease (CVD) prediction began with statistical models like the Framingham Risk Score (FRS), which used variables such as age, cholesterol, and blood pressure to estimate risk. As machine learning evolved, supervised learning models like Decision Trees, Random Forests, and Support Vector Machines were introduced, offering improved accuracy by capturing non-linear relationships in data. Logistic regression remained a common baseline for comparison. Neural networks were later applied to uncover complex patterns in large datasets, while feature selection efforts aimed to optimize predictions using demographic, lifestyle, and clinical data. Hybrid approaches combining traditional statistical methods with machine learning further enhanced prediction reliability. Recently, ensemble methods like Gradient Boosting and Boost, along with deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have advanced CVD prediction, enabling early detection and personalized treatment by leveraging vast medical data.

II. LITERATURE SURVEY

1. 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. 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. [2].
2. 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. [3]
3. Early and accurate detection and diagnosis of heart disease using intelligent computational model. [Yar Muhammad] , MuhammadTahir1 , Maqsood Hayat1,KilTo Chong.[4]
4. Same Heart Disease Different ML Models An expert medical diagnosis system was developed using Decision Tree , Naive Bayes and Neural Network by Palaniappan and Awang.[6]
5. Different Disease and Different ML Models:According to Afifa Akhtar, Susmita Roy Tithi and Fahimul Aleem in a study, a specific model provides the most accurate results for a specific type of heart disease.[5]
6. Based on the data available, there is variation in prediction accuracy. Less Data Less Accuracy and Different Feature contribute differently. These are solved using Feature Selection Algorithms

Disease Name	Best Algorithm	Score
CAD	Naive Bayes	94%
Sinus Bradycardia	Decision Tree	95%
Sinus Tachycardia	All except KNN	95%
Myocardial Infarction	Decision Tree	96%
Right Bundle Branch Block	Logistic Regression	96%

TABLE1: BEST ALGORITHMS

III. METHODOLOGY

The project aims to develop a machine learning model that accurately identifies individuals at high risk of heart disease. The model will be trained on a relevant heart disease dataset, which provides valuable insights into patient demographics, clinical measurements, and lifestyle factors.

A. Data collection

Source: The dataset is uploaded which presumably containing clinical features related to cardiovascular health. Common features in such datasets include age, sex, cholesterol levels, blood pressure, and exercise-induced angina.

Purpose- The dataset will be used to build a predictive model for identifying the likelihood of cardiovascular disease.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

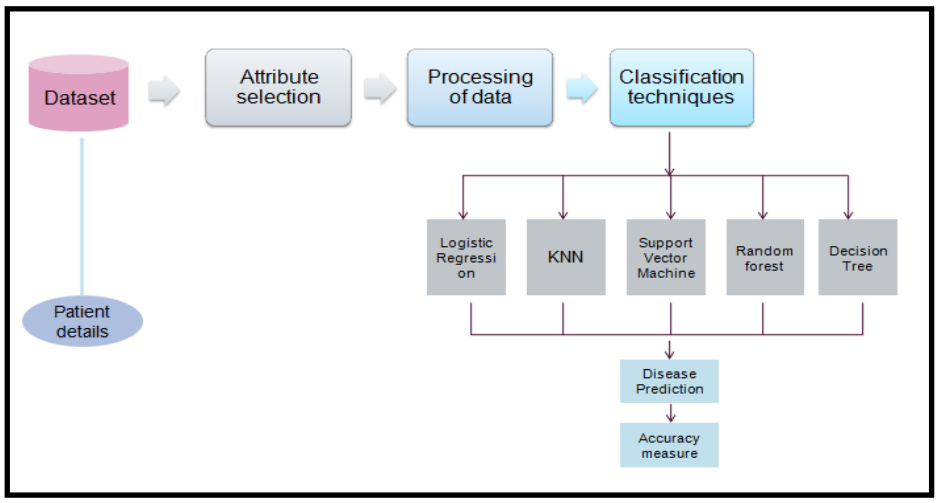


Fig.1 Process Flow diagram

B. Feature Engineering

Data Preprocessing:

1. **Load Data**: Import the dataset into a DataFrame using pandas (e.g., pd.read_csv()), ensuring the dataset is loaded correctly and checking for any inconsistencies in formatting or missing data.

2. **Exploratory Data Analysis (EDA):**Perform an initial analysis of the dataset properties to understand its structure:
   - **Check for Missing Values:** Use .info() to identify any missing values across the dataset columns. Missing data handling (imputation or removal) may be required depending on the nature and extent of the gaps.
   - **Obtain Summary Statistics:** Use .describe() to gather summary statistics for numeric columns such as mean, standard deviation, min, max, and percentiles to get a sense of the data distribution.
   - **Identify Correlations:** Use .corr() to compute pairwise correlations between numerical features. Visualize these correlations using heat maps or pair plots to understand which features are strongly related to the target variable or to each other.
3. **Target Distribution:** Plot the distribution of the target variable to assess whether the data is imbalanced (e.g., using bar plots for classification problems). Techniques like oversampling, undersampling, or using different metrics (e.g., ROC-AUC) may be needed if imbalance is present.

Feature Selection:

- **Drop Irrelevant or Redundant Features:** Identify features that have little to no impact on the target variable or those that are highly correlated with other features, and drop them to reduce dimensionality. This step can improve model performance and reduce overfitting.

- **Retain Features with High Correlation to Target:** Select features that have a strong correlation with the target variable and low correlation with each other. Feature importance methods, such as using tree-based models like Random Forest, can also help in selecting important features.

Feature Scaling:

- Standardize numerical features using StandardScaler (from scikit-learn) to have zero mean and unit variance. This is particularly crucial for models like Support Vector Machines (SVM) and K-Nearest Neighbors (KNN), which are sensitive to the scale of the input data.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

B. Model Selection and Training

Train-Test Split: Split the dataset into training and testing sets, typically with a 70-30 or 80-20 ratio. Ensure that the split is stratified if the target variable is imbalanced, to maintain the same class distribution in both sets.

Machine Learning Models:
Train multiple algorithms to identify the best performing one:

- K-Nearest Neighbors (KNN): Iterate over different values of k (the number of nearest neighbors) using cross-validation or a validation set to determine the optimal value of k that minimizes bias-variance trade-off.
- Support Vector Machine (SVM): Evaluate the performance of different kernel functions (linear, polynomial, radial basis function [RBF], and sigmoid) and adjust hyperparameters like the regularization parameter (C) to fine-tune the model.
- Decision Tree (DT): Tune the max_features, max_depth, and min_samples_split parameters to avoid overfitting and improve generalization. Visualizing the tree can also help in understanding model decisions.
- Random Forest (RF): Train the model with varying numbers of estimators (trees) and adjust parameters such as max_depth and min_samples_split to avoid overfitting. Use ensemble learning to improve accuracy and robustness.
- Logistic Regression (LR): Use as a baseline model due to its simplicity, interpretability, and efficiency. Although not always the best for complex patterns, it provides a good starting point for performance comparison.

Model Saving: Once models are trained and evaluated, use Python's pickle or joblib libraries to save the trained models for later use in predictions or deployments. This avoids retraining every time you need to make predictions, which can save significant computation time.

C. Evaluation Metrics
Metrics Used:

- **Accuracy:** Proportion of correct predictions out of the total prediction.

$$Accuracy = \frac{Accurate\ predictions}{Total\ number\ of\ predictions}$$

- **Confusion Matrix:** Provides detailed insights into true positives, true negatives, false positives, and false negatives which helps in cross validation.
- **Classification Report:** Includes Precision, Recall, and F1-Score for each class.

1. Precision is the proportion of correctly predicted positive cases (CVD presence) out of all cases predicted as positive.

$$Precision = \frac{True\ positive}{True\ positive + False\ positive}$$

2. Recall is the proportion of correctly predicted positive cases out of all actual positive cases.
3. The F1-Score is the harmonic mean of precision and recall. It provides a balanced measure of a model's performance, particularly when there is class imbalance.
4. False Positives (FP) implies Cases where the model incorrectly predicts CVD when the patient does not actually have it and False Negatives (FN) implies the Cases where the model misses a CVD diagnosis even though the patient actually has it.

- **Cross-Validation:** Use cross-validation scores to evaluate models' robustness and reduce overfitting risks. Model Comparison is done to compare the models based on the above metrics to determine the best-performing algorithm.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. IMPLEMENTATION

1. Data Collection:

- The user interacts with a form in the frontend. This form collects inputs corresponding to the model's required features (e.g., age, cholesterol level, etc.).
- Once the form is submitted, the input data is sent to the backend API for processing and prediction using HTTP requests.

2. Feature Engineering:

The backend system takes the raw input data provided by the user and processes it to ensure it matches the format and structure used during the model training phase. This ensures consistency and compatibility with the trained models. Several preprocessing steps are performed:

- Scaling/Normalization: Numerical inputs are standardized using a technique like StandardScaler to bring all features onto a uniform scale, ensuring that features with larger magnitudes don't disproportionately influence the model.
- Handling Categorical Variables: Any categorical inputs, such as gender or other non-numerical data, are encoded (e.g., using one-hot or label encoding) to make them usable by machine learning algorithms.
- After preprocessing, the cleaned data is reshaped into the required format and passed into the selected models for further analysis and prediction

3. Model Selection:

The system loads several pre-trained machine learning models, each chosen for its unique strengths:

- KNN (K-Nearest Neighbors): Chosen for its simplicity, intuitive approach, and effectiveness in handling smaller datasets, where local patterns can be crucial.
- SVM (Support Vector Machine): Used for its robustness in high-dimensional data scenarios and its ability to handle various patterns with different kernel functions (linear, polynomial, radial).
- Decision Tree: Selected for its interpretability and the ability to provide insights into feature importance, making it more explainable in medical contexts.
- Random Forest: Leveraged as an ensemble technique that combines multiple decision trees to improve accuracy, reduce overfitting, and enhance reliability.
- Logistic Regression: Employed as a baseline model for linear classification tasks, offering simplicity and speed for initial predictions.
- Model selection is based on cross-validation performance during training, ensuring that the models chosen provide a balance of accuracy, precision, recall, and interpretability for the given dataset.

4. Prediction:

- Once the input data is processed, it is fed into one or multiple models for prediction. If an ensemble approach is used, the predictions from all models are combined to provide a consensus output.
- The output includes the predicted class (e.g., CVD = 1 for cardiovascular disease or No CVD = 0 for no disease) along with the corresponding probabilities when applicable, giving users a sense of confidence in the prediction.
- On the frontend, the prediction results are displayed in a clear and concise manner within the "Result" section,
- enabling users to easily understand the output.
- User-friendly features are integrated, such as error messages for invalid or incomplete inputs and explanatory notes about the results, ensuring a seamless and intuitive user experience

© 2025 IJIRCCE | Volume 13, Issue 1, January 2025| DOI: 10.15680/IJIRCCE.2025.1301047

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

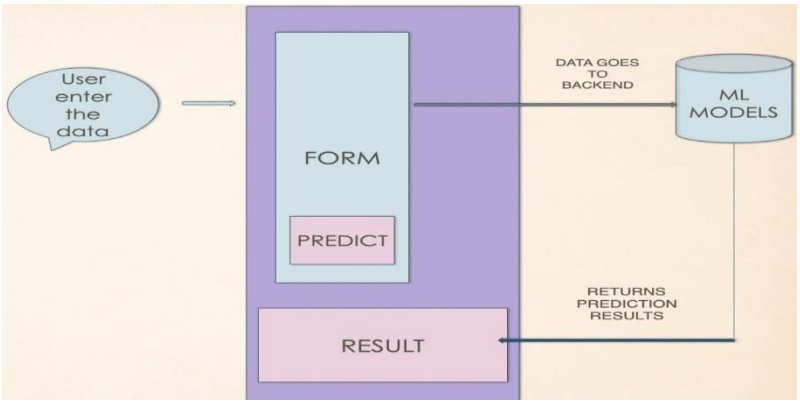


Fig2: Prediction Model

V. RESULTS AND DISCUSSION

Model	Accuracy				Precision	Recall	F1-Score	Comments
KNN	83%	82%	85%	83%	Performs well for simple relationships but sensitive to feature scaling.			
SVM	86%	85%	87%	86%	Handles non-linear boundaries effectively but can be computationally intense.			
Decision Tree	81%	80%	82%	81%	Easy to interpret but prone to overfitting with complex data.			
Random Forest	92%	91%	90%	92%	Best model overall due to ensemble technique reducing overfitting.			
Logistic Regression	84%	83%	85%	84%	Best model overall due to ensemble technique reducing overfitting.			

TABLE2: EVALUATION RESULTS

Model Performance:
Random Forest consistently outperforms others with the highest accuracy (89%) and F1-score (89%). Whereas SVM follows closely, benefiting from its ability to handle complex boundaries.

Metrics Explanation:

- **Accuracy:** Accuracy measures the overall correctness of the model by evaluating how many predictions (both true positives and true negatives) were correct out of the total predictions. In this project, Random Forest achieved the highest accuracy of approximately 92%, indicating its ability to consistently provide reliable predictions for both CVD-positive and CVD-negative cases.
- **Precision:** Precision measures the proportion of correctly predicted positive cases out of all cases predicted as positive. It helps minimize false positives, which is crucial in CVD detection to avoid incorrectly diagnosing healthy individuals as at risk. Random Forest exhibited a high precision (~91%), demonstrating its reliability in identifying true cases without overestimating risks
- **Recall:** Recall focuses on identifying all actual positive cases by measuring the proportion of true positives detected out of all actual positive cases. It helps minimize false negatives, which is critical for ensuring at-risk patients are not overlooked. The SVM and Random Forest models both achieved high recall scores (~90%),

© 2025 IJIRCCE | Volume 13, Issue 1, January 2025| DOI: 10.15680/IJIRCCE.2025.1301047

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

making them effective at capturing most CVD cases.

- **F1-Score:** The F1-score provides a balanced measure by combining precision and recall, especially useful in imbalanced datasets where one class (e.g., no-CVD cases) dominates. Random Forest achieved the highest F1-score (~92%), highlighting its ability to handle trade-offs between false positives and false negatives effectively

Use Case: Random Forest and SVM are suitable for final deployment due to their robust performance. Logistic Regression can serve as a lightweight fallback option.

Limitations and Challenges:

- Data Quality and Imbalance: The dataset may contain noisy, incomplete, or irrelevant data, which can negatively impact the accuracy and reliability of the models. Furthermore, class imbalance, where the majority of cases belong to one class (e.g., no-CVD), can cause the models to struggle in accurately identifying minority classes like actual CVD cases, leading to potential misdiagnoses.
- Model Generalizability and Complexity: Machine learning models, particularly advanced ones like Random Forest and SVM, might not generalize well to diverse populations due to variations in demographics, medical history, and health trends. Additionally, these models often act as black boxes, making their predictions difficult for medical professionals to interpret and trust, which poses a challenge for adoption in clinical settings.
- Optimization and Ethical Concerns: The process of selecting the right features and fine-tuning model hyperparameters is resource-intensive and complex, requiring careful validation to avoid overfitting or underfitting. Ethical concerns also arise, such as the risk of false positives or false negatives impacting patient safety, and the need for thorough validation before deploying these models in real-world medical applications.

A. Website Snapshot

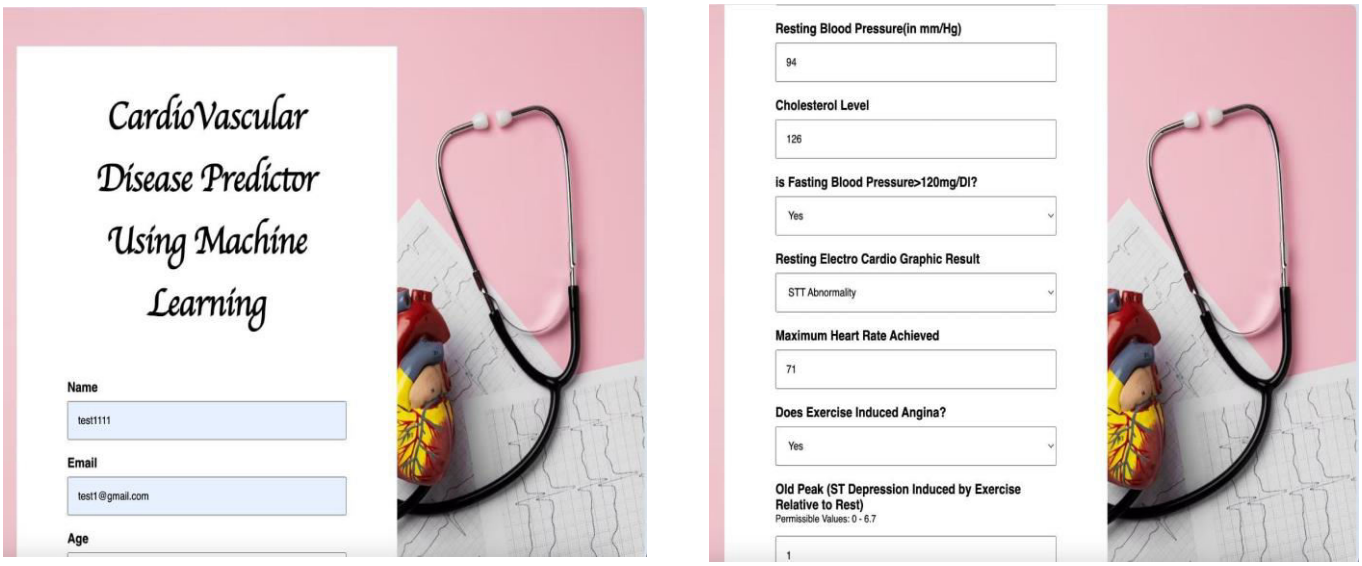


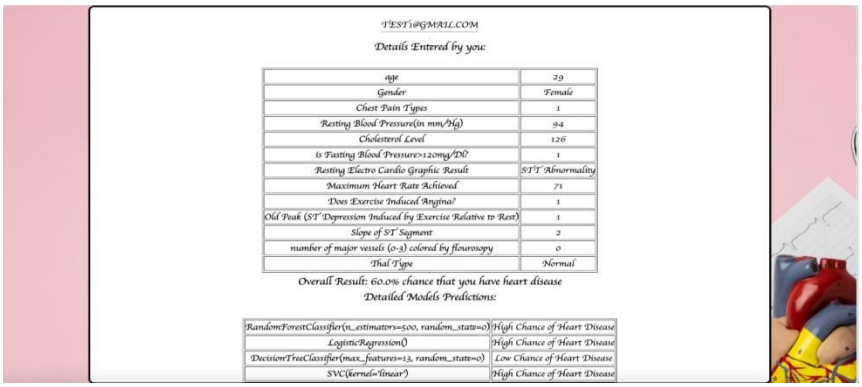
Fig2: Input Form

© 2025 IJIRCCE | Volume 13, Issue 1, January 2025| DOI: 10.15680/IJIRCCE.2025.1301047

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



TEST1@GMAIL.COM

Details Entered by you:

age	29
Gender	Female
Chest Pain Types	1
Resting Blood Pressure(in mm/Hg)	94
Cholesterol Level	126
is Fasting Blood Pressure>120mg/DL	1
Resting Electro Cardio Graphic Result	ST-T Abnormality
Maximum Heart Rate Achieved	71
Does Exercise Induced Angina?	1
Old Peak (ST Depression Induced by Exercise Relative to Rest)	1
Slope of ST Segment	2
number of major vessels (0-3) colored by flourosopy	0
Thal Type	Normal

Overall Result: 60.0% chance that you have heart disease

Detailed Models Predictions:

RandomForestClassifier(n_estimators=500, random_state=0)	High Chance of Heart Disease
LogisticRegression()	High Chance of Heart Disease
DecisionTreeClassifier(max_features=13, random_state=0)	Low Chance of Heart Disease
SVC(kernel='linear')	High Chance of Heart Disease

Fig3: Report Form

V1. CONCLUSION AND FUTURE ENHANCEMENTS

This project demonstrates the effectiveness of machine learning in predicting cardiovascular diseases. By utilizing five different machine learning models, the system aims to provide a more accurate and reliable prediction of CVD risk. The approach not only enhances diagnostic accuracy but also contributes to a more accessible and efficient healthcare system. With the ability to process large datasets and provide real-time predictions, the system offers great potential for early detection.

Future Enhancements:

- **Data Acquisition and Enrichment**: Integrate with CRM systems or customer data platforms to obtain a more comprehensive view of customer behavior and demographics. Explore external data sources like social media sentiment analysis or public web data to capture additional customer insights.
- **Model Exploration and Improvement:** Experiment with deep learning architectures like recurrent neural networks (RNNs) or convolutional neural networks (CNNs) if the data exhibits sequential or spatial patterns, respectively. Consider ensemble methods like stacking or blending to combine predictions from multiple models, potentially leading to more robust and accurate results.
- **Advanced Customer Segmentation:** Develop churn prediction models for specific customer segments based on demographics, service usage patterns, or other relevant factors. This enables more targeted customer retention strategies.
- **Explainable AI Integration:** Integrate Explainable AI (XAI) techniques to provide users with clear explanations of how the model arrives at its churn predictions. This can be particularly valuable for customer- facing applications.

REFERENCES

1. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)#:~:text=Cardiovascular%20diseases%20(CVDs)%20are%20the,%2D%20and%20middle%2Dincome%20countries
2. Bo Jin ,Chao Che, Zhen Liu, Shulong Zhang, Xiaomeng Yin,And Xiaopeng Wei, “Predicting the Risk of Heart Failure WithEHR Sequential Data Modeling” ,IEEE Access 2018.
3. Ashir Javeed, Shijie Zhou, Liao Yongjian, Iqbal Qasim,Adeeb Noor, Redhwan Nour4, Samad Wali And Abdul Basit ,“An Intelligent Learning System based on Random SearchAlgorithm and Optimized Random Forest Model forImproved Heart Disease Detection” , IEEE Access 2017.
4. Yar Muhammad, Muhammad Tahir, Maqsood Hayat & Kil To Chong Early and accurate detection and diagnosis of heart disease using intelligent computational model.[https://www.nature.com/articles/s41598-020-76635-9]
5. Sushmita Roy Tithi , Afifa Aktar , Fahimul Aleem , Amitabha Chakrabarty ‘’ECG data analysis and heart disease prediction using machine learning algorithms’’. Proceedings of 2019 IEEE Region 10 Symposium.
6. Sellappan Palaniappan , Rafiah Awang “Intelligent Heart Disease Prediction System Using Data Mining Techniques ” IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.8, August 2008.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details