

HEART DISEASE RISK ASSESSMENT USING ML

¹ B.Santhosh, ² Divya, ³ Singireddy Pavani, ⁴ Singarapu Srija, ⁵ Rajurapu Naveen

¹Assistant Professor, ^{2,3,4,5}Students

Department of ECE

Siddhartha Institute of Technology & Sciences, Narapally

bhushaboinasanthosh.ece@siddhartha.co.in, 25tq5a0405@siddhartha.org.in,

24tq1a0460@siddhartha.co.in, 24tq1a0459@siddhartha.co.in, 24tq1a0458@siddhartha.co.in

ABSTRACT

Cardiovascular disease is the leading cause of death worldwide, and coronary heart disease accounts for a large share of these deaths. Many heart attacks occur in people who were unaware that they were at risk, even though warning signs such as high blood pressure, high cholesterol, abnormal ECG patterns, and reduced exercise tolerance were present beforehand. This paper presents a heart disease risk assessment system that uses machine learning to estimate a person's likelihood of heart disease from routinely measured clinical attributes and signals, and presents the result in a form that patients and doctors can understand.

The system uses clinical attributes widely recorded in cardiology practice, including age, sex, type of chest pain, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiogram result, maximum heart rate achieved during exercise, exercise-induced angina, ST depression, slope of the peak exercise ST segment, number of major vessels seen on fluoroscopy, and thalassemia test result. From an electronics perspective, several of these inputs come directly from biomedical instruments, and the system can also accept heart rate readings from a pulse sensor and basic parameters from a single-lead ECG module for continuous monitoring.

The data is cleaned and preprocessed by handling missing and inconsistent values, one-hot encoding categorical attributes, and standardising numerical measurements. Several classifiers are trained and compared, including logistic regression, k-nearest neighbours, a support vector machine, naive Bayes, a decision tree, random forest, gradient boosting, and extreme gradient boosting. Hyperparameters are tuned with stratified cross-validation, and the models are evaluated using accuracy, precision, recall, F1 score, and the area under the ROC curve, with recall given particular weight.

The best model assigns each person a probability of heart disease and a low, medium, or high risk category. Shapley value explanations show which attributes raised or lowered the risk for that individual, for example a high ST depression or a low maximum heart rate. A web dashboard allows a clinician or health worker to enter patient data, view the risk and its explanation, compare repeated assessments over time, and read general guidance on lifestyle factors such as smoking, diet, and physical activity.

The system is intended to support clinical decision making and preventive care, not to replace diagnosis by a cardiologist. It can help primary care doctors and health camps decide which patients need further cardiac investigation. Future work includes integration with wearable ECG devices for continuous risk tracking, validation on

larger and more diverse patient populations, and the addition of signal-level features extracted from full ECG recordings.

I. INTRODUCTION

Heart disease develops over many years through the gradual build-up of risk factors such as hypertension, high cholesterol, diabetes, smoking, obesity, and physical inactivity. In India, cardiovascular disease tends to appear at younger ages than in many other countries, and a considerable number of patients experience their first heart attack before the age of fifty. Because the condition often progresses silently, many people seek medical attention only after a serious cardiac event.

The clinical assessment of heart disease risk involves several tests. Blood pressure and cholesterol are measured routinely, an electrocardiogram records the electrical activity of the heart, and an exercise stress test observes how the heart responds to physical effort. Interpreting all of these results together requires experience, and in busy outpatient clinics or rural health centres there may be little time or specialist expertise available. Risk scores based on a few factors are used in some settings, but they do not make full use of the information collected.

Machine learning can combine many clinical attributes to estimate risk more precisely. Datasets collected from cardiology departments record patient measurements together with the eventual diagnosis confirmed by angiography, and models trained on these records can learn the patterns that separate patients with and without significant coronary disease. For electronics and communication engineers, the problem has a natural connection to biomedical instrumentation, since attributes such as ECG findings, heart rate, and blood pressure are obtained from electronic sensors and signal processing circuits.

This paper describes a heart disease risk assessment system that trains and compares several machine learning classifiers on clinical data, selects the most suitable model with an emphasis on recall, and explains each prediction in terms of the patient's own measurements. The system also accepts readings from a pulse sensor and an ECG module so that basic parameters can be updated without manual entry. A dashboard presents the results to clinicians. The aim is to help identify people at risk earlier so that preventive treatment and lifestyle changes can begin in time.

II. LITERATURE SURVEY

The assessment of cardiovascular risk has a long history in epidemiology. Large long-term population studies identified age, sex, blood pressure, cholesterol, diabetes, and smoking as major risk factors, and these findings were converted into point-based risk scores that estimate the probability of a cardiac event over ten years. Such scores remain widely used in clinical guidelines because they are simple and have been validated in several populations. However, studies applying them to South Asian populations have reported that they tend to underestimate risk, since these populations develop heart disease at younger ages and at lower levels of conventional risk factors. Researchers have also noted that the scores use only a small number of variables and

assume mostly linear, additive effects. These limitations have encouraged the investigation of data-driven methods that can incorporate more clinical information, capture interactions between factors, and be recalibrated for the population in which they are used.

A widely studied benchmark for machine learning in cardiology is a collection of patient records gathered from several hospitals, which records fourteen commonly used attributes including chest pain type, resting blood pressure, cholesterol, resting ECG result, maximum heart rate, exercise-induced angina, ST depression, and fluoroscopy results, along with angiographic diagnosis. A large number of studies have applied logistic regression, naive Bayes, decision trees, k-nearest neighbours, support vector machines, and artificial neural networks to this data, typically reporting accuracy between seventy-five and ninety percent. Researchers have observed that attributes derived from the exercise test and fluoroscopy, such as ST depression, number of major vessels, and thalassemia result, are among the strongest predictors. They have also pointed out that the dataset is relatively small and contains some missing values, so results can vary considerably with the choice of preprocessing, validation strategy, and random split.

Ensemble learning and feature selection have been shown to improve heart disease prediction in many studies. Random forest, gradient boosting, and extreme gradient boosting have frequently achieved the highest accuracy and area under the ROC curve on the benchmark data, largely because they model nonlinear relationships and interactions between attributes. Feature selection methods such as chi-square testing, recursive feature elimination, and importance-based ranking have been used to remove redundant attributes, which often improves the performance of simpler models and makes the results easier to interpret. Hybrid approaches that combine several classifiers through voting or stacking have also been reported to give small but consistent gains. Researchers have stressed that model evaluation should use stratified cross-validation and should report recall and F1 score along with accuracy, since failing to identify a patient with heart disease is far more serious than a false positive that leads to further tests.

The electrocardiogram has been studied extensively as a source of information for automated cardiac analysis. Signal processing research has developed methods for removing baseline wander, power-line interference, and muscle noise from ECG recordings using filters, and for detecting the QRS complex and other waves using derivative-based and wavelet-based algorithms. Features such as heart rate, heart rate variability, QRS duration, and ST segment deviation have been extracted and used for arrhythmia detection and ischaemia assessment. More recent studies have used low-cost single-lead ECG modules and optical pulse sensors connected to microcontrollers, showing that useful measurements can be obtained outside hospital settings. Authors have suggested that combining such sensor-derived parameters with clinical risk factors could support continuous monitoring and earlier warning, although they caution that consumer-grade sensors are more susceptible to motion artefacts and require careful signal quality checks.

Interpretability has become a central concern in the application of machine learning to heart disease. Clinicians are unlikely to act on a prediction they cannot understand, and regulatory guidance increasingly requires that medical decision-support tools be transparent. Studies have applied global feature importance, partial dependence plots, and local explanation methods such as Shapley values to heart disease models, showing which attributes drive predictions for the population as a whole and for individual patients. These explanations have generally confirmed that models rely on clinically meaningful factors, which increases confidence in their use. Other work has described web-based and mobile applications that deliver risk predictions to doctors and patients along with explanations and lifestyle advice. The system proposed in this paper follows this direction by combining ensemble classification on clinical attributes, optional sensor inputs, and patient-level explanations in a dashboard designed for clinical use.

III. SYSTEM ANALYSIS

The heart disease risk assessment system receives clinical attributes for a patient and returns an estimate of the likelihood of heart disease along with an explanation. The inputs include age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG result, maximum heart rate, exercise-induced angina, ST depression, slope of the ST segment, number of major vessels, and thalassemia result. Resting heart rate can be captured from an optical pulse sensor, and basic ECG parameters can be obtained from a single-lead ECG module after filtering and R-peak detection. The inputs are validated against realistic ranges, missing values are imputed, categorical attributes are one-hot encoded, and numerical values are standardised. A trained classifier produces a probability of heart disease, which is mapped to low, medium, or high risk using thresholds chosen to keep recall high. Shapley values identify the attributes that most influenced the result, and the assessment is stored so that changes over repeated visits can be tracked on the dashboard.

Existing System

In current practice, the assessment of heart disease risk depends largely on the judgement of the treating doctor, who reviews blood pressure, cholesterol, ECG, and other results during a consultation. In busy outpatient departments and rural health centres, consultations are short, and specialist cardiologists are often not available, so subtle combinations of risk factors may not be recognised. Where formal risk scores are used, they rely on a small number of variables and were developed for populations different from the Indian population, which can lead to underestimation of risk. ECG recordings are interpreted visually, and low-cost monitoring devices used at home or in camps are rarely connected to any analysis system. Patient records are often kept on paper or in separate hospital systems, so trends in blood pressure, cholesterol, and heart rate over time are difficult to follow. Many patients therefore receive a diagnosis of heart disease only after experiencing chest pain or a cardiac event, when earlier intervention might have reduced damage.

Disadvantages of Existing System

- Risk assessment depends heavily on individual judgement during short consultations.
- Specialist cardiologists are often not available in rural and primary care settings.
- Standard risk scores use few variables and may underestimate risk in Indian patients.
- Interactions between multiple risk factors are not captured systematically.
- ECG recordings are interpreted visually without automated support.
- Readings from low-cost monitoring devices are not connected to any analysis.
- Patient measurements are not tracked over time in a single place.
- Patients receive little explanation of which factors contribute to their risk.
- Many patients are identified only after a cardiac event has occurred.
- Health camps lack a quick tool for deciding who needs further investigation.

Proposed System

The proposed system is a machine learning based heart disease risk assessment platform that combines clinical attributes with optional sensor readings. Patient records are cleaned, missing values in attributes such as number of major vessels and thalassemia result are imputed, categorical attributes are one-hot encoded, and numerical measurements are standardised. Logistic regression, k-nearest neighbours, a support vector machine, naive Bayes, a decision tree, random forest, gradient boosting, and extreme gradient boosting are trained and tuned with stratified five-fold cross-validation. Models are compared using accuracy, precision, recall, F1 score, and ROC-AUC, and the model with the best balance of recall and overall performance is selected. A pulse sensor and a single-lead ECG module connected to a microcontroller can supply heart rate and basic ECG parameters after filtering. The output is a probability and a risk category, supported by Shapley value explanations of the patient's main contributing factors. A web dashboard records assessments, shows trends over visits, and provides general lifestyle guidance.

Advantages of Proposed System

- Combines many clinical attributes into a single consistent risk estimate.
- Ensemble models capture interactions between risk factors more effectively.
- Model selection emphasises recall so that fewer at-risk patients are missed.
- Explains each prediction in terms of the patient's own measurements.
- Accepts heart rate and ECG parameters from low-cost sensor modules.
- Tracks repeated assessments so that changes in risk can be observed.
- Helps primary care doctors decide who needs specialist investigation.
- Supports quick screening during health camps and outreach programmes.
- Can be recalibrated with local patient data to suit the population served.
- Built with open-source Python tools and runs on ordinary hardware.

IV. METHODOLOGY

The methodology begins with the collection of a heart disease dataset containing clinical attributes and confirmed diagnoses from several hospital sources, which are

merged and harmonised. Duplicate records are removed, invalid values such as zero cholesterol or zero blood pressure are treated as missing, and missing values are imputed using the median for numerical attributes and the most frequent value for categorical attributes, computed on training data only. Chest pain type, resting ECG, slope, and thalassemia are one-hot encoded, and continuous attributes are standardised. Exploratory analysis with distribution plots, box plots, and correlation heatmaps examines the relationship of each attribute with the diagnosis. The data is split with stratification into training and test sets, and eight classifiers are tuned using grid search with five-fold cross-validation. The best model is evaluated on the test set with accuracy, precision, recall, F1 score, ROC-AUC, and a confusion matrix, and its decision threshold is adjusted to favour recall. For the sensor path, ECG signals are band-pass filtered and R-peaks are detected to compute heart rate before entry into the model.

SYSTEM ARCHITECTURE

The architecture of the heart disease risk assessment system is organised into layers that lead from patient data and sensor readings to explained risk estimates. The Data Input Layer accepts clinical attributes through a web form and receives measurements from the sensor module. The Sensor Interface Layer connects the pulse sensor and single-lead ECG module to a microcontroller and transmits readings to the server. The Signal Conditioning Layer filters ECG signals, removes noise and baseline wander, and detects R-peaks to compute heart rate. The Preprocessing Layer validates ranges, imputes missing values, encodes categorical attributes, and scales numerical ones. The Model Layer holds the trained classifiers and the selected production model. The Risk Scoring Layer converts the predicted probability into a risk category. The Explanation Layer computes Shapley values for the patient. The Storage Layer keeps assessment history, and the Dashboard Layer presents risk, explanations, trends, and guidance to clinicians and patients.

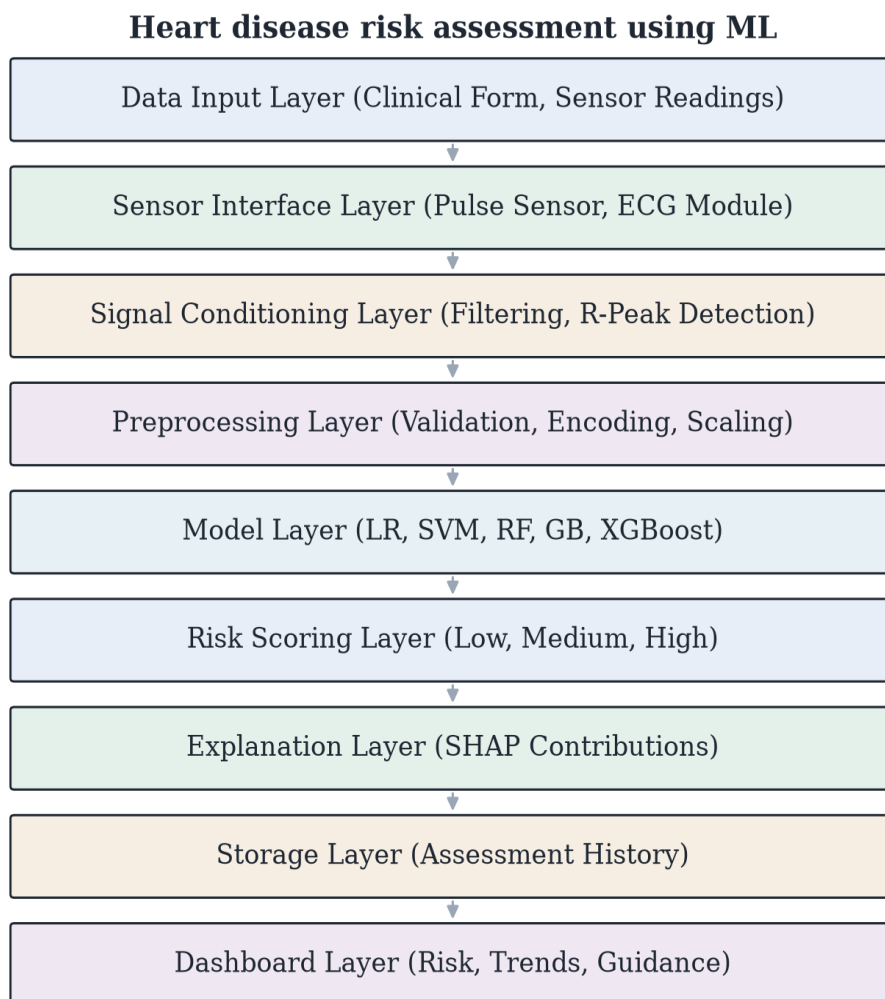


Fig. 1 System Architecture

V. RESULT AND OUTPUT



VI. CONCLUSION

This paper has presented a heart disease risk assessment system that uses machine learning to estimate the likelihood of heart disease from clinical attributes commonly recorded in cardiology practice. By combining many measurements into a single consistent estimate, the system offers more complete support than simple risk scores and helps identify patients who need further investigation before they experience a serious cardiac event.

Among the classifiers compared, ensemble methods such as random forest, gradient boosting, and extreme gradient boosting provided the strongest performance, with high recall and area under the ROC curve. Careful preprocessing, including treatment of invalid zero values, imputation using training data only, and appropriate encoding of categorical attributes, contributed significantly to reliable results. Adjusting the decision threshold in favour of recall reflects the clinical priority of not missing patients with heart disease.

The explanation component makes the system suitable for clinical use by showing which of a patient's measurements raise or lower the predicted risk. This allows doctors to check the reasoning of the model and helps patients understand which factors they can influence through treatment and lifestyle changes. The optional sensor path, which obtains heart rate and basic ECG parameters from low-cost modules after filtering and peak detection, connects the prediction system with biomedical instrumentation and supports repeated measurement outside the hospital. The final diagnosis always remains with the treating physician.

Future work can extend the system with wearable ECG devices for continuous monitoring, allowing risk to be updated automatically as new readings arrive. Signal-level features such as heart rate variability and ST segment changes extracted from full ECG recordings could be added to the model. Validation on larger and more diverse patient populations, particularly Indian cohorts, and integration with hospital information systems would be needed before wider clinical deployment.

REFERENCES

- [1] Kumar, R. D., Prudhviraaj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In *Handbook of Artificial Intelligence and Wearables* (pp. 145-158). CRC Press.
- [2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In *The International Conference on Artificial Intelligence and Smart Environment* (pp. 557-564). Cham: Springer Nature Switzerland.
- [3] Sv satyakrishna, shirisha rang, bhargavi nalacheruve. (2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm
- [4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.
- [5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in *Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT)*, Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.
- [6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks," in *Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment*, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.
- [7] R. D. Kumar, V. N. S. Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in *Blockchain for Smart Cities*, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.
- [8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.
- [9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," *Journal of Machine and Computing*, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.
- [10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in *Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE)*, Apr. 2025.

[11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.