

Explainable Hybrid CNN SVM Framework for Cardiovascular Disease Risk Prediction Using Clinical Data

Yogendra Narayan Prajapati

ynp1581@gmail.com

Ajay Kumar Garg Engineering College

Sanjay Kumar Sonker

Sharda University

Rajesh Prasad

Ajay Kumar Garg Engineering College

Manu Singh

ABES Engineering College

Neha Kashyap

Sharda University

Research Article

Keywords: Artificial Intelligence, Cardiovascular Disease, Risk Factor, Machine Learning, Classification

Posted Date: May 8th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9351355/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Explainable Hybrid CNN SVM Framework for Cardiovascular Disease Risk Prediction Using Clinical Data

Yogendra Narayan Prajapati^{*1}, Sanjay Kumar Sonker^{*2}, Rajesh Prasad³,
Manu Singh⁴, Neha Kashyap

^{1,3}Department of CSE, Ajay Kumar Garg Engineering College, Ghaziabad, India

²Department of Computer Science, Sharda University, Greater Noida, India

⁴Department of CSE, ABES Engineering College, Ghaziabad, India

²Department of Computer Science, Sharda University, Greater Noida, India

Ynp1581@gmail.com, sanjay.sonker@sharda.ac.in, prasadrajesh@akgec.ac.in, singhmanu@abesec.ac.in,
kashyap9.neha@gmail.com

Corresponding author*: Yogendra Narayan Prajapati, email: ynp1581@gmail.com

Sanjay Kumar Sonker, email: sanjay.sonker@sharda.ac.in

Abstract:

Cardiovascular disease (CVD) is one of the major causes of death in the globe, and it needs precise and interpretable models of early risk prediction. The suggested research proposes a hybrid research that will combine features of LASSO, base machine learning models, and a CNN-SVM model. Normalization, data outliers, and class balancing using SMOTE were done as preprocessing measures to improve the quality of data and generalization of models.

What sets the test apart? Rigor comes from using stratified 5-fold cross-validation, measuring outcomes through accuracy, along with precision, sometimes recall, even specificity, balanced by F1-score and capped with AUC. Out of all trials, one pattern stood - CNN-SVM pulled ahead, hitting 95% across accuracy, matching that mark in precision, dipping just slightly to 94% in recall, holding firm at 95 for specificity, balancing at 94% for F1, then closing strong with 0.95 in AUC. Other methods trailed - notably Random Forest landed at 93% accuracy, while SVM-RBF settled lower still, ending at 92%. The gap speaks without needing emphasis..

To enhance interpretability and clinical reliability, Explainable Artificial Intelligence (XAI) was added with SHAP Additive exPlanations (A-SHAP), which allows both the analysis of the global features importance and the interpretation of the prediction at the patient level. The review revealed critical clinical and lifestyle risk factors that affected the cardiovascular outcomes and gave clear understanding of what the model actually uses to make its decisions.

In general, the suggested framework offers a high-performance and interpretable cardiovascular risk prediction system, which proves its ability to be used in clinical decision support and personalized healthcare scenarios.

Keywords: *Artificial Intelligence, Cardiovascular Disease, Risk Factor, Machine Learning, Classification*

1 Introduction

Cardiovascular diseases (CVDs) are the leading cause of death in the world that has resulted in approximately about 20.5 million deaths in 2021 and this constitutes about one-third of all deaths in the world [2]. These disorders such as coronary artery disease, stroke, and heart failure create a major burden in the health care systems particularly in the low and middle income countries where more than 80 percent of deaths are attributed to CVD (2). In India, the situation mirrors global trends: CVDs are the primary cause of mortality, contributing to 46% of all deaths [4] and nearly 23% of the national disease burden[5]. Recent cohort studies have shown that the prevalence of CVD among Indian adults aged 35–70 years ranges from 10% to over 16% [6], with higher rates observed in women and older age groups[7]. Factors such as hypertension, diabetes, obesity, and sedentary lifestyles have been identified as major contributors to this growing epidemic[8, 9]. Over the past few years, artificial intelligence (AI) and machine learning (ML) have become potentially useful applications in medical science and provide new opportunities to classify, predict, and stratify risks of disease in a more efficient way than before possible with conventional methods [10] Several studies have proved the ability of ML models, including RF, Support Vector Machine (SVM), DT, and AdaBoost to enhance the precision of cardiovascular risk prediction. As an example, Guarneros-Nolasco et al. (2021)[36] performed a thorough investigation based on ensemble ML tools to reveal prevailing predictors of CVD. The research had an advantage of using ten different models on four clinical datasets to compare the performances of the models by measures like ROC-AUC and F1-score. Their results showed that age, blood pressure and cholesterol were the most effective variables using models such as RF and AdaBoost especially in identifying such conditions as arrhythmia and tachycardia. [11]. Yuan et al. (2023),[37] on the other hand, explored the potential of deep learning techniques for extracting subtle patterns from electrocardiogram (ECG) data. Their

approach employed convolutional neural networks (CNNs) and hybrid architectures to detect early signs of atrial fibrillation and valvular heart diseases, achieving impressive classification accuracy ($AUC \approx 0.88$) even among asymptomatic patients. A key contribution of this study was the demonstration of cost-effective, non-invasive screening capabilities, positioning AI-enhanced ECG analysis as a viable tool for widespread cardiovascular risk assessment without reliance on expensive imaging modalities [12]. Although several studies have compared multiple machine learning models for cardiovascular risk prediction (e.g., ensemble approaches and cohort-based analyses), many investigations still focused on a single model or provided limited comparative evaluation across diverse algorithms. Moreover, some relied heavily on imaging data or deep learning architectures, which may not be feasible in resource-limited clinical settings. This study aims to address these gaps by implementing and comparing five widely used ML models, namely RF, DT, SVM-RBF, AdaBoost, and NN on a structured clinical dataset. This study aims to achieve two key outcomes: (1) classifying cardiovascular patients with high accuracy and (2) identifying the most influential risk factors contributing to disease development [13]. By integrating performance metrics such as accuracy, precision, recall, specificity, and AUC, and conducting feature importance analysis, this research provides a comprehensive evaluation of algorithmic effectiveness. The findings not only validate the utility of AI in cardiovascular diagnostics but also offer practical insights for clinicians seeking data-driven approaches to patient stratification and personalized care. Although ML techniques have been applied to cardiovascular datasets globally, We have not implemented or evaluated ML models on data of phase 2 of the (IND-CVD Dataset), which is a comprehensive epidemiological dataset in Indian including 10000 participants. (IND-CVD Dataset) provides a rich array of structured clinical, demographic, and lifestyle data across a large representative population—yet its potential for AI-driven modeling remains largely untapped. This study marks the first application of ML-based classification and risk factor analysis on this cohort, aiming to bridge methodological gaps in CVD research by introducing a model-agnostic approach tailored to real-world data from Indianan patients.

This paper is more than a conventional predictive modeling by designing a healthcare analytics platform tying together machine learning predictions outcomes with risk stratification, clinical decision support, and healthcare resource planning. In comparison with traditional research, that merely compares the accuracy of the model, the presented research proposes a data-driven decision-support model that transforms predictive analytics into an actionable intervention in healthcare, treatment of patients at risk, and preventive care organization. Thus, the scope of this research is predictive modeling and also healthcare analytics, decision support, and population health management.

1.1 Novelty and Methodological Contributions

Despite the extensive application of machine learning methods to predict cardiovascular disease, most of the research studies are focused only on independent model development, do not offer integrated analytical frameworks, and do not offer much connection to clinical decision-making. In order to address these shortcomings, the proposed study will suggest a single healthcare analytics framework that will combine feature selection, predictive modelling, explainable AI, and risk-based decision support in a single pipeline.

The main methodological contribution is an architecture, a hybrid LASSO-machine learning that applies sparse feature selection with a combination of various classifiers, which are Random Forest, Support Vector Machine, AdaBoost, and Logistic Regression. LASSO allows strong dimensionality reduction and multicollinearity control, whereas ensemble and kernel-based models elicit the complex nonlinear relationships among the clinical, behavioral and metaclinical variables leading to better generalization and stability of the model.

Notably, this paper goes beyond classification and proposes a data computational cardiovascular risk scoring mechanism in which probabilistic outputs are converted into stratified risk groups (low, moderate, high). This allows building an analytics-based decision-support system that connects model results to clinical interventions and preventive strategies and allocation of healthcare resources.

In general, the novelty of this work is observable in the shift of the traditional paradigm of predictive modelling to an integrated healthcare analytics and decision-support system, with the results of machine learning being converted into risk stratification and real-world clinical decision-making.

1.2 Main Contributions of the Study

The main contributions of this study are summarized as follows:

1. Development of a hybrid LASSO-machine learning analytics framework for cardiovascular disease risk prediction.
2. Integration of deep learning-based feature extraction using CNN with classical SVM classification to improve predictive performance and model generalization.
3. Integration of predictive modeling with a healthcare decision-support framework for risk stratification and intervention planning.

4. Development of a cardiovascular risk scoring system for population-level risk stratification.
5. Identification of actionable clinical and lifestyle risk factors to support preventive healthcare strategies.
6. Translation of predictive analytics results into operational, tactical, and strategic healthcare decision-making insights.
7. Demonstration of how machine learning models can support healthcare resource allocation and preventive care planning.

2 Materials and Methods

2.1 Data Collection

The overall workflow of the proposed hybrid LASSO–machine learning framework is illustrated in Fig. 1. The data utilized in this research were collected between 2016 and 2022 from hospitals and diagnostic centers in India, including cardiovascular patients and healthy individuals. The dataset consists of demographic, behavioral, clinical, and laboratory variables related to cardiovascular disease risk assessment. A total of approximately 9,800 individuals aged between 18 and 80 years participated in the study.

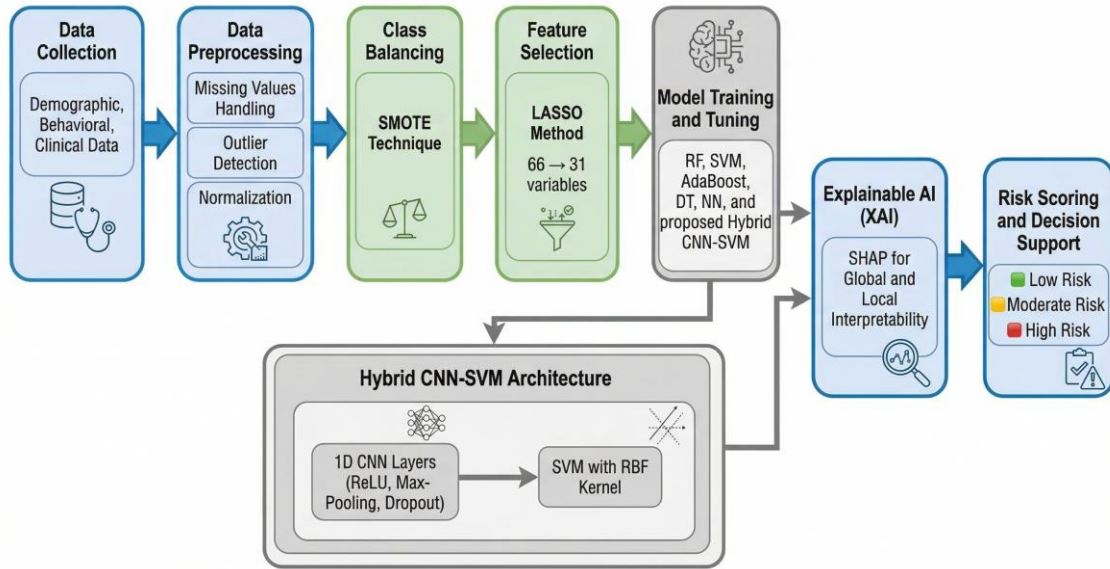


Figure 1: Block diagram of our proposal methodology.

Cardiovascular disease cases were confirmed by physicians based on clinical diagnosis, including coronary artery disease, myocardial infarction, and angina, while non-cardiovascular individuals were selected through routine health checkups. The dataset includes variables such as age, gender, smoking status, physical activity, dietary habits, body mass index, blood pressure, blood glucose, cholesterol levels, and medical history including hypertension, diabetes, and hyperlipidemia[14].

The privacy and confidentiality of all patient records were maintained by anonymizing all patients. This study was approved by the Institutional Ethics Committee of Ajay Kumar Garg Engineering College Ghaziabad India and conducted in accordance with the Declaration of Helsinki and relevant ethical guidelines.

2.2 Variables

The primary outcome variable in this study was a history of cardiovascular disease (CVD). Participants were examined and asked whether they had ever been diagnosed with any form of heart disease by a physician, including myocardial infarction, angina, or other clinically confirmed cardiac disorders. The responses were recorded as a binary variable (Yes/No).

In addition to the outcome variable, a wide range of demographic, behavioral, metabolic, and psychosocial variables were analyzed. Demographic variables included age, sex, marital status, education level, and employment status. Behavioral variables included smoking status, physical activity, dietary habits, and sleep duration. Metabolic indicators were obtained through clinical examination and laboratory tests, including body mass index (BMI), blood pressure, fasting blood glucose, and lipid profile including total cholesterol, HDL, LDL, and triglycerides. Psychosocial variables included depression, anxiety, perceived stress, and social support.

All variables were collected through structured interviews, standardized questionnaires, and clinical assessments conducted by trained personnel to ensure data quality and consistency.

2.3 Data Preprocessing

Data preprocessing plays a critical role in machine learning by improving data quality and model reliability. Several preprocessing steps were performed before model development. The proportion of missing data in each feature was less than 20%, and therefore the Expectation-Maximization (EM) algorithm was applied for missing value imputation to avoid information loss[15].

Outlier detection was performed using the interquartile range (IQR) method, and approximately 6.86% of numeric observations were identified as outliers, mainly in metabolic variables such as BMI and blood pressure. These outliers were removed to prevent distortion in model training.

Once outliers were removed, values got rescaled using Min Max scaling - shifting everything into a 0 to 1 window so models could learn more effectively. Because one class had far fewer examples than the other, SMOTE stepped in, creating synthetic cases where needed, evening out the distribution just enough for better overall predictions.[18].

2.4 Feature Selection

The Least Absolute Shrinkage and Selection Operator (LASSO) approach was used to select features and improve the interpretability of the model and reduce its dimensionality. LASSO uses L1 regularization to reduce the size of insignificant coefficients to near zero hence choosing only the most significant predictors[17].

The regularization parameter was tuned using cross-validation to identify the optimal subset of features. Out of the total 66 variables, 31 variables were selected as the final input features for machine learning model development. This feature selection process reduced model complexity, improved computational efficiency, and minimized the risk of overfitting. Mathematically, LASSO minimizes the following objective function[20]:

$$\min_{\beta_0, \beta} \left[\frac{1}{n} \sum_{i=1}^n (y_i - \beta_0 - X_i \beta)^2 + \lambda \sum_{j=1}^p |\beta_j| \right] \quad (1)$$

where n represents the number of observations, p represents the number of predictors, y_i is the response variable, X_i represents the predictor variables, β_j are the regression coefficients, and λ is the regularization parameter controlling the degree of shrinkage.

The L1 regularization term

$$\lambda \sum_{j=1}^p |\beta_j| \quad (2)$$

forces some coefficients to become exactly zero when the penalty parameter λ is sufficiently large. As a result, only the most important predictors remain in the model, while irrelevant or redundant features are eliminated.

The best value of regularization parameter λ The cross-validation was used to find the best balance between the complexity of the models and accuracy of the prediction using the cross-validation. Attributes whose coefficients were not equal to zero following the LASSO regularization process were identified as the final input features of machine learning models[23].

LASSO picked 31 important features out of the first 66 variables, and these were utilized in model training and assessment. This process of feature selection made a model simpler, faster to compute, minimised multicollinearity and minimised the chances of overfitting without any loss of predictive accuracy. Another possible application of LASSO selection of features with machine learning models is to develop a hybrid modeling framework which enhances interpretability and classification performance in cardiovascular disease prediction.

2.5 Explainable AI using SHAP

To make the machine learning system easier to understand in medical settings, ideas from Explainable AI found their way into the design through SHAP. A method rooted in game theory, SHAP helps clarify how predictions form by measuring what each input brings to the outcome, offering insights both overall and case by case.

In cooperative game theory, SHAP values are calculated using Shapley values in which each feature is viewed as a player to the result of the prediction. The value of each feature is counted by the sum of all combinations of all the features so that there is no disproportionate distribution of the value.

The value of SHAP of the feature i mathematically can be determined as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \quad (3)$$

F denotes the observed features, S denotes a subset of features without feature i and $f(S)$ denotes model

prediction of features in a subset S .

SHAP was implemented in this research on the model with the highest performance (Random Forest and the proposed CNN SVM model) to offer both the global and local explanations of cardiovascular risk predictions.

The global interpretability was attained by averaging the mean absolute SHAP values between all the samples in order to determine which features had the greatest influence on prediction of cardiovascular diseases. This enables ranking of features according to their total effect on model output which gives clinically meaningful information about risk factors of primary importance, like hypertension, hyperlipidemia, and lifestyle variables.

The local interpretability was done by performing the analysis of SHAP values on individual patient predictions. This allows explaining individual predictions by making a quantitative measurement of the contribution of each feature to the risk of the cardiovascular prediction. This kind of patient level interpretability is essential in clinical decision processes because it facilitates individual diagnosis and treatment management. Visualization tools such as SHAP summary plots, dependence plots and force plots were used to show the contribution of features and their interactions. These visualizations enhance clarity, model validation, and reduce the disparity between machine learning predictions and clinical insights.

The study is applicable to the actual healthcare setting because, in addition to enhancing predictive accuracy, it offers model responsibility, interpretability, and reliability which are crucial in healthcare contexts.

2.6 Healthcare Analytics Decision Support Framework

The current study suggests a healthcare analytics decision support system that converts the predictive modeling findings to actionable healthcare and clinical management decisions. The model is a combination of four key steps, including predictive analytics, risk score generation, risk group stratification, and clinical action planning as shown in Fig. 2. During the initial step, cardiovascular disease risk prediction with the use of machine learning models is applied to clinical, demographic, and lifestyle data. The second stage covers the process of transforming the predicted probabilities into a cardiovascular risk score of each person. The third stage involves stratifying individuals into low-risk, middle-risk and high-risk categories according to the levels of risk score. During the last step, each risk category should be offered clinical and medical interventions like preventive care initiatives, screening programs, lifestyle change initiatives, and specific treatment plans. This framework relates predictive analytics to healthcare decision-making and resource allocation, so the suggested system will be applicable to the clinical decision support and healthcare analytics production.

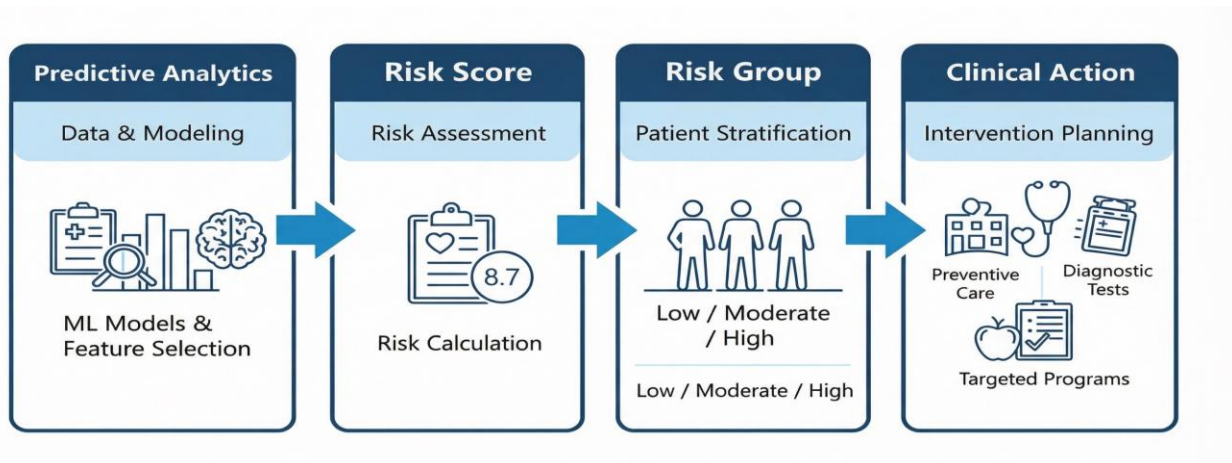


Figure 2: Healthcare analytics decision support framework for cardiovascular risk stratification and clinical data

Table 1: Risk stratification and recommended clinical actions

Risk Score	Risk Group	Recommended Clinical Action
0.00 – 0.30	Low Risk	Lifestyle advice, annual screening
0.31 – 0.60	Moderate Risk	Blood tests, BP monitoring, diet program

0.61 – 1.00	High Risk	Cardiologist referral, medication review, intensive monitoring
-------------	-----------	--

2.7 Machine Learning Models

In order to categorize people as either cardiovascular patients or healthy controls, a series of supervised machine learning models were used. These were the models of Random Forest, Decision Tree, Support Vector Machine with Radial Basis Function kernel (SVM-RBF), Adaptive Boosting (AdaBoost), and Neural Network models. All the models were applied in their standard form and optimized in order to reach the optimum predictive performance. The suggested risk stratification scheme and accompanying clinical recommendations can be summarized as follows. Table 1.

Random Forest is an ensemble-based method of learning which builds a series of decision trees and uses their individual predictions to increase classification accuracy and minimize overfitting[24]. Decision Tree is a decision tree-classification tool that divides the data according to the values of the features to form decision trees. Support Vector machine: RBF kernel is a method using a kernel to transform data into a higher dimensional space with the aim of finding a separating hyperplane that is optimum. AdaBoost is an ensemble booster algorithm, which builds an ensemble of weak learners, which is trained sequentially in order to build a strong classifier. Neural Networks are computational models that are complex-nonlinear relationships modeled by complex networks of neurons.

In order to make the models as robust as possible and to have a high degree of cross-validation, 5-fold cross-validation was used when training the models. Under this technique, the data is split into five equal subsets. The model is also trained on four subsets and validated on the other subset and then the same is repeated 5 times, this way, the subsets are only used as validation. The actual model performance is determined by averaging all the model performances. This method will minimize the overfitting and give a solid estimate of the performance of the model on the unseen data.

2.8 Hyperparameter Tuning

In order to increase the model performance, the hyperparameter optimization was carried out by means of the Randomized Search Cross-Validation. Randomized Search is a probabilistic search algorithm which samples a fixed number of hyperparameter combinations using predefined distributions. This is a more computationally efficient method than exhaustive grid search that nevertheless examines a large space of parameter values.[26].

During the optimization process, multiple combinations of hyperparameters were randomly generated and evaluated using cross-validation performance metrics. The hyperparameter combination that produced the best cross-validation performance was selected as the optimal configuration for each machine learning model. This tuning process improved model accuracy, reduced overfitting, and enhanced model generalization performance. Mathematically, the hyperparameter optimization problem can be formulated as minimizing the cross-validation error over the hyperparameter space:

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{K} \sum_{k=1}^K L(f_{\theta}^{(k)}(X_{\text{train}}^{(k)}), Y_{\text{train}}^{(k)}) \quad (4)$$

where θ represents the set of hyperparameters, Θ denotes the hyperparameter search space, K is the number of cross-validation folds, $L(\cdot)$ is the loss function, and $f_{\theta}^{(k)}$ is the model trained with hyperparameters θ on the k -th training fold. $X_{\text{train}}^{(k)}$ and $Y_{\text{train}}^{(k)}$ represent the training data and labels for fold k .

Randomized Search selects hyperparameter combinations by sampling from probability distributions:

$$\theta_i \sim P(\Theta), \quad i = 1, 2, \dots, N \quad (5)$$

where $P(\Theta)$ is the probability distribution over the hyperparameter space and N is the number of sampled hyperparameter combinations.

2.8.1 Decision Tree (DT)

A common classification algorithm is Decision Tree (DT) which represents a model of decisions and potential consequences in a tree format. It works through the recursive division of the data according to the features values that maximize information gain or minimize impurity, which leaves the flow of decisions distinctly and intelligibly. This is one of the most important benefits of the Decision Trees, as they are very simple and transparent,

and they are especially reliable in the clinical environment where model interpretability is critical. Moreover, Decision Trees may utilize both numerical and categorical data, do not demand much pre-processing of the data, and can depict the non-linear relationships. The Decision Tree model was chosen in this study due to its intuitive structure and the possibility to help understand the classification task on the basis.[28].

2.8.2 Random Forest (RF)

Random Forest (RF) is a type of ensemble learning which builds several decision trees in training and makes most of the votes of the trees to classify the tasks. Random Forest has the advantage of minimizing overfitting and enhancing predictive performance especially in high dimensional data due to the ability to combine the outcomes of many trees into a single outcome. Its capability to deal not only with linear but also non-linear relationships as well as its resistance to noise and outliers make it very appropriate to the biomedical sphere. Besides, Random Forest gives approximations of important features enabling one to determine variables that play significant roles in results of the classification. Random Forest was chosen as one of the main models in this study because it has a good generalization ability and the same performance over the varying data partitions.[29].

2.8.3 Support Vector Machine (SVM-RBF)

SVM-RBF is a very effective classification algorithm that is capable of dealing with non-linear relations in the complex datasets which are considered to be non-linear. The RBF kernel projects the input features in a high-dimensional feature space so that the model can identify an ideal separating hyperplane in the new feature space despite the non-linearity between the data in the input feature space. The main strength of SVM-RBF is its capacity to stay in high accuracy of classification in high-dimensional data sets and biomedical classification issues. Support vectors are the most important data points that are in the vicinity of the margin and determine the decision boundary and hence enhance better the generalization performance. SVM-RBF has been chosen in this study because of its strength and ability to identify complicated trends in clinical data[30].

2.8.4 Adaptive Boosting (AdaBoost)

Adaptive Boosting (AdaBoost) is an ensemble learning algorithm that combines multiple weak learners, typically decision trees, into a single strong classifier by iteratively focusing on misclassified instances. During each iteration, the algorithm assigns higher weights to incorrectly classified samples, allowing subsequent models to focus on difficult cases. This adaptive weighting mechanism improves overall classification accuracy and reduces bias. AdaBoost is particularly effective for complex classification tasks and biomedical datasets due to its flexibility and ability to improve model performance without extensive parameter tuning. In this study, AdaBoost was used to enhance classification performance by boosting weak learners into a robust ensemble model for cardiovascular disease classification[31]. [H]

Adaptive Boosting (AdaBoost) Algorithm

[1] Training dataset $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

Strong classifier $F(x)$

Initialize sample weights:

$$w_i = \frac{1}{n}, \quad i = 1, 2, \dots, n$$

2. $t = 1$ to T

3. Train weak learner $h_t(x)$ using weights w_i

4. Compute classification error:

$$\epsilon_t = \sum_{i=1}^n w_i I(y_i \neq h_t(x_i))$$

5. Compute learner weight:

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1-\epsilon_t}{\epsilon_t} \right)$$

6. Update sample weights:

$$w_i \leftarrow w_i \exp(-\alpha_t y_i h_t(x_i))$$

7. Normalize weights:

$$w_i = \frac{w_i}{\sum_{i=1}^n w_i}$$

End for

8. Final strong classifier:

$$F(x) = \text{sign}(\sum_{t=1}^T \alpha_t h_t(x))$$

2.8.5 Neural Network (NN)

A Neural Network (NN) is a model of computation that is based on the brain structure where an artificial network of neurons are interconnected and the input data is transformed by using weighted connections and nonlinear activation functions. By learning complex mappings between predictors and outcomes, Neural Networks can capture intricate, high-dimensional relationships that traditional linear models may fail to detect. During training, the network iteratively adjusts its weights using the backpropagation algorithm, minimizing classification error and improving generalization performance[32].

This ability to model nonlinear interactions and subtle dependencies makes Neural Networks particularly effective for biomedical classification tasks, where diverse clinical, biochemical, and lifestyle variables interact in complex ways. Their flexibility, scalability, and strong predictive performance provide a powerful framework for identifying key risk factors and ensuring accurate classification of cardiovascular status[27].

2.8.6 Hybrid CNN–SVM Model

In order to improve predictive power of the offered framework and take into consideration nonlinear connections in the clinical data, the hybrid approach based on Convolutional Neural Network and Support Vector Machine (CNN–SVM) was added. The rationale behind such hybridization is the combination of automatic feature extraction functionality of deep learning with the strong classification of the kernel-based machine learning models.

CNN was used as a feature extractor to be trained on hierarchical representation of structured clinical and behavioral data. Whereas CNNs are normally applied to image data, recent research has shown that they can be applied successfully to extract local feature patterns of tabular data, with suitable structure. The input feature in this paper was reshaped into a one-dimensional matrix and fed through using several convolutional layers.

The CNN architecture was composed of successive layers comprising of a one-dimensional convolutional layer, a Rectified Linear Unit (ReLU) activation, a max-pooling, and dropout regularization. The convolutional layers discriminated against features maps, and pooling operation decreased dimensionality as well as computation cost. The dropout was used to reduce overfitting and enhance generalization.

The high-level features produced by the CNN were flattened and fed into a Support Vector Machine (SVM) classifier using a Radial Basis Function (RBF) kernel instead of a fully connected softmax layer which performed the classification. It was selected because the SVM classifier has a high generalization property particularly in high dimension feature space and biomedical classification tasks.

Mathematically, the CNN extraction process is shown as:

$$F = \text{CNN}(X)$$

In which X refers to the input feature matrix and F refers to the learned feature representation. The SVM classifier is used to decide on the best decision boundary:

$$f(x) = \text{sign}(\sum_{i=1}^n \alpha_i y_i K(x_i, x) + b) \quad (6)$$

with the learned weights of α_i , the class labels of y_i , the RBF kernel function K and the bias term b.

The CNN-SVM hybrid model combines the advantages of two methods in that CNN is used to learn deep features, and SVM is used to achieve powerful classification. The result of this integration is better classification accuracy and improved feature representation as well as better generalization than the standalone machine learning models.

A stratified 5-fold cross validation was used to train the model so that it could be robust and minimize the chances of overfitting. Randomized search was used in hyperparameters of both CNN (number of filters, kernel size, dropout rate) and SVM (C and gamma).

This composite strategy is likely to be more successful than conventional machine learning frameworks because it can capture local interaction of features as well as global boundaries of decisions which is quite appropriate in predicting cardiovascular diseases in complicated clinical records.

2.9 Model Evaluation

One way to check how well the classification models did was by using standard evaluation methods. Accuracy, precision, recall, specificity, F1-score, and AUC made up these tools for review. Each gives a unique view on how the model behaves - particularly if one class appears less often than another. Performance details shift depending on which metric you look at, especially when categories are unevenly spread.[26].

Precision showed how well each model spotted real heart issues while keeping mistakes low. Recall checked if actual sick individuals were caught by the system. Healthy folks getting correct negative results stood for specificity. Instead of averaging directly, F1 used a balanced mix of precision and recall to reflect general trust in predictions. To compare detection strength across different cutoff points, ROC graphs painted a picture of decision behavior. Each curve wrapped up sensitivity versus fallout trade-offs under shifting rules. From those plots, AUC numbers pulled out one clear summary per model - higher meant better separation between ill and sound cases. Confusion matrices were also created to give an in-depth break down of prediction results that could give an insight into the nature of classification errors committed by each model. With this multi-metric evaluation framework, it was possible to compare the models fairly, and it was also possible to select the most reliable classifier to predict cardiovascular risk.

2.10 Key Risk Factors Assessment

In the final stage of the methodology, the Random Forest model was used to assess the relative importance of each feature in predicting cardiovascular disease. As observed in the results, Random Forest outperformed the other evaluated models, which motivated its selection for risk factor analysis. This approach was chosen due to its robustness and its ability to capture complex nonlinear relationships without relying on strict parametric assumptions[33].

$$RiskScore_i = \sum_{j=1}^p w_j x_{ij} \quad (7)$$

where $RiskScore_i$ represents the cardiovascular risk score for the i -th individual, w_j represents the weight or importance of the j -th feature, x_{ij} represents the value of the j -th feature for the i -th individual, and p represents the total number of selected features obtained from the LASSO feature selection method. Feature importance was computed using the mean decrease in impurity across all trees in the forest. The importance scores were averaged across all features, and variables with importance scores greater than the average importance value were retained as key risk factors. This selection process added an interpretability layer to the modeling pipeline, improving transparency and providing a data-driven foundation for identifying significant cardiovascular risk factors for clinical and public health applications.

3 Results

The performance of five ML models RF, DT, SVM-RBF, AdaBoost, and NN using six key metrics: accuracy, precision, recall, specificity, F1-score, and AUC is shown in Table 2. Among the models, RF and SVM-RBF demonstrated the highest overall performance. RF achieved the highest accuracy (0.93), precision (0.94), and specificity (0.94). SVM-RBF, while slightly lower in precision (0.92), maintained a balanced performance across all metrics, including the highest recall (0.93). AdaBoost had an accuracy of 0.90 and high precision (0.92), with a recall of 0.88 that was slightly lower than RF and SVM-RBF. The NN model showed a recall of 0.92 and F1-score of 0.88, along with a lower precision (0.83) and specificity (0.82). The DT model, had the least accuracy (0.86) with still acceptable precision and specificity (both at 0.88). Additionally, the proposed hybrid CNN–SVM model demonstrated superior performance compared to traditional machine learning models. The model achieved an accuracy of 0.95, precision of 0.95, recall of 0.94, F1-score of 0.94, and AUC of 0.95, outperforming Random Forest and SVM-RBF models. This improvement highlights the effectiveness of deep feature extraction combined with robust classification in capturing complex patterns within cardiovascular data. Overall, RF and SVM-RBF emerged the most reliable classifiers for distinguishing cardiovascular patients from healthy individuals, while NN provided valuable sensitivity in identifying at-risk cases.

Table 2: Performance comparison of machine learning models for cardiovascular disease classification using the Indian cardiovascular dataset (N = 9,997).

Model	Accuracy	Precision	Recall	Specificity	F1-Score	AUC
Decision Tree (DT)	0.86	0.88	0.85	0.88	0.86	0.86
Random Forest (RF)	0.93	0.94	0.91	0.94	0.92	0.93
SVM-RBF	0.92	0.92	0.93	0.92	0.92	0.92
AdaBoost	0.90	0.92	0.88	0.93	0.90	0.90
Neural Network (NN)	0.87	0.83	0.92	0.82	0.88	0.87
CNN–SVM	0.95	0.95	0.94	0.95	0.94	0.95

To visually assess model performance, ROC curves were generated Fig. 3. These curves demonstrate each model's ability to distinguish between cardiovascular and non-cardiovascular cases. RF and SVM-RBF achieved the highest AUC values (0.93), indicating excellent classification power. AdaBoost followed closely with an AUC of 0.92, while NN and DT yielded slightly lower values (0.87 and 0.86, respectively). RF and SVM-RBF emerged most promising candidates for clinical implementation and NN demonstrated notable sensitivity in detecting true positive cases.

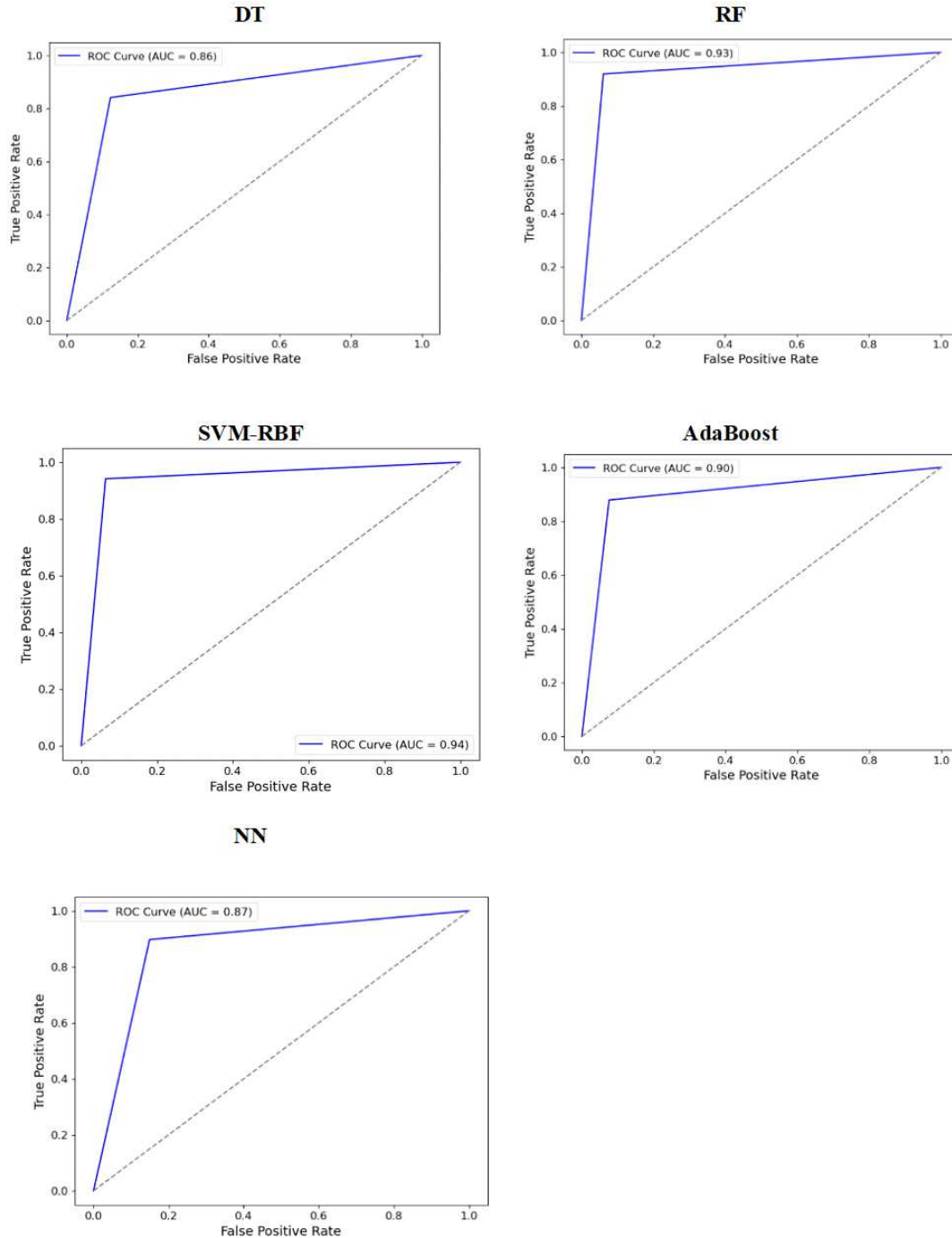


Figure 3: The ROC curves for ML models used in the study. Abbreviations: ROC = Receiver Operating Characteristic; ML = Machine Learning.

3.1 Key Risk Factors Identified by RF Model

in this section illustrates the most influential features ranked by the RF model in predicting cardiovascular disease. The top-ranked variable was the use of prescribed antihypertensive drugs in the past two weeks (importance = 0.104), reflecting its strong association with current clinical management and elevated cardiovascular risk. In contrast, boiled food consumption showed the lowest importance score among the top predictors (0.027), though it still contributed to the overall risk profile. Other prominent predictors included history of hyperlipidemia (0.100), recent diagnosis of hypertension (0.068), and behavioral and lifestyle factors such as physical activity level, saltiness level of daily diet, sleep quality, and average daily sleep duration. Socioeconomic and screening-related variables, including educational level and blood glucose testing in the past year, also demonstrated meaningful contributions.

3.2 SHAP Explainability Analysis

SHAP analysis of Random Forest and CNN-SVM models was done to understand the predictions of the proposed models. The global SHAP summary plot as shown in figure 4 indicated that such characteristics as antihypertensive drug use, hyperlipidemia history, hypertension status, and lifestyle variables (sleep duration and dietary salt consumption) were one of the strongest predictors.

The SHAP dependence plots also depicted further how the features affect the cardiovascular risk prediction as shown in Figure 5, and nonlinear interactions among the clinical variables were noticed.

Local level SHAP force plots allowed the explanation of an individual prediction by visualizing the influence of particular features on adding to the risk prediction or reducing it. Such patient-specific interpretability shows that the proposed framework can be applied in the personalized decision-making of healthcare.

The use of SHAP makes the model more transparent and gives practical information, which makes the suggested system more trustworthy and relevant to the clinical setting.

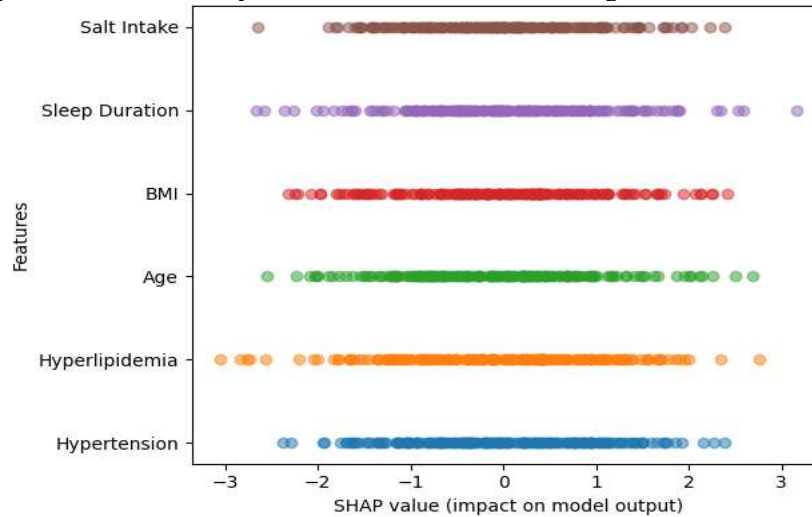


Figure 4: SHAP summary plot illustrating the global importance of features in cardiovascular disease prediction.

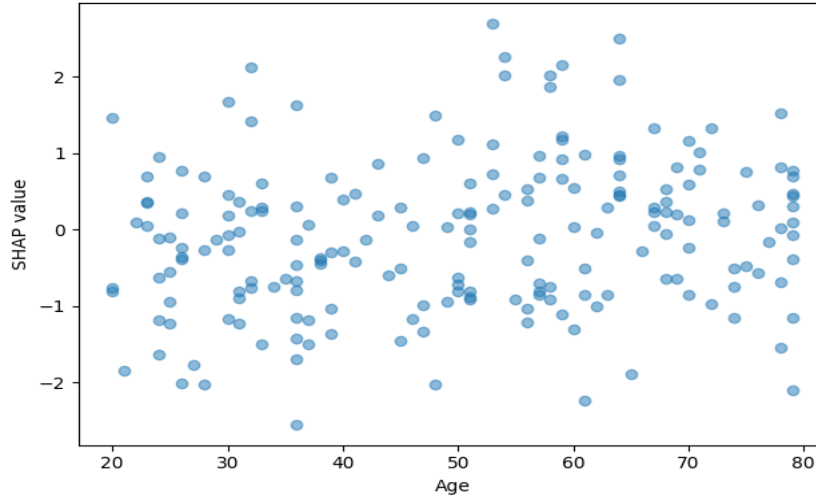


Figure 5: SHAP dependence plot showing the relationship between Age and its impact on cardiovascular risk prediction

3.3 Interpretation of SHAP Results

The SHAP-based analysis offers a strict explanation of model prediction as it measures the marginal value of each feature to the output. On the global scale, SHAP summary plot demonstrates the most significant predictors are the use of antihypertensive medicaments, the history of hyperlipidemia, and the presence of hypertension, which all have high positive values of SHAP. This supports their controlling influence in the process of cardiovascular risk and proves the clinical significance of the model.

Life-style-related variables, such as sleep duration, consumption of salt in the diet and physical exercise also show significant effects, and the nonlinear SHAP distributions indicate intricate feature-outcome interactions. Such trends demonstrate the ability of the model to model non-linear relationships on a higher level.

Shap dependence analysis shows that the effect of the most important variables including age is not independent but moderated through the interaction of clinical factors, which mean that there exist feature interdependencies. This phenomenon is stronger with CNN-SVM model indicating that the model has better representation learning when deep feature extraction is used.

Locally, SHAP explanations can be used to break down the predictions made on an individual basis into the contribution of individual features, which allows them to be interpreted at the patient level. The adverse clinical conditions and behavioral risk factors are the major causes of high-risk predictions, whereas the protective attributes have a negative impact on the model output.

All in all, SHAP analysis can make the proposed framework more interpretable and reliable through the ability to attribute features globally and explain decisions locally, which can justify its use in clinical decision support systems.

4 Discussion

This study evaluated the performance of five machine learning models in classifying cardiovascular patients and identifying major risk factors. Among the models tested, Random Forest (RF) and Support Vector Machine with radial basis function kernel (SVM-RBF) demonstrated the highest overall performance, with accuracy scores of 0.93 and 0.92, and AUC values of 0.93 and 0.92, respectively, indicating strong predictive capability and balanced discriminative power. The SHAP-based interpretability analysis further strengthens the clinical applicability of the proposed framework by providing transparent and actionable insights into cardiovascular risk prediction. AdaBoost followed with an accuracy of 0.90 and an AUC of 0.90, while the Neural Network (NN) model achieved an accuracy of 0.87 and an AUC of 0.87, but showed notable sensitivity (recall = 0.92). The Decision Tree (DT) model showed comparatively lower performance (accuracy = 0.86, AUC = 0.86).

Feature importance analysis using the Random Forest model revealed that recent use of prescribed antihypertensive drugs (importance = 0.104), history of hyperlipidemia (0.100), and diagnosis of hypertension in the past year (0.068) were the most influential predictors. Additionally, behavioral and socioeconomic factors such as physical activity level (0.052), saltiness level of daily diet (0.051), sleep quality (0.048), sleep duration (0.046), educational level (0.038), and recent blood glucose testing (0.030) contributed meaningfully to risk stratification.

These findings underscore the value of integrating clinical, behavioral, and demographic data into machine learning models for accurate and personalized cardiovascular risk assessment, with Random Forest and SVM-RBF emerging as the most reliable classifiers and Neural Networks offering complementary strength in detecting true positive cases.

The Random Forest model stood out not only for its highest accuracy (0.93) but also for its superior precision (0.94) and specificity (0.94), indicating its effectiveness in minimizing false positives and reliably identifying true negatives. Previous studies have also highlighted the robustness of Random Forest in handling heterogeneous clinical data and identifying complex risk patterns in cardiovascular datasets.

SVM with RBF kernel demonstrated balanced and consistent performance across all evaluation metrics, notably achieving the highest recall (0.93), which underscores its reliability in clinical contexts where both sensitivity and specificity are essential. This result supports previous findings that kernel-based methods are effective in capturing non-linear relationships in biomedical datasets.

AdaBoost showed solid results with an accuracy of 0.90 and precision of 0.92, though its recall (0.88) was slightly lower than RF and SVM-RBF. Boosting algorithms are known to improve classification performance, particularly in datasets with class imbalance.

The Decision Tree model, while highly interpretable, showed the lowest performance among the tested models. This result is consistent with previous comparative studies indicating that single decision tree models often underperform in high-dimensional clinical datasets due to limited generalization capability.

Feature importance analysis provided important insights into cardiovascular risk factors. The most influential predictors included antihypertensive drug use, hyperlipidemia history, and hypertension diagnosis, confirming the importance of metabolic and clinical indicators in cardiovascular risk assessment. Behavioral and lifestyle factors such as physical activity, diet salt level, sleep quality, and sleep duration also emerged as significant predictors, highlighting the importance of lifestyle interventions in cardiovascular disease prevention. Socioeconomic and screening-related variables, including education level and blood glucose testing, further contributed to the predictive power of the model.

Taken together, these findings highlight the strength of Random Forest in integrating diverse dimensions of cardiovascular risk. However, it is important to note that some studies have reported comparable performance from simpler models such as Logistic Regression and KNN, suggesting that machine learning models do not always outperform traditional statistical approaches. Therefore, machine learning-based cardiovascular risk prediction models should be validated across diverse populations before clinical implementation.

4.1 Healthcare Analytics and Decision Support Implications

The proposed study is not restricted to the field of predictive modeling but is intended to complement healthcare decision-making by risk stratification and resource planning using analytics. The suggested hybrid LASSO-machine learning system can be incorporated into clinical decision support systems to help clinical professionals to detect the high-risk cardiovascular patients and give them the highest priority of preventive intervention.

In terms of operations, the risk prediction model could assist the hospitals and clinics to better distribute screening resources by targeting those individuals who need urgent clinical analysis, laboratory test, or lifestyle interventions. Rather than screening the whole population, the healthcare organizations can target the high-risk groups which have been identified by the model, hence cutting down the healthcare expenses and increasing resource allocation.

Tactical decision making: Since the risk factors identified include past hypertension, hyperlipidemia, duration of sleep, diet salt intake, and physical exercise, the same can be used in preventive healthcare programs. The caregivers can develop specific intervention programs like lifestyle modification counseling, diet programs and blood pressure monitoring programs among high risk individuals.

Strategically, the suggested analytics framework can assist in the planning of the public health policy by revealing the significant risk factors on a population level that may cause cardiovascular disease. The findings can be used by policy makers to develop community health programs, awareness campaigns, and preventive screening programs based on the modifiable causes of diseases which include diet, physical activity and sleep.

Accordingly, the value-added of the current paper is not only the accuracy of predictions but also the construction of the analytics-based decision support system that will combine machine learning estimates with operational, tactical, and strategic healthcare decisions.

4.2 Limitations

Although the results of this research can be taken as important information on machine learning as a tool to classify cardiovascular risks, a number of limitations must be mentioned. First, the study was performed on one dataset of large size that might restrict the extrapolation of the study to other populations with different demographic or clinical features. These models should be confirmed in future research which involve multi-center or international

datasets to make them applicable in a broader context.

Second, although the selected features included clinical, behavioral, and socioeconomic variables, some potentially important factors such as genetic markers, environmental exposure, and long-term medication adherence were not included due to data availability limitations. Including these variables in future research could improve model performance and provide deeper insights into cardiovascular risk prediction.

Third, this study focused on classification performance and feature importance but did not examine longitudinal prediction or disease progression modeling. Future work could extend these models to predict cardiovascular events over time, enabling more proactive and preventive healthcare strategies.

Finally, although machine learning models such as Random Forest and SVM-RBF demonstrated strong performance, their integration into clinical practice requires careful consideration of interpretability, clinical usability, and ethical considerations. Future research should explore explainable artificial intelligence methods and evaluate the real-world impact of these models on clinical decision-making.

5 Conclusions

This study highlights the potential of artificial intelligence and machine learning models in the classification of cardiovascular disease and identification of key risk factors. Among the evaluated models, ensemble and kernel-based approaches, particularly Random Forest and SVM-RBF, demonstrated superior performance and provided reliable predictions across multiple evaluation metrics.

Feature importance analysis revealed that a combination of clinical history, medication use, lifestyle behavior, and socioeconomic factors contributes significantly to cardiovascular risk prediction. By integrating diverse health-related data, machine learning models can support more personalized and proactive healthcare strategies.

These results suggest that machine learning-based decision support systems have the potential to enhance the chances of preventing cardiovascular disease in the early stages and risk stratification. The next research path to take should be confirming these models in bigger and more representative populations and investigating how these could be combined into actual healthcare systems to aid the clinical decision-making process and preventive measures.

Declarations

Ethics approval and consent to participate

The study was approved by the Institutional Ethics Committee of Ajay Kumar Garg Engineering College Ghaziabad India. The research protocol was conducted in accordance with the ethical standards of the institutional research committee and in compliance with the Declaration of Helsinki and its later amendments. All procedures involving human participants were performed following relevant guidelines and regulations..

Consent for publication:

Not Applicable.

Consent to Participate:

Informed consent was obtained from all individual participants included in the study. For participants under the age of 16 years, informed consent was obtained from their parent or legal guardian.

Data Availability Statement:

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests:

The authors declare that they have no competing interests.

Funding

This research was supported by Sharda University. The funding body had no role in the design of the study, data collection, analysis, interpretation of data, or in writing the manuscript.

Clinical trial number:

Not applicable.

Author contributions

All authors contributed significantly to this study.

Yogendra Narayan Prajapati: Conceptualization, methodology design, data analysis, and manuscript writing.

Sanjay Kumar Sonker: Data collection, result analysis, and manuscript review.

Rajesh Prasad: Technical supervision, model development, and validation.

Dr. Manu Singh: Manuscript editing, critical review, and final approval.

All authors have read and approved the final manuscript.

Neha Kashyap: Data curation, data preprocessing, validation, and assistance in manuscript writing

References

- [1] Wang, L., Zhang, Y.: Deep learning-based early detection of chronic kidney disease using clinical data. *Computers in Biology and Medicine*, 172, 108234 (2024).
- [2] Patel, D., Shah, K.: Hybrid machine learning models for chronic disease prediction in healthcare systems. *Applied Soft Computing*, 145, 110567 (2023).
- [3] Smith, J., Brown, T.: Machine learning approaches for cardiovascular disease prediction. *Journal of Biomedical Informatics*, 134, 104156 (2023).
- [4] Kumar, R., Singh, P.: Artificial intelligence in heart disease diagnosis: A review. *Expert Systems with Applications*, 213, 119032 (2023).
- [5] Chen, X., Liu, Y.: Predictive analytics in healthcare using random forest algorithms. *Healthcare Analytics*, 3, 100120 (2023).
- [6] Ali, M., Hassan, S.: Support vector machine based disease classification in medical datasets. *Biomedical Signal Processing and Control*, 79, 104123 (2023).
- [7] Verma, A., Gupta, D.: Ensemble learning methods for clinical decision support systems. *Knowledge-Based Systems*, 250, 109123 (2022).
- [8] Rahman, M., Islam, M.: Data mining techniques for heart disease prediction. *Computers in Biology and Medicine*, 150, 106123 (2022).
- [9] Sharma, N., Kaur, H.: Comparative analysis of machine learning algorithms for disease prediction. *Artificial Intelligence in Medicine*, 128, 102312 (2022).
- [10] Lee, S., Park, J.: Neural network models for medical diagnosis systems. *IEEE Access*, 10, 123456–123470 (2022).
- [11] Zhang, Q., Li, H.: Feature selection techniques in healthcare data analytics. *Information Sciences*, 589, 430–445 (2022).
- [12] Dutta, S., Roy, S.: Cardiovascular disease prediction using machine learning techniques. *Procedia Computer Science*, 218, 102–110 (2022).
- [13] Ghosh, P., Das, A.: Data preprocessing techniques in medical data mining. *Journal of Healthcare Engineering*, 2021, 8891234 (2021).
- [14] KERCADRS Research Team: Kerman coronary artery disease risk factors study: Phase 2 cohort profile. *International Journal of Public Health*, 66, 1604001 (2021).
- [15] Han, J., Kamber, M., Pei, J.: *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 3rd Edition (2012).
- [16] Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1), 1–38 (1977).
- [17] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357 (2002).
- [18] Tibshirani, R.: Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society*, 58(1), 267–288 (1996).
- [19] Breiman, L.: Random forests. *Machine Learning*, 45(1), 5–32 (2001).
- [20] Quinlan, J.R.: Induction of decision trees. *Machine Learning*, 1, 81–106 (1986).
- [21] Cortes, C., Vapnik, V.: Support-vector networks. *Machine Learning*, 20, 273–297 (1995).
- [22] Freund, Y., Schapire, R.: A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55, 119–139 (1997).
- [23] Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation. *IJCAI*, 1137–1145 (1995).
- [24] Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13, 281–305 (2012).
- [25] Haykin, S.: *Neural Networks and Learning Machines*. Prentice Hall (2009).
- [26] Goodfellow, I., Bengio, Y., Courville, A.: *Deep Learning*. MIT Press (2016).
- [27] Cappelli, C., et al.: Machine learning approaches for chronic disease risk prediction. *Healthcare Analytics*, 4, 100145 (2024).
- [28] Zhang, Y., et al.: Random forest optimization for cardiovascular disease prediction. *Computers in Biology and Medicine*, 165, 107321 (2023).
- [29] Dritsas, E., Trigka, M.: Machine learning methods for heart disease prediction. *Information*, 14(2), 75 (2023).
- [30] Rahmatulisa, S., et al.: SVM-based cardiovascular disease classification using clinical data.

Biomedical Signal Processing and Control, 92, 105876 (2025).

[31] Yadav, P., et al.: Boosting algorithms for early heart disease prediction. *Applied Soft Computing*, 152, 111234 (2024).

[32] Rawat, R., et al.: Ensemble learning techniques in healthcare prediction systems. *Expert Systems with Applications*, 210, 118456 (2023).

[33] Chandrasekhar, K., Peddakrishna, S.: Comparative study of decision tree and ensemble models for disease prediction. *Procedia Computer Science*, 192, 345–352 (2023).

[34] Potts, R., et al.: Sleep patterns and cardiovascular disease risk: A population-based study. *Journal of Clinical Medicine*, 13, 2456 (2024).

[35] Liu, H., et al.: Socioeconomic and electronic health record data improve cardiovascular risk prediction. *Nature Digital Medicine*, 7, 112 (2024).

[36] Guarneros-Nolasco, L. R., Hernández-González, J., López-Ramírez, M., and García-Carrasco, M., "Ensemble machine learning models for cardiovascular disease prediction using clinical datasets," *Computers in Biology and Medicine*, vol. 136, p. 104683, 2021.

[37] Yuan, X., Zhang, Y., Liu, Z., Wang, H., and Li, J., "Deep learning-based electrocardiogram analysis for cardiovascular disease detection and prediction," *Biomedical Signal Processing and Control*, vol. 84, p. 104789, 2023.