

Cardiovascular Disease Prediction: An Ensemble Machine Learning Approach

¹Kunal Patil,²Nimesh Patil,³Pratik Sirsath,⁴Sanjay Sonkawade, ⁵Anand Ingle

¹²³⁴ B.E Student, MGM college of Engineering and Technology

⁵ Professor, MGM college of Engineering and Technology

Abstract— cardiovascular disease (CVD) is one of the major causes of morbidity and mortality worldwide, posing burden on healthcare systems. Accurate and early prediction of cardiovascular disease is critical for enabling preventive measures with the increasing availability of medical data, machine learning techniques have emerged as effective tools for disease prediction and risk assessment. This paper presents an ensemble machine learning approach for predicting cardiovascular disease using a combination of clinical, demographic, and lifestyle-related features such as age, gender, blood pressure, cholesterol levels, body mass index, and smoking status. There are some Several individual classification models, such as Logistic Regression, Decision Tree, Support Vector Machine, and Random Forest, are developed and evaluated. To improve prediction performance and reduce model variance, ensemble techniques such as majority voting and boosting are employed to integrate the outputs of multiple base learners. The proposed ensemble model is assessed using standard evaluation metrics including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). Experimental results shows that the ensemble-based model achieves superior performance compared to another classifiers, offering better robustness and generalization capability. The proposed approach can assist healthcare professionals in early diagnosis, risk stratification, and decision support, ultimately contributing to the reduction of cardiovascular disease-related complications.

1 INTRODUCTION

Cardiovascular disease (CVD) is a disease which comes from group of disorders affecting the heart and blood vessels and remains one of the major causes of death worldwide. According to global statistics, millions of people die each year due to heart-related problems such as coronary artery disease, heart failure, stroke, and hypertension. The increasing prevalence of uneven lifestyles, unhealthy dietary habits, smoking, obesity, diabetes, and stress has further contributed to the faster

rise in cardiovascular diseases. So the Early detection and accurate prediction of CVD play a vital role in reducing mortality rates and improving patient outcomes through timely medical and lifestyle modifications.

Traditional methods for diagnosing cardiovascular disease mostly depends on clinical expertise, medical examinations, and laboratory tests. While these methods are effective, But they can be time-consuming, costly, and subject to human mistakes. This rising volume of healthcare data generated from electronic health records, wearable devices, and diagnostic systems has made it increasingly difficult for healthcare professionals to manually analyse and interpret data efficiently. This has created a strong demand for automated and intelligent systems capable of assisting doctors in disease prediction and decision-making about the disease.

Machine learning (ML) has emerged as a powerful tool in the healthcare domain due to its ability to analyse complex datasets, identify hidden patterns, and make accurate and quicker predictions. Machine learning algorithms such as Logistic Regression, Decision Trees, Support Vector Machines (SVM), k-Nearest Neighbours (k-NN), and Random Forest have been widely applied for cardiovascular disease prediction. These models can process number of risk factors simultaneously and provide predictive insights that support early diagnosis. However, the performance of individual models often depends on the behaviour of the dataset, feature distribution, and algorithm-specific assumptions, which may lead to limited accuracy or generalization issues.

2 RELATED WORKS

Cardiovascular disease prediction has been an active research space in the healthcare and machine learning. Several studies have explored in detail a wide range of machine learning models to get better the accuracy and efficiency of heart disease diagnosis. Recent researches are primarily focused on old statistical models and single classifiers to identify risk factors associated with cardiovascular Situation. i.e, Logistic Regression and Decision Tree models have been widely used because of their interpretability and computational efficiency. However, these approaches often shows lower performance when faced with complex, nonlinear relationships present in high-dimensional clinical data.

Support Vector Machines (SVM) and k-Nearest Neighbours (k-NN) have also been applied to get the presence of cardiovascular disease. SVM, in particular, has shown robustness with high-dimensional feature area by maximizing the margin between classes. Detrital. applied SVM to heart disease datasets and reported satisfactory performance, although sensitivity to parameter selection limited generalizability. Similarly, k-NN demonstrated reliable accuracy in smaller datasets but suffered from increased computational overhead as the number of objects grew.

With improvement in ensemble learning, researchers have increasingly leveraged techniques that get multiple models together to enhance predictive performance. Random Forest, an ensemble of decision

trees built through bootstrap aggregation (bagging), has regularly outperformed its constituent-based learners by reducing overfitting and get better generalization. Studies such as those by Parvin et al. demonstrated that Random Forest models sent superior classification accuracy for CVD prediction when compared to single-tree approaches.

Boosting methods, such as AdaBoost and Gradient Boosting Machines (GBM), have also been found for heart disease prediction. Boosting trains one after one to weak learners, typically decision stumps, by focusing on previously misclassified objects. Research by Subasi et al. shows that boosting-based models get higher precision and sensitivity than traditional models, especially on imbalanced cardiovascular datasets.

In addition to classical ensemble techniques, hybrid and stacking models have been proposed to further improve performance. For instance, several studies integrated feature selection techniques with ensemble classifiers to eliminate irrelevant attributes, thereby enhancing model accuracy and reducing complexity. Singh et al. introduced a hybrid model combining feature ranking with voting-based ensembles, achieving competitive results on benchmark heart disease datasets. Similarly, stacking-based frameworks that integrate multiple heterogeneous base learners have been reported to get better prediction accuracy by learning meta-level patterns that separate classification.

3 LITERATURE SURVEY

TITLE	AUTHORS	METHODOLOGY	RESULTS
Involving machine learning techniques in heart disease diagnosis: a performance analysis	B. S. Shukur and M. M. Mijwil	Machine learning techniques including logistic regression, random forest, artificial neural network, support vector machines, and k-nearest neighbors are applied to diagnose heart disease using the Cleveland Clinic dataset for performance comparison.	Support vector machines demonstrate the highest diagnostic accuracy of 96%, highlighting the significant role of machine learning in assisting healthcare professionals in heart disease diagnosis and improving decision-making.
Heart Disease Prediction Using Novel Quine McCluskey Binary Classifier (QMBC)	R. Kapila, T. Rangunathan, S. Saleti, T. J. Lakshmi, and M. W. Ahmad	The QMBC model combines seven machine learning models (logistic regression, decision tree, random forest, K-nearest neighbor, naive Bayes, support vector machine, multilayer perceptron) with feature selection and Principal Component Analysis.	The QMBC model outperforms existing methods in heart disease prediction, offering superior accuracy by leveraging an ensemble of models and feature extraction techniques for efficient and reliable predictions
Comprehensive analysis of supervised algorithms for coronary artery heart disease detection	S. Dhanka, V. K. Bhardwaj, and S. Maini	Logistic Regression and XGBoost models are applied to the Statlog heart disease dataset. Hyperparameters are optimized using Random SearchCV. Performance is compared between non-optimized and optimized models.	Optimized Logistic Regression and XGBoost models significantly improve CAHD detection accuracy, with XGBoost achieving the highest performance, demonstrating their potential in early diagnosis and risk assessment for coronary artery heart disease
Risk assessment of coronary heart disease based on cloud-random forest	J. Wang, C. Rao, M. Goh, and X. Xiao	The C-RF model combines a cloud model and random forest by weighting evaluation attributes using a cloud-based algorithm, constructing new CART-based decision trees, and evaluating performance on the Framingham dataset.	The C-RF model enhances CHD risk assessment with superior classification accuracy, reduced error rates, and higher AUC compared to CART, SVM, CNN, and RF, showcasing improved prediction performance.

<p>Heart Disease Prediction Using Stacking Model With Balancing Techniques and Dimensionality Reduction</p>	<p>A. Noor, N. Javaid, N. Alrajeh, B. Mansoor, A. Khaqan, and S. H. Bouk</p>	<p>PaRSEL uses a stacking model combining PAC, RC, SGDC, and XGBoost at the base layer with LogitBoost at the meta layer, incorporating dimensionality reduction and balancing techniques</p>	<p>PaRSEL achieves superior accuracy, precision, and AUC-ROC, effectively addressing imbalanced and high-dimensional data for heart disease prediction while providing interpretability using SHAP to analyze feature influence.</p>
--	--	---	--

3 SYSTEM ANALYSIS AND DESIGN

This chapter determines the complete **system analysis and design** of the derived cardiovascular disease prediction system. It includes the system workflow, functional requirements, non-functional requirements, system architecture, data flow, and design modules. The main goal of this system is to **predict the risk of cardiovascular disease** using **ensemble machine learning models** to get better prediction accuracy and reliability.

3.1 Problem Definition

Cardiovascular disease is one of the major causes of death worldwide. Many patients never show symptoms at an early stage, which makes early diagnosis difficult. Old diagnosis requires medical tests, expert consultation, and duration.

Hence, this system is designed to:

- Collect patient health credentials (biometric data)
 - Process and clean the data
 - Apply ensemble machine learning algorithms
- Predict whether a patient has a risk of cardiovascular disease

3.3 Objectives of the System

The main instances are:

- To develop a system that predicts cardiovascular disease risk using machine learning.
- To improve accuracy using an **ensemble learning approach** instead of a single model.
- To provide quick and reliable prediction outcomes.
- To assist doctors and patients in early detection and prevention.
- To create a user-friendly system for prediction and reporting.

3.4 Biometric modalities.

Biometric modalities are the measurable patient health parameters used for prediction.

Common input features are used:

- Age
- Gender
- Blood Pressure (Systolic/Diastolic)
- Cholesterol level
- Blood glucose level
- BMI (Body Mass Index)
- Heart rate
- Smoking habit
- Physical activity
- ECG results

3.5 Non-Functional Requirements

The non-functional requirements of the CVD (cardiovascular disease) prediction system show the quality, performance, and operational constraints of the ensemble machine learning model. These needs ensure that the system is accurate, reliable, efficient, secure, and suitable for real-world healthcare applications.

The system must provide high performance by generating cardiovascular disease predictions within a short response time. Even when processing large patient datasets, the ensemble learning approach should maintain stable performance without significant delays. Efficient computation is essential to support timely clinical decision-making.

Accuracy and reliability are critical non-functional requirements of the system. The ensemble machine learning model should deliver in the form of consistency accurate predictions and outperform individual machine learning models. The system must be evaluated using standard performance metrics such as accuracy,

precision, recall, F1-score, and ROC-AUC to ensure dependable results.

Scalability is an important need, as the system should be capable of handling an increasing number of patient records without degradation in performance. The architecture should support and help the addition of new machine learning models in the ensemble and allow future enhancements such as real-time prediction abilities.

Security is essential due to the sensitive nature of medical data. The system must make sure the secure storage, transmission, and processing of patient information. Proper authentication, verification and authorization mechanisms should be deployed, and encryption techniques should be used as security to protect confidential health records.

Usability is another main requirement, as the system should be easy to use for healthcare professionals. The interface should be simple, clean and intuitive, allowing doctors to input patient data and interpret prediction results without requiring technical expertise. The predicted CVD risk should be clearly represented to support clinical decision-making.

The system should ensure high availability and minimal downtime so that predictions can be accessed whenever required. Even if one model in the ensemble fails somewhere, the system should continue to function correctly using the remaining models. Proper error handling mechanisms should be implemented to maintain system stability.

Maintainability is required to ensure long-term usability of the system. The ensemble framework should allow easy updating, retraining, or replacement of machine learning models. A modular design and proper documentation will help developers maintain and improve the system efficiently.

Interpretability is important in healthcare applications, as clinicians (Doctors) need to understand the factors influencing predictions. The system should provide insights into feature importance, helping doctors identify key risk factors contributing to cardiovascular disease. This improves trust and transparency in model predictions.

Finally, the system must go with the healthcare regulations and ethical standards. Patient data should be handled responsibly, by ensuring privacy and fairness. The ensemble machine learning model should minimize bias and support ethical decision-making in cardiovascular disease prediction.

3.6 Feature Extraction

Applying human visual property in the recognition of faces, people can identify face from very far distance, even the details are vague. It means the symmetry characteristic is enough to be recognized. Human face is made up of eyes, nose, mouth and chin etc. There are some differences in shape, size and structure of those organs so the faces are differed in multiple ways and we can describe them with the structure of the organs so as to recognize them. One common method is to extract the shape of the eyes, nose, mouth and chin and then distinguish the faces by distance and scale of those organs.

3.7 Face Recognition

Feature extraction is a crucial step in the cardiovascular disease (CVD) prediction system, as it transforms raw medical data into meaningful features (like data cleaning) that can be effectively used by ensemble machine learning models. Proper feature extraction improves model accuracy, reduces noise, and enhances the overall performance of the prediction system.

In the proposed system, features are extracted from patient clinical and demographic data such as age, gender, blood pressure, cholesterol levels, blood glucose, body mass index (BMI), smoking status, physical activity, and family history of heart disease. These parameters represent important risk factors associated with cardiovascular situations and provide a strong structure for prediction.

Before feature extraction, data pre-processing techniques such as handling lost values, normalization, and encoding of categorical variables are applied. Continuous attributes like blood pressure and cholesterol are scaled to a standard range, while categorical features such as gender and smoking habits are converted into numerical form using proper encoding methods. This ensures compatibility with machine learning algorithms used in the ensemble learning.

Derived features are also created to improve predictive power. For example, ratios or combined indicators such as cholesterol-to-HDL ratio, BMI categories, and age-risk groups are generated from existing data. These derived features help the ensemble model capture complex relationships among health parameters that may not be evident from raw features alone.

Feature selection techniques are applied after extraction to get the most relevant features and remove redundant or irrelevant attributes. Methods such as correlation analysis, feature importance from tree-based models, and statistical tests are used to retain only those features that significantly contribute to CVD

prediction. This reduces model complexity and improves efficiency.

The extracted features are then provided as input to the ensemble machine learning models, such as Random Forest, Gradient Boosting, or Voting Classifiers. Effective feature extraction makes sure that the ensemble model learns meaningful patterns from patient data, leading to accurate and reliable cardiovascular disease risk prediction.

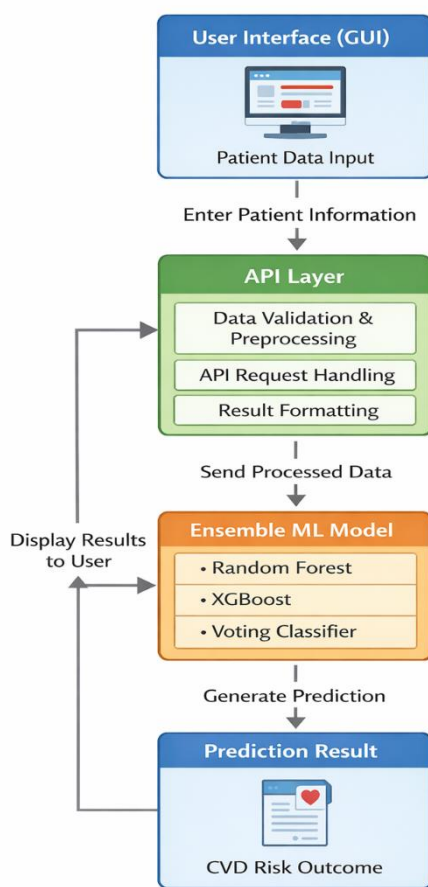
cholesterol level, blood glucose, and other related medical parameters. The GUI provides a simple, easy and interactive form for data entry and ensures that seem easy for use.

Once the patient data is submitted, it is forwarded to the **API Layer**. The API acts as a bridge between the GUI and the machine learning models. At this step, the API performs **data validation** to check for missing or incorrect values and ensures that the collected data is in the correct format. It also handles **data preprocessing**, such as visualization, sampling normalization and encoding, to make the data suitable for the prediction models.

After preprocessing, the API forwards the processed data to the **Ensemble Machine Learning Model**. The ensemble consists of multiple algorithms such as Random Forest, XGBoost, and a Voting Classifier. Each model separately analyzes the input features and generates its own prediction. The ensemble mechanism combines these separate predictions to get the more accurate and reliable CVD risk assessment.

The ensemble model then sends the final prediction back to the **API Layer**, where the result is formatted into a readable and structured response. This may include the predicted CVD risk level (low, medium, or high) and confidence information.

API and GUI Architecture for CVD Prediction System



The flowchart represents the complete working process of the cardiovascular disease (CVD) prediction system using an ensemble machine learning approach. It shows how user input is processed through the API and machine learning models to generate the last prediction.

The process begins at the **Graphic User Interface (GUI)**, where the healthcare professional or user enters patient details such as age, blood pressure,

5 PERFORMANCE EVALUATION

Existing Security includes SSL Certification, User Ids & Passwords, One Time Passwords (OTPs) to customer’s mobile. But still there are various security aspects and threats. So, enhancing the existing security is a must. This technology evolution is inevitable. Now there are Laptops and mobiles have biometric verification inbuilt, for Logon with Finger Print Sensing. This same verification method can be integrated to Banking applications as well. Utilizing biometrics for internet banking will be considerably more accurate than current methods of Verification Pins and passwords. This Biometric can be an additional authentication thus enhancing the existing security.

6 CONCLUSIONS

The cardiovascular disease (CVD) prediction system using an ensemble machine learning approach provides an effective and reliable solution for early risk assessment. By combining multiple machine learning models, the system improves prediction accuracy and reduces the limitations of individual algorithms. The use

of ensemble techniques ensures robust and consistent performance across different patient datasets.

The integration of feature extraction, preprocessing, and model selection enables the system to identify important clinical risk factors such as age, blood pressure, cholesterol levels, and lifestyle attributes. These features play a crucial role in generating meaningful predictions and enhancing the overall effectiveness of the model.

The development of the API and GUI ensures seamless interaction between users and the machine learning backend. The API enables secure and efficient data processing, while the GUI provides a simple and user-friendly interface for healthcare professionals. This design makes the system practical for real-world clinical environments.

Transactions on Electronic Circuits and Systems

8. Quinghan Xiao-(2007)'Spoofing Techniques' IEEE Computational Intelligence.
9. S. Shilaskar and A. Ghatol, "Feature selection for medical diagnosis," *International Journal of Computer Applications*, vol. 1, no. 4, pp. 1-6. 24, No. 3.
10. World Health Organization (WHO), *cardiovascular diseases (CVDs)- Fact sheet*, WHO Press, Geneva
11. R. S. Deo. "Machine Learning in medicine" *Circulation*, vol. 132, no.20, pp. 1920-1930

REFERENCES

1. K. J. Roth et al., "Global burden of cardiovascular diseases and risk factors," *Journal of the American College of Cardiology*, vol. 76, no. 25, pp. 2982-3021.
2. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York. Gunajit Sarma and Pranav Kumar-(2002) 'Biometric Authentication'-International Journal of Pure Applied Sciences and Technology-ISSN 2229-6107, pp 67-68.
3. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann. John Trader-(2014) 'Impact of Biometrics in Banking'-IEEE Transaction-Vol 54.
4. L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32.
5. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794.
6. Marius Tico and Pauli Kuosmanen (2003) 'Novel Fingerprint Representation, -IEEE Transactions on Pattern Analysis and Machine Intelligence', Vol 25, No.8.
7. Meraoumia and Bouridane-(2013) - 'Multimodal Biometrics System'-IEEE